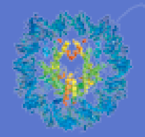


PACIFIC SYMPOSIUM ON
BIOCOMPUTING 2011

The Pacific Symposium on Biocomputing (PSB) 2011 is an international, multidisciplinary conference for the presentation and discussion of current research in the theory and application of computational methods in problems of biological significance. Presentations are rigorously peer reviewed and are published in an archival proceedings volume. PSB 2011 will be held on January 3 – 7, 2011 in Kohala Coast, Hawaii. Tutorials and workshops will be offered prior to the start of the conference.

PSB 2011 will bring together top researchers from the US, Asia Pacific, and around the world to exchange research results and address pertinent issues in all aspects of computational biology. It is a forum for the presentation of work in databases, algorithms, interfaces, visualization, modeling, and other computational methods, as applied to biological problems, with emphasis on applications in data-rich areas of molecular biology.

The PSB has been designed to be responsive to the need for critical mass in sub-disciplines within biocomputing. For that reason, it is the only meeting whose sessions are defined dynamically each year in response to specific proposals. PSB sessions are organized by leaders of research in biocomputing's "hot topics". In this way, the meeting provides an early forum for serious examination of emerging methods and approaches in this rapidly evolving field.

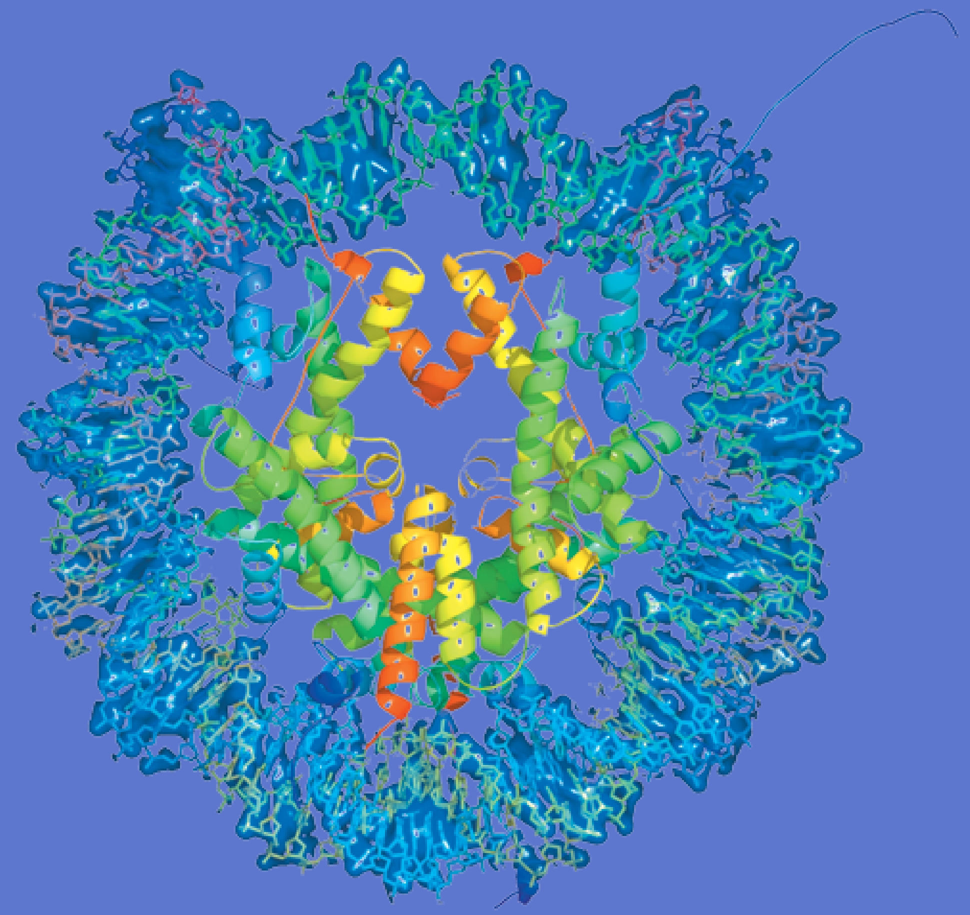
World Scientific
www.worldscientific.com
8021 hc

ISBN-13 978-981-4335-05-8(ebk)
ISBN-10 981-4335-05-3(ebk)
9 789814 335058

R. B. Altman
A. K. Dunker
L. Hunter
T. Murray
T. E. Klein

PACIFIC SYMPOSIUM ON
BIOCOMPUTING 2011

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2011



Edited by

Russ B. Altman, A. Keith Dunker,
Lawrence Hunter, Tiffany Murray & Teri E. Klein

Cover image:
This image depicts a molecular model of the Nucleosome (PDB ID: 1aoi, Luger et al. (1997) Nature 389, 251–260) — The nucleosome is the organising principle behind higher ordered chromatin structure. The histone core of the nucleosome exemplifies the many molecular mechanisms that have evolved to regulate access to the DNA in chromatin.

Image by D. Rey Banatao,
Pacific Symposium on Biocomputing.

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2011

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2011

Kohala Coast, Hawaii, USA
3–7 January 2011

Edited by

Russ B. Altman
Stanford University, USA

A. Keith Dunker
Indiana University, USA

Lawrence Hunter
University of Colorado Health Sciences Center, USA

Tiffany Murray
Stanford University, USA

Teri E. Klein
Stanford University, USA

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2011

2011 marks the 16th Pacific Symposium on Biocomputing. The impact of two major biomedical research trends are clearly seen in this year's conference. First, the national push towards "translational research" for moving discovery from bench to bedside is manifest in the sessions on integration of biological & clinical data and on personal genomics. Second, the revolution in DNA sequencing similarly impacts our sessions on data integration, genome-wide association studies, microbiomes, personal genomics and many of the others too! Thus, we present a conference in which biocomputation is at the forefront of work aimed at bringing the fruits of the genome projects to practical applications. Other sessions focus on the emerging fields of synthetic biology and multiscale modeling. It is an exciting time for the application of computational and informatics technologies to the key problems facing biomedical science.

Indeed, computation has become a recognized component of virtually all major biomedical research efforts, as a cadre of scientists dually trained in biology & medicine as well as computer science, statistics and engineering approach problems of data analysis, fusion, and the generation of new knowledge. The NIH recently renewed the "National Centers for Biomedical Computation" program (http://www.ncbcs.org/). This program grew out of the 1999 "BISTI Report" (Biomedical Information Science and Technology Initiative) which recommended (1) a National Centers program, (2) a program on the principles of information storage, curation, analysis and retrieval, (3) the provision of additional resources for investigators creating and apply biomedical computing tools, (4) the creation of a scaleable national computational infrastructure. The Centers have a dual role of performing outstanding research in methods for biomedical computation, while also disseminating software and data they produce to others via training sessions, workshops and collaborative research relationships. A key function of the centers is to provide a milieu in which biocomputing professional can develop. The PSB meeting is proud to also contribute to the creation of a cadre of skilled professional scientists and engineers, but providing pre-meeting tutorials, travel support for students and post-doctoral fellowships, and opportunities for "bottom up" organization of new sessions.

We would like to thank our keynote speakers. Dr. Vijay Pande, Associate Professor of Chemistry will talk about recent progress in large scale simulation of biological macromolecules. Our keynote in the area of Ethical, Legal and Social Implications of Technology will be Ellen Wright Clayton, the Rosalind E. Franklin Professor of Genetics & Health Policy, and Professor of Law and Pediatrics, at Vanderbilt University. Professor Clayton is a leader in law and genetics.

PSB provides sessions focusing on emerging areas in biomedical computation. These sessions are often conceived at the meeting as people discuss the opportunities for new and exciting sessions. The efforts of a dedicated group of leaders has produced an outstanding set of sessions, with associated introductory tutorials. These organizers provide the scientific core of PSB, and their sessions are as follows:

Computational Methods Integrating Diverse Biological and Clinical Data for Translational Science
Gurkan Bebek, Mark Chance, Mehmet Koyuturk, Nathan D. Price

Genome-wide association mapping and rare alleles: from population genomics to personalized medicine
Francisco M. De La Vega, Carlos D. Bustamante, Suzanne M. Leal

Microbiome studies: Understanding how the dominant form of life affects us
James Foster, Jason Moore

Multi-scale Modelling of Biosystems: from Molecular to Mesoscale
Julie Bernauer, Samuel Flores, Xuhui Huang, Seokmin Shin, Ruhong Zhou

Personal Genomics
Can Alkan, Emidio Capriotti, Fereydoun Hormozdiari, Eleazar Eskin, Maricel G. Kann

Reverse Engineering and Synthesis of Biomolecular Systems
Gil Alterovitz, Silvio Cavalcanti, May Wang, and Marco F. Ramoni

With regards to the *Reverse Engineering and Synthesis of Biomolecular Systems* session, we were saddened by the unexpected death of our friend and colleague Dr. Marco Ramoni. Marco was an internationally known computer scientist and Bayesian theorist whose contributions ranged from new understanding of the genetic mechanisms of stroke and asthma to developing novel methodologies. He was a senior faculty member of the Children's Hospital Informatics Program and an Associate Professor of Pediatrics and Medicine at Harvard Medical School and Associate Director of Bioinformatics in Harvard Partners Center for Genetics and Genomics. He will be missed.

We are also pleased to present three workshops, in which investigators with a common interest come together to exchange results and new ideas in a format that is more informal than the peer-reviewed sessions. For this year, the workshops and their organizers are:

Mining the Pharmacogenomics Literature
Kevin Bretonnel Cohen, Yael Garten, Udo Hahn, Nigam H. Shah

Identification of Aberrant Pathway and Network Activity from High-Throughput Data
Michael Ochs, Rachel Karchin, Habtom Ressom, Robert Gentleman

Validation and Modeling of Electron Cryo-microscopy Structures of Biological Nanomachines
Wah Chiu, Helen Berman, Steven Ludtke, Gerard Kleywegt

Finally, we are happy to welcome a new group for a birds-of-a-feather meeting on "Systems Pharmacogenomics." This session is sponsored by the NIH Pharmacogenomics Research Network Statistical Analysis Resource (P-STAR), which is lead by Dr. Marylyn Ritchie under NIH HL065962.

Tiffany Murray continues expertly to manage the peer review process and assembly of the proceedings. We thank the National Institutes of Health and the International Society for Computational Biology (ISCB) for travel grant support. We are particularly grateful to BJ McKay-Morrison at ISCB for her assistance. We also acknowledge the many busy researchers who reviewed the submitted manuscripts on a very tight schedule. The partial list following this preface does not include many who wished to remain anonymous, and of course we apologize to any who may have been left out by mistake.

We look forward to a great meeting once again.

Aloha!

Pacific Symposium on Biocomputing Co-Chairs,
September 29, 2010

Russ B. Altman
Departments of Bioengineering, Genetics & Medicine, Stanford University

A. Keith Dunker
Department of Biochemistry and Molecular Biology, Indiana University School of Medicine

Lawrence Hunter
Department of Pharmacology, University of Colorado Health Sciences Center

Teri E. Klein
Department of Genetics, Stanford University

Thanks to the reviewers…

Finally, we wish to thank the scores of reviewers. PSB requires that every paper in this volume be reviewed by at least three independent referees. Since there is a large volume of submitted papers, paper reviews require a great deal of work from many people. We are grateful to all of you listed below and to anyone whose name we may have accidentally omitted or who wished to remain anonymous.

Gary Bader
Ranjit Bahadur
Gurkan Bebek
Julie Bernauer
Greg Bowman
Sharon Browning
Mike Brudno
John Bunge
Anita Burgun
Jun Cai
Ali Cakmak
Sukmohinder Chahal
Jack Chen
Xian-Ming Chen
Cheng Cheng
Jacques Chomilier
Bradley Coe
Sarah Cohen Boulakia
Markus Covert
Phuong Dao
Payel Das
Mariza De Andrade
James Faeder
Michael Feig
Juan Fernandez-Recio
Pawel Gajer
Jean Garnier
Christopher Haiman
Farhad Hormozdiari
Bryan Howie
Joanna Howson
Xuhui Huang
Yin-Fu Huang
Jui-Hung Hung
Thomas Ioerger
Andy Itsara
Asif Javed
Hui Jiang
Eun Yong Kang
Ulas Karaoz
Parminder Kaur
Alon Keinan
Sadik Khuder
Ken Kidd
Sungwan Kim
Carl Kingsford
Rob Knight
Mehmet Koyutürk
Charu Gupta Kumar
Sadhir Kumar
Douglas Lauffenburger
Jose Lazano
Seunghak Lee
Jing Li
Qian-Zhong Li
Xin Li
Liming Liang
Pu Liu
Ivan Viktorovich Maly
Elena Marchiori
Eden Martin
Andrew McPherson
Jason Mezey
Karen Nelson
Mohammed Orloff
Badri Padhukasahasram
Andrew Pakstis
John Pani
Vishal Patel
Lynette Phillips
Mihai Pop
Nathan D. Price
Kriti Puniyani
Jeroen Raes
Ben Raphael
Soumya Raychaudhuri
David Reif
Marylyn Ritchie
Masaki Sasai
Dan Schaid
Ursel Schuette
Josiah Seaman
Dong Shen
Jana Shen
Seokmin Shin
Daniel Silva
Mitchell Sogin
Jean-Marc Stayert
Yan Sun
Min-Han Tan
Nelson Tang
Nicholas Tatonetti
Cheryl Thompson
Duygu Ucar
Jerome Waldispuhl
Kai Wang
Kesheng Wang
George Weinstock
Dahlia Weiss
Scott Weiss
Zhen Xia
Lei Xie
Gokhan Yavas
Seungtai Yoon
Deniz Yorukoglu
Ye Yuzhen
GQ Zhang
Zemin Zhang
Hoghyu Zhao
Jasmine Zhou

MICROBIOME STUDIES: UNDERSTANDING HOW THE DOMINANT FORM OF LIFE AFFECTS US

MULTI-SCALE MODELLING OF BIOSYSTEMS: FROM MOLECULAR TO MESOSCALE

PERSONAL GENOMICS

REVERSE ENGINEERING AND SYNTHESIS OF BIOMOLECULAR SYSTEMS

MINING THE PHARMACOGENOMICS LITERATURE

IDENTIFICATION OF ABERRANT PATHWAY AND NETWORK ACTIVITY FROM HIGH-THROUGHPUT DATA

VALIDATION AND MODELING OF ELECTRON CRYO-MICROSCOPY STRUCTURES OF BIOLOGICAL NANOMACHINES

INTEGRATIVE –OMICS FOR TRANSLATIONAL SCIENCE

GURKAN BEBEK
*Case Center for Proteomics and Bioinformatics,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: gurkan@case.edu*

MEHMET KOYUTÜRK
*Department of Electrical Engineering and Computer Science,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: koyuturk@eecs.case.edu*

MARK R. CHANCE
*Case Center for Proteomics and Bioinformatics,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: mark.chance@case.edu*

NATHAN D. PRICE
*Department of Chemical & Biomolecular Engineering,
University of Illinois at Urbana-Champaign
600 South Mathews Avenue, Urbana, IL 61801, USA
Email: ndprice@illinois.edu*

1. Introduction

Translational research aims to bridge basic life sciences and medicine by incorporating results obtained from basic science to advance clinical applications, as well as driving basic science based on insights gained from clinical experience [3]. The recent focus on translational science has been prompted by the dramatic decline in the output of novel therapies, regardless of increased efforts and investments in research and development [4]**.** Improving the translation process to close the gap between input (investments and time) and the output (therapies, biomarkers etc.) through research practices and efforts has grabbed attention from scientific agencies and institutions, as well as researchers and patient care providers.

The revolutionary improvements in –omics technologies present a great opportunity to improve human health. However, the translation of discoveries made in the laboratory bench to bedside is an arduous and lengthy process. The complexity that is introduced by the large scale and high dimensionality of high-throughput biological datasets and the specific challenges posed by the applications (e.g., complex diseases) are growing barriers adding to the translational challenge. Most importantly, models and methods that are needed to integrate and translate this research towards clinical applications are hard to come by.

The translational bridge requires computational methods to integrate large and disparate datasets in innovative ways. These methods would enable translational research through integration of various –omics (genomic sequences, gene expression, protein expression and modifications, protein-DNA interactions, protein-protein interactions, metabolome, etc.) and clinical datasets. This session targets computational approaches aimed for finding molecular mechanisms and therapies for disease, computational methods and algorithms for the analysis of molecular and clinical measurements, systems biology approaches utilizing diverse –omics datasets for understanding diseases, and therapies, and methods relating and representing molecular or subcellular phenotypes with relevance to the clinical measurements/characteristics.

2. Session Summary

This session includes an invited talk, six reviewed papers contributed as oral presentations and a tutorial. The studies presented in this session focus on the development of computational methods for integrating diverse biological and clinical data for translational science.

2.1. *Accepted Session Papers*

Understanding the relationship between phenotype and genotype of living systems is a fundamental problem in biology, and is also key to translational science. **Wu et al.** focus on the problem of predicting phenotype from genotype in RNA viruses (e.g., HIV, influenza, West Nile). Their approach is based on representing RNA sequences as clauses in disjunctive normal form (DNF) of binary variables and finding a minimal DNF clause that represents all sequences that are associated with the phenotype of interest. They show that this approach outperforms other classification algorithms in predicting viral phenotype (e.g., drug resistance in HIV) and can provide a compact set of sequence features that are associated with the phenotype.

Similarly, in an attempt to understand the relationship between genotype and phenotype in humans, **Yu *et al.*** propose an integrative approach to comprehensively study a complex human disease; obstructive sleep apnea (OSA). They build on existing knowledge on the genetic bases of OSA and integrate this knowledge with large scale SNP data from affected and control populations and gene expression data from various tissues, in the context of human protein-protein interactions (PPIs). Seeding a search of the human PPI network with known OSA genes, they identify sub-networks that are dysregulated at the mRNA-level in OSA samples. Furthermore, they identify sub-networks enriched in proteins whose coding genes have significant p-values in a GWAS for OSA. Integration of these sub-networks lead to the discovery of potential association of OSA with Phosphoinositide 3-kinase and the STAT family of proteins, which were previously unknown.

Discovery of the genetic bases of complex human disease requires analysis of very high-dimensional genomic data, which can be understood better in the light haplotypes, however haplotype discovery is a rather intensive computational problem. Motivated by these considerations, **Otten and Dechter** develop algorithms for parallelizing haplotype search algorithms. Using a novel strategy to predict problem size from the scoring function, they improve load balancing to obtain significant speed-up for very large problem sizes.

The paper by **Turcan** ***et al.*** is aimed developing algorithms for mining novel functional gene sets to address clinical problems when there is limited information available based on the underlying physiology. The authors propose to integrate gene expression data sets from multiple and diverse sources, and combine them to identify expression biclusters exhibiting consistent changes across training data sets, providing candidate gene sets likely to be informative under various clinical phenotypes.

Sorani *et al.* describe a novel systems biology approach to analyze clinical trials. This is a new approach taken to investigate clinical trials, which relates to translational science extensively. The authors identify that high-profile trials have distinctive network characteristics. They also analyze multi-level models that integrate levels of granularity of trial conditions, interventions, and sponsors, and look into dynamic models of network evolution over time.

Lee and Gonzalez describe a data integration platform for disease gene prioritization. The platform developed is tested on Alzheimer's disease. The paper presents an integrated method for gene prioritization analysis based on heterogeneous resources, and the authors evaluate the performance of the algorithm in comparison to other methods, demonstrating that the proposed method performs better than multi-source gene prioritization systems currently available.

3. Acknowledgments

We would like to thank all the anonymous reviewers who contributed with their invaluable time, and expertise to evaluate submitted papers, and the authors who submitted their work to the Integrative-omics for Translational Science Session. MRC, GB and MK would like to mention support from National Institutes of Health grants P30-CA043703 and UL1-RR024989. NDP acknowledges funding from the NIH-NCI Howard Temin Pathway to Independence Award in Cancer Research

TOWARDS INTEGRATIVE GENE PRIORITIZATION IN ALZHEIMER'S DISEASE

JANG H. LEE[1] AND GRACIELA H. GONZALEZ[2]

[1] *School of Computing, Informatics, and Decision Systems Engineering, Arizona State University,Tempe, AZ 85287-8809*

[2] *Department of Biomedical Informatics, Arizona State University,425 N. 5th Street, Phoenix, AZ 85004*

Many methods have been proposed for facilitating the uncovering of genes that underlie the pathology of different diseases. Some are purely statistical, resulting in a (mostly) undifferentiated set of genes that are differentially expressed (or co-expressed), while others seek to prioritize the resulting set of genes through comparison against specific known targets. Most of the recent approaches use either single data or knowledge sources, or combine the independent predictions from each source. However, given that multiple kinds of heterogeneous sources are potentially relevant for gene prioritization, each subject to different levels of noise and of varying reliability, each source bearing information not carried by another, we claim that an ideal prioritization method should provide ways to discern amongst them in a true integrative fashion that captures the subtleties of each, rather than using a simple combination of sources. Integration of multiple data for gene prioritization is thus more challenging than its single data type counterpart. What we propose is a novel, general, and flexible formulation that enables multi-source data integration for gene prioritization that maximizes the complementary nature of different data and knowledge sources in order to make the most use of the information content of aggregate data. Protein-protein interactions and Gene Ontology annotations were used as knowledge sources, together with assay-specific gene expression and genome-wide association data. Leave-one-out testing was performed using a known set of Alzheimer's Disease genes to validate our proposed method. We show that our proposed method performs better than the best multi-source gene prioritization systems currently published.

1. Introduction

Of particular relevance to researchers trying to track the molecular basis of disease is to be able to increase the selectivity and sensitivity when predicting the potential association of a phenotype or function with specific genes, an area referred to as "gene prioritization". Genome sizes of species of interest are typically large, and gene prioritization is an effective means for data reduction. By ranking genes in terms of their relevance to a disease, and with an appropriate thresholidng, a select set of genes can be generated by gene prioritization. Time and cost considerations in disease research usually favor a reduced gene set which enables more focused research and facilitates more effective use of the limited resources.

Over the years, many methods have been proposed for this purpose, with molecular biologists usually favoring those that focus on the statistical analysis and consequent ranking of lists of genes from the output data of high-throughput experiments. Thus, significance analysis of microarrays (SAM), analysis of variance (ANOVA), empirical Bayes t-statistic, between group analysis (BGA), and other methods are used with the help of biostatisticians, and are sometimes provided with commonly used commercial and open-source bioinformatics tools such as Illumina's Genome Studio or caBIG's geWorkbench. Knowledge about the significant genes is sometimes provided by the tools or by sought out separately by researchers only as a way to annotate the genes, but is not used to prioritize them. Researchers have to pick and choose using their own intuition and experience.

Integrating multiple kinds of heterogeneous data and knowledge sources is a challenging problem for which formulation of a flexible and general approach is sought. A number of approaches employing protein interaction as a single knowledge source[8,19,22] have been published. Other systems, the best of which are Kohler *et al*'s[12] GeneWanderer and Aerts *et al*'s[2] Endeavour, use heterogeneous knowledge and data sources. GeneWanderer was shown to outperform many existing network-based gene prioritization algorithms.[31] It assumes a set of seed genes known to be disease genes as input and proposes a method where nodes in a protein interaction network are randomly visited (restarting the walk randomly during the process), ranking candidates with respect to their relevance to the given seed gene set. Aerts *et al* proposed Endeavour, a similarity-based approach that uses heterogeneous data to calculate the similarity between a set of candidate genes and a set of 'training' or seed genes. It was successfully employed in various biological studies. Candidate genes are ranked independently by using a selection of knowledge sources. An N-dimensional order statistics is used for combining the multiple rankings. de Bie *et al*[6] used similarity measures and kernels corresponding to each data source and integrated rankings from multiple sources by weighting kernels. Li *et al*[13] employed GO-derived gene similarity networks and a PPI network, applied random walk with restart to each and combined the multiple rankings by using a discounted rating system.

Albeit intended on a genomic scale, most of the currently available knowledge sources and experimental platforms have rather low sensitivity. For example, current PPI databases are estimated to capture only 10% of true interactions.[9] Often times data and knowledge sources are orthogonal, with pieces of information absent in one being provided in another. Thus, distinct sources tend to have a complementary nature such that a holistic perspective on genes can be gained by appropriately complementing and integrating distinct sources. Existing approaches for multiple sources take data and knowledge sources separately, whereby their complementarity can be easily lost. Also, many involve rather high computational cost or assume specific types of data and limit the applicability to other data types.

Given a known group of genes associated with a specific disease as a "seed", we hypothesized that the degree of association of a candidate gene with the seed genes signifies its relevance to the disease. All knowledge about the genes was represented in a single network, which can be appropriately configured based on types of data, availability and reliability. Here, we used protein-protein interactions (PPIs), Gene Ontology annotations, gene expression data and SNP data from a Genome-Wide Association Study for validating our approach. Application to a large number of diseases of distinct kinds showed uniform performance level and hence no bias for particular kinds of diseases. We report the results of this general experiment, as well as a more extensive evaluation using genes related to Alzheimer's Disease (AD).

2. Material and methods

PPI and Gene Ontology associations were used as knowledge sources in building an integrated gene-gene association network used for gene prioritization. This is what we called the base scheme (BS) for purposes of evaluation. Additionally, gene expression and GWAS data were used as empirical data sources and incorporated in the prioritization by adding a value (level of significance) to each node in the integrated network above. This is what we called the incorporated scheme (IS). In the following subsections, we outline how the associations for each component of the network are defined and integrated, and present two experimental setups (the base scheme and the incorporated scheme) to validate the approach.

2.1. *Establishing Gene Ontology associations*

The Gene Ontology (GO) consists of a directed graph of terms organized under three main categories: biological process, cellular component and molecular function. Genes are annotated with those terms that apply to them. Resnik[21] defined similarity between two GO terms t_0, t_1 under the same category as

$$sim(t_0, t_1) = \mathrm{IC}_{ms}(t_0, t_1) = \max \mathrm{IC}(t_p) \tag{1}$$

where $t_p \in parents(t_0, t_1)$, and IC(t) is the information content of term t which is defined as $\mathrm{IC}(t) = -logP(t)$ with $P(t)$ being the probability of occurrence of the term across a genome.

Couto *et al*[5] defined similarity between two genes g_0, g_1 with respective terms $t_a \in \{terms(g_0)\}$ and $t_b \in \{terms(g_1)\}$ as

$$sim(g_0, g_1) = \max_{a,b} sim(t_a, t_b)\mathrm{IC}(t_a)\mathrm{IC}(t_b) \tag{2}$$

Term similarity is a normalized quantity ranging between 0 and 1. We used GO annotations[11] of the human genome, which included a total of 14,685 genes annotated with biological process terms, with a total term occurrence count of 60,792 for an average of 4.140 terms per gene. In establishing a gene-gene association based on GO annotations, we varied the similarity threshold from 0.30 to 0.70 in increments of 0.10 to retain gene pair similarity only above or equal to the given threshold, obtaining five nested sets of associations.

2.2. *Protein-protein interactions*

Three protein interaction databases were employed, to match those used by Kohler *et al* in [12] and allow a fair comparison: HPRD,[20] STRING[16] and NCBI yeast protein interactions. HPRD is a manually annotated protein interaction data set: the one we used had 2,125 homomeric interactions and 36,631 heteromeric interactions. The STRING database contains information from four sources (genomic context, high-throughput experiments, coexpression, and derived from text), including direct (physical) and indirect (functional) associations. We used version 8.3, which covers 2.6 million proteins from 630 organisms. Each interaction in STRING is assigned a significance score (non-linear) in the range between 150 and 1000. In addition, known protein interactions in yeast were downloaded from NCBI.[17] Each yeast protein was mapped to a human ortholog using InParanoid.[18] Only interactions involving protein pairs that have a 100% match score to human orthologs were retained (a total of 39,665).

Interacting proteins were each mapped to coding genes and then a set of interacting genes were obtained. Some common interactions in the databases derive from single experimental evidence and hence there exists a degree of duplicity among the three databases. The three PPI networks were combined into a single network by counting edges only once irrespective of their duplicity:

$$\{e'(g_1, g_2)\} = \cup\{e_{N_i}(g_1, g_2)\}, 1 \leq i \leq N \tag{3}$$

with $e'(g_1, g_2)$ being the edge between nodes g_1 and g_2 in the combined network and N being the total number of PPI networks. Five distinct sets of associations were obtained by using nested sets of interactions with different STRING significance score thresholds (300, 400, 500, 600 and 700).

2.3. *Gene expression*

For this paper, we used microarray expression data sets by Webster *et al*,[23] comprised of control and AD case samples. Genes showing significantly distinct levels between normal and disease cells were identified by using differential expression analysis. Wilcoxon rank sum test was applied to expression levels from the two groups of samples and a P-value of each gene's differential expression was obtained. The P-value threshold was set to 0.05. The significance of a gene G, $S(G)$, from differential expression was calculated as:

$$S(G) = -log(\text{P-value}) \tag{4}$$

2.4. *Genome-wide association study*

SNP genotyping is performed on genomes from normal and disease samples. Certain SNP may show distinct presence in one group vs the other e.g., allele A constitutes 80% of disease samples at a certain locus while it constitutes 30% in normal samples. A P-value can be calculated for each SNP and hence for a corresponding gene if the locus of the SNP is within or close to the gene, which would imply the gene is strongly relevant to a specific disease. If a SNP is too distant from genes (more than 20kb away upstream or 5kb downstream), then it was not included in our experiments. Similar to expression data, disease significance P-values were calculated and assigned to genes by using Eq 4.

2.5. *Network representation*

To construct the networks used for the base (BS) and incorporated (IS) schemes, the PPI and GO associations described above were used as edges, with genes mapped to nodes. If more than one knowledge source associated two genes g_1 and g_2, then the edge is weighted according to the multiplicity of the number of associating sources. Thus, if N sources were associating the two genes then weight$(e(g_1, g_2)) = N$.

Gene g may be completely missing or may not have a P-value above a threshold in the outcome of some experimental data, and have P-values above thresholds only in N_e number of effective sources. Given a significance $S_i(g)$ from empirical data source i $(1 \leq i \leq N_e)$ for a given disease, gene g's overall empirical significance is calculated as

$$S(g) = \sum_{i=1}^{N_e} S_i(g) \tag{5}$$

That is, the sum of all significance values is assigned as a combined significance score for the gene (its aggregate experimental significance).

2.6. *Base scheme*

Given a set of training seed genes $\{S_i\}$, candidate gene C was scored as follows:

$$score(C, S) = \sum^{\forall S} e(C, S_i) \tag{6}$$

where $e(C, S_i)$ is a non-zero value if an edge exists between C and S and 0 otherwise. Either only the edge presence between C and S can be recognized for scoring, or its weight from the aggregate network can be considered, i.e.,

$$e(C, S)_{BS1} = \mathbf{1}\{e(C, S)\} \tag{7a}$$

$$e(C, S)_{BS2} = \text{weight}(e(C, S)) \tag{7b}$$

with $\mathbf{1}$ being an indicator function corresponding to edge presence. If only the presence of an edge is considered, then Eq 7a is used together with Eq 6. This will be referred to as base scheme 1 (BS1). If edge weight is considered, then Eqs 7b and 6 are used which will be referred to as base scheme 2 (BS2). Candidate genes are ranked according to their scores.

2.7. *Empirical data incorporation scheme*

The network topology used in the empirical data incorporation scheme (IS) is the same as the one in the base scheme. Candidate gene C can have an edge to jth seed gene T_j of an overall empirical significance $S(T_j)$. Then T_j's contribution to the score of C is calculated as

$$e(C, T_j) + kS(T_j) \tag{8}$$

where k is a scaling factor, the value of which is to be set according to data reliability. If an edge does not exist between them, then T_j's contribution is 0. The contribution from each training gene T_j, $1 \leq j \leq |T|$, in the training set to candidate gene C is added up for its combined score:

$$score(C) = k_1 S(C) + \sum_{j=1}^{|T|} [e(C, T_j) + k_2 S(T_j)] \tag{9}$$

where k_1 and k_2 are scaling factors and $|T|$ the total number of training genes. The ranking of the candidate genes corresponds to the combined scores of the candidate genes.

2.8. *Validation*

The disease gene sets from Kohler *et al*[12] were used. Leave one out testing was performed by holding out one disease gene as a true test gene to be (ideally) recalled from the disease gene set by taking the remainder genes as a training gene set, and this was repeated for each gene over all disease gene sets. Sensitivity and specificity values were calculated as defined in (2). Specifically, ranking results were aggregated and the number of true test genes above a given ranking threshold was counted as true positives. The number of test genes below the threshold, non-test genes below the threshold and non-test genes above the threshold were respectively counted as false negatives, true negatives and false positives. As frequently done in literature, a narrowed-down set of genes (e.g., 100) in closest proximity to the true test gene along its chromosome is given as a candidate set. We also show the ranking obtained over all genes in the genome.

Current knowledge sources may involve degrees of incompleteness and incorrectness. This would correspond to false positive and negative edges in networks. Facing this, we randomly perturbed 10% of network edges by randomly reassigning them in an experiment. Eight such instances of randomly perturbed networks were generated and the base scheme was applied to each of them.

Table 1. AD gene prioritization

	Base		Endeavour	GeneWanderer		Incorp.	
Gene	Rank	Rk100	Rk100	Rank	Rk100	Rank	Rk100
APOC1	93	2	5	275	7	1	1
APOE	1	1	4	17	4	1	1
APP	382	1	4	264	1	156	1
CLU	7	1	9	102	2	17	1
CR1	437	2	44	1158	3	352	2
GAB2	202	1	31	496	3	452	2
MSRA	-	100	24	6511	11	-	100
PICALM	444	1	8	978	3	95	1
PSEN1	1	1	2	14	1	1	1
PSEN2	7	1	4	84	1	25	1
PVRL2	7	1	47	67	4	15	2
RELN	439	1	43	957	5	413	1
TOMM40	1261	10	86	3319	18	34	2

3. Results

Genes implicated in AD were collected from the literature (1, 10, 15, 14, 24, 25) (Table 1). For comparison of performance, gene prioritization based on random walk with restart (RWR) as described by Kohler *et al* (12) was implemented. In RWR, nodes are navigated in a random fashion starting from a gene randomly selected from a given set of seed genes. Gene ranking in RWR is according to the visit frequency at the conclusion of iteration following a convergence criteria. In addition, Endeavour^2 was downloaded from the authors' website. It randomly selects 99 genes other than true test gene to produce a 100 gene candidate set together with the test gene. Even though the candidate gene sets used for Endeavour are different from the ones used for base scheme and RWR, we reasoned the set size is sufficiently large from a statistical sense to facilitate sound comparisons and show the rankings under the column name of Rk100.

The base gene prioritization scheme was applied to the AD gene set. The same set was also used for Endeavour and GeneWanderer. When gene APOC1 was left out as a true test gene to be recalled and the other genes were used as a training seed gene set (row 1 in Table 1), there were 92 other genes from the human genome which ranked more significantly (column Base-Rank in Table 1). When the candidate gene set was reduced to the 100 genes of closest proximity (Loc100 set), APOC1 ranked 2nd highest (column Base-Rk100). Endeavour's ranking of the gene was 5th out of 100 genes and RWR's ranking was 275th among

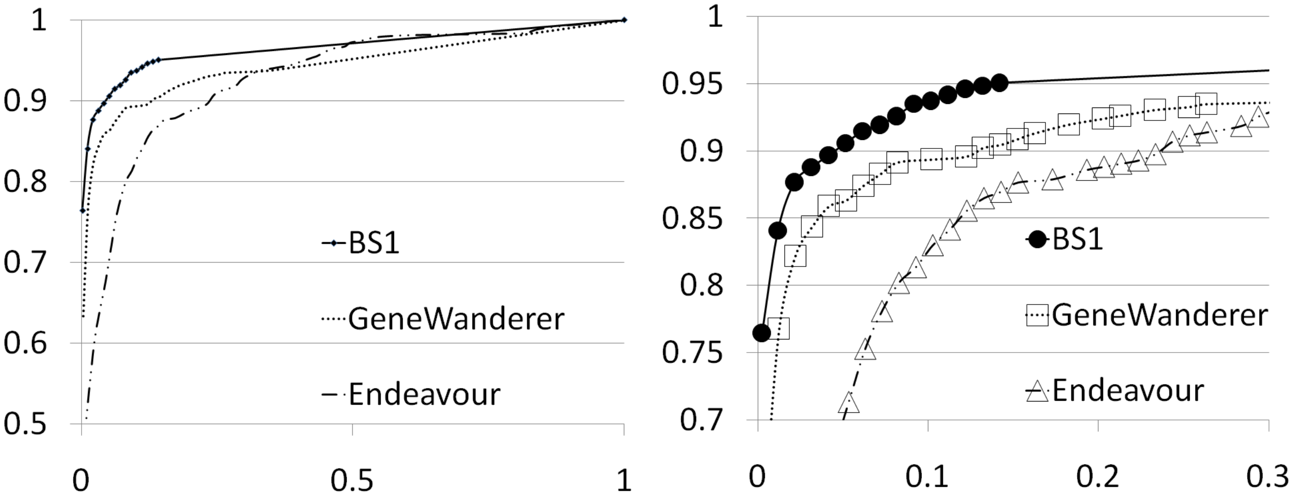


Fig. 1. ROC curves of specificity vs. 1-sensitivity (a) Base scheme has a larger AUC than Endeavour and RWR. (b) Close-up of higher sensitivity range

Table 3. AUC difference between base scheme 1 and base scheme 2; BS1 - BS2

GO $\backslash PPI$	300	400	500	600	700
30	-5.603	-12.101	-12.751	-14.561	-14.451
40	-0.265	+5.767	+9.010	+8.708	+4.035
50	+1.815	+5.567	-0.048	+6.935	+7.971
60	+0.986	+1.065	-0.157	+0.390	+0.000
70	+0.532	+0.464	+0.000	+0.165	+0.398

Units in 10^{-4}

entire genome and 7th among Loc100 genes. Each subsequent row can be read in a similar fashion. Thus, the base gene prioritization scheme ranked the AD set genes more significantly than RWR (signed rank test P-value=6.836×10^{-3}.) and Endeavour (P-value= 2.148×10^{-2}).

In order to assess the applicability of the base scheme (BS1) to other diseases besides AD, we applied it to disease gene set of Li *et al*[13] (Li10) which was derived from the complete Kohler *et al* set. It includes 36 diseases and genes implicated therein. The receiver operating characteristic (ROC) curve of the base scheme BS1 is shown in Fig 1 together with the curves of Endeavour and RWR for the same set. AUC value of the base scheme was 0.9655 while, for Endeavour and RWR, the AUC values respectively were 0.9287 and 0.9442. The reasonable AUC value means the base scheme is applicable to other diseases in general as well. Base schemes 1 and 2 were compared over the Li10 set and their AUC values showed a marginal difference possibly suggesting edge multiplicity does not greatly contribute in distinguishing true test gene from the other candidate genes (Table 3). Subsequently, we used only base scheme 1 and will refer to it as the base scheme.

Knowledge sources such as PPI or GO may entail some levels of false and missing annotations. In order to evaluate the influence of such noise on the performance of the base scheme, 10% of the edges in the combined network were randomly rewired. Eight such instances of the perturbed networks were generated, and then the base scheme was applied. In all cases, AUC values decreased by small degrees, but consistently from that of the un-perturbed network; average AUC value was 0.96070 and standard deviation 0.00223 (Fig. 2 and Table 4). Only a slight degradation in the AUC of the perturbed network means our base scheme is robust with respect to a noticeable amount of possible mis-curations in the knowledge sources and corresponding noise in the network.

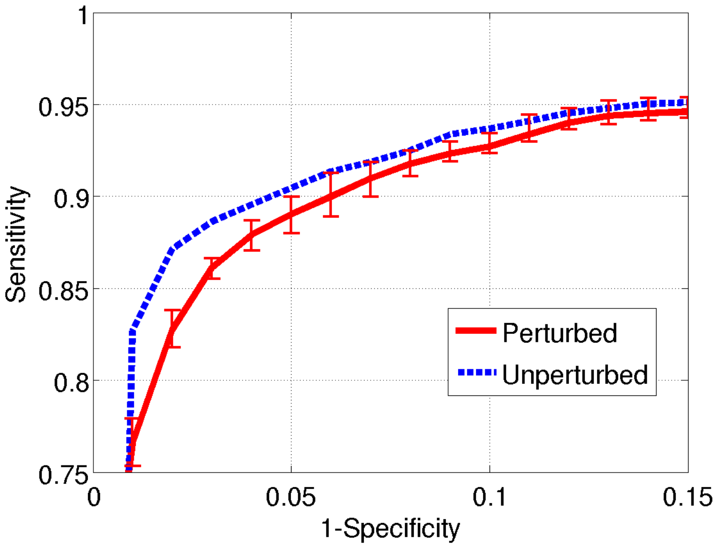


Fig. 2. ROC's from perturbed and unperturbed networks

Table 4. AUC values from perturbed networks

Instance	1	2	3	4	5	6	7	8	Average	St.dev.
AUC	0.96119	0.95875	0.96216	0.95869	0.96533	0.95993	0.96101	0.95908	0.96070	0.00223

Table 5. AUC values from application to different disease categories

Type	Cancer	Monogenic	Polygenic	Average
BS1	0.95727	0.96677	0.98025	0.96810
RWR	0.95414	0.90535	0.94978	0.95890
Endeavour	0.87947	0.94471	0.88191	0.90203

Diseases were categorized as belonging to one of three types by Kohler *et al*: cancer, monogenic and polygenic. Cancer and polygenic categories each included 12 diseases, and monogenic 86 diseases. We chose the 6 largest disease gene sets from each category to form categories balanced in count and applied the base scheme, Endeavour and RWR to each. AUC values were similar across disease categories (Table 5), thus suggesting that the base scheme is not biased to a particular category of diseases. Higher AUC values were produced by BS throughout the different categories.

The contribution of individual knowledge sources was assessed by using either PPI or GO associations alone and by comparing the resulting AUC values with the ones obtained with aggregate sources. Specifically, 5 sets of GO associations were produced with distinct thresholds of 0.30 to 0.70 in increments of 0.10, and also 5 sets of PPIs with thresholds 300 to 700 in increments of 100. A total of 35 networks resulted; 5 with only GO associations as edges, 5 PPI only, and aggregate networks in 25 different combinations of GO and PPI thresholds. The Base Scheme was applied to the Li10 set for each of the networks. The AUC value monotonically increased as GO or PPI thresholds were lowered (resulting in more network edges) (Figs 3(a), 3(b)). The highest AUC value was produced with the aggregate network of least stringent threshold combination (PPI 300 and GO 0.30).

The PPI network alone shows reasonable AUC values under varying thresholds (bottom-most curve of Fig. 3(a)). Aggregation with GO network consistently improves the AUC values. However, GO networks alone show rather low AUC values especially at high thresholds, but aggregation with PPIs, even at the highest threshold, drastically improves AUC values. Clearly, aggregation of networks from distinct knowledge sources is an effective way of comprehensively utilizing their respective information content, and our base scheme

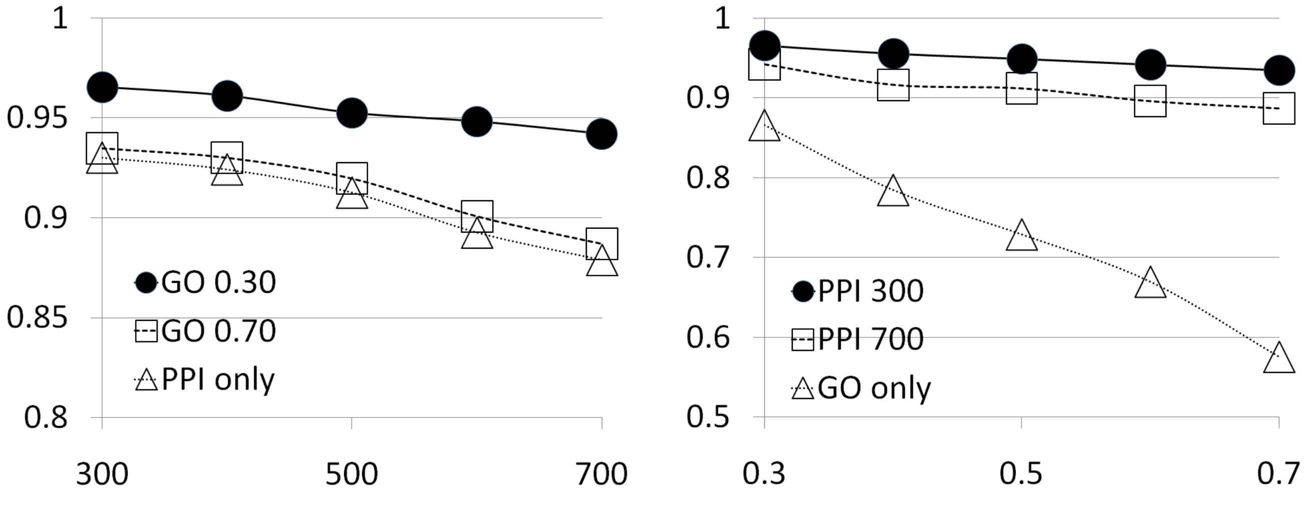


Fig. 3. AUC values from different knowledge source combinations (a) AUC vs. PPI threshold (b) AUC vs. GO threshold

indeed utilizes the higher information content.

3.1. *Incorporation of empirical data*

Alzheimer's Disease GWAS and differential expression data were incorporated in the gene prioritization process (Table 1 column Incorp.) as explained in the incorporated scheme. Improvement over the base scheme was rather marginal (P-value=0.1934). This may be attributable to a rather low reproducibility of significant genes between experiments, especially expression data.[7,27,28] A number of approaches have been suggested for an appropriate interpretation and extraction of useful information from experimental data including shifting of focus towards groups of genes rather than on individual genes.[29] A new formulation of the incorporated scheme is left as a future work, which considers the difference in nature of experimental data.

4. Discussion and conclusion

Two different knowledge sources were each represented in a network and unified in a model that allows for additional sources to be added in a similar fashion. Each independent knowledge source is likely incomplete and missing many associations between genes.[9] The proposed knowledge integration method (base scheme) complements incomplete knowledge sources to produce a more comprehensive view of genes. For example, among well known AD genes, APOE has edges to genes APP, CLU, PSEN1 and PSEN2 in PPI network and lacks an edge to PICALM (Fig 4). The GO network does not have the APOE-APP edge but contains the APOE-PICALM edge. We compared our proposed method to two of the best multi-source gene prioritization algorithms. Endeavour utilizes knowledge sources separately and tended to produce the lowest AUC values among the compared algorithms. The method proposed here effectively integrates individual knowledge sources to overcome the incompleteness of each.

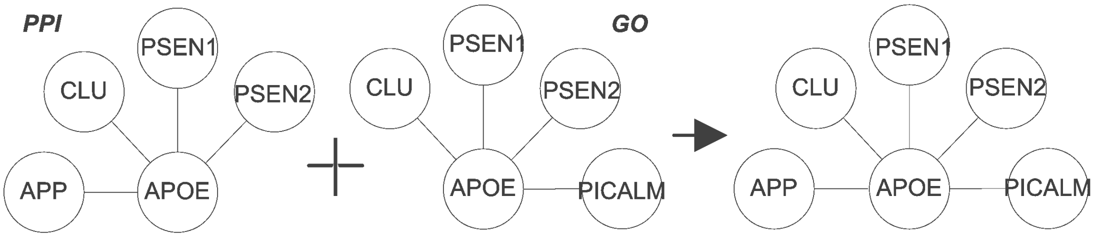


Fig. 4. Network aggregation

The base scheme alone showed better performance than Endeavour and RWR. Rankings based on combined networks were consistently better than rankings based on individual networks. There is a degree of overlap between the two knowledge sources (PPI and GO), since the same information from literature is frequently used to annotate genes. Still there is information content in one source which is not captured in the other. The edge formation by similarity criterion in the GO network can associate genes that are highly related in pathways or from biological perspectives which do not directly interact through their protein products and hence is missed in a PPI network. The described schemes rely on the association between genes to infer disease genes from known genes. The effectiveness of this approach was shown through a series of experiments. The information from knowledge sources and experimental data vary in reliability, degree of curation and level of acceptance. For example, many protein interactions have been verified over time and are well accepted, while high throughput interaction data tends to involve a high rate of false positives.

Our Gene ontology annotation of genes reflects a relatively high level of verification and curation. On the other hand, experimental data is subject to a high level of noise and variance and has not been extensively and thoroughly verified. Hence a network was not directly formed from experimental evidence at this stage, and only node significance was adjusted in accordance with the experimental significance. Our schemes are robust against false positives and missing knowledge as shown in the perturbation experiment. Future work will be directed at incorporating empirical data from experiments in a way that is more consistent with the way knowledge sources are used. While particular knowledge sources and experimental data were used for illustration, the described schemes are sufficiently general to be used with other data types as well. After the preparation of our manuscript, a gene prioritization method[30] was noted for its use of diverse data with a Bayesian approach. While a readily accessible version of their algorithm was unavailable, it will be interesting to perform a comparative study involving it.

5. Acknowledgement

Authors are grateful to Dr. Matthew Huentelman for AD GWAS data. This work was partially funded by the NIH/NIA under grant P30 AG019610-09.

References

1. R. Abraham *et al.* A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Med Genomics*, 1:44 (2008).
2. S. Aerts *et al.* Gene prioritization through genomic data fusion. *Nature Biotechnology*, **24(5)**, 537-44 (2006).
3. M. Ashburner *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet*, **25**, 25-29 (2000).
4. J. Chen, B. J. Aronow and A. G. Jegga. Disease candidate gene identification and prioritization using protein interaction networks. *BMC bioinformatics*, **10** (2008).
5. Couto *et al.* Implementation of a functional similarity measure between gene-products. *Technical report DI/FCUL TR 03-29* (2003).
6. T. de Bie *et al.* Kernel-based data fusion for gene prioritization. *Bioinformatics*, **23(13)**, i25-32 (2007).
7. L. Ein-Dor *et al.* Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171-178 (2005).
8. G. Gonzalez *et al.* Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. *Pacific symposium on biocomputing* (2007).
9. G.T. Hart *et al.* How complete are current yeast and human protein-interaction networks? *Genome Biol.*, **7**, 120 (2006).
10. D. Harold *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet*, **41(10)**, 1088-93 (2009).
11. ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/.
12. S. Kohler *et al.* Walking the interactome for prioritization of candidate disease genes. *the American journal of human genetics*, **82(4)**, 949-58 (2008).
13. Y. Li *et al.* Integration of multiple data sources to prioritize candidate genes using discounted rating system. *BMC Bioinformatics*, **11** Suppl 1:S20 (2010).
14. H. Li *et al.* Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol*, **65(1)**, 45-53 (2008).
15. J. Lambert *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet*, **41(10)**, 1094-9 (2009).
16. C. von Mering *et al.* String 7 - recent developments in he integration and prediction of protein interactions. *Nucleic acids research*, **35**, D358-62 (2003).
17. NCBI ftp://ftp.ncbi.nih.gov/gene/GeneRIF/interactions.gz (2010).
18. G. Ostlund *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196-203 (2010).
19. M. Oti *et al.* Predicting disease genes using protein-protein interactions. *Journal of medical genetics*, **43(8)**, 691-8 (2006).
20. T.S.K. Prasad *et al.* Human protein reference database - 2009 update. *Nucleic acids research*, **37**, D767-72 (2009).
21. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *IJCAI* (1995).
22. A. Sharma *et al.* Gene prioritization in type 2 diabetes using domain interactions and network analysis. *BMC genomics*, **11** :84 (2010).
23. J.A. Webster *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *American journal of human genetics*, **84(4)**, 445-58 (2009).
24. E. Reiman *et al.* GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron*, **54(5)**, 713-20 (2007).
25. P. Kramer *et al.* Alzheimer disease pathology in cognitively healthy elderly: A genome-wide study. *Neurobiol Aging*, May 6 (2010).

26. S. Frantz. An array of problems. *Nat. Rev. Drug Discov.*, **4**, 362-363 (2005).
27. G. L. Miklos and R. Maleszka. Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.*, **22**, 615-621 (2004).
28. E. Marshall. Getting the noise out of gene arrays. *Science*, **306**, 630-631 (2004).
29. D. Yang. Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics*, **24**, 265-271 (2008).
30. B. Linghu *et al.* Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol. **10** (9):R91 (2009).
31. S. Navlakha and C. Kingsford. The power of protein interaction networks for associating genes with diseases, *Bioinformatics*, **26(8)**, 1057-63 (2010).

SYSTEMS BIOLOGY ANALYSES OF GENE EXPRESSION AND GENOME WIDE ASSOCIATION STUDY DATA IN OBSTRUCTIVE SLEEP APNEA

YU LIU

Center for Proteomics & Bioinformatics, Case Western Reserve University (CWRU), Cleveland, Ohio, 44106, USA, Email: yxl442@case.edu

SANJAY PATEL

Division of Pulmonary, Critical Care and Sleep Medicine, CWRU, Cleveland, Ohio, 44106, USA, Email: srp20@case.edu

ROD NIBBE
SEAN MAXWELL

Center for Proteomics & Bioinformatics, CWRU, Cleveland, Ohio, 44106, USA, Email: rkn6@case.edu; stm@case.edu

SALIM A. CHOWDHURY

Department of Electrical Engineering & Computer Science, CWRU, Cleveland, Ohio, 44106, USA, Email: sxc426@case.edu

MEHMET KOYUTURK

Department of Electrical Engineering & Computer Science, CWRU, Cleveland, Ohio, 44106, USA, Email: mxk331@case.edu

XIAOFENG ZHU

Department of Epidemiology and Biostatistics, CWRU, Cleveland, Ohio, 44106, USA, Email:xiaofeng.zhu@case.edu

EMMA K. LARKIN

Division of Allergy, Pulmonary and Critical Care, Vanderbilt University Medical Center, 1215 21st Ave S., Nashville, Tennessee, 37232, USA, Email: emma.larkin@vanderbilt.edu

SARAH G BUXBAUM

Jackson Heart Study, Jackson State University, Jackson, MS 39213, USA Email: sarah.g.buxbaum@jsums.edu

NARESH M. PUNJABI

Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205,USA, Email: npunjabi@jhmi.edu

SINA A. GHARIB

Center for Lung Biology, Division of Pulmonary and Critical Care Medicine, University of Washington, Seattle, WA 98109, USA. Email: sagharib@u.washington.edu

SUSAN REDLINE

Department of Medicine, CWRU, Cleveland, Ohio, 44106, and Depart of Medicine, Brigham & Women's Hospital and Beth Israel Deaconess Medical School, Harvard Medical School, Boston, MA, 02115 Email: sredline@partners.org

MARK R. CHANCE

Center for Proteomics & Bioinformatics, Department of Genetics, Case Western Reserve University , Cleveland, Ohio, 44106, USA, Email: mark.chance@case.edu

The precise molecular etiology of obstructive sleep apnea (OSA) is unknown; however recent research indicates that several interconnected aberrant pathways and molecular abnormalities are contributors to OSA. Identifying the genes and pathways associated with OSA can help to expand our understanding of the risk factors for the disease as well as provide new avenues for potential treatment. Towards these goals, we have integrated relevant high dimensional data from various sources, such as genome-wide expression data (microarray), protein-protein interaction (PPI) data and results from genome-wide association studies (GWAS) in order to define sub-network elements that connect some of the known pathways related to the disease as well as define novel regulatory modules related to OSA. Two distinct approaches are applied to identify sub-networks significantly associated with OSA. In the first case we used a biased approach based on sixty genes/proteins with known associations with sleep disorders and/or metabolic disease to seed a search using commercial software to discover networks associated with disease followed by information theoretic (mutual information) scoring of the sub-networks. In the second case we used an unbiased approach and generated an interactome constructed from publicly available gene expression profiles and PPI databases, followed by scoring of the network with p-values from GWAS data derived from OSA patients to uncover sub-networks significant for the disease phenotype. A comparison of the approaches reveals a number of proteins that have been previously known to be associated with OSA or sleep. In addition, our results indicate a novel association of Phosphoinositide 3-kinase, the STAT family of proteins and its related pathways with OSA.

1. Introduction

Although its precise functions are not entirely known, sleep is important for numerous physiological and cognitive functions. Sleep disorders can have a range of consequences, from minor to severe, such as untimely drowsiness, motor vehicle collisions, and workplace accidents as well as increase risk of hypertension, diabetes and mortality. Of the more than 70 known sleep disorders, obstructive sleep apnea (OSA) is one of the most common[1,2]. OSA is a complex disorder caused by a repetitive collapse of the upper airway during sleep, disrupting breathing and sleep. Repetitive episodes of obstruction cause intermittent drops in blood oxygen and increases in carbon dioxide levels, which can lead to frequent arousals from sleep. OSA is a major cause of chronic sleep deprivation and excessive daytime sleepiness. It is estimated that up to 5% of adults in Western countries are likely to have OSA syndrome[3]. Treatments for OSA include behavioral therapies (such as changing sleeping positions), use of mechanical devices, and surgery to increase the patency of the airway. However, after decades of research the molecular mechanisms underlying OSA remain unclear.

OSA is unlikely to be a simple condition associated with a few genes or proteins; instead, it is likely a manifestation of multiple interconnected aberrant pathways and numerous molecular abnormalities[4]. In addition, it is a risk factor for many other diseases and many other diseases increase the risk of OSA. For example, OSA is associated with inflammatory states[5-8] and oxidative stress[9,10]. While obesity is one of the strongest risk factors for OSA[11], other co-morbidities include insulin resistance, hypertension, and cardiovascular disease[12-14].

Multiple studies indicate an important genetic basis for OSA, and genetic factors alone can explain approximately 30-40% of the variance of the apnea hypopnea index (AHI), a quantitative measure of OSA, defined by the number of apneas and hypopneas per hour of sleep[15,16]. OSA is also mediated by environmental factors, most obviously through those that link it to related traits such as obesity[17], but which may include influences associated with irritant exposures, alcohol use and sleep deprivation. Efforts to identify genetic variants related to OSA, include family as well as genome-wide, case-control studies and are an important attempt to provide diagnostic and/or prognostic information related to the disease. In linkage analysis of families with an affected OSA member, Larkin et al. have identified several chromosomal regions linked to the AHI[18]. Some, but not all, of the genetic pathways were believed to be obesity dependent[18]. In another study, based on a pre-selected gene set and SNP data, the same group found five genes significantly associated with OSA and the AHI[19]. Many additional genes in these and related genome-scale studies are likely relevant to mediating disease, but due to the multiple hypotheses testing problem when thousands of genes are analyzed, only a few genes with very significant p-values are allowed to pass the relevant filters for significance. The problem of identifying as many biologically significant genes as possible in such an analysis remains very important.

Network modeling of protein-protein interactions provides a relatively new context to study disease and identify disease-related genes. The effectiveness of network-based approaches to the identification of multiple disease markers has been demonstrated in the context of various diseases, such as colon cancer[20]. The aim of this study is to uncover protein-protein sub-networks associated with OSA by integrating data from multiple high-dimensional studies, both to demonstrate the power of systems biology data integration in developing novel mediators of OSA and to use the novel data available in the field to explore and validate new computational approaches. To achieve these goals, we applied two approaches, 1: candidate gene approach integrated with adipose tissue microarray data; 2: genome-wide approach integrated with adipose microarray data. The first approach is based on the method proposed by Nibbe et al.[20]: we use 56 seed proteins to drive a search of a protein-protein interaction (PPI) to discover rank-ordered sub-networks associated with OSA. In this method, proteins known or suspected to be associated with OSA and related co-morbidities are used to seed a search of a well-annotated, human PPI for candidate sub-networks, which are subsequently scored with gene expression data to derive candidate sub-networks underlying the disease phenotype. We demonstrate the utility of this approach, using a biased seed set, to provide interesting candidate sub-networks for further exploration in the etiology of OSA.

In a second unbiased approach, we mapped p-values obtained from a case-control GWAS study based on OSA phenotypes to nodes of an adipose tissue-specific interactome constructed from gene expression data[21-23]. Subsequently, we used Cytoscape based tools to identify sub-networks significantly associated with OSA. A novel feature of the study is that nodes that were highly significant along with nodes that were not the most significant in the GWAS analysis both provided important contributions to discovering the sub-networks that are of potential biological significance for the phenotype. This indicates an approach for extending the value of GWAS data to other complex phenotypes. By incorporating data from both approaches, sub-networks were identified that included targets known to be associated with OSA or sleep in general and also indicated that PI3K, STAT family, and related pathways may have important functional roles in OSA.

2. Material and Methods

2.1 *Network construction*

Two methods to construct PPI networks were used in this study. First, 56 seed genes/proteins were selected based on knowledge of the underlying biology and prior genetics studies of OSA[19]. The list of genes is provided as supplementary material (Supplementary Table 1), and can be reached at website (http://proteomics.case.edu/news_events.aspx?newsid=38). Most of the genes are known to be in one or more pathways representing intermediate phenotypes for OSA: craniofacial morphology, obesity, inflammation, and ventilatory control pathways, or across multiple pathways, through biologic pleiotropy. A traditional association study has been conducted on this set of proteins and has been recently published[19]. Ingenuity Pathway Analysis (IPA) software (Ingenuity® Systems, www.ingenuity.com) was used to construct networks by the following steps[24]:a) seed proteins are combined into networks that maximize their specific connectivity, which is their interconnectedness with each other relative to all molecules they are connected to in the Ingenuity Knowledge Base; b) additional proteins from the Ingenuity Knowledge Base are added to specifically connect two or more smaller networks by merging them into a larger one. The networks were limited to 70 nodes each to permit ease of computational scoring of sub-networks using mutual information (see below). The overall network score is based on the number of seed proteins they contain. The two top scoring networks were used in the analysis (See Results and Discussion). Note that IPA will cluster a protein complex or protein family into a single node if a number of components or family members are present in the network. For scoring purposes (see below), the expression value of the family or complex is represented by the maximum expression value among its components.

Second, an interactome specific to adipose tissue, which has been previously constructed by combining gene expression data from adipose tissues and PPI information from public databases and published papers[22,23,25,26], was optimized following curation with recent next-generation sequencing data[23]. Briefly, mRNA expression levels from Su et al[25,26] and Wang et al[23], which are used to determine the significance of a gene to the network, were estimated by combining results from microarray (chip based) experiments along with next generation sequencing results from selected references[23,25,26]. Protein nodes with mRNA levels below a defined threshold were considered as absent (the threshold for data from next generation sequencing is 20 reads; in the case of data from microarray experiments, the threshold for normalized expression level is 200[25,26]). Interactions between two proteins supported by at least three databases and two experiments were added to the interactome[22]. The adipose specific interactome in SIF format is provided as supplementary data (Supplementary Table 2, http://proteomics.case.edu/ news_events.aspx?newsid=38). Network-Analyzer is used to compute the network properties, such as the average shortest path length and the node degree distribution[27].

2.2 *Gene expression data processing*

Experimentally derived mRNA expression data for subcutaneous and visceral fat tissues were measured by cDNA microarray using the Affymetrix Human Gene 1.0 ST Array on intra-operative samples from 10 OSA patients and 8 controls undergoing elective ventral hernia repair surgery. Adipose tissue was chosen for expression studies since it is accessible and because of the central role of obesity in the pathogenesis of OSA. The information about these samples, such as sample IDs, AHIs, are provided as supplementary data (Supplementary Table 3, http://proteomics.case.edu/news_events.aspx?newsid=38). Expression values were generated using the aroma package from bioconductor[28]. Robust multichip average (RMA) and quantile normalization methods were used for background correction and normalization. In an initial analysis, two subcutaneous and three visceral samples (i.e., five out of 36 samples) had much larger variances than other samples (GEGF ID 14, 15, 16, 21, 22, all of them are control samples, see Supplementary Table 3, http://proteomics.case.edu/news_events. aspx?newsid=38), these were treated as outliers, and removed.

2.3 *Subnetwork scoring and detecting using mutual information (MI)*

Once a network enriched in seed proteins is constructed, we identify dysregulated sub-networks within this network using mRNA expression data. The aim of this procedure is to find sets of genes that exhibit coordinate differential expression, in that they can discriminate case and control samples when their expression profiles are considered together. For this purpose, we use an information-theoretic measure of coordinate dysregulation that was developed by Chuang et al.[29] and was previously used to detect dysregulated subnetworks in breast cancer metastasis[29] and late stage colorectal cancer[20]. This measure of sub-network dysregulation is powerful in that it provides a multivariate assessment of the coordination between multiple genes in their differential expression.

Namely, for a given set of proteins $S=\{g_1, g_2, \ldots, g_k\}$, let e_i denote the mRNA expression level of $g_i \in S$. Then the *subnetwork activity* of S is defined as $e_S = \sum_{i=1}^{k} e_i/\sqrt{k}$, that is the aggregate mRNA-level expression of the proteins in the sub-network. Subsequently, mutual information is used to measure the dependence of two discrete random variables: in this case the health status *vs.* subnetwork activity of S. Denoting health status vector as c (i.e., $c(j)$ denotes the health status of the jth sample) and quantized subnetwork activity of S as $\hat{e}_S$, (i.e., $\hat{e}_S(j)$ denotes the aggregate expression of the gene products in S in the jth sample), the dysregulation of S is defined as $I(c,\hat{e}_S)=H(c)-H(c|\hat{e}_S)$. Here, H(c) denotes the Shannon entropy of random variable c (that is the uncertainty on the health status of a sample) and $H(c|\hat{e}_S)$ denotes the entropy of random variable c after the observation of random variable $\hat{e}_S$ (that is the uncertainty on the health status of a sample given the subnetwork activity of S in that sample). Consequently, the mutual information (MI) $I(c,\hat{e}_S)$ is a measure of the expression levels of all genes in the subnetwork in discriminating OSA patients from control. To this end, a high MI score for a sub-network is an indicator of the coordinate mRNA-level dysregulation of the proteins in the subnetwork, i.e., although the gene coding for each protein in the sub-network may not be significantly differential expressed in OSA, the total mRNA-level expression of these proteins exhibits significant difference between OSA patients and control. This information theoretic formulation of coordinate dysregulation has been shown to be effective in identification of subnetwork markers that were powerful in prediction of breast and colon cancer metastasis [29, 51].

While Chuang et al. originally used a greedy algorithm to identify subnetworks with high MI [29], we exhaustively searched for subnetworks of the IPA network to identify sets of genes with high MI. This is because the network obtained from IPA analysis is already filtered to obtain a concise network of proteins that are functionally associated with proteins that are already known to play a role in sleep apnea. Consequently, an exhaustive search for reasonably sized subnetworks (we search for subnetworks composed of up to 6 proteins in this study) is feasible on this network, which is guaranteed to find all subnetworks with a maximum MI, as opposed to a greedy algorithm[20].

2.4 *Analyzing adipocyte interactome using SNP association scores from GWAS*

The Candidate Gene Association Resource (CARe) project initiated by the National Heart, Lung, and Blood Institute, conducted analyses of genetic variation in cardiovascular, pulmonary, hematological, and sleep-related traits in nine community-based cohorts[21]. Polysomnography data, providing objective measurements of OSA, were only available for a subset of these cohorts, and of these, a genome-wide assay (Affymetrix 6.0) was only performed in the African American participants (n=647) in the Cleveland Family Study, which provided p-values for the associations between 867,496 SNPs with OSA (defined as an AHI > 15 for identifying cases).

We then map these p-values to proteins/nodes in the adipose tissue-specific interactome map as follows. For each protein g_i in the network, the most significant p-value that is associated with a SNP located in the coding region of g_i is designated as the p-value of the association of g_i with OSA. In other words, letting *p(s)* denote the p-value of the association of SNP s with OSA, we define $p_i = \min_{s \in R_i} p(s)$ where R_i denotes the set of SNPs that reside within the coding region of $g_{i.}$[30,31]. Subsequently, we apply a Cytoscape tool, jactivemodule, to extract sub-networks of the adipose tissue-specific interactome map that are enriched in proteins with high total significance of association with OSA[31].

jactivemodule is a subnetwork search algorithm that was originally developed to identify *active subnetworks* in a network of interactions, where an active subnetwork refers to a connected subgraph of the interactome that has high total significance of differential mRNA-level expression with respect to a particular perturbation[32]. It takes as input p-values associated with each protein in the network, converts these p-values to z-scores (so that a higher z-score indicates more significant differential expression), and greedily identifies subnetworks with high aggregate *z-score.* More precisely, the score of a subnetwork $S=\{g_1, g_{2,} \ldots, g_k\}$ is defined as $A(S) = \sum_{i=1}^{k} z_i/\sqrt{k},$ where z_i denotes the z-score corresponding to p-value $p_{i.}$

Although this method was originally developed to identify differentially expressed subnetworks, it can as well be used to identify disease-associated subnetworks since the p-values of differential expression can be replaced by p-values of association with the disease. Motivated by this insight, we use this algorithm to identify subnetworks that are implicated in OSA by GWAS. Observe also that, a high-scoring sub-network is not necessarily one that is enriched in proteins with very significant p-values, but it can also be comprised of many proteins with moderately significant p-values. Consequently, this method has the potential of uncovering groups of proteins that exhibit seemingly insignificant association with OSA when considered individually, but exhibit strong association when considered together. Since such subnetworks are connected by a network of interactions by the construction of the algorithm, they are likely to be functionally associated and therefore might be underlying a potential genetic interaction that underlies the manifestation of the disease. The details of procedure can be found at the documentation of cytoscape (www.cytoscape.org) and in the literature[33].

Finally, MCODE and BiNGO were applied to analyze the sub-networks detected, e.g., detecting the functional modules and identifying the enrichment of GO category[34,35].

3. Results and Discussion

3.1 *Generating and analyzing networks from seed genes/proteins*

We used IPA to generate networks using the 56 seed proteins related to sleep disorders. The top two scoring networks were used for further analysis. Among 70 proteins in each network, network 1 (Left figure in Figure 1) contains 32 seed proteins. The enriched functions for this network as identified by IPA include neurological disease, organismal injury and abnormalities, and genetic disorders. Network 2 (Right figure of Fig. 1) includes 16 seeds, and the associated functions are genetic disorder, neurological disease, and respiratory disease.

A quantitative method to detect and score sub-networks within the networks was applied to identify sub-networks that are highly discriminative for the OSA phenotype based on transcriptional dysregulation using mRNA expression data from subcutaneous and visceral fat tissues[20]. To limit the computational overhead of the calculation while using exhaustive search, we constrained the search where sub-networks were limited to six nodes. This analysis of network 1 provided 108 sub-networks of 6 nodes using expression data from subcutaneous fat tissue, and 97 sub-networks of 6 nodes from visceral tissue that had the maximum possible values of MI. In case of network 2, 9 sub-networks are detected for subcutaneous tissue, and 8 for visceral tissue. Further analyses focus on these sub-networks.

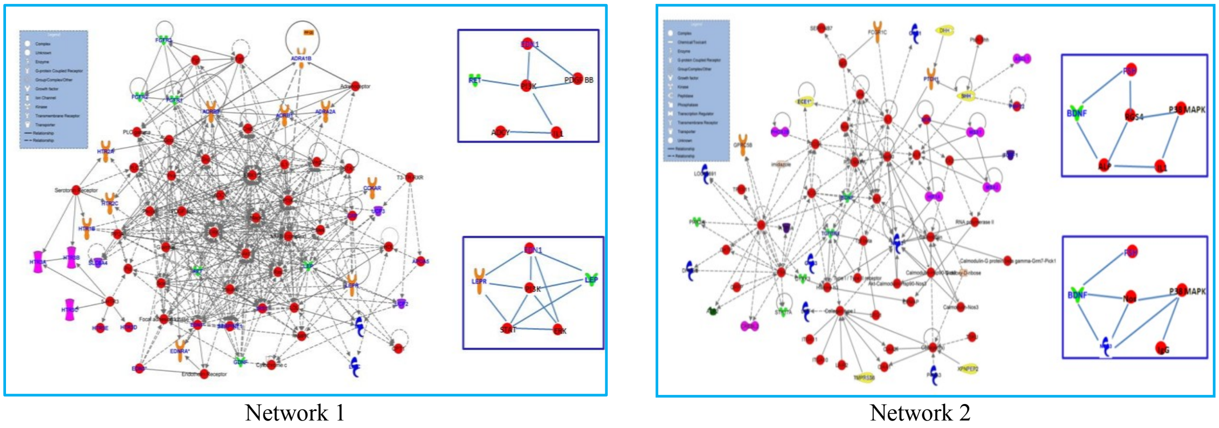


Fig. 1 Networks generated using IPA with highest score (proteins name in blue indicates seed proteins), subnetworks with 6 nodes are identified by MI scores for subcutaneous and visceral fat tissues. Larger and high resolution picture can be found at http://proteomics.case.edu/news_events. aspx?newsid=38

In order to analyze the sub-networks, we calculated the frequency of occurrence of proteins in these sub-networks. We assume that the proteins that appear most frequently will likely be significant in terms of defining differences between the OSA phenotype and control. To reduce the incidence of false positives, we focused on the proteins that are in the top 6 in frequency for both tissues, which are listed in Table 1.

Table 1a Protein detected in subnetworks from network 1(Figure 1) and its frequency in the exhaustive search

Protein (subcutaneous fat)	Frequency	Probability in detected subnetwork*	Protein (visceral fat)	Frequency	Probability in detected subnetwork*
PDGF BB	55	50.9%	ERK	53	54.6%
EDN1	43	39.8%	EDN1	34	35.0%
IL1	43	39.8%	STAT	31	31.9%
PI3K	38	35.1%	PI3K	26	26.8%
RET	27	25.0%	LEP	26	26.8%
ADCY	25	23.1%	LEPR	24	24.7%

Table 1b Protein detected in subnetworks from network 2 (Figure 1) and its frequency in the exhaustive search

Protein (subcutaneous fat)	Frequency	Probability in detected subnetwork*	Protein (visceral fat)	Frequency	Probability in detected subnetwork*
P38 MAPK	5	55.6%	BDNF	4	50.0%
RGS4	4	44.4%	P38 MAPK	4	50.0%
FSH	4	44.4%	NOS3	3	37.5%
BDNF	4	44.4%	FSH	3	37.5%
IL1	3	33.3%	Nos	3	37.5%
ALP	3	33.3%	IgG	3	37.5%

* Calculated by Frequency/(total number of sub-networks with maximum MI)

Notably, 14 out of 24 proteins in table 1 are not seed proteins, and potentially indicate novel findings discovered by our approach. A number of proteins listed in table 1 are associated with OSA or other sleep phenotypes based on previous studies. For example, Endothelin 1 (EDN1), a potent vasoconstrictor implicated in hypertension, is both a seed protein and is ranked as second most frequent node for both tissues in the network 1 analysis (Table 1). Studies using knockout mice show that EDN1 is associated with respiratory distress[36], and more recently, association studies suggests that a missense coding SNP in EDN1 is linked with OSA in a European American sample[19]. The phosphorylation of ERK (Extracellular Signal-Regulated Kinase), the most frequently identified protein in visceral fat from network 1, is correlated with sleep patterns in flies[37]. PDGF BB (subunit of platelet-derived growth factor) the most frequently identified protein in subcutaneous fat from network 1, is a growth factor that regulates cell growth and division. There is evidence for the role of PDGF BB in disordered breathing from the responses of rats to hypoxia[38-40]. Follicle-stimulating hormone (FSH), seen in both fat analyses of network 2, is a hormone found in humans and other animals. Recent studies show that the concentration of FSH has a significant correlation with the obstructive apnea index in cerebrospinal fluid.[41]

Aside from many proteins that are directly related to OSA or sleep, the sub-networks also contain proteins that are known to be involved in processes related to sleep, but have not been reported to have specific associations with OSA. P38 MAPK (a frequently observed sub-network member from the analysis of network 2) is a member of the mitogen-activated protein kinases (MAPK) that play crucial roles in signaling the inflammatory response and are involved in pathways that respond to oxidative stress[42,43]. As indicated above, both processes are known to be related to OSA[4]. Another protein, Phosphatidylinositol 3-kinases (PI3K) is ranked in the top four in both tissues (Table 1, network 1). PI3Ks are a group of lipid kinases that catalyze the phosphorylation of phosphatidylinositols and phosphoinositides. They are composed of one 85 kDa regulatory subunit and one 110 kDa catalytic subunit. PIK3R genes (such as PIK3R1, PIK3R2, PIK3R3, PIK3R5, etc), encode the p85 regulatory subunit, while PIK3C genes (such as PIK3C3, PIK3CA, PIK3CB, PIK3CD, etc), code for the p110 catalytic subunit. It has been reported that PI3K is associated with fatty acid-induced insulin resistance[44], and although OSA and insulin resistance may be causally related, the exact mechanism linking them has not been fully elucidated [4]. Another top gene, STAT, encodes a family of transcription factors. In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases, and then translocated to the cell nucleus where they act as transcription activators. In a recent report, STAT4 was found to be involved in metabolic processes, especially in insulin resistance and inflammation in adipose tissue.[45]

3.2 *Analyzing interactome in adipose*

An interactome relevant to adipose tissue was generated from a combination of public interaction databases and gene expression profiles and contains 2909 proteins and 8323 interactions (Supplementary Table 2). Analyses of the topological parameters of the network show that it possesses typical properties of realistic networks,[46,47] such as small-world properties (the average shortest path length is 4.5). The node degree distribution fits a power law distribution.

We searched this interactome for OSA related sub-networks by mapping p-values from the GWAS study to proteins of the interactome [21]. Then, cytoscape and its plugin jactivemodule are applied to detect sub-networks that are significant. The jactivemodule combines the network structure and associated p-value of each protein to extract potential meaningful sub-networks. A subnetwork with 203 proteins and 324 interactions is identified with a significant score (7.09, Figure 2, subnetworks with score > 3.0 are considered as significant[32]). Similar to the whole interactome, this sub-network shows some typical properties, such as small-world and power-law distribution of node degree. Note that many of the nodes have modest p-values (low z-scores), and would not be seen as significant in a conventional GWAS analysis. For example, the p-value of hepatocyte growth factor-regulated tyrosine kinase substrate (HGS) is 0.62, but its interacting partners (neurofibromin 2 (NF2), signal transducing adaptor molecule (STAM and, STAM2)) have p-value less than 0.006, thus, it is included in the subnetwork. Other similar examples are minichromosome maintenance complex component 7 (MCM7, p-value: 0.97) and SHC (Src homology 2 domain containing) transforming protein 1 (SHC1, p-value: 0.35).

Another cytoscape plugin MCODE was applied to explore the protein complexes or other modules present in the sub-network identified by jactivemodule. MCODE detects densely connected regions in a network that may represent functional modules. It is based on vertex weighting by local neighborhood density and outward traversal from a locally dense seed protein to isolate the dense regions. The top two clusters identified by MCODE are listed in Figure 3. The cluster with best score has ten proteins that are densely connected. Nine out of ten components are proteasome subunits. The proteasome is a large, multimeric protein complex with regulatory and catalytic functions. It is responsible for degrading damaged, misfolded, nonfunctional and potentially toxic proteins. Notably, it has been reported that the proteasome pathway and proteasomal activity are associated with OSA and hypoxia, a central feature of OSA [48,49].

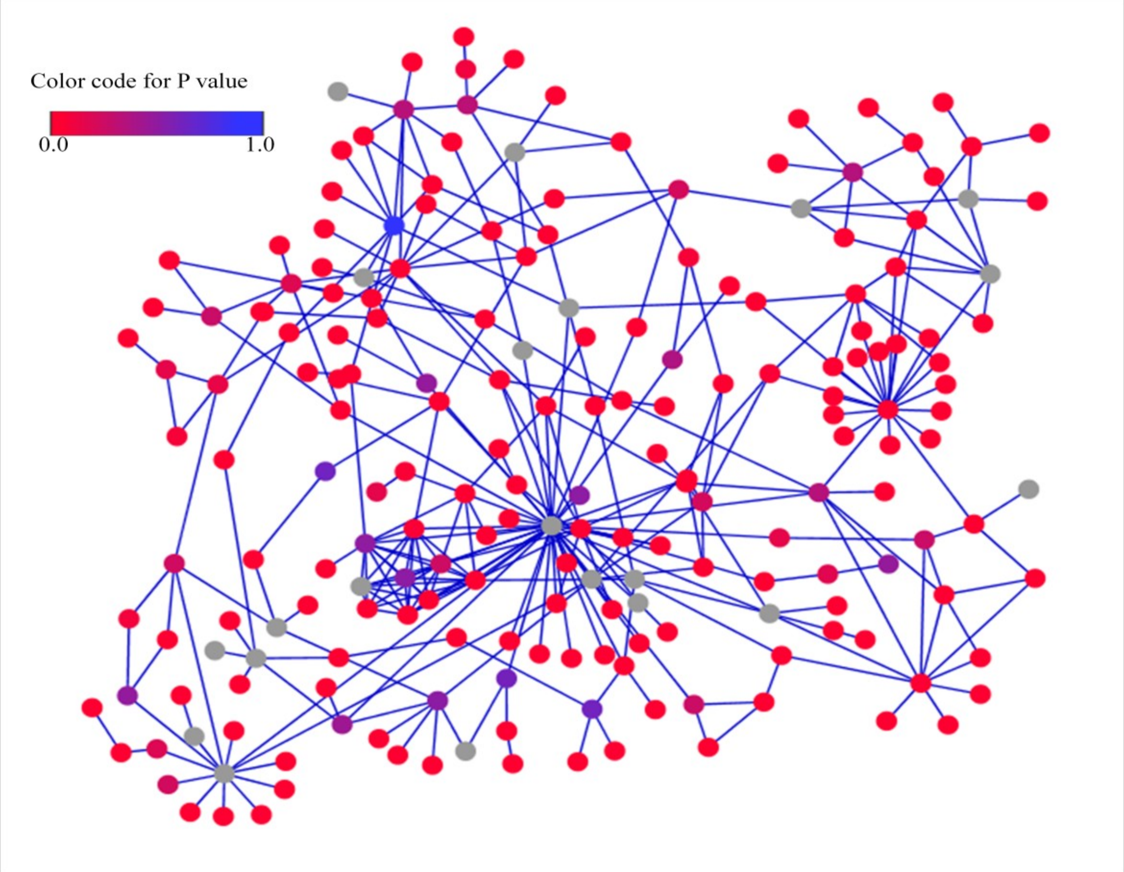


Figure 2 Network identified by jactivemodule using p-values from GWAS study, color represents the p-values and nodes with grey color indicate that the p-values are missing from GWAS. High resolution picture with the node lable can be found at http://proteomics.case.edu/news_events. aspx?newsid=38

To determine which Gene Ontology (GO) functional categories are statistically overrepresented in the sub-network, we further applied the BiNGO program to the sub-network of Figure 2. The detected functions include axon extension, spliceosome assembly, protein catabolic process, insulin receptor signaling pathway, and negative regulation of tyrosine phosphorylation of STAT3 proteins. Recent studies suggest that STAT3 tyrosine phosphorylation is critical for interleukin protein production in the inflammatory response [45]. Also, STAT family members are implicated in several processes relevant to tumor growth, providing an additional link aside from PI3K between OSA and cancer.

As there is an association between OSA and diabetes[50], the functional enrichment for the insulin receptor-signaling pathway deserves closer investigation. Three proteins in the sub-network are responsible for the enrichment of this function: PIK3R1, IRS2, and IGF1R. PIK3R1 (phosphoinositide-3-kinase, regulatory subunit 1) phosphorylates the inositol ring of phosphatidylinositol at the 3-prime position and plays an important role in the metabolic actions of insulin; IRS2 (insulin receptor substrate 2) mediates effects of insulin by acting as a molecular adaptor between diverse receptor tyrosine kinases and downstream effectors; IGF1R (insulin-like growth factor 1 receptor) binds insulin-like growth factor with a high affinity and modulates insulin's actions.. Notably, these three proteins plus YWHAG (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide) densely connect, forming a cluster in the subnetwork that is also detected by MCODE (Figure 3).

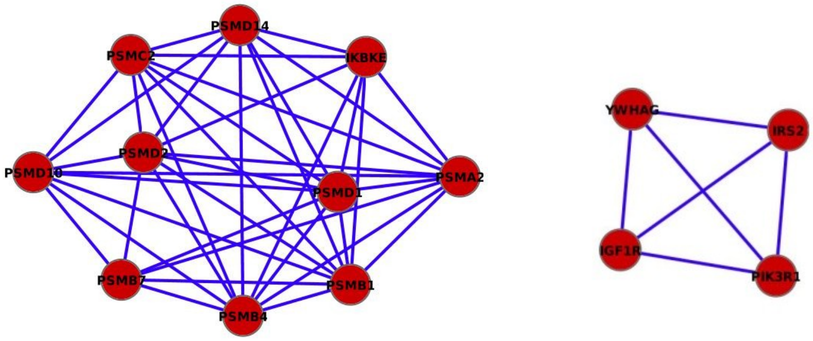


Figure 3 Densely connected subnetworks identified using MCODE, those represent potentially functional module or protein complex

3.3 *Comparison and limitation of approaches*

In this study, we took two systems biology approaches to detect subnetworks which are likely associated with OSA. Because of the nature of two approaches (the first one is biased and based on prior knowledge of OSA; the second one is unbiased), it is hard to compare them, and it is not surprising that the results are different. These two approaches use SNP data from GWAS and gene expression data from microarray experiments respectively, and treat them independently. Also the data are from two different sources (SNP data derived from CARe project[21], and gene expression data from other sources (Patel, S, et al, unpublished data).

One limitation of our approach is that the method for detection of subnetworks using MI is computationally extensive, and can only be applied on small networks. Further efforts are necessary to improve its efficacy. Another limitation is the method to derive the significance level of proteins based on the SNP data. Usually, there are multiple SNPs located within the regions for each gene. Although several methods have been proposed to condense this informaiton[30,31,52], we applied the simple and most commonly used one: consider the most significant p-value among SNPs as p-value of proteins as other methods may provide conflicting results.

4. Conclusion

Our integrated analysis of mRNA expression from adipose tissues, PPI networks, and SNP data from genome-wide association studies provides a novel approach for combining data from disparate sources to identify candidate pathways for potential validation studies. Some of the associations identified may reflect pathways that predispose to OSA, while others may indicate pathways that are perturbed by OSA-related stresses which contribute to co-morbidities such as diabetes. The results of this initial study suggest that the PI3K, the STAT protein family, and insulin signaling may be associated with OSA. Further investigation is needed to elucidate the exact role of these genes and their gene products in OSA. In addition, our approach outlines a novel application of SNP data in sub-network discovery relevant to disease that is consistent with other well-accepted methodologies. Thus, we suggest this approach could be generally applied to the analysis of GWAS data that is available for over 100 other diseases.

5. Acknowledgments

This research is supported by the Case Center for Transdisciplinary Research on Energetics and Cancer, U54-CA-116867 (PI: Berger, Nate) and Case Comprehensive Cancer Center, P30-CA-043703 (PI: Gerson, Stanton), from the National Cancer Institute, UL1-RR024989 (PI: Davis, Pamela) from the National Center for Research Resources (Clinical and Translational Science Awards), Case Proteomics Center for HIV/AIDS and Drug Abuse, P20-DA26113 (PI: Chance, Mark) from the National Institute for Drug Abuse, CFS HL 463680 (PI: Redline, Susan) and by NSF CAREER Award CCF-0953195 (PI: Koyuturk, Mehmet). NHLBI's Candidate Gene Association Resource has supported extensive genotyping and has created a genotype-phenotype database using data and samples from nine cohorts, including the Atherosclerosis Risk in Communities (ARIC) study, the Cardiovascular Health Study (CHS), the Cleveland Family Study, the Cooperative Study of Sickle Cell Disease (CSSCD), the Coronary Artery Risk Development in Young Adults (CARDIA) study, the Framingham Heart Study (FHS), the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Sleep Heart Health Study (SHHS). Detailed information on the grants and contracts that have supported CARe and its parent cohorts can be found at http://public.nhlbi.nih.gov/GeneticsGenomics/home/care.aspx.

References

1 Reite, M., Ruddy, J. & Nagel, K. *Concise guide to evaluation and management of sleep disorders*. (American Psychiatric Publishing, Inc., 2002).
2 Vgontzas, A. N. & Kales, A. *Annu Rev Med* **50**, 387-400, (1999).
3 Young, T., Peppard, P. E. & Gottlieb, D. J. *Am J Respir Crit Care Med* **165**, 1217-1239 (2002).
4 Arnardottir, E. S., et al. *Sleep* **32**, 447-470 (2009).
5 Waradekar, et al. *Am J Respir Crit Care Med* **153**, 1333-1338 (1996).
6 Donadio, V. *et al. J Sleep Res* **16**, 327-332, (2007).

7 Bravo Mde, L. *et al. Sleep Breath* **11**, 177-185 (2007).

8 Minoguchi, K. *et al. Am J Respir Crit Care Med* **172**, 625-630, (2005).

9 Schulz, R. *et al. Am J Respir Crit Care Med* **162**, 566-570 (2000).

10 Dyugovskaya, L., Lavie, P. & Lavie, L. *Am J Respir Crit Care Med* **165**, 934-939 (2002).

11 Young, T. *et al. N Engl J Med* **328**, 1230-1235 (1993).

12 Ip, M. S. *et al. Am J Respir Crit Care Med* **165**, 670-676 (2002).

13 McNicholas, W. T. & Bonsigore, M. R. *Eur Respir J* **29**, 156-178, (2007).

14 Logan, A. G. *et al. Eur Respir J* **21**, 241-247 (2003).

15 Strohl, K. P., Saunders, N. A., Feldman, N. T. & Hallett, M. *N Engl J Med* **299**, 969-973 (1978).

16 Redline, S. *et al. Am J Respir Crit Care Med* **151**, 682-687 (1995).

17 Riha, R. L. *Respiration* **78**, 5-17, (2009).

18 Larkin, E. K. *et al. Ann Hum Genet* **72**, 762-773,(2008).

19 Larkin, E. K. *et al. Am. J. Respir. Crit. Care Med.* in press (2010).

20 Nibbe, R. K. *et al. Mol Cell Proteomics* **8**, 827-845, (2009).

21 Musunuru, K. *et al. Circ Cardiovasc Genet*, in press (2010).

22 Bossi, A. & Lehner, B. *Mol Syst Biol* **5**, 260, (2009).

23 Wang, E. T. *et al. Nature* **456**, 470-476, (2008).

24 Szabo, P. M. *et al. Oncogene* **29**, 3163-3172, (2010).

25 Su, A. I. *et al. Proc Natl Acad Sci U S A* **99**, 4465-4470, (2002).

26 Su, A. I. *et al. Proc Natl Acad Sci U S A* **101**, 6062-6067, (2004).

27 Assenov, Y. et al. *Bioinformatics* **24**, 282-284, (2008).

28 Bengtsson, H., et al. (Dep. of Statistics, Univ. of California, Berkeley, February 2008).

29 Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D. & Ideker, T. *Mol Syst Biol* **3**, 140, (2007).

30 Wang, K., Li, M. & Bucan, M. *Am J Hum Genet* **81**, (2007).

31 Baranzini, S. E. *et al. Hum Mol Genet* **18**, 2078-2090, (2009).

32 Ideker, T., et al. *Bioinformatics* **18 Suppl 1**, S233-240 (2002).

33 Cline, M. S. *et al. Nat Protoc* **2**, 2366-2382, (2007).

34 Bader, G. D. & Hogue, C. W. *BMC Bioinformatics* **4**, 2 (2003).

35 Maere, S., Heymans, K. & Kuiper, M. *Bioinformatics* **21**, 3448-3449, (2005).

36 Kuwaki, T. *et al. Am J Physiol* **270**, R1279-1286 (1996).

37 Foltenyi, K., Greenspan, R. J. & Newport, J. W. *Nat Neurosci* **10**, 1160-1167, (2007).

38 Alea, O. A. *et al. Am J Physiol Regul Integr Comp Physiol* **279**, R1625-1633 (2000).

39 Vlasic, V., Simakajornboon, N., Gozal, E. & Gozal, D. *Pediatr Res* **50**, 236-241 (2001).

40 Gozal, D. *et al. J Neurochem* **74**, 310-319 (2000).

41 Capatina, C., et al. *Endocrine Abstracts* **22**, P631 (2010).

42 Ryan, S., et al. *Biochem Biophys Res Commun* **355**, 728-733, (2007).

43 Hu, J. Y. *et al. Cell Mol Life Sci* **67**, 321-333, (2010).

44 Kruszynska, Y. T. *et al. J Clin Endocrinol Metab* **87**, 226-234 (2002).

45 Samavati, L. *et al. Mol Immunol* **46**, 1867-1877, (2009).

46 Uetz, P. *et al. Nature* **403**, 623-627, (2000).

47 Ito, T. *et al. Proc Natl Acad Sci U S A* **97**, 1143-1147 (2000).

48 Gozal, D. *et al. J Neurochem* **86**, 1545-1552, (2003).

49 Taylor, C. T., et al. *Proc Natl Acad Sci U S A* **97**, 12091-12096, (2000).

50 Shaw, J. E., et al. *Diabetes Res Clin Pract* **81**, 2-12, (2008).

51 Chowdhury S.A., Nibbe, R.K., Chance, M.R., Koyuturk, M. *Proc. RECOMB*, LNCS 6044, 80-95, (2010)

52 Yu, K, et al. *Genet Epidemiol*. 33(8):700-9, (2009)

FINDING MOST LIKELY HAPLOTYPES IN GENERAL PEDIGREES THROUGH PARALLEL SEARCH WITH DYNAMIC LOAD BALANCING

LARS OTTEN and RINA DECHTER

Bren School of Information and Computer Sciences
University of California, Irvine, CA 92697, U.S.A.
{lotten,dechter}@ics.uci.edu

General pedigrees can be encoded as Bayesian networks, where the common MPE query corresponds to finding the most likely haplotype configuration. Based on this, a strategy for grid parallelization of a state-of-the-art Branch and Bound algorithm for MPE is introduced: independent worker nodes concurrently solve subproblems, managed by a Branch and Bound master node. The likelihood functions are used to predict subproblem complexity, enabling efficient automation of the parallelization process. Experimental evaluation on up to 20 parallel nodes yields very promising results and suggest the effectiveness of the scheme, solving several very hard problem instances. The system runs on loosely coupled commodity hardware, simplifying deployment on a larger scale in the future.

1. Introduction

Given a general pedigree expressing ancestral relations over a set of individuals, the haplotyping problem is to infer the most likely ordered haplotypes for each individual from measured unordered genotypes. This has previously been cast as solving an optimization problem over a appropriately constructed Bayesian network,[6] for which powerful inference algorithms can be exploited. Yet practical problems remain infeasible as more data becomes available, for example through SNP sequencing, suggesting a shift to parallel or distributed computation.

This paper therefore explores parallelization of combinatorial optimization tasks over such Bayesian networks, which are typically generalized through the framework of graphical models. Specifically, we consider one of the best exact search algorithms for solving the MPE/MAP task over graphical models, AND/OR Branch and Bound (AOBB). AOBB, which exploits independencies and unifiable subproblems, has demonstrated superior performance for these tasks compared with other state-of the art exact solvers (e.g., it was ranked first or second in several competitions[13]).

To parallelize AOBB we use the established concept of parallel tree search[8] where the search space is explored centrally up to a certain depth and the remaining subtrees are solved in parallel. For graphical models this can be implemented straightforwardly by exploring the search space of partial instantiations up to a certain depth and solving the remaining conditioned subproblems in parallel. This approach has already proven successful for likelihood computation in Superlink-Online, which parallelizes cutset conditioning for linkage analysis tasks.[16] Our work differs in focusing on optimization (e.g., MPE/MAP) and in exploiting the AND/OR paradigm, leveraging additional subproblem independence for parallelism. Moreover, we use the power of Branch and Bound in a central search space that manages (and prunes) the set of conditioned subproblems.

The main difference however is that, compared to likelihood computation, optimization presents far greater challenges with respect to load balancing. Hence the primary challenge in search tree parallelization is to determine the "cutoff", the *parallelization frontier*. Namely, we need a mechanism to decide when to terminate a branch in the central search space and send the corresponding subproblem to a machine on the network. There are two primary issues: *(1)* Avoid *redundancies*: caching of unifiable subproblems is lost across the independently solved subproblems, hence some work might be duplicated; *(2)* Maintain *load balancing* among the grid resources, dividing the total work equally and without major idling periods. While introducing redundancy into the search space can be counterproductive for both tasks, load balancing is a far greater challenge for optimization, since the cost function is exploited in pruning the search space. Capturing this aspect is essential in predicting the size of a subproblem and thus the focus of this paper.

The contribution of this work is thus as follows: We suggest a parallel BaB scheme in a graphical model context and analyze some of its design trade-offs. We devise an estimation scheme that predicts the size of future subproblems based on cost functions and learns from previous subproblems to predict the extent of BaB pruning within future subproblems. We show that these complexity estimates enable effective load distribution (which was not possible via redundancy analysis only), and yield very good performance on several very hard practical problem instances, some of which were never solved before. Our approach assumes the most general master-worker scenario with minimal communication and can hence be deployed on a multitude of grid setups spanning hundreds, if not thousands of computers worldwide. While our current empirical work is tested on up to 20 machines so far, its potential for scaling up are very promising.

Related work: The idea of parallelized Branch and Bound in general is not new, but existing work often assumes a shared-memory architecture or extensive inter-process communication,[3,7,8] or specific grid hierarchies.[1] Earlier results on estimating the performance of search predict the size of general backtrack tress through random probing.[10,12] Similar schemes have been devised for Branch and Bound algorithms, where the algorithm is ran for a limited time and the partially explored tree is extrapolated.[4] Our method, on the other hand, is not sampling-based but only uses parameters available a priori and information learned from past subproblems which is facilitated through the use of depth-first branch and bound to explore the master search space.

2. Background

Our approach is based on the general framework of graphical model reasoning:

Definition 2.1 (graphical model). *A* graphical model *is given as a set of variables $X = \{X_1, \ldots, X_n\}$, their respective finite domains $D = \{D_1, \ldots, D_n\}$, a set of cost functions $F = \{f_1, \ldots, f_m\}$, each defined over a subset of X (the function's* scope*), and a combination operator (typically sum, product, or join) over functions. Together with a marginalization operator such as $\min_X$ and $\max_X$ we obtain a* reasoning problem*.*

For instance, the *MPE* problem (most probable explanation) is typically posed over a Bayesian Network structure, representing the factorization of a joint distribution into conditional probabilities, with the goal of finding an assignment with maximum probability. In the area of constraint reasoning, a *weighted CSP* is defined as minimizing the sum of a set of cost functions over the variables.

Definition 2.2 (primal graph, induced graph, induced width). *The* primal graph *of a graphical model is an undirected graph, $G = (X, E)$. It has the variables as its vertices and an edge connecting any two variables that appear in the scope of the same function. Given an undirected graph G and an ordering $d = X_1, \ldots, X_n$ of its nodes, the width of a node is the number of neighbors that precede it*

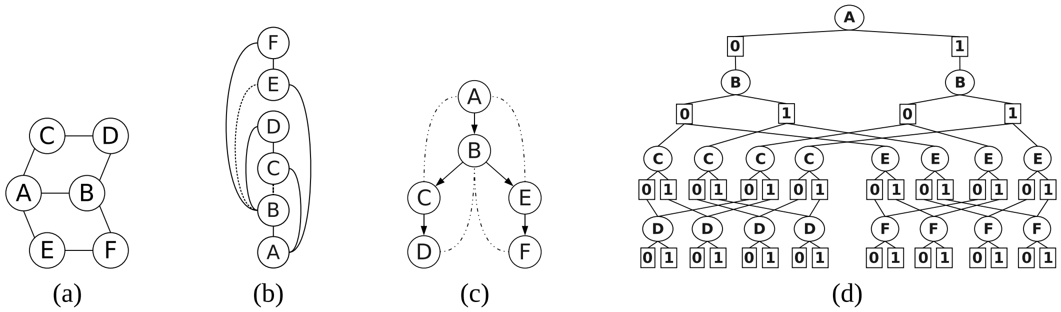


Fig. 1: (a) Example primal graph with six variables, (b) its induced graph along ordering $d = A, B, C, D, E, F$, (c) a corresponding pseudo tree, and (d) the resulting context-minimal AND/OR search graph.

in d*. The* induced graph G' *of* G *is obtained as follows: from last to first in* d*, each node's preceding neighbors are connected to form a clique (where new edges are taken into account when processing the remaining nodes). The* induced width w^* *is the maximum width over all nodes in the induced graph along ordering* d*.*

Figure 1(a) depicts the primal graph of an example problem with six variables. The induced graph for the example problem along ordering $d = A, B, C, D, E, F$ is depicted in Figure 1(b), its induced width is 2. Note that different orderings will vary in their implied induced width; finding an ordering of minimal induced width is known to be NP-hard, in practice heuristics like *minfill*[11] are used to obtain approximations.

2.1. *Encoding Pedigrees as Bayesian Networks*

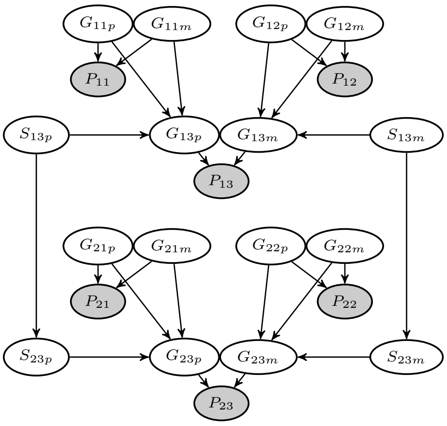


Fig. 2: Example fragment of a Bayesian network encoding of a general pedigree.

Expressing a particular pedigree as a Bayesian Network utilizes three building blocks: (1) For each individual and each locus, the two haplotypes are represented by two variables, with the possible alleles as their domain and a probability distribution conditioned on the variables representing the parents' haplotypes at this locus. (2) The measured, unordered genotypes are captured as phenotype variables, which are conditioned on the corresponding pair of haplotypes. (3) Auxiliary binary selector variables are linked across loci, to capture recombination events.

Figure 2 shows a simple example of such a Bayesian network, the displayed fragment includes three individuals (two parents and their child) and two loci. For instance, G_{13p} is the paternal haplotype of individual 3 (the child) at locus 1. It depends on the father's haplotypes G_{11p} and G_{11m}, where the inheritance is determined by the selector variable S_{13p} i.e., $G_{13p} = G_{11p}$ if $S_{13p} = 0$ and $G_{13p} = G_{11m}$ if $S_{13p} = 1$. Together with the maternal haplotype G_{13m}, G_{13p} determines the genotype in P_{13}. The value of the inheritance selector S_{23p} for the paternal haplotype of individual 3 at locus 2 is dependent on the selector S_{13p} for locus 1, where the actual probabilities are recombination fractions between these two loci, provided as input.

With this construction, the joint distribution of the Bayesian network captures the probability over all haplotype configurations. Given a set of evidence (i.e., measurements for some or all of the unordered genotypes), the solution to the common problem of finding the most probable explanation (MPE) will yield the most likely haplotypes.[6]

2.2. *AND/OR Search Spaces*

The concept of AND/OR search spaces has been introduced as a unifying framework for advanced algorithmic schemes for graphical models to better capture the structure of the underlying graph.[5] Its main virtue consists in exploiting conditional independencies between variables, which can lead to exponential speedups. The search space is defined using a *pseudo tree*, which captures problem decomposition:

Definition 2.3 (pseudo tree). *Given an undirected graph $G = (X, E)$, a* pseudo tree *of G is a directed, rooted tree $\mathcal{T} = (X, E')$ with the same set of nodes X, such that every arc of G that is not included in E' is a back-arc in $\mathcal{T}$, namely it connects a node in $\mathcal{T}$ to an ancestor in $\mathcal{T}$. The arcs in E' may not all be included in E.*

AND/OR Search Trees : Given a graphical model instance with variables X and functions F, its primal graph (X, E), and a pseudo tree $\mathcal{T}$, the associated *AND/OR search tree* consists of alternating levels of OR and AND nodes. OR nodes are labeled X_i and correspond to the variables in X. AND nodes are labeled $\langle X_i, x_i \rangle$, or just x_i and correspond to the values of the OR parent's variable. The structure of the AND/OR search tree is based on the underlying pseudo tree $\mathcal{T}$: the root of the AND/OR search tree is an OR node labeled with the root of $\mathcal{T}$. The children of an OR node X_i are AND nodes labeled with assignments $\langle X_i, x_i \rangle$ that are consistent with the assignments along the path from the root; the children of an AND node $\langle X_i, x_i \rangle$ are OR nodes labeled with the children of X_i in $\mathcal{T}$, representing conditionally independent subproblems. It was shown that, given a pseudo tree $\mathcal{T}$ of height h, the size of the AND/OR search tree based on $\mathcal{T}$ is $\mathcal{O}(n \cdot k^h)$, where k bounds the domain size of variables.[5]

AND/OR Search Graphs : Different nodes may root identical and can be merged through *caching*, yielding an *AND/OR search graph* of smaller size, at the expense of using additional memory during search. A mergeable node X_i can be identified by its *context*, the partial assignment of the ancestors of X_i which separates the subproblem below X_i from the rest of the network. Merging all context-mergeable nodes yields the *context minimal* AND/OR search graph.[5]

Proposition 2.1. *Given a graphical model, its primal graph G, and a pseudo tree $\mathcal{T}$, the size of the context-minimal AND/OR search graph is $\mathcal{O}(n \cdot k^{w^*})$, where w^* is the induced width of G over a depth-first traversal of $\mathcal{T}$ and k bounds the domain size.*

Example 2.1. Figure 1(c) depicts a pseudo tree extracted from the induced graph in Figure 1(b) and Figure 1(d) shows the corresponding context-minimal AND/OR search graph. Note that the AND nodes for B have two children each, representing independent subproblems and thus demonstrating problem decomposition. Furthermore, the OR nodes for D (with context $\{B, C\}$) and F (context $\{B, E\}$) have two edges converging from the AND level above them, signifying caching.

Weighted AND/OR Search Graphs : Given an AND/OR search graph, each edge from an OR node X_i to an AND node x_i can be annotated by *weights* derived from the set of cost functions F in the graphical model: the weight $l(X_i, x_i)$ is the sum of all cost functions whose scope includes X_i and is fully assigned along the path from the root to x_i, evaluated at the values along this path. Furthermore, each node in the AND/OR search graph can be associated with a *value*: the value $v(n)$ of a node n is the minimal solution cost to the subproblem rooted at n, subject to the current variable instantiation along the path from the root to n. $v(n)$ can be computed recursively using the values of n's successors.[5]

2.3. *AND/OR Branch and Bound*

AND/OR Branch and Bound is a state-of-the-art algorithm for solving optimization problems over graphical models. Assuming a minimization task, it traverses the context-minimal AND/OR graph in a depth-first manner while keeping track of a current upper bound on the optimal solution cost. It interleaves forward node expansion with a backward cost revision or propagation step that updates node values (capturing the current best solution to the subproblem rooted at each node), until search terminates and the optimal solution has been found.[5]

3. Setup and Parallel Scheme

We assume a very general parallel framework in which autonomous hosts are loosely connected over some network – in our case we use ten dual-core desktop computers, with CPU speeds between 2.33 and 3.0 GHz, on a local Ethernet, thus allowing experiments with up to 20 parallel nodes. We impose a *master-worker* hierarchy on the computers in the network, where a special *master* node runs a central process to coordinate the *workers*, which cannot communicate with each other. This general model is chosen to accommodate a wide range of parallel resources, where direct node communication is often either prohibitively slow or entirely impossible; it also facilitates flexible deployment on geographically dispersed, heterogeneous resources in the future.

The setup is similar to Superlink-Online,[16] which has been very successful in using large-scale parallelism in likelihood algorithms for genetic linkage analysis, or SETI@home,[2] which uses Internet-connected PCs around the world to search through enormous amounts of radio data. Like Superlink-Online, our system is implemented on top of the *Condor* grid middleware.[17]

3.1. *Parallel AND/OR Branch and Bound*

We include here only a brief outline of the master process and refer to Ref. 15 for details and pseudo code. As a Branch and Bound scheme, exploration and propagation alternate as follows:

Master Exploration. The master process explores the AND/OR graph in a depth-first manner guided by the start pseudo tree T_c. Upon expansion of a node n it consults a heuristic lower bound $lb(n)$ to make pruning decisions, where the computation of the upper bound $ub(n)$ can take into account previous subproblem solutions. If $lb(n) \geq ub(n)$, the current subtree can be pruned. Exploration is halted when the parallelization frontier is reached. The master then sends the respective subproblem, given by the subproblem root variable and its context instantiation, to a worker node.

Master Propagation. The master process also collects and processes subproblem solutions from the worker nodes. Upon receipt of a solved subproblem, its solution is assigned as the value of the respective node in the master search space and recursively propagated upwards towards the root, updating node values identical to sequential AOBB.

With a fixed number of workers p, the master initially generates only the first p subproblems; worker nodes solve subproblems using sequential AOBB[13] and send the solution back to the master, where it is propagated; the central exploration is then resumed to generate the next subproblem.

Example 3.1. Consider again the AND/OR search graph in Figure 1(d). Given a start pseudo tree having A and B, we can illustrate the parallelization scheme through Figure 3: the search space of the master process is marked in gray, and each of the eight independent subproblems rooted at C or E can be solved in parallel.

The central decision is obviously where to place the *parallelization frontier*, i.e., at which point to cut off the master search space. Preliminary experiments, conducted with globally enforced fixed-depth cutoff, have shown that the parallel scheme carries great potential.[15] It also became evident, however, that the issue of load balancing is crucial for the overall performance (while structural redundancy, for instance, does not seem to have a major impact). In particular, the scheme needs to ensure that the workload is evenly distributed over all processing units, each of which should be utilized equally. Secondly, it is critical to minimize overhead resulting from network communication and resource management.

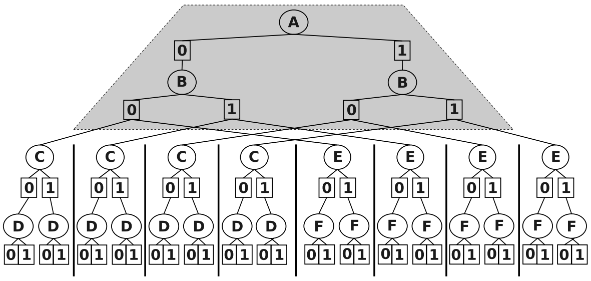


Fig. 3: Parallelization scheme applied to the example problem: master search space (gray) and eight independent subproblems.

In the fixed cutoff experiments we observed great variance in subproblem complexity with relative differences of up to three orders of magnitude. In the following section we will therefore focus on estimating subproblem complexity ahead of time[15] . With this the master can dynamically decide at which point a given subproblem is “simple enough” for parallelization (to avoid excessively hard tasks) and also avoid very easy subproblems, whose solution time will be dominated by the distributed system overhead.

4. Predicting Subproblem Size Using the Cost Function

In this section we derive a scheme for estimating the size of the explored search space of a conditioned subproblem using parameters associated with the problem’s cost function, allowing us to enforce an upper bound on the complexity of subproblems.

When considering a particular subproblem rooted at node n, we propose to estimate its complexity $N(n)$ (i.e., the number of node AOBB explores to solve it) as a function of the heuristic lower bound $L(n)$ as well as the upper bound $U(n)$, which can be computed based on earlier parts of the search space or through an approximation algorithm like local search; we will also use the height $h(n)$ of the subproblem pseudo tree.

4.1. *Main Assumptions*

We consider a node n that roots the subproblem $P(n)$. If the search space below n was a perfectly balanced tree of height D, with every node having exactly b successors, clearly the total number of nodes is $N = (b^{D+1} - 1)/(b - 1) \approx b^D$.

However, even if the underlying search space is balanced, the portion expanded by BaB, guided by some heuristic evaluation function, is not: the more accurate the heuristic, the more focused around the optimal solution paths the search space will be. In state-based search spaces it is therefore common to measure effectiveness in post-solution analysis via the *effective branching factor* defined as $b = \sqrt[D]{N}$ where D is the length of the optimal solution path and N is the actual number of nodes generated.[14]

Inspired by this approach, for a subproblem rooted at n we adopt the idea of approximating the explored search space by a balanced tree and express its size through $N(n) = b(n)^{D(n)}$. However, in place of the optimal solution path length (which corresponds to the pseudo tree height in our case), we propose to interpret $D(n)$ as the average leaf node depth $\bar{D}(n)$ defined as follows:

Definition 4.1 (Average leaf node depth). *Let $l_1, \dots, l_j$ denote the leaf nodes generated when solving subproblem $P(n)$. We define the* average leaf node depth *of $P(n)$ to be $\bar{D}(n) := \frac{1}{j}\sum_{k=1}^{j} d_n(l_k)$, where $d_n(l_i)$ denotes the depth of leaf node i relative to the subproblem root n.*

We next aim to express $b(n)$ and $\bar{D}(n)$ as functions of the subproblem parameters $L(n)$, $U(N)$, and $h(n)$ (using other parameters is subject to future research).

4.2. *Estimating the Effective Branching Factor*

For the sake of simplicity we assume an underlying, "true" effective branching factor b that is constant for all possible subproblems. We feel this is a reasonable assumption since all subproblems are conditioned within the same graphical model. We thus model $b(n)$ as a normally distributed random variable and take its mean as the constant b , which we found to be confirmed in experiments. An obvious way to learn this parameter is then to average over the effective branching factors of previous subproblems, which is known to be the right statistic for estimating the true average of a population.

Estimating b for new Subproblem $P(n)$: Given a set of already solved subproblems $P(n_1), \dots, P(n_r)$, we can compute $\bar{D}(n_i)$ and derive effective branching degrees $b(n_i) = \sqrt[\bar{D}(n_i)]{N(n_i)}$ for all i. We then estimate b through $b^* = \frac{1}{r}\sum_{i=1}^{r} b(n_i)$.

4.3. *Deriving and Predicting Average Leaf Depth*

With each subproblem $P(n)$ rooted at a node n we associate a lower bound $L(n)$ based on the heuristic estimate and an upper bound $U(n)$ derived from the best solution from previous subproblems[a]. Both $L(n)$ and $U(n)$ are known before we start solving $P(n)$. We can assume $L(n) < U(n)$, since otherwise n itself could be pruned and $P(n)$ was trivially solved. We denote with $lb(n')$ and $ub(n')$ the lower and upper bounds of nodes n' within the subproblem $P(n)$ at the time of their expansion and similarly assert that $lb(n') < ub(n')$ for any expanded node n'.

[a] We assume a graphical model with addition as the combination operator. Adaption to multiplication is straightforward.

Since the upper bound is derived from the best solution found so far it can only improve throughout the search process. Furthermore, assuming a monotonic heuristic function (that provides for any node n' a lower bound on the cost of the best solution path going through n'), the lower bounds along any path in the search space are non-decreasing and we can state that any node n' expanded within $P(n)$ satisfies:

$$L(n) \leq lb(n') < ub(n') \leq U(n)$$

Consider now a single path within $P(n)$, from n down to leaf node l_k, and denote it by $\pi_k = (n'_o, \ldots, n'_{d_n(l_k)})$, where $n'_0 = n$ and $d_n(l_k)$ is again the depth of l_k with respect to n (and hence $n'_{d_n(l_k)} = l_k$). We will write lb_i for $lb(n'_i)$ and ub_i for $ub(n'_i)$, respectively, and can state that $lb_i \geq lb_{i-1}$ and $ub_i \leq ub_{i-1}$ for all $1 \leq i \leq d_n(l_k)$ (note that $lb_0 = L(n)$ and $ub_0 = U(n)$). An internal node n' is pruned iff $lb(n') \geq ub(n')$ or equivalently $ub(n') - lb(n') \leq 0$, hence we consider the (non-increasing) sequence of values $(ub_i - lb_i)$ along the path π_k; in particular we are interested in the average change in value from one node to the next, which we capture as follows:

Definition 4.2 (Average path increment). *The* average path increment of π_k *within* $P(n)$ *is defined by the expression:*

$$inc(\pi_k) = \frac{1}{d_n(l_k)} \sum_{i=1}^{d_n(l_k)} ((ub_i - lb_i) - (ub_{i-1} - lb_{i-1})) \tag{1}$$

If we assume $(ub_{d_n(l_k)} - lb_{d_n(l_k)}) = 0$, the sum reduces to $(U(n) - L(n))$. Thus rewriting Expression 1 for $d_n(l_k)$ and averaging to get $\bar{D}(n)$ as in Definition 4.1 yields:

$$\bar{D}(n) = (U(n) - L(n)) \, \frac{1}{j} \sum_{k=1}^{j} \frac{1}{inc(\pi_k)} \tag{2}$$

We now define $inc(n)$ of $P(n)$ through $inc(n)^{-1} = \frac{1}{j}\sum_{k=1}^{j} \frac{1}{inc(\pi_k)}$, with which Expression 2 becomes $\bar{D}(n) = (U(n) - L(n)) \cdot inc(n)^{-1}$, namely an expression for $\bar{D}(n)$ as a ratio of the distance between the initial upper and lower bounds and $inc(n)$. Note that in post-solution analysis $\bar{D}(n)$ is known and $inc(n)$ can be computed directly, without considering each π_j.

One more aspect that has been ignored in the analysis so far, but which is likely to have an impact, is the actual height $h(n)$ of the subproblem pseudo tree. We therefore propose to scale $\bar{D}(n)$ by a factor of the form $h(n)^\alpha$; in our experiments we found $\alpha = 0.5$ to yield good results[b]. The general expression we obtain is thus:

$$\frac{\bar{D}(n)}{h(n)^\alpha} = \frac{U(n) - L(n)}{inc(n)} \tag{3}$$

Predicting $\bar{D}(n)$ for New Subproblem $P(n)$: Given previously solved subproblems $P(n_1), \ldots, P(n_r)$, we need to estimate $inc(n)$ in order to predict $\bar{D}(n)$. Namely, we compute $inc(n_i) = (U(n_i) - L(n_i)) \cdot h(n_i)^\alpha \cdot \bar{D}(n_i)^{-1}$ for $1 \leq i \leq r$. Assuming again that $inc(n)$ is a random variable distributed normally we take the sample average to estimate $inc^* = \frac{1}{r}\sum_{i=1}^{r} inc(n_i)$.

[b] Eventually α could be subject to learning as well.

Using Equation 3, our prediction for $\bar{D}(n)$ is:

$$\bar{D}^*(n) = \frac{(U(n) - L(n)) \cdot h(n)^\alpha}{inc^*} \tag{4}$$

Predicting $N(n)$ for a New Subproblem $P(n)$: Given the estimates b^* and inc^* as derived above, we will predict the number of nodes $N(n)$ generated within $P(n)$ as:

$$N^*(n) = b^{*\ \bar{D}^*(n)} \tag{5}$$

The assumption that inc and b are constant across subproblems is clearly too strict, more complex dependencies will be investigated in the future. For now, however, even this basic approach has proven to yield good results, as we will demonstrate in Section 5.

4.4. *Parameter Initialization*

To find an initial estimate of both the effective branching factor as well as the average increment, the master process performs 15 seconds of sequential search. It keeps track of the largest subproblem $P(n_0)$ solved within that time limit and extracts $b(n_o)$ as well as $inc(n_0)$, which will then be used as initial estimates for the first set of cutoff decisions. Additionally, we perform a 60 second run of stochastic local search,[9] which returns a solution that is not necessarily optimal, but in practice usually close to it. This provides an initial lower bound for subproblem estimation and pruning.

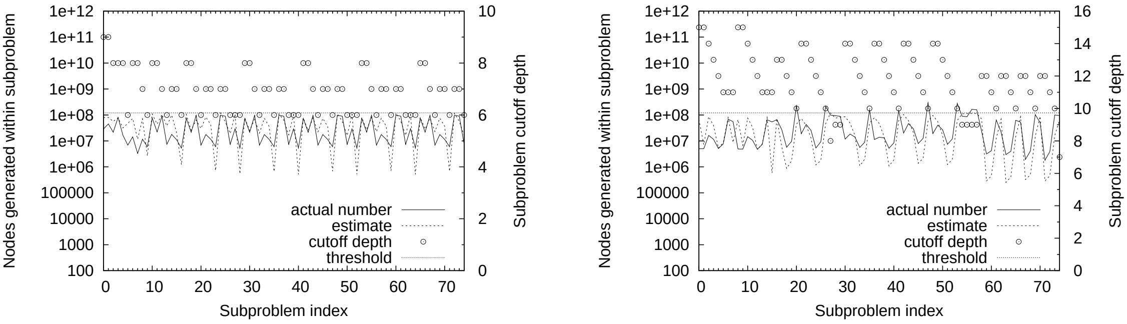
5. Experiments

We conducted experiments with our parallel AOBB scheme using the above prediction scheme to make the cutoff decision fully automatically. The cutoff threshold was set to $T = 12 \cdot 10^8$, which corresponds to roughly 20 minutes of processing time and was deemed to be a good compromise between subproblem granularity and parallelization overhead.

Overall solution times are given in Table 1. n, k, and w denote the number of variables, max. domain size, and induced width of the problem's Bayesian network. For reference we include the sequential solution time seq and the time par_{fix} of the best-performing parallel run with fixed cutoff depth from previous work.[15] seq/sls is then the time of the sequential scheme prefaced by 60 seconds

Table 1: Results of the automated parallel scheme (ped: 15 workers, mm: 10 workers).

instance	n	k	w	seq	par_{fix}	seq/sls	par^*/sls
ped7 (25/20)	1068	4	32	19,114	3,352	19,369	**2,843**
ped13 (20/20)	1077	3	32	2,752	**379**	2,856	419
ped19 (15/20)	793	5	25	*time*	27,372	*time*	**10,671**
ped31 (25/20)	1183	5	30	77,580	15,230	37,904	**3,970**
ped41 (25/20)	1062	5	33	14,643	**2,173**	14,059	2,311
ped51 (25/20)	1152	5	39	*time*	65,818	*time*	**59,975**
mm3.8.5-11	3612	2	37	9,715	1,443	3,003	**1,145**
mm3.8.5-12	3612	2	37	7,568	**1,430**	2,090	1,644
mm6.8.3-00	1814	2	31	12,595	1,797	319	**288**
mm10.8.3-11	2558	2	47	84,920	10,044	39,821	**6,906**
mm10.8.3-12	2558	2	47	5,630	1,357	2,549	**814**
mm10.8.3-13	2558	2	46	10,385	2,413	5,397	**2,208**



(a) ped31

(b) ped51

Fig. 4: Subproblem statistics for the first 75 subproblem of ped31 and ped51.

of stochastic local search providing an initial lower bound. Finally, column par^*/sls contains the overall solution time of the automated parallel scheme (similary including SLS preprocessing).

Pedigree Networks : The first set of problems consists of some very hard pedigree networks, encoded as Bayesian networks as described in Section 2.1, with the number of individuals and loci, respectively, given after the instance name in Table 1. We can see that in all cases the automatic scheme does at least as good as the best fixed cutoff, in some cases even better. Again it is important to realize that par_{fix} in Table 1 is the result of trying various fixed cutoff depths and selecting the best one, whereas par^*/sls requires no such “trial and error”. In case of pedigree31 the SLS initialization is quite effective for the sequential algorithm, cutting computation from 21 to approx. 10 hours – yet the automated scheme improved upon this by a factor of almost 10, to just above one hour. Furthermore, for ped51 and in particular ped19, both of which could not be solved sequentially, par^*/sls marks a good improvement over par_{fix} .

Mastermind Networks : While not as practically relevant, these hard problems encoding board game states can provide further insight into the parallel performance. Here we find that for most problems the automated scheme performs at least as well as the best fixed cutoff (determined after trying various depths); in general, however, we believe that the overall problem complexity is too close to the subproblem threshold, inhibiting better parallel performance.

5.1. *Subproblem Statistics*

Figures 4(a) and (b) contain detailed subproblem statistics for the first 75 subproblems generated by the automated parallelization scheme on ped31 and ped51, respectively. Each plot shows actual and predicted number of nodes as well as the (constant) threshold that was used in the parallelization decision. The cutoff depth of the subproblem root is depicted against a separate scale to the right.

As expected, the scheme does not give perfect predictions, but it reliably captures the trend. Furthermore, the actual subproblem complexities are all contained within an interval of roughly one order of magnitude, which is significantly more balanced than the results for fixed cutoff depth.[15] We also note that “perfect” load balancing is impossible to obtain in practice, because subproblem complexity can vary greatly from one depth level to the next along a single path. In particular, if a subproblem at depth d is deemed too complex, most of this complexity might stem from only one of its child subproblems at depth $d{+}1$, with the remaining ones relatively simple – yet solved separately. In light of this, we consider the above results very promising.

5.2. *Performance Scaling*

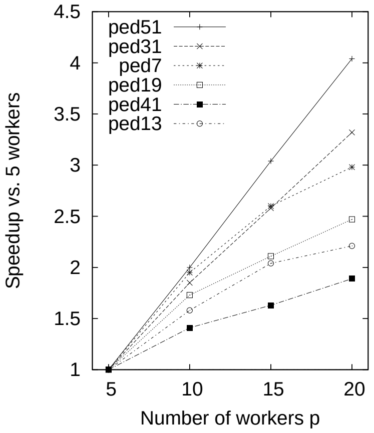


Fig. 5: Performance relative to $p = 5$ workers.

At this time we only have a limited set of computational resources at our disposal, yet we wanted to perform a preliminary evaluation of how the system scales with p, the number of workers. We hence ran the automated parallel scheme with $p \in \{5, 10, 15, 20\}$ workers and recorded the overall solution time in each case.

Figure 5 plots the relative overall speedup in relation to $p = 5$ workers. For nearly all instances the behavior is as expected, at times improving linearly with the number of workers, although not always at a 1:1 ratio. It is evident that relatively complex problem instances profit more from more resources; in particular ped51 sees a two-, three-, and fourfold improvement going to twice, thrice, and four times the number of workers, respectively. For simpler instances, we think the subproblem threshold of approx. 20 minutes is too close to the overall problem complexity, thereby inhibiting better scaling.

6. Conclusion & Future Work

This paper presents a new framework for parallelization of AND/OR Branch and Bound (AOBB), a state-of-the-art optimization algorithm over graphical models, with applications to haplotyping for general pedigrees. In extending the known idea of parallel tree search to AOBB, we show that generating independent subproblems can itself be done through an AOBB procedure, where previous subproblem solutions are dynamically used as bounds for pruning new subproblems.

The underlying parallel framework is very general and makes minimal assumptions about the available parallel infrastructure, making this approach viable on many different parallel and distributed resource pools (e.g., a set of networked desktop computers in our case).

Experiments have shown that the central requirement for good performance lies in effective load balancing. We have therefore derived an expression that captures subproblem complexity using an exponential functional form using three subproblem parameters, including the cost function. We then proposed a scheme for learning this function's free parameters from previously solved subproblems. We have demonstrated empirically the effectiveness of the estimates, leading to far better workload balancing and improved solution times when computing the most likely haplotypes on a number of hard pedigree instances.

We acknowledge that this initial estimation scheme, while justified and effective, still includes some ad hoc aspects. We aim to advance the scheme by taking into account additional parameters and by providing firm theoretical grounds for our approach. Besides extending the scheme itself, future work will also more thoroughly investigate the issue of parallel scaling, using larger grid setups than what we had access to so far (or performing simulations to that effect).

Furthermore, we plan to conduct more experiments on larger and harder problems from the haplotyping domain. In that context we are currently also working on a more in-depth analysis relating the size and structure of the pedigree and the number of loci in the problem to our scheme's performance. And while some problems may remain out of reach due to their inherent complexity, we do believe that our scheme will scale to many instances of interest; our confidence is in part based on the results obtained with the Superlink Online system,[16] which exploits a very similar strategy in the context of linkage analysis tasks and has proven very successful.

Finally, we note that in practice a small loss in accuracy can often be tolerated if it leads to significant time savings or better scaling. To that end, we intend to extend our current exact inference scheme to approximate reasoning; in particular, our parallel implementation should adapt very well to the concept of anytime search.

Acknowledgements

This work was partially supported by NSF grant IIS-0713118 and NIH grant 5R01HG004175-03.

References

1. Kento Aida, Wataru Natsume, and Yoshiaki Futakata. Distributed computing with hierarchical master-worker paradigm for parallel branch and bound algorithm. In *CCGRID*, pages 156–163, 2003.
2. David P. Anderson, Jeff Cobb, Eric Korpela, Matt Lebofsky, and Dan Werthimer. Seti@home: an experiment in public-resource computing. *Commun. ACM*, 45(11):56–61, 2002.
3. Geoffrey Chu, Christian Schulte, and Peter J. Stuckey. Confidence-based work stealing in parallel constraint programming. In Ian Gent, editor, *CP*, volume 5732 of *Lecture Notes in Computer Science*, pages 226–241, Lisbon, Portugal, September 2009. Springer-Verlag.
4. Gérard Cornuéjols, Miroslav Karamanov, and Yanjun Li. Early estimates of the size of branch-and-bound trees. *INFORMS Journal on Computing*, 18(1):86–96, 2006.
5. Rina Dechter and Robert Mateescu. AND/OR search spaces for graphical models. *Artif. Intell.*, 171(2-3):73–106, 2007.
6. Maáyan Fishelson, Nickolay Dovgolevsky, and Dan Geiger. Maximum likelihood haplotyping for general pedigrees. *Human Heredity*, 59:41–60, 2005.
7. Bernard Gendron and Teodor Gabriel Crainic. Parallel branch-and-bound algorithms: Survey and synthesis. *Operations Research*, 42(6):1042–1066, 1994.
8. Ananth Grama and Vipin Kumar. State of the art in parallel search techniques for discrete optimization problems. *IEEE Trans. Knowl. Data Eng.*, 11(1):28–35, 1999.
9. Frank Hutter, Holger H. Hoos, and Thomas Stützle. Efficient stochastic local search for MPE solving. In *IJCAI*, pages 169–174, 2005.
10. Philip Kilby, John Slaney, Sylvie Thiébaux, and Toby Walsh. Estimating search tree size. In *AAAI*, pages 1014–1019. AAAI Press, 2006.
11. Uffe Kjaerulff. Triangulation of graphs – algorithms giving small total state space. Technical report, Aalborg University, 1990.
12. Donald E. Knuth. Estimating the efficiency of backtrack programs. *Mathematics of Computation*, 29(129):121–136, 1975.
13. Radu Marinescu and Rina Dechter. AND/OR Branch-and-Bound search for combinatorial optimization in graphical models. *Artif. Intell.*, 173(16-17):1457–1491, 2009.
14. Nils J. Nilsson. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, 1998.
15. Lars Otten and Rina Dechter. Towards parallel search for optimization in graphical models. In *ISAIM*, 2010.
16. Mark Silberstein, Anna Tzemach, Nickolay Dovgolevsky, Maáyan Fishelson, Assaf Schuster, and Dan Geiger. Online system for faster multipoint linkage analysis via parallel execution on thousands of personal computers. *American Journal of Human Genetics*, 78(6):922–935, 2006.
17. Douglas Thain, Todd Tannenbaum, and Miron Livny. Distributed computing in practice: the Condor experience. *Concurrency - Practice and Experience*, 17(2-4):323–356, 2005.

DYNAMIC, MULTI-LEVEL NETWORK MODELS OF CLINICAL TRIALS

MARCO D. SORANI[1], GEOFFREY T. MANLEY[1],
J. CLAUDE HEMPHILL[1] AND SERGIO E. BARANZINI[2]
[1]Department of Neurological Surgery and [2]Department of Neurology
University of California, San Francisco
San Francisco, CA 94110, USA
Emails: soranim@pharmacy.ucsf.edu, manleyg@neurosurg.ucsf.edu,
chemphill@sfgh.ucsf.edu, sebaran@cgl.ucsf.edu

While networks models have often been applied to complex biological systems, they are increasingly being implemented to investigate clinical questions. Clinical trials have been studied extensively by traditional statistical methods but never, to our knowledge, using networks. We obtained data for 6,847 clinical trials from five "Nervous System Diseases" (NSD) and five "Behaviors and Mental Disorders" (BMD) from the clinicaltrials.gov registry. We constructed networks of diseases and interventions for visualization and analysis using Cytoscape software. To standardize nomenclature and enable multi-level annotation, we used MeSH and UMLS terms. We then constructed separate BMD and NSD networks to study dynamics over time. To assess how topology features related to clinical significance, we constructed a sub-network of Multiple Sclerosis and Alzheimer's trials and identified which trials had been published in high-profile medical journals. We found that the BMD network has evolved into a large, decentralized topology and does not distinctly reflect the five diseases by which it was defined, while the NSD network does, though other diseases and sub-phenotypes have emerged as areas of research. We also found that high-profile trials have distinctive network characteristics. Future work is needed to address mathematical questions such as scale-dependence of network features, clinical questions such as trial design optimization, and methodological questions such as data quality improvement.

1. Background

Network models can reveal complex relationships in large data sets, and network topologies have been shown to share remarkably consistent features across diverse fields of study [1-3]. These features include scale free and small world properties, preferential attachment growth dynamics, vulnerability to perturbations of network hubs, and modularity. Network approaches are particularly revealing when they can describe systems in their entirety, incorporate quantitative and computable measurements, integrate multiple levels of detail, and capture the dynamics of a system over time [4]. Network models have often been applied to biological systems. For example, studies of gene-disease associations have interrogated genetic similarities among auto-immune diseases [5] and relationships between metabolic diseases and co-morbitities [6]. Increasingly, networks are also being implemented to investigate clinical and social questions. Recent studies have investigated the dynamics of infectious disease transmission [7], workflow in the intensive care unit [8], and collaborations resulting in publication [9,10].

Clinical trials, the gold standard of clinical research, have long been studied by traditional statistical methods [11-13] but never, to our knowledge, using network models. This is ironic since clinical collaborations have long been operated and even referred to as "networks" [14,15]. However, with requirements by journal editors [16] and U.S. Federal law [17] that certain trials be registered, and with the growth of public-access registries, it is now possible to apply network models to clinical trial data. Furthermore, given the central role of clinical trials in translational research, network modeling may provide useful insights to address challenges such as designing comprehensive but non-redundant research programs and building on existing knowledge to design new trials.

In this study, we apply network models to characterize the dynamics and multi-level structure of a large set of data from clinical trials in nervous system diseases and behavioral and mental disorders obtained from the clinicaltrials.gov registry (http://www.clinicaltrials.gov). We also address challenges related to defining nomenclature and scope of clinical trial networks. We hypothesize that, as in other types of networks, hubs are functionally important, and networks grow according to a preferential attachment model. To test this, we describe features of these multi-level networks, including topological parameters, characteristics of network hubs and clusters, and the dynamics of clinical trial networks over time. We find that different disease types demonstrate divergent network topologies over time, and we observe distinct characteristics of clinically "influential" trials. These findings will be useful to assess areas of emphasis, overlap and omission in clinical research and funding programs as well as to identify relationships within and among disease phenotypes and therapeutic strategies.

2. The scope and construction of a clinical trials network

As the number of nodes in a network grows, the number of edges among them can increase exponentially, making analyses computationally intensive. There were 91,813 trials in clinicaltrials.gov categorized into 22 categories of diseases and conditions, as of 6/24/2010, and in the category of "Nervous System Diseases" alone, there were 65,462 non-unique trials in 506 specific diseases and conditions (Figure 1). To define a computationally tractable and clinically interpretable system, we downloaded 6,847 non-unique trials from five conditions categorized as "Nervous System Diseases" (NSD) and five conditions categorized as "Behaviors and Mental Disorders" (BMD). The NSD conditions were "Alzheimer Disease" (AD) (687 studies), "Brain Injuries" (505), "Multiple Sclerosis" (MS) (536), "Parkinson Disease" (PD) (629), and "Stroke" (1070); the BMD conditions were "Alcoholism" (357), "Attention Deficit and Disruptive Behavior Disorders"(444), "Bipolar Disorder" (584), "Schizophrenia" (1271), and "Smoking" (764). The NSD and BMD categories had similar total numbers of trials (3427 v. 3420) for later comparisons.

For each trial, we downloaded multiple parameters and performed several types of re-coding on the data (Table 1). We obtained National Clinical Trial (NCT) ID, Recruitment Status, Condition, Intervention, Sponsor, Study Type, Start Date and Completion Date for each trial. Intervention and Sponsor data frequently contained Unicode characters which we converted to ASCII text. Intervention attributes were nested (e.g., Drug: Aspirin) and were parsed to derive an Intervention Type field (e.g., Drug). Multiple fields contained more than one attribute per row in a pipe ("|") delimited manner (e.g., Drug: Aspirin|Drug: Codeine) and were also parsed. Other parameters such as study design and outcome are available, but they are highly heterogeneous and not readily computable.

Fig. 1

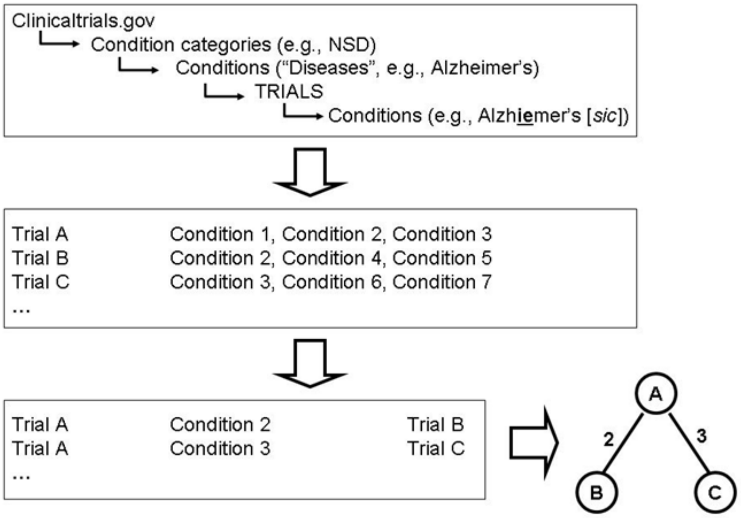

Clinicaltrials.gov
Condition categories (e.g., NSD)
Conditions ("Diseases", e.g., Alzheimer's)
TRIALS
Conditions (e.g., Alzhiemer's [*sic*])
Trial A Condition 1, Condition 2, Condition 3
Trial B Condition 2, Condition 4, Condition 5
Trial C Condition 3, Condition 6, Condition 7
...
Trial A Condition 2 Trial B
Trial A Condition 3 Trial C
...
A
2
3
B
C

Fig. 1. Schematic representation of clinical trial data converted from tabular to graph structure. Clinicaltrials.gov contains disease data in a nested, partly ambiguous hierarchy. Trials are organized into categories such as Nervous System Diseases (NSD). These categories include standardized Conditions which include specific trials. Specific trials can include multiple free-form Conditions entered by investigators--please note the hypothetical typographical error. Since the term "Conditions" is used twice, we subsequently refer to the higher Conditions as "Diseases". From the tabular layout, files of node-edge-node triplets were generated for import into Cytoscape.

Where possible, we standardized conditions to Medical Subject Headings (MeSH) (http://www.ncbi.nlm.nih.gov/mesh) and Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls) nomenclature. Clinicaltrials.gov uses the term "Condition" for both their pre-defined diseases and the multiple free-form, mixed case text conditions submitted by investigators for each trial. Intervention data are also submitted by investigators as free-form text. For example, in multiple sclerosis, there are three pre-defined Diseases: "Multiple Sclerosis", "Multiple Sclerosis, Relapsing-Remitting", and "Multiple Sclerosis, Chronic Progressive". Diseases map to MeSH terms and identifiers such as "Multiple Sclerosis, Chronic Progressive" [C10.314.350.500.200]. By contrast, there may be dozens of free-form text Conditions, some of which reflect true sub-phenotypes [18] and some of which are cases of inconsistent nomenclature, as shown in the examples related to brain injuries in Table 2. To standardize free-form text to UMLS Concept Unique Identifiers (CUI), we used the Batch SemRep web tool (http://skr.nlm.nih.gov).

We then constructed networks for visualization and analysis using Cytoscape software (http://www.cytoscape.org). First, we implemented custom Python scripts (http://www.python.org) to convert tabular data into undirected graphs defined by node-edge-node triplets (e.g., Trial-Condition-Trial) (Figure 1). Briefly, we iterated through all Conditions in all trials to identify trials that studied Conditions in common. This same algorithm was applied to Interventions and Sponsors and could be applied to any trial parameter. We also generated node attribute files defined by "node = <type>" statements (e.g., "Trial1 = NSD"). Once we constructed the networks, we analyzed their topologies visually and quantitatively. In this study, networks are displayed using the yFiles (http://www.yworks.com) Organic layout. Using the Network Analyzer plug-in [19] and custom code, we computed common topological parameters including node degree (i.e., the number of edges incident to the node) and others. We also identified hubs and clusters in the networks.

Table 1. Example record layout for a single trial. The Category, Disease and Intervention Type fields were derived from the clinicaltrials.gov primary data.

Parameter	Example value
NCT ID	NCT00167323
"Category"	BMD
"Disease"	Alcoholism
Recruitment Status	Completed
Intervention	Behavioral: Adherence therapy
"Intervention Type"	Behavioral
Study Type	Interventional
Start Date	Jul-03
Completed Date	Jul-07
Condition	Bipolar Disorder|Alcohol Use Disorder
Sponsors	University of Pittsburgh|National Institute on Alcohol Abuse and Alcoholism (NIAAA)

Table 2. Examples of non-standard nomenclature in the Condition field of trials in the "Brain Injuries" category, in order of frequency. UMLS identifiers and terms were assigned using the Batch SemRep web tool.

Standard UMLS term	UMLS CUI	Total	Free-text Condition	Subtotal
Traumatic Brain Injury	C0876926	173	Traumatic Brain Injury	148
			Mild Traumatic Brain Injury	10
			Severe Traumatic Brain Injury	9
			TBI (Traumatic Brain Injury)	6
Cerebral Palsy	C0007789	96	Cerebral Palsy	96
Brain Injuries	C0270611	82	Brain Injury	41
			Brain Injuries	33
			Brain Injuries Traumatic	5
			Acquired Brain Injury	3
Craniocerebral Trauma	C0018674	13	Craniocerebral Trauma	7
			Head Injury	6
Hypoxic-Ischemic Encephalopathy	C0752304	12	Hypoxic Ischemic Encephalopathy	7
			Hypoxic-Ischemic Encephalopathy	5
Subarachnoid Hemorrhage	C0038525	11	Subarachnoid Hemorrhage	11

3. Annotation of multiple network levels

A fundamental challenge in network modeling and systems biology is the integration of multiple "levels" of data, that is, data with different levels of granularity. We constructed multi-level networks to analyze trial Conditions, Interventions, and Sponsors. We defined four Condition levels, from the top down, as (1) the two clinicaltrials.gov categories (BMD and NSD) or "Both", (2) the 10 clinicaltrials.gov diseases or "Multiple", (3) the standardized UMLS CUIs, and (4) the free-form text Conditions. We defined three Intervention levels as (1) the Intervention type (e.g., "Drug" for the Intervention "Drug: Aspirin"), (2) the standardized CUIs of the Intervention, and (3) the free-form text Intervention. We defined two Sponsor levels as (1) sets of Sponsors and (2) individual Sponsors.

Fig. 2

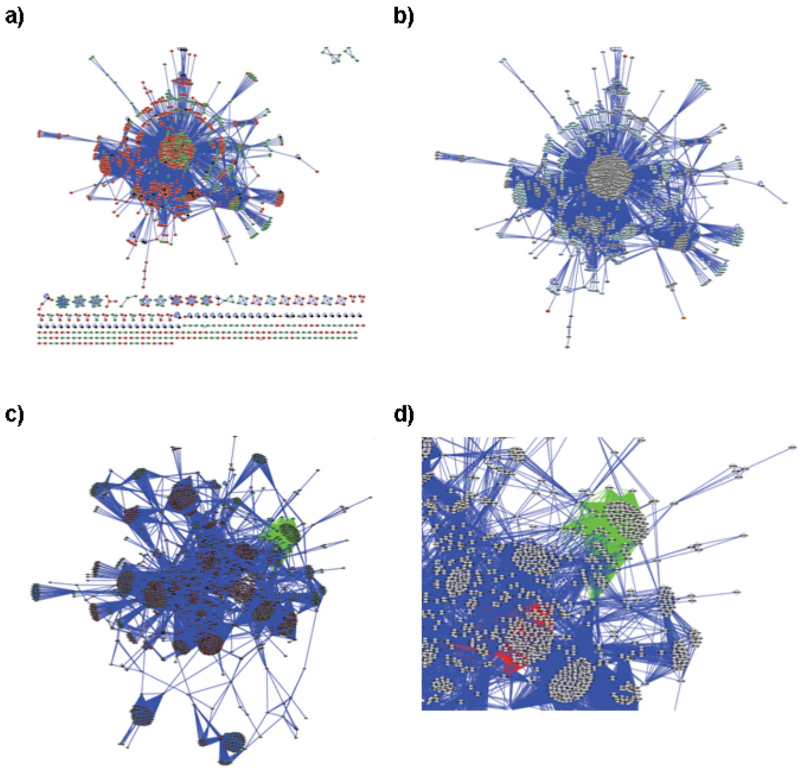


Fig. 2. Multi-level networks. Intervention networks are colored (a) by disease category, Behaviors and Mental Disorders (BMD) in red, Nervous System Diseases (NSD) in green, Both in black and (b) by intervention type ("Multiple" in gray, "Drug" in cyan). The central cluster represents placebo-controlled trials. There are more BMD trials represented than NSD trials and relatively few "Both" trials. Sponsor networks are colored (c) by category (nodes: BMD in red, NSD in green, Both in black) and (d) in an enlarged view of the upper right quadrant of (c), by the specific sponsor (edges: the National Institute of Mental Health in red, the National Institute of Neurological Disorders and Stroke in green).

To construct multi-level networks of Conditions, Interventions and Sponsors, we selected a subset of 2412 completed interventional trials (Figure 2). The Conditions network had 2202 nodes, similar to the Interventions and Sponsors networks (Table 3) but had 231,813 edges, a 4-5 fold increase compared to the other networks. A clear result of multi-level visualization is that, especially in large networks, higher levels of aggregation (e.g., disease category v. individual disease) are easier to interpret.

All networks had a large, primary connected component subdivided into more or less distinct sub-graphs which correspond to higher levels of aggregation, such as disease category. Clear clusters were also visible and correspond to lower levels of aggregation, such as placebo intervention or large sponsors.

Table 3. Similar topology parameters of completed interventional trial networks of Interventions and Sponsors. The clustering coefficient is a measure of the extent to which nodes in a graph cluster together.

Network	Nodes	Edges	Clustering coefficient	Diameter
Interventions	1372	61,749	0.74	9
Sponsors	2124	53,248	0.87	9

4. Assessment of network dynamics

Clinical trials provide an opportunity to study network dynamics because rich longitudinal data is available, and because multiple behaviors can be observed including new node and link formation as well as "death" (e.g., when a trial is withdrawn or terminated). We identified all trials in the registry that started after 1980 or completed before 2009. We did not include trials started after 2005 to control for possible lags in registration. We then constructed separate networks for BMD and NSD trials to compare their evolution over time, visually (Figure 3) and by topology metrics.

Trial network topologies diverged over time between BMD and NSD. Of note, for both the BMD and NSD networks, approximately 90% of trials in the registry began after 2000. Before 1995, there was a core BMD component and a separate co-morbidity component. Between 1996-2000, the schizophrenia\bipolar disorder core component grew, and alcohol and smoking trial clusters began to form, while attention deficit trials were a separate component. Since 2000, the BMD network has become large and decentralized and does not distinctly reflect the five major diseases by which it was defined. Before 1995, there was a greater total number of trials in NSD than in BMD, including a large stroke trial cluster, a more sparse cognitive disorders cluster, and a separate MS component. Between 1996-2000, the network clusters grew and became more integrated. Since 2000, the NSD network's clusters reflect the five major diseases by which it was defined. In addition, new disease trial clusters such as cerebral palsy have emerged as major areas of research, other areas like MS and Parkinson's Disease have spawned sub-phenotypes such as Secondary Progressive MS and Idiopathic PD, and more rare conditions have emerged. Notably, around 2005, first in BMD and then in NSD, the increase in completing trials began to exceed the number of starting trials in the registry.

Fig. 3

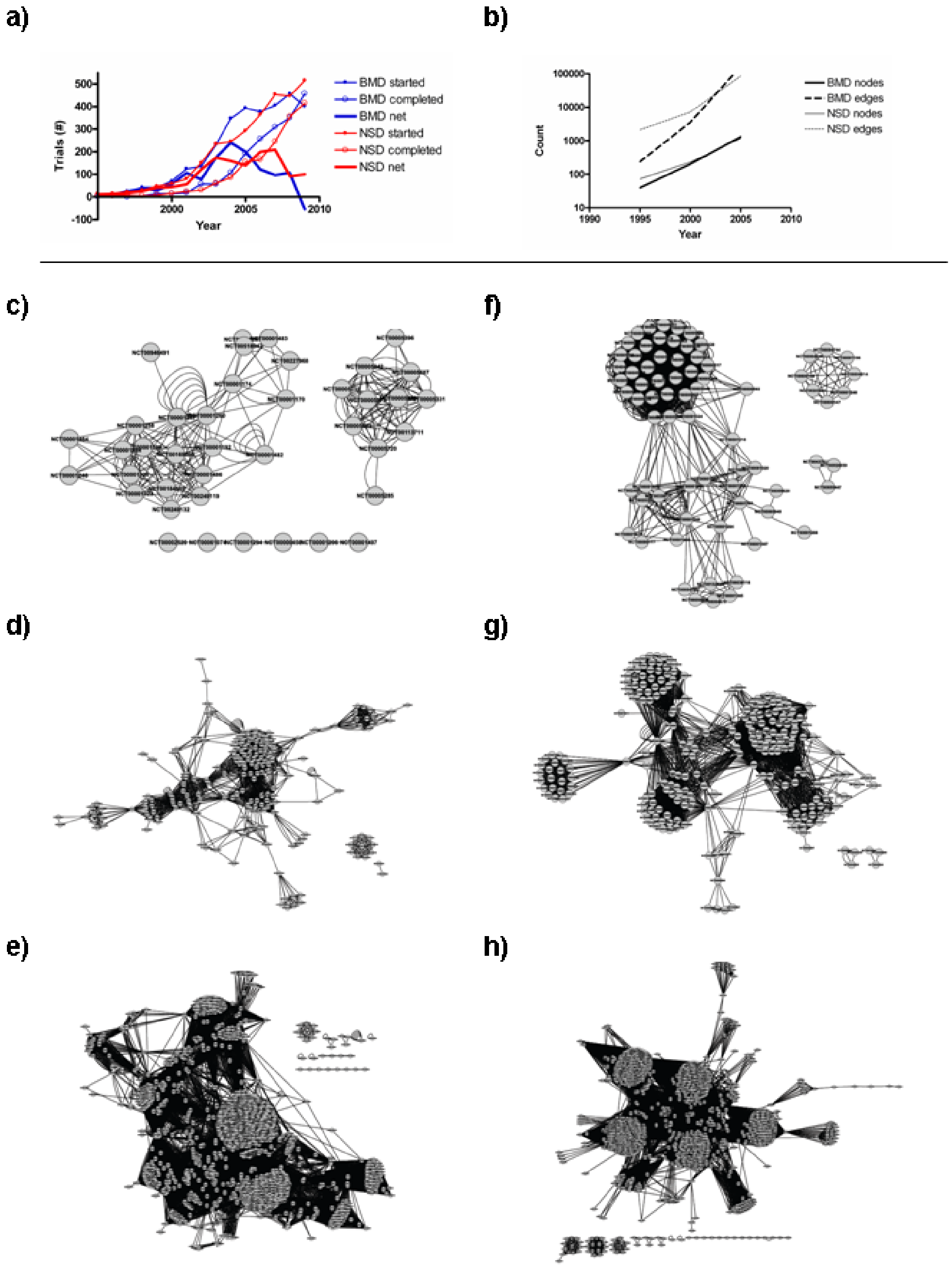


Fig. 3. Clinical trial network dynamics. (a) The number of trials started and completed in Behaviors and Mental Disorders (BMD) and Nervous System Diseases (NSD) rose slowly until about 2000 and then increased rapidly. (b) Sizes of both networks grew logarithmically, but the number of edges grew at different rates. Networks of BMD trials started (c) up until 1995, (d) between 1996-2000, and (e) between 2001-2005 are shown. Networks of NSD trials started (f) up until 1995, (g) between 1996-2000, and (h) between 2001-2005 are also shown.

These differences were reflected in network topology parameters. The BMD network's clustering coefficient increased over time (0.67, 0.85, 0.92) but was unchanged in the NSD network (0.91, 0.9, 0.93). In contrast, the BMD network's diameter and characteristic path length first rose and then fell, but both metrics rose steadily in the NSD network (4, 10, 6 v. 3, 5, 11 and 1.6, 3, 2.2 v. 1.8, 2.4, 2.5). Finally, the BMD network's density was unchanged over time (0.21, 0.16, 0.18) but fell in the NSD network (0.29, 0.17, 0.11).

5. Network characteristics of "influential" trials

We next examined how network topology features and dynamics related to clinical significance. First, we constructed a sub-network of the 479 MS and Alzheimer's trials in clinicaltrials.gov with a status of "Completed" and a reported start date. We then identified MS and Alzheimer's trials published in the Journal of the American Medical Association or the New England Journal of Medicine since 2005 to represent "influential" clinical trials. Figure 4 shows the topological characteristics of these influential studies relative to other trials in the context of Conditions, Interventions, and Sponsors networks.

Fig. 4

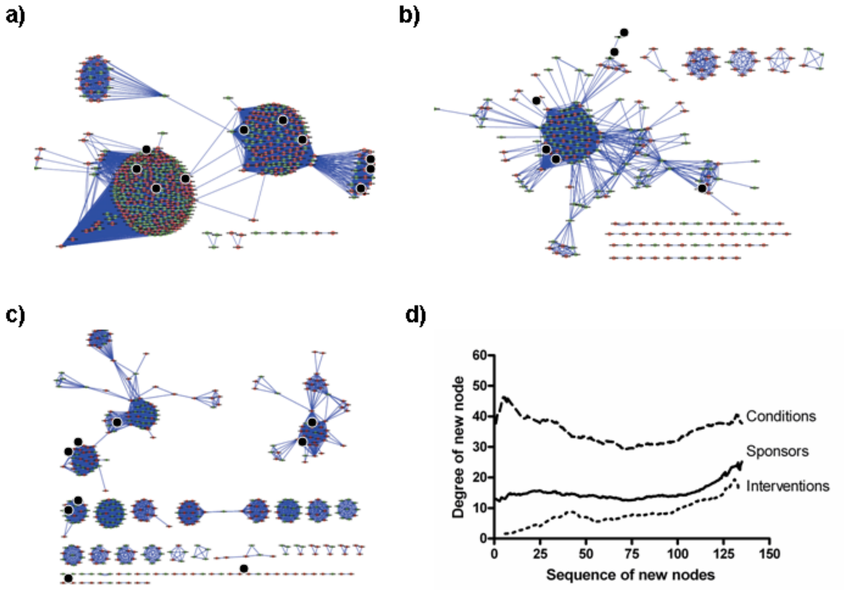


Fig. 4. Topological features of "influential" trials. (a) In the Conditions network, influential trials (black) are members of large disease clusters (AD, MS, Relapsing Remitting MS) near both prior (red) and subsequent (green) trials. (b) In the Interventions network, influential trials are in the network periphery, representing novel approaches, and are primarily near subsequent trials, supporting their influence on future studies. These include four trials in RRMS using fingolimod, interferon beta-1a, cladribine, and rituximab. (c) In the Sponsors network, influential trials are dispersed across various clusters of trials with common sponsors, which represent academic and industrial collaborations. They also occur primarily near prior trials, suggesting a culmination of previous efforts. (d) While the current degree of new nodes in the Intervention and Sponsor networks increased steadily as networks grew, the degree of new nodes in the Condition network started high and fell before rising again.

We set out to understand the dynamics of these three networks and, in particular, to describe when high-degree network hubs joined the networks. We first calculated degree distributions for the three networks in their current state. We then sorted the trials, that is the network nodes, by their start dates. Finally, we fit a coarse lowess spline to the three network series to identify trends in the data (Figure 4d). The degree of new nodes in the Intervention and Sponsor networks increased steadily, suggesting that early interventions were often abandoned and early sponsors often left the field, while later entrants formed more connections. The degree of new nodes in the Condition network started high, like a preferential attachment model, but fell before rising again, suggesting that medically relevant conditions were identified early and more have been defined recently.

We have defined influential trials using just one out of a multitude of possible ways. The increasing number of connected components in the networks (n=1, 6 and 16, for >3 nodes) supports the hypothesis that innovation may spread most easily among disease areas but may disseminate more slowly among groups studying different interventions, and information may flow with some difficulty across unconnected sponsor groups.

6. Discussion

In this study, we present the first network-based analysis, to our knowledge, of clinical trial data. Using a large set of data from clinical trials in nervous system diseases and behavioral and mental disorders obtained from the clinicaltrials.gov registry, we examine the topological parameters, network features, and longitudinal dynamics of clinical trial conditions, interventions and sponsors. We propose solutions to defining nomenclature and scope for constructing clinical trial networks. We hypothesized that, as in other types of networks, hubs and clusters are functionally important, and networks grow according to a preferential attachment model. We found that the role of network hubs was more similar among conditions and sponsors, since those hubs had functionally dominant roles, whereas, aside from the placebo cluster in the interventions network, the interpretation of functional importance was less clear. We also found that networks of different disease categories grew in divergent manners, and networks demonstrated variant models of preferential attachment.

Inconsistent data quality has previously been identified as an impediment to the construction of biological networks [20,21] and clinical databases [22,23]. One of the major challenges to studying human diseases computationally is the development of vocabularies and ontologies that realistically reflect the complex inter-relationships among phenotypes. Multiple solutions to this problem have been reported including mapping diseases to OMIM disorders [24], Medicare records [6], MeSH terms [25], ICD-9 codes and other ontologies. Clinicaltrials.gov implements MeSH terminology at the upper levels of its disease classification system but allows submission of free-form text at the lower levels. Other attributes such as interventions are unstructured but could be mapped to reference data sets such as RxNorm [26] or Drugs@FDA (http://www.accessdata.fda.gov). Still other attributes are submitted by trial sponsors with limited safeguards to ensure the accuracy and consistency of terminology.

We used clinicaltrials.gov and UMLS terms to standardize trial nomenclature and enable multi-level analysis. Multi-level analysis of trials may be useful when sponsors evaluate trials "upward" as components in an overall research program and "downward" as collections of individual patients. In both cases, issues of membership overlap and hierarchy may be encountered. Ahn et al. [27] constructed communities that incorporate overlap and hierarchical organization in biological and social networks. We addressed overlap by implementing classes such as "Both" for disease categories and "Multiple" for diseases.

Understanding the dynamic "evolution" of clinical trials from a systems perspective is similar to a phylogenetic analysis of ecosystems and may be useful in understanding the emergence, persistence, diversification and modularity of clinical research, particularly given the "noise" we have described in this type of data [28]. Many models have been proposed to describe dynamics in different types of network systems. These models include preferential attachment and linear distance dependence in internet topology [29], duplication-mutation schemes in the E. coli genetic network [30], modified preferential attachment in sexual contact networks [31], asymmetric disassembly for contraction and preferential attachment for re-growth in the New York garment industry [32], and anti-preferential attachment in protein-protein interaction [33]. In the context of clinical trials, one might expect that a growth model similar to preferential attachment might hold true for networks of conditions and sponsors, where it may be the case that "the rich get richer". However, for networks of interventions, one might expect an altogether different growth model, where it is less likely that trials will be initiated for an existing intervention, since an intervention will eventually either fail and disappear or succeed and no longer require new studies, though in both cases it might be introduced into new indications.

Studying the evolution of clinical trial networks can provide insight into mechanisms of knowledge flow, just as studying the spatial-temporal transmission of infectious disease might provide insight into mechanisms of communicability. While the flow of a virtual entity like information through a trial network is different from the flow of a physical entity like electricity through a power grid, both require sources and connectivity. An example of a knowledge source is the release of information via Pubmed, clinicaltrials.gov, or perhaps a conference, highlighting the importance of registering and reporting results of trials, including negative studies, in a complete and timely manner. An example of knowledge connectivity was described in Figure 4 where the differing connections between components in the three networks may have implications on how knowledge from influential trials is disseminated. This is significant because the average time to take a new therapeutic compound from discovery to commercialization in the U.S. is nearly 13 years, up from less than eight years in the 1960s. Opening a trial requires a median of approximately 2.5 years simply to begin patient accruals, not to complete the trial [34].

The primary limitation of this study is the reliance on potentially ambiguous categorizations and nomenclature of study conditions. For example, there are 536 studies in the pre-defined "Multiple Sclerosis" category, 181 studies under "Multiple Sclerosis, Relapsing-Remitting", and 26 studies under "Multiple Sclerosis, Chronic Progressive". However, searching with the text term "Multiple Sclerosis" returns 585 studies, while searching for the terms independently returns 625 studies. Other recognized phenotypes such as Primary Progressive MS are not explicitly defined. Clearly, the definition of disease nomenclature is an ongoing effort.

There are several potential future directions for this work. We chose to take a disease-centric approach by focusing on relationships among trials in neurological diseases. Alternative approaches could include expanding scope, by looking at all diseases; altering scope, by looking at a different set of diseases or a specific class of interventions; or altering the network model, by focusing on relationships using some unit other than trials. There are many opportunities to study other aspects of network dynamics. For example, in studies of corporate networks, the merging and splitting of nodes can represent acquisitions and spin-offs. Similarly, merging and splitting of nodes in clinical trial networks could reflect evolving understanding of sub-phenotypes of disease or differences in drug mechanisms of action. To further integrate multiple levels of detail into network models of clinical trials, it would be extremely useful to have patient-level data, but this may be difficult to obtain since such data is typically reported at a summary level, if at all. Other issues to investigate include scale-dependence [35] and network vulnerability [36]. For example, in the same way that other types of networks such as power grids or the Internet might experience cascading failure, what might happen to the practice of medicine if the findings of a "hub" trial are called into question?

In conclusion, we have presented the first network-based analysis of public clinical trials data. We defined a large set of trials in neurological conditions using data from clinicaltrials.gov. We analyzed multi-level models that integrated levels of granularity of trial conditions, interventions, and sponsors. We also analyzed dynamic models of network evolution over time. In both cases, we performed visual and topological evaluations. We highlight opportunities to make trial nomenclature more consistent and computable, we describe divergent network topologies over time in different disease types, and we identify characteristics of clinically "influential" trials in neurology using network models.

Acknowledgments

We thank Leonard Apeltsin and Michael Keiser for helpful discussions.

References

1. A. Clauset, C. Moore, M. E. Newman, *Nature.* **453**, 7191 (2008).
2. C. T. Butts, *Science.* **325**, 5939 (2009).
3. A. Tero, S. Takagi, T. Saigusa, K. Ito, D. P. Bebber, M. D. Fricker, K. Yumiki, R. Kobayashi and T. Nakagaki, *Science.* **327**, 5964 (2010).
4. [podcast] S. Mirsky and L. Hood, *Sci. Am.* July 11, (2007).
5. S. E. Baranzini, *Curr. Opin. Immunol.* **21**, 6 (2009).
6. D. S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai and A. L. Barabasi, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 29 (2008).
7. S. Riley, *Science.* **316**, 5829 (2007).
8. J. E. Gray, D. A. Davis, D. M. Pursley, J. E. Smallcomb, A. Geva and N. V. Chawla, *Pediatrics.* **125**, 6 (2010).
9. S. A. Greenberg, *BMJ.* **339** (2009).
10. C. M. Morel, S. J. Serruya, G. O. Penna and R. Guimaraes, *PLoS Negl. Trop. Dis.* **3**, 8 (2009).

11. J. F. Scoggins and D. L. Patrick, *Contemp. Clin. Trials.* **30**, 4 (2009).
12. X. Cao, K. B. Maloney and V. Brusic, *Immunome Res.* **4** (2008).
13. A. R. Brunoni, L. Tadini and F. Fregni, *PLoS One.* **5**, 3 (2010).
14. Asthma Clinical Research Network. J. M. Drazen, E. Israel, H. A. Boushey, V. M. Chinchilli, J. V. Fahy, J. E. Fish, S. C. Lazarus, R. F. Lemanske, R. J. Martin, S. P. Peters, C. Sorkness and S. J. Szefler, *N. Engl. J. Med.* **335**, 12 (1996).
15. MAC Study Group of the Canadian HIV Trials Network. S. D. Shafran, J. Deschenes, M. Miller, P. Phillips and E. Toma, *N. Engl. J. Med.* **330**, 6 (1994).
16. F. Berenbaum, *Best Pract. Res. Clin. Rheumatol.* **24**, 1 (2010).
17. T. Tse, R. J. Williams and D. A. Zarin, *Chest.* **136**, 1 (2009).
18. G. Miller, *Science.* **328**, 5976 (2010).
19. Y. Assenov, F. Ramírez, S. E. Schelhorn, T. Lengauer and M. Albrecht, *Bioinformatics.* **24**, 2 (2008).
20. J. Y. Chen, S. Mamidipalli and T. Huan, *BMC Genomics.* **10**, Suppl 1 (2009).
21. M. L. Nahm, C. F. Pieper and M. M. Cunningham, *PLoS One.* **3**, 8 (2008).
22. J. D. Horbar and K. A. Leahy, *Control. Clin. Trials.* **16**, 1 (1995).
23. A. Lux, S. Kropf, E. Kleinemeier, M. Jurgensen and U. Thyen; DSD Network Working Group, *BMC Public Health.* **9** (2009).
24. K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A. L. Barabasi, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 21 (2007).
25. F. Barrenas, S. Chavali, P. Holme, R. Mobini and M. Benson, *PLoS One.* **4**, 11 (2009).
26. F. Parrish, N. Do, O. Bouhaddou and P. Warnekar, *AMIA Annu. Symp. Proc.* **1057** (2006).
27. Y. Y. Ahn, J. P. Bagrow and S. Lehmann, *Nature.* **466**, 7307 (2010).
28. S. Erten, X. Li, G. Bebek, J. Li and M. Koyuturk, *BMC Bioinformatics.* **10** (2009).
29. S. H. Yook, H. Jeong and A. L. Barabasi, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 21 (2002).
30. M. Middendorf, E. Ziv, C. Adams, J. Hom, R. Koytcheff, C. Levovitz, G. Woods, L. Chen and C. Wiggins, *BMC Bioinformatics.* **5 (**2004).
31. B. F. de Blasio, A. Svensson and F. Liljeros, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 26 (2007).
32. S. Saavedra, F. Reed-Tsochas and B. Uzzi, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 43 (2008).
33. W. K. Kim and E. M. Marcotte, *PLoS Comput. Biol.* **4**, 11 (2008).
34. D. Dilts, *Nat. Med.* **16**, 6 (2010).
35. N. J. Gotelli, G. Graves and C. Rahbek, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5030 (2010).
36. S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley and S. Havlin, *Nature.* **464**, 7291 (2010).

MINING FUNCTIONALLY RELEVANT GENE SETS FOR ANALYZING PHYSIOLOGICALLY NOVEL CLINICAL EXPRESSION DATA

SEVIN TURCAN[*], DOUGLAS E. VETTER

Department of Biomedical Engineering, Tufts University, 4 Colby St., Medford, MA, 02155, USA
Department of Neuroscience, Tufts School of Medicine, Boston, MA 02111, USA

JILL L. MARON

Department of Pediatrics, Tufts Medical Center, 800 Washington St., Boston, MA 02111, USA

XINTAO WEI, DONNA K. SLONIM[†]

Department of Computer Science, Tufts University, 161 College Ave., Medford, MA, 02155, USA

Gene set analyses have become a standard approach for increasing the sensitivity of transcriptomic studies. However, analytical methods incorporating gene sets require the availability of pre-defined gene sets relevant to the underlying physiology being studied. For novel physiological problems, relevant gene sets may be unavailable or existing gene set databases may bias the results towards only the best-studied of the relevant biological processes. We describe a successful attempt to mine novel functional gene sets for translational projects where the underlying physiology is not necessarily well characterized in existing annotation databases. We choose targeted training data from public expression data repositories and define new criteria for selecting biclusters to serve as candidate gene sets. Many of the discovered gene sets show little or no enrichment for informative Gene Ontology terms or other functional annotation. However, we observe that such gene sets show coherent differential expression in new clinical test data sets, even if derived from different species, tissues, and disease states. We demonstrate the efficacy of this method on a human metabolic data set, where we discover novel, uncharacterized gene sets that are diagnostic of diabetes, and on additional data sets related to neuronal processes and human development. Our results suggest that our approach may be an efficient way to generate a collection of gene sets relevant to the analysis of data for novel clinical applications where existing functional annotation is relatively incomplete.

1. Introduction

Genome-wide expression studies are producing large quantities of experimental data characterizing a growing range of human diseases. Yet the biological interpretation of results obtained from these experiments is still a challenge, and clinical applications remain relatively elusive. Typically, microarray data are analyzed at the single gene level to identify transcripts with statistically significant differences between phenotypes, and a functional analysis is then performed on the gene list. Originally, such functional annotation was performed manually[1,2], but soon many tools to automate the process were developed[3-6].

[*] Current address: Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10065; current email: turcans@mskcc.org.

[†] Corresponding author; email: Donna.Slonim@tufts.edu. This work was supported by R01HD058880 of the National Institutes of Health to DKS.

More recently, analysis at the level of gene sets has emerged as a powerful alternative to individual-gene analyses to reflect the functional relationship between genes in a set. Mootha *et al.* initially demonstrated the power of using pre-defined gene sets in a case where no *individual* gene's expression was significantly different between normal and diabetic patients[7]. Since then, many gene set analysis methods have been developed[8-14]. The goal of all gene set analysis methods is to identify functionally related genes that display coordinated expression changes. Typically, gene set analysis methods can be distinguished by their statistical criteria for differential expression, null hypotheses, and p-value calculations[15].

However, all analytical methods incorporating gene sets depend on the knowledge of sets or pathways relevant to the underlying physiology. For fields such as diabetes and cancer, there has been considerable effort toward manual and computational curation of relevant gene function[16]. The Gene Ontology[17] contains controlled descriptions of gene function that are frequently used to define gene sets. Pathway databases such as KEGG[18], BioCyc[19], and BioCarta (www.biocarta.com) can also be used to generate gene sets. However, for many complex physiological processes, there is still a need to identify relevant groups of functionally linked genes. Recent work studying gene expression in human development suggests that this area is one in which additional annotation is needed[20].

Clustering approaches have long been used to find meaningful patterns in gene expression data and to identify functional gene sets from microarray data[7,21-23]. However, such methods do not necessarily generalize to inform the analysis of novel data sets since functionally related genes may be co-expressed only in a subset of conditions, and such gene sets would be missed by traditional clustering methods. Biclustering methods have emerged as an alternative to traditional clustering methods in such cases. Biclustering[24] finds subgroups of genes that exhibit similar expression patterns over a subset of conditions. Many biclustering algorithms have been proposed[25,26]. More sophisticated biclustering algorithms search for *coherent* expression changes within subsets of conditions[27-29]. Coherence of a bicluster refers to coordinated changes of the genes' expression patterns across a subset of conditions (as in Figure 1). Gene sets with coherent expression patterns in a data set may be functionally linked to the phenotype of interest.

Here, we describe a novel approach to identifying candidate gene sets using new criteria for selecting coherent biclusters across multiple experiments somewhat related to the desired clinical application. Previous efforts have looked for coherent functional modules showing enrichment in a particular gene expression data set, often by incorporating network, pathway, or clinical information[30-32]. Our method differs from these approaches in that we identify gene sets showing coherent expression patterns across multiple related studies, and then assess the general relevance of our candidate sets by using them for gene set analysis of *novel* clinical data. In this sense, our work is closest to that of Liu *et al.*[33], who find processes dysregulated across many related experiments. However, their work still requires pre-defined gene sets relevant to the phenotype being studied. The goal of our method is to systematically identify novel gene sets that generalize well for the analysis of new data in fields where molecular annotation is sparse, such as development or neuronal function. We use careful dataset selection, biclustering, and filtering to identify novel candidate gene sets, and we observe that several of these show coherent differential expression patterns in clinical test data sets from different yet related physiological processes.

This method works even when the training data sets come from different tissues or species than the test data, allowing us to find clinically-applicable gene sets using existing data from model organisms. Several of the gene sets differentially expressed in the test data show enrichment for informative Gene Ontology terms, but many others have no significant overlap with previously known functional categories. Nonetheless, they can be useful as diagnostics and can help direct future translational research into gene-gene and gene-disease relationships, particularly in medical fields where the underlying molecular physiology is not yet well understood.

2. Methods

2.1. *Algorithm overview*

We start by integrating publicly available gene expression data from several studies that are related, but not too closely related, to each other and to the test data set we wish to analyze. We apply a biclustering algorithm that finds coherent changes within and across studies (Figure 1) to the combined training data. Subsequently, we filter out biclusters that do not meet certain quality criteria. We consider the remaining biclusters as candidate gene sets, which we use for the analysis of human clinical gene expression test data distinct from the data used for gene set discovery. Details of each of these steps in our method are discussed below.

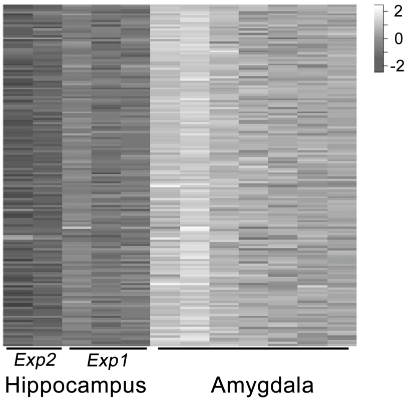


Fig 1. **Heatmap of a representative bicluster that shows coherent change across samples.** Samples from two studies on the hippocampus show lower gene expression when compared to samples from amygdala. Within each tissue type, coherent changes in expression are also apparent.

2.2. *Data acquisition and normalization*

We downloaded single channel Affymetrix microarray data (as .CEL files) from the Gene Expression Omnibus (GEO) (Table 1). The Affymetrix CEL files for each medical area of interest were imported into the R statistical software (v2.8.1; http://www.R-project.org), and all training data for that area were normalized at once. Normalization was performed with the AffyPLM package in BioConductor (v2.4), using RMA background correction, quantile normalization, and the Tukey biweight summary method. After normalization, the variances of all probes were computed across all samples, and the 50% of the probes with the lowest variance were removed, eliminating probes that are not expressed in the relevant tissues or whose expression does not vary enough to be informative for our purposes.

2.3. *Biclustering*

Next, we biclustered the normalized, filtered gene expression data using the Iterative Signature Algorithm (ISA)[27,34]. We have found that ISA identifies more coherent and potentially biologically relevant biclusters than several other biclustering methods[35,36]. Briefly, ISA starts with a random initial set of genes. All samples are scored for coherence with respect to this gene set and samples are chosen for which the score exceeds a predefined condition threshold (t_C). Next, all genes are scored across the selected samples and a new set of genes is selected based on a predefined gene threshold (t_G). The entire procedure is repeated until it converges. We used the BiCAT implementation[35] of the ISA algorithm with $t_G = 2$ and $t_C = 1$, parameters recommended for the identification of coherent patterns in a prior study[37].

Table 1 – Selected gene expression data sets for gene set discovery.

Data Set	GEO Accession #	Title	Tissue	Samples
Metabolic (Human)	GSE5090	Polycystic ovary syndrome patients vs control subjects	Adipose	PCOS patients, controls
	GSE9105	Effect of acute physiologic hyperinsulinemia	Vastus lateralis	240 mins of insulin infusion
	GSE474	Obesity and fatty acid oxidation	Vastus lateralis	Lean, obese
Developmental (Mouse)	GSE6882	Embryonic ovary development	Ovary	Embryonic
	GSE8065	Early postnatal development of the small intestine	Intestine	Postnatal
	GSE12769	Testis developmental time course	Testis	Postnatal
	GSE13103	Early mouse embryo eye development	Optic fissure	Embryonic
Neuronal (Mouse)	GSE9803	Striatal gene expression data	Striatum	wild-type
	GSE4040	Gene expression in murine hippocampus	Hippocampus	wild-type
	GSE4034	Gene expression in amygdala and hippocampus	Amygdala, Hippocampus	wild-type

2.4. *Selecting biclusters as candidate gene sets*

Although we chose the ISA biclustering approach because the algorithm is able to find coherent biclusters that include samples from multiple experiments, there is no guarantee that the resulting biclusters have the generalizable-coherence property that we want for our candidate gene sets. In addition, ISA often identifies multiple overlapping biclusters. While some degree of overlap between gene sets might accurately represent genes involved in more than one cellular process, a high degree of overlap of both genes and samples likely occurs when different random starting points of the iterative algorithm converge to similar solutions. Additionally, some of the resulting biclusters can be noisy and their genes' expression patterns only poorly correlated with each other. Therefore, we subject the biclusters to several quality measures before selecting certain ones as candidate gene sets.

First, we remove any biclusters that do not show coherent expression changes across samples from two or more experiments. That is, if the samples selected for a bicluster do not come from at least two different source data sets, we discard the gene set as being less likely to generalize to new conditions and tissues. Our experience suggests that this criterion, given an appropriate choice of training data, is most responsible for the applicability of these discovered gene sets in new contexts (data not shown).

We next assess the overlap between the gene sets defined by the biclusters. If any pair of gene sets G and H overlap such that at least 80% of the genes in G are in H *and* at least 80% of the genes in H are in G, we select only the bicluster with fewer genes. We reason that the smaller bicluster contains a core group of genes with a stronger functional association with the phenotype.

To enforce expression homogeneity within the biclusters, we use a recently proposed measure of bicluster quality, the average correlation value (ACV)[38], to score biclusters for homogeneity. The ACV measures the average pairwise expression correlation between all pairs of genes in a cluster. The maximum ACV score of 1.0 denotes a highly correlated bicluster. ACV has been shown to be more robust than the widely-used mean squared residue score[25]. We discard biclusters with ACV < 0.5 (though results are quite robust to varying this threshold). Biclusters that remain after all of these filtering steps are considered as candidate gene sets.

Finally, we note that normalization in meta-analyses is an important challenge, since many experiment-specific factors may persist even after normalization, and over-normalization may suppress real signal. In order to assess normalization bias in our resulting biclusters, we calculate a score called the chip correlation value (CCV). The CCV is measured by calculating the correlation between sample averages for genes in a given bicluster with the sample averages over the entire gene expression matrix. Although biclusters are not discarded based on their CCV scores, it should be noted that extreme correlations might reflect insufficient normalization.

2.5. *Applying candidate gene sets to analyze test data*

If our novel gene sets show coherent expression changes in a new setting, we can assume that their genes have some functional relationship, even if the exact nature of that relationship is unknown. Any gene-set data analysis method can be applied to assess coherent expression changes in test data; here, we choose Gene Set Enrichment Analysis (GSEA)[16]. GSEA is a statistical framework that determines if members of a given gene set show collective expression changes linked to sample phenotypes by calculating a Kolmogorov-Smirnov running sum called the enrichment score (ES). We report the normalized enrichment score (NES) because this measure accounts for the gene set size, thus allowing for comparison between different experiments. The magnitude of the NES reflects the degree of enrichment for a given gene set. We accept a gene set as differentially expressed using an FDR q-value cut-off of 25%, as suggested by the GSEA authors[16]. For time series data (the developmental data sets), we used the Pearson metric for ranking genes. For the maternal blood data set[20] (see Results), we used the GSEA-preranked option on genes ranked by the closer-to-zero (i.e., approximately the less-significant) of two t-scores, one comparing paired antepartum and postpartum maternal blood samples, and the other comparing paired neonatal cord blood and postpartum maternal blood samples.

Subsequently, in order to gain biological insight into the biclusters, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID)[39,40] (the April, 2008 release) to identify functional annotation terms significantly over-represented in the gene sets. A functional term is considered to be significantly enriched if its Benjamini-Hochberg-adjusted *p*-value, as reported by DAVID, is less than 0.05.

2.6. *Orthology*

In some cases, we derived biclusters based on gene expression data in model organisms and evaluated their utility for interpreting human gene expression data from clinical samples. In these cases, mouse-derived biclusters were mapped to their human gene symbols using DAVID's Gene ID Conversion Tool. Further, probe sets from human Affymetrix Chips are collapsed to their gene symbols using GSEA. In such cases, the gene symbols are used instead of their Affymetrix probe set identifiers.

3. Results

We applied this approach to three different functional areas to highlight its utility for functional interpretation of clinical data. We start by applying our method to the well-studied metabolic field and follow with two other areas where annotation is relatively sparse: neuronal function and development. Table 2 summarizes the characteristics of the resulting biclusters from each field.

Table 2 – Characterization of resulting biclusters.

Study	# of genes			# of conditions			ACV	CCV
	min	mean	max	min	mean	max	mean ± stdev	mean ± stdev
Metabolic	12	63.8	154	5	10.1	17	0.74 ± 0.12	-0.23 ± 0.30
Neuronal	7	122.6	436	3	9.0	19	0.95 ± 0.03	0.12 ± 0.49
Developmental	4	528.8	893	6	9.1	12	0.94 ± 0.03	0.07 ± 0.37

3.1. *Metabolic data set*

Metabolic disorders include a broad array of medical conditions such as diabetes, obesity, hypertension, and insulin resistance. We compiled gene expression data from publicly available metabolic studies involving human tissue samples hybridized to Affymetrix GeneChip HG-U133A arrays. The initial experiments include adipose tissue samples from polycystic ovary syndrome (PCOS) patients compared with control subjects (GSE5090), vastus lateralis muscle samples during acute physiologic hyperinsulinemia (GSE9105), and vastus lateralis muscle samples from obese and lean subjects. PCOS is a common endocrine disorder that is associated with metabolic abnormalities including insulin resistance, increased risk for diabetes mellitus, obesity and hyperlipidemia[41].

The entire metabolic data set consisting of 53 samples and 11,141 genes was used as input for biclustering. Overall, ISA identified 15 biclusters for the metabolic data. Filtering resulted in 11 biclusters selected as candidate metabolic gene sets. One bicluster was discarded based on low ACV; three biclusters were filtered because of high degree (>80%) of overlap (Figure 2). In such cases, the biclusters with fewer genes were selected because they were likely to be more specific. On average, the selected biclusters contain 64 genes and 10 conditions with more than 73% correlation between genes. Further, average CCV is relatively low (-0.23 ± 0.3) suggesting that the clusters are not due to normalization artifacts (Table 2).

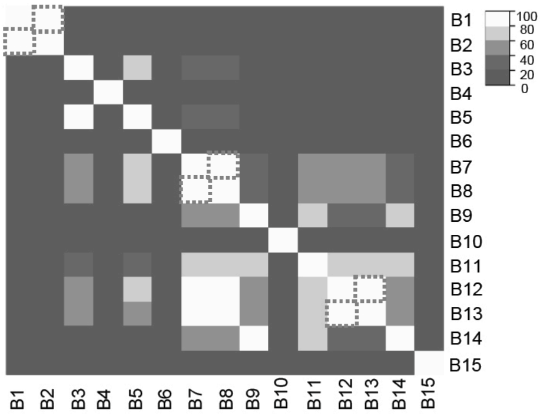


Figure 2. **Metabolic bicluster overlap before filtering.** A heatmap of overlap between biclusters from the metabolic study is shown. Biclusters with >80% overlap with each other are outlined in dashed boxes. In such cases, the bicluster with fewer genes is chosen as a candidate gene set. Note that biclusters 7 and 13 are both retained because the high overlap is in one direction only. In such cases, it is possible that both gene sets represent interesting biological functions.

We then applied these candidate metabolic gene sets in a GSEA analysis of data from Mootha, *et al.* comparing smooth muscle gene expression in diabetic patients and healthy controls[7]. Recall that this is the data set that was first used to demonstrate the GSEA approach; there are no individually differentially expressed genes, and gene sets related to oxidative phosphorylation were shown to be downregulated in diabetics in this data. However, no gene sets were shown to be significantly *upregulated* in diabetes[7]. In our experiments on the same data, out of our eleven candidate biclusters, three were significantly upregulated (FDR q-value < 0.25) in smooth muscle from diabetic patients: *bicluster9*, *bicluster11* and *bicluster14*. The GSEA results for differential expression of these gene sets are summarized in Table 3A, and full functional enrichment results are listed in supplementary table S1 (http://bcb.cs.tufts.edu/genesetPSB11/).

In an attempt to interpret the functional role of these gene sets, we evaluated the enriched biclusters using functional annotation tools in DAVID. However, these differentially expressed biclusters either showed no statistically significant overlap with current ontology classes (*bicluster11*) or overlapped only with broad GO terms such as *developmental process* (*bicluster14*) or *multicellular organismal process* and *biological regulation* (*bicluster9*).

We had originally expected that any gene sets we discovered in our metabolic data would overlap heavily with existing functional annotation, reflecting the wealth of research about the molecular mechanisms of diabetes and obesity. However, we instead discovered new gene sets that exhibited coherent changes across diverse experiments and that also showed significant coordinated upregulation in diabetics. While the exploratory q-value cutoff suggested for GSEA analysis[16] allows for a one-in-four false-positive rate, all three of the gene sets identified in this analysis had much lower q-values. Thus, although any of these findings *might* be a false-positive, it is unlikely (probability ≤ 0.0005) that all three of them are. We believe these results suggest that there may be previously unrecognized functional links among the members of each of these gene sets, warranting further study. In clinical applications where diagnosis is difficult or early diagnosis is critical, such gene sets might also be useful as diagnostic tools even before their functional roles are understood.

Table 3. Differential expression of candidate gene sets in test data.

A) Metabolic biclusters

Species	Tissue	Bicluster #	# of genes	ES	NES	NOM p-val	FDR q-val
Homo Sapiens	Smooth Muscle	Bicluster14	31	0.57	1.71	0.01	0.08
		Bicluster9	39	0.54	1.64	0.03	0.07
		Bicluster11	32	0.50	1.60	0.04	0.08

B) Neuronal biclusters

Species	Tissue	Bicluster #	# of genes	ES	NES	NOM p-val	FDR q-val
Homo Sapiens	Dorsolateral prefrontal cortex	Bicluster4	128	0.65	1.52	0.00	0.21
		Bicluster12	65	0.53	1.39	0.05	0.22
		Bicluster1	197	0.58	1.38	0.13	0.19
		Bicluster3	219	0.51	1.37	0.10	0.17

C) Developmental biclusters

Species	Tissue	Bicluster #	# of genes	ES	NES	NOM p-val	FDR q-val
Homo Sapiens	Blood	Bicluster4	239	0.31	1.54	0.000	0.005

3.2. *Neuronal data set*

Motivated by an interest in the impact of loss of nicotinic activity on cochlear synapse formation[42], we collected gene expression data from substructures of the mouse central nervous system: striatum (GSE9803), hippocampus (GSE4040) and amygdala (GSE4034). Gene expression data from only wild-type mice were considered and all studies utilized Affymetrix Mouse430.2 GeneChips. This neuronal data set included 32 samples and 22,550 genes. ISA initially identified 33 biclusters for the neuronal data[42]; filtering resulted in 25 candidate neuronal gene sets, whose characteristics are summarized in Table 2.

We applied the neuronal candidate gene sets to analyze human gene expression data from postmortem brains (specifically, dorsolateral prefrontal cortex) of adults with Down syndrome (DS) and healthy control subjects (GSE5390). In this data set *bicluster4*, *bicluster12*, *bicluster1* and *bicluster3* were upregulated in DS patients (Table 3B).

Bicluster4 showed statistically significant enrichment for the GO biological process term, *lipid metabolic process*, and several PANTHER terms including *lipid, fatty acid and steroid metabolism; mRNA transcription regulation*; *voltage-gated K channel;* and *transferase*. *Bicluster1* is enriched for several GO categories including *nervous system development, myelination,* and *regulation of action potential*. Enriched GO terms for bicluster 3 include *developmental process, localization, cell adhesion and death*. Enriched PANTHER categories for this bicluster include *neuronal activities, receptor mediated endocytosis, cytoskeletal protein, cell junction protein,* and *cadherin*. On the other hand, *bicluster12* did not exhibit statistically significant overlap with any functional annotation terms.

Cadherins are proteins involved in calcium-ion-mediated cell adhesion. Abnormalities in myelination, cell adhesion, and lipid classes have been implicated in DS[43-45]. In addition, these results are consistent with our recent observation of increased oxidative stress, and apparent downstream disruption of ion signaling and cell structural integrity, in the DS fetus[46]. The functional roles of genes in these novel gene sets mined from diverse neuronal tissues in healthy mice may therefore help inform ongoing translational efforts to develop novel therapies for Down syndrome.

3.3. *Developmental data set*

We collected gene expression data representing mouse developmental time courses in various tissues, all hybridized to Affymetrix Mouse430.2 GeneChips. We only considered data from wild-type animals; treated samples and mutant strains were excluded. The data were derived from ovary (GSE6882) and optic fissure (GSE13103) during embryonic development, and intestine (GSE8065) and testis (GSE12769) during postnatal development. Overall, this data set contained 24 samples and 22,550 genes.

Initially, ISA identified 25 biclusters on this data set. Filtering resulted in 10 biclusters to be considered as candidate developmental gene sets, which are characterized in Table 2. We then applied these developmental biclusters to re-analyze expression data from our previous study of maternal and fetal gene expression[20]**.** This study confirmed the detection of fetal mRNA in maternal whole blood by SNP analysis after identifying candidate fetal transcripts that were upregulated in both antepartum maternal blood (at 37-40 weeks' gestation) and umbilical cord blood compared to postpartum maternal blood. We used the GSEA "preranked" feature so that we could rank the genes based on their *less* significant performance in these two different comparisons (antepartum to postpartum, and antepartum to neonatal; see Methods).

In this analysis we found that developmental *bicluster4* (Table 3C) was significantly upregulated (FDR q-value < 0.005) in both the antepartum mothers and the babies' cord blood compared to the postpartum mothers, and therefore would be considered likely to include fetal transcripts in maternal circulation. *Bicluster4* showed statistically significant overrepresentation of several GO terms, including *digestion, lipid transport,* and *lipid binding.* SP_PIR (Protein Information Resource) terms such as *intestine, glycoprotein, neuropeptide, and inflammatory response* were also overrepresented. Given that myelin membrane synthesis relies upon lipid and sterol metabolism[47], expression of these genes may reflect the maturing neurological system of the near term fetus, necessary for coordinating the complex sequence of actions needed for feeding and breathing; or it may simply reflect direct preparation for digestion. In our previous analysis of this data[20], we saw evidence of putative fetal expression of genes related to several functional processes likely to be needed at birth: immunity, sensory perception, lung maturation, and neurological function. However, no functional over-representation of digestive or metabolic proteins was detected as a set. Indeed, a painstaking manual annotation effort revealed hints that such proteins were among the likely fetal transcripts, but their significance was unclear. In contrast, the present work likely suggests that the healthy term fetus is preparing to feed.

The fact that such transcripts are detectable in maternal circulation helps support the proposal to use transcriptional analysis of maternal blood as a non-invasive approach to monitor fetal development. Translational applications of this work might include detecting potential feeding disorders before birth by identifying dysregulation of this gene set in individual fetuses.

4. Discussion

4.1 Implications

Our understanding of functional relationships among sets of genes is still in its infancy. Discovery of coherent gene sets that work together in different biological processes or disease states may help further annotate genomes by assigning function to unknown genes or discovering previously unsuspected relationships. Our method allows us to identify gene sets likely to have a common functional role in a given tissue or disease state. We found that many candidate gene sets selected in this way show statistically significant differential expression in new test data sets, suggesting that such gene sets may generalize well across tissues and relevant disease states.

Many gene set discovery methods rely upon annotation tools that utilize ontology or pathway databases. A potential issue with such functional enrichments is the dependency of *p*-values on bicluster sizes[48]. Smaller yet functionally-relevant biclusters may go unnoticed due to their insignificant enrichment *p*-values. Our approach of searching for coherent biclusters spanning conditions from multiple experiments allows us to extract biological phenotype features that generalize well across different tissues and species, even in the absence of enrichment for known functional pathways. Thus, this approach may be a way to generate a collection of gene sets relevant to the analysis of data from novel areas, where existing functional annotation is relatively incomplete.

The question of whether the enriched biclusters exhibit known functional coherence is itself of interest. The rationale behind using metabolic disease samples in our first experiment was to determine whether our method would capture meaningful functional annotation in a field where such annotation is relatively plentiful. Although one metabolic bicluster (Bicluster4) was enriched for expected metabolic terms such as UDP-glycosyltransferase activity and carbohydrate metabolism (Supplemental Table S1), we found several metabolic gene sets that were *not* statistically enriched for any informative pathway terms. This lack of enrichment may be due to the relatively small size of the metabolic biclusters. Importantly, despite the lack of enrichment, several of these biclusters were significantly differentially expressed in the test data. Furthermore, inspection of these biclusters revealed several genes with previously assigned roles in metabolic disorders. For example, consider *bicluster9*, which we found to be significantly upregulated in smooth muscles of diabetic individuals. The Phenopedia[49] component of the Human Genome Epidemiology database (HuGE Navigator)[50] suggests that several of the genes in this bicluster, including ADRA1A, ADRB1, APOC3, CACNA1A, MTHFR and TH, are disease susceptibility genes associated with cardiovascular diseases and obesity. However, no previous relationship between most of these genes was detected in the literature. These results suggest that our approach may help capture novel links among genes and between genes and phenotypes.

Equally important, several of our test data sets were from a different species than that of the original data used for biclustering. This is particularly important for biological processes such as development that rely on mammalian model systems. For example, for the developmental data set, candidate gene sets were acquired from several murine tissues: ovary, intestine, testis and optic fissure. Yet, orthologous gene sets were found to be upregulated during human development. Similarly, neuronal biclusters derived from mouse brain tissues provided information about expression in the dorsolateral prefrontal cortex of Down syndrome patients.

4.2 Future work

Future work will include obtaining a wider range of gene sets based on larger collections of training data, and exploring the impact of varying training set size or other parameters. Biclusters identified with ISA depend on the initially chosen set of genes and the threshold parameters t_G and t_C. By varying the threshold parameters and running ISA with different initial conditions, it is possible to generate a representative set of biclusters and to determine the method's sensitivity to these changes. Additionally, it is preferable to identify smaller biclusters that consist of tightly linked genes. This goal can be realized by either refining our smaller discovered biclusters or by clustering the larger ones into smaller subsets. The impact of using different biclustering methods should also be explored further. To expand the training data sets, integration of data from different microarray platforms and multiple species, though non-trivial, is feasible[51,52] and desirable. Furthermore, it is important to determine how best to select training data to facilitate discovering new gene sets for the analysis of particular test data sets. Future work might explore the effectiveness of this approach as a function of, for example, distances between MeSH terms describing the training and test data. Finally, future experiments are needed to identify and validate new functional relationships between genes that are suggested by our results.

References

1. V. R. Iyer *et al.*, *Science* **283**, 83 (Jan 1, 1999).
2. X. Wen *et al.*, *Proc Natl Acad Sci U S A* **95**, 334 (Jan 6, 1998).

3. B. R. Zeeberg *et al.*, *Genome Biol* **4**, R28 (2003).
4. D. A. Hosack, G. Dennis, Jr., B. T. Sherman, H. C. Lane, R. A. Lempicki, *Genome Biol* **4**, R70 (2003).
5. P. Khatri, S. Draghici, G. C. Ostermeier, S. A. Krawetz, *Genomics* **79**, 266 (Feb, 2002).
6. P. Khatri, S. Draghici, *Bioinformatics* **21**, 3587 (Sep 15, 2005).
7. V. K. Mootha *et al.*, *Nat Genet* **34**, 267 (Jul, 2003).
8. J. H. Hung *et al.*, *Genome Biol* **11**, R23 (2010).
9. W. T. Barry, A. B. Nobel, F. A. Wright, *Bioinformatics* **21**, 1943 (May 1, 2005).
10. L. Tian *et al.*, *Proc Natl Acad Sci U S A* **102**, 13544 (Sep 20, 2005).
11. J. J. Goeman, S. A. van de Geer, F. de Kort, H. C. van Houwelingen, *Bioinformatics* **20**, 93 (Jan 1, 2004).
12. S. W. Kong, W. T. Pu, P. J. Park, *Bioinformatics* **22**, 2373 (Oct 1, 2006).
13. U. Mansmann, R. Meister, *Methods Inf Med* **44**, 449 (2005).
14. I. Dinu *et al.*, *BMC Bioinformatics* **8**, 242 (2007).
15. J. J. Goeman, P. Buhlmann, *Bioinformatics* **23**, 980 (Apr 15, 2007).
16. A. Subramanian *et al.*, *Proc Natl Acad Sci U S A* **102**, 15545 (Oct 25, 2005).
17. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (May, 2000).
18. M. Kanehisa, S. Goto, *Nucleic Acids Res* **28**, 27 (Jan 1, 2000).
19. P. D. Karp *et al.*, *Nucleic Acids Res* **33**, 6083 (2005).
20. J. L. Maron *et al.*, *J Clin Invest* **117**, 3007 (Oct, 2007).
21. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863 (Dec 8, 1998).
22. P. Tamayo *et al.*, *Proc Natl Acad Sci U S A* **96**, 2907 (Mar 16, 1999).
23. M. Kankainen, G. Brader, P. Toronen, E. T. Palva, L. Holm, *Nucleic Acids Res* **34**, e124 (2006).
24. J. Hartigan, *J. Am. Stat. Assoc.* **67**, 123 (1972).
25. Y. Cheng, G. M. Church, *Proc Int Conf Intell Syst Mol Biol* **8**, 93 (2000).
26. S. C. Madeira, A. L. Oliveira, *IEEE/ACM Trans Comput Biol Bioinform* **1**, 24 (Jan-Mar, 2004).
27. J. Ihmels, S. Bergmann, N. Barkai, *Bioinformatics* **20**, 1993 (Sep 1, 2004).
28. A. Tanay, R. Sharan, M. Kupiec, R. Shamir, *Proc Natl Acad Sci U S A* **101**, 2981 (Mar 2, 2004).
29. X. Gan, A. W. Liew, H. Yan, *BMC Bioinformatics* **9**, 209 (2008).
30. M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, T. Muller, *Bioinformatics* **24**, i223 (Jul 1, 2008).
31. A. Keller *et al.*, *Bioinformatics* **25**, 2787 (Nov 1, 2009).
32. I. Ulitsky, R. Shamir, *Comput Syst Bioinformatics Conf* **7**, 249 (2008).
33. M. Liu *et al.*, *PLoS Genet* **3**, e96 (Jun, 2007).
34. J. Ihmels *et al.*, *Nat Genet* **31**, 370 (Aug, 2002).
35. S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, E. Zitzler, *Bioinformatics* **22**, 1282 (May 15, 2006).
36. X. Wei, PhD Dissertation, Computer Science, Tufts University (2010).
37. K. O. Cheng, N. F. Law, W. C. Siu, A. W. Liew, *BMC Bioinformatics* **9**, 210 (2008).
38. L. Teng, L. Chan, *Journal of Signal Processing Systems* **50**, 267 (2007).
39. G. Dennis, Jr. *et al.*, *Genome Biol* **4**, P3 (2003).
40. W. Huang da, B. T. Sherman, R. A. Lempicki, *Nat Protoc* **4**, 44 (2009).
41. M. Urbanek, S. Sam, R. S. Legro, A. Dunaif, *J Clin Endocrinol Metab* **92**, 4191 (Nov, 2007).
42. S. Turcan, D. K. Slonim, D. E. Vetter, *PLoS One* **5**, e9058 (2010).
43. K. E. Wisniewski, B. Schmidt-Sidor, *Clin Neuropathol* **8**, 55 (Mar-Apr, 1989).
44. G. Lubec *et al.*, *J Neural Transm Suppl* **57**, 161 (1999).
45. B. W. Brooksbank, M. Martinez, *Mol Chem Neuropathol* **11**, 157 (Dec, 1989).
46. D. K. Slonim *et al.*, *Proc Natl Acad Sci U S A* **106**, 9425 (Jun 9, 2009).
47. M. H. Verheijen *et al.*, *Proc Natl Acad Sci U S A* **106**, 21383 (Dec 15, 2009).
48. G. Li, Q. Ma, H. Tang, A. H. Paterson, Y. Xu, *Nucleic Acids Res* **37**, e101 (Aug, 2009).
49. W. Yu, M. Clyne, M. J. Khoury, M. Gwinn, *Bioinformatics*, (Oct 30, 2009).
50. W. Yu, M. Gwinn, M. Clyne, A. Yesupriya, M. J. Khoury, *Nat Genet* **40**, 124 (Feb, 2008).
51. L. Shi *et al.*, *Nat Biotechnol* **24**, 1151 (Sep, 2006).
52. J. Tsai *et al.*, *Genome Biol* **2**, SOFTWARE0002 (2001).

GENOTYPE PHENOTYPE MAPPING IN RNA VIRUSES - DISJUNCTIVE NORMAL FORM LEARNING

CHUANG WU, ANDREW S. WALSH and RONI ROSENFELD

School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA
*E-mail: chuangw@cs.cmu.edu, awalsh@cs.cmu.edu, *Roni.Rosenfeld@cs.cmu.edu*

RNA virus phenotypic changes often result from multiple alternative molecular mechanisms, where each mechanism involves changes to a small number of key residues. Accordingly, we propose to learn genotype-phenotype functions, using Disjunctive Normal Form (DNF) as the assumed functional form. In this study we develop DNF learning algorithms that attempt to construct predictors as Boolean combinations of covariates. We demonstrate the learning algorithm's consistency and efficiency on simulated sequences, and establish their biological relevance using a variety of real RNA virus datasets representing different viral phenotypes, including drug resistance, antigenicity, and pathogenicity. We compare our algorithms with previously published machine learning algorithms in terms of prediction quality: leave-one-out performance shows superior accuracy to other machine learning algorithms on the HIV drug resistance dataset and the UCIs promoter gene dataset. The algorithms are powerful in inferring the genotype-phenotype mapping from a moderate number of labeled sequences, as are typically produced in mutagenesis experiments. They can also greedily learn DNFs from large datasets. The Java implementation of our algorithms will be made publicly available.

1. INTRODUCTION

RNA viruses (including retroviruses), such as HIV, Influenza, Dengue and West Nile, impose a very significant disease burden throughout the world. Because of their very short generation time and low replication fidelity,[1] RNA viruses exhibit extensive variability at the nucleic acid and protein level which results in fast adaptation rate, and great ability to evade the immune system and antiviral drugs.[2,3] For example, HIV drug resistance has developed to all available drugs,[4] and some drug resistance mutations are probably present before the start of therapy;[5] Influenza resistance to Neuraminidase Inhibitor is rare but the resistance mutations are emerging and the resistance becoming more prevalent.[6]

Genotype-phenotype function learning is important first step in elucidating the mechanisms responsible for various viral phenotypes. It is also a crucial step towards inferring the phenotype from sequence alone, which has broad uses in clinical decision making (e.g. antiviral drug choice based on drug resistance) and in public policy (e.g. vaccine formulation based on immunogenicity and cross-reactivity).

We believe that by appropriately exploiting domain knowledge, computational methods can efficiently and correctly learn genotype-phenotype mapping. This can be combined with the large and rapidly growing sequence datasets to reduce the amount of required biological experimentation. The fast replicating RNA viruses provide us a large pool of RNA virus sequences data. There were 229,451 sequences in the HIV Sequence Database at Los Alamos National Laboratory by the end of 2007, an increase of 17% since the year before (http://www.hiv.lanl.gov/). This abundant data suggests a great opportunity for computational models to unveil the underlying mechanisms of phenotype changes. Specifically, by making best usage of the domain knowledge the models could capture the genotype-phenotype correlations and improve prediction performance. We believe that computational tools will be essential as exploratory and interpretation systems to support clinical decisions concerning the prediction of the phenotypes.[7]

RNA virus phenotypes typically result from multiple alternative mechanisms. Each mechanism is sufficient to explain the phenotypes and is constituted of a small number of key residues; yet each key residue alone may correlate weakly with the observed viral phenotype. This is biologically plausible and can be seen in a variety of biological evidences. For example, drug resistance is often a steric-structural problem, and the physical interactions with inhibitors involve more than one part of the target molecule, e.g. Protease Inhibitors (PIs) bind to four or more binding pockets in the protease substrate cleft of HIV viruses;[8] a variety of active site properties are playing roles in the binding determination, such as residue types, hydrophobicity, charges, secondary structure, cavity volume, cavity depth and area etc. Therefore, multiple, alternative potential mechanisms exist. Each mechanism involves only a small number of mutations since it has to be "discovered" by the virus via random mutations. Thus overall only a small number of key residues are involved. Therefore, a short Disjunctive Normal Form (DNF, "OR" of "AND") would be an appropriate bias over the hypothesis space under these assumptions. DNF is a disjunction of conjunctions where every variable or its negation is represented once in each conjunction. DNFs of proteins provide a mapping between key residues and their phenotype, and are an informative abstraction of key residues for the construction. Another advantage of DNF is that it is a natural form of knowledge representation for humans to interpret.

The learning of DNFs is a machine learning technique to infer Boolean function relevant with a class of interest. It has been extensively used in electric circuit design, information retrieve,[9] chess game,[10] and so on. The learnability of DNFs has been a fundamental and hard problem in computational learning theory for more than two decades. Because of the combinatorial complexity, exhaustive search algorithms for finding solutions require huge computational resources. Our group has been developing algorithms for accelerating and optimizing the DNF learning for RNA virus phenotypes, based on biologically plausible assumptions. We are also concerned with the amount of data available and the learning efficiency of the algorithms. In this study, we develop fast exhaustive DNF learning algorithms under biologically plausible assumptions. The algorithms can learn DNFs either from only a few mutagenesis experiments or from large high-throughput datasets. The learning quality is evaluated by examining the biological interpretation and prediction quality of the functions on a variety of RNA virus datasets representing different phenotypes.

2. Related computational work

Existing work on computational and statistical inference of genotype-phenotype relationship focuses on population genetics, using linkage analysis and association studies. Linkage analysis is not applicable to our case because crossover is not a significant force in the evolution of most RNA viruses. Similarly, association studies are not applicable here because they can only detect single-locus associations, or else require exceedingly large datasets: for a typical scenario where up to a few dozen labeled sequences are available and the phenotype depends on 2-4 key residues that interact in a complex fashion, there is not enough power in statistical tests to identify these residues. More specifically, tests like those described in[11–13] look for association between each individual residue position and the phenotype. But if the phenotype is determined by a complex interaction among, say, four residue positions, then there will be only moderate association between any one of these positions and the phenotype label, and this association may not be reliably detected with the limited number of labeled sequences that are usually available. This is a weakness shared by all methods that look for phenotypic association with individual residue positions (call these "*position-specific association methods*"). This deficiency on the real data was described in.[14] Although position-specific association methods can be expanded to look for phenotypic associations with any pair or triplet of residues etc., the exponential growth in the number of covariates further reduces the power of the tests. An even more serious limitation of these methods is that they assume that the labeled data were independently sampled, a patently false assumption in most cases of interest.

Rule induction algorithms, such as simultaneous covering by decision tree algorithm,[15] and ordered list of classification rules induction[16] can also mine if-then rules, but they only discover small number of rules for efficient prediction or classification purposes. Sequence analyses using logic regression[17] and Monte Carlo Logic regression[18] adaptively identify weighted logic terms that are associated with phenotypes. These approaches do not explore the whole hypothesis space to identify all possible solutions; hence it is not guarantee to learn the global optimal solution.

Many state-of-the-art machine learning approaches have been applied to RNA virus genotype-phenotype mapping, such as support vector machine (SVM) regression,[19] decision tree classification,[20] statistical models,[21] neural network,[22,23] recursive partitioning,[24] linear stepwise regression,[25] support vector regression,[8] least-squares regression,[8] and least angle regression.[8] These models learn from a training data set and then test their performance using a test data set. The effort focuses on the prediction accuracy in a cross validation manner, but these approaches lack the intention to learn biologically meaningful and interpretable functions. Nonetheless, we will comprehensively compare our DNF learning algorithms with these approaches on prediction quality.

We are not aware of any statistical or computational methods designed specifically to infer genotype-phenotype relationship in RNA viruses or other situations dominated by point mutations and small to moderate datasets.

3. Disjunctive Normal Form (DNF) learning algorithms

Disjunctive Normal Form (DNF) is a disjunction of conjunctions, where the conjunctions vary over positive and negative literals. Any given boolean function $f : \{0,1\}^d \rightarrow \{0,1\}$ can be written in an equivalent DNF. For example, a DNF formula is of the form:

$f(x_1, x_2, x_3) = x_1 \wedge x_3 + x_1 \wedge \neg x_2 \wedge x_3 + x_2$

where '$\wedge$' denotes 'AND', '+' denotes 'OR', '$\neg$' denotes negation, and 'x' is a binary literal. This example formula is a 2-term 3-DNF which contains two conjunctive terms (called clauses hereafter) with a maximum clause length of 3 literals. The size of the DNF formula is defined as the number of clauses it contains. A DNF formula represents a logic if-then rule, which is true only if the logic calculation of inputs is true. To adapt biological sequence data, the binary literals are extended to positional category variables. For example, an extended literal 'x = 5A' means 'x = Ind(the sequence item at the 5th position is 'A')', where 'Ind()' is an indicator function, and 'A' is the string representation of amino acid or nucleotide acids. This extension enables us to assign labels to any biological sequences. When a function assigns a positive label to an input sequence, we say that the function 'covers' the sequence. The goal of our DNF learning algorithm is to learn the shortest DNF(s) that cover all the positively labeled data, and do not cover any of the negatively labeled data.

Finding the minimum size DNF formula is a well-known NP-Complete problem;[26,27] hence there is no polynomial time learning algorithm. Existing practical solutions usually sacrifice completeness for efficiency. The existing heuristic or approximation approaches can be categorized into deterministic[9,28,29] and stochastic algorithms.[10,30] The deterministic methods include bottom-up schemes (learning clauses first and building DNFs in a greedy way) and top-down schemes (converting DNF learning to a Satisfiability problem). Stochastic methods randomly walk through the solution space to search for clauses but are not guaranteed to yield optimal solutions.

In this study, we aim to find the minimum size DNFs by making the assumption that only small numbers of key residues are involved in determining the functions. The assumption is biologically plausible and can be seen in a variety of RNA virus phenotypes:[14] **Drug resistance:** In Influenza, resistance to the M2 ion channel blockers amantadine and rimantadine is associated with two mutations in the M2 protein;[31] **Immunogenicity:** In HIV-1, decreased immunogenicity has been shown to be caused by three mutations in the gag protein;[32] **Pathogenicity:** In Avian Influenza, dramatically increased pathogenicity was found to be associated with a small number of mutations in the polyprotein cleavage site;[33] **Antigenicity:** In Influenza A, the investigation of the differences between the vaccine strain (A/Panama/2007/99) and the circulating (A/Fujian/411/02-like) virus showed that two mutations in the hemaglutinin protein are responsible for the antigenic drift.[34]

Using this assumption, our method:

(1) Converts DNF learning to learning k-DNF where $k \geq 1$ is the maximum size of conjunctive clauses. (Standalone DNF learning algorithm)
(2) Exhaustively learns monotone DNF(s) after feature selection (Monotone DNF learning algorithm)
(3) Greedily learns DNFs for hard problem settings.
(4) Extracts biologically meaningful solutions and better predicts phenotypes from genotypes.

3.1. *Standalone DNF learning*

Valiant[35] showed that for every constant $k \geq 1$, k term DNF can be PAC learned in polynomial time by k-CNF, i.e. CNFs with at most k literals in each clause. K-term DNF learning is essentially a combinatorial problem. The standalone DNF learning algorithm first learns a set of conjunctive clauses deterministically with the maximum clause length of k (table 1), and then constructs DNFs from the clause pool. The construction process becomes a typical SET-COVER problem (table 2) after converting each clause into a set of sequences it covers. In response, the DNF learning algorithm is equivalent to finding the minimum number of sets that cover all the positive sequences. Although the SET-COVER problem is again NP-Complete, by limiting the maximum clause length, for typical RNA virus problem settings the number of clauses is usually manageable and the SET-COVER can be exhaustively completed. Typically, the number of possible clauses of size$k$ is up to L^k, where L is the sequence length. The actual number of clauses that appear in the dataset is much smaller than this number, especially for biologically conserved datasets. After equivalence filtering (see Section 3.4), given the datasets we evaluated, the number of learned clauses is usually about several hundreds.

The standalone algorithm can be extensively used to infer DNFs from small (a couple of sequences) to medium size (hundreds of sequences) datasets, or large conserved datasets.

Table 1. Clause learning algorithm

```
Clause Learning Algorithm (S, k):
Input:
          A set S of already-available labeled sequences
          $k$: assumed upper-bound length of clauses (a small positive integer)
Steps:
          1. Enumerate all combination of literals to form conjunction clauses
          2. Record the set of (positive and negative) sequences that each clause covers
          ($n_j^+, n_j^-$)
Output:
          The set of clauses $C$ and the corresponding sequence index sets ($N^+, N^-$)
```

Table 2. DNF learning algorithm

```
Disjunctive Normal Form Learning Algorithm(C, $n^+$):
Input:
          A set C of clauses
          $n^+$: the set of positive sequence index to be covered by the clauses
Steps:
          1. Euivalence filtering (see Section 3.4)
          2. Among the clauses that cover only the positive sequences, find a minimum set of
          clauses that cover all the positive sequences:
            2a. start from the clauses that cover the positive sequences which are rarely covered
          by other clauses
              2b. repeat 2a recursively until all the positive sequences $n^+$ are covered
Output:
          The set of the shortest DNFs
```

3.2. *Monotone DNF learning after feature selection (MtDL)*

In machine learning, feature selection is a technique of selecting a subset of relevant features to build robust learning models or for prediction purposes. Here we use feature selection to choose the set of features that we believe the solution DNFs are based on, and then construct DNFs within the selected feature space. Note that the feature selection is only used to narrow the feature space, but does not infer any mapping functions. The monotone DNF learning algorithm then exhaustively builds DNFs based on the selected features. By doing this, the size of the solution space is greatly narrowed, as a result MtDL does not need to limit the maximum length of clauses and can search the hypothesis space more thoroughly than the standalone DNF learning algorithm.

The Monotone DNF learning algorithm after feature selection (MtDL) is explained in table 5. The learning algorithm first enumerates all possible literals within the selected features, and then combines them into conjunctive clauses. MtDL differs from the standalone algorithm in that it does not limit the maximum size of clauses but completely considers all possible combinations. Take a typical example: L features are selected after feature selection, where L is usually much smaller than the sequence length. In step 1 there should be at most $M = 20 * L$ possible literals in the case of protein sequences. Because the literals from the same position will not appear in the same conjunctive clause, we do not need to consider all 2^M combinations, and instead only combinatorially choose up to L literals from M. Hence the total number of clauses is at most $N = (M \text{ choose } L)$. In reality, depending on the divergence and the amount of the data, the actually number of possible literals is always much smaller than this number. Furthermore, in step 4, the N clauses will be pre-filtered by removing the clauses that cover any negative sequences. When the clause pool is ready, in step 5 the algorithm incrementally constructs the combination of clauses to be candidate DNFs and examines the coverage of sequences. MtDL starts from 1 clause, and checks the next larger number if no solution is found. The algorithm terminates when the DNFs cover all positive sequences but not any of the negative sequences. In the result section we show that in practice MtDL runs fast on real RNA virus datasets.

Feature selector: The choice of feature selector is critical to the MtDL algorithm. The best feature selector for MtDL needs to guarantee that the selected feature space is a superset of the DNF solution space, and as small as possible for fast calculation. The Combinatorial Filtering (CF) algorithm we developed[14] works seamlessly with MtDL as a feature selector. CF() efficiently identifies the smallest set of positions that completely explains the differences between classes, thus MtDL can definitely learn DNFs based on these positions. In the following evaluation, MtDL always runs together with CF(). Notwithstanding, other feature selection methods, such as LASSO, Logistic Regression with regularization, or dimension reduction methods like PCA, are also good candidate selectors, but in these cases, the coverage threshold might need to be set (section 3.5).

Table 3. Monotone DNF learning algorithm

Monotone DNF Learner (F, S):
Input:
F: A set of selected features (by CF(), for example)
S: the labeled training datasets
Steps:
1. Construct {L}, the list of literals in the features (e.g. 5A).
2. Throw out L that does not cover any positive sequences.
3. Combinatorial construct {Clauses}, the list of conjunctive clauses from {L}, (e.g. $5A \wedge 8C$). The possible combinations are $|L|$ chooses $1, 2, .., |F|$.
4. Throw out the conjunctive clauses that cover any negative sequences.
5. Incrementally construct {DNF}, the list of disjunctive normal form that covers all positive sequences but no negative sequences: starts from 1 clause, construct DNFs from {Clauses}, try the next larger number if no solution learned.
Output:
The set of the shortest DNFs

3.3. *Greedy versions of both algorithms*

As we will show in the result section, with typical RNA virus datasets, the standalone and monotone DNF learning algorithms learn DNFs efficiently. In cases of very large dataset, both algorithms are modified to greedy versions to learn DNFs rapidly. The greedy versions only differ from the exhaustive versions in the DNF construction step. Instead of exhaustively combining all clauses to construct DNFs, the greedy algorithms iteratively select the clause that covers the largest number of the uncovered positive sequences until all the positive sequences are covered.

3.4. *Equivalence filtering*

Computationally equivalent clauses cover the same set of sequences while differing in their composition literals. They are equivalent in DNF functions in the sense that replacing one clause with its equivalent clauses will not change the predictions of the DNF on the same training set. Equivalent clauses are very common in RNA virus datasets; therefore, during DNF learning process equivalent clauses are filtered and only one of them is used as the representative to construct DNFs. By using equivalence filtering the DNF learning running time is greatly reduced. Note that the equivalence filtering is only for computational efficiency purpose. After learning DNFs, all clauses that have equivalent clauses will be expanded to recover all the DNFs.

3.5. *Avoiding over-fitting and robustness to noise*

The following two techniques are used to avoid over-fitting and make the algorithms robust to noise:

(1) **DNF pruning**: similar to the pruning of decision tree, after learning DNFs the clauses that only cover a small number of sequences may be pruned if removing them results in an increase of the prediction accuracy on the test dataset. The advantages of pruning are:
   - Avoiding over-fitting, because irrelevant clauses are removed.
   - The DNFs are shorter, which makes them easier to understand and more biologically meaningful.
   - Robust to noise, because pruned DNFs ignore the clauses/literals rendered meaningless by noise.

(2) **Threshold setting**: set thresholds of the fractions of the sequences that the learned DNF(s) cover. The DNF learning algorithms can be easily modified to terminate when at least a fraction p of the positive sequences are covered, and at most a fraction n of the negative sequences can be covered by the DNFs. The thresholds p and n are determined in a cross-validation way. Similar to pruning, the threshold setting method can also avoid over-fitting, learn shorter DNFs and be robust to noise. One advantage of threshold setting over pruning method is that threshold setting usually achieves better prediction quality.

3.6. *Extension of literals*

The literals can also be extended to negation of one amino acid or a subset of the amino acids.

3.7. *Extension to multiple class data*

The algorithm is applicable to multiple class data by running multiple times with each time one of the classes is made positive class and the rest are merged as negative class.

4. RESULTS

4.1. *Demonstrating DNF learning algorithms consistency*

The DNF learning algorithms will first be validated on simulated protein sequences with hypothetical target functions. We will use this stage to validate the algorithms' consistency. When generating simulated sequences, we match the position-specific amino acid distributions to those of a real protein datasets, and then generate random phenotypic target functions (making sure they did not label the entire dataset with the same value). We use 732 HIV-1 gp160 protein sequences (downloaded from LANL), and assumed a variety of target functions (e.g. $(70a \wedge 9l \wedge 11p \wedge 70t) + (62l \wedge 45m \wedge 36y \wedge 9l) + (62P \wedge 53V \wedge 36s) + (83i \wedge 45I) = +)$. Each target function contains a number of clauses, and it is used to label the sequences accordingly. Notice that this method enables us to generate as many sequences as we want so we can test the algorithm convergence under a variety of conditions. We repeated this process many times, and in all of these cases both standalone algorithm and MtDL algorithm converged to the target functions with moderate number of sequences.

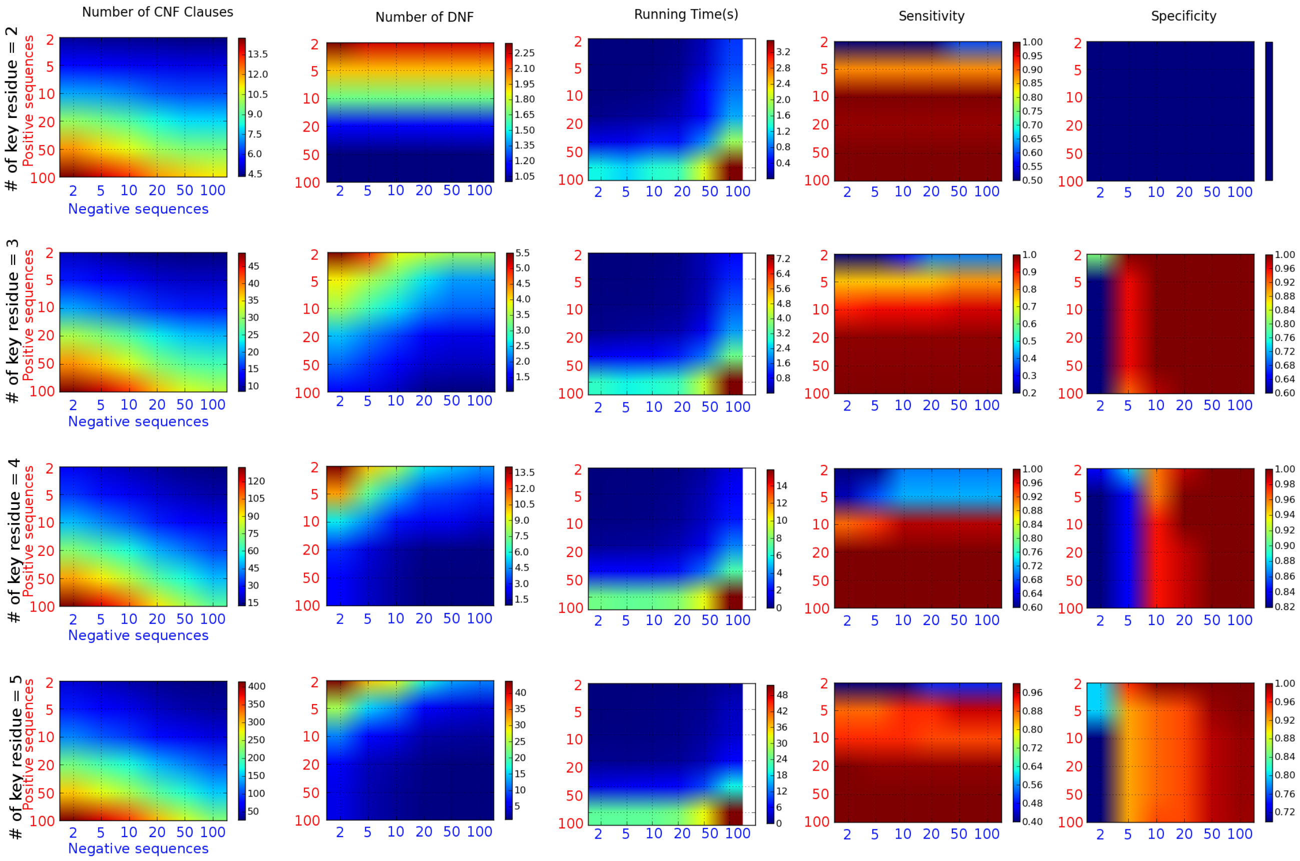


Fig. 1. **The evaluation of MtDL algorithm on simulated sequences.** From left to right, the number of CNF clauses, the number of DNFs, running time, prediction sensitivity and specificity are plotted as functions against the number of key residues assumed in the target function (rows), and the number of positive sequences and negative sequences (vertical columns and horizontal rows of small colored squares). The numerical values of the colors are shown in the colorbar. Take the top left chart for example, when the key residues are assumed to be 2 in the target function, with say 100 positive and 2 negative sequences used, the number of CNF clauses is about 12 (red color means higher value as indicated in the colorbar).

4.2. *Measuring inference efficiency (convergence rate as function of dataset size)*

To assess the efficiency of our learning method, we would like to understand the relationship between the complexity of the function to be learned and the number of training examples needed to converge to it. Namely, we would like to know the convergence rate as a function of the amount and type of available sequences and the complexity of the genotype-phenotype mapping. This is important because the available number of sequences vary for different datasets. Based on the convergence rate we can assess the likelihood of convergence, and whether (and how much) further experimentation will be needed.

To do this, we used 588 aligned sequences of HIV protease protein downloaded from the Stanford HIV database, with an aligned length of 99. We then randomly generated putative binary target functions, each depending on a small number (2..5) of literals. For each such target function, the 588 sequences were labeled accordingly. The DNF learning algorithm was then run 20 times, each time assuming a different target function to produce a statistically robust result. As an illustration, we will show the evaluation results of MtDL in the following sections, and the simulation result of the standalone algorithm is similar.

Convergence Rate: As described in the introduction, we are concerned with the relation between the amount of available data and the convergence of the algorithm. It is therefore most meaningful to compare the convergence of the hypothesis space under our algorithm with this method. Figure 1 shows the number of DNFs learned by the algorithm as a function of the number of literals in the target function (# of key residues, top-to-bottom), and the number of positive and negative sequences (vertical and horizontal rows of small colored squares, respectively, with values of 2, 5, 10, 20, 50, 100 sequences each). Blue color indicates convergence (i.e. one single DNF is learned given the amount of data, and the learned DNF is exactly the same as the target function), and red color indicates alternative DNFs exist. The number of DNFs reduces with more available sequences, and in all of the cases, MtDL algorithm converges with only about 50 positively labeled and 50 negatively labeled sequences.

Another important factor in measuring efficiency is the running time due to the combinatorial nature of DNF learning. Recall that in MtDL, the number of candidate clauses in step 4 is bounded by 2^L and the number of DNFs is bounded by 2^{2^L}, where L is the number of literals. L increases when more sequences are available, and this explains why, in the "running time" column the red color, which indicates the longest running time, is at the right bottom corner. Note that the longest running time is still on the order of seconds in the simulation.

The prediction sensitivity and specificity showed that the algorithms converge with only moderate numbers of sequences. Interestingly, the prediction quality chars are symmetric in that if we flip the labels of the data, the prediction accuracy will remain the same. This is important because although our DNF learning algorithms identify DNFs that only cover the whole positive space, the sequences in both classes contribute equally to the learning.

4.3. *Retrospectively validating DNF learning algorithms when ground truth is known*

We retrospectively validated the MtDL algorithm by testing it on datasets with real viral protein sequences where the genotype-phenotype mapping is already known and assumed to be correct. We compiled a number of datasets covering several RNA viruses with varying degree of average sequence identity (SI) and a variety of phenotypes, including Avian Flu High/Low pathogenicity (4 mutations in HA proteins changed the pathogenicity from low to high in H5N2 Influenza HA, SI: 95%),[33] Influenza H3N2 antigenicity shift (2 mutations in HA shifted the antigenicity of Influenza H3N2, SI: 93%),[34] SIV Env neutralizability (2 mutations in SIV Env proteins determined the neutralizability of SIV, SI: 99%),[36] FIV tropism in CRFG cells (2 mutations in FIV polymerase PA subunit made it unable to replicate in CRFK cells, SI: 95%).[37] These conclusions were made from mutagenesis experiments that were chosen empirically or by domain knowledge. We applied our MtDL+CF (using CF as the feature selector) algorithm on the same set of mutagenesis sequences to predict the key residues for the phenotype changes. For all the tests performed, our algorithm converged to the correct answer(s). In contrast, the conventional position-specific association method we selected as comparison,[11] can only predict positions of importance, but our MtDL+CF algorithm explicitly learns the actually mapping functions. Even so when only comparing the positions identified, the conventional method only correctly identifies the positions in one of the datasets, and yields high false positive and false negative rates in the other four datasets (Table 4). This demonstrates our arguments in section 2 that if the phenotype is determined by a complex interaction, the traditional methods cannot detect all the key residues correctly.

4.4. *The utility of DNF learning algorithms on large datasets*

To demonstrate the applicability of MtDL to learn biologically meaningful results, MtDL+CF was applied to a large, divergent dataset, the HIV drug resistance dataset (download from Stanford HIV database). The dataset was retrieved from the Stanford HIV database, including seven Protease Inhibitor drugs and eleven Reverse Transcriptase Inhibitor drugs. We ran the MtDL algorithm on all the drug datasets and learned very short and interpretable DNFs (Table 5). For example, a protein is resistant to NFV when position 9 is I or (position 63 is I and position 9 is not I and

Table 4. **Comparing DNF learning with position-specific association methods** Retrospective comparisons of the DNF learning algorithm on a variety of datasets

Data set name (# pos/# neg seq)	**Golden standard (identified mutations)**	**DNF(s) learned by MtDL**	**Positions identified by Traditional method**
H3N2 hema Antigenicity shift (490pos/421neg)	145H, 146Q	145H+146Q	18, 67, 122, 145, 146
H5N2 hema Pathogenicity (4pos/11neg)	275K, 275T, 323K, 324R, 325K	323K+324R+325K	275, 323, 324, 325
FIV tropism (3pos/7neg)	30E, 32K	30K∧32E	32
SIV Envelope Neutralizability (8pos/5neg)	179N, 337R	179N+337R	331, 348

position 87 is not N) .

The purpose of these concise DNFs is three folds: 1) to identify the key positional determinants of drug resistance, e.g. position 89I for RTV, position 36 and 45 for APV,etc.. These positions have been identified by experiments and reported in literatures; 2) quantitatively describe how the residues in these positions combine to produce resistance; and 3) proposed new interpretation of HIV drug resistance mechanisms that have potential to be validated by domain experts.

Table 5. DNFs learned from HIV drug resistance dataset

Drug type	Drug	DNF = sensitivity/specificity; in the DNFs, lowercases mean negation
PI	NFV	$9l + (63l \wedge 9i \wedge 87n) = 0.902/0.834$
	RTV	$(81i \wedge 81v) + 83i + (70a \wedge 70l \wedge 89l \wedge 70t) = 0.981/0.988$
	LPV	$(9l \wedge 53i) + (9l \wedge 45m \wedge 9h \wedge 9m) = 0.961/0.965$
	APV	$(36s \wedge 45m \wedge 36y \wedge 9l) = 0.787/0.961$
	IDV	$(70a \wedge 9l \wedge 11p \wedge 70t) + (62l \wedge 45m \wedge 36y \wedge 9l) + (62P \wedge 53V \wedge 36s) + (83i \wedge 45I) = 0.965/0.982$
	SQV	$(9r \wedge 83i) + (62q \wedge 53i \wedge 89l \wedge 70l) + (45i \wedge 89l \wedge 70t \wedge 70a) + (76V \wedge 89l \wedge 81v \wedge 9l) = 0.899/0.963$
	ATV	$(70a \wedge 9l) + (89l \wedge 76V \wedge 81a) = 0.833/0.910$
NRTI	DDI	$150M + (68s \wedge 68t \wedge 68d \wedge 68n) + (74v \wedge 42n \wedge 74t) = 0.738/0.985$
	AZT	$(73v \wedge 34 - \wedge 214t \wedge 214d) = 0.848/0.988$
	D4T	$(209l \wedge 214d \wedge 34t \wedge 214t) + (68t \wedge 34m \wedge 214d \wedge 214t) + (68T \wedge 117I \wedge 66N) = 0.797/0.956$
	TDF	$(34i \wedge 66N \wedge 214t \wedge 183v) + (68g \wedge 19r \wedge 214Y \wedge 68t) + (34V \wedge 68t \wedge 214f \wedge 214t) = 0.784/0.980$
	ABC	$183V + (214d \wedge 121p \wedge 209l \wedge 82k) + (82k \wedge 66d \wedge 214Y \wedge 180c) = 0.940/0.944$
NNRTI	NVP	$(102k \wedge 102r) + (189g \wedge 102n) + (180y \wedge 100e) = 0.868/1.0$
	DLV	$(210t \wedge 102N) + (180y \wedge 100q \wedge 226F \wedge 210t) = 0.915/0.994$
	EFV	$(102r \wedge 102k) + (102s \wedge 189g) = 0.871/0.997$

4.5. *Improved prediction performance of DNF learning algorithms on the HIV drug resistance problem*

To demonstrate the prediction power of our DNF learning algorithms, we examine the prediction quality of both standalone and MtDL learning algorithms on two well-known Biology datasets respectively: the HIV drug resistance dataset and the UCI promoter gene dataset (details in section 4.6). Many state-of-the-art machine learning models, such as Support Vector Machine, Decision Trees, Neural Networks, Nave Bayes etc, have been tested on the datasets to learn genotype-phenotype mapping, and the five-fold cross-validation prediction quality was reported.[8] In the HIV drug resistance dataset, drug resistance levels are defined as fold-increased resistance compared to the wild type virus strain; therefore, for classification models the numerical resistance values were converted to multiple class labels by setting resistance thresholds. We use the thresholds suggested on the website, and select the sequences with significant resistant values and susceptible values while ignoring those with weak resistance or weak susceptible labels. The five-fold cross-validation prediction accuracies of Protease Inhibitor are shown in table 6. The standalone DNF learning algorithm outperforms other machine learning algorithms in 4 out of the 7 PI datasets (Table 6). The result suggests that by exploiting domain-knowledge to reduce the running time, the exhaustive

Table 6. Comparing Standalone DNF learning with published machine learning algorithms on HIV Protease Inhibitor datasets. The numbers of positively labeled and negatively labeled sequences in the datasets are shown, as well as and the prediction accuracies of 1) Standalone DNF, 2) Z-score,[11] 3) Naive Bayes (from Weka), 4) SVM (svm light software, default parameters), 5) Decision Tree (Weka, ID3 algorithm), 6) Winnow (Weka). The highest accuracy of each drug is highlighted in bold

Prediction accuracy (%)	NFV	SQV	IDV	RTV	APV	LPV	ATV
#pos/#neg sequences	194/211	119/321	115/279	154/244	47/308	103/142	42/111
Standalone DNF	93.5	**91.8**	**91.7**	96.1	**96.1**	88.2	**93.3**
Z-score	74.6	87.3	**91.7**	87.4	92.3	90.5	88.8
NaiveBayes	**95.1**	75.1	78.4	93.2	87.3	92.7	73.1
SVM (svm light)	77.2	74.2	83.4	92.2	87.5	86.3	72.6
DT	94.0	89.0	90.1	**98.6**	91.8	**98.6**	78.5
Winnow	91.1	84.7	89.9	94.6	91.1	94.6	85.9

Table 7. Comparing MtDL+CF with published machine learning algorithms on Promoter Gene dataset

System	Errors	Comments
MtDL + CF	4/106	No domain knowledge required
KBANN	4/106	A hybrid ML system that uses domain knowledge to initial the network structure
BP	8/106	Std backprop with one hidden layer
O'Neill	12/106	Ad hoc technique from the bio. lit.
Nearest neighbor	13/106	k-nearest neighbor, $k = 3$.
ID3	19/106	Quinlans decision-tree builder

algorithms achieve better prediction performance, and DNF turns out to be a reasonable bias on the hypothesis space as genotype-phenotype mapping functions for HIV drug resistance.

4.6. *Improved prediction performance on the UCI Promoter Gene dataset*

Another dataset we use to evaluate our DNF learning algorithms' prediction power is the popular UCI's promoter gene dataset, which has been studied with many machine learning models. The task is to predict promoters from DNA sequences of nucleotides, A, C, G, or T. The dataset contains 53 promoter sequences and 53 non-promoter DNA sequences. In Biology, the promoters are characterized by special motifs at certain positions from the transcription starting location, e.g. "cttgac" motif at +37 position indicates a promoter region. However, deriving all such domain theories is impractical and not meaningful. Computationally machine learning algorithms showed promising prediction performance on this dataset (Table 7). Among them the knowledge-based artificial neural network (KBANN)[38] achieves the best accuracy of 4 out of 106 errors in a held-out test manner.

The KBANN model is a hybrid system of both Explanation-based learning (EBL)(a system that corporate pre-existing knowledge) and Empirical learning system (learning solely from training examples). In[38] they argue that the hybrid system should be superior, in terms of classification accuracy, to empirical learning systems. On the Promoter Gene dataset, KBANN learns a neural network model and translates a set of domain theories to initial the neural network structure. The error rate is the number of wrongly predicted examples in a leave-one-out cross-validation (LOOCV) manner. Three other machine learning algorithms, standard back propagation, Quinlan's ID3, O'Neill's ad hoc partial pattern matching, and the "nearest neighbor" are compared in Table 7.

We employed the same LOOCV method on MtDL algorithm, and selected CF() as the feature selector. Although the prediction performance of MtDL+CF is the same as the best one KBANN, MtDL+CF does not require any pre-existing domain knowledge, which is not always available.

5. CONCLUSIONS

We developed two efficient DNF learning algorithms under the assumption that DNF is an appropriate bias over the hypothesis space for RNA virus phenotype datasets. The assumption is biologically plausible and very important to our algorithms, because it reduces the hypothesis space greatly to make the computational hard problem solvable. We also demonstrated the learning efficiency and consistency on simulated sequences, showed the strength of the methods in learning biological meaningful mapping functions and showed superior prediction accuracies to positional-specific methods and other machine learning methods.

We aimed to learn the minimum size DNFs even though the exact learning is NP-complete. Compared to existing heuristic algorithms that only focus on learning time and learnability, we exploit the domain knowledge and develop efficient exhaustive algorithms to learn the shortest DNFs. We also applied a number of techniques to accelerate the DNF learning process, including setting the maximum length of clauses in standalone algorithm, using feature selector (CF) in MtDL to narrow down the searching space, equivalence filtering of the clauses, and extending both algorithms to greedy versions. This enables the algorithms to run over very large datasets. Notwithstanding, as shown in the result section, the DNF learning algorithms are also powerful in extracting DNFs from only a small numbers of sequences.

We focus on the learning from mutagenesis data where the data is highly reliable and the alignment is well defined. In the cases of noisy data and low quality alignment when combinatorial algorithms usually suffer more than statistical models, we use pruning and thresholds to make the algorithms robust to noise.

Our goal in this work has been to aid biological investigation by learning the genotype-phenotype mapping. Since this is our focus, we compared our method to other methods designed to do the same. Our algorithms explicitly learn genotype-phenotype mappings that are interpretable to humans, so that the mapping functions can not only predict phenotypes from genotypes along, but also unveil biologically meaningful explanations. The algorithms can learn DNFs from different sizes of data: ranging from a few sequences to large high-throughput datasets, and show superior prediction performances. In contrast, given the limited data, the positional-specific association methods would be ineffective if they were to be applied to the full set of protein positions because there is not enough statistical power for the inference. Given full size of dataset, our DNF learning algorithms outperformed other published machine learning algorithms on two common datasets.

We successfully demonstrated the learning efficiency and the prediction power of our DNF learning algorithms on RNA virus datasets, and the algorithms can be extensively used on other domains where similar assumptions hold.

6. ACKNOWLEDGEMENTS

We are grateful to Dr. Elodie Ghedin, Dr. Ted Ross and Dr. Kelly Cole for useful discussions.

References

1. J. W. Drake and J. J. Holland, *PNAS* **96**, 13910 (1999).
2. D. W. Klein, L. M. Prescott and J. Harley, *Microbiology* (Dubuque, 1993).
3. E. Domingo and J. Holland, *Annu. Rev. Microbiol* **51**, 151 (1997).
4. D. Pillay and M. Zambon, *BMJ* **317**, 660 (3 1998).
5. R. W. Shafer, *http://hivinsite.ucsf.edu/* (2004).
6. M.-A. Rameix-Welti, V. Enouf, F. Cuvelier, P. Jeannin and S. van der Werf, *PLoS Pathogens* **4**, 1 (2008).
7. A. Carvajal-Rdriguez, *Recent Patents on DNA and Gene Sequences* **1**, 63 (2 2007).
8. S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag and R. W. Shafer, *PNAS* **10**, 53 (2006).
9. S. N. Sanchez, E. Triantaphyllou, J. Chen and T. Liao, *Computer & Operations Research* **29**, 1677 (2002).
10. U. Ruckert and S. Kramer, 648 (2003).
11. S. S. Hanenhalli and R. B. Russell, *JMB* **303**, 61 (2000).
12. L. A. Mirny and M. S. Gelfand, *JMB* **321**, 7 (2002).
13. F. Pazos, A. Rausell and A. Valencia, *Bioinformatics* **22**, 1440 (2006).
14. C. Wu, A. Walsh and R. Rosenfeld, *IEEE-BIBE* (2010).
15. J. R. Quinlan, Induction of decision trees, in *Readings in Machine Learning*, eds. J. W. Shavlik and T. G. Dietterich (Morgan Kaufmann, 1990) Originally published in.
16. P. Clark and R. Boswell, *Proceedings of the Fifth European Working Session on Learning* **482**, 151 (1991).
17. C. Kooperberg, I. Ruczinski, M. L. LeBlanc and L. Hsu, *Genetic Epidemiology* **21**, 626 (2001).

18. C. Kooperberg and I. Ruczinski, *Genetic Epidemiology* **28**, 157 (2005).
19. N. Beerenwinkel, T. Lengauer, J. Selbig, B. Schmidt, H. Walter, K. Korn, R. Kaiser and D. Hoffmann, *IEEE* **16**, 35 (11 2001).
20. N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn and J. Selbig, *PNAS* **99**, 8271 (6 2002).
21. G. DiRienzo1, V. DeGruttola1, B. Larder and K. Hertogs, *Statistics in Medicine* **22**, 2785 (2003).
22. S. Draghici and R. B. Potter, *Bioinformatics* **19**, 98 (1 2003).
23. D. Wang and B. Larder, *The Journal of Infectious Diseases* **188**, 653 (3 2003).
24. A. D. Sevin, V. DeGruttola, M. Nijhuis, J. M. Schapiro, A. S. Foulkes, M. F. Para and C. A. B. Bouche, *The Journal of Infectious Diseases* **182**, 59 (2000).
25. K. Wang, E. Jenwitheesuk, R. Samudrala and J. E. Mittler, *Antiviral Therapy* **9**, 343 (2004).
26. Brayton (1985).
27. Gimpel (1965).
28. E. Triantaphyllou and A. L. Soyster, *Computers & Operations Research* **23**, 783 (1999).
29. Kamath, *Mathematical Programming* **57**, 215 (2005).
30. U. Ruckert, S. Kramer and L. D. Raedt, 405 (2002).
31. R. B. Belshe, M. H. Smith, C. B. Hall, R. Betts, and A. J. Hay, *Journal of virology* **62**, 1508 (5 1988).
32. S. Andre, B. Seed, J. Eberle, W. Schraut, N. Bultmann and J. Haas, *Journal of Virology* **74**, 2628 (2000).
33. M. Garcia, J. M. Crawford, J. W. Latimer, E. Rivera-Cruz and M. L. Perdue, *Journal of General Virology* **77**, 1493 (1996).
34. H. Jin, H. Zhou, H. Liu, W. Chan, L. Adhikary, K. Mahmood, M. Lee and G. Kemble, *Viology* **336**, 113 (2005).
35. L. G. Valiant, *Proceedings of the sixteenth annual ACM symposium on Theory of computing table of contents* , 436 (1984).
36. K. Cole and T. Ross, *Current HIV Res.* **5**, 505 (2007).
37. E. J. Verschoor, L. A. Boven, H. Blaak, A. Vliet, M. C. Horzinek, and A. Ronde, *American Society for Microbiology* **69**, 4752 (8 1995).
38. G. G. Towell, J. W. Shavlik and M. O. Noordewier, *AAAI-90 Proceedings* (1991).

GENOME-WIDE ASSOCIATION MAPPING AND RARE ALLELES: FROM POPULATION GENOMICS TO PERSONALIZED MEDICINE

FRANCISCO M. DE LA VEGA
Life Technologies
Foster City, CA 94403, USA
Email: Francisco.delavega@lifetech.com

CARLOS D. BUSTAMANTE
Department of Genetics, Stanford University
Stanford, CA, USA
Email: cdbustamante@stanford.edu

SUZANNE M. LEAL
Department of Molecular and Human Genetics, Baylor College of Medicine
Houston, TX 77030, USA
Email sleal@bcm.edu

Genome-wide associations studies (GWAS) have been very successful in identifying common genetic variation associated to numerous complex diseases [1]. However, most of the identified common genetic variants appear to confer modest risk and few causal alleles have been identified [2]. Furthermore, these associations account for a small portion of the total heritability of inherited disease variation [1]. This has led to the reexamination of the contribution of environment, gene-gene and gene-environment interactions, and rare genetic variants in complex diseases [1, 3, 4]. There is strong evidence that rare variants play an important role in complex disease etiology and may have larger genetic effects than common variants [2].

Currently, much of what we know regarding the contribution of rare genetic variants to disease risk is based on a limited number of phenotypes and candidate genes. However, rapid advancement of second generation sequencing technologies will invariably lead to widespread association studies comparing whole exome and eventually whole genome sequencing of cases and controls. A tremendous challenge for enabling these “next generation” medical genomic studies is developing statistical approaches for correlating rare genetic variants with disease outcome.

The analysis of rare variants is challenging since methods used for common variants are woefully underpowered. Therefore, methods that can deal with genetic heterogeneity at the trait-associated locus have been developed to analyze rare variants. These methods instead analyzing individual variants analyze variants within a region/gene as a group and usually rely on collapsing. They can be applied to both in cases vs. controls and quantitative trait studies are needed. The paper of Bansal et al. in this volume describes the application of a number of statistical methods for testing associations between rare variants in two genes to obesity. The authors considered the relative merits of the different methods as well as important implementation details, such as the leveraging of genomic annotations and determining p-values.

Knowledge of haplotypes can increase the power of GWAS studies and also highlight associations that are impossible to detect without haplotype phase (e.g. loss of heterozygosity).

Even more complicated phase-dependent interactions of variants in linkage equilibrium have also been suggested as possible causes of missing heritability. In their work, Hallsorsson et al. formulate algorithmic strategies for haplotype phasing by multi-assembly of shared haplotypes from next-generation sequencing data. These methods would allow testing haplotypes harboring rare variants for association and potentially increase their explanatory power.

Since single SNP tests are often underpowered in rare variant association analysis, Zeggini and Asimit propose a locus-based method that has high power in the presence of rare variants and that incorporate base quality scores available for sequencing data. Their results suggest that this multi-marker approach may be best suited for smaller regions, or after some filtering to reduce the number of SNPs that are jointly tested to reduce loss of power due to multiple-testing adjustments.

Finally, the paper of Zhou et al., presents a penalized regression framework for association testing on sequence data, in the presence of both common and rare variants. This method also introduces the use of weights to incorporate available biological information on the variants. Although these tactics improve both false positive and false negative rates, they represent an incremental development and there is still significant room for improvement.

With the development of sequencing technologies and methods to detect complex trait rare variant associations many new and exciting discovery are imminent. The analysis of rare variants is still in its infancy and the next few years promises to produce many new methods to meet the special demands of analyzing this type of data.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: *Proc Natl Acad Sci U S A* 2009, **106**(23):9362-9367.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al*: *Nature* 2009, **461**(7265):747-753.
3. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: *PLoS Biol* 2010, **8**(1):e1000294.
4. McClellan J, King MC: *Cell* 2010, **141**(2):210-217.

AN APPLICATION AND EMPIRICAL COMPARISON OF STATISTICAL ANALYSIS METHODS FOR ASSOCIATING RARE VARIANTS TO A COMPLEX PHENOTYPE

VIKAS BANSAL*, ONDREJ LIBIGER*, ALI TORKAMANI*

The Scripps Translational Science Institute (VB, OL, AT) and Department of Molecular and Experimental Medicine,
The Scripps Research Institute (AT); 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037 USA
Email: vbansal@scripps.edu; olibiger@scripps.edu; atorkama@scripps.edu

NICHOLAS J. SCHORK†

The Scripps Translational Science Institute and Department of Molecular and Experimental Medicine, The Scripps
Research Institute; 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037 USA
Email: nschork@scripps.edu

The contribution of collections of rare sequence variations (or 'variants') to phenotypic expression has begun to receive considerable attention within the biomedical research community. However, the best way to capture the effects of rare variants in relevant statistical analysis models is an open question. In this paper we describe the application of a number of statistical methods for testing associations between rare variants in two genes to obesity. We consider the relative merits of the different methods as well as important implementation details, such as the leveraging of genomic annotations and determining p-values.

1. Introduction

1.1. *Rare variants and the 'hidden heritability' of complex traits*

Genome wide association (GWA) studies have been pursued for many diseases and phenotypes. Although the results of these studies have been mixed, with some studies identifying more compelling associations than others, virtually all of these studies have resulted in the discovery of variants that collectively only explain a small fraction of the heritable component of the diseases and phenotypes they have considered [1]. This fact has not only raised important questions about the degree to which common variants, which are typically of focus in GWA studies, influence phenotypic expression, but also the best way to identify factors not detectable via current common-variant-based GWA study protocols [2] that contribute to a 'hidden heritability' behind phenotypic expression.

Recently it has been argued that collections of rare variants could contribute to phenotypic expression over-and-above common variants [3-4]. The intuition behind this argument is that although each rare variant may have a small overall effect on phenotypic expression, collectively these variants may have a moderate or even more pronounced effect [3-4]. Rapid developments in high-throughput DNA sequencing technologies are likely to facilitate searches for rare variants that may influence phenotypic expression, but are not the only item necessary for a successful study of rare variants. Also needed are appropriate study designs and subject sampling methods, data analysis methods, and ways of validating or conceptualizing the biological influence of

[*] VB, OL and AT contributed equally to this paper.
[†] To whom correspondence should be addressed: nschork@scripps.edu

multiple rare variants on phenotypic expression once they are found to be associated with a phenotype.

In this paper we describe a number of different statistical methods for testing the hypothesis that collections of rare variants are associated with a qualitative phenotype in a case/control sampling setting. These methods build off the notion of ‘collapsing’ a number of rare variants into a single set whose collective frequency is contrasted between case and control groups [5-6]. Many approaches involve regression or regression-like models in which dummy variables indicating the presence (i.e., individuals assigned a dummy variable value of 1.0) or absence (0.0) of a variant are used. For the collapsed set of variants, an individual is ultimately assigned a value of 1.0 if they have any of the rare variants among a larger set and 0.0 otherwise. This collapsed dummy variable can then be tested for association by testing the regression coefficient associated with the dummy variable [7]. Other regression approaches consider the effects of each individual variant, no matter how rare, as well as collapsed sets of variants [8]. We apply these and other methods to a case/control study of obesity and compare the results of the application of each. We also consider extensions of the proposed statistical analysis methods.

Before describing the data set, statistical methods, and the results of their application, however, we provide brief descriptions of two overarching frameworks for the study of the collective effects of rare variants on phenotypic expression: one leveraging functional genomic annotations and one considering the collective effects of variants in defined contiguous genomic regions.

1.2. *Collapsing variants based on functional annotations*

Testing collections of rare variants for association to a phenotype requires some way of grouping or collapsing variants into a coherent set; i.e., defining the set whose collective frequency is tested for association. This can be approached by defining a set based on functional annotations associated with the genomic regions harboring the variants to be tested for association. For example, one could test the collective frequency differences of coding variants, non-synonymous coding variants, variants in known transcription factor binding sites, or conserved sites, between cases and controls. Such groupings could lead to easily interpreted biological associations but, ultimately, would only be as good and reliable as the annotations used.

1.3. *Moving window analysis*

An alternative to defining sets of collapsed variants based on functional genomic annotations is to consider all the variants in a genomic subregion defined by its size and test these variants for association. Such subregions could then be systematically tested over the entire genomic region of interest. By starting at one end of a genomic region of interest, testing variants within the ‘window’ defined by the subregion, and then moving the window to an adjacent subregion, testing that subregion, and continuing this process until the entire region is covered would provide a test of the hypothesis that some subregions within the broader region of interest harbor collections of variants associated with a phenotype. This moving window approach can be repeated with different window sizes, including overlapping windows, but at the cost of increased type I error due to the multiple tests.

1.4. *Accommodating other sources of variation and assessing statistical significance*

In any test of genetic association there are a few things that need to be considered. For example, stratification issues need to be accommodated or controlled for. This can be done by ensuring that the subjects used in a study are matched for genetic background or the statistical test used is appropriately adjusted for potential stratification [9]. In addition, in order to assess the statistical significance of an association study involving multiple variants within a genomic region, appropriate control for multiple comparisons must be made [10]. Finally, accommodating covariate effects (e.g., gender, age, other genetic factors, ancestry information, etc.) in association analysis is important, but may not be trivial for many statistical models. Thus, gauging the ability of different statistical analysis models to accommodate covariates may be of particular importance in rare variant analysis settings.

2. Sequencing the MGLL and FAAH genes in obese and control individuals

2.1. *DNA sequencing and sample selection*

Genomic intervals covering two genes that encode the endocannabinoid metabolic enzymes, FAAH and MGLL, were sequenced in 289 individuals of European ancestry using the Illumina GA sequencer. Ancestry was determined using a panel of ancestry informative markers and individuals with an outlying genetic background were removed from the analysis. Sequencing was done using 36 base pair reads. The median coverage was 60X across the individuals sequenced. The program MAQ was used for alignment and variant calling, resulting in 1410 high quality single nucleotide variants (SNVs; 228 in the FAAH gene and 1182 in the MGLL gene) which were used for association analysis. The sequenced regions were captured using long range PCR and represented a total of 188,270 nucleotides. The 289 individuals included 147 normal controls (Body Mass Index (BMI) <30) and 142 extremely obese cases (BMI >40).

2.2. *Genomic annotations, window definitions, and multiple comparisons*

We leveraged genome annotations from the UCSC genome browser to identify sets of variants that reside in functionally-relevant regions of the genome. We identified sets of variants that reside within 5 different functional elements within the MGLL and FAAH genes: non synonymous SNVs (‘NS’), H3K27 acetylation sites, Fox2 interaction sites, Amidase protein domains, and all transcription factor binding sites (‘TFBS’). Variants within these elements were collapsed and tested for association with obesity. For the moving window analyses, we considered window sizes of 5 kb over the two genes. In order to accommodate multiple comparisons we identified the effective number of independent variants based on linkage disequilibrium (LD) using the method discussed by Nyholt [11]. This number provides a very rough approximation for the number of tests to be corrected for and was found to be 584 for our data. We assumed a nominal type I error rate of 0.05 in assessing statistical significance of variant associations, so our approximate multiple comparisons corrected p-value was 0.05/584 = 0.000086 (-log(p-value) = 4.06). Obviously, more sophisticated strategies for correcting for multiple comparisons, including possibly permuting cases and controls and repeating the entire moving window and functional annotation-based collapsed set analyses, need to be investigated.

3. Statistical methods for rare variant associations

We briefly describe 11 methods that can be used to test the hypothesis that collections of rare variants are associated with a phenotype. We also consider 9 high-dimensional regression and data mining procedures that can be used to simultaneously test the association of all individual variants, rare and common, as well as collapsed sets of variants. We did not consider covariates in these analyses. Space limitations preclude an in-depth discussion of each method so we provide references and only the main intuitions behind each method.

3.1. *Single locus and general collapsed variant test-based methods*

The following very brief descriptions of the methods we considered. Many of the papers describing these methods include discussions of possible extensions or alternative formulations of each method. We chose what we believe is the strategy that best represents the approaches described in those papers.

Single-locus tests (SL). We considered the use of Fisher's exact test to assess the association between each SNV and morbid obesity case/control status. We pursued single locus tests as a contrast for the multilocus-based collapsed variant tests since the power to detect an association involving a rare SNV is low.

Li and Leal Collapsing Method (LL). Li and Leal [6] proposed a collapsing method for testing for association with multiple rare variants. Briefly, the method collapses the genotype information across multiple (rare) variants into a single variable for each individual. This new variable can then be tested for association with a phenotype using a chi-square test or the Fisher exact test. Given a collection of variants (grouped together based on function or position in a genomic region), we considered the subset of variants with a low minor allele frequency (MAF <0.02). Additionally, variants with virtually no difference in allele frequency between the cases and controls (Fisher test p-value >=0.6) were also removed. Using the remaining variants, a binary variable was defined for each individual as 1 if the individual had the rare allele for any of the variants and 0 otherwise. Fisher's exact test was used to compute the significance of the difference in allele frequency of this binary variable between the cases and controls. The p-value of the statistic was computed by permuting the case-control status of the individuals and determining the fraction of permutations for which the statistic was lower than or equal to the observed statistic.

Madsen and Browning Method (MB). We implemented the groupwise association test described by Madsen and Browning [12]. Given a group of variants, this method tests for the presence of an excess of rare SNVs in the cases as compared to the controls. Each SNV is given a weight based on its minor allele frequency in the controls. A score is calculated for each individual using the individual genotypes and the weights of each variant. The sum of ranks of scores of the cases is used as the statistic (similar to Wilcoxon rank test). We computed the p-value for each statistic using a maximum of 1000 permutations. The test was performed using the ‘general disease’ model described by Madsen and Browning [12]. This model only allows for the analysis of rare variants and does not accommodate the effects of common variants.

Subset Selection Method (SS). Recently, Bhatia et al. [13] have proposed an extension of the Collapsing method of Li and Leal. Instead of collapsing across all rare variants in a set, the method searches for a subset of variants which maximally discriminate between the cases and controls. The method described by Bhatia et al. [13] uses a greedy algorithm to identify a subset of variants for which their collective occurrence or union has a large difference in frequency between the case and control individuals. This model only allows for the analysis of rare variations (MAF < 0.02) and does not accommodate the effects of common variations. Fisher’s exact test was used to assess the significance of sets of variants at any point in the search for the optimal set.

Distance-based diversity (Dis). Distances between the diploid sequences of all pairs of individuals in the study were calculated as one minus the identity-by-state similarity across the variant loci in a set. The dispersion of (i.e., variation among) the sequences within and between case and control groups was then compared using the ‘betadisper’ function of the ‘vegan’ package (version 1.17-0) in the R computing environment [14]. This function essentially implements Anderson's [15] PERMDISP2 procedure for the analysis of multivariate homogeneity of group dispersions [15]. Tests of the hypothesis that there is greater diversity among the cases or controls was assessed empirically via a permutation test implemented in the function ‘permutest’ in the PERMDISP2 package.

Omnibus haplotype frequency test (PHap). We considered the omnibus haplotype test strategy outlined by Fallin et al. [16] and Zhao et al. [17] and implemented in PLINK [18] for sets of variants in contiguous regions. This approach essentially tests the hypothesis that haplotype frequency profiles are equal between cases and controls, where the haplotypes harbor the variants of interest.

Power-based diversity statistic Gst (Div). We tested the hypothesis that for any set of variants the cases and controls would differ in terms of the diversity they exhibited across those variants. To conduct an appropriate test of this hypothesis, we implemented the procedure for assessing population differentiation based on the measure Gst described in equation 8 of Jost [19].

Sequence similarity statistic leveraging MDMR (Sim). We considered the use of the multivariate distance matrix regression (MDMR) and Generalized Analysis of Molecular Variance (GAMOVA) approaches discussed by Wessel et al. [20] and Nievergelt et al. [21] to test the hypothesis that the multilocus genotype profiles encompassing a set of variant loci exhibited by the cases were more similar amongst themselves than with the controls. Distances between pairs of sequences were calculated by subtracting the average value of identity-by-state similarity across loci in each window from one. The approach was implemented by O. Libiger and M. Zapala in Python (script available at http://polymorphism.scripps.edu/~cabney/). Permutation tests were used to assess statistical significance of any differences in similarity.

Ridge regression (Ridge). We used ridge regression to test the hypothesis that individual variants and collapsed sets of variants (made into a dummy variable, as described in section 1.1 above) were associated with obesity level. We used the approach outlined by Malo et al. [22] for this analysis. The method of Hoerl, Kennard, and Baldwin [23] was used to estimate the ridge parameter.

Logic regression (Logic). We also considered logic regression to identify combinations of variants that were associated with obesity. We used the implementation of logic regression that is available in the R computing environment package 'LogReg' [24]. We fit two logic trees and performed a null-model permutation test to assess significance of the association between identified sets of variants and case/control status.

Set based analysis (PSet). We considered variant set-based tests similar in orientation to Fisher's combined p-values methodology [25]. We use the method implemented in the PLINK software package for this analysis [18]. PLINK default parameters were used throughout the analysis. Statistical significance was assessed via a permutation test.

Table 4.1. P-values for tests of the association of multiple variants within five functional genomic regions in the FAAH and MGLL genes.

			FAAH		
	NS	**H3K27**	**TFBS**	**FOX2**	**Amidase**
# of variants	5	29	4	14	5
Dispersion (Dis)	0.59	0.05	0.77	0.99	0.61
Diversity (Div)	0.43	0.42	0.81	0.33	0.46
MDMR Similarity (Sim)	0.19	0.21	0.05	0.14	0.41
Li & Leal (LL)	0.60	0.03	0.60	1.00	0.50
Subset Selection (SS)	1.00	0.01	0.60	0.75	0.60
Madsen & Browning (MB)	1.00	0.01	0.33	1.00	0.75
Logic Regression (LR)	0.23	0.18	0.39	0.22	0.48
Ridge Regresssion (RR)	0.35	0.09	0.06	0.33	0.54
PLINK Haplotype (Phap)	NA	0.92	NA	0.34	0.61
PLINK Set Analysis (Pset)	1.00	1.00	0.02	1.00	1.00
			MGLL		
	NS	**H3K27**	**TFBS**	**FOX2**	**Amidase**
# of variants	9	100	11	3	0
Dispersion (Dis)	0.28	0.99	0.02	0.72	NA
Diversity (Div)	0.77	0.65	0.73	0.64	NA
MDMR Similarity (Sim)	0.81	0.07	0.67	0.29	NA
Li & Leal (LL)	1.00	1.00	1.00	0.75	NA
Subset Selection (SS)	0.60	0.43	1.00	1.00	NA
Madsen & Browning (MB)	0.75	0.30	0.02	0.20	NA
Logic Regression (LR)	0.35	0.67	0.02	0.49	NA
Ridge Regresssion (RR)	0.71	0.50	0.01	0.61	NA
PLINK Haplotype (Phap)	NA	0.81	0.07	NA	NA
PLINK Set Analysis (Pset)	1.00	0.43	0.05	1.00	NA

3.2. *High-dimensional regression methods*

As noted, we also considered the analysis of the data using high-dimensional regression and data mining procedures. These procedures could essentially consider all the variants, both in isolation or in collapsed sets, as predictors of the phenotype and were not used in moving window analyses.

Lasso (L). We considered the use of Lasso-based regression [26] using 'bridge' regression with the penalty parameter set to 1.0 and all other parameters set to their default value [27], as implemented in the 'gpsbridge' function of the R/GPS interface developed by Jerome Friedman for the R computing environment [14]. 10-fold cross validation was performed to select the best model.

Generalized path seeking regression (GPS). We employed 'bridge' regression with all parameters set to their default value [27]), as implemented in the 'gpsbridge' function of the R/GPS interface developed by Jerome Friedman for the R computing environment [14]. 10-fold cross validation was performed to select the optimal model and penalty value.

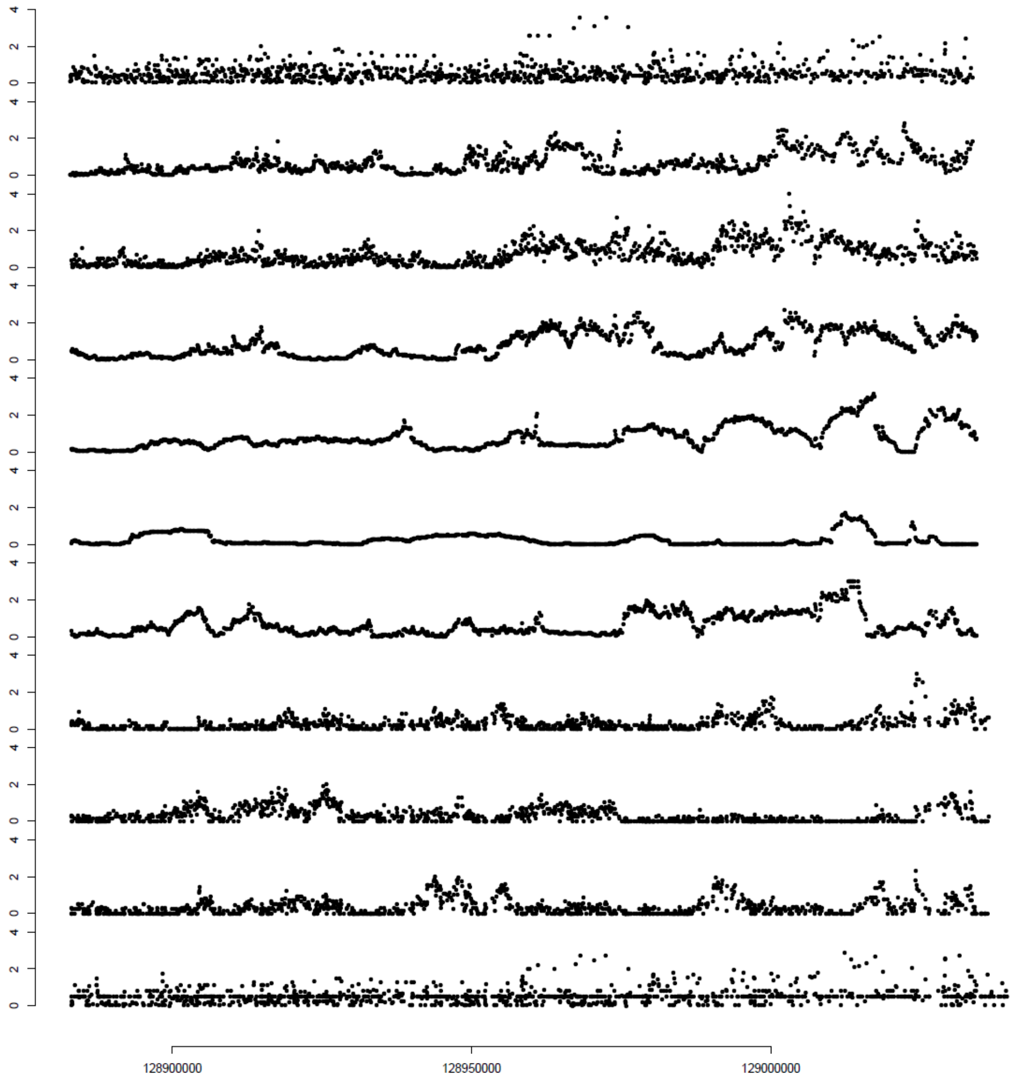


Figure 4.2. Moving window analysis of the MGLL gene using 11 different methods. Note that the y axis provides the –log(p-value) for the association for all variants in the 5 kb window whose midpoint is given on the x axis.

Stepwise Regression (SR). We performed stepwise linear model selection via the Akaike Information Criterion (AIC) for choosing associated variants and collapsed variant sets using the function ‘stepAIC’ from the package ‘MASS’ developed for the R computing environment [14].

Classification and regression trees (SPM-CART). We considered the CART method originally described by Breiman et al. [28] and implemented in the Salford systems data mining software suite (http://salford-systems.com/) to identify predictors of obesity.

Multiple adaptive regression trees (SPM-TreeNet). We also used the TreeNet procedure originaly described by Friedman et al. [29] and implemented in the Salford systems data mining software suite (http://salford-systems.com/).

Multivariate adaptive regression splines (SPM-MARS). We implemented the MARS procedure originally developed by Friedman [30] and implemented in the Salford systems data mining software suite (http://salford-systems.com/).

Random Forests (SPM-RF). We explored the use of the Random Forests procedure introduced by Breiman [31] and implemented in the Salford systems data mining software suite (http://salford-systems.com/).

Conjunctive rule learner (Weka CRL). We considered the conjunctive rule learner algorithm as described by Witten and Frank [32] and implemented in Weka [33] with no ranking.

Representative tree (Weka REPTree). We used the representative tree algorithm as described by Witten and Frank [32] and implemented in Weka [33].

4. Results

4.1. *Collapsed variants based on functional annotations*

We first considered the significance of the difference of variants within the five functional elements derived from annotations for the FAAH and MGLL genes discussed in section 2.2. Table 4.1 provides the p-values associated with 10 multilocus data analysis methods described in section 3.1 (we did not consider single locus analyses here). From Table 4.1 it can be seen that, with the exception of an analysis of collapsed variants within all transcription factor binding sites (the ‘TFBS’ column in Table 4.1) for the MGLL gene, there is not consistent evidence for association among the different methods.

4.2. *Moving window analysis*

We considered the application of the 11 different analysis methods to a moving window analysis of the MGLL and FAAH genes. The analysis explored adjacent windows of size 5000 bases for the both the MGLL and FAAH genes. The –log(p-value) computed for each test is plotted on the y-axis of Figure 4.2 against the midpoint of each window. The different panels (i.e., contour plots) reflect different analysis methods, which are, from bottom to top: standard single locus analysis using Fisher’s exact test (SL); Li and Leal’s [6] method (LL); the Madsen and Browning [12] weighted average statistic (MB); the optimal subset selection method [13]; SS); the sequence distance-based diversity statistic based on the method of Anderson [15] (Dis); the sequence diversity statistic based on the power statistic of Jost [19] (Div); the sequence similarity based statistic discussed by Wessel et al. [20] and Nievergelt et al. [21] (Sim); the ridge regression statistic [22]) (Ridge); the Logic Regression [24]) statistic (Logic); the omnibus haplotype frequency test implemented in the PLINK software package [16-18] (Phap); and the set based analysis method implemented in the PLINK software package [18] (Pset). As noted in section 2.2, a –log(p-value) of 4.06 provide some correction for an overall multiple comparisons type I error rate of 0.05. It does not appear that any of the windows produces a -log(p-value) that would be significant after multiple comparisons corrections. In addition, many of the contour plots do not appear to track together, suggesting that the various data analysis methods do not produce correlated test statistics or evidence for association. Although there is some suggestion of consistency of a signal in the 'rightmost' region of the MGLL gene, its significance is debatable. Similar conclusions were drawn from the analysis of the FAAH gene (data not shown).

4.3. *Correlations between statistics*

We assessed the correlations between the test statistics obtained over the moving window analyses of the two genes. We did not include single locus analyses or the set (Pset) and haplotype analysis (Phap) methods implemented in PLINK as part of this analysis. This provides some indication as to whether or not the different statistical methods are capturing the same signals. Table 4.3 provides the Spearman non-parametric correlation coefficients between the test statistics computed over the windows.

Table 4.3. Spearman correlations between test statistics from the moving window analyses.

Method	Dis	Div	Sim	Ridge	Logic	LL	MB	SS
Dis		0.13	0.40	0.29	0.27	-0.01	-0.10	0.03
Div	0.13		-0.11	-0.09	-0.22	-0.01	-0.20	-0.11
Sim	0.40	-0.11		0.34	0.37	-0.06	-0.03	0.04
Ridge	0.29	-0.09	0.34		0.69	-0.08	-0.17	0.16
Logic	0.27	-0.22	0.37	0.69		-0.10	-0.23	0.00
LL	-0.01	-0.01	-0.06	-0.08	-0.10		0.25	0.49
MB	-0.10	-0.20	-0.03	-0.17	-0.23	0.25		0.21
SS	0.03	-0.11	0.04	0.16	0.00	0.49	0.21	

The shaded cells within Table 4.3 reflect significant correlations (p<0.05). It should be recognized that the majority of test statistics computed in the window-based analyses are not themselves statistically significant. Therefore, the value of the test statistics that went into the calculation of the correlations may reflect noise which clearly will affect the correlation strength between the test statistics. Despite this, some of the test statistics do exhibit correlations and therefore may be essentially capturing the same types of collective effects. For example, Ridge and Logic regression are highly correlated, as are the subset selection (SS) and Li and Leal's [6]; (LL) method. Many methods are not correlated, suggesting that they may either suffer from flaws, have low power, or are more powerful to detect different types of effects. Obviously, simulation studies could be used to sort this out.

4.4. *High-dimensional regression analysis*

We also considered the use of the nine high-dimensional regression and data mining procedures listed in section 3.2, as well as the ridge regression procedure discussed in section 3.1, to simultaneously evaluate the association of each SNV, rare and common, in addition to the 10 collapsed sets of variants within each of the two genes described in sections 1.2 and 4.1 to obesity. The various procedures tested are designed to identify the minimal set of factors that are predictive of a dependent variable and hence may have an ability to capture or identify variants causally associated with obesity. Table 4.4 lists the five most significant factors identified from the 10 different procedures in addition to providing the adjusted R-squared and the root mean squared error characterizing the fit of the model that includes those 5 factors. Note that individual SNVs are denoted by a number (e.g., 166) and the gene within which they reside (MGLL or FAAH) whereas collapsed sets of variants are denoted by their labels as defined in section 1.2. From Table 4.4 it can be seen that although some factors appear in the list of five factors for different methods (e.g., individual SNV 166 appears on the list for ridge regression (RR), the Lasso (L), and the GPS method, most of the factors identified for any method are unique to that method or just a few of the methods. This suggests that the different methods are likely to disagree about which factors are the most strongly associated with a phenotype. This may be a function of the purpose and design of these methods, which is for making reliable predictions and not necessarily detecting the strongest associations among a large set of potential predictors.

Table 4.4. Top 5 chosen genomic predictors of obesity for different regression analysis methods.

RR	L	GPS	SR	SPM-CART
166 (MGLL)	166 (MGLL)	166 (MGLL)	124 (FAAH)	1036 (MGLL)
677 (MGLL)	677 (MGLL)	677 (MGLL)	8 (FAAH)	1009 (MGLL)
581 (MGLL)	76 (MGLL)	76 (MGLL)	136 (FAAH)	H3K27 (MGLL)
76 (MGLL)	581 (MGLL)	428 (MGLL)	223 (FAAH)	1136 (MGLL)
90 (FAAH)	90 (FAAH)	90 (FAAH)	200 (FAAH)	H3K27 (FAAH)
adj. R2: 0.008	**adj. R2: 0.008**	**adj. R2: 0.011**	**adj. R2: <0**	**adj. R2: 0.066**
RSE: 10.56	**RSE: 10.56**	**RSE: 10.54**	**RSE: 10.62**	**RSE: 10.25**
SPM-TreeNet	**SPM-MARS**	**SPM-RF**	**Weka CRL**	**Weka REPT**
H3K27 (MGLL)	1036 (MGLL)	H3K27 (MGLL)	1058 (MGLL)	1036 (MGLL)
H3K27 (FAAH)	1009 (MGLL)	1036 (MGLL)	H3K27 (MGLL)	
1036 (MGLL)	654 (MGLL)	634 (MGLL)	56 (FAAH)	
1136 (MGLL)		H3K27 (FAAH)	210 (MGLL)	
1009 (MGLL)		632 (MGLL)	173 (FAAH)	
adj. R2: 0.066	**adj. R2: 0.076**	**adj. R2: 0.038**	**adj. R2: 0.033**	**adj. R2: 0.025**
RSE: 10.25	**RSE: 10.19**	**RSE: 10.4**	**RSE: 10.43**	**RSE: 10.47**

5. Conclusions and Future Directions

Studies investigating the role of rare variants in phenotypic expression and disease susceptibility will be pursued routinely in the not-so-distant future as sequencing technologies improve in efficiency. The ability to exploit these technologies will depend critically on an ability to assemble and organize sequence data as well as an ability to draw reliable inferences concerning the statistical (and biological) significance of differences in combinations of sequence variants between individuals with and without a particular phenotype. We have considered a number of different approaches for relating collections of rare sequence variants to a phenotype. We compared these methods on actual sequence data obtained from two genes in a study of morbidly obese and control subjects. Some of these methods (e.g., Logic, MDMR, Dis) are computationally intensive, which may complicate their utility in very large studies. Although we did not find overwhelming evidence for an association with obesity, our studies suggest that different analysis methods, not surprisingly, do not necessarily agree on the strength of associations.

This raises important questions as to why this is so and whether or not some statistical methods may be more powerful for detecting certain types of association over other approaches. In addition, if it is the case that one or another of the proposed methods is better at picking up a certain type of association signal (e.g., most methods are likely to be better for detecting multiple independent acting variants whereas a few, such as similarity based methods [20], may be better at detecting synergistically-acting variants) then a researcher might consider analyzing their data with different analysis methods and possibly different window sizes. This in turn raises questions about false positive rates due to the use of multiple analysis methods and the pursuit of multiple comparisons. In addition, the robustness of the methods to outliers, their level accuracy, ultimate power in various settings, and their ability to accommodate covariates all need to be explored. Many of these questions can be addressed by exploring both the theoretical derivation of different methods as well as their behavior in contrived, simulated data settings [34]. Such activity will be crucial if progress is to be made in understanding the contribution of rare variants to the genetic basis of complex phenotypes.

Acknowledgments

This work was supported in part by the following research grants: U19 AG023122-05, R01 MH078151-03, N01 MH22005, U01 DA024417-01, P50 MH081755-01, R01 AG030474-02, N01 MH022005, R01 HL089655-02, R01 MH080134-03, U54 CA143906-01, UL1 RR025774-03], the Price Foundation and Scripps Genomic Medicine. Ondrej Libiger is also supported by a grant from Charles University (GAUK number 134609). The Obesity data were generated as a result of a collaboration between the Scripps Translational Science Institute and Sanofi-Aventis.

References

1. T. A. Manolio, L. D. Brooks, F. S. Collins, *J Clin Invest* **118**, 1590 (2008).
2. T. A. Manolio *et al.*, *Nature* **461**, 747 (2009).
3. W. Bodmer, C. Bonilla, *Nat Genet* **40**, 695 (2008).

4. N. J. Schork, S. S. Murray, K. A. Frazer, E. J. Topol, *Curr Opin Genet Dev* **19**, 212 (2009).
5. S. Morgenthaler, W. G. Thilly, *Mutat Res* **615**, 28 (2007).
6. B. Li, S. M. Leal, *Am J Hum Genet* **83**, 311 (2008).
7. A. P. Morris, E. Zeggini, *Genet Epidemiol* **34**, 188 (2010).
8. C. J. Hoggart, J. C. Whittaker, M. De Iorio, D. J. Balding, *Plos Genetics* **4**, (2008).
9. A. L. Price *et al.*, *Nat Genet* **38**, 904 (2006).
10. J. D. Storey, in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. (2010).
11. D. R. Nyholt, *Am J Hum Genet* **74**, 765 (2004).
12. B. E. Madsen, S. R. Browning, *PLoS Genet* **5**, e1000384 (2009).
13. G. Bhatia *et al.*, *PLoS Computational Biology*. **in press**, (2010).
14. W. N. Venables, B. D. Ripley, *Modern applied statistics with S*. Statistics and computing (Springer, New York, ed. 4th, 2002), pp. xi, 495 p.
15. M. J. Anderson, *Biometrics* **62**, 245 (2006).
16. D. Fallin *et al.*, *Genome Res* **11**, 143 (2001).
17. J. H. Zhao, D. Curtis, P. C. Sham, *Hum Hered* **50**, 133 (2000).
18. S. Purcell *et al.*, *Am J Hum Genet* **81**, 559 (2007).
19. L. Jost, *Mol Ecol* **17**, 4015 (2008).
20. J. Wessel, N. J. Schork, *Am J Hum Genet* **79**, 792 (2006).
21. C. M. Nievergelt, O. Libiger, N. J. Schork, *PLoS Genet* **3**, e51 (2007).
22. N. Malo, O. Libiger, N. J. Schork, *American Journal of Human Genetics* **82**, 375 (2008).
23. E. Hoerl, R. W. Kennard, K. F. Baldwin, *Communications in Statistic - Simulation and Computation* **4**, 105 (1975).
24. C. Kooperberg, I. Ruczinski, M. L. LeBlanc, L. Hsu, *Genet Epidemiol* **21 Suppl 1**, S626 (2001).
25. J. Hoh, J. Ott, *Nat Rev Genet* **4**, 701 (2003).
26. R. Tibshirani, *Journal of the Royal Statistical Society Series B-Methodological* **58**, 267 (1996).
27. J. H. Friedman. Fast sparse regression and classification. Stanford University Technical Report (2008).
28. L. Breimann, J. H. Friedmann, R. A. Olshen, C. J. Stone, Wadsworth, Ed. (Pacific Grove, 1984), pp. 385.
29. J. Friedman, T. Hastie, R. Tibshirani, *Annals of Statistics* **28**, 337 (2000).
30. J. H. Friedman, *Annals of Statistics* **19**, 1 (1991).
31. L. Breiman, *Machine Learning* **45**, 5 (2001).
32. I. H. Witten, E. Frank, *Data mining : practical machine learning tools and techniques*. Morgan Kaufmann series in data management systems (Morgan Kaufman, Amsterdam ; Boston, MA, ed. 2nd, 2005), pp. xxxi, 525 p.
33. E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten, *Bioinformatics* **20**, 2479 (2004).
34. V. Bansal, O. Libiger, A. Torkamani, N.J. Schork. Statistical Analysis Strategies for Association Studies Involving Rare Variants. Nature Reviews: Genetics. **In press** (2010).

HAPLOTYPE PHASING BY MULTI-ASSEMBLY OF SHARED HAPLOTYPES: PHASE-DEPENDENT INTERACTIONS BETWEEN RARE VARIANTS*

BJARNI V. HALLDÓRSSON[1,†], DEREK AGUIAR[2,3,†] AND SORIN ISTRAIL[2,3]

[1] *School of Science and Engineering, Reykjavik University,*
Reykjavik, Iceland

[2] *Department of Computer Science, Brown University,*
Providence, RI, USA

[3] *Center for Computational Molecular Biology, Brown University,*
Providence, RI, USA

In this paper we propose algorithmic strategies, Lander-Waterman-like statistical estimates, and genome-wide software for haplotype phasing by multi-assembly of shared haplotypes. Specifically, we consider four types of results which together provide a comprehensive workflow of GWAS data sets: (1) statistics of multi-assembly of shared haplotypes (2) graph theoretic algorithms for haplotype assembly based on conflict graphs of sequencing reads (3) inference of pedigree structure through haplotype sharing via tract finding algorithms and (4) multi-assembly of shared haplotypes of cases, controls, and trios. The input for the workflows that we consider are any of the combination of: (A) genotype data (B) next generation sequencing (NGS) (C) pedigree information.

(1) We present Lander-Waterman-like statistics for NGS projects for the multi-assembly of shared haplotypes. Results are presented in Sec. 2. (2) In Sec. 3, we present algorithmic strategies for haplotype assembly using NGS, NGS + genotype data, and NGS + pedigree information. (3) This work builds on algorithms presented in Halldórsson *et al.*[1] and are part of the same library of tools co-developed for GWAS workflows. (4) Section 3.3.1 contains algorithmic strategies for multi-assembly of GWAS data. We present algorithms for assembling large data sets and for determining and using shared haplotypes to more reliably assemble and phase the data. Workflows 1-4 provide a set of rigorous algorithms which have the potential to identify phase-dependent interactions between rare variants in linkage equilibrium which are associated with cases. They build on our extensive work on haplotype phasing,[1–3] haplotype assembly,[4,5] and whole genome assembly comparison.[6]

Keywords: haplotype assembly; haplotype inference; rare variants; phasing; phase inference

1. Introduction

> *Improving data quality is crucial, because if a human genome cannot be independently assembled then the sequence data cannot be sorted into the two sets of parental chromosomes, or haplotypes. This process – haplotype phasing – will become one of the most useful tools in genomic medicine.* – J. Craig Venter, 2010[7]

A genome-wide association study (GWAS) is a leading approach to find genetic determinants associated with a particular phenotype.[8–10] GWAS proceed by identifying a set of individuals carrying a disease or trait (cases) and a set of individuals that do not (controls).

*corresponding authors are Bjarni V. Halldórsson bjarnivh@ru.is and Sorin Istrail sorin@cs.brown.edu
†Authors contributed equally to this work

The cases and controls are then genotyped for a large number of common single nucleotide polymorphism (SNP) genetic variants which are then statistically tested for association to some disease or trait. GWAS have been successful in identifying many common genetic risk variants to many diseases,[8,11,12] but many associations appear to have no known connection to biological mechanisms and thus cannot be targeted for clinical intervention. Furthermore, some of these studies reveal a paradoxically encouraging and, at the same time, disappointing theme for complex traits: a set of SNPs are found to be highly statistically significant (and are often replicated in subsequent studies) yet individually, and in aggregate, these SNPs only explain a very small proportion of genetic variance.[13]

The problem of interpreting the low explicative and predictive power of these variants has been deemed the "missing heritability" problem. Many hypotheses have been presented to explain the missing heritability.[14] Most echo caveats frequently associated with GWAS such as difficulties with defining phenotypes of cases and controls, cryptic population stratification, common variation that is often left out of GWAS (copy number variation or gaps in SNP coverage), or environmental factors.[14,15] We concern ourselves with two explanations that have received much attention recently: *phase-dependent interactions* and *rare variation.*

Knowledge of haplotypes can greatly increase the power of GWAS studies and also highlight associations that are impossible to detect without haplotype phase (e.g. loss of heterozygosity). Even more complicated phase-dependent interactions of variants in linkage equilibrium have also been suggested as possible causes of missing heritability.[14] The actual haplotypes in the typed region can be obtained only at considerably higher experimental cost or via computational haplotype phasing for which most algorithms fail to work on genome-wide data. For these reasons, GWAS have generally ignored phase-dependent interactions or associations.

Although the significance of phase-dependent interactions is yet to be determined, rare variation is now accepted as playing a significant role in many common diseases[16–18] as well as rare diseases.[19,20] SNP arrays used for GWAS are designed to tag common variants only, thus rare variant associations are ignored. However, with cost of next-generation sequencing decreasing rapidly and the sequencing of tens of thousands of individuals already underway,[21,22] GWAS are likely to develop novel approaches for association. Anticipating this data will soon be available, we have developed algorithms to simultaneously identify rare variation and determine the haplotype phase of a number of individuals using sequence reads.

The class of algorithms that use sequence reads to infer the haplotypes of a diploid organism are called *haplotype assembly* algorithms.[5,23] Early formulations focused on assembling the haplotypes from the reads of one individual. Because most bases on a read are identical regardless of the chromosome of origin, the reads can be mapped to a reference genome. After mapping, reads are translated into *haplotype fragments* containing only the polymorphic (SNP) sites. A fragment *covers* a site if the corresponding read contains the SNP. Fragments that cover more than one SNP site provide valuable phase information, that is, if two SNPs co-occur on one fragment then they exist on the same haplotype. Thus, the input to the haplotype assembly problem is an $m \times n$ SNP matrix M whose m rows correspond to fragments $f_1, ..., f_m$ and each fragment f_i covers at least 2 of the n SNP sites. Formally, we define a fragment f_i as a vector of $\{0, 1, -\}$ where 0 and 1 represent the major and minor alleles at some site and '-'

represent a lack of information either because the read does not cover the SNP site or there was a technical failure (e.g. in mapping or sequencing). Two fragments i and j *conflict* if they cover a common SNP and have different alleles at that site (Fig. 1).

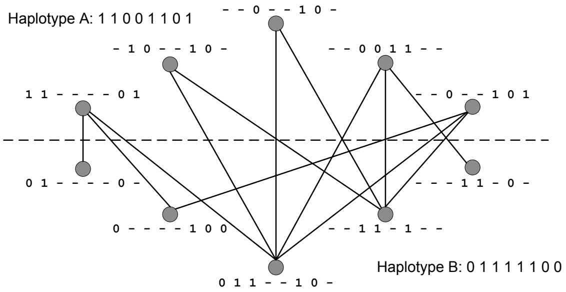


Fig. 1. Fragment conflict graph. Major and minor alleles are denoted by 0 and 1 respectively. Fragments from haplotype A (11001101) appear above the dotted line while fragments from haplotype B (01111100) appear below. The bipartition which separates the two sets of fragments is denoted by the dotted line. The haplotypes may be reconstructed by combining the shores of the bipartite graph.

Most haplotype assembly algorithms refer to an abstraction on M called the *fragment conflict graph*. The *fragment conflict graph*, $G_F(M) = (V, E)$, has nodes $f_i \in V \ \forall i$ and edges $\{f_i, f_j\} \in E$ if f_i and f_j conflict $\forall i, j$. Figure 1 demonstrates the translation from fragments to the fragment conflict graph. If the data is free of errors then, for each connected component in the fragment conflict graph, the vertices can be divided into two independent sets, that is, the graph is bipartite. Therefore, haplotype assembly of one individual can be expressed as: *Given the fragment conflict graph $G_F(M)$, find the underlying bipartite graph whose shores define the haplotypes of the individual.*

We present a novel approach to GWAS with sequence data of assembling the haplotypes of cases and controls using paired end sequencing reads and long range sharing information. Multi-assembly of GWAS sequence data has the power to enhance the discrepancies between cases and controls by phasing haplotypes using shared haplotype tracts. By assembling the cases and controls together, we can avoid missing marginal SNP variation at the level of misassembly that are associated with rare SNP variants. If the pedigree structure is known or long-range sharing information can be inferred, we can strengthen the multi-assembly by using the combined fragment coverage on the shared haplotype. First, we give a formula that relates a number of statistics/parameters with the coverage of SNPs on the haplotypes of many individuals. We present an efficient algorithm for finding the shared haplotype of two individuals in the fragment conflict graph. In addition, we present an efficient algorithmic strategy to resolve errors and assemble fragment conflict graph components that is capable of assembling genome-wide data. We employ methods that have been previously used for haplotype assembly as well as methods that have been applied to haplotype phasing.[1]

2. Multi-Assembly of Shared Haplotypes

The input for the haplotype assembly of multiple individuals problem is the same $m \times n$ matrix M in the case of one individual with an additional annotation on the fragments denoting the person of origin. In this section, we estimate the coverage needed to assemble haplotypes of multiple individuals. Consider the parameters:

G The length of the genome
S The number of SNPs in the genome
L The average length of a read
N The number of reads
c Coverage $= \frac{LN}{G}$

For these calculations we assume the distance between SNPs and sequence reads follow a Poisson distribution ($\lambda = 400$ and $\lambda = 10000$ respectively). A read is considered a "good-read" if it covers at least two SNPs. The intuition behind good-reads is that they determine the haplotype phase of two or more SNPs (they are on the same haplotype). The probability that a paired read covers at least 2 SNPs is $\geq \left(\frac{L}{400}\right)^2$. L is assumed to be ≤ 400. Let s_i be a SNP on chromosome i in position p. s_i is covered if a read starts in the interval $(p - L, p]$ on chromosome i. The expected number of reads starting in the interval $(p - L, p]$ on chromosome i is $\geq \frac{LN\left(\frac{L}{400}\right)^2}{2G} = \left(\frac{c}{2}\right)\left(\frac{L}{400}\right)^2$. The probability that no reads start in $(p - L, p]$ is $e^{-\left(\frac{c}{2}\right)\left(\frac{L}{400}\right)^2}$.

$$P\left(> 0 \; reads \; start \; in \; (p - L, p]\right) = 1 - e^{-\left(\frac{c}{2}\right)\left(\frac{L}{400}\right)^2} \tag{1}$$

Thus, the number of the SNPs covered by good-reads is approximately

$$\left(1 - e^{-\left(\frac{c}{2}\right)\left(\frac{L}{400}\right)^2}\right) 2S \tag{2}$$

Enhanced coverage due to sharing. The coverage needed greatly depends on the probability of a good-read. A high good-read probability may be obtained through targeted sequencing, mate pairs, or larger read lengths. When multiple individuals are considered, the coverage needed may be greatly reduced if the haplotype sharing is high. Sequence reads from different individuals, but on a shared haplotype, can be considered as originating from the same chromosome and assembled together. This increases the effective coverage in Equation 2. For example, three unrelated individuals have 6 unique haplotypes for an effective genome size of 6G. A trio of individuals consisting of a child, mother, and father have 4 unique haplotypes for an effective genome size of 4G. Thus, for the same amount of reads you can achieve 50% more coverage on trios than unrelated individuals.

Building on the Lander-Waterman type of statistical analysis,[24] we can estimate two important statistical parameters of haplotype assemblies that guide our algorithms: (1) What is the coverage needed so that we cover X% of SNPs on both haplotypes of a single individual with a good sequence read? (2) What is the coverage needed so that we cover X% of SNPs on both haplotypes of a trio of individuals with a good sequence read? Table 1 shows estimates of coverage needed to cover a percentage of SNPs for a single individual and trios for different parameters.

3. Algorithmic Strategies

Finding haplotype assemblies for a single individual has been considered by several researchers.[2,25–27] This can be formulated as an approximate bipartition problem, where the bipartition stems from the fact that an individual is expected to have two chromosomes of each type and the approximation stems from the fact that some of the graph edges or vertices

Table 1. Length of genome $G = 3.2 \times 10^9$. Number of SNPs $S = \frac{G}{400} = 8 \times 10^6$.

Number of Individuals	Read Length	Coverage	Mean % of SNPs Covered
1	100	1	3
1	100	2	6
1	100	4	12
1	100	10	27
1	100	20	46
1	200	1	12
1	200	2	22
1	200	4	39
1	200	10	71
1	200	20	92
3 (trio)	200	1	17
3 (trio)	200	2	31
3 (trio)	200	4	53
3 (trio)	200	10	85
3 (trio)	200	20	98

are spurious. Spurious edges occur when there is a genotyping error, some sort of error in the lab protocol or an error in the mapping of reads.

Extensions of the single individual haplotype assembly include those that employ genotype data.[28–30] In these algorithms, genotype data is used to correct errors after sequence reads are mapped. However, genotype data is prone to errors and probe common SNPs only which are not helpful regarding rare and other non-probed variation. For multiple individuals, genotype data can be used to infer evolutionary relationships between haplotypes where pedigree data is not available.[1,31,32]

3.1. *Optimization Formulations*

The minimum fragment removal formulation introduced in Lancia *et al.*[5] and minimum error correction formulation (sometimes referred to as minimum letter flip) introduced in Lippert *et al.*[4] are two optimization formulations useful for the purposes of generalizing to multiple individuals. For k individuals, Li *et al.* show that a fragment conflict graph is feasible if and only if it is $2k$-colorable. However, in the case of identical by descent haplotype sharing, there are less than $2k$ unique haplotypes. Given some haplotype sharing is likely to exist, we can rewrite the optimization problems for multiple individuals.

(1) *Minimum Fragment Removal for k Individuals (k-MFR)*: Given a SNP matrix M of fragments from k individuals, remove the minimum number of fragments (rows) such that the resulting fragments can be combined to form *at most* $2k$ haplotypes.

(2) *Minimum Error Correction for k Individuals (k-MEC)*: Given a SNP matrix M of fragments from k individuals, correct the minimum number of errors in fragments such that the resulting fragments can be combined to form *at most* $2k$ haplotypes.

A correction of an error is defined as a flip from 0 to 1 or 1 to 0. A *gapless fragment* is a fragment covering a contiguous set of SNPs. MFR, MEC, k-MFR, and k-MEC using gapless fragments are tractable and useful problems when the read length is long enough to cover multiple SNPs. However, given the smaller read length sizes of next-generation sequencing, haplotype assembly is most effective with mate paired reads. MFR, MEC, k-MFR, and k-MEC using gapped fragments have been shown to be NP-hard.[5,33]

3.2. *Problem Formulations*

Problem 3.1. *Given a set of reads from k individuals, determine the minimum number of fragments to be removed such that the remaining fragments can be assembled into $2k$ haplotypes.*

Li *et al.*[33] give an IP formulation and a parameterized algorithm which is exponential in the number of individuals. The problem formulation is somewhat simplistic as it does not assume that it is known from which individual the reads are from.

Problem 3.2. *Given a set of reads from k individuals, with each read labeled with an individual, determine the minimum number of fragments to be removed such that the remaining fragments can be assembled into $2k$ haplotypes and the individual associated with the haplotypes and fragments agree.*

The equivalent IP formulation can be seen by adding the following constraint to the Li *et al.*[33] formulation: reads labeled with an individual must be included in the assembly of that individual. For problem instances with no errors, the integer program has a very nice decomposition, since the set of constraints for each individual require it to perform a bipartite graphs. It is also likely to be quite efficient since finding bipartite graphs is easy. However, real data contains errors from miscalls and erroneous read mappings.

Problem 3.3. *Given a set of reads, each labeled with an individual, find the minimum number of haplotypes such that (1) each individual is phased with exactly two haplotypes, (2) a minimum number of fragments are removed and (3) the individual associated with the haplotypes and fragments agree.*

For general graphs, this problem is NP-hard.[5,33] We suggest a heuristic algorithm which exploits the specific signatures of sequence read errors that we can find in the data and correct. Errors in the fragment conflict graph fall into three categories.

Category 1: A fragment would otherwise conflict with another fragment from the opposite chromosome but, due to an error, is consistent with fragments on the opposite chromosome but conflicts with fragments from the chromosome of origin.

Category 2: A fragment would otherwise not be included in the fragment conflict graph but acquires an error.

Category 3: Due to an error, a fragment conflicts with fragments from both haplotypes of the individual.

Category 1 has little effect on the fragment conflict graph. We would interpret the fragment as belonging to the wrong haplotype but this does not remove the bipartiteness of the graph.

Category 2 and Category 3 can remove bipartiteness of the graph and make the general MFR and MEC problem hard. However, given a high coverage, these two cases produce regular signatures in the fragment conflict graph; namely high degree nodes that conflict with fragments of both haplotypes.

An algorithmic strategy based on the architecture of these errors was implemented in Java and works well on simulations derived from HapMap and Hudson simulated data.[34] The algorithm begins by attempting to create a breadth first search tree of the fragment conflict graph. When the algorithm encounters a level of the BFS tree that does not fit in the biparition, it computes the 3-cliques (small conflicting sub-graphs of the fragment conflict graph) in the current BFS level and subsequent levels until zero 3-cliques are found by the addition of a BFS level. It then removes the fragment belonging to the most 3-clique conflicts. As a tie breaker, the algorithm removes the node with the highest degree. Since an erroneous fragment conflicts with fragments from both chromosomes it should belong to many 3-cliques and/or have a high degree. Also, because the number of conflicting fragments in a dataset is usually small, the algorithm runs in speed comparable to BFS.

3.3. *Assembly when Haplotype Sharing is Known*

If sharing information is unknown, assembling multiple individuals can help identify Category 1 errors. If haplotype sharing information is known or can be inferred, assembling multiple individuals simultaneously provides additional information on the coverage of haplotypes. The sharing of haplotypes between individuals could be known from pedigree data[35] or inferred.[1,31,32,36]

Problem 3.4. *We are given a set of individuals and for each pair of individuals, haplotype sharing information is known or can be inferred. We are also given a set of paired end sequencing reads for each individual. Output a pair of haplotypes for each individual such that each individual sharing a haplotype do so.*

If haplotype sharing information is unknown, we begin by inferring pedigree information using the tract finding algorithm[1] or similar methods.[31,32] We then build the fragment conflict graph. *The only edges that are informative are edges between fragments from the same individual and other individuals who share a haplotype identical by descent (IBD).[35]* If a segment of a haplotype is shared identical by state (IBS) then it is likely to conflict in other places on the haplotype and can yield a feasible but erroneous assembly. In addition, if we interpret these non conflicts as IBD, then can obtain the wrong coverage estimate on the haplotype which is essential for phasing the assembly.

3.3.1. *Haplotype sharing algorithm*

When there are no errors in the reads of an individual the fragment reads will form a bipartite graph.[5] The fragments belonging to one of the two shores of the bipartite graph will form one of the haplotypes and the fragments belonging to the other shore will form the other haplotype. In the case when the bipartite graph is disconnected then each connected component may be considered separately. Given fragments from two individuals which are known to share a haplotype, we propose the following algorithm for the joint haplotype assembly of the two individuals. Let the two individuals be denoted by i and j. In our algorithm α corresponds to the shared haplotype and β corresponds to the non-shared haplotype. We note that $\alpha_i = \alpha_j$, while $\beta_i \neq \beta_j$.

Algorithm 3.1.

```
def Branch( s )
   For each edge with an endpoint in the $\alpha_s$ and other endpoint, $e$,
   in $H$, identify the connected component, $t$, of $G_i$ or $G_j$ that contains $e$.
      Label the color of $t$ that is connected to $e$ as $\beta$.
      Label the other color of $t$ as $\alpha$.
      $H \longleftarrow H - t$
      Branch( t )
   For each tree, $t \in H$ that has a color $c$ that has edges connecting to $\alpha_s$ and $\beta_s$
      Label $c$ as $\beta$
      Label the other color as $\alpha$
      $H \longleftarrow H - t$
      Branch( t )

Determine Sharing
   Construct the fragment conflict graph, $G$. Let $G_i$ and $G_j$ be
   the restriction of $G$ to $i$ and $j$.
      Color each component $G_i$ and $G_j$ with two colors.
      If no such coloring exists, the algorithms fails.
   $H \longleftarrow G$
   While $H \neq \emptyset$
      Find connected components $t, s$, s.t. $s \in G_j \cap H$ and $t \in G_i \cap H$ or $s \in G_i \cap H$
      and $t \in G_j \cap H$ and a color $\beta$ of $s$ with an edge to both colors of $t$.
         If no such tree exists, choose $s$ arbitrarily from $G_i$ or $G_j$
            Arbitrarily label the colors of $s$ as $\alpha$ and $\beta$.
      $H \longleftarrow H - s$.
      Branch( s)
```

The algorithm is motivated by the key observation that a haplotype cannot be shared if one of its fragments is connected to a color that is shared. We may observe a connection to a color that is shared either from the fact that the color is labeled as shared or it is connected to both colors.

Lemma 3.1. *The algorithm runs in $O(n+mn)$ time, where n is the number of fragments and m is the number of edges between the fragments.*

Proof. The initial step of coloring of a bipartite graph of each individual can be done in time $O(n+m)$. The edges that lie between two individuals can then be labeled with the component and color that they belong to. We then loop over each component of i in an outer loop and each component of j in an inner loop, followed by a loop over each component of j in an outer loop and each component of i in the inner loop. We determine whether there exists an edge from the component in the outer loop to both colors of the component in the inner loop. We observe that each edge will be visited at most as many times as there components in G. The number of components is upper bounded by n, for a total upper bound of nm. □

This algorithm presents an approach that may be generalized to more complex patterns of haplotype sharing.

3.4. *Phasing Components in the Fragment Conflict Graph*

Even with error-free data we aren't guaranteed to be able to assemble and phase the data. *Long runs of homozygosity form disconnected components in the fragment conflict graph.* Runs of homozygosity, which are paradoxically simple to phase, cause problems when assembling haplotypes. If the run of homozygosity is longer than the mate pair length no read can connect the two components as there wont be any conflicts in homozygous regions (Fig. 2). The more connected the graph is, the easier it is to phase because you have to eventually phase the shores of each component into two haplotypes. The number of valid haplotype phasings may therefore be large once the haplotypes of each individual have been assembled; if the haplotype assembly of a single individual consists of k disconnected bipartite components then there are 2^{k-1} unique ways to map the shores to haplotypes. Varying the mate pair read length, increasing the read length, adding coverage, or adding more individuals who may share a haplotype IBD help connect components together.

Fragments from haplotypes that are identical by descent can be considered when constructing bipartitions for both individuals. If two components need to be phased and one haplotype is shared then we'd expect the shared haplotype to have twice the coverage of the non-shared haplotype in both components, thus we phase the two shores with greater coverage from different components together. For example, Fig. 2 shows fragments from two haplotypes of two individuals one of which (10000001) is shared. The phasing of the two components is ambiguous but we know that the shared haplotype is likely to have approximately 50% more coverage. Therefore, it is more likely to phase the components such that we maximize the difference of cardinality between the phasings. For Fig. 2 the first phasing (10000001/00000000) yields $|6 - 3| = 3$ while the second phasing (10000000/00000001) yields $|5 - 4| = 1$. When phasing disconnected components where sharing is not known, the resulting phasing should try to minimize the difference of cardinality in the overall phasing.

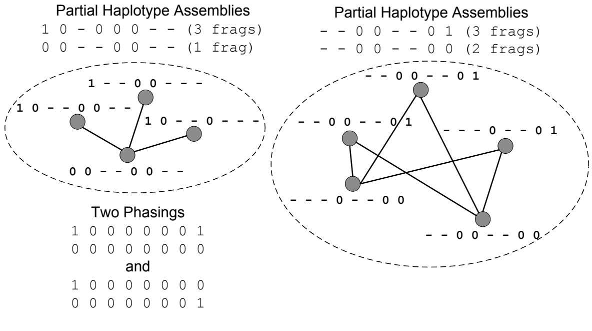


Fig. 2. Fragment conflict graph separated by a run of homozygosity. We assume the maximum distance between fragments is 2 SNPs.

4. Results on Simulated Data

We ran simulations on individuals of the CEU and JPT populations from HapMap[c]. First we sampled individuals randomly and then isolated a subset of the haplotype (30 SNPs for

[c] CEU denotes Utah residents with Northern and Western European ancestry and JPT denotes Japanese individuals from Tokyo, Japan.

visualization purposes). We placed the SNPs from the phased HapMap haplotypes a uniform distance from each other (500bp). Genome length is calculated by *number of SNPs* × *distance between SNPs*. The distance between sequence reads is calculated using a Poisson distribution and is varied under different models because most NGS technologies are capable of varying the distances between reads (e.g. Solexa or SOLiD). The average read length and coverage are also varied. Figures 3 and 4 show simulations on two unrelated individuals (one from CEU colored green, one from JPT colored red) while Fig. 5 shows simulations from two related individuals.

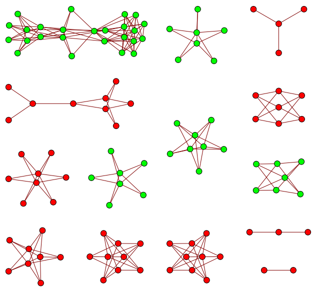


Fig. 3. Fragment conflict graph for unrelated individuals with coverage $c = 4$, read length $L = 50$, and distance between reads is Poisson with $\lambda = 2000$. Green vertices denote fragments originating from the CEU individual.

Figure 3 has many disconnected components due to the low probability of a good-read and regular distance between reads. Figure 4 shows the effect of changing the read length, coverage, and distance between reads. Read length, coverage, and variation of mate pair length correlate strongly with connectivity of the fragment conflict graph. In Fig. 5 two related individuals are shown with the same parameters used in Fig 3. It is clear the more sharing existing in the population, the easier it is to assemble and phase the data.

We also used our haplotype assembly simulator to test the accuracy and scalability of our minimum fragment removal heuristic. The first dataset we tested is the same 30 SNP segment from the HapMap CEU individual; the second dataset is a Hudson simulated chromosome of length 3434 SNPs. We decided to use the ratio of the number of erroneous fragments removed to the number of non-erroneous fragments removed as our metric. After the fragment conflict graph is generated, it may be advantageous to remove non-erroneous fragments to minimize our objective function. Nevertheless, this ratio is a good indicator of the quality of the output. For 1000 runs of the 30 SNP dataset we observed an overall ratio of 6.73; and for 100 runs of the 3434 SNP dataset we observed a ratio of 5.72. Further improvements to this type of algorithmic strategy for this problem is the subject of future work.

We've presented statistical estimates of coverage needed to cover a percentage of SNPs on a genome. These estimates could provide valuable insight when deciding sequence coverage per individual in association studies employing NGS technology. We've suggested a practical algorithmic strategy that exploits the high coverage possible with next-generation sequencing technology and the structure of errors in the fragment conflict graph. This algorithm produces promising results on the simulated fragment conflict graphs. We have presented an algorithm

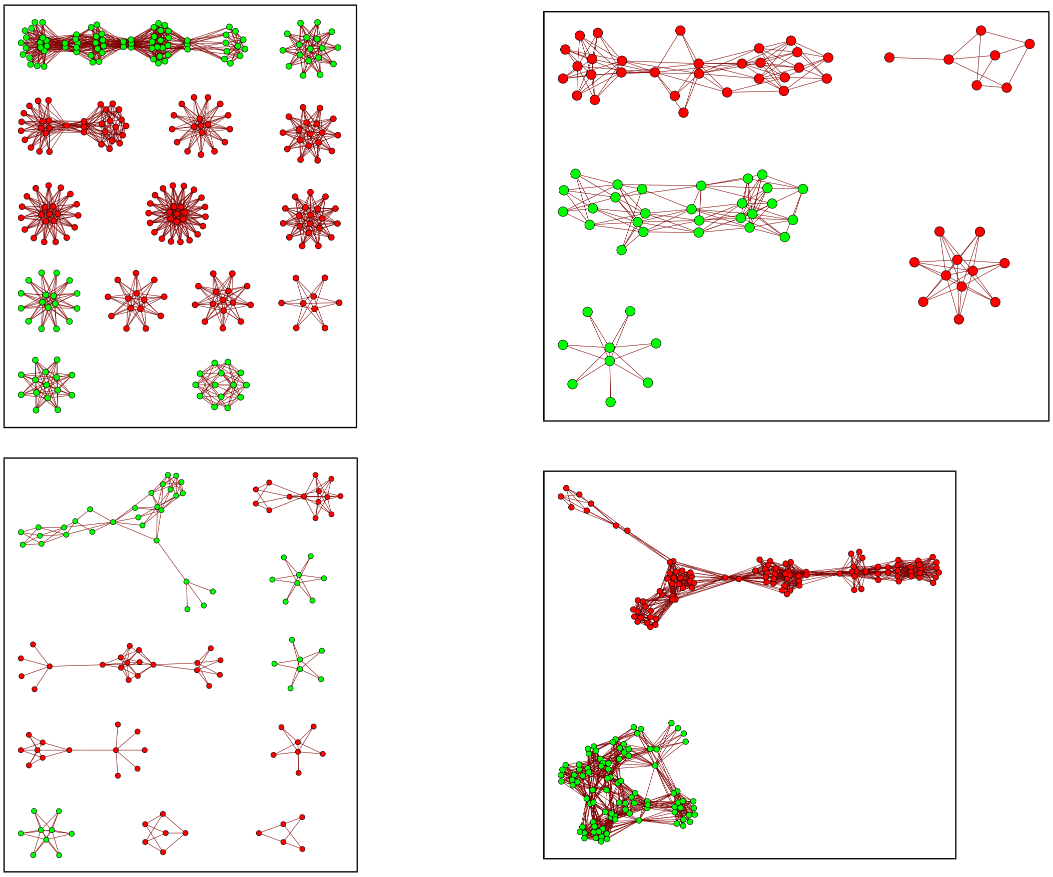


Fig. 4. Fragment conflict graph for two unrelated individuals. Green vertices denote fragments originating from the CEU individual. The baseline for each graph is: Read length $L = 50$; Coverage $c = 4$; distance between reads is Poisson with $\lambda = 2000$. From bottom left clockwise: (1) Distance between reads is Poisson with $\lambda = \{1000, 2000, 5000, 10000\}$ which is selected uniformly at random. (2) Coverage is changed to $c = 10$. (3) Read length is changed to $L = 1000$. (4) Coverage is $c = 10$, read length is $L = 1000$, and distance between reads is varied from $\{1000, 2000, 5000, 10000\}$.

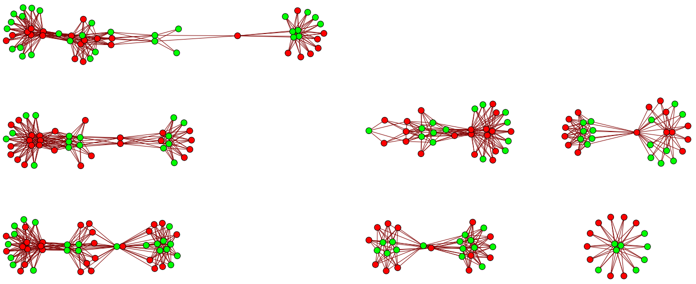


Fig. 5. Fragment conflict graph for two individuals sharing one haplotype. Green and red vertices denote fragments originating from different individuals. Read length $L = 50$. Coverage $c = 4$. The distance between reads is Poisson with $\lambda = 2000$.

for finding and exploiting haplotype sharing in the fragment conflict graph to enable the reliable phasing of disconnected components. We've also shown through simulation how various genomic and experimental parameters impact the quality of the haplotype assembly.

5. Acknowledgments

We thank the National Science Foundation for their support of this research.

References

1. B. V. Halldórsson, D. Aguiar, R. Tarpine and S. Istrail, *International Conference on Research in Computational Molecular Biology (RECOMB)* **6044**, 158 (2010).
2. B. V. Halldórsson *et al.*, *Lecture Notes in Computer Science* (2004).
3. R. Sharan, B. V. Halldórsson and S. Istrail, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3**, 303 (2006).
4. R. Lippert, R. Schwartz, G. Lancia and S. Istrail, *Brief Bioinform* **3**, 23 (March 2002).
5. G. Lancia, V. Bafna, S. Istrail, R. Lippert and R. Schwartz, *Proceedings of the 3rd European Symposium on Algorithms, (EAS01) Springer Lecture Notes in Computer Science* **2161**, 182 (2001).
6. S. Istrail *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 1916 (2004).
7. J. C. Venter, *Nature* **464**, 676 (April 2010).
8. R. Sladek *et al.*, *Nature* **445**, 881 (February 2007).
9. *Nature* **447**, 661 (June 2007).
10. R. J. Klein *et al.*, *Science* **308**, 385 (April 2005).
11. D. A. Hafler *et al.*, *N Engl J Med* **357**, 851 (2007).
12. N. J. Samani *et al.*, *N Engl J Med* **357**, 443 (August 2007).
13. M. N. Weedon *et al.*, *Nature Genetics* **40**, 575 (April 2008).
14. E. E. Eichler *et al.*, *Nature reviews. Genetics* **11**, 446 (June 2010).
15. J. McClellan and M.-C. King, *Cell* **141**, 210 (April 2010).
16. T. Walsh and M.-C. King, *Cancer Cell* **11**, 103 (February 2007).
17. J. C. Cohen *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **103**, 1810 (February 2006).
18. A. A. Dror and K. B. Avraham, *Annual Review of Genetics* **43**, 411 (2009).
19. H. Stefansson, D. Rujescu, S. Cichon, O. P. H. Pietilainen *et al.*, *Nature* **455**, 232 (Sep 2008).
20. J. Sebat *et al.*, *Science* **316**, 445 (April 2007).
21. N. Siva, *Nature biotechnology* **26**, p. 256 (March 2008).
22. Wellcome trust launches effort to sequence 10,000 human genomes `http://www.genomeweb.com//node/943774?hq_e=el&hq_m=751896&hq_l=1&hq_v=672d790b6d`.
23. R. Schwartz, *Communications in Information and Systems* **10**, 23 (2010).
24. E. S. Lander and M. S. Waterman, *Genomics* **2**, 231 (April 1988).
25. V. Bansal, A. L. Halpern, N. Axelrod and V. Bafna, *Genome Research* **18**, 1336 (August 2008).
26. D. He *et al.*, *Bioinformatics* **26**, i183 (June 2010).
27. V. Bansal and V. Bafna, *Bioinformatics* **24**, i153 (August 2008).
28. Y. Wang *et al.*, *Computational Biology and Chemistry* **31**, 288 (August 2007).
29. J. Wang, M. Xie and J. Chen, *Algorithmica* **56**, 283 (March 2010).
30. S. Kang, I. Jeong, M. Choi and H. Lim, *Lecture Notes in Computer Science* **5059/2008**, 45 (2008).
31. S. Purcell *et al.*, *American journal of human genetics* **81**, 559 (September 2007).
32. S. R. Browning and B. L. Browning, *American journal of human genetics* **86**, 526 (April 2010).
33. Z. Li, L. Wu, Y. Zhao and X. Zhang, *Acta Mathematicae Applicatae Sinica* **22**, 405 (2006).
34. R. R. Hudson, *Bioinformatics* **18**, 337 (2002).
35. A. Kong, G. Masson, M. L. Frigge *et al.*, *Nat Genet* **40**, 1068 (Sep 2008).
36. M. J. Minichiello and R. Durbin, *Am J Hum Genet.* **79**, 910 (Nov 2006).

AN EVALUATION OF POWER TO DETECT LOW-FREQUENCY VARIANT ASSOCIATIONS USING ALLELE-MATCHING TESTS THAT ACCOUNT FOR UNCERTAINTY

E. ZEGGINI* and J.L. ASIMIT
Wellcome Trust Sanger Institute, Hinxton, CB10 1HH, UK
**E-mail: Eleftheria@sanger.ac.uk*

There is growing interest in the role of rare variants in multifactorial disease etiology, and increasing evidence that rare variants are associated with complex traits. Single SNP tests are underpowered in rare variant association analyses, so locus-based tests must be used. Quality scores at both the SNP and genotype level are available for sequencing data and they are rarely accounted for. A locus-based method that has high power in the presence of rare variants is extended to incorporate such quality scores as weights, and its power is compared with the original method via a simulation study. Preliminary results suggest that taking uncertainty into account does not improve the power.

Keywords: Allele-Matching; Rare variants;Locus-based method; Quality scores; Sequencing

1. Introduction

There is an increasing interest in the role of rare variants in multifactorial disease etiology, while the evidence that rare variants are associated with complex traits is steadily expanding. Although any individual rare variant exists in low frequencies, the frequency with which any rare variant is present makes them collectively common. Under the multiple rare variant hypothesis (MRV), the effects of multiple rare variants with moderate to high penetrance combine to increase the risk of most common inherited diseases [1]. At the other extreme is the common disease common variant (CDCV) hypothesis, which states that most common complex diseases are due to a few common variants with moderately small effects [2]. The most likely scenario is that a combination of both common and rare variants contribute to disease risk.

In most genome-wide association (GWA) studies only variants with minor allele frequency (MAF) greater than 1-5% are followed up, and the focus tends to be on identifying common disease variants that are associated with complex diseases. However, this approach is limited since only 5-10% of the heritable component of disease is explained by the many genetic variations identified as having strong evidence of disease association in GWA studies. This suggest that a fruitful direction is to search for associations with multiple rare variants [3].

By design, SNP genotyping panels often focus on common SNPs, so that they only contain a relatively small number of rare variants. This leads to a common issue in rare variant analyses, in that on most platforms there is an insufficient number of rare variants (Table 1).

There appears to be a clear difference in the effects of rare variants in comparison to SNPs of higher frequency, with rare variants having stronger effects. According to the odds ratios (OR) for common and rare variants identified in published studies, most common-disease associated variants have ORs between 1.1 and 1.4 with only a few above 2, while the majority of the identified rare variants to date have an OR greater than 2 and a mean of 3.74 [1]. In

Table 1: Approximate low frequency/ rare variant GWAS platform content.

Platform	Affymetrix 500k	Affymetrix 6.0	Illumina 370k	Illumina 550k	Illumina 610k	Illumina 1.2M
MAF$<$ 0.05	55k	106k	9k	32k	35k	62k
MAF$<$ 0.01	17k	35k	1k	7k	8k	22k

addition, causality may more easily be fine-tuned by identifying rare variants. For most GWA-identified loci, there is difficulty in assigning causality since high LD complicates the use of association mapping to precisely determine which variant is functionally relevant. There are even more complications when elucidating the effects of SNPs that map to genomic regions with no clear role. The problem may be simplified by searching for disease-associated rare variants in known functional genomic regions, such as genes. In addition, it might be easier to at least infer causality at a locus that contains both common and rare disease-associated variants.

In the analysis of the association of rare variants and disease, there is a loss of power due to genotype misspecification. Quality scores are available for genotype and sequence-derived data, but in rare variant analyses, the information is not usually put to use. In addition, the 1000 Genomes reference set contains variants with MAF as low as .01, which makes the imputation of rare variants now possible. A probability distribution for the genotype at each variant may be estimated using the imputation method of choice. We propose methods for rare variant analyses that take advantage of the extra information contained in quality scores derived from sequencing and probability distributions resulting from imputation.

In section 2 we introduce an Allele Matching Empirical Locus-specific Integrated Association test (AMELIA), which is a nonparametric and robust test that accounts for genotype uncertainty. It is an extension of a Kernel-Based Association Test (KBAT) [4], which has been demonstrated to have high power in the presence of rare variants. In section 3 the powers of AMELIA and KBAT are briefly compared in a short simulation study, while a concluding discussion is provided in section 4.

2. Allele-Matching Tests

Before providing the details of AMELIA, we first discuss the original method, KBAT. The kernel-based association test (KBAT) [4] tests for a joint association of multiple SNPs (correlated or independent) with a categorical phenotype, without any assumptions on the directions of individual SNP effects. In simulation studies done by the authors, KBAT was found to generally have more power than other multi-marker approaches (Zglobal[5] and MDMR[6]), especially in the presence of rare causal SNPs. First, similarity scores $y_{l(ij)}$ between individuals i and j in group l (e.g. 1=cases, 2=controls) are determined by using a kernel, such as the Allele Match (AM) kernel, which is the count of common alleles between the genotypes of two individuals. Let g_i be the genotype score at a specific SNP, which is conveniently defined as the number of reference alleles at the SNP, since knowledge of the risk allele is irrelevant. At a given SNP, for individuals $i \neq j$ in group l with respective genotypes $g_l(i)$ and $g_l(j)$, the similarity score is defined by

$$y_{l(ij)} = \begin{cases} 4, \text{ if } g_{l(i)} = g_{l(j)} \\ 2, \text{ if } g_{l(i)} = 1, g_{l(j))} \in \{0, 2\} \text{ or } g_{l(j)} = 1, g_{l(i)} \in \{0, 2\} \\ 0, \text{ otherwise} \end{cases}, \quad (1)$$

By defining the kernel in this way, there is no need to have knowledge of the risk allele at each SNP. Similarity scores that depend on knowledge of the risk allele are also explored in [4]. This is general to any number of $L \geq 2$ groups, where group l consists of n_l individuals.

The similarity scores $y_{l(ij)}$ between individuals i and j in group l are modelled using a one-way ANOVA model at each SNP:

$$y_{l(ij)} = \mu + \alpha_l + \varepsilon_{l(ij)}, \quad i < j = 1, \ldots, n_l; \quad l = 1, 2,$$

where μ is the general effect for pairs of individuals, α_l is the group specific treatment effect, and to test for disease association the null hypothesis is $H_0 : \alpha_1 = \alpha_2$. The single SNP test statistic at marker k is the ratio of the between group sum of squares SSB_k and the within group sum of squares SSW_k, and the K-marker KBAT test statistic is

$$\frac{\sum_{k=1}^{K} SSB_k}{\sum_{k=1}^{K} SSW_k}. \quad (2)$$

Rather than summing over the K single SNP test statistics (ratios), the K-marker test statistic takes the form of (2), which was found to have a higher power when the SNPs are correlated (see [4]). Clearly the similarity scores $y_{l(ij)}$ are not independent Normal random variables, so that neither the single SNP test statistics nor the KBAT test statistic (2) may be approximated by an F-distribution. Thus, permutation is required to obtain the p-value for each locus.

Our extensions that incorporate genotype uncertainty due to quality scores at the SNP and genotype level or imputation are introduced as AMELIA. Here, we focus on the incorporation of the two levels of quality scores. Quality scores of SNPs and genotypes can be accounted for by using weights. Phred quality scores at both the SNP and genotype level are transformed into the probability of a correct call as follows, $1 - 10^{-q/10}$), where q is the quality score. This transformation is employed in order to account for the fact that the phred quality scores are not linear and to avoid down-weighting SNPs that are actually of acceptable quality. For example, quality scores of 30 and 90 both translate to probabilities near 1, and by using the phred quality scores as weights the SNP with score 30 would contribute little weight when it is not really of poor quality.

First, (transformed) genotype quality scores are incorporated into the analysis by fitting a weighted ANOVA model at each SNP k, where the weight for the pair of individuals (i,j) in group l is a function of the genotype quality scores $q^k_{l(i)}$ and $q^k_{l(j)}$, with the simplest weight function being $w^k_{l(ij)} = q^k_{l(i)} + q^k_{l(j)}$. Note that for a more suggestive notation for the quality incorporation into the analysis we use $q^k_{l(i)}$ to denote the transformed genotype quality score. In the original method, KBAT, each of the similarity scores contributes a unit weight to the SNP-level test statistic. However, with the simple weighting scheme that we consider, similarity scores for which both genotype calls have a high probability of being correct are assigned a weight above 1, while those with two poor scores are down-weighted to contribute a weight below 1. At marker k the weighted sum of squares within groups $wSSW_k$ and between groups $wSSB_k$ may be computed as follows, where for simplicity we have dropped the k superscript, and $\bar{T}_{l\cdot}$ is the weighted group mean of the similarity scores, $\bar{T}_{\cdot\cdot}$ is the weighted grand mean, and $m_l = n_l(n_l - 1)/2$ is the number of similarity scores in group l:

$$wSSW = \sum_{l=1}^{L} \sum_{i=2}^{m_l} \sum_{j<i} w_{l(ij)}(y_{l(ij)} - \bar{T}_{l\cdot})^2 \tag{3}$$

$$wSSB = \sum_{l=1}^{L} m_l(\bar{T}_{l\cdot} - \bar{T}_{\cdot\cdot}) \tag{4}$$

Components of SNP test statistic k in the sums of the K-marker test statistic can be weighted by the SNP quality score(s) of SNP k. In the case that there is a common SNP quality score Q_k across all individuals (score at a SNP is based on reads from all individuals), the weight for SNP k in the sums is simply the (transformed) single SNP quality score Q_k. If the quality scores at a SNP differ among individuals (score at a SNP based on multiple reads from single individual), then the weight may be taken as the sum of these scores at the SNP. In the latter case, the K-marker test statistic is

$$\frac{\sum_{k=1}^{K} Q_k SSB_k}{\sum_{k=1}^{K} Q_k SSW_k}. \tag{5}$$

In this form, SNPs that have a low probability of being a true variant contribute a lower weight than the others.

2.1. *Implementation*

In order to increase the speed of the permutations, as suggested in [4], the similarity scores between all possible pairs of individuals are computed, regardless of which cohort they belong to. Then, in the permutation stage, the similarity scores for the permuted case-control samples may be quickly extracted without further computation. However, for large cohorts ($N > 1000$), this causes both AMELIA and KBAT to be memory-intensive, requiring additional memory allocation to run. For example, when $N = 1000$ there are 499,500 similarity scores between all possible pairs of individuals, which requires manipulation of a 499,500 × 499,500 array. The time requirement for both methods also increases with the number of SNPs since a test statistic must be computed at each SNP for each permutation.

3. Simulation Study

A brief simulation study has been run to compare the powers of KBAT [4] and our version of AMELIA that accounts for quality scores. Genotype and quality score data are simulated based on data from the pilot study of 1000 Genomes (68 individuals). More specifically, we use the `haplosim` function of the `hapsim` [7] R package to simulate a population of haplotypes that possess the same allele frequencies and pairwise LD structure as a specified chromosomal region from the 1000 Genomes data. This approach produces realistic data that includes variants with MAFs down to .01. A cohort of N individuals is formed by randomly pairing up

$2N$ haplotypes sampled from a population of 40000 simulated haplotypes. SNP and genotype quality scores were generated by randomly sampling with replacement from the scores observed in the 1000 Genomes data. In the simulations considered there is only one causal SNP, which has a MAF close to a certain frequency, and is chosen randomly among the possible SNPs that satisfy this criterion. More complicated simulations involving multiple causal SNPs are to be explored in the near future.

Case-control status is generated by using a multiplicative model for the genotype relative risks to compute the probability of disease given the genotype at the causal SNP and its relative risk (RR) (for details see [4]). This probability is then used to generate a Bernoulli random variable that ascertains an individual as a case when its value is 1, and a control otherwise. For this reason, it is necessary to over-sample (say, $5N$) the number of individuals to ensure that the desired number of cases is attained.

In order to obtain the p-value in an efficient manner, we first obtained p-values based on 1000 permutations. If this p-value was below .02, additional permutations were run to update the p-value on the basis of 10,000 permutations. This procedure of updating the p-value continues up to a maximum of 1,000,000 permutations, if necessary.

In order to compare the two tests in a scenario similar to that of [4], rather than testing the whole region we also test regions of 11 SNPs formed from the causal SNP and 10 randomly selected SNPs among the 20 SNPs that form a neighborhood around the causal SNP (10 upstream and 10 downstream from the causal SNP) (termed the neighborhood region).

3.1. *Results*

In this brief simulation study, a 150 KB region from chromosome 1 of the 1000 Genomes data was considered, which contains 342 SNPs. This region was chosen slightly arbitrarily, but also because it has a genome-average recombination rate of approximately 1Mb/cM. All SNPs were retained, except for those with a SNP quality of 0. We assumed a single low frequency causal SNP (MAF=.02, RR=2), and 500 cases and 500 controls were simulated over 1000 replications.

Table 2: Power results (5% level of significance) for AMELIA and KBAT when there is one rare causal SNP and there are 500 cases and 500 controls.

region	AMELIA	KBAT
whole	.0871	.0953
neighborhood	.1731	.2161

When jointly testing all SNPs within a region there is a slight loss of power with the use of AMELIA in comparison to KBAT. However, both methods have a relatively low power when there are many SNPs in the region. In a comparable scenario examined in [4], where the region contains only 10 SNPs and the causal SNP has a MAF of .108 with RR=1.25 the power of KBAT was .323. In our neighborhood simulations comparing AMELIA and KBAT we obtain powers of similar magnitude (see Table 2). Thus the low powers for the entire region tests are likely due to the fact that our region of 150kb contains almost 350 SNPs, which are all jointly tested. This illustrates a caveat of this multi-marker testing approach.

In order to examine type I error, a null simulation in which we set the relative risk as 1 is also examined. However, we only consider the neighborhood region due to the extremely low power observed for the entire region. At the 5% level both methods are found to be quite conservative, with AMELIA and KBAT having respective type I errors of .00502 and .00401.

4. Discussion

In the short simulation study presented here, a decrease in power has been observed by incorporating quality scores of SNPs and genotypes as in AMELIA, with the difference largest for a small number of SNPs. The relatively low power of the two methods may be due to the fact that almost 350 SNPs are being tested jointly, of which there is only one causal SNP. This may suggest that this multi-marker approach may be best suited for smaller regions, or after some filtering to reduce the number of SNPs that are jointly tested. For example, when the focus is on low-frequency variants, the analysis may include only those with a MAF below a certain threshold, such as 0.05. It is noted that the replications that were identified only by KBAT tend to have a causal SNP with a high SNP quality score. In such situations it may be that by allowing for uncertainty that is not present, power to detect the signal is inadvertently diluted. In the simple simulations examined, the power of AMELIA appears to be lower than KBAT, and both tests are conservative with similar error rates. We are extending our methods further to achieve greater power.

References

1. Bodmer W and Bonillna C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics* **40**: 695-701.
2. Pritchard JK and Cox NJ. (2002). The allelic architecture of human disease genes: common disease common variant ... or not? *Human Molecular Genetics* **11**: 2417-2423.
3. Schork NJ, Murray SS, Frazer KA, and Topol EJ. (2009). Common vs rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development* **19**: 212-219.
4. Mukhopadhyay I, Feingold E, Weeks DE, and Thalamuthu A. (2010). Association Tests Using Kernel-Based Measures of Multi-Locus Genotype Similarity Between Individuals. *Genetic Epidemiliogy* **34**: 213-221.
5. Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, and Thibodeau SN. (2005). Nonparametric Tests of Association of Multiple Genes with Human Disease. *American Journal of Human Genetics* **76**: 780-793.
6. Wessel J and Schork NJ. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. American Journal of Human Genetics 79: 792-806.
7. Montana G. (2005). HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics* **21**: 4309-4311.

PENALIZED REGRESSION FOR GENOME-WIDE ASSOCIATION SCREENING OF SEQUENCE DATA

H. ZHOU*[1,2], D.H. ALEXANDER[3], M.E. SEHL[4],
J.S. SINSHEIMER[2,3,5], E.M. SOBEL[2], AND K. LANGE[2,3,6]

[1] *Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203*
Departments of [2] *Human Genetics,* [3] *Biomathematics,* [4] *Medicine,* [5] *Biostatistics, and* [6] *Statistics*
University of California, Los Angeles, CA 90095
* *E-mail: hua_zhou@ncsu.edu*

Whole exome and whole genome sequencing are likely to be potent tools in the study of common diseases and complex traits. Despite this promise, some very difficult issues in data management and statistical analysis must be squarely faced. The number of rare variants identified by sequencing is apt to be much larger than the number of common variants encountered in current association studies. The low frequencies of rare variants alone will make association testing difficult. This article extends the penalized regression framework for model selection in genome-wide association data to sequencing data with both common and rare variants. Previous research has shown that lasso penalties discourage irrelevant predictors from entering a model. The Euclidean penalties dealt with here group variants by gene or pathway. Pertinent biological information can be incorporated by calibrating penalties by weights. The current paper examines some of the tradeoffs in using pure lasso penalties, pure group penalties, and mixtures of the two types of penalty. All of the computational and statistical advantages of lasso penalized estimation are retained in this richer setting. The overall strategy is implemented in the free statistical genetics analysis software Mendel and illustrated on both simulated and real data.

Keywords: GWAS; penalized regression; rare variant; deep sequencing

1. Introduction

Deep resequencing is emerging as a new and potent means for mapping Mendelian disease genes.[1,2] The initial successes raise the question of whether the search for rare variants is apt to be as promising a route to mapping genes for common complex diseases and traits. In our opinion, the answer is likely to be in the affirmative, but too few studies have been completed to form a strong opinion. The recent finding of an association between copy number variation and autism is one argument in favor of the rare variant common disease hypothesis.[3] This association is not too surprising given the correlation between autism and paternal age, which is known to increase the risk of deleterious mutations. The paternal age argument applies to other psychiatric traits such as schizophrenia[4] and bipolar disorder.[5] The rare variant hypothesis is also more plausible on evolutionary grounds than the common variant hypothesis because genetic variants with a negative impact on fitness should in theory be driven to extinction. The lessons classical population genetics teaches about the balance between selection and mutation are still relevant today. Thus, there is good reason to explore the statistics of rare variation detection in anticipation of sequence based genetic studies.

Resequencing will deliver both rare and common variants. It would be counterproductive to discard the common variants because in reality there is no sharp dividing line between common and rare. Thus, statistical methods that can analyze both rare and common variants simultaneously are preferable. Furthermore, some form of model selection is absolutely necessary because the number of SNP predictors in most studies far exceeds the number of participants. The rare variants uncovered in resequencing will exacerbate the excess of predictors over responses. The recent papers[10–13] have stressed the role of penalized estimation in statistical genetics. Lasso penalties[18–20] have the interesting capacity to force many parameter estimates to zero. Model selection with a predetermined number of predictors can be achieved by tuning the strength of the lasso penalty. If model fitting is carried out by coordinate ascent or descent, then lasso penalized estimation is exceptionally fast.[12,25]

A particular rare disease predisposing allele may be present in only a handful of patients. Hence, statistical tests that capture only marginal effects are doomed to low power. This sad fact suggests focusing on disease gene discovery rather than disease variant discovery. One of the most attractive strategies for combining signals is to group variants by gene or pathway. Li and Leal[7] proposed a group-wise test exploiting both multivariate and collapsing strategies that possesses higher power than a simple multivariate test or simple collapsing. Madsen and Browning[8] extended the method by incorporating weights (dependent on allele frequency) into the group-wise statistics and approximating p-values by permutations within each group. Both methods consider rare variants with minor allele frequencies falling below a pre-specified threshold and exclude more common variants from analysis. The pooling strategy of Price et al[9] circumvents the issue of arbitrarily chosen frequency threshold by calculating a group-wise statistic under a variety of thresholds. Higher power is achieved at the cost of an increased computational burden.

These methods have certain drawbacks. Environmental predictors are excluded from analysis even though they may contribute significantly to an association. Multiple testing remains an issue. More importantly, existing methods are sensitive to the classification of variants. If all types of variants (causal, protective, or neutral) coexist, then the various signals can cancel one another and potentially compromise statistical power. Our recent paper[11] explores a remedy that groups variants by gene or pathway membership in penalized regression. The encompassing multiple regression framework allows simultaneous consideration of genetic and environmental predictors and overcomes the unfortunate cancelations of causal and protective variants. Here we continue our exploration of group penalties, with emphasis on weighted penalties that keep both common and rare variants in play. In accord with the notion that variants with large deleterious effects should be rarer than variants with small deleterious effects, lower weights should be assigned to variants with lower population frequencies.[8,9]

In pursuing group effects, we have attempted to retain the following advantages of lasso penalized estimation: (a) it applies to both ordinary regression (quantitative traits) and logistic regression (case-control studies), (b) it puts genetic and environmental predictors on the same footing, (c) it keeps both rare and common SNP predictors in play, (d) it partially circumvents the vexing issue of multiple comparisons, (e) it is computationally very efficient, (f) it offers a principled approach to model selection when the number of predictors exceeds the number of study participants, (g) it identifies protective variants as well as deleterious variants, and (h) it is amenable to finding interactions among predictors. We have previously demonstrated that Euclidean group penalties preserve these advantages.[11] Group penalties make it easier for related predictors to enter a model once one of the predictors does. Lasso penalties are retained to discourage the inclusion of neutral mutations in disease susceptibility genes. When disease genes harbor one or more borderline-rare variants with substantial risk, a mixture of lasso and group penalties performs well.

The major innovation in the current paper is the imposition of weights modulating lasso and group penalties. Ideally the weights should be chosen to reflect prior biological knowledge. In reality, we need better systems for rating the potential severity of point mutations. There is a severity hierarchy extending from non-synonymous mutations to synonymous mutations and ultimately to frameshift and protein truncating mutations. A non-synonymous mutation in a highly conserved codon is more important than the corresponding mutation in a less conserved codon. If both copies of a gene are disabled, this is a clear sign of trouble. If several genes in a common pathway are disabled or disregulated, the pathway as whole may be compromised. Integration of prior knowledge in penalized regression is an obvious priority, but until sequence data becomes more widely available, it is probably premature to pursue such elaborations.

The remainder of the paper is organized as follows. Section 2 describes the penalized regression framework with mixed lasso and group penalties, suggests a few plausible weighting schemes, and explains how both group penalties and weights can be implemented. Fortunately, the coordinate descent algorithms found successful in lasso penalized regression require trivial changes. Coordinate descent is exceptionally quick and permits optimal tuning of the penalty constant by cross-validation. Section 3 applies the mixed penalty method with weights to simulation examples. Section 4 provides a detailed description of the user interface to our implementation of penalized model selection in our statistical genetics program Mendel. We illustrate the mechanics of problem definition using the breast cancer data analyzed in our previous paper.[11] Finally, the discussion mentions some strengths and weaknesses of model selection under mixed penalties and suggests potentially helpful extensions.

2. Methods

Genome-wide association testing is one application field challenging current model selection procedures. All generalized linear models involve an $n \times 1$ response vector y and an $n \times p$ predictor matrix X. If the number of predictors p far exceeds the number of responses n, then some form of model selection is mandatory. Indeed, the ability to estimate parameters consistently requires the ratio p/n to tend to 0. Traditional model selection techniques include forward and backward stepwise regression and minimization of AIC (Akaike) and BIC (Bayesian) information criteria; the latter two lead to a combinatorial search over a space with 2^p possible submodels. For this reason statisticians have substituted penalized estimation for combinatorial search. Generally the objective function being minimized is a convex combination of a loss function (or negative loglikelihood) and a penalty function. Penalty functions act like priors in Bayesian statistics and must be carefully constructed to steer parameter estimates in productive directions. The following reasons are cause for optimism in applying penalization estimation in statistical genetics:

(a) Speed. Standard algorithms often choke when confronted with genomic-scale data. Efficient algorithms such as coordinate descent have been devised for solving convex optimization problems.[10,12,25]

(b) Flexibility. The modeling of complex biological phenomena is naturally embedded in the design of the loss and penalty functions. In association studies, biological meaningful units such as genes and pathways can be examined by introducing group penalties.[11] In copy number variation (CNV) reconstruction, copy number should change infrequently along a chromosome. Such smoothness is enforced by the fused lasso penalty.[15]

(c) Theoretical Justification. Recent advances in theoretical statistics justify the use of penalized estimation in high dimensional settings. Model selection consistency is especially relevant to association testing. Under certain regularity conditions, the predictors singled out by penalized estimation have a high probability of coinciding with the true predictors.[16,17]

(d) Empirical Justification. There are many success stories of penalized regression methods in natural language processing, remote sensing, financial engineering, and other application areas outside genetics.

2.1. *Penalized Regression with Weights*

In lasso penalized linear regression[12,18,19] estimates of the intercept μ and the regression coefficients β_j are derived by minimizing the objective function

$$f(\theta) = \frac{1}{2}\|y - \mu - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $\theta = (\mu, \beta)$, $\|z\|_2 = (\sum_j z_j^2)^{1/2}$ is the Euclidean (ℓ_2) norm, and $\|z\|_1 = \sum_j |z_j|$ is the taxicab (ℓ_1) norm. The sum of squares $\|y - \mu - X\beta\|_2^2$ represents the loss function minimized in ordinary least squares; the ℓ_1 contribution $\|\beta\|_1$ is the lasso penalty function. Its multiplier $\lambda > 0$ is the penalty constant. The order in which predictors enter a model as λ decreases is roughly determined by their impact on the response. Exceptions to this rule occur for correlated predictors. Because the intercept is felt to belong to any reasonable model, the lasso penalty omits it, and the intercept freely moves off zero.

Logistic regression is handled by replacing the sum of squares by the negative loglikelihood. The loglikelihood amounts to

$$L(\theta) = \sum_{i=1}^{n} \left[y_i \log p_i + (1 - y_i)\log(1 - p_i)\right], \tag{1}$$

where the success probability p_i for response i is defined by

$$p_i = \frac{e^{\mu + x_i^t\beta}}{1 + e^{\mu + x_i^t\beta}}. \tag{2}$$

Here the response y_i is 0 (control) or 1 (case), and x_i^t is the ith row of the predictor matrix X. To put the regression coefficients on an equal penalization footing, all predictors are centered around 0 and scaled to have approximate variance 1. There is a parallel development of lasso penalized regression for generalized linear models.[20] In each case the objective function is written as

$$f(\theta) = L(\theta) - \lambda\|\beta\|_1$$

as the difference between the loglikelihood and the lasso penalty. Because we now maximize $f(\theta)$, we subtract the penalty.

To construct a weighted lasso penalty, we assign a positive weight s_j to each predictor j and substitute the sum $\sum_j s_j|\beta_j|$ for $\|\beta\|_1$. A larger weight s_j corresponds to a higher penalty and discourages the j-th predictor from entering the model. Conversely, a smaller weight s_j exerts less penalty and encourages selection of the corresponding predictor. Eliminating a weight ($s_j = 0$) forces the j-th predictor to be retained in the model. In association testing, there are several sources of prior knowledge pertinent to assigning lasso weights:

(a) Genotyping Error. Variants that cannot be typed reliably should be penalized more.

(b) Allele Frequencies. In a different context, Madsen and Browning[8] propose the weight $s = \sqrt{p(1-p)}$ for a variant with population frequency p. This scheme assigns smaller penalties to rarer variants as suggested by classical population genetics theory. The more extreme weights $s = p(1-p)$ risk giving rare variants too much influence.

(c) Properties of Point Mutations. Several programs predict the functional effects of non-synonymous changes. The SIFT software of the Venter Institute,[22] PolyPhen-2,[23] and MAPP[24] represent a good start in quantifying the risk entailed by coding mutations.

(d) Conservation Across Species. Conservation scores are particularly valuable for assigning weights to noncoding mutations not covered by SIFT.

Integrating the weights derived from different types of information is a challenge. For the sake of simplicity, we adopt the allele frequency weights $s = 2\sqrt{p(1-p)}$ in our examples. The factor of 2 makes the weights scale between 0 and 1.

Yuan and Lin[21] have suggested Euclidean penalties as a natural way to group predictors. The lasso penalty $\|\beta\|_1$ and the ridge penalty $\|\beta\|_2^2$ separate parameters. If a parameter enters a model, then it does not strongly encourage or inhibit other associated parameters from entering the model. Euclidean penalties act more subtlely. Let G denote a group label and t_G a corresponding group weight. The objective function

$$f(\theta) = L(\theta) - \lambda \sum_G t_G \|\beta_G\|_2$$

incorporates a Euclidean penalty on each group. Here β_G is the subvector of the regression coefficients corresponding to group G. For the purposes of this paper, we take all $t_G = 1$. In studies with good candidate genes or pathways, it makes sense to reduce t_G for a candidate group. Groups with a single predictor are allowed. Singleton groups are advisable for dispersed variants far from any gene.

Euclidean group penalties run the risk of letting in response-neutral predictors. As soon as one predictor from a group enters a model, it opens the door for other predictors from the group to enter the model. For this reason we favor a mixture of group and lasso penalties.[11] Lasso penalties maintain the pressure for neutral mutations to be excluded, even if they occur in causal genes or pathways. There is no need to group SNPs that occur outside coding or obvious regulatory regions. Simultaneous imposition of lasso and Euclidean penalties has further advantages. In addition to enforcing model parsimony and selecting relevant parameters, both penalties improve the convergence rate in minimizing the objective function. Because the penalties are convex, they also increase the chances for a unique minimum point when the loss function is non-convex.

2.2. *Algorithms*

Traditional algorithms such as Newton's method and scoring falter on high-dimensional, non-smooth problems. Cyclic coordinate ascent-descent is a better choice. Block relaxation, a generalization of cyclic coordinate descent, cycles through disjoint blocks of parameters and updates one block rather than one coordinate at a time. Meier et al[26] use block relaxation to fit logistic regression with purely group penalties. The extreme efficiency of cyclic coordinate descent-ascent in penalized estimation stems from the low cost of the univariate updates and the fact that most parameters never budge from their initial value of 0. Here we summarize cyclic coordinate ascent-descent for linear and logistic regression with mixed lasso and group penalties. Full algorithmic details appear in our previous papers.[10–12] Adding weights imposes trivial changes to the algorithms.

In coordinate ascent we increase $f(\theta)$ by moving one parameter at time. If a slope parameter β_j is parked at 0, when we seek to update it, its potential to move off 0 is determined by the balance between the increase in the loglikelihood and the decrease in the penalty. The directional derivatives of these two functions measure these two opposing forces. The directional derivative of $L(\theta)$ is the score $\frac{\partial}{\partial \beta_j} L(\theta)$ for movement to the right and the negative score $-\frac{\partial}{\partial \beta_j} L(\theta)$ for movement to the left. An easy calculation shows that the directional derivative of $\lambda \|\beta_G\|_2$ is λ in either direction when $\beta_G = \mathbf{0}$. In this case note that $\|\beta_G\|_2 = |\beta_j|$. If $\beta_G \neq \mathbf{0}$, then the partial derivative of $\lambda\|\beta_G\|_2$ with respect to β_j is $\lambda\beta_j/\|\beta_G\|_2$. Hence, the directional derivatives both vanish at $\beta_j = 0$. In other words, the local penalty around 0 for each member of a group relaxes as soon as the regression coefficient for one member moves off 0.

2.2.1. *Logistic Regression*

In logistic regression the penalized loglikelihood with group and lasso penalties is

$$f(\theta) = L(\theta) - \lambda_{\mathrm{L}} \sum_j s_j |\beta_j| - \lambda_{\mathrm{E}} \sum_G t_G \|\beta_G\|_2,$$

where j ranges over all variants and G ranges over all groups. In practice, we fix the ratio of λ_{L} to λ_{E} and define $\lambda = \lambda_{\mathrm{L}} + \lambda_{\mathrm{E}}$. Formulas for the score vector $\nabla L(\theta)$ and the expected information matrix $\mathrm{E}[-d^2 L(\theta)]$ are well known[11] and need not be repeated here. The expected and observed information matrices coincide in logistic regression.

In penalized maximum likelihood estimation, coordinate ascent is implemented by replacing the loglikelihood by its local quadratic approximation based on the relevant entries of the score and observed information. The penalty terms are likewise approximated locally by linear or quadratic functions in the parameter being updated. The one-dimensional updates are not exact, but they can be computed easily by Newton's method. To update a slope parameter β_j, one resets $\beta_j = 0$ and commences maximization. If the directional derivatives to the right and left of 0 are both negative, then no progress can be made, and β_j remains at 0. Otherwise, maximization is confined to the left or right half-axis, whichever shows promise. Because the objective function is concave, the two directional derivatives at 0 cannot be simultaneously positive. Newton's method almost always converges within five iterations. At each iteration one should check that the objective function is driven uphill. If the ascent property fails, then the simple remedy of step halving is available.

2.2.2. *Linear Regression*

In ordinary linear regression, the objective function to be minimized is

$$f(\theta) = \frac{1}{2}\|y - \mu - X\beta\|_2^2 + \lambda_{\mathrm{L}} \sum_j s_j |\beta_j| + \lambda_{\mathrm{E}} \sum_G t_G \|\beta_G\|_2.$$

Coordinate descent for linear regression also yields to Newton's method. Owing to the discontinuities in the penalties, once again iteration is confined to the left or right half-axis, provided either passes the directional derivative test. In contrast to unpenalized linear regression, minimization takes more than a single iteration. This complication just reflects the fact that the group penalty is neither linear nor quadratic.

2.3. ***Selection of Tuning Constants***

In principle, cross validation can be invoked to determine the optimal values λ_{L} and λ_{E}. As we show in our simulation, setting them equal works well. Given a fixed ratio of the two penalties, the total penalty $\lambda = \lambda_{\mathrm{L}} + \lambda_{\mathrm{E}}$ can be adjusted to deliver a predetermined number of genes or SNP variants. Because the number of non-zero predictors entering a model is generally a decreasing function of λ, a bracketing and bisection strategy is effective in finding a relevant λ.[10] Of course, the smaller the number of predictors desired, the faster the overall computation proceeds. If computing time is not a constraint, it is helpful to optimize the objective function over a grid of points and monitor how new predictors enter the model as λ decreases. Another way to choose λ is to minimize either the BIC or AIC criterion as a function of λ. Recall that the purpose of convex relaxation is to avoid the combinatorial search entailed by the traditional application of the AIC and BIC criteria. Guiding the choice of λ by these criteria is a better tactic.

3. Analysis of Simulated Data

Our first simulation example, admittedly a toy example, involves 1000 controls and 1000 cases under different scenarios reflecting heterogeneity in both minor allele frequencies (MAF) and relative risks (RR). We assume 10 participating genes, each with 5 rare variants. Across the variants the MAFs are simulated from the Wright-Fisher distribution under balancing selection

$$f(p) \propto c p^{\alpha_s - 1}(1-p)^{\alpha_n - 1} e^{\sigma(1-p)},$$

where c is a scaling constant such that $\int_0^1 f(p)\,dp = 1$ and σ is a selection coefficient. We take $\alpha_s = 0.2$, $\alpha_n = 0.002$, and $\sigma = 15$.[27] For $i = 1, \ldots, 5$, gene i has i causal rare variants. Therefore, the model has 15 causal rare variants dispersed over 5 genes and 35 neutral rare variants dispersed over 10 genes. All neutral variants have relative risk (RR) 1; causal variants'

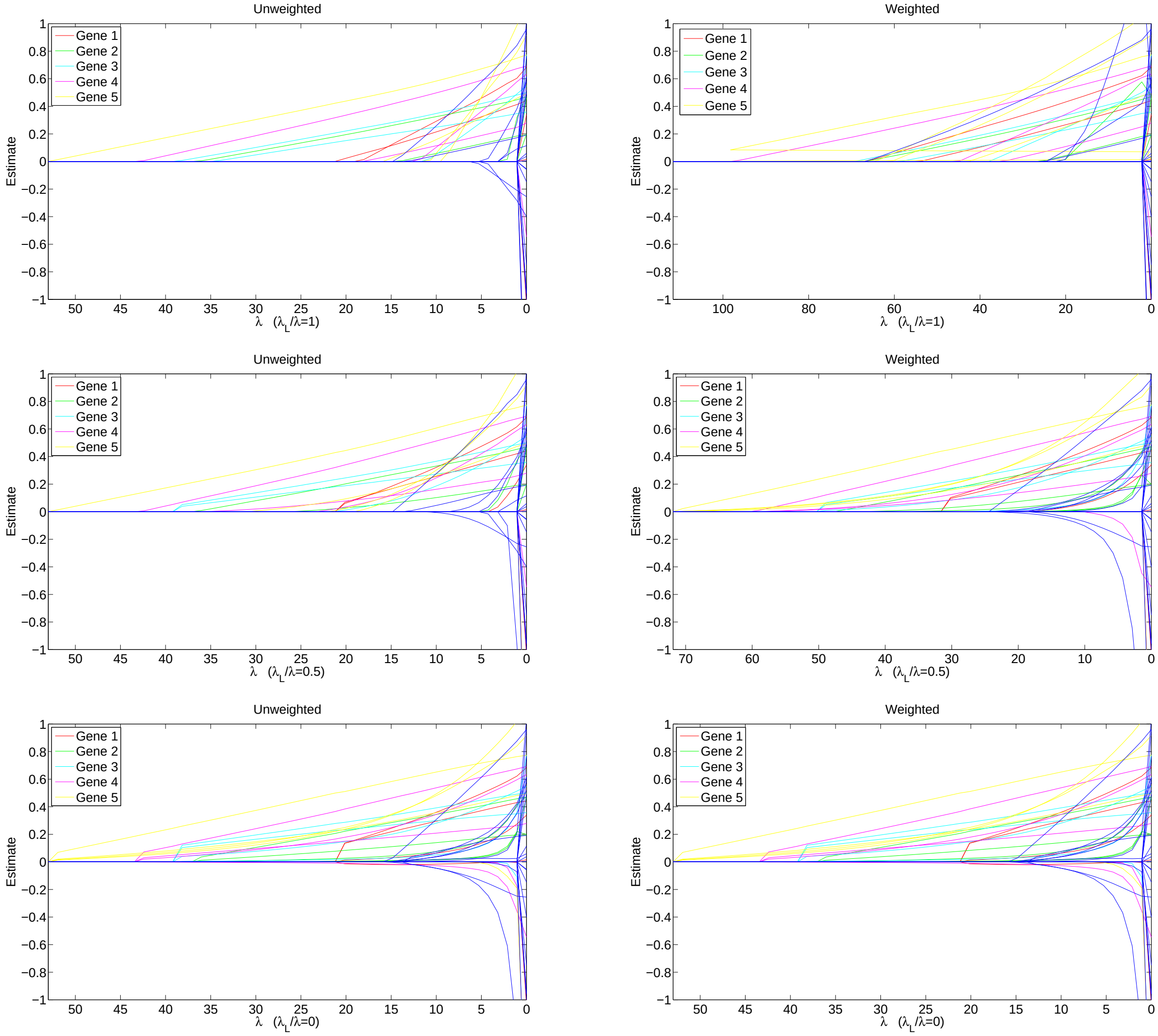


Fig. 1. Solution paths of parameter estimates under lasso penalties (top row), mixed penalties (middle row), and group penalties (bottom row). Left column: $s_j \equiv 1$ and $t_G \equiv 1$ (no weighting). Right column: $s_j = 2\sqrt{p_j(1-p_j)}$ and $t_G \equiv 1$.

RRs are drawn uniformly from the interval [1.2,5]. The wild-type penetrance f_0 is set at 0.01. For more details on data simulation algorithm, see our previous paper.[11] Figure 1 shows the solution paths of lasso, mixed penalty, and group penalty estimates with and without weights $s_j = 2\sqrt{p_j(1-p_j)}$, where p_j is the MAF estimated from the controls. All group weights are set to 1. The pure lasso penalty ($\lambda_{\mathrm{L}}/\lambda = 1$) picks up significant variants sequentially. The pure group penalty ($\lambda_{\mathrm{L}}/\lambda = 0$) picks up genes (groups) 1, 2, and 3 sequentially. The mixed group plus lasso penalty ($\lambda_{\mathrm{L}}/\lambda = 0.50$) achieves a good compromise between the two.

To discern the effects of weighted and unweighted penalized estimation, we repeat the same simulation 100 times and plot ROC curves for selected variants and genes in Figure 2. Each point of the ROC curves records the true and false positive rates of the selected

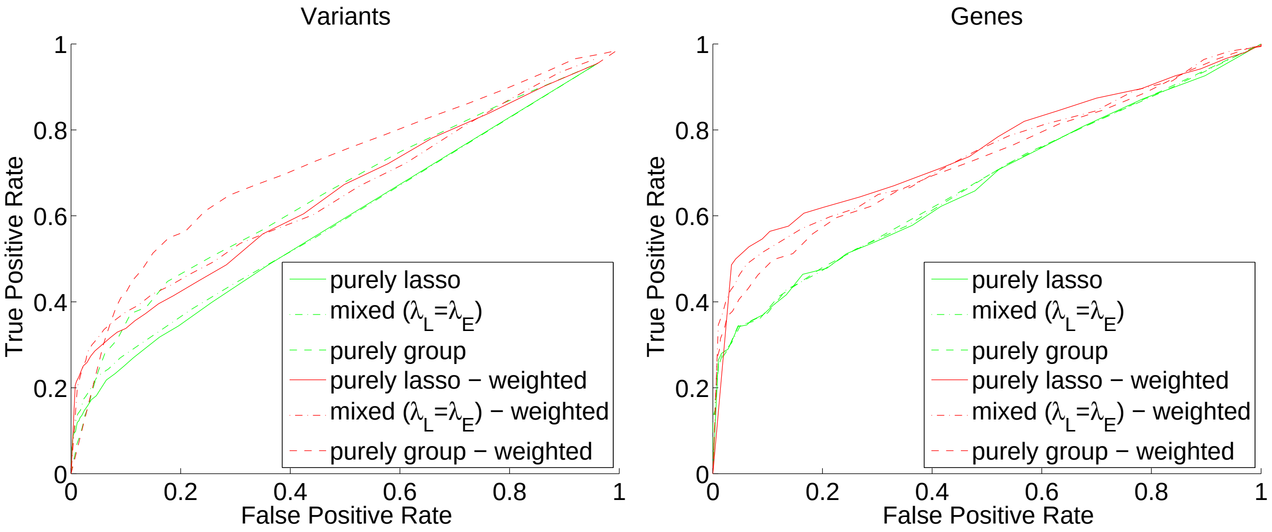


Fig. 2. ROC curves based on 100 simulations using the setup of Figure 1.

variants (left panel) and genes (right panel) at a specific λ value. A true positive for selection of a gene is defined as choosing any true variant within that gene. In all three situations, adding weights improves the selection of causal variants and genes. Indeed, the ROC curves shift visibly toward the upper left. Also notice that for acceptable false positive rates (less than 0.05) the mixed-weight penalty provides the best true positive rates for selection of both variants and genes.

4. Software Implementation and Illustration of Real Data

The methods we have described are implemented in the statistical genetics software Mendel[6] and will be freely available in its next public release, version 10.5 or higher. Mendel is available for Linux, MacOS, and Windows at `http://www.genetics.ucla.edu/software`. Within Mendel the SNP association option handles GWAS (genome-wide association study) data, both simple marginal p-value calculations and the above lasso based analyses.

We previously applied mixed penalized logistic regression to a familial breast cancer dataset[11] with SNPs assigned to genes involved in double strand break repair. We now take advantage of these data to illustrate the mechanics of our implementation in Mendel. The data originate from genotype samples of participants enrolled in the UCLA Family Cancer registry. We performed penalized logistic regression in which the response, breast cancer status (affected versus unaffected), is coded as a binary outcome. Our sample contains 399 Caucasian participants, of whom 196 were affected and 203 were unaffected. Covariates include age, Ashkenazi Jewish heritage, and education level. We imputed missing non-genetic predictors using the mean value for a continuous variable and the most frequent category for a categorical variable. Overall 148 SNPs from 17 genes in the DSBR pathway were typed and grouped by gene. Missing SNP data were imputed using the SNP Imputation option of Mendel.[6] For a complete description of the data, results, and insights gained from mixed penalized analysis, see our companion paper.[11] Mendel takes less than five seconds on a standard desktop computer to complete all analyses on this dataset. On a more challenging dataset with 10,000 SNPs and 2,200 individuals, Mendel completes all marginal and lasso analyses in under 30 seconds.

The input files used for the breast cancer and other analyses adhere to the usual Mendel conventions. In particular, the compressed SNP genotype data file conforms to the standard binary format adopted by both Plink and Mendel. SNP group designations and weights are optional. If they are desired, then they should be deposited in the SNP definition input file alongside the name, chromosome, and base pair position of each SNP. If no group is specified for a SNP, it is considered to be a singleton group. If no weight is specified for a SNP, then the SNP is assigned the default weight $2\sqrt{p(1-p)}$, where p is its MAF. The user may specify a value for the ratio $\lambda_{\mathrm{L}}/\lambda$ by invoking the keyword LASSO_PROPORTION in the Control file. Mendel reads all optional parameter settings from the Control file. To provide flexible modeling, the user can force any predictor or group to be retained in the lasso model by assigning to the keywords RETAINED_PREDICTOR or RETAINED_GROUP the corresponding predictor or group name. If a retained group is specified, then all predictors within that group are retained. For example, the Control file snippet

```
Analysis_option = SNP_Association
Model = 2
Quantitative_trait = BC
Marginal_analysis = True
Lasso_analysis = True
Desired_predictors = 50 :: marginal
Desired_predictors = 20 :: lasso
Lasso_proportion = 0.5
Predictor = Grand :: BC
Predictor = Age :: BC
Transform = standardize :: Age
Retained_predictor = rs11571476
Retained_predictor = Age
Retained_group = XRCC4
```

instructs Mendel to perform SNP association analysis using cases and controls. The value 2 for the keyword Model implies logistic regression; the default value 1 implies ordinary linear regression. The third command in the above Control file indicates that affection status pertains to the trait BC. Both a marginal and lasso analysis will be performed, with the top 50 marginal predictors and the top lasso set of 20 predictors reported in a Summary output file. Marginal results on all predictors are always reported in another output file intended for plotting. For this analysis run, the ratio $\lambda_{\mathrm{L}}/\lambda$ ratio is set to 0.5. If the keyword LASSO_PROPORTION is not specified, the ratio has its default value of 1. All defined SNPs are always included as predictors unless specifically excluded in a SNP exclusion file. In this example two non-SNPs are named as predictors for the trait BC, a mandatory grand mean and an optional variable Age. The Transform keyword specifies that the Age variable will be normalized prior to analysis; we recommend normalization for all quantitative predictors. Finally, the above Control file specifies that the two predictors rs11571476 and Age and all predictors in the group XRCC4 should be retained in the lasso model.

As mentioned, most results are presented in a Summary output file. At the top of this file appear the results for each predictor individually. For example, the first few rows of marginal results might be

```
PREDICTOR        MARGINAL    REGRESSION  STANDARD   HARDY-      MINOR   GENOTYPING  GROUP
  NAME           P-VALUE      ESTIMATE    ERROR    WEINBERG    ALLELE    SUCCESS    NAME
                                                   P-VALUE   FREQUENCY    RATE
Grand Mean               -   -0.03509         -          -          -          -        -
Age             0.2347E-04    0.43700    0.10660          -          -          -        -
rs9634161          0.00760         -          -     0.19917    0.15539    1.00000   RAD52
rs16889040         0.00768         -          -     0.49854    0.25815    1.00000   RAD21
rs4986763          0.01123         -          -     0.20101    0.37469    1.00000   BRIP1
rs16888997         0.01298         -          -     0.67786    0.25815    1.00000   RAD21
rs16888927         0.01932         -          -     0.17591    0.26817    1.00000   RAD21
rs1120476          0.02024         -          -     0.48503    0.43233    1.00000   XRCC4
```

To decrease computation time, regression estimates are only calculated for predictors with marginal p-values more significant than 0.001. This default threshold can be reset by the user. A table of false discovery rates for the marginal p-values appears after the single predictor summary.

The results of the lasso analysis are listed after the marginal results in the Summary file. For example, the first few rows of lasso results might be

```
PREDICTOR        MARGINAL  LEAVE-ONE-OUT  REGRESSION   HARDY-    MINOR  GENOTYPING  GROUP
  NAME           P-VALUE       INDEX      ESTIMATE   WEINBERG   ALLELE   SUCCESS    NAME
                                                     P-VALUE  FREQUENCY   RATE
Age             0.2347E-04  0.1645E-05    0.50391           -         -        -        -
rs9634161          0.00760     0.00166   -0.42841     0.19917   0.15539   1.00000   RAD52
rs2061783          0.35871     0.01508    1.46004  0.2611E-10   0.03509   1.00000   XRCC4
rs10514249         0.02687     0.02757   -0.80396     0.86985   0.43985   1.00000   XRCC4
rs2075685          0.34106     0.05623   -0.38380     0.05712   0.42105   1.00000   XRCC4
rs2887531          0.50526     0.08627   -0.24576     0.11833   0.23183   1.00000   RAD52
rs11571476         0.05510     0.08633    0.30271     0.68282   0.42481   1.00000   RAD52
```

Since our example Control file specified that group XRCC4 should be retained, all members of that group will be included in the complete lasso output set. The lasso output is sorted by the leave-one-out index, which is simply the p-value of the likelihood ratio test of the full regression model, using all predictors in the lasso set, versus the model leaving out the specified predictor. Because of the prior selection of predictors, the leave-one-out index is not a legitimate p-value.

5. Discussion

This paper presents penalized estimation as a framework for association testing in the presence of both common and rare variants. Our results partially vindicate the twin strategies of mixed group penalties and penalty weights acting at either the single predictor or the group level. Penalty weights provide a flexible way of incorporate prior biological knowledge and have the potential to increase power in association mapping. Even choosing to weight individual variants by their population frequencies makes a difference in sorting through the confusion of causal genes and neutral variants within them. Although our recommended tactics improve both false positive and false negative rates, they represent an incremental improvement rather than a panacea. In our opinion, there is still room for further improvement. More progress is apt to come through more nuanced weights or propensity scores cumulating risks across the whole spectrum of variants within a gene or pathway. Replacing variant predictors by group-wise propensity scores may serve to reduce the number of predictors and the need for differential penalty weights altogether.

Acknowledgments

We thank Dr. Patricia Ganz and the UCLA Family Cancer Registry. Funding for this work was provided in part from USPHS (GM53275, MH59490, CA87949, CA16042) and a Young Investigator Award from the American Society of Clinical Oncology (M.E.S.).

References

1. E. Hodges, Z. Xuan, V. Balija, M. Kramer, M. N. Molla, S. W. Smith, C. M. Middle, M. J. Rodesch, T. J. Albert, G. J. Hannon and W. R. McCombie, *Nature Genetics* **39**, 1522 (2007).
2. E. H. Turner, C. Lee, S. B. Ng, D. A. Nickerson and J. Shendure, *Nature Methods* **6**, 315 (2009).
3. D. Pinto et al. *Nature* doi:10.1038/nature09146
4. A. Sipos, F. Rasmussen, G. Harrison, P. Tynelius, G. Lewis, D.A. Leon and D. Gunnell, *BMJ*, 329 (2004).
5. E.M. Frans, S. Sandin, A. Reichenberg, P. Lichtenstein, N. Langstrom and C.M. Hultman, *Arch Gen Psychiatry*, **65** (2008).
6. K. Lange, R. Cantor, S. Horvath, M. Perola, C. Sabatti, J. Sinsheimer, E. Sobel, *AJHG*, **69** (2001).
7. B. Li and S. M. Leal, *AJHG*, **83**, 311 (2008).
8. B. E. Madsen and S. R. Browning, *PLoS Genet* **5**, p. e1000384 (2009).
9. A. L. Price, G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples, L.-J. Wei and S. R. Sunyaev, *AJHG*, **86**, 832 (2010).
10. T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel and K. Lange, *Bioinformatics* **25**, 714 (2009).
11. H. Zhou, S. Sehl, J. Sinsheimer and K. Lange, *Bioinformatics*, in press.
12. T. T. Wu and K. Lange, *Ann. Appl. Stat.* **2**, 224 (2008).
13. K. Ayers and H. Cordell, *Genetic Epidemiology*, to appear.
14. P. C. Ng and S. Henikoff, *Nucleic Acids Research* **31**, 3812 (2003)
15. Z. Zhang, K. Lange, R. Ophoff and C. Sabatti, *Ann. Appl. Stat.* **41** (2010).
16. P. Zhao and B. Yu, *J. Mach. Learn. Res.* **7**, 2541 (2006).
17. P. Ravikumar, M. J. Wainwright and J. Lafferty, *Ann. Stat.* (in press)
18. D. L. Donoho and I. M. Johnstone, *Biometrika* **81**, 425 (1994).
19. R. Tibshirani, *J. Roy. Statist. Soc. Ser. B* **58**, 267 (1996).
20. M. Y. Park and T. Hastie, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 659 (2007).
21. M. Yuan and Y. Lin, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 49 (2006).
22. P. Kumar, S. Henikoff and P. C. Ng, *Nature Protocols* **4**, 1073 (2009).
23. I. Adzhubei, S. Schmidt, L. Peshkin, V. Ramensky, A. Gerasimova, P. Bork, A. Kondrashov and S. Sunyaev, *Nat. Methods* **7**, 248 (2010).
24. E. Stone and A. Sidow, *Genome Research* **15**, 978 (2005).
25. J. Friedman, T. Hastie, H. Höfling and R. Tibshirani, *Ann. Appl. Stat.* **1**, 302 (2007).
26. L. Meier, S. van de Geer and P. Bühlmann, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**, 53 (2008).
27. J. Pritchard, *AJHG* **69**, 124 (2001).

MICROBIOME STUDIES: PSB 2011 SPECIAL SESSION INTRODUCTION[*]

JAMES A. FOSTER[†]

Department of Biological Sciences, University of Idaho, Moscow, ID 83844-3051 USA
Initiative for Bioinformatics and Evolutionary STudies (IBEST)
BEACON Center for the Study of Evolution in Action
Email: foster@uidaho.edu

JASON MOORE[†]

Institute for Quantitative Biomedical Sciences, Departments of Genetics and Community and Family Medicine,
Dartmouth Medical School Lebanon, NH 03756 USA
Email: jason.h.moore@dartmouth.edu

Recent advances in sequencing technologies have made is possible, for the first time, to take a thorough census of the microbial species present in a given environment. This presents a particularly exciting opportunity since bacteria and archea comprise the dominant forms of life on earth, and since they are vital to human health and to the wellbeing of our environment. However, the bioinformatics for interpreting these very large sequence datasets are not fully developed. This session presents recent work supporting the computational analysis of microbiome data.

1. Introduction to Microbiome Studies

1.1. *Producing Hard Copy Using MS-Word*

Microbes, including both eubacteria and archaea, are the dominant forms of life on earth, in absolute numbers, biomass, and diversity of ecosystems. During more than half of the Earth's 3.5 billion year biological history, only microbes were present. Microbial physiology is a dominant factor in carbon cycling, greenhouse gas emission, and oxygen production. In the human body there are ten times more microbial cells than human cells, and there are two orders of magnitude more microbial gene products than human gene products. Unfortunately, over 97% of microbes cannot be cultivated with current techniques, which has significantly biased the choice of model systems, as well as microbial genome sequencing and bioinformatics. Even our evolutionary and ecological theories and software were developed with macro-biology in mind, and often appear to be ill suited to studying the microbial world. Consequently, we have until recently been unable to fully understand and appreciate some of the most important ecological systems on earth.

Fortunately, new sequencing and bioinformatics technologies, such as tagged barcoding, community genomics, pyrosequencing, and metagenomics have made it possible to study the structure and dynamics of microbial communities. Some natural microbiomes that have recently been characterized include surveys of human microbiomes and their relationship to human health

[*] This work is partially supported by NIH P20RR16448 from the NCRR IDEA/COBRE program and by NSF "Evolution in Action" STC DBI 0939454

[†] Work partially supported by NIH R01 LM009012 and P20 RR018787.

in the human gut, skin, mouth, and reproductive tracts and ecological surveys of soil, air, and water to understand the effects of climate change and pollution.

A crude estimate of the number of publications in microbiome studies (using a Pubmed search for “microbiome OR metagenome OR community genomics OR microbial ecology”) has exploded in recent years, growing from 248 in 2000 to 1102 in 2009, with a total of 7117 hits to date. Publications to date are on track to double or triple in the coming year.

During this time, funding for microbiome studies has significantly increased, including signature areas such as the Human Microbiome Project from NIH. Community resources have become widely used (such as RDP, Greengenes/Silva, CAMERA, VAMPS, HMP DACC) and comprehensive bioinformatics tool suites are (such as mg-rast, mothur, catchall, and unifrac) being developed and widely used. Moreover, “the personal microbiome” may prove to be as important as “personal genomics”, the theme of PSB 2010.

These were the considerations that led to this PSB special session on “microbiome studies”. Our intention is to bring the expertise of the PSB community to bear on this increasingly important new field.

2. Papers in this session

One major challenge posed by microbial sequence data is to infer function and ecology of complex microbial communities from very large sequence datasets. This is the challenge addressed in this year’s session.

Most of the papers in this session present tools or frameworks to facilitate interpreting community function or composition from 16S fragments extracted from the environment and analyzed directly. The 16S gene codes for the small subunit of the ribosome, which is essential to DNA replication. Therefore this molecule is strongly conserved even in very ancient lineages. Woes introduced the use of the 16S gene as a phylogenetic marker for microbes, thereby showing that the archea comprise a distinct third kingdom of life. It has since become standard practice to use the similarity of several hypervariable regions in the 16S genes of microbes to identify and distinguish populations of microbes.

Holmes et al., “Visualization and Statistical Comparisons of Microbial Communities using R packages on Phylochip Data” introduces packages for the very popular R statistical analysis package which interpret data from the PhyloChip. This is a microarray with 16S targets for 8743 distinct bacteria and archea. Thus this paper represents a technology for identifying microbial communities using microarray technologies. The utility of this tool is limited by the set of targets on the Phylochip, of course. However, the R packages will be useful for any data from similar microarrays, including potentially customized arrays for specific complex screenings.

So called “metagenomic” or “barcode pyrosequencing” techniques have been developed to avoid the bias inherent in microarray design, and the even stronger bias that comes from culturing sequences and building clone libraries. (Note, “metagenomics” has multiple meanings, and we use it here to refer to 16S fragment analysis.). One approach, popularized by Roche, is to attach DNA “barcodes” to primers and then to perform very large scale PCR in micro-droplets that contain nano-beads with complementary sequences attached. This makes it possible to generate over million reads between 200bp and 500bp long, which is ideal for the hypervariable regions of 16S. Other technologies exist and are emerging that can generate far more reads, and these technologies are progressing rapidly.

This much data creates a major bioinformatics challenge. Several software packages or web services are emerging to provide bioinformatics support for this type of data. Two are presented in this session. Eran et al., "A FRAMEWORK FOR ANALYSIS OF METAGENOMIC SEQUENCING DATA" presents a software framework that allows scientists to build custom workflows for their "next generation" data analysis—with a particular emphasis on 16S microbial community sequence data.

The paper by Moore et al., "Human microbiome visualization using 3D technology", addresses the problem of making sense of microbiome data analysis visually. Scientists often need to "play with" possible interpretations of very large datasets, searching for more precise hypotheses to be verified or just getting a handle on what the data are like. This paper presents a possible framework for this challenge. This differs from the many existing "pipelines" in that it enables the user to directly customize their software to support individual workflows.

The paper by Bunge, "Estimating the Number of Species With CatchAll," presents the newly expanded CatchAll package for microbial community analysis. Most statistical techniques for inferring species richness and other ecological measure of diversity are nonparametric and based on relatively small samples from relatively small populations, having been derived to interpret mega-biome data such as that from forests or reefs. But techniques that work for hundreds of species often fail to work robustly for thousands, especially when there are typically dozens to hundreds of "rare" species in the sample. CatchAll provides the classical measures of diversity, but also adds some novel parametric estimates that seem to work very well for microbiome analyses. (Note: be sure to see Dr. Bunge's tutorial on ecological diversity estimations as well.)

In summary, we are pleased with the orientation of this new PSB special session toward practical solutions to the very large data interpretation problems arising from next generation sequencing and microbiome studies.

ESTIMATING THE NUMBER OF SPECIES WITH CATCHALL

JOHN BUNGE
Department of Statistical Science, 1198 Comstock Hall, Cornell University, Ithaca, NY 14853 , USA
E-mail: jab18@cornell.edu

In many situations we are faced with the need to estimate the number of classes in a population from observed count data: this arises not only in biology, where we are interested in the number of taxa such as species, but also in many other fields such as public health, criminal justice, software engineering, etc. This problem has a rich history in theoretical statistics, dating back at least to 1943, and many approaches have been proposed and studied. However, to date only one approach has been implemented in readily available software, namely a relatively simple nonparametric method which, while straightforward to program, is not flexible and can be prone to information loss. Here we present CatchAll, a new, platform-independent, user-friendly, computationally optimized software package which calculates a powerful and flexible suite of parametric models (based on current statistical research) in addition to all existing nonparametric procedures. We briefly describe the software and its mathematical underpinnings (which are treated in depth elsewhere), and we work through an applied example from microbial ecology in detail.

Keywords: species richness; finite mixture model; abundance.

1. Introduction

In many applied settings we encounter the need to estimate the number of classes in a population based on observed sample count data. For example, in biology it is common to collect a sample of organisms and sort them into taxa – we will use the term "species" for these taxa, recognizing that this may not be exact in some cases – count the number of representatives of each species in the sample, and from this data estimate the total number of species, both seen and unseen, in the underlying population. This is called the "species richness." There are many examples from other fields as well, such as veterinary medicine, where one may wish to estimate the number of farms with animals having a certain disease, or software engineering, where interest is in the number of potential types of errors in a complex software program.[1] Statisticians have been interested in this problem since the time of R. A. Fisher,[2] and many approaches have been studied theoretically and tested empirically, ranging from frequentist to Bayesian and from parametric to nonparametric.[3] However, to date only one class of statistical methods has been implemented in readily available software, namely the (frequentist) coverage-based nonparametric estimators of Chao and colleagues.[4] (This is not the only possible class of nonparametric estimators; see Section 3 below.) These are provided in, e.g., SPADE[5] and EstimateS,[6] and in some broader-use bioinformatics packages such as mothur[7] and QIIME;[8] see Section 2 for details.

The coverage-based nonparametric estimators are mathematically simple and computationally straightforward, and these estimators, known as Good-Turing, Chao1, the Abundance-Based Coverage Estimator ACE and its variants, and Chao-Bunge, can be accurate in some situations. (The associated standard errors are more complex computationally; see Section 2.2 below.) However, it is known that these estimators are typically downwardly-biased in high-diversity situations[9] such as arise in modern high-throughput DNA sequencing studies, for instance. Furthermore, they are sensitive to inclusion/exclusion of outliers, i.e., species that appear with high abundance in the sample, so it is standard practice to truncate species abundance counts at 10 when using these estimators, that is, to ignore species with sample counts higher than 10, adding the number of such species to the estimate *ex post facto*. In addition, these estimators do not admit goodness-of-fit testing or other diagnostic assessments, and it is not clear how to graph or visualize the results.

In contrast, recent statistical research has elucidated a class of parametric *finite mixture models*[9] which are accurate in high-diversity populations (when the model is correct), are relatively insensitive to outliers, and permit a broad array of diagnostic and goodness-of-fit assessments, both quantitative and graphical. The basic idea is to "mix" several component parametric models together (i.e., to form a convex combination of them) so that one component fits the rare species and another the abundant ones (possibly using one or more additional components for improved fit to the sample count data). Estimators based on these models are not simple to compute, though, requiring multidimensional numerical search routines to obtain maximum likelihood estimates of the parameters (based on the expectation-maximization or EM algorithm), and model-selection procedures which are partly statistical and partly heuristic. In addition, computation of standard errors is quite involved, requiring numerical computation of inverse Fisher information matrices that can involve thousands of lines of code.

We originally explored the use of these models in biological applications by building a proof-of-concept system on a cluster in Cornell's Center for Advanced Computing using Maple,[10] but while functional the system was very slow, sometimes taking a week to complete an analysis. We analyzed several hundred datasets using this system (many from microbial ecology), and based on this experience we re-engineered our algorithms and rebuilt the system using a combination of C# and C. The result is CatchAll, a freely downloadable, user-friendly, platform-independent (Windows/Macintosh/Unix, single-processor/cluster, GUI/batch) software program which computes the full suite of finite-mixture models and all known nonparametric coverage-based estimates. CatchAll then compares all of these results; selects the best in each category and the "best-of-the-best"; and returns recommended estimates to the user along with associated standard errors, confidence intervals, and goodness-of-fit assessments. For the GUI version there is also an Excel-based module which produces publication-quality graphics displaying the the fit of the parametric models to the data, and the comparative performance of the various estimators. CatchAll usually computes a complete analysis in a minute or two on a single-processor machine.

Our purpose here is not to enter into the mathematical, statistical and computational details of CatchAll, which are discussed elsewhere[9],[11] but rather to describe a complete case study resulting in an estimate of total species richness in a particular setting. In Section 2 we discuss the data and its analysis, and Section 3 we draw conclusions and mention some future directions for expansion of CatchAll. In the Appendix we give a brief outline of our most important algorithm, which computes maximum likelihood estimates for the parametric models.

2. Analysis of a microbial diversity dataset

The International Census of Marine Microbes (ICoMM) is a large-scale research project on microbial diversity, intended "to (1) catalogue all known diversity of single-cell organisms inclusive of the Bacteria, Archaea, Protista and associated viruses, (2) to explore and discover unknown microbial diversity, and (3) to place that knowledge into appropriate ecological and evolutionary contexts."[12] Part of the ICoMM activity consists of taking samples of marine microbial organisms for (among other purposes) diversity evaluation. Essentially, a sample of water is taken and microbial 16S rRNA sequences are extracted. These sequences are then clustered into "operational taxonomic units" or OTUs; in our example below two sequences are assigned to the same OTU if they share 97% sequence identity, but the 97% value is conventional rather than theoretically based and can be varied at the discretion of the investigator. The OTU frequencies are then counted: some OTUs contain only one member sequence (the "singletons"), others two, others three, and so on. Finally we reorganize this information as "frequency count" data, consisting of the number of OTUs having one member or element; the number having two; the number having three, and so on. We note that each stage of this process, from sampling to sequence alignment and comparison to clustering, is nontrivial and subject to variation across labs and differing interpretation of results,[13] but for our purposes here we will assume that the frequency-count data is obtained in an unambiguous and closed-ended manner.

2.1. *Example dataset*

The sample data analyzed below was collected on January 7, 2005, as part of the ICoMM subproject "Application of the 454 technology to active-but-rare biosphere in the oceans: large-scale basin-wide comparison in the Pacific Ocean," by Koji Hamasaki and Akito Taniguchi of The University of Tokyo; for full details on this sub-project see the ICoMM Microbial Oceanographic Biogeographic Information System MICROBIS.[14] The complete frequency-count data is shown in Table 1. There were 19854 sequences grouped into 3018 OTUs, with (in particular) 2013 singleton OTUs; the maximally abundant sample OTU contained 1784 sequences. Figure 1 shows the data with the best fitted parametric curve (we explain this in Section 2.2). We retain the original scale (rather than, say, a log-log scale) to show the steep descent from the left, followed by the long slow decay to the right (in fact the plot is truncated at a maximum frequency of 254, while the actual data extends to 1784). This shape is typical of high-diversity data, which is often encountered in microbial diversity studies.

2.2. *CatchAll analysis of example data*

The basic idea of parametric species richness estimation is to fit a curve to the frequency count data and to project this curve upwards and to the left, to an abscissa of zero, so as to obtain an estimate of f_0. The estimate of the total number of species, unobserved + observed, is then $f_0 + f_1 + f_2 + ...$ The curve is a mixed-Poisson distribution based theoretically on an underlying species abundance distribution. It is fitted to the data via maximum likelihood, and the same procedure also yields standard errors, fitted values, goodness-of-fit statistics, etc.[9] CatchAll fits an ordered suite or family of five parametric curves: the (ordinary or unmixed)

Table 1. ICoMM frequency count dataset "ABR 0005 2005 01 07." i = frequency; f_i = # of sample OTUs with frequency i.

i	f_i	i	f_i	i	f_i	i	f_i
1	2013	23	1	56	1	165	1
2	416	25	3	57	1	173	1
3	173	27	2	59	1	191	1
4	85	28	1	71	1	195	1
5	63	29	3	73	1	201	1
6	43	30	1	76	1	202	1
7	39	31	3	80	1	208	1
8	18	32	1	84	1	223	1
9	24	33	1	85	1	225	1
10	8	34	1	93	1	233	1
11	17	35	1	94	1	254	1
12	8	36	1	114	1	319	2
13	6	38	2	119	1	328	1
14	3	40	1	122	1	548	1
15	4	42	1	123	2	560	1
16	6	43	1	131	1	675	1
17	9	46	1	148	1	1036	1
18	2	48	1	150	1	1361	1
20	6	53	2	154	1	1526	1
22	4	54	1	155	1	1784	1

Poisson, which unrealistically stipulates equal species abundances and is useful mainly as a lower-bound benchmark for the true richness; the (single) geometric; and mixtures of two, three, and four geometrics. (For technical reasons these are called mixtures of exponentials in the output and display.) The Poisson is mathematically the zero-order model in this scheme, followed by first- (single geometric), second- (mixture of two geometrics), third- and fourth-order mixture models. The idea, as noted above, is to mix (form a convex combination of) several component sub-models, one component fitting the steep decline of the frequency count data on the left, another fitting the shallow decline on the right, and possibly others fitting intermediate parts of the data.

At this point a second issue arises. Any parametric curve is defined by a finite number of parameters (1, 1, 3, 5 and 7 in our models of order 0, 1, 2, 3, and 4, respectively), and consequently has finite flexibility. In most datasets it is not possible for any parametric curve to fit the entire extent of the data from f_1 to f_{max} (where f_{max} is the number of species, i.e., the frequency count, at the largest sample frequency). It is therefore standard practice to truncate the data on the right at some frequency, which we call τ; the statistical analysis is then based on $f_1, f_2, \ldots, f_\tau$, and the number of species with frequencies greater than tau (i.e., $f_{\tau+1} + f_{\tau+2} + \ldots + f_{\text{max}}$) is added to the estimate *ex post facto*. As noted above, in the coverage-based nonparametric methods τ is fixed at 10 (we return to this issue below). The parametric methods are more flexible, and we wish to base the statistical analysis on as much of the frequency count data as possible, that is, to use the largest possible τ for which we can still obtain a good fit of the model. CatchAll therefore fits every model at every value of

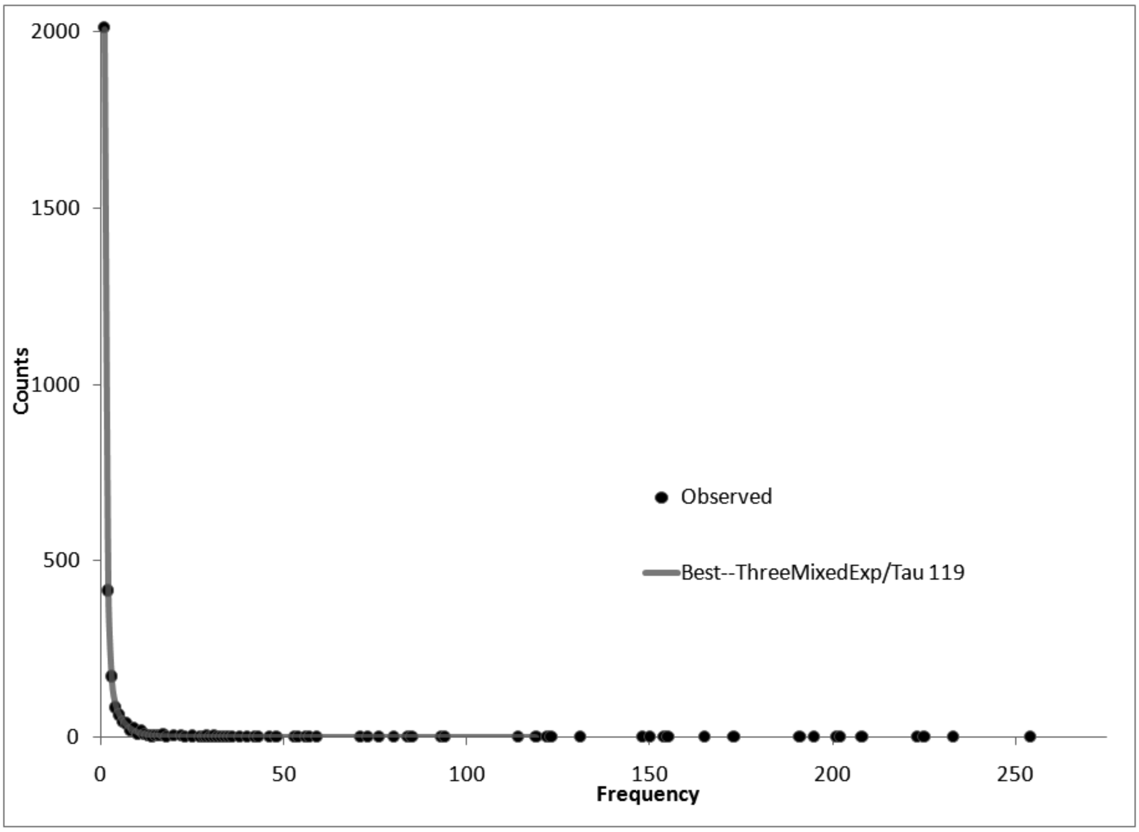


Fig. 1. Frequency count distribution of sample ICoMM data, with best fitted parametric curve.

τ and compares the results. Essentially we select the best-fitting parametric model at each fixed τ using the small-sample-size-adjusted Akaike Information Criterion, AICc, and then select across τ's using the p-values from the following two Pearson χ^2 statistics. "GOF0" is the p-value of a Pearson χ^2 goodness-of-fit test based on the "raw" or unadjusted frequency counts, and "GOF5" is the p-value of the χ^2 test after concatenating adjacent frequencies to obtain a minimum cell count of 5. We use both because the p-value of the χ^2 test is based on an asymptotic (large-sample) approximation to the distribution of the test statistic, and the aforementioned concatenation of adjacent cells is standard practice to obtain sufficiently large cell counts in sparse tables such as we typically have in our data (Table 1). Thus GOF0 should be regarded as a diagnostic "divergence statistic" signalling divergence from the null hypothesis (which states that the model is correct), while GOF5 represents the p-value from a legitimate statistical hypothesis test of model fit. In either case smaller p-values represent evidence against the fit of the model, and larger p-values represent evidence in favor.

The final model-selection algorithm, in outline form, is as follows.

Model selection algorithm

(1) (Statistical.) Eliminate model*τ combinations for which GOF5 < 0.01.

(2) (Statistical.) For each τ, select the model with minimum AICc (Akaike Information Criterion, corrected when necessary for small sample sizes).

(3) (Heuristic.) Eliminate model*τ combinations for which SE $>$ estimate/2.

Table 2. CatchAll analysis summary for ICoMM dataset. "Selection" = status of model, "Model" = order of parametric model or designation of nonparametric method, τ = upper frequency cutoff, "Est." = estimated total species richness, "SE" = standard error of estimate, "LCB" = lower 95% confidence bound, "UCB" = upper 95% confidence bound, "GOF0" = unadjusted χ^2 p-value, "GOF5" = adjusted χ^2 p-value.

Selection	Model	τ	Est.	SE	LCB	UCB	GOF0	GOF5
Best	3	119	15369	1322	13037	18243	0.0102	0.3199
2a	3	71	16032	1615	13231	19600	0.0704	0.0932
2b	4	254	16245	1833	13111	20352	0.0004	0.0785
2c	2	13	16604	1607	13802	20134	0.0004	0.0105
NP1	Chao1	2	7888	330	7283	8580		
NP2	ACE1	10	13519	777	12104	15156		
Parm $\tau_{\max}$	3	1784	13476	810	12005	15188		0.0000
NP $\tau_{\max}$	ACE1	1784	12227106	4012967	6532741	22887348		

(4) (Heuristic.) Then:

- Best model: Select the largest τ for which GOF0 ≥ 0.01.
- Model 2a: Select the τ with maximum GOF0.
- Model 2b: Select the largest τ.
- Model 2c: Select τ as close as possible but ≤ 10.

(5) (Heuristic.) If all model*τ combinations are eliminated, relax the restrictions in (3) and (4) and iterate.

We then report the "best-of-the-best" parametric model, along with three competing models 2a–2c, which are unordered in terms of preference.

The results of the ICoMM data analysis are shown in Table 2. The best analysis overall is given in the first row of the table. The fitted model is a mixture of three geometric components, one fitting the steep decline of the data on the left, one fitting the middle, and one fitting the shallow decline to the right. This is the curve shown in Figure 1, although the three components are of course not visible separately. The estimated total number of species is 15369 (i.e., the estimate of f_0 is 12351 so that $15369 = 12351 + 3018$). The standard error associated with the estimate of species richness is 1322. We do not form the "Wald" or Gaussian 95% confidence interval consisting of the estimate $\pm 1.96 * \mathrm{SE}$; rather, we use an asymmetric confidence interval based on a lognormal approximation due to Chao,[15] which is more realistic in this context. (In the parametric modeling setting the Chao interval is an approximation to the profile likelihood confidence interval, which though optimal is more complicated computationally and will appear in a later version of CatchAll.) The last two columns display the goodness-of-fit statistics GOF0 and GOF5. Both p-values exceed 0.01, indicating good fit of the model to the data.

Note that $\tau = 119$ for the best selected model, which, while still some distance from the maximum frequency of 1784, represents the use of the first 53 of the 80 frequencies existing in the data (66%). Thus the fitted curve in Figure 1 extends through 119 on the horizontal axis. The competing models 2a, 2b, and 2c (these are unordered in terms of desirability) represent various good but suboptimal compromises vis-à-vis goodness-of-fit, large τ, and other factors.

Model 2a again has 3 components and better GOF0 (than the best model), but smaller τ; 2b has four components and higher τ but GOF0$<$ 0.01, and 2c has low τ and low GOF0. Nevertheless their estimates, SEs and confidence intervals do not differ too much from those of the best model.

CatchAll also computes all known coverage-based nonparametric richness estimates, including that of Chao and Bunge.[4] The best of these are reported in the results table. Table 2 first shows "NP1" (meaning the first reported, not best selected) nonparametric analysis which is the "Chao1" statistic, a simple lower bound estimator with $\tau = 2$. This is useful as a cross-check or benchmark. Next, as "NP2" (the second reported nonparametric analysis), we report either ACE or its high-diversity variant ACE1, selected according to a criterion based on the coefficient of variation of the frequency count data.[4] These both have τ fixed at 10 (as noted above), as recommended in the original statistical research. In this connection we note that we have also re-engineered the standard error computation algorithms for the coverage-based nonparametric methods, yielding improved precision relative to existing software for these methods. For the ICoMM example data we see that ACE1 returns both an estimate and an SE that are lower than those of the parametric models, in accordance with the downward bias in high-diversity situations mentioned above.

We also report the best parametric and the best nonparametric analysis with τ fixed at the maximum frequency in the data, i.e., using the entire frequency count dataset. Table 2 shows that, while the parametric analysis at maximum τ differs little from the best selected parametric analysis (reflecting the relative insensitivity to outliers referred to above), the coverage-based nonparametric analysis "drifts off to infinity" along with its SE, when the large outlying frequencies are included.

To display the behavior of the various models and estimators as the larger frequencies are added to the data, the GUI version of CatchAll provides a bubble plot, shown in Figure 2. The figure displays the increase of the nonparametric estimates and their SEs as a function of increasing τ, compared to the relatively stable behavior of two of the parametric estimates (3rd- and 4th-order mixtures) as functions of τ. The bubble areas are proportional to 1/2 the associated standard error at each point. (Figure 2 has been reduced for simplicity to show only part of the τ-range and only two of the five parametric models.) It is clear that, while the estimators agree reasonably well at $\tau = 10$ (as was seen in Table 2), the nonparametric coverage-based estimates (and their error terms) diverge to infinity as τ increases, whereas the parametric estimates decrease slightly. Thus the coverage-based nonparametric methods can produce non-overlapping, hence contradictory, confidence intervals from the same dataset, depending on which large "outlying" frequencies are included in the analysis. The cause of this behavior has not yet been mathematically ascertained (although it is universally observed), and is a topic for further theoretical research.

The final selected analysis, referring again to the first row of Table 2, is the 3rd-order model at $\tau = 119$, and is indicated by an arrow in Figure 2.

Several existing and widely-used programs also compute (some of) the coverage-based nonparametric estimates. The Chao-Bunge estimator[4] is produced by SPADE;[5] ACE/ACE1 are produced by SPADE, EstimateS[6] and mothur;[7] and Chao1 is produced by SPADE, Es-

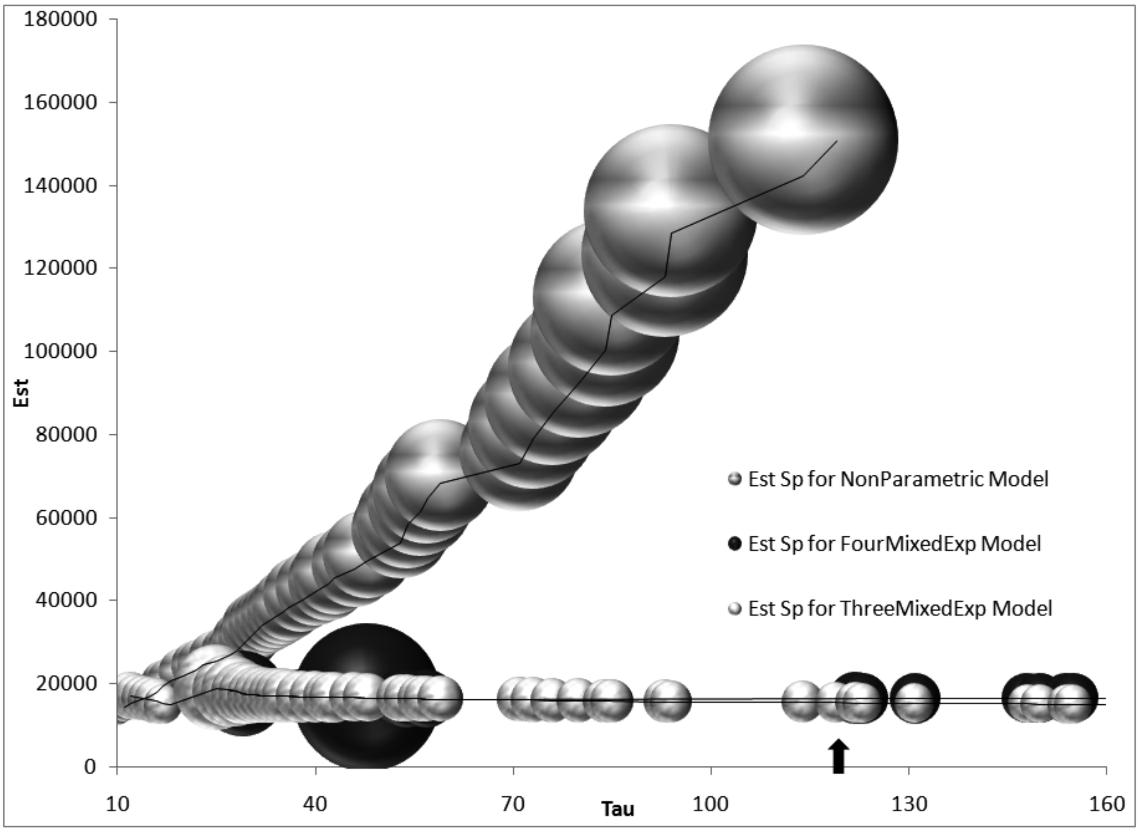


Fig. 2. Total species richness estimates as a function of τ in ICoMM data. “Est Sp for NonParametric Model” = total richness estimated using ACE or ACE1, “Est Sp for FourMixedExp Model” = total richness estimated using 4th-order parametric mixture model, “Est Sp for ThreeMixedExp Model” = total richness estimated using 3rd-order parametric mixture model.

timateS, mothur and QIIME[8] (accompanied in all cases by standard errors and confidence bounds). Thus the user will essentially find replicates of these nonparametric estimates in the cited programs and in CatchAll. There are two notable differences, however. First, CatchAll is unique in computing these nonparametric estimates at *every* value of τ, so as to reveal their behavior as more frequency counts are included in the data. Second, the standard error and confidence interval computations in CatchAll are based on new algorithmic representations of the underlying mathematics, and are considerably more accurate and precise than the usual algorithms used for this purpose.

3. Conclusions and future directions

We have presented CatchAll, a software program for parametric and nonparametric statistical estimation of total species richness, along with visualization and comparison of competing analyses. CatchAll is user-friendly — it requires no input other than the input data file specification from the user, that is, no options need to be set; it is freely downloadable, from http://www.northeastern.edu/catchall/; it is platform-independent and will run under Windows or the Macintosh operating system or Unix, on single- or multiple-processor machines, in GUI or in batch mode; and it is fast, completing most analyses in a minute or two on a modestly-specified machine. It is the first program to implement parametric species richness modeling in a comprehensive, integrated and accessible fashion, and it also computes all existing coverage-based nonparametric estimates (with improved standard errors). A full manual is also provided.

In terms of future developments, we will next extend CatchAll to include a completely novel species richness estimation method based on fitting a linear model to ratios of adjacent frequency counts.[1] We will then incorporate objective Bayesian methods using reference and Jeffreys priors for the number of species.[16] Finally we will implement nonparametric maximum likelihood estimation, another new approach in the species richness problem.[9] These are computationally intensive procedures which will take some time to program. We welcome comments and suggestions regarding the current or potential future versions of CatchAll, which may be addressed to the author.

Acknowledgments

We deeply appreciate the advice and assistance of the following collaborators. Linda Woodard is the principal CatchAll programmer and Consultant in Cornell University's Center for Advanced Computing. Sean Connolly programmed the Excel display module and is a 2010 graduate of Cornell. Linda Amaral Zettler of the Marine Biological Laboratory is Secretariat of ICoMM. Slava Epstein of Northeastern University is co-investigator on the CatchAll grant. This research was conducted using the resources of the Cornell University Center for Advanced Computing, which receives funding from Cornell University, New York State, the National Science Foundation, and other leading public agencies, foundations, and corporations. The development of CatchAll is funded by NSF grant DEB-0816638 to JB. We thank James Foster of the University of Idaho for proposing the submission of this paper.

Appendix A. Computing the maximum likelihood estimates

We outline the expectation-maximization (EM) algorithm for computing the maximum likelihood estimates of the parameters in the mixture-of-two-exponentials (-geometrics) model, when the frequency count data is truncated on the right at τ. Extending the algorithm to higher numbers of components (three and four) is straightforward though not simple.

The observed data consists of the (nonzero) frequency counts $f_1, f_2, \ldots$, where f_i denotes the number of species observed i times in the sample. The relevant part of the log-likelihood of the data under the model is[9]

$$\sum_{i=1}^{\tau} f_i \log\left(u\left(\frac{1}{t_1}\right)\left(\frac{t_1}{1+t_1}\right)^i + (1-u)\left(\frac{1}{t_2}\right)\left(\frac{t_2}{1+t_2}\right)^i \right), \tag{A.1}$$

where $t_1, t_2 > 0, u \in (0,1)$. Our objective is to find (t_1, t_2, u) to maximize (A.1) given $f_1, f_2, \ldots$. We initialize u as $u^{(0)} = 1/2$, and t_1, t_2 as

$$t_1^{(0)} = \frac{\sum_{i=1}^{\lfloor 2\tau/3 \rfloor} i f_i}{\sum_{i=1}^{\lfloor 2\tau/3 \rfloor} f_i} - 1, \qquad t_2^{(0)} = \frac{\sum_{i=\lfloor \tau/3 \rfloor+1}^{\tau} i f_i}{\sum_{i=\lfloor \tau/3 \rfloor+1}^{\tau} f_i} - 1.$$

Now suppose we are at the kth step, $k = 0, 1, \ldots$, so that we have values $t_1^{(k)}, t_2^{(k)}, u^{(k)}$. Define

$$z_i^{(k)} := \frac{u^{(k)} \left(\frac{1}{t_1^{(k)}}\right) \left(\frac{t_1^{(k)}}{1+t_1^{(k)}}\right)^i}{u^{(k)} \left(\frac{1}{t_1^{(k)}}\right) \left(\frac{t_1^{(k)}}{1+t_1^{(k)}}\right)^i + (1-u^{(k)}) \left(\frac{1}{t_2^{(k)}}\right) \left(\frac{t_2^{(k)}}{1+t_2^{(k)}}\right)^i},$$

$i = 1, 2, \ldots$.
<u>Update u</u>:

$$u^{(k+1)} = \frac{\sum_{i=1}^{\tau} f_i z_i^{(k)}}{\sum_{i=1}^{\tau} f_i}$$

<u>Update t_1, t_2</u>:

$$t_1^{(k+1)} = \frac{\sum_{i=1}^{\tau} f_i i z_i^{(k)}}{\sum_{i=1}^{\tau} f_i z_i^{(k)}} - 1;$$

$$t_2^{(k+1)} = \frac{\sum_{i=1}^{\tau} f_i i (1 - z_i^{(k)})}{\sum_{i=1}^{\tau} f_i (1 - z_i^{(k)})} - 1.$$

<u>Update z</u>: $z_i^{(k)} \to z_i^{(k+1)}$.
Iterate to convergence. This yields MLEs (t_1, t_2, u), which are the key quantities required for all estimates, standard errors, fitted values, and goodness-of-fit statistics.[9]

References

1. I. Rocchetti *et al.*, forthcoming in *Annals of Applied Statistics* (2010).
2. R. A. Fisher *et al.*, *Journal of Animal Ecology* **12**, 44 (1943).
3. J. Bunge and M. Fitzpatrick, *Journal of the American Statistical Association* **88**, 364 (1993).
4. A. Chao and J. Bunge, *Biometrics* **58**, 531 (2002).
5. T. J. Shen *et al.*, *Ecology* **84**, 798 (2003).
6. R. Colwell, http://purl.oclc.org/estimates.
7. P. Schloss, *Applied and Environmental Microbiology* **75**, 7537 (2009).
8. J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, *Nature Methods* **7**, 335 (2010).
9. J. Bunge and K. Barger, *Biometrical Journal* **50**, 971 (2008).
10. http://www.maplesoft.com/ .
11. J. Bunge *et al.*, in preparation (2010).
12. http://icomm.mbl.edu/
13. S. M. Huse, D. M. Welch, H. G. Morrison, M. L. Sogin, *Environmental Microbiology* **12**, 1889 (2010).
14. http://icomm.mbl.edu/microbis/
15. A. Chao, *Biometrics* **43**, 783 (1987).
16. K. Barger and J. Bunge, forthcoming in *Journal of Bayesian Analysis* (2010).

A FRAMEWORK FOR ANALYSIS OF METAGENOMIC SEQUENCING DATA

A. MURAT EREN
Department of Computer Science, University of New Orleans, 2000 Lakeshore Drive,
New Orleans, LA 70148, USA
Email: aeren@uno.edu

MICHAEL J. FERRIS
Departments of Pediatrics and Microbiology Immunology and Parasitology, Louisiana State University Health Sciences Center
New Orleans, LA 70112, USA
Email: mferris@chnola-research.org

CHRISTOPHER M. TAYLOR
Department of Computer Science, University of New Orleans, 2000 Lakeshore Drive,
New Orleans, LA 70148, USA
Email: taylor@cs.uno.edu

The human body is home to a diverse assemblage of microbial species. In fact, the number of microbial cells in each person is an order of magnitude greater than the number of cells that make up the body itself. Changes in the composition and relative abundance of these microbial species are highly associated with intestinal and respiratory disorders and diseases of the skin and mucus membranes. While cultivation-independent methods employing PCR-amplification, cloning and sequence analysis of 16S rRNA or other phylogenetically informative genes have made it possible to assess the composition of microbial species in natural environments, until recently this approach has been too time consuming and expensive for routine use. Advances in high throughput pyrosequencing have largely eliminated these obstacles, reducing cost and increasing sequencing capacity by orders of magnitude. In fact, although numerous arithmetic and statistical measurements are available to assess the composition and diversity of microbial communities, the limiting factor has become applying these analyses to millions of sequences and visualizing the results. We introduce a new, easy-to-use, extensible visualization and analysis software framework that facilitates the manipulation and interpretation of large amounts of metagenomic sequence data. The framework automatically performs an array of standard metagenomic analyses using FASTA files that contain 16S rRNA sequences as input. The framework has been used to reveal differences between the composition of the microbiota in healthy individuals and individuals with diseases such as bacterial vaginosis and necrotizing enterocolitis.

1. Background

Understanding the composition of microbial communities is important since microbes drive global nutrient cycles and there is a significant correlation between human microbial community composition, health and disease [1, 2]. Although they are not visible to the naked eye, microbes are ubiquitous in nature. Microbial cells constitute a large portion of the Earth's biomass [3] and the human body is colonized by bacteria in the gastrointestinal tract, oral cavity, skin, airway passages and urogenital system [4]. The 16S rRNA gene sequence has been widely used to detect bacterial species in natural specimens and to establish phylogenetic relationships among them. All bacteria possess this gene, which has highly conserved regions that are needed to construct phylogenies and are useful targets for PCR amplification and pyrosequencing analyses of microbial communities. The 16S rRNA gene also has hypervariable regions that are diverse enough to identify individual species [5]. Because of the large amount of sequence information associated with PCR amplification and pyrosequencing of 16S rRNA genes from microbial communities, a variety of statistical methods and extensive computational aid is needed for the analysis of the data. The primary goal of our work is to bring the analysis of large amounts of microbial community sequence data within reach of scientists who have only basic computer skills.

2. Framework

There are several computational methods available to process microbial community 16S rRNA gene sequence data in order to understand and compare bacterial populations within them. Most of these were not designed to manipulate large pyrosequencing files. Preparing individual scripts in order to manipulate large sequencing files for each analysis is a difficult solution that requires extensive programming skills and experience to maintain. We present a software framework that overcomes many of these challenges of metagenomic sequencing data analysis and provides researchers with an easy way to analyze and interpret their data.

2.1. *Motivation*

Software packages that are available to researchers to process 16S rRNA gene sequence data can be divided into two groups: those that are hosted on a server and used via web interfaces, and those that are downloaded and run locally. Both approaches have their benefits and their limitations. Online ribosomal sequence analysis applications and pipelines, such as Microbial Community Analysis (MiCA) [6] and the Ribosomal Database Project (RDP) pipeline [7], require researchers to upload their data over the Internet and work using web interfaces that are designed to be easy to operate. However online analyses usually have stringent limitations on the number of sequences that can be analyzed (or number of runs or permutations), primarily due to the fact that scarce resources, such as CPU time, memory size and network bandwidth, must be shared by many researchers in any centralized approach. Another limitation of this approach is that the software cannot be customized and enhanced for specialized analysis since it is running on another group's server. On the other hand software that can be downloaded and run locally such as MOTHUR [8] and QIIME [9], permits researchers to use their own computational resources without requiring them to upload their data to another server. However, since most of these applications necessitate the use of command line interfaces to perform function calls, the learning curve for these tools is steep and a significant investment of time is required to learn and operate them.

Another aspect of available 16S rRNA analysis software that limits its utility is the "pipeline" approach. Pipeline approaches are a model of computing where a set of applications are connected to each other such that output from one application becomes input to one or more applications in the subsequent stage. A pipeline approach is not an efficient structure for an application that is designed to analyze sequencing data. Applications in a pipeline cannot use previous applications' resources; these resources may need to be re-allocated or re-computed at every stage of the pipeline. This redundancy is not efficient use of computational resources and negatively impacts overall performance. In addition, the process of file upload, analysis and download, which may be repeated at different stages, is time consuming since the user must wait for output and must often upload results again for the next stage of analysis. Lastly, the preponderance of intermediate results from different stages of the pipeline that the user must manage is a large burden that can easily lead to mistakes due to human error.

Our goal is to design an extensible, easy-to-use software framework that is liberated from these issues as much as possible by offering a hybrid solution. During its development, our framework has been tested and used by microbial community researchers studying the microbiota associated with various diseases such as bacterial vaginosis and necrotizing enterocolitis. Researchers using the framework were empowered to analyze their own samples, test hypotheses, and produce publication quality figures in order to communicate their results.

2.2. *Technical Features*

The framework is developed on the Pardus Linux distribution using the Python programming language and open source scientific computing tools and libraries such as SciPy (http://scipy.org) and matplotlib (http://matplotlib.sourceforge.net/). A reliance on open source development tools and libraries will allow us to easily extend the framework and make it portable to non-Linux-based environments.

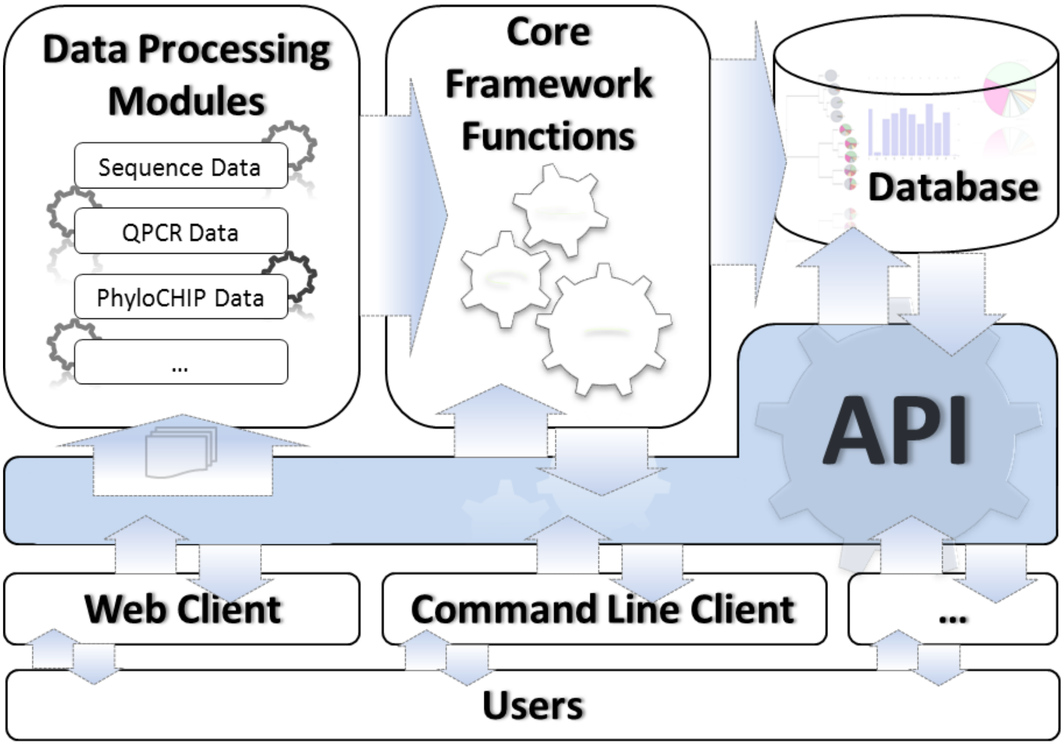


Figure 1: Architectural overview of the framework.

Figure 1 shows an architectural overview of the framework with two major components: A multi-threaded server application that runs in the background performing data processing and core framework functions and interfaces for users to interact with the server.

2.2.1. *Server*

The server performs all manner of computational tasks and figure generation. The multi-threaded design of the server allows it to run multiple analyses concurrently and handle queries simultaneously. The server exposes its functions via an application programming interface (API). This makes it possible for different types of clients to be written and interact with the server seamlessly (Figure 1). This flexibility also allows our framework to be used in both the graphical, user-friendly manner or invoked by scripts for automated analysis of large numbers of data sets.

The server has more than one data processing module, and a set of core functions that is separated from the data. This modularity allows us to extend the server's core functions and analysis capabilities to different types of inputs, such as quantitative PCR data.

2.2.2. *Client*

Any client that can communicate via UNIX domain socket or TCP/IP protocols can query and submit tasks to the server through the API. The default client of the framework is a set of Django (http://www.djangoproject.com) powered web interfaces. The web client allows users to connect to and use the framework via their web browser. Thus, users can interact with the default client of the framework using any operating system and Internet browser they choose.

2.3. *Implementation Status and Limitations*

The framework is still under development and currently the server analyzes 16S rRNA sequences only using RDP's naïve Bayesian classifier [10] and performs all analyses based on genus level taxonomy assigned by RDP. However, the modular nature of the framework allows us to extend its capabilities easily and we are currently working on implementing other data processing modules for phylogenetic analysis based solely on sequence similarity.

Currently the server is being used in house by several biological researchers for a number of active research projects. The most demanding project that has been analyzed on the framework included 166 samples with more than 2 million sequences. The framework server is installed on a Linux server as we work towards our first stable release. We are working to port the server to other platforms such as Mac OSX in order to distribute it more widely upon release.

It is also important to note that the classification of sequence data (currently performed via the RDP classifier) is independent and orthogonal to the downstream analysis and visualization tools. In fact, any data set that contains names and associated abundance values can be slipped into the framework and processed through the downstream analysis and visualization. As a concrete example of this, we have implemented a facility for quantitative PCR data to be loaded into the framework and analyzed in a similar manner to classified 16S rRNA sequencing data. We intend to extend this facility to microarray data as well.

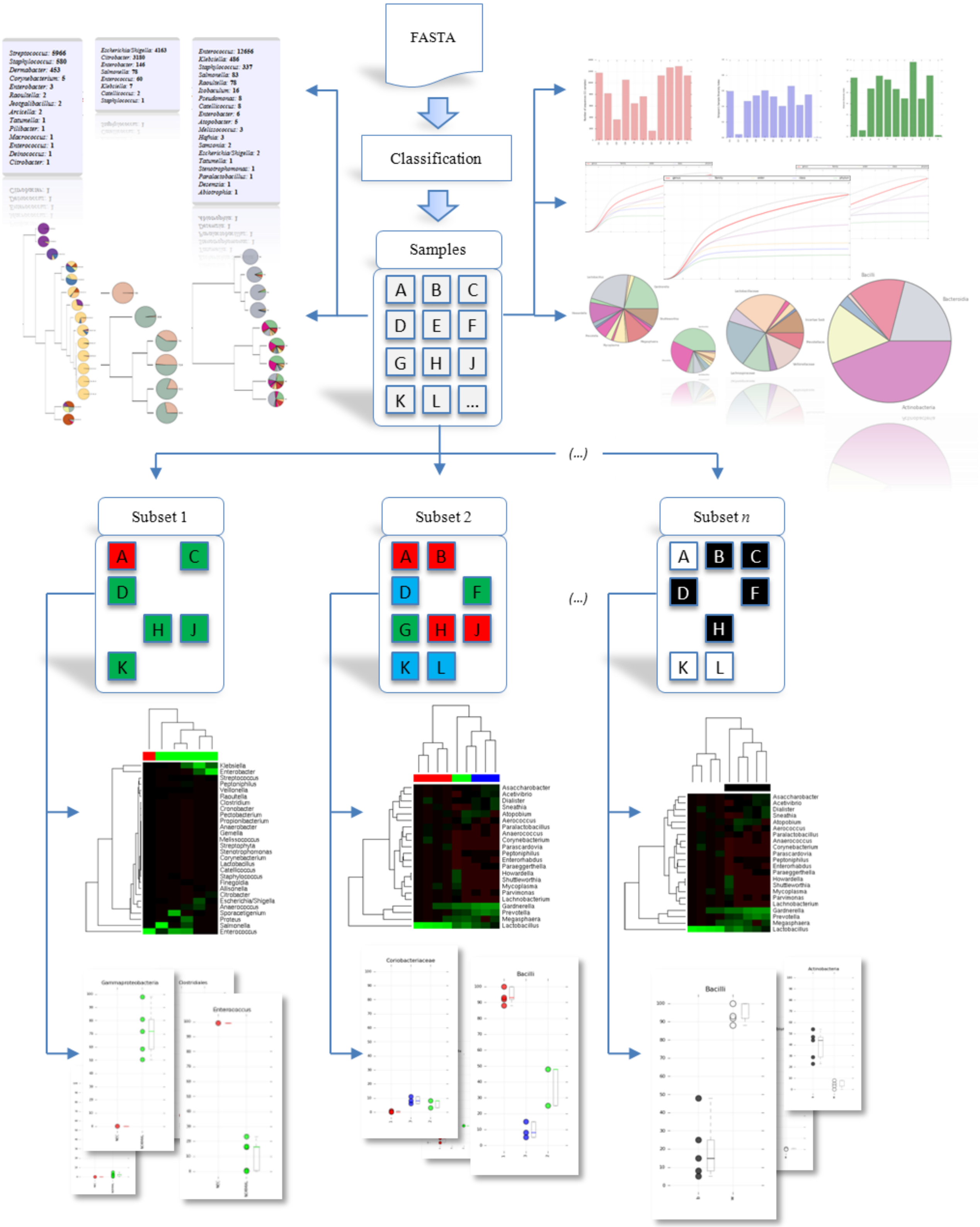


Figure 2: Basic workflow of the framework. Analysis begins with the submission of a FASTA formatted 16S rRNA sequence file.

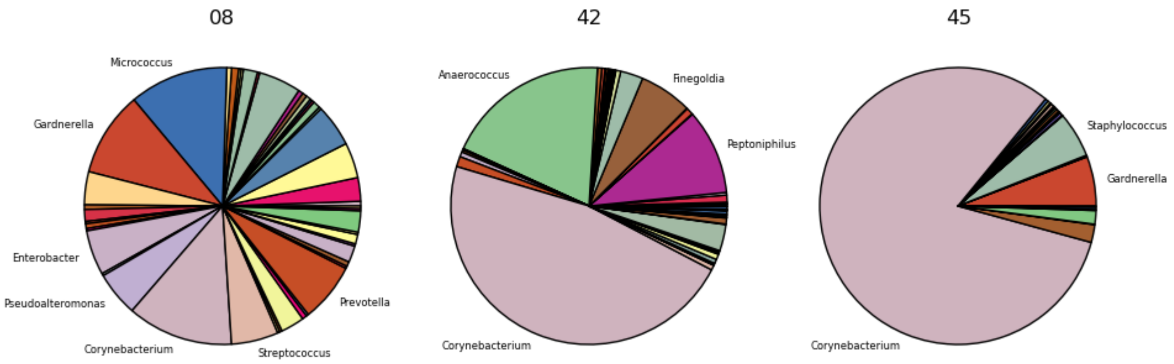


Figure 3: Example pie chart figures show bacterial composition at the genus level of three random samples from a bacterial vaginosis study analyzed using the framework.

3. Workflow

Ease of use and extensibility are key design concerns for the framework. Hence, most of the analysis tasks are performed without requiring any *a priori* knowledge to be provided by the researcher. The basic workflow of the framework is illustrated in Figure 2. Readers are encouraged to visit http://meren.org/framework/ to view an example analysis performed with the framework.

An analysis begins by submitting a FASTA formatted file containing 16S rRNA gene sequences. The file can contain multiple FASTA files originating from multiple environmental or clinical specimens. The framework then employs RDP's naïve Bayesian classifier [10] for rapid assignment of sequences to the taxonomic groups at the phylum, class, order, family and genus levels and the framework proceeds to perform unsupervised preliminary analyses on the samples acquired from the RDP classifier results. These analyses include:

- Calculations of total and percent abundance of bacteria in every sample,
- Bar chart representation of the number of sequences acquired for each sample,
- Bar chart representation of Shannon and Simpson's diversity indices,
- Pie chart representations of samples based on their bacterial compositions at each taxonomic level ranging from phylum to genus (Figure 3),
- Rarefaction curves to illustrate the degree of diversity covered by each sample (Figure 4),
- Hierarchical clustering dendrograms that illustrate how samples clustered based on their bacterial composition at different taxonomic levels (Figure 7).

Once this set of unsupervised alpha-diversity analyses is completed, researchers can assign keys to desired samples and create subsets of samples for further investigation. The user defines subsets by assigning samples to groups, and then assigns a color to each of those groups for visualization. There is no limit on the number of subsets the user may define. The framework automatically ignores samples that are present in the original library if they are not assigned into any groups in a defined subset.

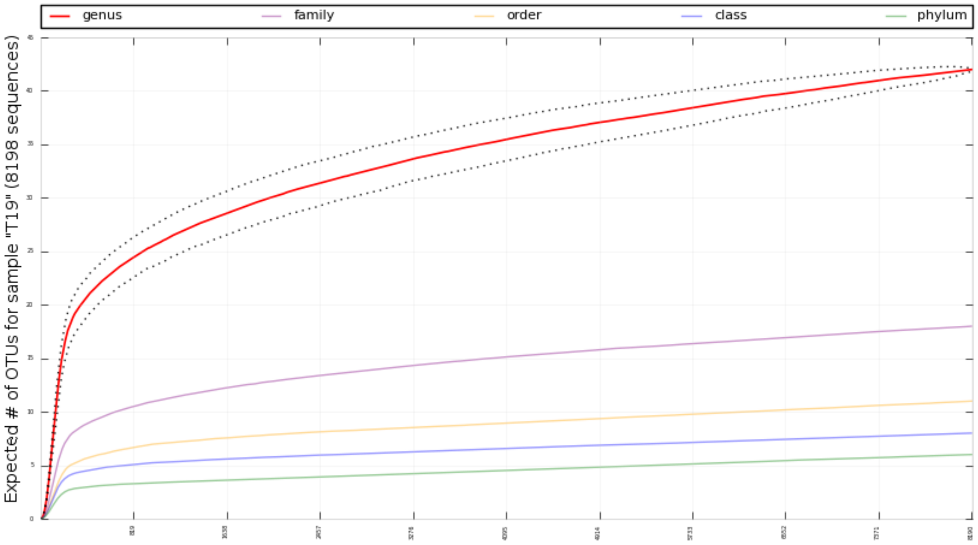


Figure 4: In this example set of rarefaction curves, species richness and expected number of OTUs are shown at different taxonomic levels of a sample that was analyzed using the framework.

When the newly defined subset of samples is submitted for analysis, dot plots of every operational taxonomic unit (OTU) at each taxonomic level ranging from phylum to genus are generated. Box plots are attached alongside the dot plots to illustrate the abundance of each individual OTU across subsets of samples (Figure 5). Complete linkage clustering analysis is performed to assess similarities between microbial communities based on the percent abundance of the taxa they contain. These clustering results are displayed as dendrograms along with heatmaps illustrating the abundance of taxa in each sample (Figure 6). Heatmaps can be refined further to eliminate very low abundance OTUs or to use logarithmic values.

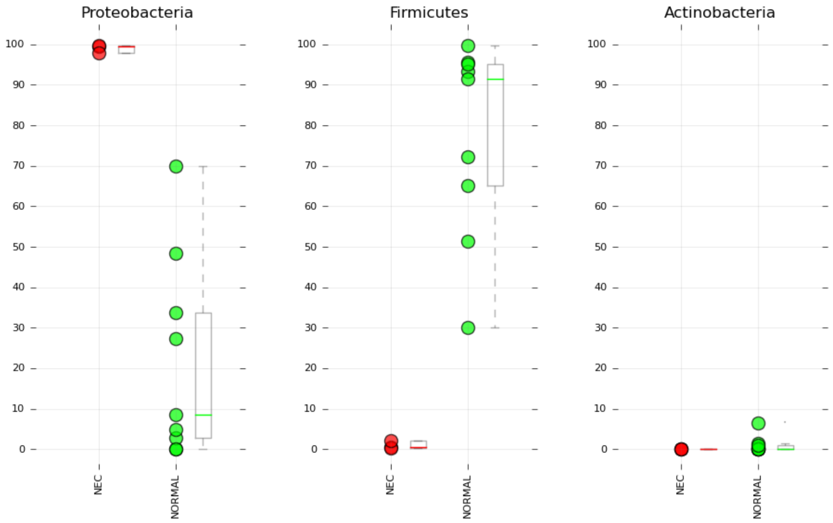


Figure 5: Example dot plot of a subset of samples assigned to two categories, NEC (green) or NORMAL (red) showing differences in the percent abundance of three different OTUs at the phylum level.

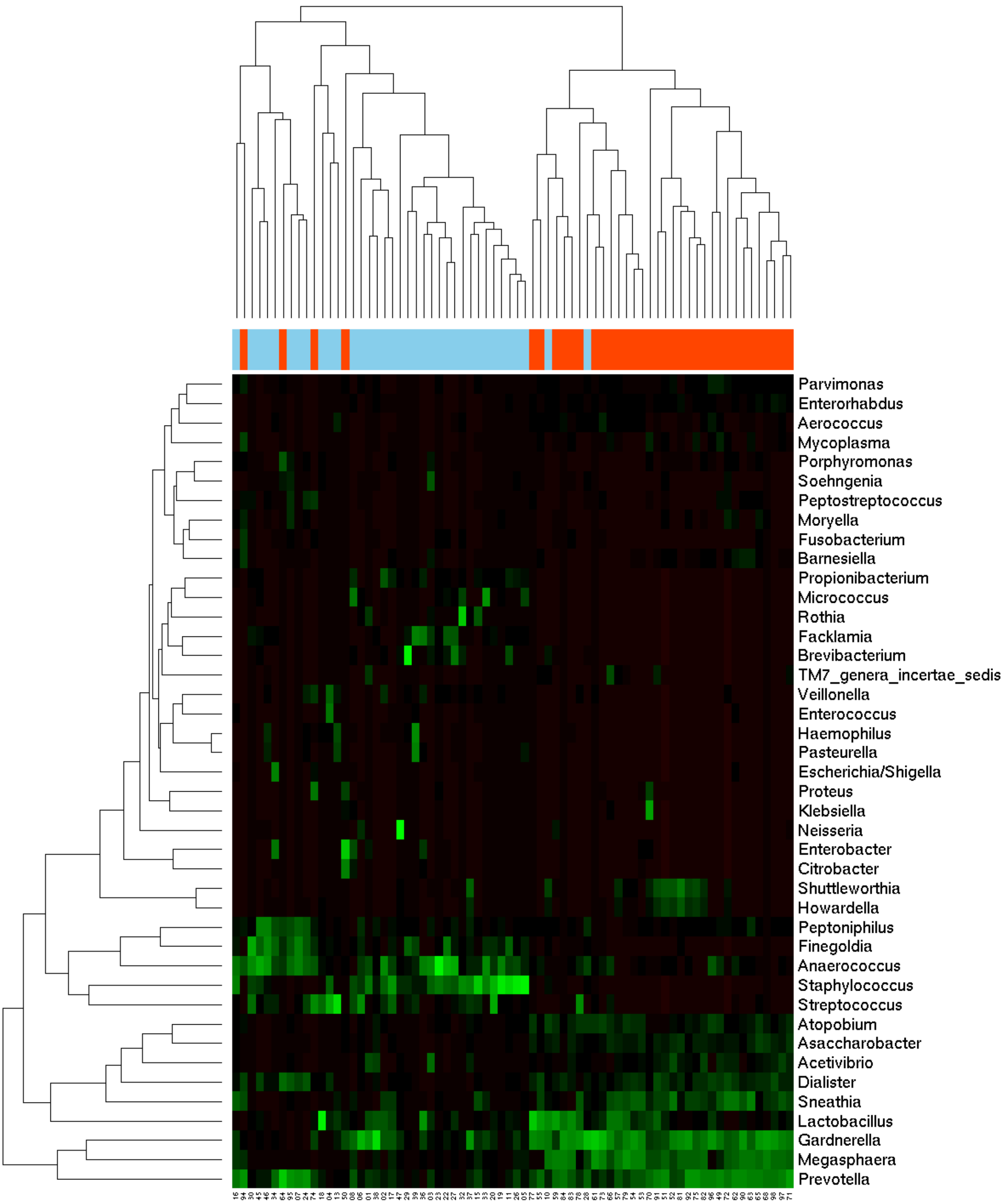


Figure 6: Example heatmap generated by the framework showing how a subset of samples clustered based on their microbial flora at the genus level. Within this particular subset of samples, the cyan color represents penile skin swab samples collected from male patients and the red color represents vaginal swab samples gathered from female patients. The vaginal swab samples largely cluster together on the left of the heatmap, while the penile skin swab samples cluster together on the right side of the heatmap.

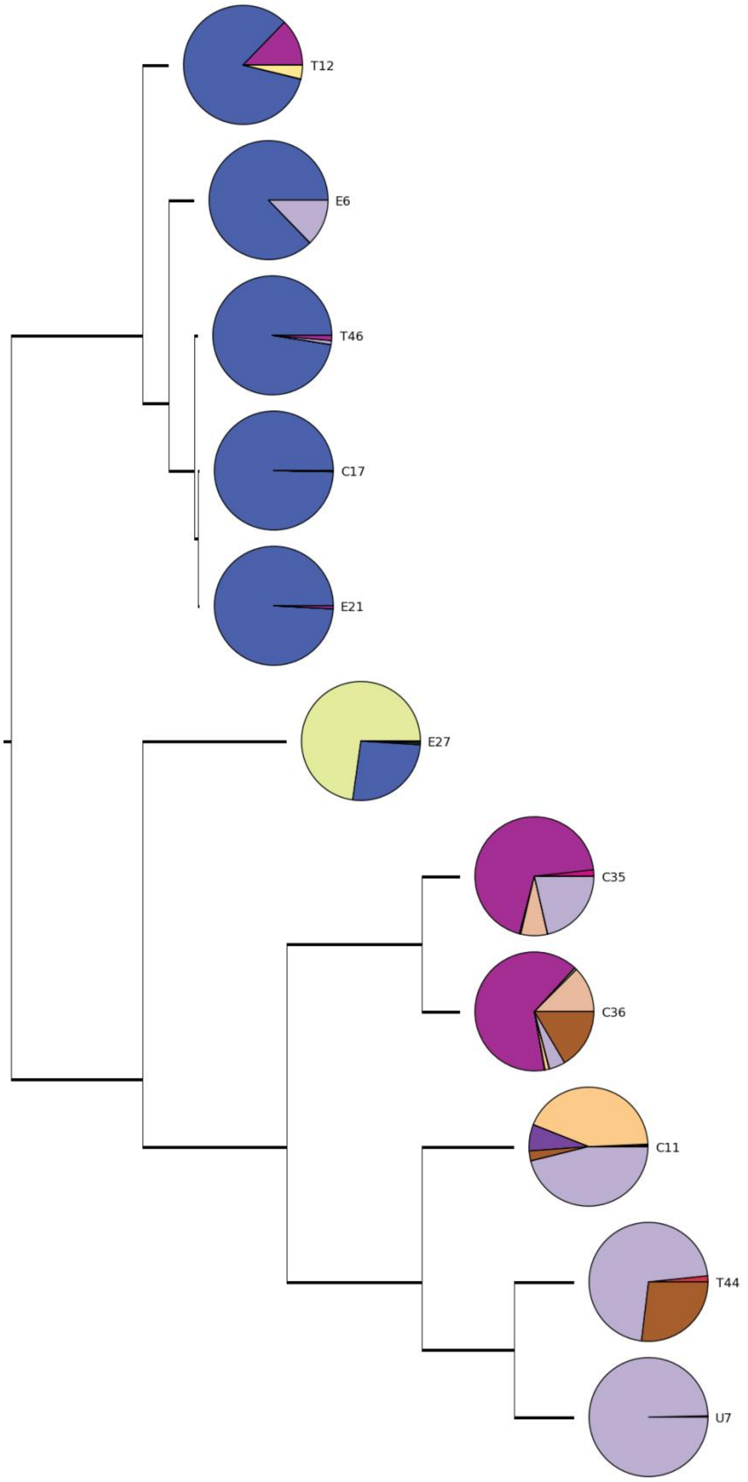


Figure 7: Example dendrogram generated by the framework showing how samples from a necrotizing enterocolitis study were clustered based on their microbial composition at the family level. Smaller versions of the pie chart representations of samples attached to the tree provide additional visual evidence for clustering results.

4. Discussion and Future Work

Many existing tools for metagenomic sequence analysis force biologists to learn application specific details to run various tests on their 16S rRNA sequence data. This burden may cause researchers to eschew new methods and tools for analysis in favor of those that they have already worked to become familiar with. A framework that provides ease of use and seamless integration of new methods as they appear will encourage researchers to try new methods.

We plan to enhance the framework with a variety of additional components in the future. We are currently working to add phylogeny based beta diversity analysis methods, such as UniFrac [11]. It is also worth noting that the longer read lengths being produced by the Illumina Genome Analyzer IIe have made deep sequencing of entire metagenomes feasible. This will allow researchers to go beyond simple classification based on 16S rRNA and on to analysis of complete metagenomes. Our framework provides the infrastructure for further development of features to address assembly, classification and processing of broader metagenomic sequencing data while maintaining the ease of use through web-based client interfaces.

Finally, our framework provides an important separation between classification of metagenomic sequencing data, and analysis and visualization of the classified data. We have currently implemented a front-end that uses the RDP classifier to interpret pyrosequencing reads of 16S rRNA into their taxonomic categories. We intend to enhance the utility of the framework by developing other front-end classifiers that may use the NCBI taxonomy or perform classification based solely on edit distance of sequences to further explore intra-genus and intra-species diversity. We are also working on a facility to utilize the analysis and visualization features of the framework on other data types such as quantitative PCR and microarray data, which can be slipped into the framework past the classification front-end.

5. Conclusion

Although there are a variety of tools currently available for metagenomic sequence analysis, they impose unnatural paradigms or restrictive limitations on biological researchers who may have only rudimentary computer skills. Pipeline approaches force users to select analyses to perform on their data instead of performing a comprehensive analysis by default. They also place a burden on the user for maintenance and routing of intermediate results that can lead to errors. Web based applications have advantages in terms of ease of use, but can be restrictive in the quantity of data they allow to be analyzed and the amount of user interaction required to perform an analysis. None of these approaches, by themselves, provide a viable alternative for microbial community researchers to analyze their data without scaling a significant learning curve.

Our framework provides a scalable, hybrid approach to the problem of metagenomic sequence analysis. Researchers can run the framework on their own computational resources and are not faced with limitations on the quantity of sequences or number of analyses they can perform. They are also able to use familiar web-based interfaces to access the server and do not need to shepherd analyses through a pipeline and manage intermediate results. All of the standard analysis methods are run at the push of a button and the user is presented with an intuitive interface to group samples for further directed analyses. This flexibility and ease of use has allowed microbial community researchers to perform their own analyses and generate publication quality figures to communicate their results with relative ease.

Acknowledgements

This work is supported by funding from the Research Institute for Children in New Orleans.

We thank Marcela Zozaya-Hinchliffe, Johana Norori, Lynn Tran, and Duna Penn for sample sequences used in these studies and helpful suggestions throughout the development of this work.

References

1. D. A. Sandoval, R. J. Steeley, *Science* **328**(5975):179-80 (2010).
2. Y. Wang, J. D. Hoenig, K. J. Malin, S. Qamar, E. O. Petrof, J. Sun, *et al*, *ISME J.*, doi:**10.1038**/ismej.2009.37 (2009).
3. W. B. Withman, D. C. Coleman, W. J. Wiebe, *Proc. Natl. Acad. Sci. USA* **95**, 6578–6583 (1998).
4. NIH HMP Working Group, J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, *et al*, *Genome Res*. **19**(12):2317-23 (2009).
5. G. J. Olsen, C. R. Woese, *FASEB J.*, **7**:113–123 (1993).
6. C. Shyu, T. Soule, S. J. Bent, J. A. Foster, L. J. Forney, *Microb. Ecol.* **53f4**, 562 (2007).
7. J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, *et al*, *Nucleic Acids Res.* **D**, 141 (2009)
8. P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, *et al*, *Appl. Environ. Microbiol.* **75**(23):7537-41 (2009).
9. J. G. Caporaso, J. N. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, *et al*, *Nature Methods* doi:**10.1038** / nmeth.f.303 (2010).
10. Q. Wang, G.M. Garrity, J.M. Tiedje, J.R. Cole, *Appl Environ Microbiol.*, **73**(16):5261-7 (2007).
11. M. Hamady, C. Lozupone, R. Knight, *ISME J.*, doi:**10.1038**/ismej.2009.97 (2009).

VISUALIZATION AND STATISTICAL COMPARISONS OF MICROBIAL COMMUNITIES USING R PACKAGES ON PHYLOCHIP DATA

SUSAN HOLMES*
Statistics Department, Stanford University,
Stanford, CA 94305, USA
**E-mail:susan@stat.stanford.edu*
www-stat.stanford.edu/~susan/

ALEXANDER ALEKSEYENKO
Center for Health Informatics and Bioinformatics,
NYU School of Medicine,
New York, NY USA

ALDEN TIMME
Statistics Department, Stanford University,
Stanford, CA 94305, USA

TYRRELL NELSON
Department of Civil and Environmental Engineering
Stanford University, Clark Center E-250
318 Campus Drive, Stanford CA, 94305, USA

PANKAJ JAY PASRICHA
Division of Gastroenterology and Hepatology
Stanford University Medical Center
Alway Building, Room M211
300 Pasteur Drive,
Stanford, CA 94305, USA

ALFRED SPORMANN
Department of Civil and Environmental Engineering
Stanford University, Clark Center E-250
318 Campus Drive, Stanford CA, 94305, USA

This article explains the statistical and computational methodology used to analyze species abundances collected using the LNBL Phylochip in a study of Irritable Bowel Syndrome (IBS) in rats.

Some tools already available for the analysis of ordinary microarray data are useful in this type of statistical analysis. For instance in correcting for multiple testing we use Family Wise Error rate control and step-down tests (available in the `multtest` package). Once the most significant species are chosen we use the hypergeometric tests familiar for testing GO categories to test specific phyla and families.

We provide examples of normalization, multivariate projections, batch effect detection and integration of phylogenetic covariation, as well as tree equalization and robustification methods.

Keywords: Hypergeometric Test; PhyloChip; projections; Quality Control; R; Phylogenetic Tree

1. Introduction

We present here some examples of using robust multivariate methods for the specific challenges of microbiome studies. We use as a running example a comparative study of microbiological communities in healthy and IBS rats sampled at different locations in the intestine. The results of the biological analysis have been submitted elsewhere,[1] we concentrate here on the statistical and computational challenges involved in such a project.

1.1. *IBS in humans and rats*

It is believed that alterations in the microflora of humans with IBS comes from changes in colonic fermentation patterns as has been described in King et al.[2] Recently, some research groups have been able to use culture-independent methods and deep high throughput 16S ribosomal RNA gene sequencing to demonstrate significant differences in the microbiome of IBS patients.[3,4] The complexity induced by high individual variation of the microbiome suggested that a good starting point in this comparative study would be a rodent model that mimics the human condition. We have as our working hypothesis that the enteric microflora of adult rats with colonic hypersensitivity would differ from that of controls. We use a comprehensive and relatively simple way of studying the microflora using a 16S rRNA gene DNA microarray called the Phylochip.[5] The Phylochip has the advantage over high-throughput sequencing assays in that it is designed to detect presence and abundance of individual species. A major drawback of utilizing the Phylochip platform for this project was that the chip design was not specific to the intestinal microbiome and as a consequence there is a very unequal resolution in certain phyla, representing unequal knowledge about prokaryotic constituents of these phyla.

1.2. *The data and software platform*

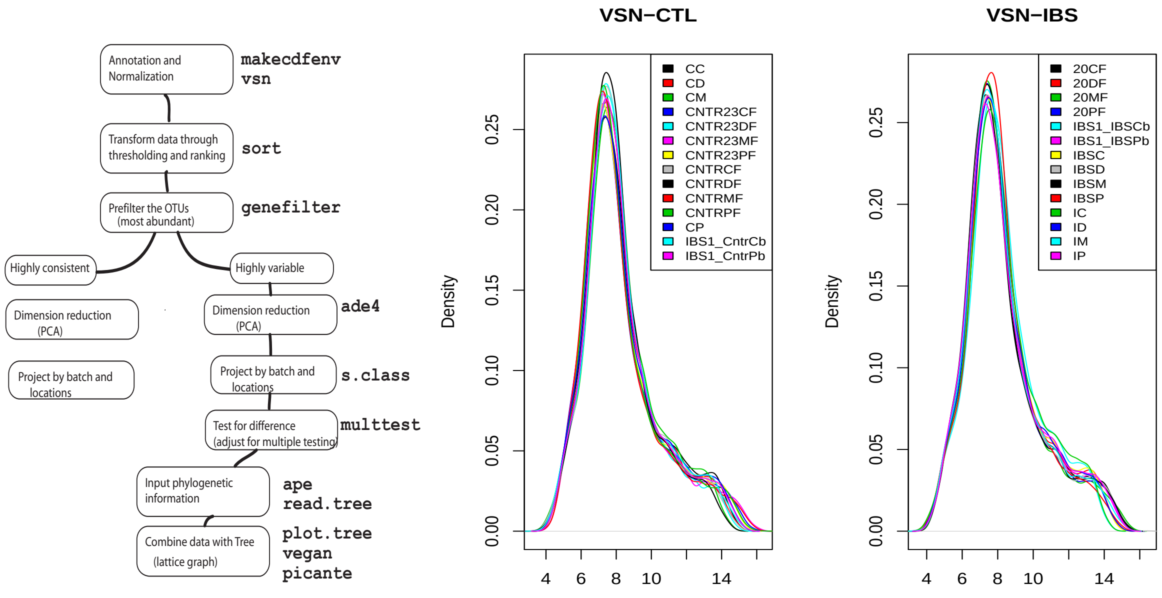
Data were collected on the microbial community of different sections of the large bowel of rats with colonic hypersensitivity induced by neonatal acetic acid irritation. This microarray consists of 500,000 oligonucleotide probes capable of identifying 8743 of bacteria and archaea and provides a comprehensive census for presence and relative abundance of most known prokaryotes in a massive parallel assay. This array uses the the GeneChip (Affymetrix Corporation) technology, thus we could use the Bioconductor[6] suite of tools for annotation[7] and normalization of the data in the same way as is usual for microarray studies.[8] We then used multivariate methods to visualize comparisons between different groupings of the data enabling us to enhance our quality control of the experimental protocol.

We then separated the data into consistently present species and those presenting higher variability. Previous computational approaches include the use of the weighted `unifrac` (Wasserstein distance[9]) between communities.[10] Here we take a geometrical approach to the visualization and detection of various multidimensional biases and changes in variability, as well as the combination of phylogenetic and low rank information. This is more akin to Purdom[11] who also combines phylogenetic and abundance data, but for PCR sequenced phylotypes. Figure 1(a) shows a diagram of the data analysis workflow we chose to follow.

2. Details of the Data Analysis Procedures

2.1. *Prefiltering and Normalization of the Microarray Data*

We created and used a custom-tailored package containing the annotation of all the probes on the Phylochip using the `makecdfenv`[7] package. As with standard expression data, the data need to be preprocessed to ensure that the variance was independent of the level of abundance as described in Durbin et al[12] and implemented in the `vsn` package[8] in the Bioconductor[6] suite of `R`[13] packages. Figure 1 (b) shows the densities of each of the arrays in the two groups after variance stabilizing normalization.



(a) Different stages of Data Analysis

(b) Density after variance stabilizing transformations.

Fig. 1: Tools were transposed from the standard microarray analyses

3. Batch Effect Detection using projections on Principal Planes

A standard principal component analysis was done on the centered and scaled abundance data. In the first set of data, we had originally 24 samples, 12 from IBS, 12 from healthy controls that we wanted to compare, the 12 samples for each group came from 4 locations in the large intestine, however the first apparent differences came from batch groups. We had a first batch of samples corresponding to analyses that were done on day 1 consisted of 6 arrays (3 IBS/3CTL), a second batch 18 arrays (9IBS and 9CTL), done on a second date with a different protocol and array batch. We used the additional ability provided by the projection of supplementary group means and variance as in the function `s.class` in the `ade4`[14] package to explore these batch effects in the laboratory methods used to generate the data. The ellipses are computed using the means, variances and covariance of each group of points on both axes, and are drawn with these parameters: the center of the ellipse is centered on the means, its width and height are given by the variances, and the covariance sets the slope of the main axis of the ellipse. In Figure 2, on the left we see the first two batches although both balanced with regards to IBS and healthy rats were extremely different in variability and overall multivariate location. In order to explore this further, a third batch was generated with the same arrays as batch 2 but the same experimental protocol as batch 1. We see that the third group faithfully overlaps with batch 1 thus showing that the batch effect was not due to a difference in arrays but to the experimental protocol. This shows the utility of PCA in quality control. After finding this particular effect we redid part of the data collection procedure, using only the protocol used in batches 1 and 3, we analyzed 24 samples. We also added 8 samples from mucosal linings, 4 from IBS , 4 for control in each of the 4 intestinal locations. We combined the data into a 32 column matrix of abundance of 8364 species. Since the abundance data were extremely variable and we had seen the sensitivity of the data to varying conditions and protocols we decided to pair the data by location and type. For each pair we had an IBS and a CTL rat, for a sample collected in the location and in the same way, we used the pairing design to minimize the biases from experimental artifacts.

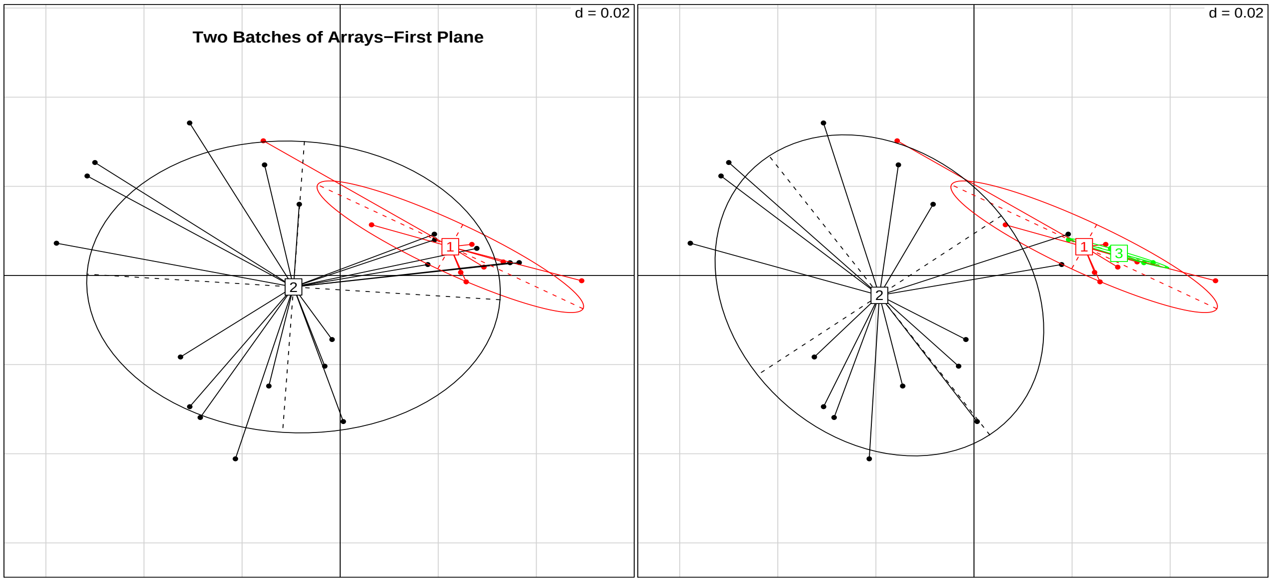


Fig. 2: On the left the first plane of the PCA shows the first set of data with two batches and on the right the third set of arrays was added.

3.1. *Ranking and Thresholding*

In order to deliver a more robust statistical analysis, we ranked the species abundances within each array: the ranks go from 1 (small) to 8364 (large). This is a standard non parametric statistical procedure that enhances the stability of the results because a few outliers cannot bias the analyses. We considered that there were not more than 2000 species present so we set a threshold at 6000 (this is conservative as for instance a recent study in humans places the estimate of numbers of species in the human gut at between 1,000 and 1,200[15]). We thus suppose that all ranks smaller than 6000 were just noise and set them all to be equal to 6000. This avoids finding large differences in ranks for species that are only present at the noise level. We restrict the first part of our analysis here to the species that appeared present in almost all 32 arrays, ie those that had a ranking larger than 6000 in all but one of the arrays. We can see the distribution patterns with varying thresholds from 5000 to 8400 in Table 1. As in microarray studies, it is important to prefilter the species so that only those yielding consistent

#Arrays	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
# > 5000	3997	241	144	91	64	60	55	43	43	37	45	46	41	28	28	38	23
# > 6000	5180	207	136	71	62	48	32	39	38	31	34	25	25	24	24	24	22
# > 7000	6492	120	82	40	32	27	31	13	33	22	22	18	19	12	13	20	17
# > 8000	7737	70	35	25	9	9	10	12	13	12	15	14	9	11	9	11	7
# > 8400	8235	36	24	21	14	6	6	11	10	5	8	5	5	5	6	2	6

#Arrays	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
# > 5000	26	26	38	42	41	33	40	47	56	54	47	59	80	88	167	2766
# > 6000	24	20	22	20	26	28	46	43	41	41	45	40	46	72	109	1989
# > 7000	14	18	18	25	18	20	26	22	18	21	19	26	30	59	83	1204
# > 8000	11	7	11	11	11	10	9	16	9	12	13	12	18	23	38	415
# > 8400	7	2	8	4	5	7	6	13	10	5	5	5	6	16	18	112

Table 1: Tables showing the number of species present at a given level of abundance as measured by ranks in 0,1,2,. . . ,32 arrays. We can see in particular that there are about 2,000 species present at least at the rank 6000 in all 32 arrays and about 415 which are highly abundant (> 8000) in all arrays.

signals enter the analysis. In particular, this is important for various testing procedures we will use later (testing for differences between IBS and CTL), where having extra non-meaningful species costs us extra power requiring us to perform more tests than necessary. Table 1 is the basis of most of the prefiltering presented in the paper.

4. Incorporating and adjusting the phylogenetic information

4.1. *Difficulty with the Original Tree: heterogeneous levels of resolution*

We entered the complete phylogeny of 16sRNA provided by GreenGenes into the **R**[13] package ape.[16] We can see in the left tree of Figure 3 that the phylogenetic tree of all the bacteria tested for on the microarrays is not ultra-metric. That is, not every species is at the same distance from the root. When looking at the phylogenetic tree (Figure 3), it is evident that some areas of the tree have much greater resolution than others. The problem with this is that some species of bacteria are probed multiple times by the array. Therefore, they have more chances than other bacteria of showing significance under the null hypothesis. For example, of the 158 bacteria found to be significantly over- or under-abundant in IBS rats at the $\alpha = 0.05$ level in the first dataset (excluding the mucosal samples), nine are C. leptum, ten are R. hansenii, and ten are P. ruminicola. One of the questions that must be answered is whether or not higher resolution in certain areas of the phylogenetic tree caused these species to be over-represented among the bacteria of interest. We will see below that the hypergeometric test provides a way to control the phylogenetic bias at the higher-order level, but there is a lot of information lost when we look only at phyla. In an attempt to conserve information while correcting for this oversampling in certain regions, we also propose a method for collapsing the tree by merging the tips of related species with similar microarray intensities.

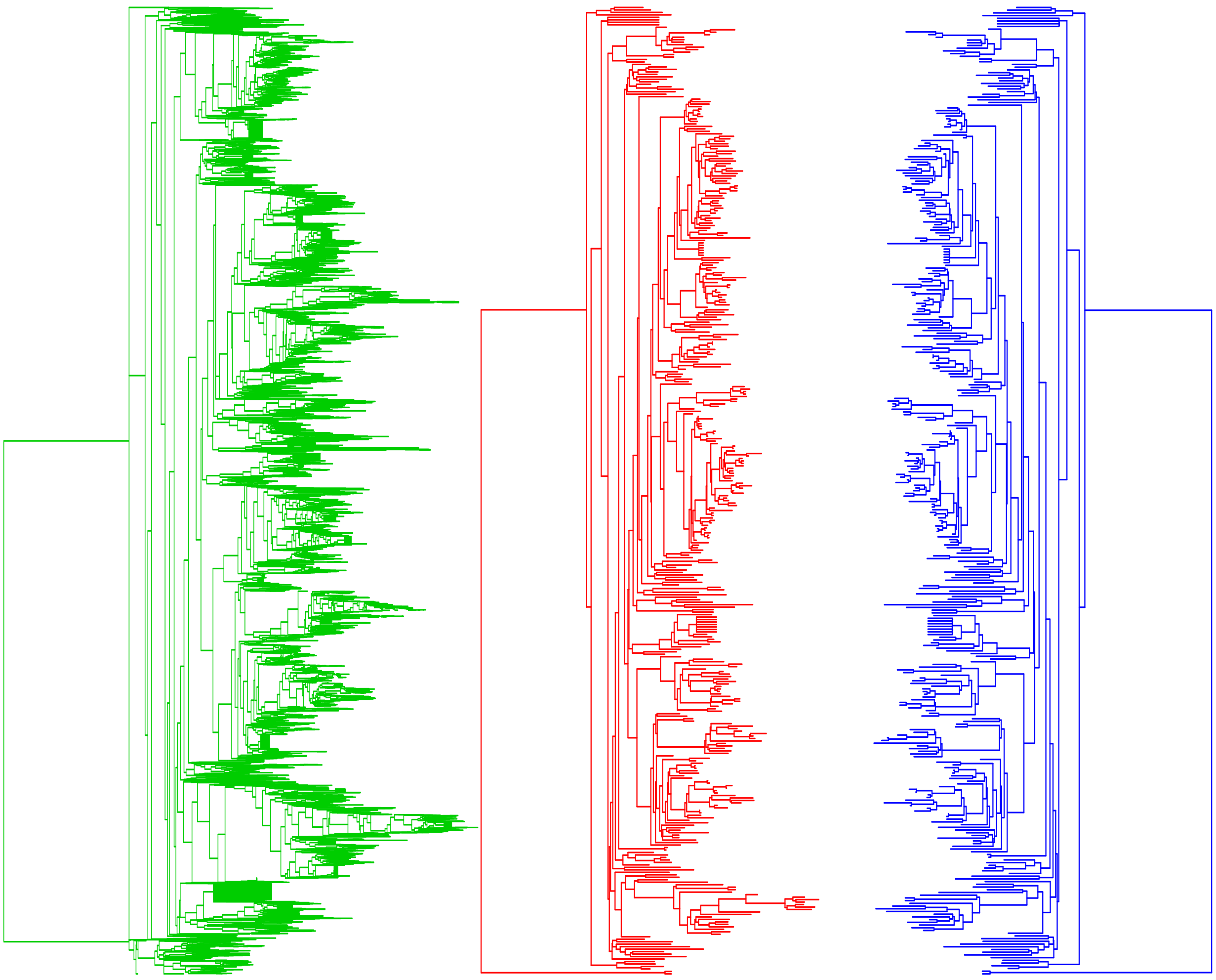


Fig. 3: On the left, we have the tree of all operational taxonomic units (otus) present on the Phylochip, we can observe that the distance to the root of many of the otus is variable, thus indicating a heterogeneous degree of resolution. The two trees on the right are filtered trees representing only the 400 most abundant species. The blue tree on the right was computed by using the collapsing algorithm presented in this section, we see that the long right clade at the bottom of the middle tree has disappeared.

The idea is to control for over-resolution by merging tips of the clades that are more resolved, creating a more level playing field for the multiple testing. We used the length from the root of the tree as the main parameter for collapsing tips. That is, for any two species further from the root than the given maximum distance, we try to merge the two tips. However, merging is only done if the microarray data from the two species are similar enough to be merged. Tips are only merged if there is a low enough variance across the bacteria for each microarray measurement. What is a low enough variance, however, is difficult to define. For the purposes of the analysis here, we used a bootstrap procedure[17] that estimated the $q = 0.9$-quantile for a random collection of groups of size n bacteria. This served as the cutoff of what could be considered a small enough variability within that clade. A collection of n

bacteria is merged only when all their tips are farther than the maximum length from root and $p = 80\%$ of the 32 variances across the collection (one for each microarray) are below the computed thresholds. These were arbitrary thresholds that we have only evaluated empirically by running the algorithm with varying values for n, q and p.

4.2. *Consistently Abundant Species and their place on the Tree*

Here we chose about the top 100 most consistently abundant species following the choice of a threshold of about 8400 as in Table 1. Here we show how we can use the enhanced plotting facilities in R through the Lattice compatible packages, we can plot the complete tree, identifying the part of the tree which is covered by a subset of species.

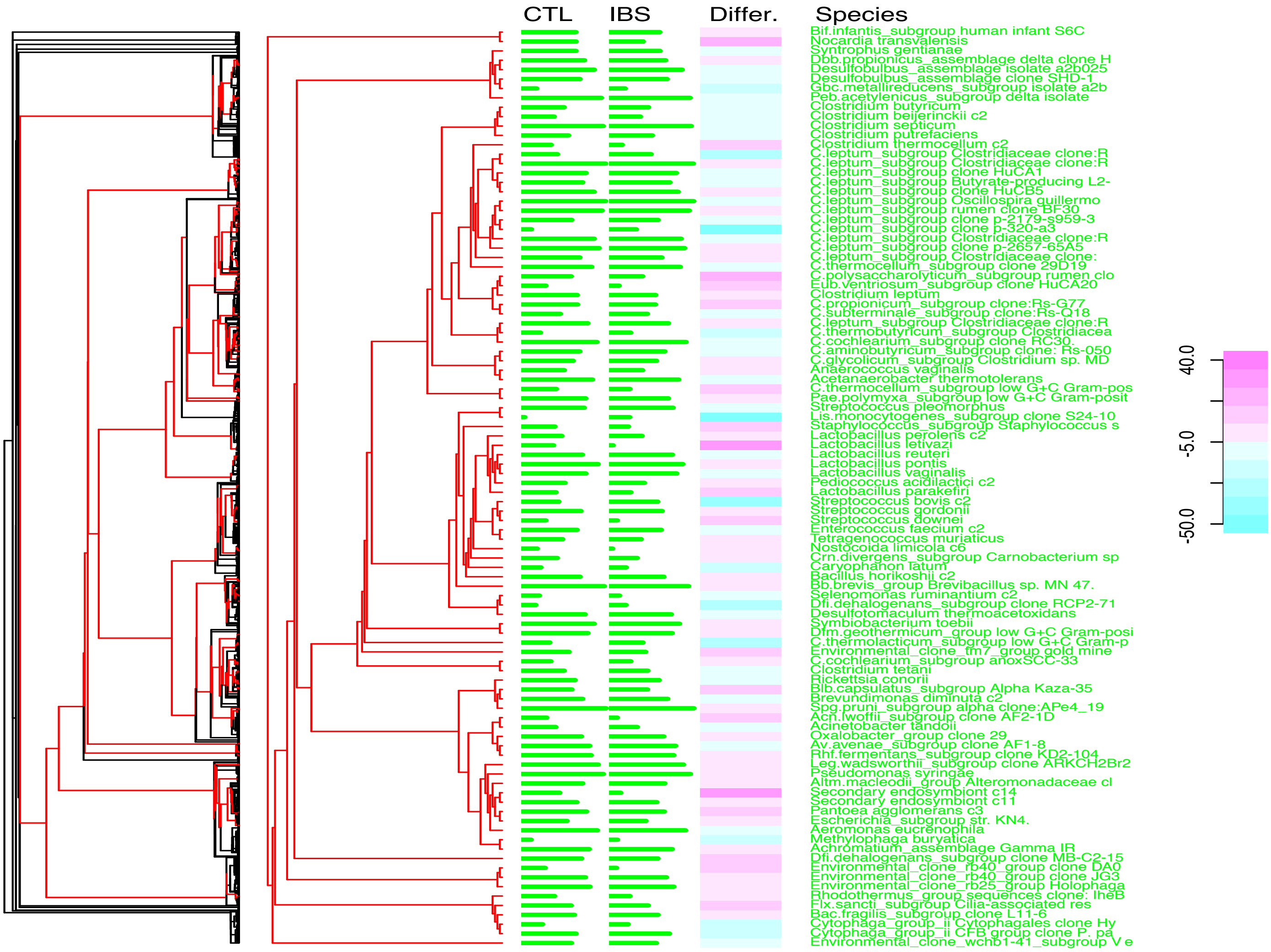


Fig. 4: The left tree shows the complete tree on all species in black with the subtree of most abundant species in red, this subtree is the one plotted on the next panel. Values of abundance in CTL and IBS rats are plotted in the next two columns, the pink/blue scaled variables are the truncated rank differences between the two groups.

4.3. *Highly variable species*

We concentrate now on the species which are abundant enough to be considered consistently present (more than 15 out of 32 arrays over 7800) but that also show high variability (standard deviation above 150). These values were arbitrarily chosen to retain about 100 species. There were actually 99 such species for which we had complete annotation information. We then took the results of the PCA analysis and combined them with the tree information by using the loadings on the first two components (which account for 55% variance) and plotted them alongside the phylogenetic sub tree of the species we had retained as most variable. This plot is much easier to read than the projections of long species names in the two dimensional principal plane. We have colored in red the species that are more abundant in the mucosal samples.

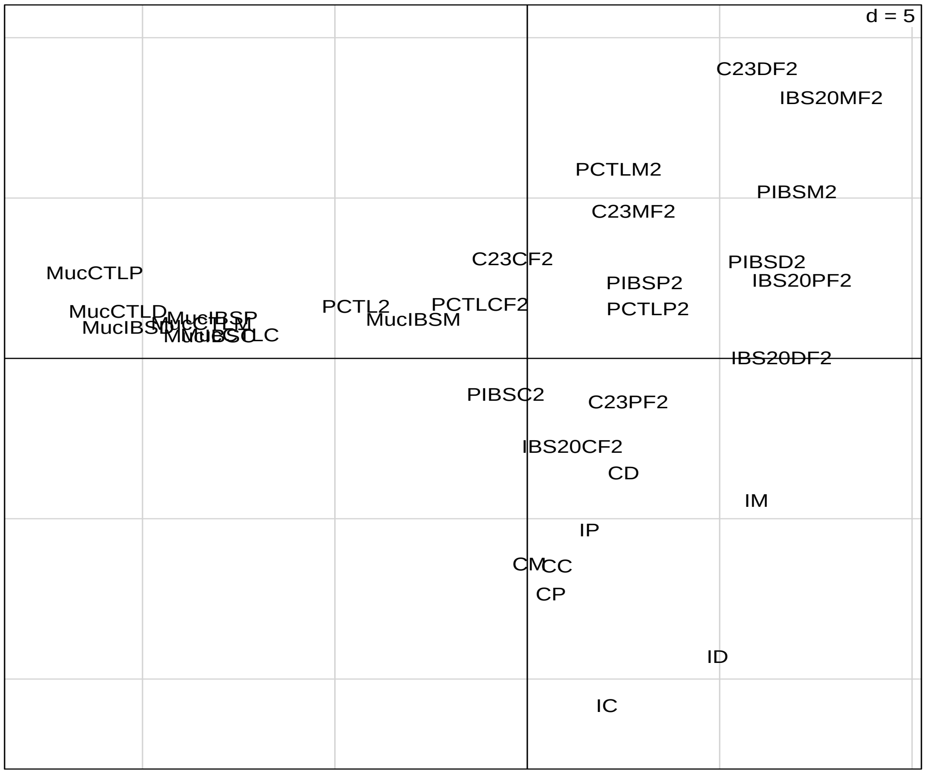


Fig. 5: Principal component analysis of top most abundant and variable species, we see the Mucosal location is the explanation for the first component, all the mucosal samples have negative loadings on this factor.

4.4. *Multiple Testing for finding differentially expressed species*

The first set of analyses showed that the main differences were batch effects and differences between the mucosal and other samples, so we decided to proceed by pairing the data by location, batch and mucosal types, thus removing the extra variance due to these factors. Thus we proceed into the testing phase using a paired design and we will use corrections made on the paired t-test rather than the ordinary one. We will use truncated paired differences

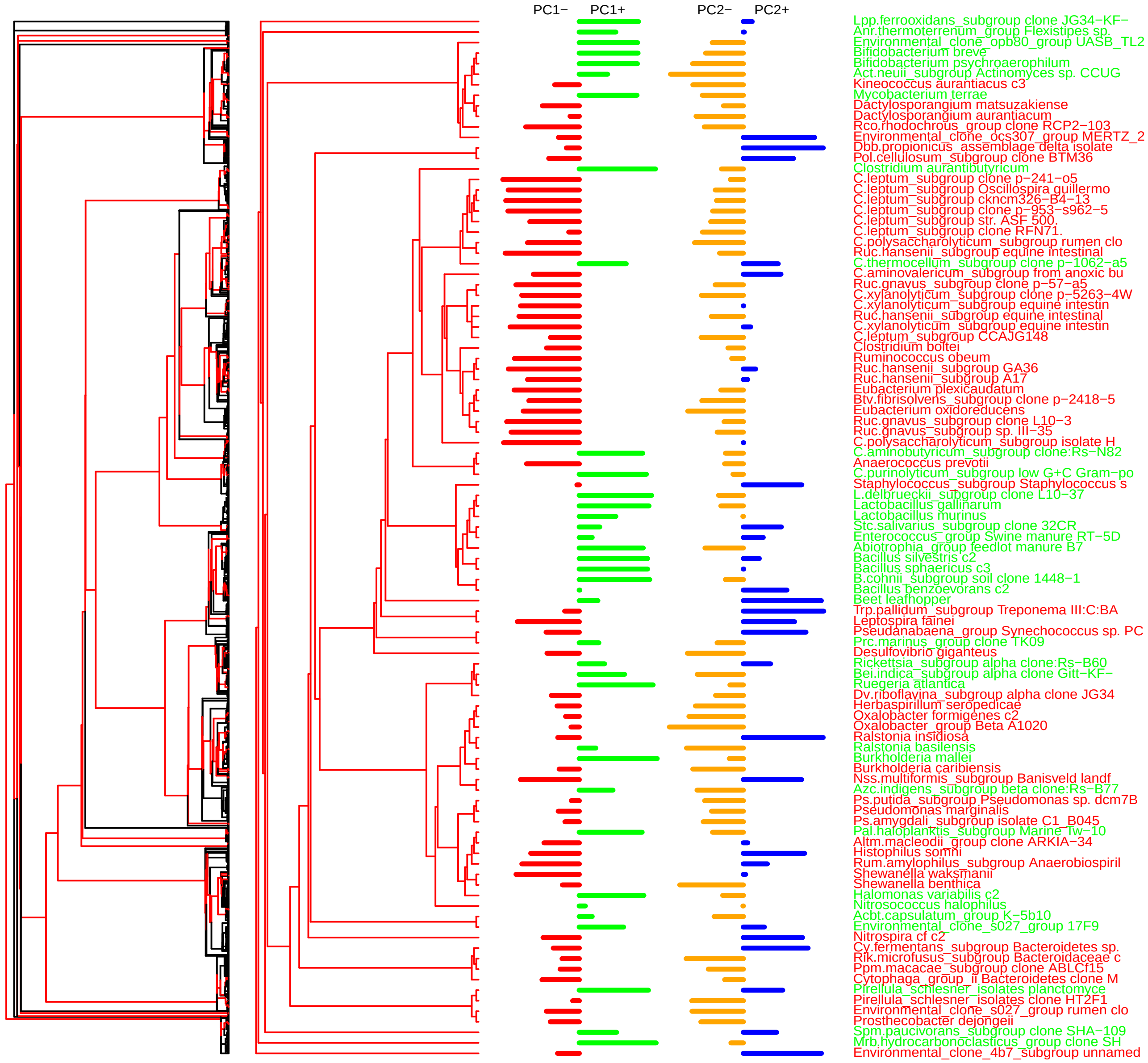


Fig. 6: Complete tree with the subtree of most variable among the consistently abundant species and the loadings on the first two principal components.

in ranks as input to standard multiple testing programs for finding the adjusted p-values. To control for false discovery due to multiple testing, p-values were adjusted according to the Benjamini-Hochberg procedure, which is able to control for FDR given some assumptions on the expression levels of the bacteria on the microarray. We used the `multtest` package from `Bioconductor`.[6]

4.5. *Significant differences projected onto the Tree*

In order to visualize the parts of the phylogenetic tree most influenced by changes in species abundance between groups we retained the most significantly changed species (up in IBS or up in CTL) on the tree and used the facilities available through the `ape`[18] and the `lattice`[19] packages.

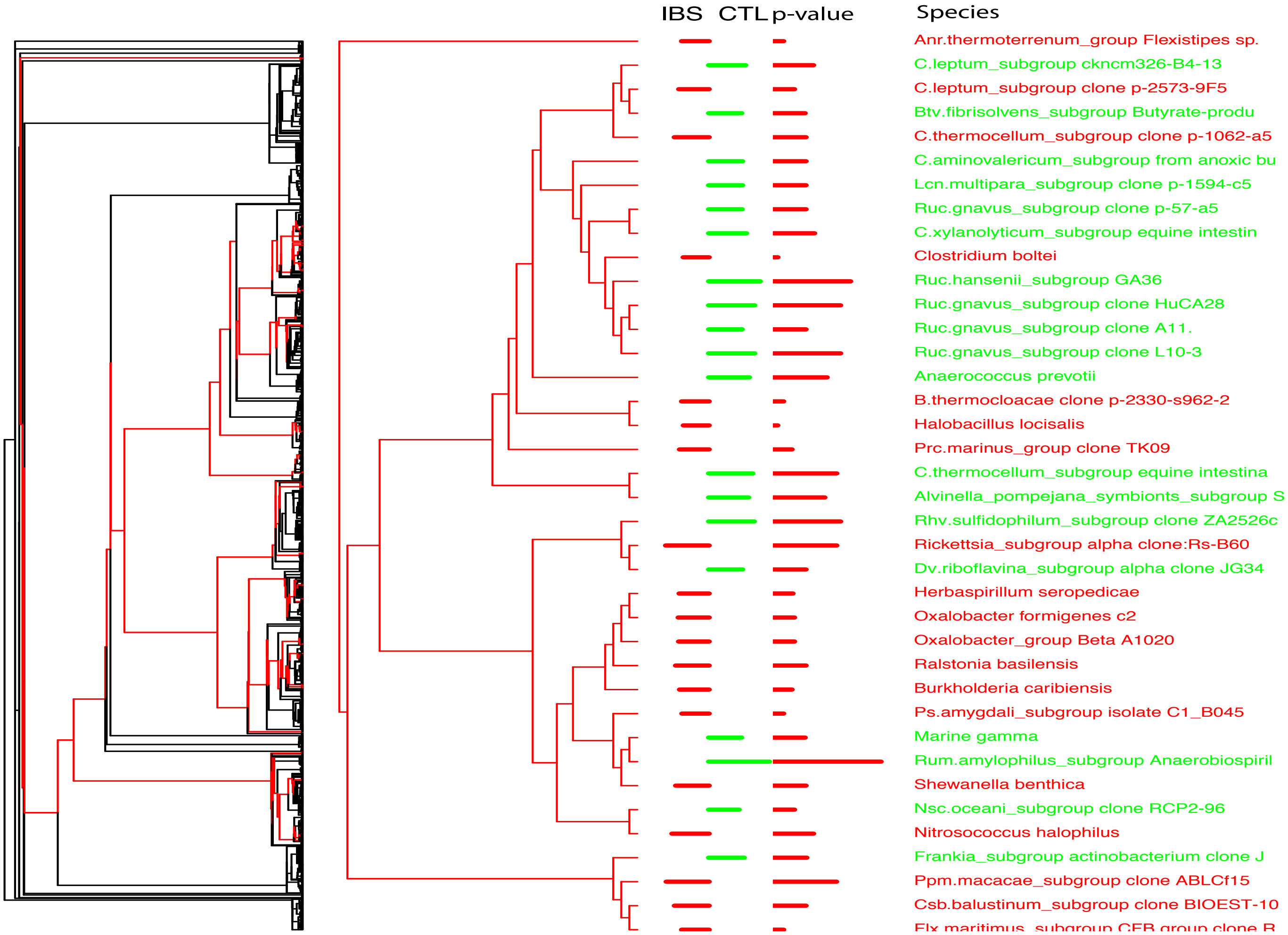


Fig. 7: The left tree shows the complete tree on all species in black with the subtree of set of species that show the most significantly differences between CTL and IBS in red in the second panel. Values of abundance in CTL and IBS rats are plotted in the next two columns, the next column shows the $-\log(pvalue)$, so the largest bars represent the most significantly different species.

4.6. *Category Based Comparisons*

We chose as the list of most significant species those that had adjusted p-values lower than 0.05 in the multiple testing procedure detailed above. We created two lists, one for which the ranked abundances were larger in the IBS, the other for which the ranked abundances were larger in the CTL group. We wanted to find specific families or phyla that are over-represented in either of the lists. This is a similar situation as that of testing significance of Gene Ontology categories for expression studies. We recall that in both situations the relevant test is the hypergeometric and that Fisher's exact test and the hypergeometric test formulation are equivalent.[20] We define the set of prefiltered species (**species universe**) as those that passed the threshold test of being present (> 6000) in at least 31 of the arrays (see Table 1). The chosen species (universe and significant) are then binned by phyla or families, these categories replace the Gene Ontology categories used in microarray studies. We are looking for overrepresentation of certain families or phyla. This method is especially relevant here as the chip does not have equal representation of different families and phyla.

The results and details of the hypergeometric tests can be consulted in Nelson et al, 2010[1] where we conclude in particular that the IBS had significantly more Bacteriodetes and on the other hand there is an overrepresentation of Firmicutes in the healthy controls. At the family level, the results showed that the families of Oxalobacteraceae, Prevotellaceae, Burkholderiaceae, Sphingobacteriaceae were significantly overrepresented in IBS rat. Conversely, the most significantly enriched family in control rats were Lachnospiraceae, including Ruminococcus sp., followed by Erysipelotrichaeceae and Clostridiaceae.

5. Summary

Some methods developed for standard microarray studies can be useful in Phylochip studies, examples shown here include variance stabilization, prefiltering, multiple testing and hypergeometric tests.

Batch effects can be detected through multivariate projections using methods such as PCA complemented with the projections of the relevant means, variance and covariance ellipses on the principal planes. We concluded that the best way to counter batch effects was then to use paired differences between subjects if a comparative design is available.

High between subject variability in bacterial abundances suggests the use of ranks is more effective than the original intensities. This method is known to be robust in the sense that if some of the abundance values are on very different scales, their effect on the overall outcome can be minimized by replacing the original values by the ranks within each array. We have provided an example of such an approach here.

Finally the integration of complex phylogenetic structure is possible through the conjoint use of the many available packages in R for doing phylogenetics and community analysis. We have provided an example of a complex combination of plotting trees and results from PCA.

Acknowledgments

We thank Todd DeSantis and Cindy Wu for enlightening conversations about the Phylochip and all participants in the R and Bioconductor projects for their packages. This work was supported by a grant from the Bio-X Interdisciplinary Initiative Program IIP4-32 to P.J. Pasricha and A. Spormann and grant NIH-5-R01GM086884 to S. Holmes. We thank the reviewers who made useful comments that helped improve the paper.

References

1. T. A. Nelson, S. Holmes, A. V. Alekseyenko, M. Shenoy, T. DeSantis, C. Wu, G. L. Anderson, J. Sonnenburg, P. J. Pasricha and A. Spormann, Phylochip microarray analysis reveals altered gastrointestinal microbial communities in a rat model of colonic hypersensitivity, Submitted.
2. T. King, M. Elia and J. Hunter, *The Lancet* (Jan 1998).
3. E. Malinen, T. Rinttilä, K. Kajander and J. Mättö, *The American journal of Gastroenterology* **100**, 373 (Jan 2005).
4. A. Kassinen, L. Krogius-Kurikka and H. Mäkivuokko, *Gastroenterology* (Jan 2007).
5. K. H. Wilson, W. J. Wilson, J. L. Radosevich, T. Z. DeSantis, V. S. Viswanathan, T. A. Kuczmarski and G. L. Andersen, *Appl. Environ. Microbiol.* **68**, 2535 (2002).
6. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang and J. Zhang, *Genome Biology* **5**, p. R80 (Jan 2004).
7. R. A. Irizarry, L. Gautier, W. Huber and B. Bolstad, makecdfenv_1.18 (2009), `http://cran.r-project.org/doc/packages/makecdfenv.pdf`.
8. W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka and M. Vingron, *Bioinformatics* **18 Suppl 1**, S96 (Jan 2002).
9. S. N. Evans and F. A. Matsen, *arXiv* **q-bio.PE** (Jan 2010).
10. M. Hamady, C. Lozupone and R. Knight, *The ISME Journal* (Jan 2009).
11. E. Purdom, *Annals of Applied Statistics* .
12. B. Durbin, J. Hardin, D. Hawkins and D. Rocke, *Bioinformatics* **19**, 1360 (2003).
13. R. Ihaka and R. Gentleman, *Journal of Computational and Graphical Statistics* **5**, 299 (1996).
14. D. Chessel, A. Dufour and J. Thioulouse, *R News* **4**, 5 (2004).
15. J. Qin, R. Li, J. Raes, M. Arumugam, K. er Solvsten Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. ce Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. ong Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. orben Hansen, D. L. Paslier, A. Linneberg, H. B. Nielsen, E. P. tier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. ang Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. gang Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. D. a nd Francisco Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, M. Consortium, P. Bork, S. D. Ehrlich and J. Wang, *Nature* **464**, 59 (Mar 2010).
16. E. Paradis, Ape (analysis of phylogenetics and evolution) v1.8-2 (2006), `http://cran.r-project.org/doc/packages/ape.pdf`.
17. B. Efron, R. Tibshirani and R. Tibshirani, *An introduction to the bootstrap* (Chapman & Hall/CRC, 1993).
18. E. Paradis, J. Claude and K. Strimmer, *Bioinformatics* **20**, 289 (2004).
19. D. Sarkar, *lattice: Lattice Graphics*, (2009). R package version 0.17-26.
20. I. Rivals, L. Personnaz, L. Taing and M.-C. Potier, *Bioinformatics* **23**, 401 (Feb 2007).
21. S. Kembel, P. Cowan, M. Helmus, W. Cornwell, H. Morlon, D. Ackerly, S. Blomberg and C. Webb, *Bioinformatics* **26**, 1463 (2010).
22. K. S. Pollard, H. N. Gilbert, Y. Ge, S. Taylor and S. Dudoit, *multtest: Resampling-based multiple hypothesis testing*, (2010). R package version 2.4.0.
23. J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, R. G. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens and H. Wagner, *vegan: Community Ecology Package*, (2010). R package version 1.17-0.

HUMAN MICROBIOME VISUALIZATION USING 3D TECHNOLOGY*

JASON H. MOORE

Institute for Quantitative Biomedical Sciences, Departments of Genetics and Community and Family Medicine,
Dartmouth Medical School, Lebanon, NH 03756
Email: jason.h.moore@dartmouth.edu

RICHARD COWPER SAL.LARI

Institute for Quantitative Biomedical Sciences, Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756
Email: richard.cowper.sal.lari@Dartmouth.edu

DOUGLAS HILL

Institute for Quantitative Biomedical Sciences, Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756
Email: douglas.hill@Dartmouth.edu

PATRICIA L. HIBBERD

Department of Pediatrics, Division of Global Health, Massachusetts General Hospital, Harvard Medical School,
Boston, MA 02114
Email: patricia.hibberd@gmail.com

JULIETTE C. MADAN

Department of Pediatrics, Division of Neonatology, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756
Email: juliette.c.madan@hitchcock.org

High-throughput sequencing technology has opened the door to the study of the human microbiome and its relationship with health and disease. This is both an opportunity and a significant biocomputing challenge. We present here a 3D visualization methodology and freely-available software package for facilitating the exploration and analysis of high-dimensional human microbiome data. Our visualization approach harnesses the power of commercial video game development engines to provide an interactive medium in the form of a 3D heat map for exploration of microbial species and their relative abundance in different patients. The advantage of this approach is that the third dimension provides additional layers of information that cannot be visualized using a traditional 2D heat map. We demonstrate the usefulness of this visualization approach using microbiome data collected from a sample of premature babies with and without sepsis.

1. Introduction

1.1. *The Human Microbiome*

The primary goal of the human microbiome project is to understand the role that symbiotic microorganisms play in determining health and disease [1,2]. This is a staged effort that includes

* This work is supported by the Hitchcock Foundation Joshua B. Burnett M.D. Research Career Development Award, the Hearst Foundation, the Dartmouth Center for Clinical and Translational Science, the Norris-Cotton Cancer Center, the Institute for Quantitative Biomedical Sciences and NIH grants LM009012 and RR018787.

1) construction of draft assemblies of reference genomes, 2) creation of reference microbiome data sets, 3) determination of the full human microbiome and 4) determination of the global diversity of the human microbiome. Significant progress toward these goals has been made. For example, Wu et al. [3] carried out a phylogenetic analysis of 56 microbes demonstrating the need for a comprehensive encyclopedia of microbial genomes. A recent study reports the results of the initial sequencing of 178 microbial genomes that will help provide a reference for human microbiome studies [4]. Costello et al. [5] assayed the spatial and temporal variation of the human microbiome from up to 27 different sites in seven to nine subjects. This study demonstrated that bacterial flora varied significantly across body sites and time. The rapid advances in the development of high-throughput sequencing technologies will make it feasible to accomplish many of these goals over the next few years.

1.2. *The Fecal Microbiome of Preterm Infants*

The ultimate goal of these baseline genomic studies is to provide the framework for relating microbial diversity and composition to clinical endpoints. One important application of this technology is to determine whether the human microbiome will be useful for predicting outcomes in infants born prematurely. Colonization of the neonatal intestine happens rapidly after birth, is dependent on delivery method [6], but may be delayed in infants born prematurely. Further, premature infants are more likely to be colonized by pathogenic bacteria [7-9]. Our working hypothesis is that the fecal microbiome of preterm infants will be useful for predicting their clinical course and might provide potential time points for intervention to ameliorate disease risk. Previous work in this area has focused, for example, on neonatal necrotizing enterocolitis (NEC) in preterm infants. This is an inflammatory disorder that may lead to death and has an incidence of one to three per 1000 live births. A study by Wang et al. [10] sequenced 16S rRNA from the fecal samples of 20 preterm infants and found that those with NEC had less diversity and a higher abundance of *Gammaproteobacteria*. Although not conclusive, this study provides a baseline for beginning to think about how the microbiome might influence susceptibility to NEC and other clinical endpoints in preterm infants such as sepsis.

1.3. *A Role for Visualization in Human Microbiome Studies*

The biocomputing challenges of microbiome analysis are both diverse and numerous. This is partly due to the volume of sequence data that is generated and the hierarchical complexity of the microbial data itself. Examples of prior biocomputing work in this area include the development of algorithms for identifying human gut-specific protein families [11], reference genome databases [4], computational inference of function using genomic context [12] and efficient taxonomic profiling [13]. These studies and others are providing the computational methodologies that will be necessary to accurately and efficiently analyze human microbiome data.

Despite these advances, there are still many biocomputing needs. For example, a typical data set might include a list of hundreds of bacterial species that are hierarchically organized into different groups, including genus, families, orders classes and phyla. One goal is to relate abundance of bacteria at these different taxonomic levels to clinical endpoints. This is further complicated by information about genes and pathways that are present in each of the bacterial species and how these relate to various clinical endpoints. The genomic information of the host can also be added to the analysis. The ultimate challenge is to put these many different layers of information together in a statistical or machine learning analysis to identify the clinically useful patterns. Although not yet routine, this type of biocomputing analysis will be in high demand in the near future.

The working hypothesis of the present study is that the inherent hierarchical complexity of human microbiome data, and the need to relate these many layers of information to clinical endpoints, will necessitate the development of intuitive user interfaces for visual exploration and analysis. In other words, the Excel-based spreadsheet paradigm will not provide the level of human-computer interaction that is necessary to both understand a complex data set and inform data mining and machine learning analyses. The goals of the present study were to develop a methodology for visualizing multiple dimensions of human microbiome information using 3D technology that is both intuitive and interactive. We present here a three-dimensional (3D) heat map methodology and software that builds on the familiarity and success of the conventional two-dimensional (2D) heat map and the power of commercial video game development engines and 3D technology. The ultimate goal of these studies is to provide comprehensive visual analytics methodology and software for facilitating human microbiome analysis.

2. Methods

2.1. *A 3D Heat Map*

Heat maps have become a popular and useful method for visualizing high-dimensional data (http://en.wikipedia.org/wiki/Heat_map) and were introduced more than fifty years ago by Sneath [14] for biological problems. Eisen et al. [15] popularized the heat map for visualizing the results of clustering genomics data. A heat map consists of a 2D grid or matrix of colored squares where each square represents an observation of a random variable and the color of the square is proportional to the value of that observation. It is common to order the squares by additional categorical data such as tissue of origin and gene on the two axes. Our working hypothesis is that adding an additional dimension (z-axis) to the traditional 2D heat map will provide the opportunity to visualize additional layers of information that will enhance the visual discovery process. To test this hypothesis we developed a 3D heat map methodology and software package using a commercial video game engine. We apply it here to human microbiome visualization.

There are many reasonable platforms for developing 3D visualization software. OpenGL (http://www.opengl.org) with a C++, Java or scripting front end and a user-interface toolkit has all the necessary elements. A virtual reality modeling language, such as X3D (http://www.web3d.org/x3d), with scripting capabilities and free viewers, could also be used. The "Processing" visual programming environment (http://processing.org) provides a rapid prototyping environment with the ability to use Java libraries. Each choice has advantages and disadvantages. We chose here a video game development environment because game engines are explicitly designed around interactivity and immersion in a 3D environment. The ability to interactively explore a heat map visualization as you would a video game environment was an important feature. We chose the Unity3D (http://unity3d.com) development tool because it uses Mono, the open-source, cross-platform .NET implementation, so we would not be limited to code libraries supplied by the vendor. For a reasonable licensing fee we could distribute royalty-free tools that run on Windows and Macintosh machines. Unity makes GUI code easy to write, enabling rapid prototyping, and the work-flow for incorporating assets from other tools such as Maya and Photoshop is straightforward. An additional advantage is that Unity can use Direct3D on Windows machines, which allows users to employ off-the-shelf drivers to see 3D heat maps in stereo on suitable equipment. Using OpenGL we would have to explicitly code the view from each eye to produce stereo. The ability to easily see 3D heat maps in stereo is important given the widespread availability of 3D televisions and computer monitors.

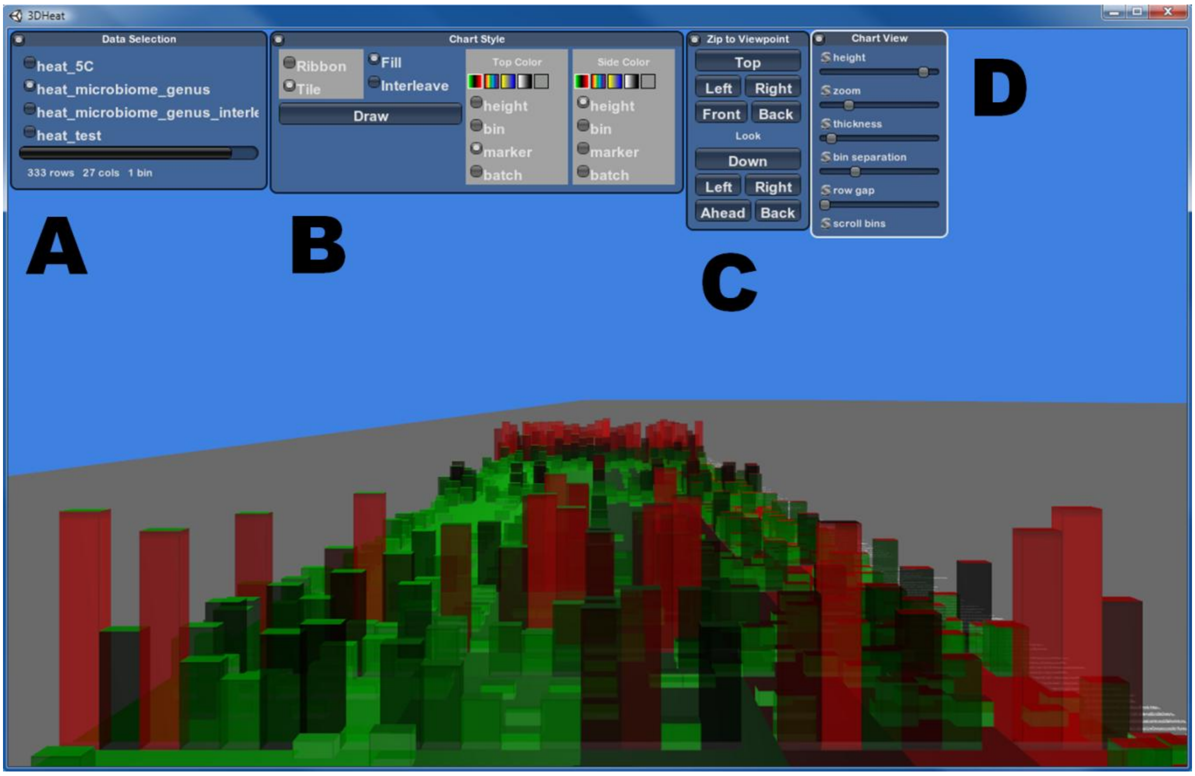


Fig. 1. Screenshot of the 3D heat map application showing menus for data selection (A), chart style (B) viewpoint (C) and chart sizing options (D). Each menu can be minimized or hidden.

A potential disadvantage of the Unity framework is that it makes low-level control of the Graphics Processing Unit (GPU) more difficult. If our primary objective had been to render massive amounts of data, we would have chosen a toolkit that allowed finer control over what is stored on the GPU, to minimize transfers between the CPU and GPU. Our principle goal though is insight through exploration and interaction. Unity allows us to render 120,000 data points (1,440,000 triangles) at 47 frames per second on a Mac Pro with two 2.8 GHz Quad-Core Intel Xeon processors and an ATI Radeon HD 2600 graphics card holding 256 MB of video memory. This scales reasonably well to about 500,000 data points corresponding to a dataset with 10,000 rows and 50 columns. This allows smooth motion and very good responsiveness in exploring datasets of moderate size.

Figure 1 provides a screenshot of the 3D heat map application with the various menus that control what data is being viewed, the style of the heat map, the viewpoint and features of the heat map itself. The menus can be minimized or hidden to make full use of the screen. The chart style menu provides options to view the heat map as ribbons where each data point is connected as in a time series or in the traditional tile or square view. There is an additional option (shown) to fill the tiles or ribbons to make solid objects. This menu also allows the user to map different color schemes for different data to the tops and sides of the 3D objects. The 3D heat map application is freely available by request from the authors or for download from http://Sourceforge.net/3dheatmap.

Figure 2 provides an example of our 3D heat map with some hypothetical data. The leftmost panel shows the data in 2D with a pattern visible as defined by the sorting of the columns (x-axis) and the rows (y-axis). The middle and rightmost panels show the same in 3D with bars on a z-axis that are proportional to intensity. Note that the sides of the bars are colored on a yellow to blue scale. This is an example of how the extra dimension can be used to visualize additional layers of information in parallel without needing to switch between perspectives. Further, the tops of the bars could be colored to represent yet another layer of information.

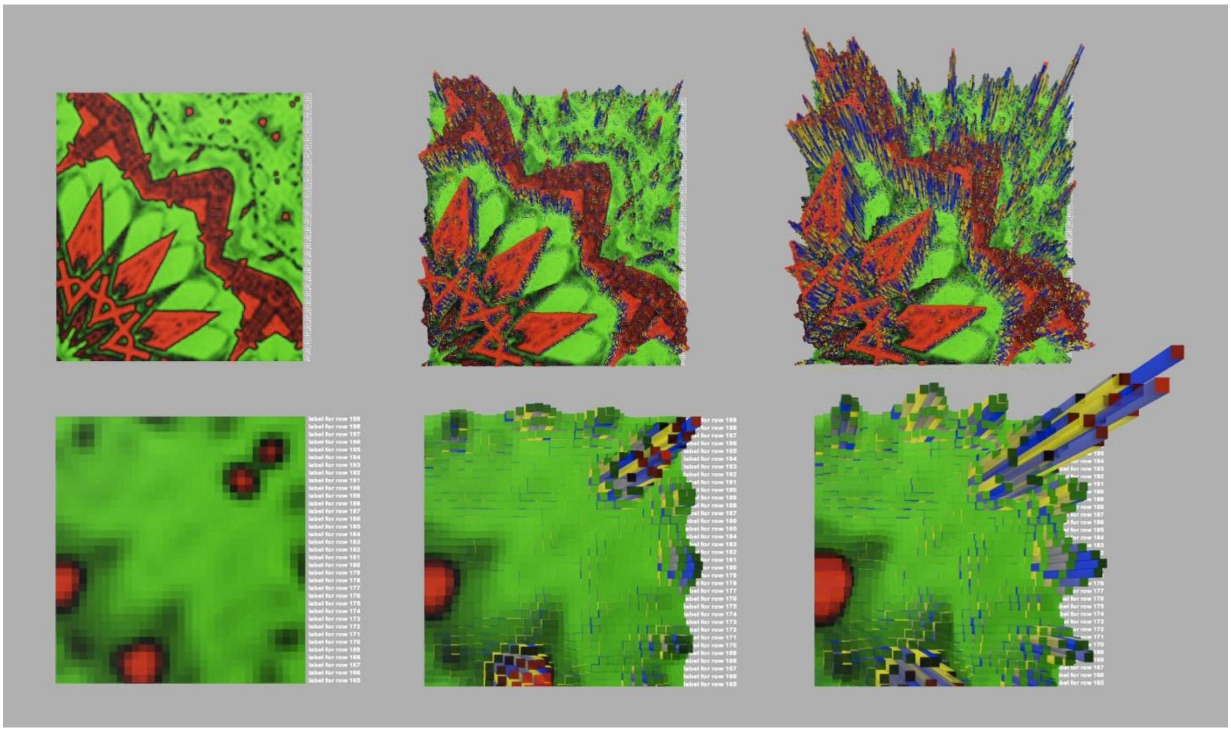


Fig. 2. 3D heat map visualization of hypothetical data. Note the additional layers of information provided by the sides of the bars when illustrated in 3D. Note that he bottom panels are a subset of the top panels corresponding to the upper right corner.

2.2. *Application of the 3D Heat Map to Human Microbiome Visualization*

We applied our 3D heat map method and software to the visualization and interactive exploration of a fecal microbiome data set from infants born prematurely. IRB approval was obtained from the Dartmouth Center for the Protection of Human Subjects in April 2009. Subjects' parents provided informed consent. Six very low birth-weight infants were enrolled within two days of birth for the study and inclusion criteria included birth weight of 501-1500 grams without major congenital or genetic anomalies. Serial stool samples were collected weekly, beginning with the first stool or meconium passed. Stool samples were aliquotted and stored at -80C and bacterial DNA was extracted using the MoBio Powersoil bacterial DNA isolation kit. DNA was quantified and then 454 pyrosequencing was performed at the Josephine Bay Paul Marine Biological Laboratories in Woods Hole, Massachusetts. High throughput sequencing was performed at the Josephine Bay Paul Center and overseen by Dr. Mitch Sogin. Pyrosequencing was targeted at the bacterial 16S the Titanium Roche α-R&D 454 amlicon informatics pipeline to analyze the bacterial community composition of samples.

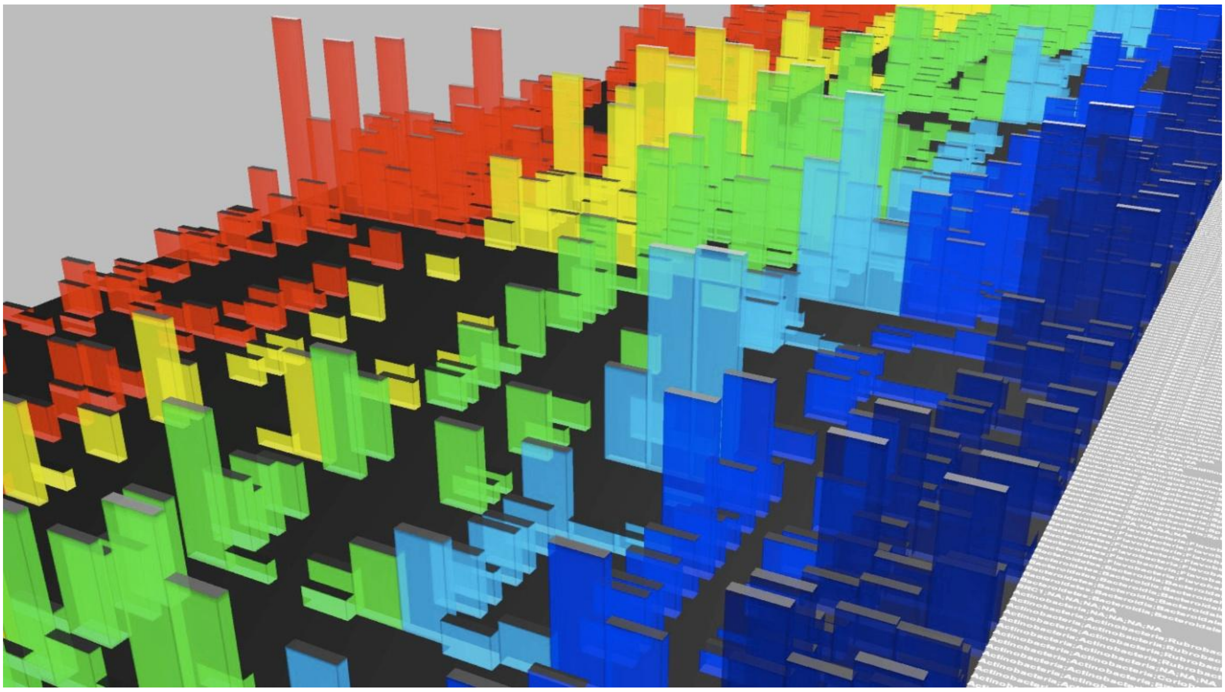


Fig. 3. 3D heat map visualization of fecal microbiome data from six premature infants. The patients and their different time points are ordered on the x-axis while bacterial species are ordered on the y-axis with the species name in white text (side). The bars in the z-axis represent relative abundance of each bacterial species for each specific patient and time point. The tops of the bars are colored in grayscale to also reflect relative abundance with lighter colors indicating higher abundance. This corresponds to the colors used in the 2D heat map. The sides of bars are color-coded by patient.

Our goal was to compare 2D and 3D heat map representations of the microbiome data from serial samples from these six patients. The 2D heat map was used to visualize relative abundance of bacteria (colored squares) with patient and time on the x-axis and bacterial species on the y-axis. For the 3D heat map we also visualized abundance of bacteria as colored squares organized by patient and time on the x-axis and bacterial specie on the y-axis. In addition, we extended bars for each colored square into the z-axis according to relative abundance with higher bars being more abundant. We added an additional layer of information about samples by coloring the four sides of the 3D bars extending into the z-axis. This demonstrates the ability to include additional layers of information in 3D space to facilitate exploration and interpretation. While it is possible to add additional symbols to a 2D heat map, it is much easier to see and explore in 3D. Symbols and other shapes would also significantly enhance the 3D visualization and would be easy to implement within the video game development framework.

3. Results

Figure 3 illustrates the 3D heat map of bacterial abundance for the six patients (x-axis and side color on each bar) over different time points for each bacterial species measured (y-axis). Note that the 3D perspective allows at least five layers of information to be visualized simultaneously. The five layers include the three axes, the top color of the bars and the side color of the bars. For example, the top color could be used to indicate the presence of sepsis in an infant at a particular time point. The ability to include clinical data with microbiome data will facilitate the visual discovery of patterns that otherwise would not be visible in a 2D heat map representation.

Not only does the 3D heat map allow multiple dimensions of information to be displayed, the video game technology allows the user to interactively explore the 3D space using the keyboard or a 3D mouse that facilitates movement in all three dimensions. The ability to 'fly' through a 3D visualization allows all of the information to be easily explored from multiple different angles. This sort of exploration and interactive visualization is not possible with a typical 3D bar plot as implemented in Microsoft Excel or other similar software packages.

Figure 4 specifically compares a 2D heat map (right panel) with the 3D heat map representation. In both panels the bacterial abundance is colored in grayscale. We kept the grayscale color-coding of abundance on the tops of the bars in the 3D heat map in addition to representing abundance on the z-axis to facilitate direct comparison to the 2D heat map. However, as mentioned above, the tops of the bars could be used to color-code additional information such as a clinical covariate. Resetting graphing parameters and assigning data to each of the dimensions can be done literally "on the fly" as the speed and direction of flight are unchanged by the update. This allows the user to explore multiple projections of the data without losing their current point of view.

Figure 5 compares a specific portion of the 2D and 3D heat maps from Figure 4. Here, the rows highlighted with the red asterisk are for a bacterial species from the Veillonellaceae family. This family belongs to the order Clostridiales and are characterized by gram-negative obligate anaerobes. The top panel of Figure 5 shows the traditional 2D heat map of the data with lighter squares indicating higher relative abundance. The bottom panel of this figure shows the same data in 3D with the patients color-coded on the sides of the bars. It is clear from the 3D heat map that the yellow and green patients have very similar patterns of bacterial abundance across the different

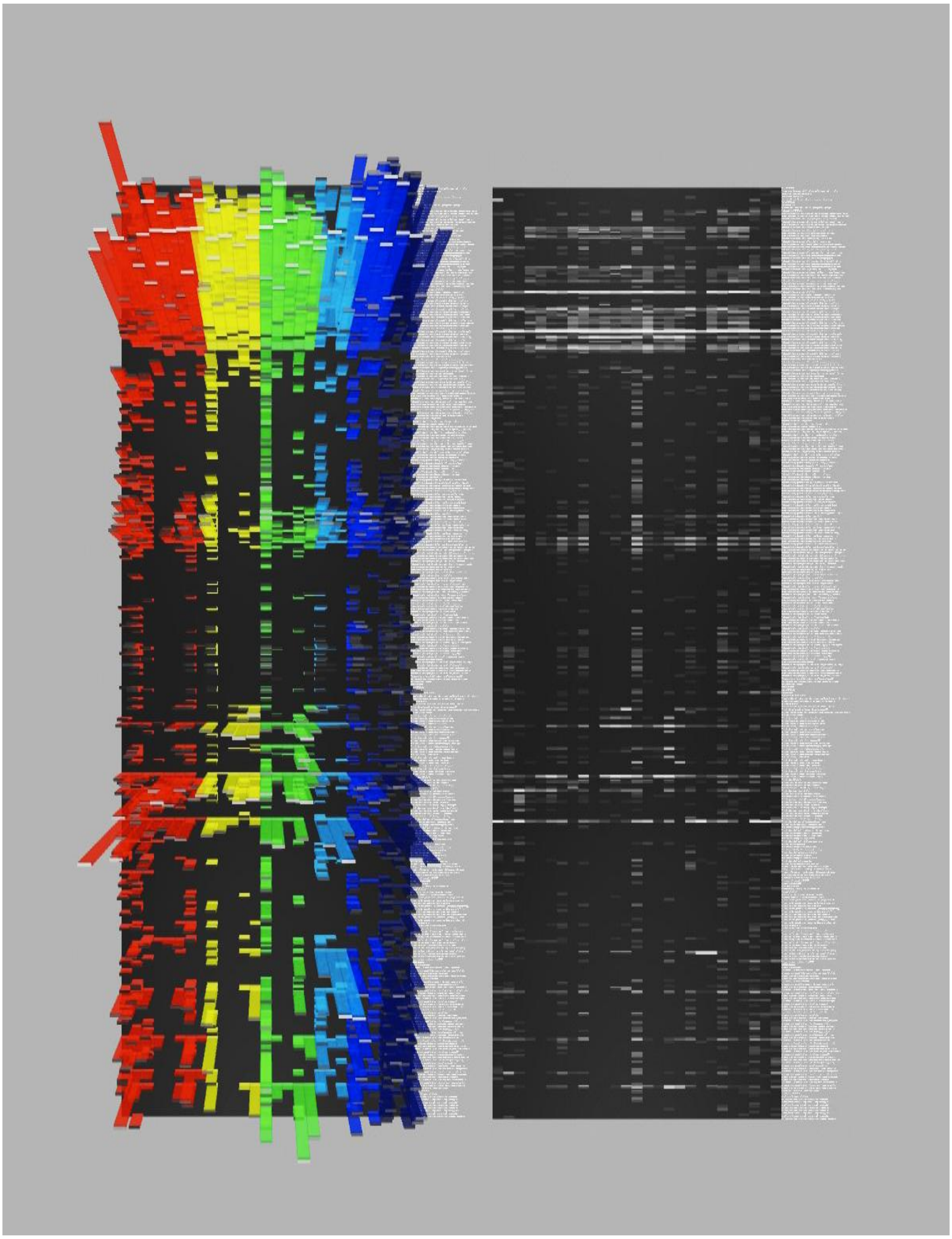


Fig. 4. Comparison of the 2D (right panel) and 3D (left panel) heat maps of bacterial abundance (grayscale in 2D and 3D and z-axis in 3D) in six premature infants. Note the additional layers of information that can be provided the z-axis, the bar top color and the bar side color.

time points. Interestingly, these two patients are twins who received the same diet (maternal breastmilk) and who had similar clinical courses without complications of prematurity. The interactive exploration provided by the 3D video game platform makes these kinds of patterns easy to identify and explore. The colors associated with the additional layers of information make the patient-specific pattern more apparent than in the 2D heat map.

4. Discussion

We have introduced here a 3D heat map method and freely available software package (3dheatmap) for interactive visualization of high-dimensional biomedical data. We have demonstrated the ability of the 3D heat map to visualize at least three more layers of information that the traditional 2D heat map. In addition, the use of a commercial video game engine has made

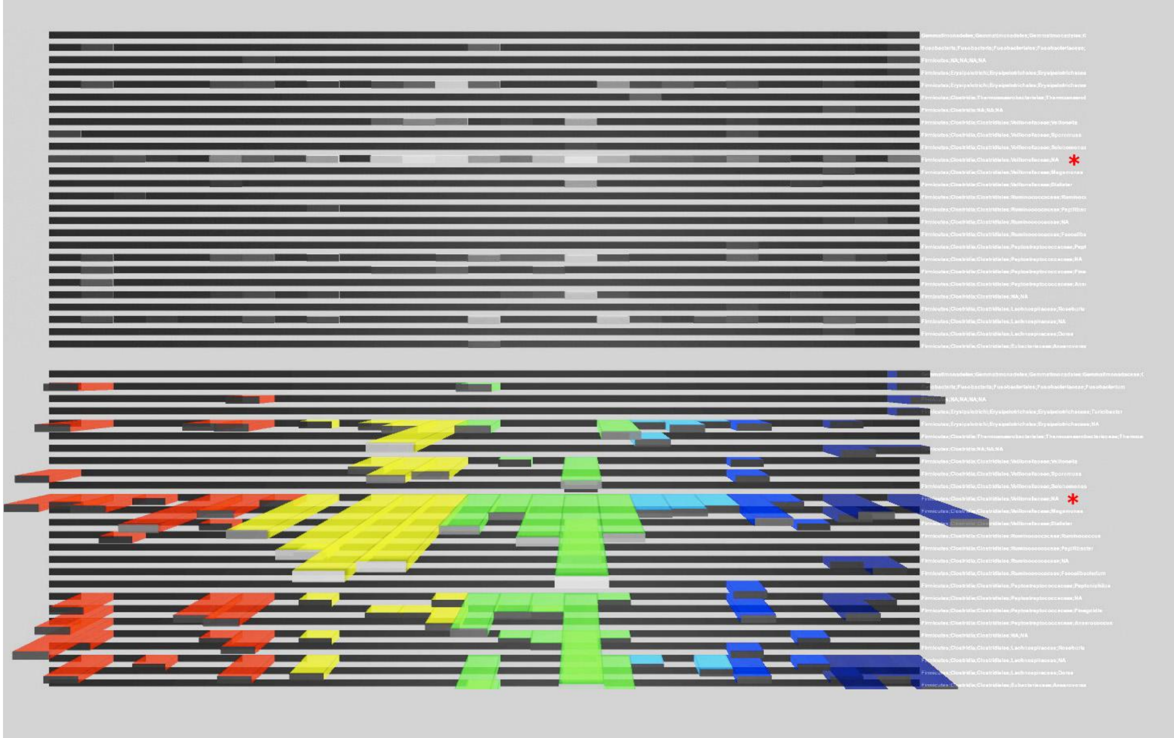


Fig. 5. A zoomed portion of the 2D (top) and 3D (bottom) heat map from Figure 3 highlighting a bacterial species from the Veillonellaceae family that has similar levels of relative abundance across time points within two patients (yellow and green).

it possible to harness the power of video games for interactive exploration of the 3D visualization. We have applied the 3D heat map method to the visualization of human microbiome data and have compared the results with that provided by a 2D heat map.

While the additional layers of information and the interactive exploration of the visualization move well beyond traditional visualization methods, there are many additional features that need to be added to move this approach from the realm of 'information visualization' to that of 'visual analytics'. Visual analytics is an emerging discipline that combines visualization methods with data analysis and human-computer interaction [16]. This is distinguished from *scientific visualization* that focuses on the mathematics and physics of visualizing 3D objects and *information visualization* that focuses on methods such as heat maps for showing high-dimensional research results. Heer et al. provide a thorough review of information visualization methods [17]. What makes visual analytics different is the integration of the visualization methods with data analysis. That is, the statistical and machine learning analyses can be launched directly from the visualization and the visualization, in turn, can be changed in a manner that is dependent on the data analysis results. This iterative and synergistic process of visualization and analysis is facilitated by computer hardware technology that makes it easy for the user to interact with the software. For example, new touch-based computer interfaces such as the Microsoft Surface Computer or the Apple iPad could replace the keyboard and mouse as the preferred interface for visual analytics. All of this combined with a 3D visualization screen or wall provides a modern visual analytics discovery environment that immerses the user in their data and research results.

Our future goals are to integrate the R statistical computing platform so that analyses can be launched directly from the 3D heat map application. It might be of interest, for example, to interactively select two different families of bacterial species within the visualization and then launch a statistical analysis comparing the relative abundance of species in the two different groups. The ability to launch analyses directly from the visualization environment opens the door to making discoveries that are inspired by visual cues rather than pre-conceived hypotheses that are dependent on existing knowledge. Of course, rigorously testing this hypothesis is not easy but Perer and Shneiderman have presented design guidelines for evaluating visual analytics software [18]. Their methodology has five phases. First, the domain expert or user is interviewed for one hour to determine their intentions. Second, there is a two hour training phase on use of the software. Third, there is a two to four week early use phase in which the users employ the software and the development team is available for troubleshooting and user support. Fourth, there is another two to four hour period of mature use where the only support that is given is for technical problems with the software. Finally, there is an outcome interview to determine whether the visual analytics software had an impact on the research of the user. Impact can be measured in many different ways but might include the generation of new ideas or hypotheses or new knowledge leading to a scientific publication. Positive impact could also be measured in terms of research efficiency. For example, the visualization approach could allow the researchers to make discoveries faster.

Human microbiome data and related research questions will continue to become more complex. This is especially true once the DNA sequence of the host is added to the mix. Visualization has an important role to play in helping the investigator become familiar with their high-dimensional data in a way that might not be possible with a spreadsheet or database. Visual analytics is an emerging discipline that harnesses the power of visualization technology, data analysis and human-computer interaction. The 3D heat map application we have presented here provides a starting point for developing such a discovery system for microbiome analysis. The use of video game engines and other 3D technology has the potential to make this technology accessible to those not skilled in bioinformatics or biostatistics.

5. Acknowledgments

We would like to thank the patients and their families for participating in the research study described here.

References

1. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Ligget, R. Knight, J. I. Gordon, *Nature*. **449**, 804 (2007).
2. The NIH HMP Working Group, *Genome Res.* **19**, 2317 (2009).
3. D. Wu et al., *Nature*. **462**, 1056 (2009).
4. The Human Microbiome Jumpstart Reference Strains Consortium, *Science*. **328**, 994 (2010).
5. E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, R. Knight, *Science*. **326**, 1694 (2009).
6. M. G. Dominguez-Bello, E. K. Costello, M. Contreras, M. Magris, G. Hidalgo, N. Fierer, R. Knight, *PNAS*. **107**, 11971 (2010).
7. D. A. Goldman, J. Leclair, A Macone, *J Pediatr.* **93**, 288 (1978).
8. I. H. Gewolb, R. S. Schwalbe, V. L. Taciak, T. S. Harrison, P. Panigrahi, *Arch Dis Child Fetal Neonatal Ed.* **80**, F167 (1999).
9. A. Schwiertz, B. Gruhl, M. Lobnitz, P. Michel, M. Radke, M. Blaut, *Pediatr Res.* **54**, 393 (2003).
10. Y. Wang et al., *The ISME J.* **3**, 944 (2009).
11. K. Ellrot, L. Jaroszewski, W. Li, J. C. Wooley, A. Godzik, *PLoS Comp Bio.* **6**, e1000798 (2010).
12. G. Vey, G. Moreno-Hagelsieb, *Mol BioSys.* **6**, 1247 (2010).
13. F. Schreiber, P. Gumrich, R. Daniel, P. Meinicke, *Bioinformatics.* **26**, 960 (2010).
14. P. H. A. Sneath, *J Gen Micro.* **17**, 201 (1957).
15. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *PNAS.* **95**, 14863 (1998).
16. J. Thomas, K. Cook, *Illuminating the Path: Research and Development Agenda for Visual Analytics.* IEEE Press (2005).
17. J. Heer, M. Bostock, V. Ogievetsky, *Comm ACM.* **53**, 59 (2010).
18. A Perer, B. Scneiderman, *IEEE Comp Graph App.* **29**, 39 (2009).

COMPARING BACTERIAL COMMUNITIES INFERRED FROM 16S rRNA GENE SEQUENCING AND SHOTGUN METAGENOMICS

Neethu Shah [1], Haixu Tang [1], Thomas G. Doak [2] and Yuzhen Ye [1]

[1]: School Of Informatics and Computing, [2]: Biology Department, Indiana University
Bloomington, IN 47408, U.S.A
E-mails: neetshah{hatang, tdoak, yye}@indiana.edu

16S rRNA gene sequencing has been widely used for probing the species structure of a variety of environmental bacterial communities. Alternatively, 16S rRNA gene fragments can be retrieved from shotgun metagenomic sequences (metagenomes) and used for species profiling. Both approaches have their limitations—16S rRNA sequencing may be biased because of unequal amplification of species' 16S rRNA genes, whereas shotgun metagenomic sequencing may not be deep enough to detect the 16S rRNA genes of rare species in a community. However, previous studies showed that these two approaches give largely similar species profiles for a few bacterial communities. To investigate this problem in greater detail, we conducted a systematic comparison of these two approaches. We developed PHYLOSHOP, a pipeline that predicts 16S rRNA gene fragments in metagenomes, reports the taxonomic assignment of these fragments, and visualizes their taxonomy distribution. Using PHYLOSHOP, we analyzed 33 metagenomic datasets of human-associated bacterial communities, and compared the bacterial community structures derived from these metagenomic datasets with the community structure derived from 16S rRNA gene sequencing (71 datasets). Based on several statistical tests (including a statistical test proposed here that takes into consideration differences in sample size), we observed that these two approaches give significantly different community structures for nearly all the bacterial communities collected from different locations on and in human body, and that these differences cannot be be explained by differences in sample size and are likely to be attributed by experimental method.

Keywords: Bacterial community; 16S rRNA gene sequencing; shotgun metagenomics.

1. Introduction

Metagenomics is the study of microbial communities sampled directly from their natural environment, without prior culturing.[1] There has been remarkable progress in this field of research due to the recent advances of Next Generation Sequencing (NGS) technologies.[2] Since over 99.8% of the microbes in some environments cannot be cultured,[3] metagenomics offers a path to the study of their community structures, phylogenetic composition, species diversity, metabolic capacity, and functional diversity. A motivation for the field is medical: human microbial flora have long been recognized to be important to human disease and health, and the human gastrointenstinal tract is one of the most thoroughly surveyed bacterial ecosystems in nature,[4] although this ecosystem remains incompletely characterized and its diversity poorly defined.[5] It is essential to evaluate not only the species diversity of microbial communities but also to analyze how the species structures of those communities change over time and space.[6] The National Institute of Health has initiated the Human Microbiome Project (HMP) with the mission of generating resources enabling comprehensive characterization of the human microbiota and the analysis of its role in human health and disease (http://nihroadmap.nih.gov/hmp/).[7]

16S rRNA gene profiling has been applied to the analysis of the genetic diversity of complex bacterial populations since the middle 1990s,[8] and is one of the primary steps in any metagenomics project. The application of 16S rRNA profiling has recently been boosted by advances in DNA sequencing techniques and the application of barcoded pyrosequencing.[9] NGS technologies—including 454 and Illumina sequencers—use 16S rRNA amplification primers targetting hypervariable regions, although it is still arguable which regions are best for species profiling: 16S rRNA genes contain nine hypervariable regions (V1–V9) that demonstrate considerable and differential sequence diversity among different bacteria. Although no single hypervariable region is able to distinguish among all the bacteria,[10] hypervariable regions V2 (nuceotides 137–242), V3 (nucleotides 433-497) and V6 (nucleotides 986–1043) contain the maximum heterogeneity and provide the maximum discriminating power for analyzing bacterial groups[10] . Barcoded pyrosequencing can produce large 16S rRNA datasets that contain hundreds of thousands of 16S RNAs fragments,[11] enabling deep views into hundreds of bacterial communities simultaneously, and have revealed much greater species diversity in many environments (e.g., soil, ocean water, and human bodies) than previously anticipated.

16S rRNA based analysis of metagenomic samples is complicated by several artifacts, including chimeric sequences caused by PCR amplification and sequencing errors.[12] According to a study by Ashelford K.E *et al*, at least 1 in 20 16S rRNA sequences currently in public repositories contains substantial anomalies,[13] and it was shown in one study[12] that some metagenomics projects may overestimate the species diversity because of the presence of sequencing errors and chimeric sequences.

Whole genome shotgun (WGS) sequencing of environmental DNA can also be used to study the species composition and diversity of natural bacterial communities,[14–16] and an increasing numbers of shotgun metagenomic sequencing datasets have been produced for various bacterial communities. Although shotgun metagenomic sequencing does not involve the biased amplification of 16S rRNA genes, the relative organism abundances inferred from metagenomic sequences vary significantly depending on the DNA extraction and sequencing protocol utilized.[17] Furthermore, shotgun metagenomic sequencing is generally not deep enough to detect rare species in complex communities.[18] Still, previous studies have shown that these two approaches give largely similar (although not identical in detail) pictures of the species structure for bacterial communities; for instance, Kalyuzhnaya et al[18] reported that the taxonomic distribution of 16S rRNA gene sequences derived from metagenomes is similar to distributions inferred from PCR-amplified libraries.[19]

Here we carry out a systematic comparison of these two approaches. We developed PHYLOSHOP, a pipeline that extracts 16S rRNA gene fragments from metagenomic sequences, reports the taxonomic assignment of the identified 16S rRNA fragments, and visualizes the taxonomy distribution. The bacterial community of a sample inferred from the identified 16S rRNA gene fragments can then be compared to the community derived from 16S rRNA gene sequencing, using the UniFrac metric,[20] which measures the phylogenetic distance between two sets of taxa, one for each community, on a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from one environment or the other. For a group of communities, a matrix of pairwise UniFrac measures can be prepared, and further subjected to Principal Coordinates Analysis (PCoA, a multivariate method that represents distance, or similarity measures, in the space of principal coordinates)[21] to study the relationship between communities. We used a P test, a commonly used phylogenetic approach to assess community differentiation.[20,22] For a given set of sequences sampled from multiple communitites, the P test estimates the minimum number of changes (switch from one community to another) required to explain the observed species distribution, and computes the significance of the difference by determining the expected number of changes, under the null hypothesis that the communities from which the sequences are sampled do not covary with phylogeny.[22] Since a significantly smaller number of 16S rRNA gene fragments can be extracted from metagenomic datasets, as compared to a 16S rRNA gene sequencing project, we also propose a new statistical test for comparing the community diversities that are inferred from collections of 16S rRNA gene fragments with vastly different numbers.

2. Methods

2.1. *PHYLOSHOP: extracting and annotating 16S rRNA gene fragments from metagenomes*

The PHYLOSHOP pipeline (Figure 1) includes the following steps.

(a) 16S rRNA sequence calling. If the given sequences are metagenomic sequences, 16S rRNA gene fragments are predicted by a HMMER search (see 2.1.1).[23,24]
(b) Chimeric sequence checking. 16S rRNA gene fragments are examined for chimeras using ChimeraSlayer and putative chimeras are removed (see 2.1.2)
(c) Mapping of 16S rRNA gene fragments. Filtered 16Sr RNA gene fragments are mapped to a phylogenetic tree of the Greengenes[25] core set of 4,938 16S rRNA genes (as of May 2009) using MEGABLAST (with a default E-valule cutoff of 1e-30). The tree and the sequences of the core set were downloaded from the Fast UniFrac website (`http://128.138.212.43/fastunifrac`). The taxonomic assignments of the core set sequences were obtained from the Greengenes website (`http://greengenes.lbl.gov`).
(d) Taxonomic assignment of 16S rRNA gene fragments. PHYLOSHOP classifies the 16S rRNA gene fragments based on their mapping to the phylogenetic tree of 16S rRNA genes.

2.1.1. *16S rRNA gene fragment prediction*

We used the bacterial 16S rRNA Hidden Markov Model (HMM) of Huang et al[23] (downloaded from `http://tools.camera.calit2.net/camera/meta_rna/`), which was constructed from 16S rRNA sequences in the European rRNA database. 16S rRNA gene fragments can then be predicted using HMM scanner (HMMER 3.0 package[26]) against a dataset of metagenomic sequences.

2.1.2. *Checking chimeric sequences*

ChimeraSlayer (`http://microbiomeutil.sourceforge.net/`) is included in PHYLOSHOP for detecting chimeric sequences in the samples used for this analysis. As chimeric sequence

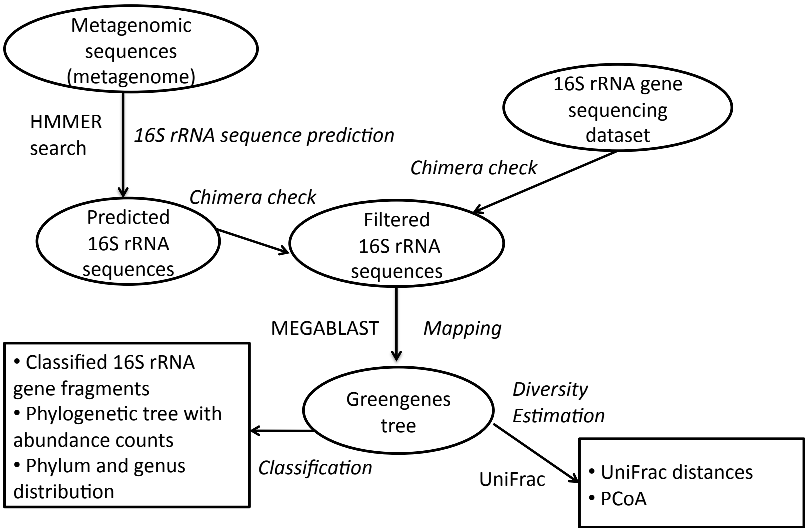

Metagenomic sequences (metagenome)
HMMER search
16S rRNA sequence prediction
16S rRNA gene sequencing dataset
Chimera check
Predicted 16S rRNA sequences
Chimera check
Filtered 16S rRNA sequences
MEGABLAST
Mapping
• Classified 16S rRNA gene fragments
• Phylogenetic tree with abundance counts
• Phylum and genus distribution
Greengenes tree
Diversity Estimation
Classification
UniFrac
• UniFrac distances
• PCoA

Fig. 1. Schematic representation of the PHYLOSHOP pipeline.

checkers do not work with very short reads (e.g., 100 bps), this option is only available for relatively long 16S rRNA gene fragments.

2.1.3. *PHYLOSHOP output*

PHYLOSHOP reports the following results, summarizing the taxonomic assignments of 16S rRNA sequences at different phylogenetic levels.

(a) Extracted 16S rRNA gene fragments, if the input is a metagenome in FASTA format.
(b) Classified 16S rRNA sequences, with an option for the user to choose the taxonomy systems—RDP,[27] NCBI or Hugenholtz.[28]
(c) Length distribution of the 16S rRNA sequences classified/extracted in a png figure.
(d) Phylum and genus disribution of the sequences mapped to the Greengenes tree.
(e) Rooted and unrooted trees in png format, showing the number of reads mapped to each identified species.

2.2. *Comparison of bacterial communities*

We used Fast UniFrac[6] to compare the structure and composition of bacterial communities.

2.3. *Statistical test of community structure differences by sampling*

A typical 16S rRNA gene sequencing dataset contains many more 16S rRNA gene fragments than those retrieved from a metagenome, so it is necessary to devise a measure that can be used to test if the observed difference in species structure between bacterial communities is statistically different, or if the difference is more likely to be caused by the dramatic difference in the numbers of 16S rRNA fragments used for inferring the bacterial communities. We propose a significance test based on multiple random sampling of subsets of 16S rRNA sequences from the larger 16S rRNA dataset. Assume there is a sample that has both a metagenomic and a 16S rRNA sequencing dataset. From the shotgun metagenomic dataset, we extract 16S rRNA gene fragments and infer the bacterial community (denote as community-m). Denote the community inferred from the 16S rRNA sequencing dataset as community-s0. From the 16S rRNA sequencing dataset, we generate n subsets of 16S rRNA sequences by random sampling and the inferred bacterial communities are denoted as community-s1, community-s2, and so on. We use the UniFrac metric to define the distance between two communities; denote the UniFrac distance between community-m and community-s0 as d_0, and the distance between the community-m and simulated community-s1, ..., community-sn as d_1, d_2, ..., and d_n. We define the P-value of the difference between the communities inferred from metagenomic sequences and from 16S rRNA sequencing dataset as the fraction of random sampling experiments that result in distance larger than d_0; this value can then be used to evaluate the significance of observed community differences, when comparing communities that have been characterized by separate methods.

2.4. *Data sets*

We analyzed 104 datasets, including 33 (32 gut and 1 oral) shotgun metagenomic datasets and 71 (42 gut and 29 oral) 16S rRNA sequencing datasets of human-associated bacterial communities; see Supplementary Tables 1–4 for the details of the datasets. The sequences were downloaded from CAMERA (`http://camera.calit2.net/`),[29] NIH Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`), and MG–RAST (`http://metagenomics.nmpdr.org/`).[30] Among these datasets, the twin study[16] has sequence datasets from both techniques—shotgun and 16S rRNA sequencing—for 18 individuals (see Table 1).

3. Results

Using PHYLOSHOP, we analyzed 33 metagenomic datasets of human-associated bacterial communities. We further compared the bacterial community structures derived from these metagenomic datasets to community structures inferred from 16S rRNA sequencing datasets, and observed clear differences in the inferred species structures associated with the different approaches (shotgun metagenomics versus 16S rRNA gene sequencing), in addition to the differences due to the different human body locations from which the samples were collected.

3.1. *Evaluation of 16S rRNA gene fragment prediction*

We first need to predict 16S rRNA gene fragments from metagneomic datasets. We compared the performance of 16S rRNA gene prediction by HMMER search[23] (implemented in the PHYLOSHOP pipeline) to predictions from the MG-RAST server, which uses BLAST searches against the Greengenes sequences. The comparison shows that HMMER searchs predicted slightly more 16S rRNA gene fragments in 11 out of the 17 metagenomic datasets shown in Figure 2. The difference is not significant, but considering that the HMMER search method is efficient and has shown high specificity and sensitivity in predicting 16S rRNA gene fragments,[23] we chose to use this method in the PHYLOSHOP pipeline. We then used 16S rRNA gene predictions from the PHYLOSHOP pipeline for the following analysis.

Table 1. Summary of the 18 gut samples that have both metagenomic datasets and 16S rRNA sequencing datasets.

Individuals	Metagenomic datasets			16S rRNA datasets		Significantly different?[d]	
	Reads[a]	Length[b]	16S rRNA[c]	Reads	Length	P test[e]	Our method[f]
TS1	217,386	238	464	25,140	126	Yes	Yes
TS2	443,526	178	658	42,186	126	Yes	Yes
TS3	510,972	201	871	17,726	126	Yes	Yes
TS4	414,754	229	731	25,705	126	Yes	Yes
TS5	490,776	205	1,108	26,608	126	Yes	Yes
TS6	535,763	221	1,207	27,007	126	Yes	Yes
TS7	555,853	243	1,310	17,469	126	Yes	Yes
TS8	414,497	243	1,036	17,170	126	Yes	Yes
TS9	499,499	250	1,024	14,787	126	Yes	Yes
TS19	498,880	165	767	43,639	126	Yes	Yes
TS20	495,039	198	1045	13,476	126	Yes	Yes
TS21	413,772	215	905	23,714	126	Yes	Yes
TS28	302,772	335	734	20,905	126	Yes	Yes
TS29	502,399	345	1,301	15,698	126	Yes	Yes
TS30	495,865	190	961	32,083	126	Yes	Yes
TS49	519,072	177	1,028	22,201	126	Yes	Yes
TS50	549,700	204	1,446	30,498	126	Yes	Yes
TS51	434,187	187	756	22,691	126	Yes	Yes

[a]: the total number of reads. [b]: the average length of reads. [c]: the total number of 16S rRNA gene fragments extracted from the metagenomic datasets. [d]: statistical significance of the difference between two communities, one inferred from the 16S rRNA sequencing dataset, and the other from the metagenomic dataset for the same individual. [e]: P-values for the P test[22] (computed using the Fast UniFrac website) are 0 for all the 18 individuals. [f]: P-values (computed using our method; see section 2.3) are $< 1e\text{-}4$ for all the 18 individuals, based on 10,000 sampling experiments for each.

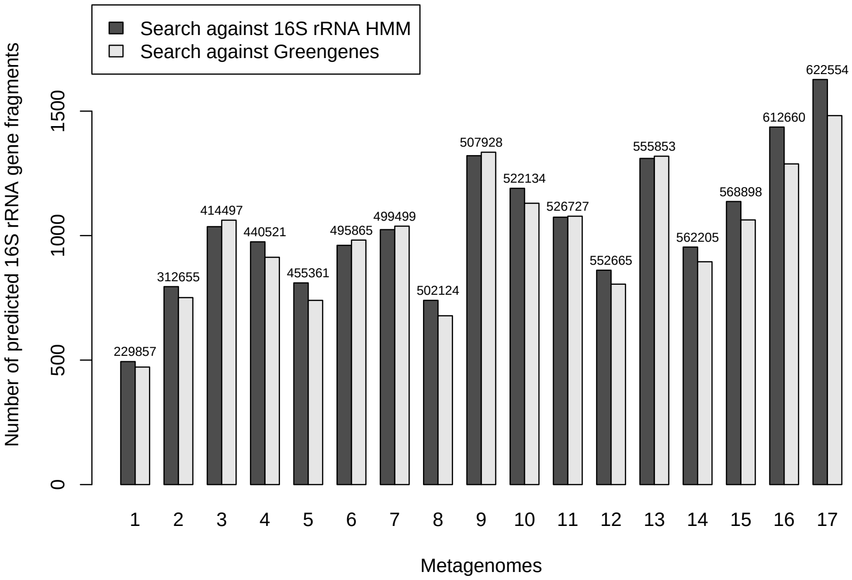
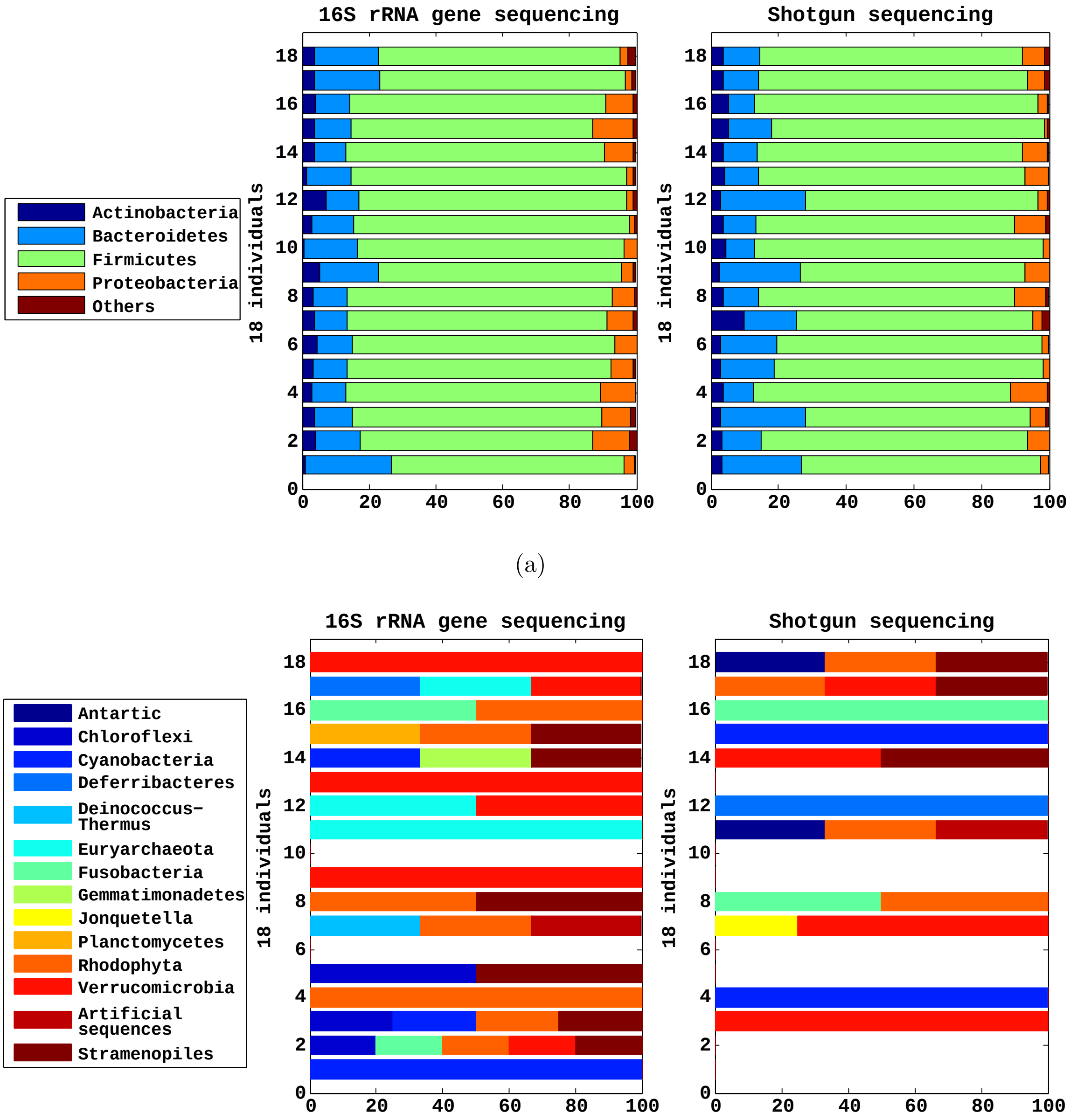


Fig. 2. Comparison of 16S rRNA prediction methods. The number of reads in each metagenome is shown above the corresponding bars.



(a)

(b)

Fig. 3. Phylum distributions of 18 gut-associated bacterial communities, inferred from 16S rRNA gene sequencing and shotgun metagenomics, in the four major (a) and other phyla (b). X-axis shows the percentage, and the phylum distribution for each individual is shown as a horizontal bar in each plot. Note that some communities (e.g., the communities in individual 6) have no reads assigned to the minor phya. The NCBI taxonomy was used, and reads assigned to "Unclassified" taxa were excluded in this analysis.

3.2. *16S rRNA gene sequencing reveals more species*

We analyzed the bacterial communities inferred from the 18 gut-associated individuals (see Table 1) that have both shotgun metagenomic and 16S rRNA gene sequencing datasets. Phylogenetic distributions of these samples show that there are clear differences in the relative abundance of the four major phylum (Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria) (Figure 3a); e.g., for individual 12 (TS21), the 16S rRNA gene sequencing dataset contains more reads from Firmicutes as compared to the metagenomic dataset. Figure 3b shows that, for most of the individuals, 16S rRNA sequencing also reveals more diverse phyla than whole genome shotgun sequencing. 16S rRNA sequencing data also found a greater diversity within genera; e.g., 35 Firmicutes genera were identified by 16S rRNA sequencing reads, whereas only 22 genera were identified by metagenomics for individual TS1 (see Supplementary Figure 1).

3.3. *Bacterial communities inferred from metagenomes are different from those inferred from 16S rRNA sequencing reads*

P tests for the 18 gut-associated samples show that, for each of these samples, the bacterial communities inferred from the metagenome and from the corresponding 16S rRNA sequencing dataset are significantly different (see Table 1). Our sampling-based tests showed similar results—the difference between the inferred communities can not be explained by the different numbers of 16S rRNA sequences. Here we use individual TS50 as an example. The TS50 metagenome includes 549,700 reads with 1,446 16S rRNA gene fragments, while its 16S rRNA gene sequencing dataset contains 30,498 16S rRNA gene fragments. The UniFrac distance (weighted) between the communities inferred from the two methods is 0.164. We simulated 10,000 subsets of 16S rRNA gene fragments from the 16S rRNA gene sequencing dataset, each containing the same number of 16S rRNA gene sequences as in the metagenome, and computed the community distances between the sampled subsets and the complete 16S rRNA gene sequencing dataset. The species structures inferred from these sampled subsets are all significantly more similar to the structure inferred from the complete 16S rRNA gene sequencing dataset (with an average UniFrac distance of 0.021; see Figure 4 for the distribution of the distances) than the complete data set is to the metagenomic dataset (0.164).

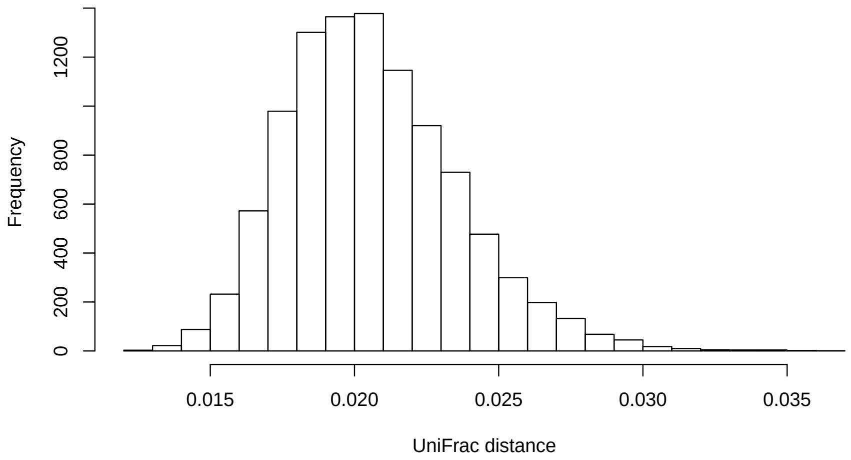


Fig. 4. Distribution of the UniFrac distance between a subset and the complete set of 16S rRNA sequencing data for the TS50 sample, based on 10,000 sampling experiments.

3.4. *Both body location and experimental technique matter*

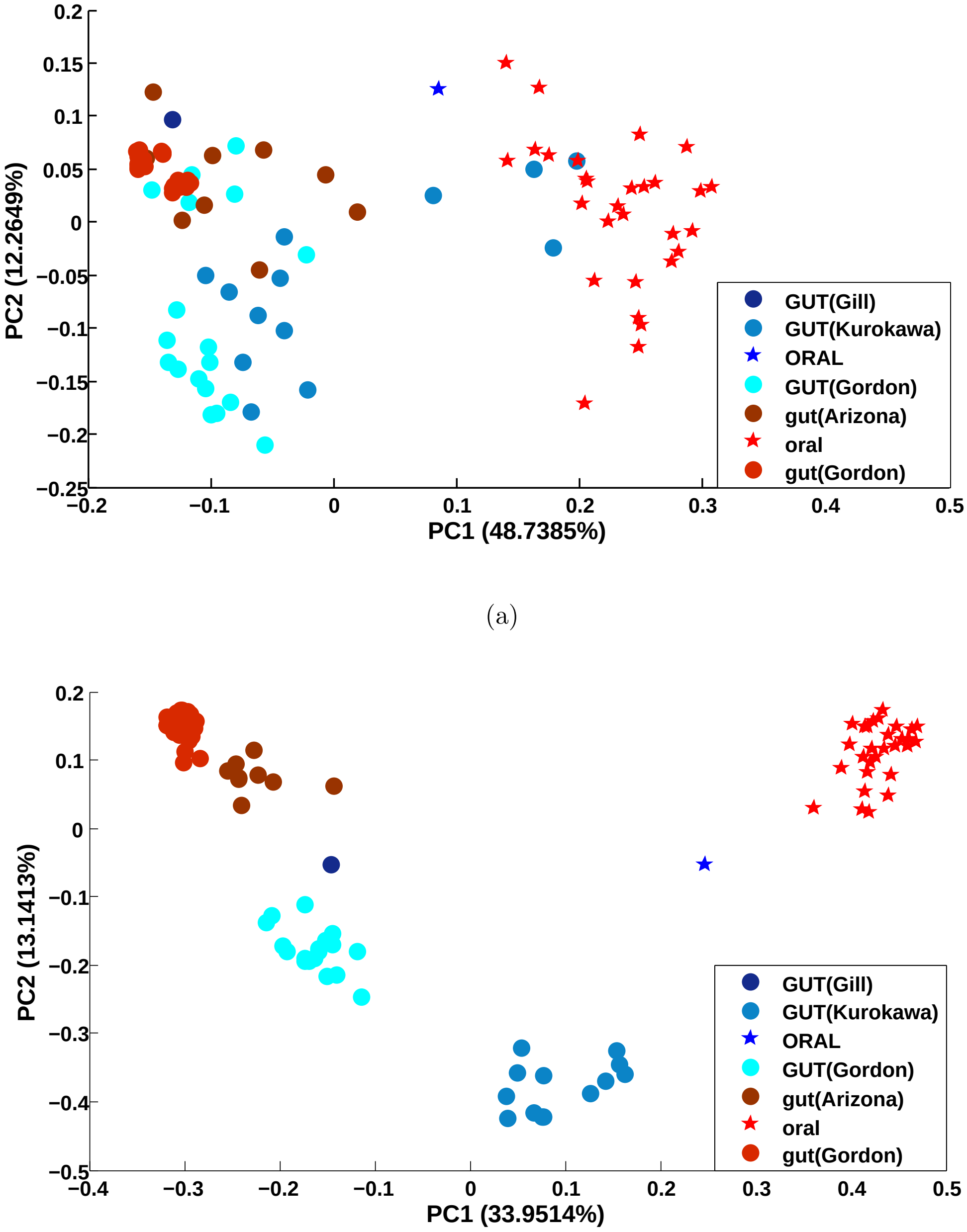
We further analyzed and compared 104 bacterial communities for different body sites inferred from metagenomes and 16S rRNA sequences, using PCoA. All of the 16S rRNA sequences (from the 16S rRNA gene sequencing, and extracted from metagenomes) were mapped to the phylogenetic tree of the core gene set of Greengenes to derive phylogenetic distributions of 16S rRNA sequences, from which UniFrac distances between any two communities can be computed. We used both weighted and unweighted UniFrac distances (weighted UniFrac weights the branches based on abundance information)[20] to derive UniFrac distance matrices. The PCoA results of the two matrices (Figure 5) show that there are at least two factors that affect community clustering: the body location, and the experimental method. The separation of the communities by experimental technique is more prominent when unweighted UniFrac distances were used (Figure 5b). For example, gut samples derived from 16S rRNA gene sequencing and whole genome shotgun sequencing (note there are 18 gut samples that have both, see Table 1) are far away from each other in the two-dimensional projection of the communities.

4. Discussion

Our comparative studies revealed significant differences in the bacterial diversities derived from 16S rRNA gene sequencing and whole genome shotgun sequencing (metagenomics) of the same sample. These differences are not due simply to the different depths of sampling in the two methods, and indicate that 16S rRNA gene sequencing can profile the bacterial communities in a greater detail than can metagenomics. Our results indicate that even when corrected for depth, conclusions derived from 16S rRNA gene sequencing and shotgun metagenome sequencing cannot be directly compared. In addition, low abundance species are best identified through 16S rRNA gene sequencing.

There can be other factors that cause the differences observed between bacterial communities inferred from 16S rRNA gene sequencing and metagenomics. For example, the 16S rRNA gene fragments derived from metagenomic datasets may cover different regions as compared to the 16S rRNA gene fragments from PCA-based pyrosequencing (which often targets a certain region of 16S rRNA gene). And it has been shown that different regions of 16S rRNA gene have different sequence diversity,[10] and therefore a certain region may serve well for profiling a certain spectrum of bacteria but not all. Ideally we could do the comparison using only the 16S rRNA gene fragments that cover the same region, but we were only able to extract a small number of such 16S rRNA gene fragments from the metagenomic datasets we tested. When bigger metagenomic datasets become available, it will be interesting and necessary to do such a comparison, using the fragments spanning the same region of 16S rRNA gene derived from different experimental techniques.

We focused on bacterial communities in this paper, but the PHYLOSHOP pipeline can easily be extended by incorporating HMMs of other phyla's RNA genes, such as archarea or fungi. Notably, the reference tree in this analysis contains only the core set of Greengenes 16sRNA genes, and thus can be further refined. Finally, the rapidly growing numbers of metagenomic samples in the public domain will provide a more comprehensive resource to



(a)

(b)

Fig. 5. Two-dimensional projection of metagenomic samples by using PCoA of the weighted (a) and unweighted (b) UniFrac distance matrices of their bacterial communities. The labels of the samples indicate the source (gut or oral), the research group involved (Gordon,[16] Gill,[31] and Kurokawa[32]), and the technique that was used (shotgun metagenomics in capital letters, and 16S rRNA gene sequencing in lower case letters). For instance, GUT (Gordon) and gut (Gordon) represent gut-associated metagenome and 16S rRNA datasets, respectively, which were produced from the same research lab. The gut (Arizona) datasets were downloaded from the NIH SRA website (accession number: SRP001377).

conduct our analysis more thoroughly and elaborately, but we suggest that for the foreseeable future metagenomic projects should be paired with 16S rRNA gene sequencing.

5. Availability

PHYLOSHOP is implemented in Python and the source codes are available for download at http://omics.informatics.indiana.edu/mg/phyloshop. The supplementary tables and figure also available in the same website.

6. Acknowledgments

The authors thank Dr. Sun Kim for useful suggestions, and thank the anonymous reviewers for valuable comments. This work was supported by National Institutes of Health grant 1R01HG004908-02.

References

1. J. Wooley and Y. Ye, *Journal of Computer Science and Technology* **25**, 71 (2010).
2. S. C. Schuster, *Nat. Methods* **5**, 16 (2008).
3. W. Streit and A. Schmitz, *Current Opinion in Microbiology* **7**, 492 (2004).
4. J. Xu, M. Mahowald, R. Ley, C. Lozupone, M. Hamady, E. Martens, B. Henrissat, P. Coutinho, P. Minx, P. Latreille, H. Cordum, A. Brunt, K. Kim, R. Fulton, L. Fulton, A. Clifton, R. Wilson, R. Knight and J. Gordon, *PLoS Biology* **5**, 1574 (2007).
5. L. Hooper and J. Gordon, *Science* **292**, 1115 (2001).
6. M. Hamady, C. Lozupone and R. Knight, *The ISME Journal* **4**, 17 (2010).
7. J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson and M. Guyer, *Genome Res.* **19**, 2317 (2009).
8. G. Muyzer, E. C. de Waal and A. G. Uitterlinden, *Appl. Environ. Microbiol.* **59**, 695 (1993).
9. M. Hamady, J. J. Walker, J. K. Harris, N. J. Gold and R. Knight, *Nat. Methods* **5**, 235 (2008).
10. S. Chakravorty, D. Helb, M. Burday, N. Connel and D. Alland, *Journal of Microbiology Methods* **69**, 330 (2007).
11. E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon and R. Knight, *Science* **326**, 1694 (2009).
12. C. Quince, A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read and W. T. Sloan, *Nat. Methods* **6**, 639 (2009).
13. K. Ashelford, N. Chuzhanova, J. Fry, A. Jones and A. Weightman, *Applied and Environmental Microbiology* **71**, 7724 (2005).
14. G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar and J. F. Banfield, *Nature* **428**, 37 (2004).
15. S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz and E. M. Rubin, *Science* **308**, 554 (2005).
16. P. Turnbaugh, M. Hamady, T. Yatsunenko, B. Cantarel, A. Duncan, R. Ley, M. Sogin, W. Jones, B. Roe, J. Affourtit, M. Egholm, B. Henrissat, A. Heath, R. Knight and J. Gordon, *Nature* **457**, 480 (2009).

17. J. L. Morgan, A. E. Darling and J. A. Eisen, *PLoS ONE* **5**, p. e10209 (2010).
18. M. G. Kalyuzhnaya, A. Lapidus, N. Ivanova, A. C. Copeland, A. C. McHardy, E. Szeto, A. Salamov, I. V. Grigoriev, D. Suciu, S. R. Levine, V. M. Markowitz, I. Rigoutsos, S. G. Tringe, D. C. Bruce, P. M. Richardson, M. E. Lidstrom and L. Chistoserdova, *Nat. Biotechnol.* **26**, 1029 (2008).
19. M. G. Kalyuzhnaya, M. E. Lidstrom and L. Chistoserdova, *ISME J* **2**, 696 (2008).
20. C. Lozupone and R. Knight, *Applied and Environmental Microbiology* **71**, 8228 (2005).
21. W. Krzanowski, *Principles of multivariate analysis. A user's perspective.* (Oxford University Press, Oxford, United Kingdom, 2000).
22. A. Martin, *Applied and Environmental Microbiology* **68**, 3673 (2002).
23. Y. Huang, P. Gilna and W. Li, *Bioinformatics* **25**, 1338 (2009).
24. R. Durbin, S. Eddy and A. Krogh, *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* (Cambridge University Press, 1999).
25. T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. Andersen, *Applied and Environmental Microbiology* **75**, 5069 (2006).
26. S. Eddy, *Genome informatics. International Conference on Genome Informatics* **23**, 205 (2009).
27. Q. Wang, G. M. Garrity, J. M. Tiedje and J. R. Cole, *Appl. Environ. Microbiol.* **73**, 5261 (2007).
28. A. C. McHardy, H. G. Martin, A. Tsirigos, P. Hugenholtz and I. Rigoutsos, *Nat. Methods* **4**, 63 (2007).
29. R. Seshadri, S. Kravitz, L. Smarr, P. Gilna and M. Frazier, *PLoS Biology* **5** (2007).
30. F. Meyer, D. Paarmann, M. DŚouza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening and R. Edwards, *BMC Bioinformatics* **9** (2008).
31. S. Gill, M. Pop, R. DeBoy, P. Eckburg, P. Turnbaugh, B. Samuel, J. Gordon, D. Relman, C. Fraser-Liggett and K. Nelson, *Science* **312**, 1355 (2006).
32. K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. Sharma, T. Srivastava, T. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi and M. Hattori, *DNA Research* **14**, 169 (2006).

MULTI-SCALE MODELLING OF BIOSYSTEMS: FROM MOLECULAR TO MESOCALE

JULIE BERNAUER

INRIA AMIB – Bioinformatique, LIX,
École Polytechnique, 91128 Palaiseau, France
Email: julie.bernauer@inria.fr

SAMUEL FLORES

Department of Bioengineering,
Stanford University, Stanford, California 94305, USA
Email: scflores@stanford.edu

XUHUI HUANG

Department of Chemistry,
Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
Email: xuhuihuang@ust.hk

SEOKMIN SHIN

School of Chemistry
Seoul National University, Seoul 151-747, Republic of Korea
City, State ZIP/Zone, Country
Email: seokmin.shin@gmail.com

RUHONG ZHOU

IBM Thomas J. Watson Research Center,
Yorktown Heights, NY 10598, USA
Email: ruhongz@us.ibm.com

1. Background

Modeling of Biosystems and their representations is one of the major challenges in current computational biology. Topics such as Structure Prediction, Dynamics and Sampling, Mesoscale Modeling, Molecular Assemblies, Structural Interactions and Systems Biology are important for better understanding biological function. They are still computationally or experimentally expensive and lead to a large amount of data, with attendant analysis challenges. In the structural genomics and systems biology era, models are thus needed at different scales, both in space and time. This session focuses on multi-scale approaches to model biosystems, and in particular those which extend simulation size and time scales.

Recent progress in protein structure prediction and protein folding dynamics has allowed in-silico experiments to reach longer timescales. Knowledge-based techniques have grown in accuracy and efficiency, partly due to the growth in solved structures, but the physics-based simulation so crucial to correct kinetics and thermodynamics is still very challenging. Adequate sampling and dynamics analysis at different size and time scales is thus of interest in the field and in this session.

Molecular modeling techniques can now model complexes of significant size, while at the macroscopic scale neuromuscular and tissue modeling has gained finer and finer resolution. In the middle stands the mesoscale -- cellular systems which we wish to model accurately and economically. In this session we will exhibit structural and dynamical modeling techniques which represent progress towards that goal.

Along with modeling molecules comes the challenge of understanding how they interact and aggregate, since a complex is often the functional unit in normal cells, or the causative agent in disease. The analysis of such assemblies is key in many therapeutic studies. Several neurodegenerative diseases are known to be correlated to protein aggregates. Understanding the phenomena of complex formation, aggregation, and misfolding requires the combination of various techniques. This session will present novel techniques and results in this field.

A yet higher level of organization exists at the system level. The role of computer modeling at this level has grown in the emerging discipline of systems biology. Different physiological functions occur at different scales, and so this field often requires multiresolution modeling techniques. Due to its crucial high-level perspective of the mesoscale, systems biology is one of the foci of our session.

The Biology, Computer Science, Chemistry and Mathematics communities have a growing interest in multiscale modeling at or near the mesoscale. This session highlights its impact on our understanding of biological systems.

2. Session Summary

This session includes four oral presentations and accepted articles, a tutorial and a discussion session. The tutorial contains two parts: a survey talk and an introduction to the multi-scale macromolecular modeling software RNABuilder.

2.1. *Oral Presentations and accepted articles*

Molecular dynamics simulations of the full triple helical region of collagen type I provide an atomic scale view of the protein's regional heterogeneity

Authors: Dale L. Bodian, Randall J. Radmer, Sean Holbert and Teri E. Klein

This article presents an all-atom explicit solvent molecular dynamics study (10 ns) of the full triple helical domain of collagen to gain insights on the role of variation in the sequence, structure and dynamics of the protein involved in fibril formation. To make this large system tractable, the authors split it into smaller overlapping fragments. A careful analysis of the trajectories obtained showed that key regions of collagen present structural heterogeneity.

Computational generation inhibitor-bound conformers of P38 MAP kinase and comparison with experiments

Authors: Ahmet Bakan and Ivet Bahar

This study focuses on the structural dynamics of an important drug target: the p38 MAP kinase. It compares the dominant changes observed in the 134 structural coordinate sets that are known for this protein, to molecular dynamics (MD) and elastic network model (ENM) in silico experiments. ENM is shown to sample the observed structural diversity well, compared to an MD simulation improved by inclusion of small organic molecules. The work describes the role of global modes in ligand binding, and suggests improvements to flexible-enzyme docking algorithms.

New conformational search method using genetic algorithm and knot theory for proteins

Authors: Yoshitake Sakae, Tomoyuki Hiroyasu, Mitsunori Miki, Yuko Okamoto

In this article, a previously described conformational search and sampling method for biomolecules is used as a base to propose a new strategy. The authors combine parallel simulated annealing using genetic crossover and knot theory to generate putative protein structures. While the previous parallel simulated annealing and genetic crossover method led to global-minimum energy protein conformations that had bad conformational properties ("knots"), the new method is shown to perform well on protein G.

Structural insights into pre-translocation ribosome motions

Authors: Samuel C. Flores and Russ Altman

Cryo-EM reconstruction and a dynamic model are used in this study to provide structural information on pre-translocation ribosome motions. The entire T.Thermophilus 16S and 23S rRNAs and most of the r-proteins are fitted in the cryo-EM map of the E.coli ribosome in the hybrid state. The fitted model exhibits a contact between P/E site tRNA and the head domain that was predicted by coevolution; it also recovers the intersubunit bridges known to be maintained during the full transition. The rotation of 16S with respect to 23S, and of the head domain with respect to the body, is modeled subject to the constraints that the ribosome pass through three experimentally observed conformations and maintain the head-tRNA contact. The results show that it is geometrically and sterically feasible for the head and tRNA to move in a coordinated fashion, and for a controversial experimentally observed intermediate to be sampled in the course of the motion. The method is applicable to the study of other large complexes.

2.2. *Tutorial*

2.2.1. *Survey Talk*

One of the emerging challenges of modern computational biology is the modeling of large biosystems. The understanding of these systems is key to solving many fundamental biological problems. Our session is focused on multi-scale techniques from molecule studies to cell or organism level. The emphasis is on: structure prediction, dynamics and sampling, mesoscale modeling, molecular assemblies, aggregation and analysis of structural interactions. These approaches may consist of a wide range of tools such as force fields, sampling, structure prediction, and dynamics methods.

In the first part of the tutorial session, we will address the issues and background relevant to the multi-resolution modeling of biological systems. We will start with an example of multi-scale modeling of human heart. After that, the following three topics will be briefly discussed: (1). Molecular Assemblies, Aggregation and Analysis of Structural Interactions. (2) Conformational sampling. (3). Multi-scale modeling of RNA structures.

2.2.2. *Introduction to RNABuilder*

In recent years we have seen an explosion in newly discovered RNA functions in the cell. However as mentioned our understanding of mechanisms of action has been hampered by a lack of structural information -- RNA's large size, flexibility, charge, folding time, and propensity for kinetic trapping challenges both experimental and computational probes of structure. The major thrust of our session is to demonstrate techniques to progress towards the mesoscale in biocomputation. To that end I will describe RNABuilder, a multi-resolution, internal coordinate dynamics code. It gives the user control over the flexibility, sterics, and forces acting on the molecule. It will be shown how this approach can be used to fold moderate-sized RNAs, or model the dynamics of large protein-RNA complexes, in a minutes to hours on a single processor. The presentation includes a short demo to show how to model a simple system.

3. Acknowledgements

Julie Bernauer acknowledges support from INRIA. Xuhui Huang acknowledges support from the Hong-Kong University of Science and technology.
The sessions chairs would like to thank all the authors and anonymous reviewers who contributed to the session.

COMPUTATIONAL GENERATION INHIBITOR-BOUND CONFORMERS OF P38 MAP KINASE AND COMPARISON WITH EXPERIMENTS

AHMET BAKAN AND IVET BAHAR

Department of Computational and Systems Biology, School of Medicine,
University of Pittsburgh, 3501 Fifth Ave, Suite 3064 BST3, Pittsburgh, PA, USA
Email: ahb12@pitt.edu and bahar@pitt.edu

The p38 MAP kinases play a critical role in regulating stress-activated pathways, and serve as molecular targets for controlling inflammatory diseases. Computer-aided efforts for developing p38 inhibitors have been hampered by the necessity to include the enzyme conformational flexibility in ligand docking simulations. A useful strategy in such complicated cases is to perform ensemble-docking provided that a representative set of conformers is available for the target protein either from computations or experiments. We explore here the abilities of two computational approaches, molecular dynamics (MD) simulations and anisotropic network model (ANM) normal mode analysis, for generating potential ligand-bound conformers starting from the apo state of p38, and benchmark them against the space of conformers (or the reference modes of structural changes) inferred from principal component analysis of 134 experimentally resolved p38 kinase structures. ANM-generated conformations are found to provide a significantly better coverage of the inhibitor-bound conformational space observed experimentally, compared to MD simulations performed in explicit water, suggesting that ANM-based sampling of conformations can be advantageously employed as input structural models in docking simulations.

1. Introduction

The p38 mitogen-activated protein (MAP) kinase, referred to as p38, is a key signaling protein activated in response to external stress; it regulates the production of proinflammatory cytokines, and as such serves as an important target for the treatment of inflammatory diseases (*1*). The structure of p38 in the presence of a variety of inhibitors/ligands has been resolved. However, the intrinsic flexibility of the enzyme has been a major challenge in accurate design and docking of potent inhibitors, and the necessity to gain a better understanding of the conformational variability of p38 has been pointed out (*2*). Our recent analysis of a set of p38 X-ray structures suggests that the structural changes observed in different ligand-bound forms of the enzyme correlate with its conformational motions intrinsically accessible in the ligand-free form (*3*). Effective generation of a representative set of conformers that would be utilized for flexible docking appears therefore as a feasible task. The development of such efficient tools for generating representative subsets of potentially bound conformers would greatly facilitate computational efforts for drug discovery, not only for this particular family, but for many target proteins, especially in the absence of sufficient structural data on their alternative conformers (*4*).

There is a multitude of approaches at different resolutions for generating conformational ensembles. Molecular dynamics (MD) simulations are broadly used in investigating specific interactions and structural changes at atomic scale, but they may be prohibitively time consuming if large-scale changes are explored (*5*). Elastic network models (ENMs), on the other hand, efficiently explore large-scale changes for large systems, but this comes at the cost of losing atomic precision, and the observed changes are restricted to the neighborhood of the global energy minimum (*6*, *7*). With the ever growing size of Protein Data Bank (PDB) (*8*), we are able to assess the utility of these methods by benchmarking them against the structural changes detected for well-studied proteins in the presence of different inhibitors.

We use here an ensemble of 134 X-ray structures resolved for p38 in different forms as the *reference for the conformational space* accessible to p38 upon binding its ligands. p38 has two small-molecule binding sites (Fig. 1A): the ATP-binding site where competitive binding of inhibitors takes place, and the lipid/compound binding site at the MAPK insert, which offers alternative targeting strategies (*9*). p38 is bound to a structurally diverse set of inhibitors in this dataset. To gain a simplified view of the dataset structural variability, we identified dominant directions of structural changes (modes) via principal component analysis (PCA) (*10*) (Fig. 1 panels B - E). PCA is a powerful technique for extracting recurrent modes of structural changes from sets of structures (*11*). Its use in assessing functional dynamics is clearly demonstrated by a recent study of substrate-bound X-ray structures of ubiquitin (*12*). Hereafter, we will refer to the modes identified by PCA as *reference* modes.

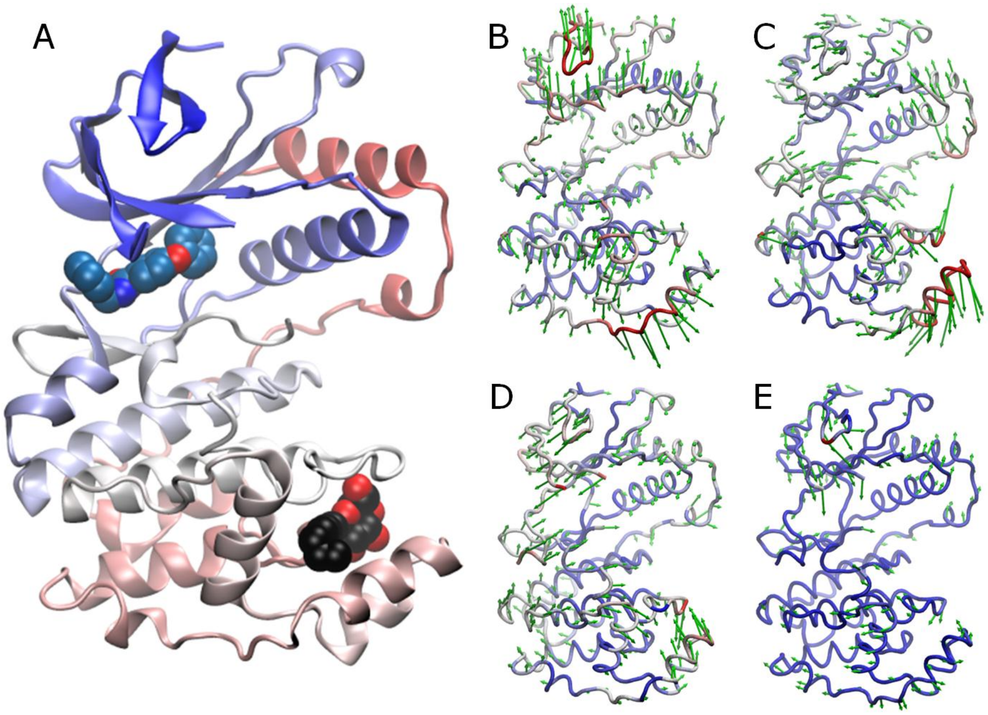


Fig. 1. p38 structure and reference modes. **A.** p38 structure (PDB id: 1ZYJ) is shown as a ribbon diagram colored by residue index from blue to red. The upper and lower lobes are referred to as the N- and C-terminal lobes. Two ligand-binding sites are distinguished: ATP-binding site, with the bound inhibitor shown in blue spheres; and the site at the MAPK insert, marked by the bound lipid (n-octyl-β-D-glucoside) in black/red space filling representation. **B-E.** Directions of PCA modes 1-4 (green arrows) retrieved from the analysis of 134 X-ray crystallographically resolved p38 structures in different forms. Coloring is based on mobility along the mode directions, red being most mobile.

We generated alternative conformations by two approaches: MD simulations, subjected to essential dynamics analysis (EDA), (*13*) and anisotropic network model (ANM) (*14-16*) analysis. MD simulations were repeated in the presence of explicit water and in solutions of water and probe molecules (small organic molecules to mimic the effects of drugs/inhibitors, as recently performed to investigate target protein druggability (*17, 18*)). We examined (i) the coverage of reference conformational space by MD and ANM, and (ii) the correspondence of the modes observed in MD-EDA and predicted by ANM to those (Figure 1 B-E) inferred from experiments.

2. Materials and Methods

2.1. Structural data

Using the human p38α sequence (GenBank id: CAG38743.1) in Biopython (http://biopython.org) protein Blast module, we retrieved from the PDB a set of 134 p38 MAP kinase isoform α structures, with 95% or more sequence identity (human or mouse proteins). Most structures contained a ligand bound to ATP binding site and/or MAPK insert (Table 1).

Table 1. Summary of p38 structural ensemble(*).

		MAPK insert				
		Unbound	*Lipid*	*Compound*		Total
ATP site	*Unbound*	4 ●	10 ♦			14
	Inhibitor	87 ●	20 ♦	8▲		115
Peptide/protein bound					5 ●	5
	Total	91	30	8	5	134

(*) Counts of different forms of p38 structures are listed. Markers refer to Figs 2 and 3.

2.2. MD simulations

We performed two types of 20 ns simulations, each repeated twice. The 1st type contained water and counter ions, in addition to p38. The 2nd was performed in a solution of water and small-organic molecules at a fixed ratio of 20:1, summarized in Table 2. The ligand-free p38 structure resolved by Wang et al.(*19*) (PDB id: 1P38) was used. Missing atoms were modeled using PSFGEN (*20*). Solvent box padding distance was at least 6 Å. Solvated system coordinates were prepared using VMD (*20*). Prior to the productive runs, probe-free systems were equilibrated for 30 ps. For probe-containing systems were subjected to 450 ps simulated annealing to achieve uniform spatial distributions of probes, followed by 300 ps equilibration. All simulations were performed using NAMD (*21*) software with CHARMM (*22*) force field.

Table 2. Description of MD simulations performed for p38 in different solvent environments.

MD Sim	*Atoms(*)*	*Water*	*Isopropanol*	*Isopropylamine*	*Acetate*	*Acetamide*
1	26454	6929	-	-	-	-
2	26454	6929	-	-	-	-
3	28515	6360	318	-	-	-
4	28563	6400	224	32	32	32

(*)protein and non-protein molecules. All simulations contained 9 sodium ions to balance the charge.

2.3. PCA/EDA of data from experiments/simulations

The PCA of ensembles of structures is an orthogonal linear transformation that projects data from Cartesian coordinate space onto a space of collective coordinates uniquely defined by the examined ensemble (*10*). The new coordinate system is such that the greatest variance in the dataset lies along the first principal component (PC) axis, shortly referred to as PC1, followed by PC2, PC3 and so on. The method of approach is identical in the EDA of MD trajectories (*13*), with the only exception that the analyzed set of conformers consists of MD snapshots, rather than the experimentally resolved structures collected from the PDB.

Both analyses are based on the cross-correlations (or covariance) observed (in experiments or simulations) between the fluctuations of C^{α}-atoms. Here, we used 324 C^{α}-atoms (residues 5-31, 36-116, 121-168,185-352) that were structurally resolved in at least 90% of the examined dataset. The approach in either case is to diagonalize the covariance matrix and examine the dominant modes of structural changes (eigenvectors) which are associated with the largest eigenvalues. Prior to PCA/EDA, structures/snapshots are superposed using the Kabsch algorithm (*23*) in an iterative procedure (*3*). Mean positions $<\boldsymbol{R}_i> = [<x_i> <y_i> <z_i>]^T$ are determined for each α-carbon *i*. The departures of α-carbons from their mean positions, $\Delta\boldsymbol{R}_i^s = [\Delta x_i^s\ \Delta y_i^s\ \Delta z_i^s]^T$ (where $\Delta x_i^s = x_i^s - <x_i>$) are organized in a *3N*-dimensional deformation vector $\boldsymbol{\Delta R}^s$ where $(\boldsymbol{\Delta R}^s)^{\mathbf{T}} = [(\Delta\boldsymbol{R}_1^s)^{\mathbf{T}}\ (\Delta\boldsymbol{R}_2^s)^{\mathbf{T}}\ \ldots.\ (\Delta\boldsymbol{R}_N^s)^{\mathbf{T}}]$), for each structure, *s*, in the dataset; and their cross-correlations, averaged over the entire set are organized in a *3N* x *3N* covariance matrix **C**. **C** may be written in terms of *N* x *N* submatrices $\boldsymbol{C}^{(ij)}$ ($1 \le i, j \le N$), each of size *3* x *3*, given by

$$\boldsymbol{C}^{(ij)} = \begin{bmatrix} \langle \Delta x_i \Delta x_j \rangle & \langle \Delta x_i \Delta y_j \rangle & \langle \Delta x_i \Delta z_j \rangle \\ \langle \Delta y_i \Delta x_j \rangle & \langle \Delta y_i \Delta y_j \rangle & \langle \Delta y_i \Delta z_j \rangle \\ \langle \Delta z_i \Delta x_j \rangle & \langle \Delta z_i \Delta y_j \rangle & \langle \Delta z_i \Delta z_j \rangle \end{bmatrix} \qquad (1)$$

Here $\langle \Delta x_i \Delta x_j \rangle$ represents the cross-correlation between the x-component of $\Delta\boldsymbol{R}_i^s$ and the y-component of $\Delta\boldsymbol{R}_j^s$ averaged over all structures ($1 \le s \le S_{tot}$) in the dataset. The trace of $\boldsymbol{C}^{(ij)}$ gives the cross-correlations between the fluctuations of residues *i* and *j* as $tr\{\boldsymbol{C}^{(ij)}\} = <\Delta\boldsymbol{R}_i \bullet \Delta\boldsymbol{R}_j>$, and that of the i^{th} diagonal block $\boldsymbol{C}^{(ii)}$ gives the mean-square fluctuations $< (\Delta\boldsymbol{R}_i)^2>$ of α-carbon *i*.

Principal/essential modes are obtained by decomposing **C** for the dataset of conformers (PDB/MD) as $\mathbf{C} = \sum_{i=1}^{m} \sigma_i\ \boldsymbol{p}^{(i)}\ \boldsymbol{p}^{(i)\mathbf{T}}$ where $\boldsymbol{p}^{(i)}$ and σ_i, are the i^{th} eigenvector and eigenvalue of **C**, respectively, and *m* is the total number of nonzero eigenvalues (m = 3N-6 if S_{tot} > 3N-6, and $m = S_{tot}$ otherwise). σ_1 corresponds to the largest variance component (i.e. $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_m$). The fractional contribution of fluctuations along $\boldsymbol{p}^{(i)}$ to the overall structural variance in the dataset is given by $f_i = \sigma_i / \boldsymbol{\Sigma}_j \sigma_j$ where the summation is performed over all *m* components.

2.4. ANM analysis and sampling of conformers using ANM modes

In contrast to PCA and EDA, the ANM analysis is performed for a *single* structure (e.g., the apo structure), not an ensemble. In the ANM, the second-order partial derivatives of the potential energy function (a sum over uniform pairwise harmonic potentials of force constant γ between all 'connected' residues in the network) are organized in the Hessian matrix **H**, which, in turn, is decomposed into *3N-6* nonzero eigenvalues λ_i and corresponding eigenvectors $\boldsymbol{u}^i$, i.e., **H** $=\sum_{i=1}^{3N-6}\lambda_i \boldsymbol{u}^{(i)}\boldsymbol{u}^{(i)\mathbf{T}}$(*14, 16*). **H** is written in terms of *N* x *N* submatrices each of size *3* x *3*. The ij^{th} submatrix is given by

$$\boldsymbol{H}^{(ij)} = \frac{\gamma\Gamma_{ij}}{\left(\boldsymbol{R}_{ij}^{0}\right)^{2}}\begin{bmatrix} X_{ij}X_{ij} & X_{ij}Y_{ij} & X_{ij}Z_{ij} \\ Y_{ij}X_{ij} & Y_{ij}Y_{ij} & Y_{ij}Z_{ij} \\ Z_{ij}X_{ij} & Z_{ij}Y_{ij} & Z_{ij}Z_{ij} \end{bmatrix} \qquad (2)$$

and $\boldsymbol{H}^{(ii)} = -\sum_{j,i\neq j}^{N}\boldsymbol{H}^{(ij)}$. Here $\boldsymbol{R}_{ij}^{0}$ is the equilibrium distance between the α-carbons *i* and *j*, and X_{ij}, Y_{ij} and Z_{ij} are its components; Γ_{ij} is the ij^{th} element of the Kirchhoff matrix **Γ**, equal to 1 if *i* and *j* are connected (or within a distance r_{cut}), zero otherwise. The ANM covariance matrix is $\mathbf{C}_{\text{ANM}} = \mathbf{H}^{-1}$ such that $1/\lambda_1$ is the counterpart of the PCA σ_1, and $\boldsymbol{u}^{(i)}$ is the counterpart of $\boldsymbol{p}^{(i)}$.

ANM conformations along mode *i* are generated using the relation $\boldsymbol{R}^{j+1} = \boldsymbol{R}^{j} \pm s\lambda_i^{-1/2}\boldsymbol{u}_i$, where *s* is a scaling parameter proportional to $(k_B T/\gamma)^{½}$ (*24*). Thus, the structural changes along the slowest/softest mode ($\boldsymbol{u}_1$) are the largest in size ($\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{3N-6}$). We generated ensembles around the initial conformation $\boldsymbol{R}^0$, using the pseudocode given in Textbox 1, with M = 3 modes and $s = K(k_B T/\gamma)^{½}$ where *K* varies as $1 \leq K \leq 7$ and $k_B T/\gamma = 2Å^2$, to obtain $15^3 = 3375$ conformers, the spread of which matches that of the reference space. The root-mean-square deviations (RMSDs) between nearest conformers along modes 1, 2 and 3, were 0.25, 0.21, and 0.14 Å, approximately. Calculations were repeated with different structures to confirm the robustness of the predicted ANM modes. Results obtained with unliganded (PDB id: 1P38 (*19*)) and inhibitor-bound p38 (PDB id: 2BAJ (*25*)) yielded practically indistinguishable results.

Textbox 1. Pseudocode for ANM sampling.

```
Initialize a list to store conformations, and append the initial conformation:
        L = [$\boldsymbol{R}^0$]
Do for ANM modes 1 ≤ i ≤ M
        Do for each conformation $\boldsymbol{R}^C$ in L
                Initialize a list to store conformations generated at this step, L_temp = [ ]
                Do for k in [-K, -(K-1), …, -1, 1, …, K-1, K]
                        $\boldsymbol{R}^k = \boldsymbol{R}^c + ks(\lambda_i^{-1/2}\boldsymbol{u}_i)$
                        Append $\boldsymbol{R}^k$ to L_temp
                Append conformations in L_temp to L
```

2.6. Comparison of dominant modes from PCA, EDA and ANM

In this section, we define the metrics for comparing the modes from ANM (predicted) and PCA (experiments); similar expressions hold for the comparison of EDA (simulations) and PCA modes, as well as EDA and ANM modes. The *overlap* between ANM and PCA modes is given by the correlation cosine $O_{ij} = \boldsymbol{p}^{(i)} . \boldsymbol{u}^{(j)}$ (*26*). The *cumulative overlap*, $CO_i^J = \left[\sum_{j=1}^{J}(O_{ij})^2\right]^{1/2}$measures how well a subset of *J* low frequency ANM modes reproduces the PCA mode *i(27)*. Note that for *J* = 3*N* -6, CO_i^J = 1 by definition, i.e., the complete set of *3N-6* ANM eigenvectors form an orthonormal basis set. Finally, the *essential subspace overlap* between the PCA and ANM subspaces spanned by top *K* modes is evaluated using $SO^K = \left[\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K}(O_{ij})^2\right]^{1/2}$ (*13*). Finally, the degrees of collectivity of the principal modes derived from either computations (ANM or EDA) or experiments (PCA) were calculated using the definition proposed by Brüschweiler (*28*).

2.7. Projection of conformations onto a reference subspace and normality test

The projection of a given conformational change $\boldsymbol{\Delta R}^s$ onto $\boldsymbol{p}^{(i)}$ is found from $c_i^s = (\boldsymbol{\Delta R}^s)^{\mathbf{T}}\, \boldsymbol{p}^{(i)}$. The points in Figs 2 and 3 represent the projections onto the subspaces spanned by PC1, PC2, PC3, and/or PC4. In the extreme case of $(\boldsymbol{\Delta R}^s)^{\mathbf{T}}$ perfectly aligned along $\boldsymbol{p}^{(i)}$, $c_i^s = ||\boldsymbol{\Delta R}^s||$, where the double bars designate the magnitude of the enclosed vector. The normality of projections of PDB structures onto the principal modes were tested using A'Agostino and Pearson's test (*29*, *30*) where skewness and kurtosis are combined into an omnibus test (using SciPy, http://scipy.org/).

3. Results and Discussion

3.1. p38 reference modes derived from X-ray crystallographic data

The ensemble of 134 p38 structures provides a rich representation of the conformational space accessible to this enzyme under a wide variety of conditions, e.g., differences in ligands, crystal conditions, or mutations. The dominant changes observed in this dataset were extracted by PCA as described above, and displayed in Fig. 1B-E. Table 3 (columns 1-5) provides more information on these modes, including the size of the associated conformational variance, σ^2, their contribution to structural variation, and their degree of collectivity. The first mode, PC1 (Fig. 1B) for example, accounts for 24.5% of the structural variability in the dataset (Table 3). The fluctuations along this mode correspond to the anti-correlated movements of the N- and C-terminal lobes of p38. Motions along the PC1 axis in the positive direction (indicated by the arrows in Fig. 1B) favor 'open' conformers, and those in the negative direction favor 'closed' forms. Movements along PC1 thus directly affect the size of the ATP/inhibitor-binding pocket in the N-terminal lobe.

Table 3. Properties of the reference (PCA) modes from experiments, and projection of MD snapshots onto them.

PCA	*PCA of PDB ensemble*				*Sim1*		*Sim2*		*Sim3*		*Sim4*	
Mode	$\sigma^{2(a)}$	%$^{(b)}$	p-value$^{(c)}$	Collectivity$^{(d)}$	$\mu^{(e)}$	$\sigma^{2(f)}$	μ	σ^2	μ	σ^2	μ	μ
1	44.5	24.5	0.26	0.49	6.9	40.1	6.4	74.8	9.9	59.7	15.4	43.6
2	37.6	20.7	0.00	0.36	6.3	30.6	9.9	46.9	11.0	43.9	8.3	31.1
3	25.6	14.1	0.00	0.58	-8.2	25.0	-0.6	20.5	0.5	20.2	0.9	14.1
4	11.1	6.1	0.63	0.52	9.4	20.8	3.2	27.1	0.8	19.9	2.7	12.9

[(a)] Variance along the reference mode in the PDB dataset. [(b)] Percent of total structural heterogeneity accounted for by the reference mode. [(c)] The probability that the projection of structures along the reference mode obeys a normal distribution. [(d)] Degree of collectivity. [(e)] Mean position of MD snapshots along the reference mode. [(f)] Variance of MD snapshots along the reference mode.

Figure 2 displays the projections of the 134 structures, each indicated by a color/shape coded symbol (described in Table 1), on the subspace spanned by PC1 and PC2. The unliganded structures (red dots in Fig. 2A) occupy the region PC1 > 0 of the subspace, consistent with their tendency to assume a relatively open form. Upon inhibitor binding, p38 tends to close down. Normality test of the projection onto PC1 (upper bars plot in Fig. 2A) shows that an approximately Gaussian distribution is obtained. The ensemble is not separated into distinct clusters in this case, suggesting that a continuous spectrum of conformers is visited rather than two distinctive 'open' and 'closed' states, with the unbound structures exhibiting a tendency to be open.

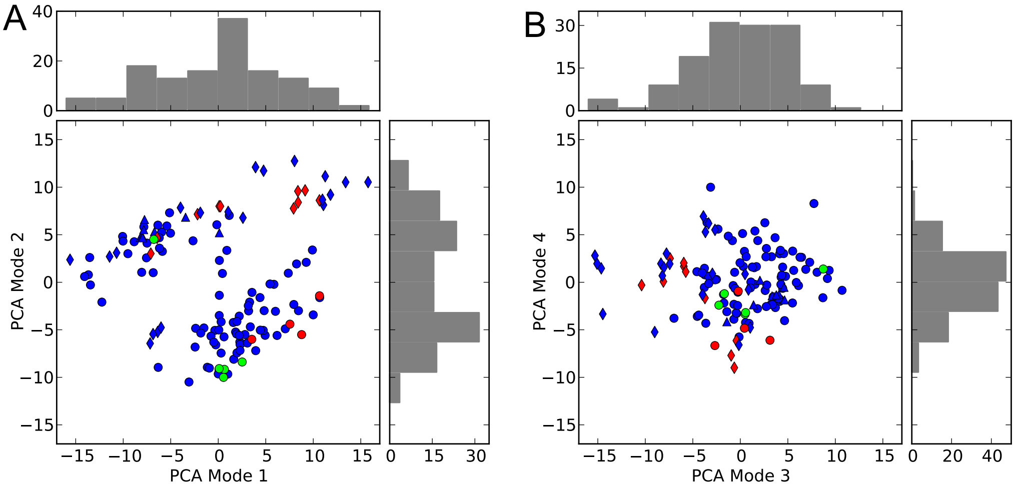


Fig. 2. Distribution of the PDB ensemble of structures on the subspaces spanned by reference modes. 134 p38 structures are projected onto PC1-PC2 (A) and PC3-PC4 (B) subspaces. Markers are described in Table 1. The distributions of structures along the individual modes are shown by the histograms.. A conformation on the positive portion of these projections corresponds to a deformation along the direction indicated by the arrows in Fig. 1B-E.

The second reference mode (PC2), on the other hand, describes the structural changes in the secondary lipid/compound binding pocket at the MAPK insert. This mode explains 20.7% of the structural variability in the dataset. As shown in Fig. 2A, this mode divides the ensemble into two groups: (i) structures with a bound ligand (lipid) molecule at the MAPK insert (red and blue *diamonds*, mostly clustered in the positive PC2 region), and (ii) structures with empty MAPK inserts (red and blue *circles*). Normality test confirms that the PDB structures exhibit a bimodal distribution along this mode. Compared to PC1, changes are slightly more localized and pronounced near the lipid-binding site at the C-terminal lobe. The collectivity of this mode is lower (0.36) compared to that of the first mode (0.49).

The structural changes along the 3rd and 4th reference modes (Fig. 1D and E) account for the respective 14.1% and 6.1% of the total variance. Both of these modes are highly collective (Table 2) as may also be seen from the uniform distribution of movements across the enzyme. Lipid-bound structures (diamonds) tend to move toward the negative direction along PC3. This behavior is particularly distinctive in inhibitor-bound structures, which results in a skewed, non-Gaussian distribution. The movements along PC4, on the other hand, exhibit a normal distribution.

In summary, the first four modes provide a description of structural changes associated with binding of ligands (ATP and/or inhibitors) at the ATP-binding sites (PC1), binding of lipids to the MAPK insert (PC2), and the collective rearrangements of the entire enzyme to accommodate different bound forms (PC3 and PC4). Notably, local changes at the binding sites are coupled to global changes in the enzyme structure (Fig. 1 panels C and D), pointing to the functional significance of the global modes favored by the p38 architecture.

3.2. Do MD snapshots provide good coverage of reference space?

Of interest is to assess how close MD- or ANM-generated conformations are to known PDB structures. In Fig. 3, we show the projection of computationally generated conformations onto the subspace spanned by top three reference modes (PC1-PC3). Panels A-C compare the snapshots from three MD runs (see Table 2), shown by the black dots, to PDB structures (indicated by the symbols in Table 1). In all three cases, we see that the conformations sampled during MD runs drift away from the large majority of the experimentally detected structures. This is also evident from the mean positions of MD snapshots along these three principal axes reported in Table 3. For example, along PC1, the mean position sampled by MD snapshots varies between 6.4 (*Sim2*) and 15.4 Å (*Sim4*), which correspond to 0.36 and 0.85 Å RMSD. Likewise, along PC2, the average positions of MD snapshots depart from the experimental dataset by up to 11.0 Å (*Sim3*). Overall, four independent runs starting from a 'central' experimental structure ended up sampling conformational subspaces that do not encompass the majority of experimental structures.

3.3. Do ANM predictions provide good coverage of reference space?

In sharp contrast to results from MD runs, Panel D in Figure 3 shows that conformers generated by deforming the starting structure along ANM modes 1, 2 and 3 are able to cover the reference subspace of conformations comprehensively. In this case ANM sampling is performed using only the slowest three modes. The present comparison clearly shows that the subspace sampled by the three softest ANM modes overlaps with the experimentally accessed subspace of conformations. This remarkable coverage of reference space (from experiments) by ANM predictions also translates into the minimum RMSD plot shown in Fig. 4. In this plot, for each PDB structure, the lowest RMSDs with respect to (i) all other PDB structures (black curve with black dots), (ii) MD snapshots in four different runs (colored as labeled), and (iii) ANM conformations generated along the softest three modes (purple) are shown. The plot for pairs of PDB structures yields an average value of 0.4 Å; ANM sampling yields 0.6 Å. MD runs, on the other hand, yield an average of 1.0Å at least. Simulations performed in the presence of probe molecules (*Sim3* and *Sim4*) yield slightly better results, although the improvement is not significant.

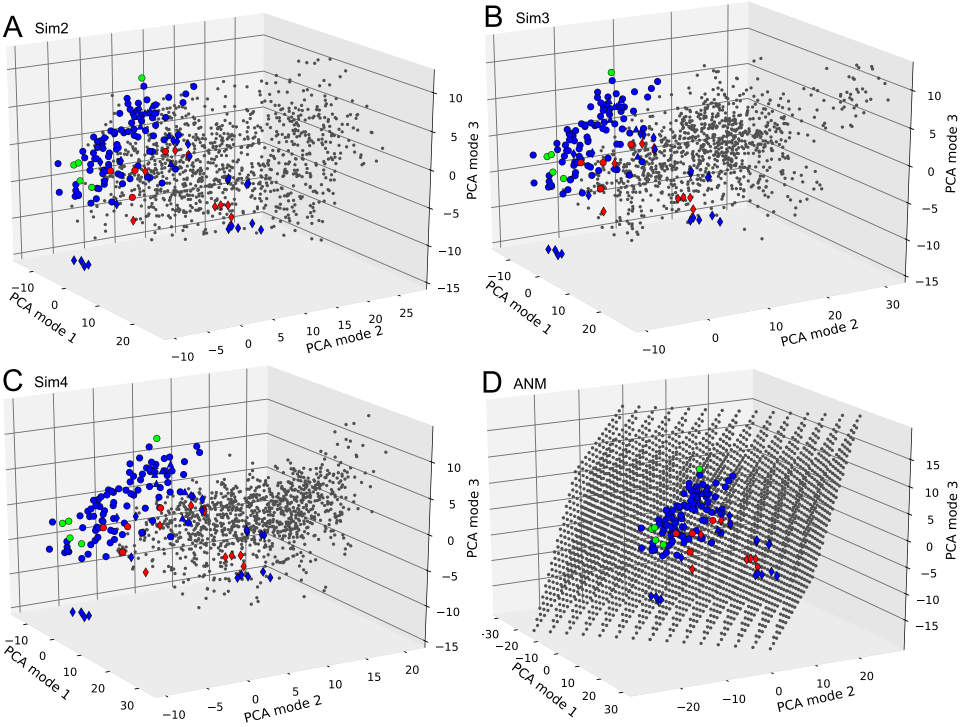


Fig. 3. Projections of MD and ANM ensembles onto the subspace spanned by the reference modes PC1-PC3. Ensembles from *Sim2*, *Sim3*, *Sim4*, and ANM are shown in panels **A**, **B**, **C**, and **D**, respectively. PDB structures are marked as in Fig. 2. Conformations generated by computations are shown by gray points. The perspective is the same in all panels for ease of comparison.

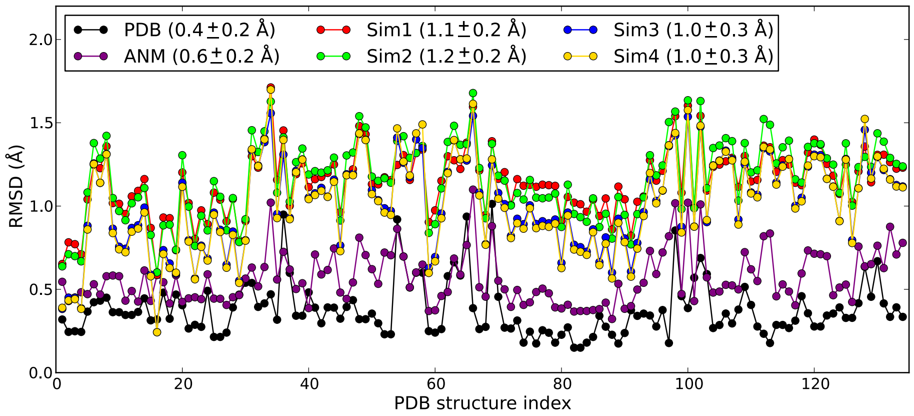


Fig. 4. Minimum RMSD from PDB structures. Results are shown for all PDB structures (indexed in alphabetical order along the abscissa). The black curve refers to the RMSDs from the PDB structures themselves (experimental data); the purple curve displays the min RMSDs achieved by ANM sampling; and other curves (labeled) refer to MD runs Mean and standard deviations are given in the legend.

3.4. Correspondence of ANM and MD-PCA Modes to Reference Modes

We performed the EDA of the four MD trajectories, and compared the essential modes derived from these runs to the reference modes obtained from the experimental dataset. Fig. 5 panels A and C show for the runs *Sim2*, *Sim3* and *Sim4* the overlap (correlation cosine) between each of the essential 10 modes (from EDA of MD) and the top 10-ranking reference modes (from PCA of

PDB structures). Panel D displays the correlation between the ANM and PCA modes. If we focus in particular on the top two pairs of modes, the lowest overlaps are observed in *Sim2* (53% or less) and similar observations (not shown) were made for *Sim1*. In both of these two cases, the protein was simulated in water. In *Sim3* and *Sim4*, the correlations with the first two reference modes increase to 63%. This is due to the existence of probe molecules which were able to find either binding pocket on p38, and assisted in the stabilization of bound conformers.

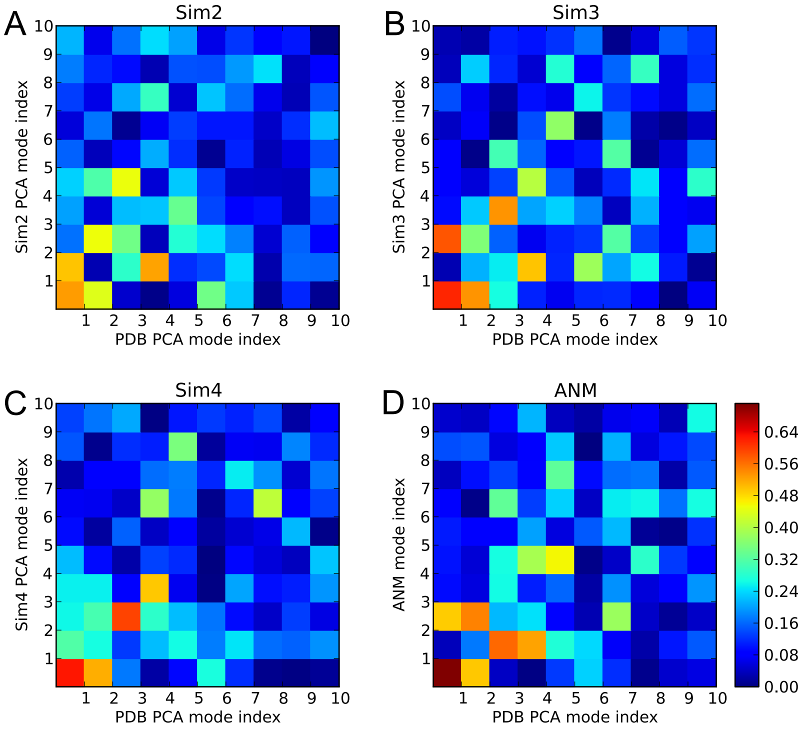


Fig. 5. Correlations between experimentally observed and computationally obtained modes of structural changes. Panels A-C show the overlap between top-ranking PCA modes (from 134 PDB structures) and the modes yielded by MD runs EDA. Panel D displays the overlaps between ANM and PCA modes.

The comparison of ANM modes with the reference modes shows, on the other hand, a much stronger correlation. The slowest mode (ANM1) alone exhibits an overlap of 75% with PC1. Thus, the most prominent conformational variation experimentally observed for p38 upon binding its ligands is in remarkable agreement with the softest mode of motion intrinsically accessible to the enzyme in the unbound state. That is, it is easiest to deform the protein along this ANM mode, or the protein is most likely to sample alternative conformations along this mode under native state conditions (presumably within time scales of the order of μs). In addition, the cumulative overlap between top reference mode and first three ANM modes reaches 87%. We also found significant overlaps between reference modes PC2, PC3, and PC4 and the low frequency ANM modes 1, 2 and 3: PC2 overlaps with ANM1 and ANM3 by 0.49 and 0.55, respectively, yielding a cumulative overlap of 0.74 with these two modes. PC3 correlates with ANM2 by 57%, and PC4 correlates with ANM2 by 52%. It is not thus surprising to see that the alternative conformations generated along ANM1-3 provided a satisfactory coverage of the experimentally observed dataset (Fig. 3D).

Apparently, 65% of the variability in the PDB ensemble is well explained by slowest three ANM modes. A measure of such correspondence is the essential subspace overlap (*13*). Using three modes, we find that the essential subspace overlap between ANM and reference dataset is 75%. This value is 62% for MD runs *Sim1* and *Sim2*, and increases to 67 and 68%, respectively for *Sim3* and *Sim4*. This suggests that the directions of structural changes observed in MD, represented by EDA modes 1-3, exhibit *some* correlations with the PCA modes 1-3 extracted from experiments, but there is a drift away from the original subspace during MD runs such that the conformations sampled by MD deviate from the reference state, leading to the relatively high RMSD values shown in Figure 4.

4. Conclusion

We presented a detailed analysis of the structural variations in a large ensemble of p38 MAP kinase X-ray structures, compared to those predicted by snapshots/models generated by MD simulations and by ANM methodology. Our results show that ANM is able to capture the modes of motion that are relevant to ligand/inhibitor/ lipid binding much better than MD simulations. The use of probe molecules that mimic the interactions of the protein with inhibitor molecules improves the ability of MD to sample the relevant modes (Fig. 5B-C). Yet, the conformers sampled by MD trajectories of tens of nanoseconds generally fell short of encompassing the space of inhibitor-bound PDB conformations (Fig. 4). MD simulations of 20 ns take about 2 weeks using 12 processors for a typical kinase. Generating a broader MD ensemble would take longer and would demand considerably larger numbers of processors, and this might not prevent the drift away from the experimental structures. The generation of conformers along ANM soft modes, on the other hand, is achieved within minutes, if not seconds, and provides an accurate sampling of experimentally detected subspace. Notably, the latter does not necessitate expensive computations, nor knowledge of multiple structures.

Our work is the most comprehensive comparative analysis of a protein kinase (p38) dynamics using multi-resolution methods. In a previous study, based on small sets of PDB structures for four different kinases, local changes in the glycine rich loop (a β-hairpin that interacts with inhibitors) of the N-terminal lobe were observed (*2*), and the fast modes were deemed to be used to capture the conformational variability in ligand-bound structures (*31*). Our current and previous (*3*) results show that up to 65% of the changes experimentally observed in p38 MAP kinases are actually collective changes explained by 2-3 softest modes. *The local motions at ligand-binding site are an integral part of these global movements*, and the structural variations observed in different ligand-bound conformers are well represented by the structural rearrangements along the global modes. These observations are in line with work of May and Zacharias (*32*) in which relaxing a protein kinase along global modes during docking simulations improved the prediction of bound conformers. Their approach circumvents the problem of dealing with conformations potentially irrelevant to ligand binding (decoys). Such decoys are generated by both computational methods, but eliminating ANM conformations with RMSD larger than a threshold to the initial structure can improve the accuracy of sampling. Furthermore, current results are important as they suggest a key coupling between global motions and local binding events, which will need to be systematically examined for a series of proteins. The method and application set forth in the present study may be readily extended to perform such a critical assessment for a large set of proteins.

Acknowledgements

Support from NIH grants 5R01GM086238-02 and 5R01LM007994-06 is gratefully acknowledged by I. Bahar.

References

1. S. Kumar, J. Boehm, J. C. Lee, *Nat. Rev. Drug Discov.* **2**, 717 (2003).
2. C. N. Cavasotto, R. A. Abagyan, *J Mol. Biol* **337**, 209 (2004).
3. A. Bakan, I. Bahar, *Proc. Natl. Acad. Sci. U. S. A* **106**, 14349 (2009).
4. M. Totrov, R. Abagyan, *Curr. Opin. Struct Biol* **18**, 178 (2008).
5. R. O. Dror, M. O. Jensen, D. W. Borhani, D. E. Shaw, *J Gen. Physiol* **135**, 555 (2010).
6. I. Bahar, T. R. Lezon, A. Bakan, I. H. Shrivastava, *Chem. Rev.* (2009).
7. I. Bahar, *J Gen. Physiol* **135**, 563 (2010).
8. H. M. Berman *et al.*, *Nucleic Acids Res.* **28**, 235 (2000).
9. J. J. Perry, R. M. Harris, D. Moiani, A. J. Olson, J. A. Tainer, *J Mol. Biol.* **391**, 1 (2009).
10. I. T. Jolliffe, *Principal Component Analysis* (Springer, New York, ed. 2nd, 2002), p. -487.
11. A. Amadei, A. B. Linssen, H. J. Berendsen, *Proteins* **17**, 412 (1993).
12. O. F. Lange *et al.*, *Science* **320**, 1471 (2008).
13. A. Amadei, M. A. Ceruso, N. A. Di, *Proteins* **36**, 419 (1999).
14. A. R. Atilgan *et al.*, *Biophys. J* **80**, 505 (2001).
15. P. Doruker, A. R. Atilgan, I. Bahar, *Proteins* **40**, 512 (2000).
16. E. Eyal, L. W. Yang, I. Bahar, *Bioinformatics.* **22**, 2619 (2006).
17. J. Seco, F. J. Luque, X. Barril, *J Med. Chem.* **52**, 2363 (2009).
18. O. Guvench, A. D. Mackerell, Jr., *PLoS. Comput Biol.* **5**, e1000435 (2009).
19. Z. Wang *et al.*, *Proc. Natl. Acad. Sci. U. S. A* **94**, 2327 (1997).
20. W. Humphrey, A. Dalke, K. Schulten, *J Mol. Graph.* **14**, 33 (1996).
21. J. C. Phillips *et al.*, *J Comput Chem.* **26**, 1781 (2005).
22. B. R. Brooks *et al.*, *J Comput Chem.* **30**, 1545 (2009).
23. W. Kabsch, *Acta Crystallographica Section A* **32**, 922 (1976).
24. C. Xu, D. Tobi, I. Bahar, *J. Mol. Biol* **333**, 153 (2003).
25. J. E. Sullivan *et al.*, *Biochemistry* **44**, 16475 (2005).
26. F. Tama, Y. H. Sanejouand, *Protein Eng* **14**, 1 (2001).
27. L. Yang, G. Song, A. Carriquiry, R. L. Jernigan, *Structure.* **16**, 321 (2008).
28. R. Bruschweiler, *The Journal of Chemical Physics* **102**, 3396 (1995).
29. R. D'Agostino, *Biometrika* **58**, 341 (1971).
30. R. D'Agostino, E. S. PEARSON, *Biometrika* **60**, 613 (1973).
31. C. N. Cavasotto, J. A. Kovacs, R. A. Abagyan, *J Am. Chem. Soc.* **127**, 9632 (2005).
32. A. May, M. Zacharias, *J Med. Chem.* **51**, 3499 (2008).

MOLECULAR DYNAMICS SIMULATIONS OF THE FULL TRIPLE HELICAL REGION OF COLLAGEN TYPE I PROVIDE AN ATOMIC SCALE VIEW OF THE PROTEIN'S REGIONAL HETEROGENEITY

DALE L BODIAN, RANDALL J RADMER[†], SEAN HOLBERT[·], TERI E KLEIN

Department of Genetics, Stanford University, 1501 S. California Avenue
Palo Alto, CA 94304
Email: teri.klein@stanford.edu

Collagen is a ubiquitous extracellular matrix protein. Its biological functions, including maintenance of the structural integrity of tissues, depend on its multiscale, hierarchical structure. Three elongated, twisted peptide chains of >1000 amino acids each assemble into trimeric proteins characterized by the defining triple helical domain. The trimers associate into fibrils, which pack into fibers. We conducted a 10 ns molecular dynamics simulation of the full-length triple helical domain, which was made computationally feasible by segmenting the protein into overlapping fragments. The calculation included ~1.8 million atoms, including solvent, and took approximately 11 months using the CPUs of over a quarter of a million computers. Specialized analysis protocols and a relational database were developed to process the large amounts of data, which are publicly available. The simulated structures exhibit heterogeneity in the triple helical domain consistent with experimental results but at higher resolution. The structures serve as the foundation for studies of higher order forms of the protein and for modeling the effects of disease-associated mutations.

1. Introduction

Collagen is a ubiquitous protein found in all multicellular organisms. Type I collagen, the most abundant protein in mammals, provides structural and functional integrity to bones, tendons, blood vessels, and other tissues. The function of collagen relies on its multiscale, hierarchical structure. The type I collagen protein is composed of two α1(I) and one α2(I) peptide chains, each composed of >1000 amino acids. The three chains, each with a left-handed polyproline II-like twist, associate into a supercoiled right-handed triple helical structure nearly 3000 Å long.[1] The heterotrimers assemble into fibrils with a characteristic packing arrangement that is observable in electron micrographs as a 670 Å repeating pattern of gaps and bands. The fibrils in turn pack into fibers or other suprafibrillar architectures in the extracellular matrix whose configuration varies with the biological tissue in which it is found.[2, 3]

The primary sequence of collagen proteins is characterized by multiple repeats of the Gly-X-Y triplet, where X and Y can be any amino acid but are often proline and hydroxyproline. The type I collagen peptide chains α1(I) and α2(I) each contain 338 uninterrupted copies of this repeat which form the triple helical domain in the heterotrimer. Disruption of the structure of the triple helical domain by mutation, particularly of any of the invariant glycine residues, is associated with disease, most commonly osteogenesis imperfecta, a set of disorders characterized by brittle bones.

Experimental studies[4-7] and computational models[8, 9] have revealed that, despite the repetitive sequence, the triple helical domain is not homogeneous and has structural and biological

[†] Current address: Department of Bioengineering, Stanford University, Stanford, CA 94305.

[·] Current address: Department of Computer Science, Stanford University, Stanford, CA 94305.

properties that vary throughout its length. Results from multiscale modeling suggest that variation in the sequence repeat impacts collagen's mechanical properties.[10, 11] Regional models have been proposed to delineate the heterogeneity of the observed properties, but the boundaries of the regions are still incompletely defined.[6, 12, 13] Structural studies have been performed to characterize the sequence-dependence of triple helix properties in detail.[14] However, because of collagen's size and fibrous structure, atomic-level analysis of the full-length protein has been difficult. The crystal structure of rat tail tendon collagen determined by fiber diffraction[15] provides a low resolution view of the trimer within a native fibril. However, the Cα-only structure does not permit detailed analysis of the atomic interactions, and does not provide a picture of the dynamic properties of the triple helix. Crystallographic and NMR studies provided atomic level detail of the triple helical conformation and NMR and molecular dynamics have revealed aspects of collagen's dynamic behavior.[14] One such finding is that less stable regions of the triple helix may be less tightly wound on average, with longer intermolecular backbone hydrogen bonds contributing to the decreased stability.[16, 17] However, the sequences investigated in these studies are limited to those of model peptides, most of which are repeating Gly-Pro-Hyp sequences, rather than native collagen sequence.

To further our understanding of the structure and dynamics of the native collagen, we carried out a 10 ns molecular dynamics simulation of the complete, heterotrimeric, triple helical domain of human type I collagen. We designed a relational database to store the raw and derived data and developed specialized analyses protocols to handle analysis of the large volume of data generated. The simulated structures capture structural variation consistent with experimental structural and biophysical studies. Our simulations support hypotheses about collagen's dynamic heterogeneity and lend insights into the properties and regions of the native collagen triple helix. Finally, the simulation structures can serve as the basis for future molecular dynamics studies on the effects of mutations and polymorphisms on collagen structural properties and their relationship to the higher order forms of the protein.

2. Materials and Methods

2.1. *Software and parameters*

Molecular dynamics simulations were performed with GROMACS version 3.3[18] and managed by the Folding@home distributed computing servers.[19] Parameters are from the AMBER-99 force field[20] supplemented by published values for hydroxyproline.[21] Data are stored in a postgres version 8.1 relational database.

2.2. *Construction of the starting structures*

The 1014 amino acid-long sequences of the peptide chains comprising the triple helical region of human type I collagen were taken from residues 179-1192 of GenBank entry NP_000079 for α1(I) and residues 91-1104 of AAB59374 for α2(I). The chains were assembled into a heterotrimer with chain order α1(I)-α2(I)-α1(I). Coordinates of an idealized triple helix were generated with GENCOLLAGEN[22] using default parameters, with all Y-position prolines converted to 4'-

hydroxyproline. The initial helical symmetry parameters for the Gly-X-Y triplets were: ϕ = -74°, ψ = 170°, ω = 180° for glycine, ϕ = -75°, ψ = 168°, ω = 180° for the residue in the X position, and ϕ = -75°, ψ = 153°, ω = 180° for Y-position residues.

The full triple helical region was split into 24 overlapping fragments to facilitate parallelization of the computation (Figure 1). The termini of all fragments were capped with neutral acetyl and N-methylamine end groups. All fragments include 85 residues per chain except the C-terminal fragment which spans 48 triple helix positions, and overlap their adjacent fragments by 42 residues. The starting residue of each fragment is offset from the previous fragment by 42 residues, giving an additional overlap of one helix position between every other fragment. For example, fragment 0 is residues 1-85, fragment 1 is 43-127, and fragment 2 is 85-169.

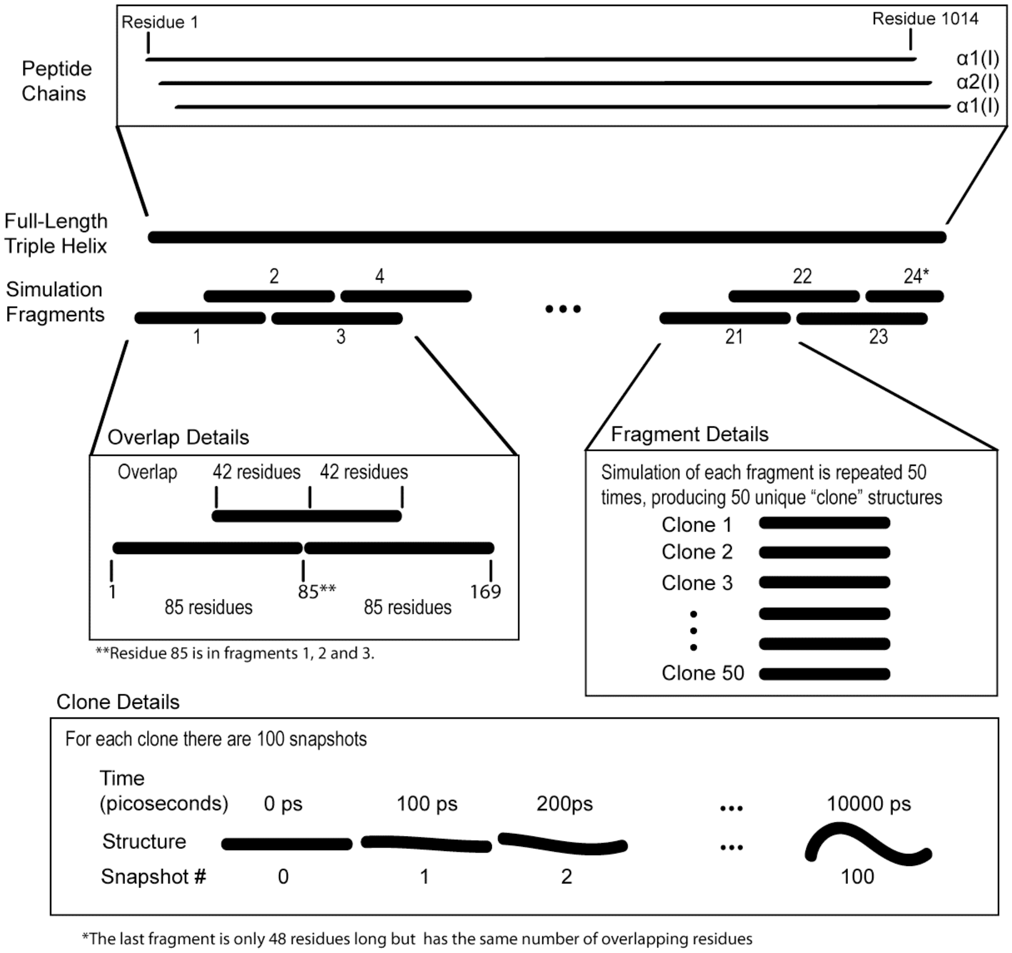


Fig. 1. Schematic representation of the molecular dynamics simulation. The full triple helical region, comprised of 1014 residues per polypeptide chain with chain order α1(I) - α2(I) - α1(I), was modeled as 24 overlapping fragments. Each fragment is associated 50 clones, trajectories of 50 simulations performed for that fragment. Simulated structures are sampled every 100 ps, yielding 101 snapshots for each clone. Snapshots are numbered sequentially with snapshot 0 the starting structure for that clone.

2.3. *Molecular Dynamics Simulations*

Periodic boundary conditions were applied and each fragment was aligned along the Z-axis and then solvated in a box of TIP3P water.[23] A box size of 45 x 45 x 380 Å was selected for all fragments, giving at least 12 Å water boundary on all sides. To equilibrate the water, position restraints applied to protein heavy atoms and simulations were performed for 100 ps at constant temperature (300°K) and pressure (1 atm).

After equilibration, the simulation was conducted at a constant temperature of 300 °K and continuous pressure of 1 atm for 10 nanoseconds with no restraints. The Nose-Hoover thermostat[24] and isotropic Berendsen barostat[25] were used for temperature and pressure control, respectively. All covalent bonds that involve hydrogen atoms were constrained with the LINCS algorithm.[26] A 2 fs time step was used for all simulations. The total linear and angular momentum were removed at every time step, for protein and water separately. Electrostatic forces were calculated using reaction field with a cutoff distance of 12 Å.

Simulations for each fragment were repeated 50 times starting from the same initial coordinates but with different initial random velocities. Each of the fifty simulations per fragment represents one 10 ns trajectory or clone (Figure 1). Coordinates were written every 100 ps for a total of 101 snapshots for each trajectory including the starting structure.

2.4. *Derived Structural Metrics*

Helical radius was defined for each residue i as the distance from its Cα atom to the centroid defined by the Cα of residues i, i-1 and i-2 on chains A, B and C respectively. Hydrogen bond lengths were calculated as the distance between the amide hydrogen atom of glycine in one chain and the carbonyl oxygen atom of the X-position residue in an adjacent chain for each Gly–X–Y triplet. Since the large values observed in some structures suggest disruption of hydrogen bonding, this measure will be referred to as the H_N-O_C distance. Backbone dihedral angles were calculated by computing the three vectors v1, v2, and v3 between C-N-Cα-C for ϕ and N-Cα-C-N for ψ and then taking the arc tangent of |v2| v1 • [v2 x v3], [v1 x v3] • [v2 x v3].

Average interchain H_N-O_C distance, ϕ angle, and ψ angle were calculated over all simulations, using ~7,500 structures from the last 5 ns of each 10 ns simulation. Autocorrelations were computed using the mean values over each of the three chains computed as a function of residue offset.

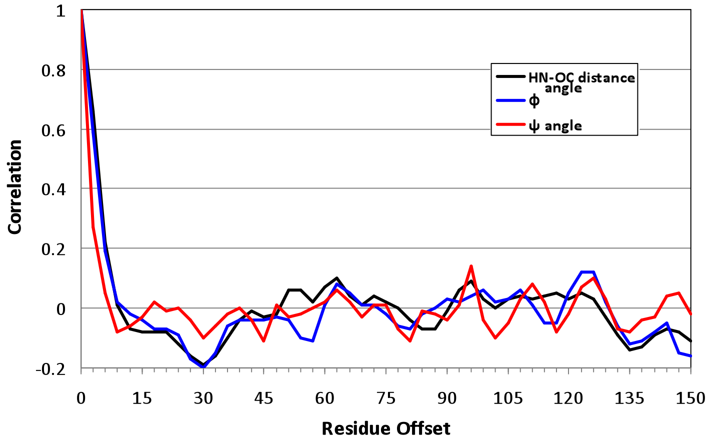


Fig. 2 Autocorrelation of glycine H_N-O_C distance, ϕ angle, and ψ angle as a function of residue offset. The maximum offset shown is 150 residues, which is a region spanned by approximately three simulated fragments.

3. Results

3.1. *Molecular Dynamics Simulation*

The complete triple helical region of heterotrimeric collagen type I, with 1014 amino acids per chain, was simulated for 10 ns. The simulation included ~1.8 million atoms, including solvent, and took approximately 11 months using the CPUs of over a quarter of a million computers. The calculation was parallelized by fragmenting the protein into 24 overlapping triple helical segments (Figure 1). Each fragment included 85 residues per chain, the shortest length needed to minimize end effect issues.[27] Fifty 10-ns trajectories (clones) were computed for each fragment, for a total of 1200 simulations.

The overlap in the fragments means that most residues occur in two different sets of 50 clones. To generate a composite set of structures for subsequent analysis in which each residue is represented once, for each residue we used only the simulation with that residue closer to the center of the fragment. This strategy eliminated potential artifacts that might have resulted from fraying at the ends of fragments, which had been observed previously in simulations of triple helical structures.[28] Angles at the overlap boundaries were calculated using residues from each side of the boundary.

To facilitate analysis and interpretation of the results, raw and summarized data were stored in a relational database along with biological annotations retrieved from the collagen database COLdb.[29] Stored data include 5,072,422 H_N-O_C distances, 15,247,566 each of ϕ and ψ backbone dihedral angles, 15,217,266 helical radii, and 1,615 biological features. The biological data encompass information from multiple scales, including: (1) thermostability of Gly-X-Y triplets, (2) features of the procollagen trimer, such as experimentally observed folding domains, (3) fibril-level features, e.g. ligand interaction sites, (4) characteristics of assembled fibers, such as gaps and bands visible in electron micrographs, and (5) patient phenotypes associated with specific mutation sites. The database and other supplementary materials are available online at simtk.org.

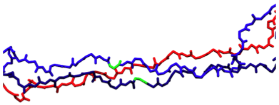
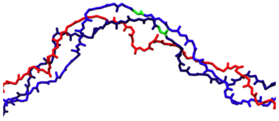
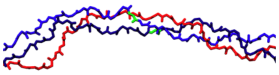
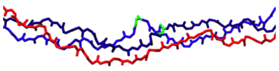
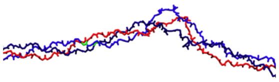
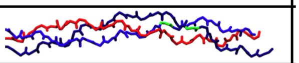
3.2. *Validation of the Simulated Structures*

Several tests were performed to ensure the validity of the resulting composite structures. First, visual inspection of the 1,200 trajectories confirmed the absence of artifacts resulting from use of periodic boundary conditions. Significantly, the solute did not directly interact with its periodic image. Second, analysis of the potential energies of the structures showed that the system was stable over the course of the simulation. The potential energy averaged over all simulations for all fragments equilibrated at approximately -980,000 KJ/mol. Third, the autocorrelations of H_N-O_C distance, ϕ angle, and ψ angle were calculated for each glycine to identify any periodicity resulting from fragmenting the protein at fixed locations (Figure 2). No significant peak was found at 42, suggesting that the fragmentation did not introduce periodicity artifacts. Interestingly, there is a small negative correlation between glycines separated by 30 residues. The source of this correlation is unknown since known repetitive features present in the structures, such as super-helical turns and typical fragment length, are either greater or less than 30 residues.

3.3. *Analysis of the Simulation Structures*

Visual inspection revealed that the composite structures generally maintained a triple helical conformation throughout their length. However, the triple helical structure was not uniform, with some regions more tightly wound on average than others. For example, Table 1 shows sample regions of type I collagen with disease-associated glycine mutations that differ in the position of the mutation on the α1 chain, whether multiple distinct substitutions have been observed at that position, and the location in the triple helix with respect to known ligand-binding sites.

Table 1. Sample simulation structures of regions of native collagen encompassing observed glycine mutation sites in the α1(I) chain. The mutation sites are highlighted in green.

Chain Location	***Gly ⇒ Mutation***	***Structural Description around mutation***	***Representative Model of ~42 residues near mutation***
106	Ala, Ser	Unstable region; very loose region	
220	Ala, Asp, Cys	Moderately loose region; bending observed	
382	Arg, Cys, Ser	Normal region with mild variation; helix loosens and tightens	
415	Cys, Asp, Ser, Ser	Very tight region with little variation; region becomes unstable near residue 437	
844	Ser, Val	Unstable region; very loose region with high variation (842-858); tight helix C-terminus to residue 858	
1009	Ser, Val	Very tight region with little variation	

The conformations were characterized by mean ϕ and ψ angles (Figure 3). The average ϕ angle ranged between -66.1° and -87.6° and the average ψ angle between 151.5° and 160°. We further characterized the structural variation by computing three additional measures: H_N-O_C distance between glycines on adjacent chains, the number of residues per turn, and helical radius. Interchain H_N-O_C distance averaged 2.6 Å, with a minimum of 1.5 Å. Average helix radius measured at glycine positions ranged from 2.8 Å to 7.3 Å (Figure 3). Visual inspection of the structures revealed that the sharp peaks in the figure correspond to regions of unwinding of the triple helix. The number of residues per turn ranged from 4.0 to 5.1, with an average of 4.9. The number of residues per turn inversely correlated with helix radius and H_N-O_C distance, and deviations from the average angles corresponded to increased helix radius and hydrogen bond length. Detailed data for these four measurement types are available for download at simtk.org. Although these measures were uniform in the ideal starting conformation, they vary along the length of the triple helix as a result of the simulation, in agreement with results from experimental studies revealing local variations in helical twist.[30, 31]

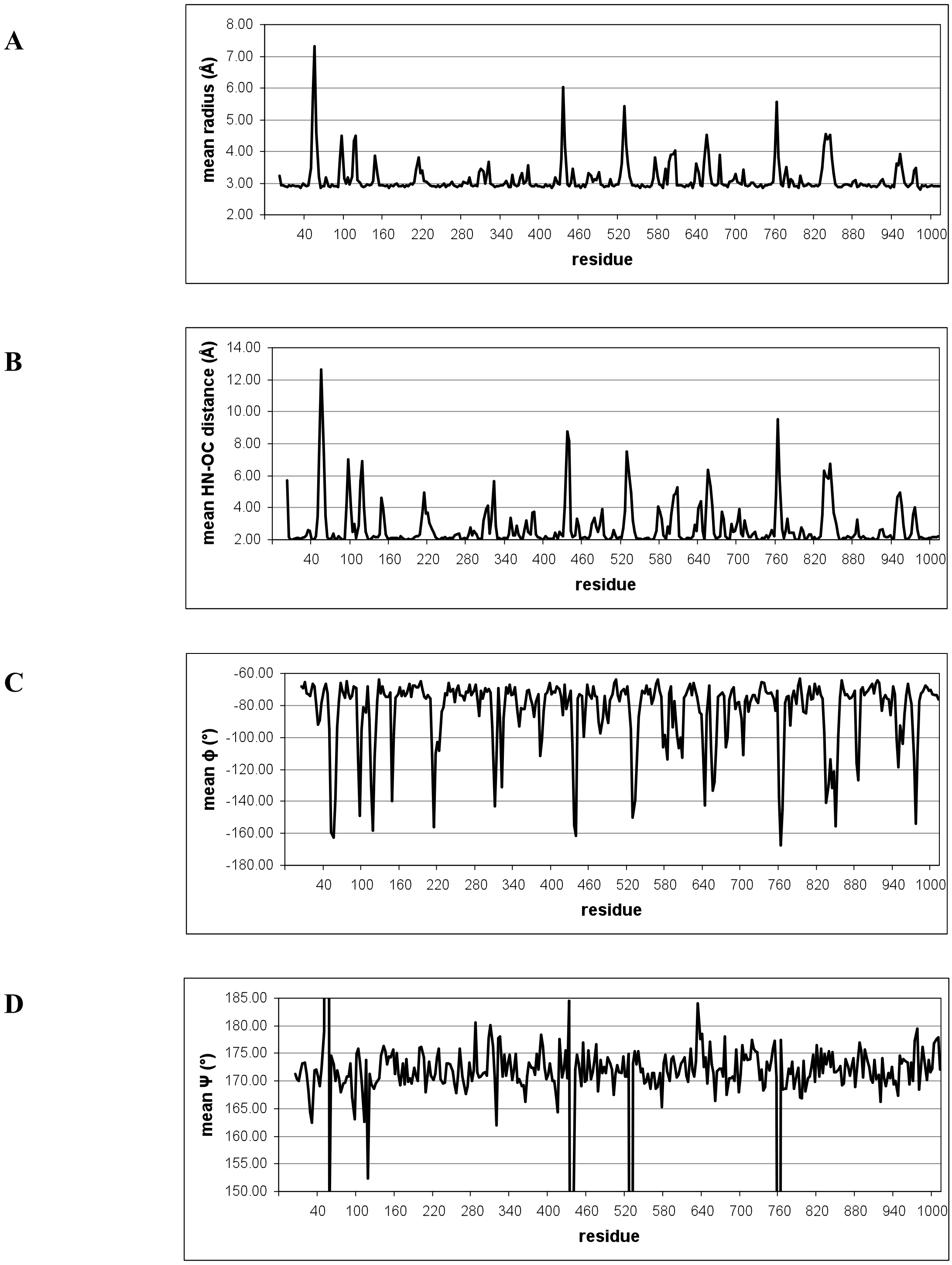


Fig. 3. Average properties at each glycine position from the second half of the simulation. (A) radius. (B) H_N-O_C distance. Long distances indicate the loss of the hydrogen bond in at least a subset of structures (next page). (C) ϕ angle. (D) ψ angle.

Table 2 contrasts ϕ and ψ angles calculated from our simulations with the corresponding angles calculated from crystal structures of model collagen peptides. The angles are organized based on the position of the residue within a triplet (Gly, X, or Y) and by their residue's type (imino, meaning the triplet contains a proline or hydroxyproline, or amino, meaning the triplet contains no proline or hydroxyproline). For most angles (10/12), the crystal structure values are well within ten degrees of the simulation values, and were not examined in more detail. The two most interesting values are for the ϕ angles of the residues in the Gly and X locations, where the values differ by 13 and 16 degrees between the simulation and the crystal structure. These differences correspond to opening of the canonical triple helical structure.

Table 2. Φ and Ψ angles for Amino Residue Pairs and Imino Residue Pairs

Angle	Residue Position	Residue Type	Crystal Structure[a]	Simulation	Difference
ϕ	Gly	Amino Pair	**-68**	**-81**	**-13**
		Imino Pair	-72	-74	-2
	X	Amino Pair	**-71**	**-87**	**-16**
		Imino Pair	-74	-67	7
	Y	Amino Pair	-66	-69	-3
		Imino Pair	-60	-55	5
ψ	Gly	Amino Pair	167	173	6
		Imino Pair	176	172	-4
	X	Amino Pair	160	155	-5
		Imino Pair	163	156	-7
	Y	Amino Pair	148	146	-2
		Imino Pair	152	151	-1

Table 3 shows details for the ϕ angles for residues in the Gly and X positions, based on the type of residue occupying the Y position. The residues in the Y position are split into two groups. The first group consists of all γ-branched residues (Asp, Phe, His, Leu, Asn) plus glutamic acid and lysine (henceforth identified using their concatenated one-letter-codes, "DEFHKLN"). The second group consists of all other observed residues that are not part of the first group (identified using their one letter codes: "AGIMOQRSTV"). The ϕ angles for the DEFHKLN group show significant deviation from the values derived from the crystal structures.

[a] From Rainey and Goh [32]

Table 3. Φ angle for Helical and Helix-Breaking Residues

Angle	Residue Position	Y Residue Type	Crystal Structure[b]	Simulation	Difference
ϕ	Gly	D,E,F,H,K,L, or N	**-68**	**-87**	**-19**
		A,G,I,M,O,Q,R,S,T, or V	-68	-77	-9
	X	D,E,F,H,K,L, or N	**-71**	**-88**	**-17**
		A,G,I,M,O,Q,R,S,T, or V	-71	-76	-5

Figure 4 shows the frequency of observed ϕ angles for all residues in the Gly position, for each observed residue type in the Y position. The seven dotted lines indicate residues types that are part of the DEFHKLN group described above. The other ten lines represent members of the group that shows a preference for triple helices (the AGIMOQRSTV group). Recall from Table 2 that the mean ϕ angle for non-imino residues in the Gly position is -68°.

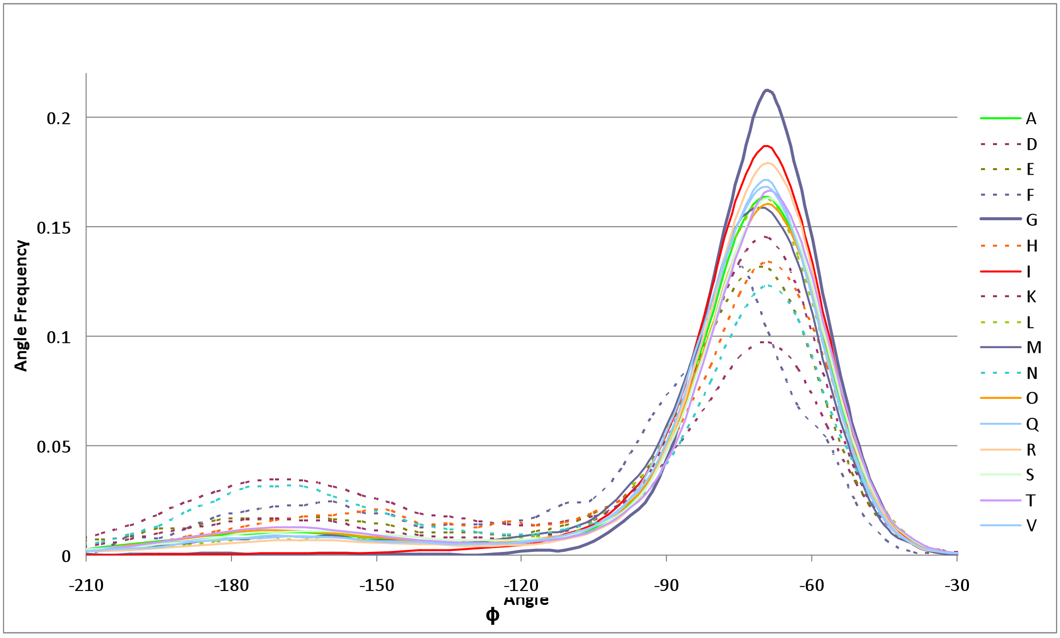


Fig. 4. Frequency of glycine ϕ angles graphed by the residue in the Y position.

Figure 5 is similar to Figure 4, but shows the frequency of observed ϕ angles for all residues in the X position. The mean ϕ angle for non-imino residues in the X position is -71°.

[b] From Rainey and Goh [32]

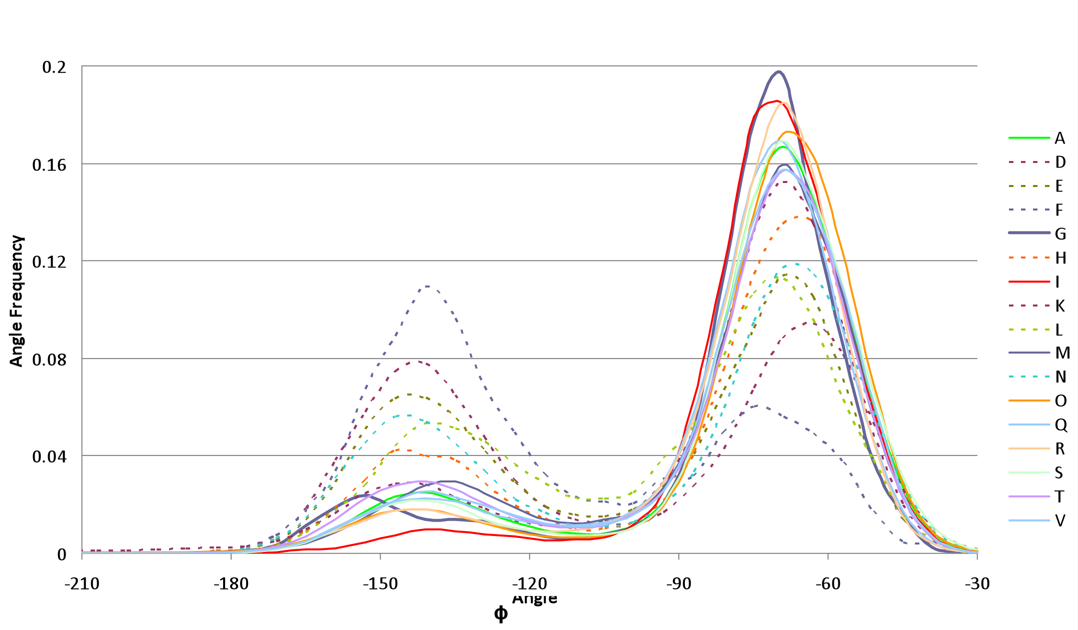


Fig. 5. Frequency of observed ϕ angles of X-position amino acids.

4. Discussion

We have conducted a 10 ns molecular dynamics simulation of the complete triple helical domain of human type I collagen as a first step towards understanding the dynamics of this critical region, its sequence dependence, and its relationship to higher order forms of the protein. Due to the large size of the protein, the calculation was made feasible by splitting the protein into overlapping fragments and the analyses were facilitated by storing the results in a customized relational database.

The simulated molecules capture structural variation in the triple helical domain consistent with experimental structural and biophysical studies and the atomic resolution of the simulations enables more precise definition of the region boundaries. Residues 1-85, identified by Makareeva, et. al.,[6] as the high stability N-anchor region, have short radius and H_N-O_C distance in our simulations. However, this tightly wound region is interrupted at residue 55 in our simulations, where we see an average radius of 7.3 Å, the highest in the protein (Figure 3). The molecular basis for the implied decrease in stability is unknown but may be due to electrostatic repulsion in a cluster of Asp residues at positions 53 and 54 in α1(I) and position 54 in α2(I). The results also suggest that the high stability region may extend to glycine 91, and that the proposed adjoining microunfolding region, captured by long H_N-O_C distances, may span glycine residues 94 - 121. The next two areas of largest unfolding in our simulations, with maximum unwinding at glycines 436 and 763, are contained within Makareeva, et. al.'s mid-flex and C-flex low stability regions. These residues may represent positions at which unfolding initiates in these areas.

Tables 2 and 3 show how dynamic, solvated collagen molecules differ from model collagen crystal structures examined by Rainey and Goh.[32] It is interesting that there is little difference between the ϕ and ψ angles shown in Table 2, despite the difference in environment and the difference in sequence. The only significant difference is in the ϕ angles for residues in the Gly and X positions that are part of triplets containing no proline or hydroxyproline residues. Data mining showed a strong correlation between these ϕ angles and the type of residue occupying the Y position of the triplet (Table 3). Five of the residues in the DEFHKLN group are γ-branched (Asp, Phe, His, Leu, Asn), which are known to be destabilizing in model collagen systems,[33] so their connection with non-helical structures is not surprising. The reason why glutamic acid and lysine also correlate with non-helical structures is under investigation.

Although the calculated structures are similar to experimentally determined structures of collagen-like peptides, the simulation did not reproduce the microfibril conformation of the triple helices observed in the crystal structure conformation of rat tail tendon collagen.[15] This is not surprising due to differences between the simulated and crystallized collagens in: primary sequence, scale (trimer and fibril), environment (water vs in situ), tissue source, and post-translational modifications. This is consistent with there being important structural differences between isolated (solvated) heterotrimers and the trimers in the more complex fibril structure, in which the collagen proteins are closely packed and associated with proteoglycans and other factors.

The complexity of the hierarchical conformations of collagen has made it difficult to determine experimentally the structure of native collagen at high resolution, and the large size of the protein has previously prohibited full atomic modeling of its structure and dynamics. We were able to accomplish molecular dynamics simulation of the full-length triple helix through technological improvements dependent on the accessibility of hundreds of thousands of computers. The resulting models are an important starting point for investigating the unique hierarchical conformations of collagen and for studying the effects of disease-associated mutations on collagen structure.

5. Acknowledgments

This work was supported by NIH grant AR051582 (to TEK) and the Osteogenesis Imperfecta Foundation and the Children's Brittle Bone Foundation (to DLB). We thank Sanghyun Park and Vijay Pande for running the calculations on Folding@Home, the Folding@Home folders for donating CPU time, and Simbios for simtk.org (GM072970).

References

1. P. H. Byers, *Online Metabolic and Molecular Bases of Inherited Disease* **22**, 5241 (2001).
2. T. J. Wess, *Adv Protein Chem* **70**, 341 (2005).
3. C. M. Kielty and M. E. Grant in *Connective Tissue and Its Heritable Disorders*, P. M. Royce and B. Steinmann, Editors.Wiley-Liss 159 (2002).
4. W. V. Arnold, A. Fertala, A. L. Sieron, H. Hattori, D. Mechling, H. P. Bachinger and D. J. Prockop, *J Biol Chem* **273**, 31822 (1998).
5. E. Makareeva, W. A. Cabral, J. C. Marini and S. Leikin, *J Biol Chem* **281**, 6463 (2006).

6. E. Makareeva, E. L. Mertz, N. V. Kuznetsova, M. B. Sutter, A. M. DeRidder, W. A. Cabral, A. M. Barnes, D. J. McBride, J. C. Marini and S. Leikin, *J Biol Chem* **283**, 4787 (2008).

7. A. Steplewski, I. Majsterek, E. McAdams, E. Rucker, R. J. Brittingham, H. Ito, K. Hirai, E. Adachi, S. A. Jimenez and A. Fertala, *J Mol Biol* **338**, 989 (2004).

8. D. L. Bodian, B. Madhan, B. Brodsky and T. E. Klein, *Biochemistry* **47**, 5424 (2008).

9. G. A. Di Lullo, S. M. Sweeney, J. Korkko, L. Ala-Kokko and J. D. San Antonio, *J Biol Chem* **277**, 4223 (2002).

10. A. Gautieri, S. Uzel, S. Vesentini, A. Redaelli and M. J. Buehler, *Biophys J* **97**, 857 (2009).

11. S. G. Uzel and M. J. Buehler, *Integr Biol (Camb)* **1**, 452 (2009).

12. H. P. Bachinger, N. P. Morris and J. M. Davis, *Am J Med Genet* **45**, 152 (1993).

13. J. C. Marini, A. Forlino, W. A. Cabral, A. M. Barnes, J. D. San Antonio, S. Milgrom, J. C. Hyland, J. Korkko, D. J. Prockop, A. De Paepe, P. Coucke, S. Symoens, F. H. Glorieux, P. J. Roughley, A. M. Lund, K. Kuurila-Svahn, H. Hartikka, D. H. Cohn, D. Krakow, M. Mottes, U. Schwarze, D. Chen, K. Yang, C. Kuslich, J. Troendle, R. Dalgleish and P. H. Byers, *Hum Mutat* **28**, 209 (2007).

14. B. Brodsky and A. V. Persikov, *Adv Protein Chem* **70**, 301 (2005).

15. J. P. Orgel, T. C. Irving, A. Miller and T. J. Wess, *Proc Natl Acad Sci U S A* **103**, 9001 (2006).

16. R. J. Radmer and T. E. Klein, *Biochemistry* **43**, 5314 (2004).

17. R. J. Radmer and T. E. Klein, *Biophys J* **90**, 578 (2006).

18. E. Lindahl, B. Hess and D. van der Spoel, *J Mol Model* **7**, 306 (2001).

19. http://folding.stanford.edu/

20. J. Wang, P. Cieplak and P. A. Kollman, *J Comput Chem* **21**, 1049 (2000).

21. S. Park, R. J. Radmer, T. E. Klein and V. S. Pande, *J Comput Chem* **26**, 1612 (2005).

22. C. C. Huang, G. S. Couch, E. F. Pettersen, T. E. Ferrin, A. E. Howard and T. E. Klein, *Pac Symp Biocomput* 349 (1998).

23. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J Chem Phys* **79**, 926 (1983).

24. D. J. Evans and B. L. Holian, *J Chem Phys* **83**, 4069 (1985).

25. H. J. Berendsen, J. P. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J Chem Phys* **81**, 3684 (1984).

26. B. Hess, H. B. Herman, J. C. B. Johannes and G. E. M. Fraaije, *J Comput Chem* **18**, 1463 (1997).

27. R. J. Radmer and T. E. Klein, unpublished results.

28. S. Park, T. E. Klein and V. S. Pande, *Biophys J* **93**, 4108 (2007).

29. D. L. Bodian and T. E. Klein, *Hum Mutat* **30**, 946 (2009).

30. J. Emsley, C. G. Knight, R. W. Farndale and M. J. Barnes, *J Mol Biol* **335**, 1019 (2004).

31. R. Z. Kramer, J. Bella, P. Mayville, B. Brodsky and H. M. Berman, *Nat Struct Biol* **6**, 454 (1999).

32. J. K. Rainey and M. C. Goh, *Protein Sci* **11**, 2748 (2002).

33. A. V. Persikov, J. A. Ramshaw, A. Kirkpatrick and B. Brodsky, *Biochemistry* **39**, 14960 (2000).

STRUCTURAL INSIGHTS INTO PRE-TRANSLOCATION RIBOSOME MOTIONS

SAMUEL COULBOURN FLORES
Bioengineering Department, Stanford University, James H Clark Center S172 MC:5448
Stanford, California 94305, USA
Email: samuelflorec@gmail.com

RUSS ALTMAN
Bioengineering Department, Stanford University, James H Clark Center S231 MC:5444
Stanford, California 94305, USA
Email: russ.altman@stanford.edu

Subsequent to the peptidyl transfer step of the translation elongation cycle, the initially formed pre-translocation ribosome, which we refer to here as R_1, undergoes a ratchet-like intersubunit rotation in order to sample a rotated conformation, referred to here as R_F, that is an obligatory intermediate in the translocation of tRNAs and mRNA through the ribosome during the translocation step of the translation elongation cycle. R_F and the R_1 to R_F transition are currently the subject of intense research, driven in part by the potential for developing novel antibiotics which trap R_F or confound the R_1 to R_F transition. Currently lacking a 3D atomic structure of the R_F endpoint of the transition, as well as a preliminary conformational trajectory connecting R_1 and R_F, the dynamics of the mechanistically crucial R_1 to R_F transition remain elusive. The current literature reports fitting of only a few ribosomal RNA (rRNA) and ribosomal protein (r-protein) components into cryogenic electron microscopy (cryo-EM) reconstructions of the *Escherichia coli* ribosome in R_F. In this work we now fit the entire *Thermus thermophilus* 16S and 23S rRNAs and most of the remaining *T. thermophilus* r-proteins into a cryo-EM reconstruction of the *E. coli* ribosome in RF in order to build an almost complete model of the *T. thermophilus* ribosome in R_F thus allowing a more detailed view of this crucial conformation. The resulting model validates key predictions from the published literature; in particular it recovers intersubunit bridges known to be maintained throughout the R_1 to R_F transition and results in new intersubunit bridges that are predicted to exist only in R_F. In addition, we use a recently reported *E. coli* ribosome structure, apparently trapped in an intermediate state along the R_1 to R_F transition pathway, referred to here as R_2, as a guide to generate a *T. thermophilus* ribosome in the R_2 state. This demonstrates a multiresolution method for morphing large complexes and provides us with a structural model of R_2 in the species of interest. The generated structural models form the basis for probing the motion of the deacylated tRNA bound at the peptidyl-tRNA binding site (P site) of the pre-translocation ribosome as it moves from its so-called classical P/P configuration to its so-called hybrid P/E configuration as part of the R_1 to R_F transition. We create a dynamic model of this process which provides structural insights into the functional significance of R_2 as well as detailed atomic information to guide the design of further experiments. The results suggest extensibility to other steps of protein synthesis as well as to spatially larger systems.

1. Introduction

The structure of the ribosome is surprisingly well conserved across the kingdoms of life and is thus biologically interesting for what its structure and, potentially, dynamics tell us about evolution. Multiple structures of bacterial (mostly *Escherichia coli* and *Thermus thermophilus*) ribosomes in complex with their tRNA substrates have been solved crystallographically [1], while others have been solved at low resolution by cryogenic electron microscopy (cryo-EM) [2]. However, the motions connecting these static conformational states of the ribosome and its bound tRNA substrates are currently the subject of intense research. In order to provide initial insight and focus further experiments, it would be useful to have at least a preliminary trajectory of motion connecting all of the states encompassing a cycle of translation, generated initially by flexible alignment or morphing. For this, we must translate all crystallographic structures into a single species, and fit all-atom models into cryo-EM density maps, and then connect the states by flexible alignment. We show how the recently announced RNABuilder modeling code and other tools can be used to accomplish this. We focus on the conformational changes of the ribosome as it undergoes a critical ratchet-like intersubunit rotation and on the associated reconfiguration of the deacylated tRNA bound at the peptidyl-tRNA binding site (P site) of the ribosomal pre-translocation complex from its so-called classical P/P configuration to its so-called hybrid P/E configuration. We model these conformational changes by a flexible alignment to ribosomal structures in three presumably sequential conformational states. The resulting dynamic model highlights important phenomena and provides a structural basis for the design of further experiments.

2. Background

2.1. *Multi-resolution modeling at the mesoscale*

The ribosome, by its sheer size, challenges conventional computational techniques and calls for new approaches. Multi-resolution modeling (MRM) refers to the treatment of different molecules, domains, spatial regions, or time spans in a system at different levels of resolution, either from the force field or kinematic perspective [3]. A wide variety of techniques fall under this paradigm. Some workers, for example, treat lipids and water using a reduced set of pseudoatoms, while modeling a protein inserted in the membrane at full-atomic resolution. Others collect statistics from experiments or short-time Molecular Dynamics (MD) simulations at fine resolution, and use these statistics to parameterize a coarse-grained force field, thus separating the resolutions in time. Still others run a fine and a coarse grained simulation simultaneously for the same system and exchange resolution from time to time, in an approach known as Resolution Exchange [3].

Mesoscale modeling refers to the structural and dynamic study of phenomena at length scales between those of single molecules (which can be modeled with MD and related methods) and those of extended tissues (which may be best modeled using continuum mechanics approaches) [4]. Some examples of this are actin filament elongation, muscle function, chromatin remodeling, and mitosis.

Internal Coordinate Mechanics (ICM) refers to a calculation in which bodies are connected to other bodies using *mobilizers* which may grant zero to six degrees of freedom [5]. In ICM, computer time is only spent computing the degrees of freedom granted by the mobilizers, whereas in MD reducing the degrees of freedom actually *increases* expense by adding constraint equations which must then be solved. The mobilizers connect bodies in a tree structure, and the calculation of position, velocity, and acceleration begins at the base body and proceeds up the tree. In an ICM framework, bodies may consist of one or more atoms, and the connections between atoms may permit dihedral angle changes but leave bond lengths and angles fixed. This characteristic makes ICM an ideal approach to MRM, because arbitrary domains, molecules, and complexes which are converged or uninteresting may be rigidified for economy and other modeling goals. Note that so far we have only described coarse graining the *kinematics;* the atoms within rigidified regions can still have force interactions with atoms in other regions.

RNABuilder is an ICM, MRM code which allows the user to instantiate molecules, match their structural coordinates to input files, control their flexibility, and apply base pairing, steric, and other atomic interaction forces. The user can further control most simulation parameters, including temperature, run time, output frequency, and type of time integrator and thermostat. A wide variety of applications are possible, such as structure prediction from base pairing contacts [6], refinement, threading [7], and flexible alignment. We have instantiated RNA chains as long as 13000 residues [5]; we have also shown that computer time can have order-N scaling with molecule size. This suggests that the presented methods are applicable to problems that approach the mesoscale.

2.2. *Progress and limitations in ribosome structure*

The past decade has seen an explosion of discoveries in ribosome structure. The state of the pre-translocation ribosomal complex that is sampled immediately following peptidyl transfer, which we refer to here as R_I, following the nomenclature recently introduced by Cate and co-workers[8], is particularly well characterized[9]. The structures of a number of other states have also been solved, if not crystallographically[1] then by cryo-EM [2]. The full story of ribosome function, however, involves the structural dynamics connecting the various observed states. Knowing these details will lead to a more fundamental understanding of protein synthesis across all kingdoms of life and promises to guide the development of novel antibiotics which function by confounding dynamics crucial to function. As a future goal, we wish to generate an all-atom trajectory of the entire translation cycle. For this we face two challenges: first, structures of the various available states have been solved using ribosomes and biomolecular components isolated from a variety of species, mostly *E. coli* and *T. thermophilus,* and second, some of these structures have been solved by cryo-EM and are available only as electron density maps. In this work we show how to address both these issues with the help of RNABuilder and other packages.

2.3. *Structure of the ribosome in the fully rotated R_F conformation*

Immediately following peptidyl transfer, the pre-translocation ribosome, initially in R_I, undergoes a ratchet-like intersubunit rotation in order to sample a rotated conformation, referred to here as R_F, that is an obligatory intermediate in the translocation of tRNAs and mRNA through the ribosome during the translocation step of the translation elongation cycle. Unlike R_I, atomic resolution structures of R_F remain elusive and, although considerable structural insight has come from cryo-EM reconstructions of R_F, these workers typically only published fits of those ribosomal and/or tRNA components and/or fragments of components needed to answer specific structural questions. As a consequence, a published and widely available all-atom model of R_F, based on the fitting of atomic resolution structures to cryo-EM reconstructions of R_F, is not currently available. In this work we report an all-atom model of R_F built by fitting atomic resolution structures to cryo-EM reconstructions of R_F [2]. This provides structural insights not obvious from inspection of the raw density of the reconstruction and promises to be useful for understanding the details of intersubunit interactions in the context of biochemical experiments[10].

2.4. *Structure of the ribosome in the intermediate R_2 state, and the R_1-R_2-R_F trajectory*

A recent structure of the ribosome in complex with P and A site-bound anticodon stem-loops (ASLs) rather than full-length tRNAs reveals a ribosome conformation that is apparently in an intermediate state of intersubunit rotation, referred to here as R_2, that lies somewhere between R_1 and R_F [8]. Interestingly, the ASLs in R_2 are positioned in a way that suggests that full-length tRNAs, should they have been present in the ribosomal complex that was crystallized, would be in their hybrid configurations. Nevertheless, the lack of full-length tRNAs in the ribosomal complex used to solve the R_2 structure and the associated lack of tRNA-ribosome interactions that would ordinarily be made between full-length tRNAs and the ribosome in R_2, leave open the question of how full-length tRNAs would be positioned within R_2. The current work addresses this question by constructing putative trajectories of motion connecting R_1, R_2, and R_F. This gives us a dynamic view of the trajectory of motion from R_1 to R_F, rather than the more typical view gained by inspection of static structures.

3. Method

3.1. *Creating the T.thermophilus R_2 structure by morphing*

As mentioned above, the crystallographically observed R_2 structure is inconvenient for our purposes because it was solved using *E. coli* ribosomes, whereas we are working with *T. thermophilus* ribosomes. We solved this problem by morphing the *T. thermophilus* R_1 structure onto the *E. coli* R_2 structure, thus copying the R_2 conformation onto the *T. thermophilus* ribosome.

To generate this morph we flexibilized the neck region (which connects the head to the body domain) as well as the base of the beak domain on the small ribosomal subunit and the base of the L1 stalk domain of the large ribosomal subunit (see **Figure 1** for a map of ribosomal domains and structural figures). We restrained the P-site tRNA into the classical P/P configuration by enforcing two Watson-Crick base pairs between the aminoacyl acceptor stem of the tRNA and the 23 rRNA nucleotides comprising the so-called P loop within the P site of the peptidyl transferase center and enforcing three Watson-Crick base pairs between the anticodon stem-loop of the tRNA and the mRNA codon. Likewise, we restrained the E-site tRNA into the classical E/E configuration using a base stacking sandwich involving the tRNA 5' terminal residue, and by a Weld constraint to an apparent tRNA-binding domain on ribosomal protein S7 (the tRNA binding domain was connected to the rest of S7 by a flexible hinge). The majority of ribosomal proteins were rigid and fixed to the corresponding domain on the 23S or 16S rRNA. The system also included the 16S and 23S rRNA as well as the tRNAs from the *E.coli* R_2 state; all of these were rigid and fixed to ground. Corresponding residues on 23S and 16S rRNA were then pulled together, causing flexible alignment of these two subunits.

3.2. *Fitting atomic resolution structures to the electron density resulting from cryo-EM reconstructions of R_F*

R_F continues to be the subject of intense structural and dynamic research. This has been primarily driven by: (i) the availability of atomic resolutions structures of R_1; (ii) the possibility that, like R_1, R_F is a potential target for novel antibiotics; and (iii) the realization that a structural and dynamic understanding of R_F is necessary in order to fully understand the mechanism of the R_1 to R_F

transition. Despite the great importance of this state, however, atomic resolution structures have not yet been fitted into the electron density map resulting from all or even the majority of cryo-EM reconstructions of this conformational state of the pre-translocation ribosomal complex. As a result, Molecular Dynamics studies that may assist structure-based drug design have no structural point from which to begin. Interpolated trajectories which could begin to elucidate the path of tRNA during the R_1 to R_F transition lack a crucial endpoint.

In published work, structural coordinates for Elongation Factor G, the hybrid P/E configured tRNA, and a small number of additional components were fitted to a electron density map of the *T. thermophilus* R_F (EMD-1315)[2]. In this work, we add the 16S, 23S, and 5S rRNAs and most of the r-proteins.

Our main task was to fit an existing all-atom *probe* to a *target* electron density map. The most general methods available allow all bonds in the probe to vary in length, angle, and dihedral during the course of the fitting. One such method is Molecular Dynamics Flexible Fitting [11, 12]. However such methods are expensive, typically requiring parallel computers and significant run time. A popular alternative approach consists of three steps as follows:

1) Begin with rigid-body fitting, under which the probe has *no* flexibility, and the entire molecule is fitted to the electron density of the target using only rigid body rotations and translations.
2) If the molecule exhibits domain motions much of the flexibility can be recovered by breaking up the model into multiple fragments and adjusting each fragment separately into the electron density map. The natural boundaries between such rigid fragments are the hinge points connecting rigid domains; multiple experiments and calculations have been done to locate these hinges in the ribosome as we will explain below.
3) As a final step, anneal the gaps between fitted fragments belonging to a single RNA or protein chain; we will describe how the last structure of the R_2 to R_F motion provides this.

For step 1, we used SITUS COLORES to do the initial rigid body fitting of the entire ribosome[13, 14]. The next step, in which the probe is divided into fragments for adjustment into the map, requires a selection of hinge points. Fortunately this topic has been well studied. The rRNA-based neck domain of the small ribosomal subunit, which connects the head to the body/platform domains of the small ribosomal subunit, has long been known to be flexible[8]. Likewise, Noller and coworkers have found that by Translation-Libration-Screw Motion Determination (TLSMD) that the beak domain of the small ribosomal subunit and the L1 stalk domain of the large ribosomal subunit have the highest displacements about their librational axes[15], indicating a hinge point at the base of each of these two domains. Thus we divided the *probe* into six rigid pieces corresponding to the body, head, and beak domains of the small ribosomal subunit (along with their attendant proteins), the L1 stalk domain of the large ribosomal subunit (along with r-protein L1), the remainder of the large ribosomal subunit, and the tRNA. Since the probe has already been rigidly fitted at this point, the fragments created as described are already close to their correct positions in the electron density map. Thus an exhaustive search is not required, only a local adjustment. For this, Chimera's[16] *Fit in Map* feature is useful.

In the last step of the fitting, we correct the unnatural bond geometries spanning the gaps between fragments. This is done by taking the last structure resulting from the R_2 to R_F motion, which conserved bond lengths and angles throughout.

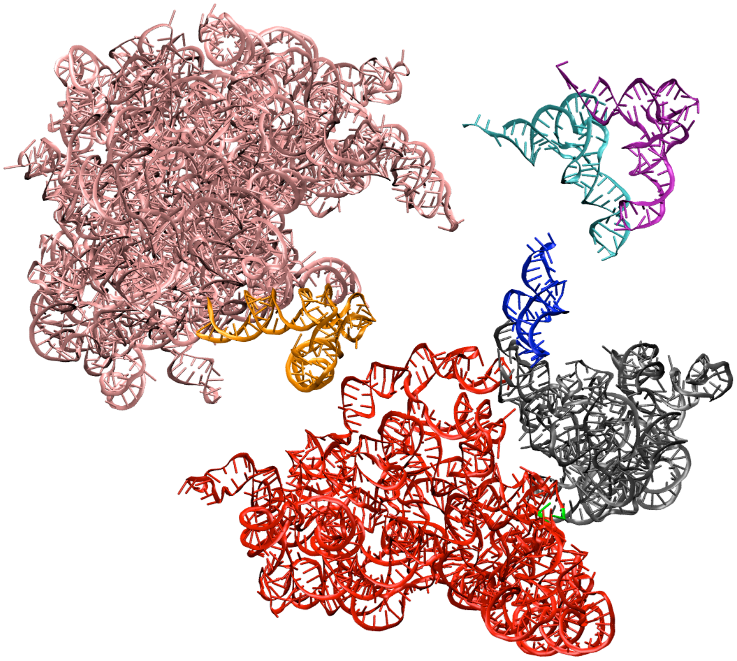


Figure 1. Choice of rigid domains for Cryo-EM fitting and creating *T.thermophilus* R2 structure (exploded view)

For fitting to Cryo-EM density: Following theoretical and experimental results described in the text, we broke up 16S into body (red), head (grey), and beak (blue). The "neck" connecting the head and body is shown in green for reference. The L1 stalk (orange) was separated from the rest of 23S (pink). The two tRNAs (cyan and purple) were independent bodies. Proteins (not shown) were attached rigidly to the corresponding 23S or 16S domain. The subunits were otherwise free to translate and rotate. For creating *T.thermophilus* R2 structure: The mentioned domains were left rigid as above, but instead of breaking 16S and 23S into fragments, we flexibilized hinge regions at the base of the L1 stalk, in the neck, and at the base of the beak.

3.3. *Generating the R_1-R_2-R_F trajectory*

In the last stage of our work we used a variation of our morphing technique to generate a controlled sequence of motions transforming the R_1 to R_2 to R_F state. We used the R_1, R_2 and R_F structures as fully-rigid, immobile templates much as before, but the aligned (or model) molecule was a ribosome which only one hinge – in the neck. The 23S was completely rigid and fixed to ground. tRNA was rigid (except for the 4 residues in the acceptor terminus) and could undergo rigid-body motion. 16S had a single flexible hinge in the neck region and could also undergo overall translation and rotation. 16S residue 1338 has been implicated in stabilizing the P/E site tRNA [10], therefore we connected this residue to tRNA residue 41 using a Sugar Edge / Sugar Edge interaction force, to approximately maintain an interaction observed in R_1, R_2, and R_F. We applied collision detecting spheres (to approximately represent steric repulsion) to the tRNA and to segments of 16S and 23S that might otherwise clash with tRNA (see Discussion).

We structurally aligned the ribosome model to R_1 to generate a starting point for our trajectory. We then aligned the model to R_2 and inspected the motion from R_1 to R_2. We then aligned the model to R_F. The trajectory of conformational change from R_1 to R_2 to R_F was the source of considerable insight as we will discuss.

Since it is not known whether full-length tRNAs in R_2 are in the classical or hybrid configuration, we generated an additional trajectory in which the tRNA remains in the classical state until 16S is fully rotated into the R_F conformation. The entire process is controlled with a single RNABuilder input file. The model is parametric in that global variables can be changed to easily generate any alternative ordering of these steps. Additional experimental information can be used to alter or constrain the motion, conversely the generated trajectory can help design focused experiments to generate further constraints.

4. Results

4.1. *Creating the T. thermophilus R_2 structure*

As mentioned a *T. thermophilus* ribosome in state R_1 was flexibly aligned to the *E.coli* ribosome in state R_2. We observed that as desired, proteins from R_1 were carried along with the RNA motion, and tRNAs moved in such a way as to maintain contact with their binding sites. The RMSD computed based on aligned glycosidic nitrogen atoms in 16S and 23S was initially 8.1Å and dropped to 2.9Å after 30 minutes of computer time. The degree of alignment can be qualitatively appreciated in (Figure 2). In a demonstration of convergence, we continued the calculation for an additional 127 minutes during which the RMSD remained nearly constant.

1.1. *Validation of the fitting of atomic resolution structures to the electron density from cryo-EM reconstructions of R_F*

In order to validate our fitting of atomic resolution structures into the electron density from cryo-EM reconstructions of R_F, we demonstrate that key molecular contacts expected or experimentally determined within R_F are recapitulated in our model. In particular, the detailed interactions of the tRNA aminoacyl-acceptor ends with the 23S rRNA within the peptidyl transferase center of the large ribosomal subunit are recovered (Figure 3).Also, the intersubunit bridges connecting the 23S rRNA with the body/platform of the 16S rRNA are recapitulated. Intersubunit bridge B4, which was predicted by Spahn (REF) and later Cate [17] to be maintained throughout the R_1 to R_F transition, indeed remains remains intact in our model (Figure 3).

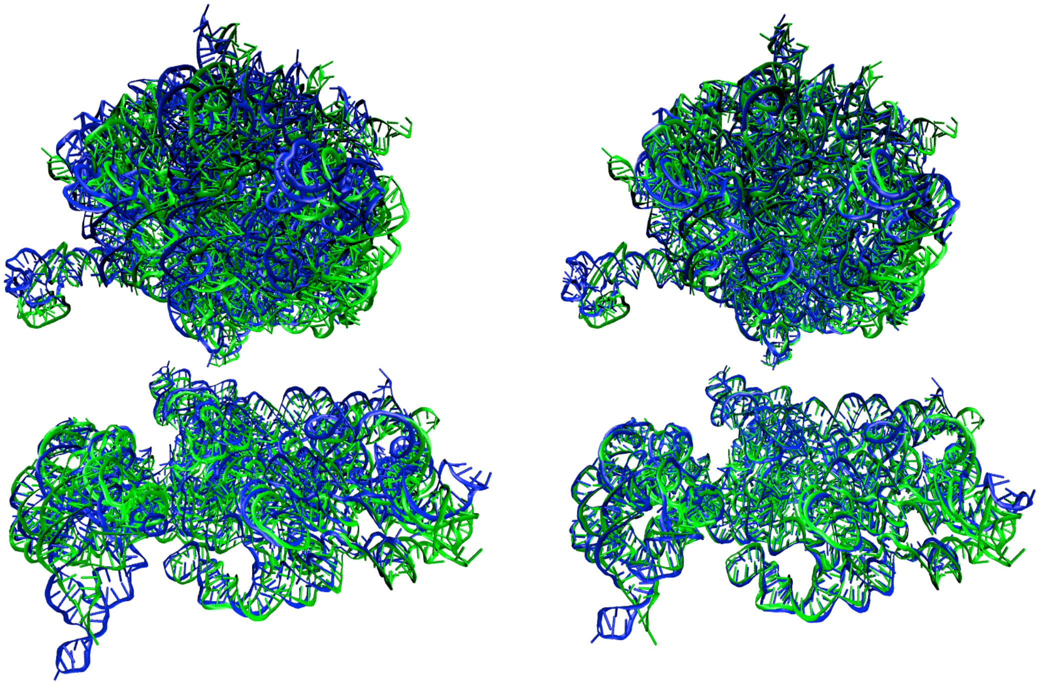


Figure 2. Morphing the *T.thermophilus* ribosome in state R1 onto the E.coli ribosome in state R2.

The *thermophilus* 23S (upper subunit) and 16S (lower subunit) are in blue. The E.coli 23S and 16S are in green. Additional *thermophilus* RNA and protein subunits are not shown. We included no E.coli subunits other than 23S and 16S. The thermophilus ribosome had hinges in the neck, base of the beak, and base of the L1 stalk. The E.coli ribosome was fully rigid and fixed to ground. tRNAs were attached to thermophilus P/P and E/E sites using base pairing and other forces and adjusted their positions as 16S and 23S moved. Thermophilus mRNA, 5S, and protein subunits were fully rigid and fixed to the corresponding 23S or 16S domain. Left panel: initial, rigid-body alignment. Note that blue and green are misaligned by as much as a helical diameter. Right panel: final alignment. Note that blue and green are now much more closely aligned. RMSD based on aligned glycosidic nitrogen atoms in 16S and 23S was initially 8.1Å (left) and converged to 2.9Å (right).

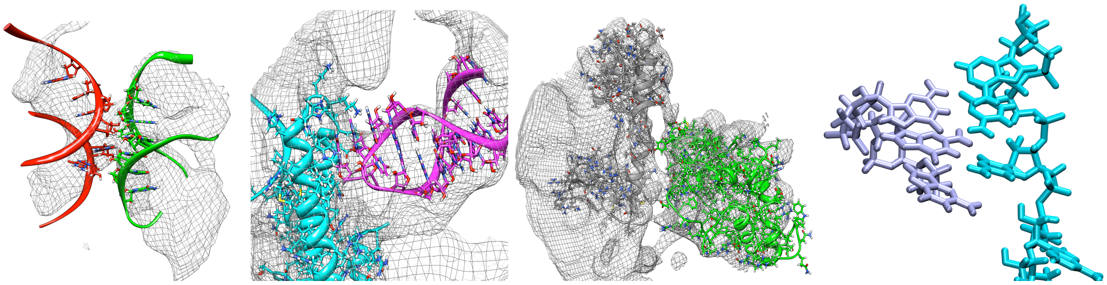


Figure 3. Intersubunit contacts.

The fitted RF model recapitulates bridge B3 (left panel). Bridge B4 was believed by earlier workers to remain in contact throughout ratcheting; the contact is maintained in our model (left-center panel). In bridge B1b/B1c (right-center panel), proteins S13 and L5 are connected by substantial regions of density, and the fitted proteins are in range to make contact. The A/P site tRNA acceptor (cyan, right panel) makes the correct base-pairing contacts with the 23S P site (purple).

4.2. *R1-R2-RF trajectory*

The trajectory of motion produced three main types of evidence which may be a source of insight. First, the residues to which we applied sterics (based on trial and error runs) may have functional importance. The aligned R_1, R_2 and R_F structures, even as static coordinates, may provide useful insight into the mechanism of motion. Our final trajectory as well as various alternative trial trajectories suggest constraints on the order and correlation of the motion of domains.

We applied sterics to H80/L80. Note that capital H indicates a helix in 23S, lowercase h indicates a 16S helix; similarly L/l indicates a loop and J/j indicates a junction in 23S/16S. Note also we are following Yusupov numbering for helices [18]. The H80 region contacted the acceptor terminus of the tRNA in some runs. In our model we left the acceptor terminus flexible so it simply sways out of the way. We also applied steric spheres to H69. When no sterics were applied here, the tRNA dropped down in some runs as the body moved away from 23S during 16S subunit rotation (Figure 4). We also applied sterics to the "gate" in 16S (h24, j29-42), beyond which the tRNA ASL should not pass (Figure 3). These and more possible interactions are listed in Table 1.

Subunit	Residues (our num.)	Residues (consensus numbering)	Region	Notes
tRNA	10-20 30-50 74-80	6-16 26-46 70-76	Inside elbow Anticodon stem-loop Acceptor terminus	Contacts H80/L80 Contacts 16S (gate and surroundings) Contacts 23S (various points)
23S	1829-1847 1781-1785 2135-2142 2316-2321	1907-1925 1850-1854 2252-2259 2433-2438	H69 H68 (int. loop) H80 H74	Contacts tRNA inside elbow Contacts tRNA acceptor terminus
16S	676-679 771-775 1208-1212 1318-1322 1474-1475	693-696 788-792 1227-1231 1337-1341 1497-1498	h23 h24 h30 j29-42 h44	 Part of gate Part of gate

Table 1 : Residues needing collision-detecting spheres to prevent steric clashes.

We used preliminary runs to determine which 16S and 23S residues tRNA would contact in its trajectory from P/P to P/E sites. We applied steric spheres to the interacting residues. The contact points are suggested for experimental validation.

We aligned the R_1 and R_2 crystallographic structures and our fitted R_F structure to each other based on 23S rRNA. We observed that all three have tRNA anticodons very near to each other (Figure 5). For this reason it was quite easy for us to generate a classical-state R_2 model without prohibitive steric clashes, while Jamie Cate's R_2 model is in the hybrid state. The classical-state R_2 structure, however, did not have the tRNA contact with 16S residue 1338, as we will discuss, whereas all three experimental structures exhibit the residue-1338 contact.

We generated R_1-R_2-R_F trajectories ordering the motion in various ways. We tried moving the head, body, and tRNA separately; while this was sterically possible, the residue-1338 contact had to be broken to do it. By moving the head and body from R_1 to R_2 without moving the tRNA, we generated a classical-state R_2 conformation, which of course did not have the 1338 contact. We found the results more credible when the head, body, and tRNA were moved at the same time, particularly since we were able to approximately maintain the contact using the base-pairing force mentioned between 1338 and residue 41. The reported final trajectory therefore has all three units moving together.

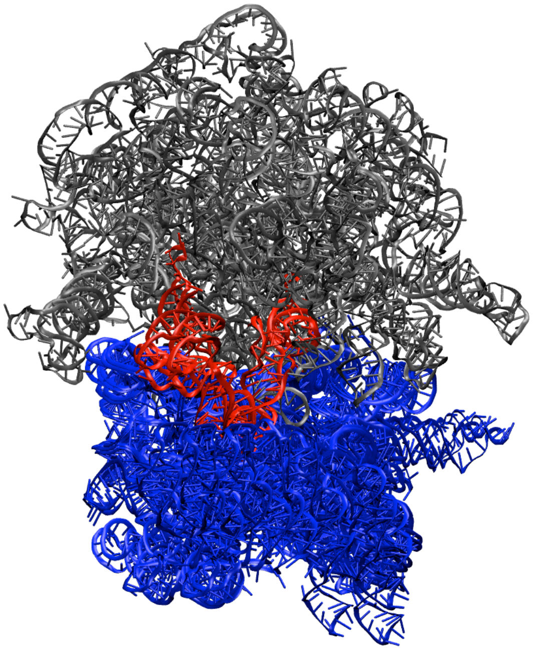


Figure 4. Putative motion from R1 to R2 to RF, with 23S fixed in space.

Composite view of the three main stages of the conformational rearrangement. The rightmost of the (red) tRNAs corresponds to the classical-state *T.thermophilus* R_1 structure (Protein Data Bank ID: 2J02 and 2J03), followed by the hybrid-state tRNA fitted by Jamie Cate to the R_2 structure (solved crystallographically in *E.Coli,* PDB ID: 3I1M & 3I1N, here modeled in *T.Thermophilus*). The leftmost tRNA was fitted by us to the CryoEM density map of the hybrid-state R_F structure (EMBL 3D-EM database ID EMD-1315). All three 16S conformations are shown superimposed (blue). The 23S subunit (grey) is kept rigid and stationary throughout. Note that the three anticodon stem-loops of the tRNAs are in nearly the same position throughout the motion, relative to 23S.

5. Discussion

In this work we modeled the trajectory of conformational change as the ribosome moves from the classical to the hybrid state, with an emphasis on the tRNA which moved from the classical P/P to the hybrid P/E configuration. Toward this end, we created a *T. thermophilus* ribosome in the R_2 state, as well as an all-atom fitted model of the ribosome in the R_F state; these are provided as supplementary materials and may be useful for further studies. The trajectory from R_1 to R_2 to R_F provides insight and suggests further experiments from three perspectives: the steric contacts, the aligned static structures, and the interpolated structures.

The regions where it was necessary to apply steric spheres are suggestive of specific experiments. In any of the contacting regions, one can imagine generating mutant ribosomes in which the steric barrier is either increased or eliminated. For instance: the part of the gate on the head domain is clearly important, since it includes residue 1338. What would happen if we eliminated the part that is on the body domain? We also noticed that H69 prevents the tRNA from moving down towards the mRNA. What would happen if it, too, disappeared? What about H80, which the acceptor terminus must brush past? What if the barrier were even higher? And residue 1338—is it really always in contact with tRNA? Could this be probed with a combination of mutagenesis and single-molecule fluorescence resonance energy transfer (smFRET) experiments?

Some interesting insight comes just from the static R_1, R_2, and R_F structures. First, after fitting all-atom subunit structures to the density map of R_F, we realized that the crucial 1338-41 contact is maintained in this state as well, thus validating the prediction of [10]. This was a key piece of evidence which encouraged us to enforce this contact throughout the motion. Second, aligning R_1,

R_2, and R_F structures based on 23S yielded an interesting finding—the P-site anticodons of the three structures are just a few angstroms apart. Thus the 16S P-site moves little with respect to 23S rRNA during the intersubunit rotation. This makes it quite easy to generate a clash-free model of R_2 with a tRNA in the classical P/P configuration. However, we believe that it is more likely for the tRNA to occupy the hybrid P/E configuration on the basis that this will maintain the 1338-41 base contact which is so clearly present in all three experimental static structures.

The trajectory itself is suggestive. First, maintaining the 1338-41 contact as we did implies a coordinated motion of the head and the tRNA. Since the anticodon does not move much, the motion is mostly a long translation of the aminoacyl acceptor end of the tRNA. Using this trajectory a donor/acceptor flurophore pair for smFRET experiments can be designed with one fluorophore attached to the tRNA and another to the head; we would predict that such a construct would exhibit a constant, static FRET value that is maintained throughout the R_1 to R_F transition. Second, we did not directly address the twisting motion of the neck, but the trajectory can be used to optimally place donor/acceptor fluorophore pairs spanning the head and body. Lastly, one can design a donor/acceptor fluorophore pair in order to determine how much time the ribosome spends in state R_2—perhaps with a donor/acceptor fluorophore pair spanning 23S rRNA and the body domain.

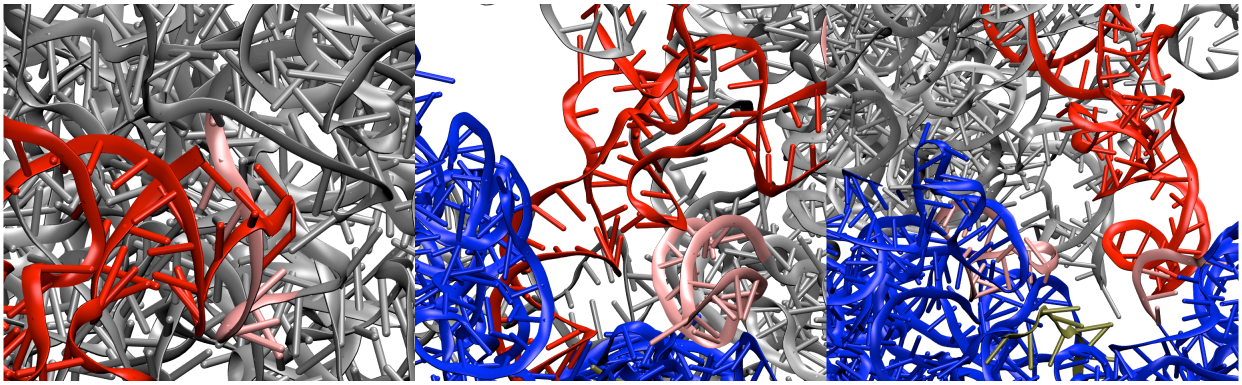


Figure 5. tRNA steric barriers.

Left panel: The acceptor terminus of tRNA (red) makes contact with H80/L80 (pink) of the 23S (grey). In nature tRNA may rotate, move away from 23S, or the floppy acceptor end may simply move out of the way. Middle panel: Contact between tRNA and H69 (pink) of 23S. In our model this forces tRNA to translocate laterally whereas without it tRNA might drop downwards to maintain contact with mRNA after body rocking. Right panel: Contact of the anticodon stem-loop with P-site barriers formed by h24 (pink) and the "gate" loop (residues 1335-1339) (also pink) in 16S (blue). tRNA cannot move fully into the P/E configuration until body and head have rotated.

Acknowledgements and availability

We thank Ruben Gonzalez for numerous discussions, guidance, and contributions to the text. We also thank Tom Cech and Martin Laurberg for illuminating discussions. Lastly, we thank Jamie Cate and Sean Connell for providing advice and unpublished structural coordinates. We provide starting coordinates, RNABuilder input files, the all-atoms fitted RF structure, the *T.thermophilus* R_2 structure, and the full trajectory for both R_1- R_2- R_F motions at https://simtk.org/home/ribosome-paper. RNABuilder can be downloaded at http://simtk.org/home/rnatoolbox.

References

1. Berk V, Cate JH: **Insights into protein biosynthesis from structures of bacterial ribosomes**. *Current opinion in structural biology* 2007, **17**(3):302-309.
2. Connell SR, Takemoto C, Wilson DN, Wang H, Murayama K, Terada T, Shirouzu M, Rost M, Schuler M, Giesebrecht J *et al*: **Structural basis for interaction of the ribosome with the switch regions of GTP-bound elongation factors**. *Molecular cell* 2007, **25**(5):751-764.
3. Ayton GS, Noid WG, Voth GA: **Multiscale modeling of biomolecular systems: in serial and in parallel**. *Current opinion in structural biology* 2007, **17**(2):192-198.
4. Glotzer SC PW: **Molecular and mesoscale simulation methods for polymer materials**. *Annu Rev Mater Res* 2002, **32**:401-436.
5. Flores S, Sherman, M, Bruns, C, Eastman, P, Altman, RB: **Fast flexible modeling of macromolecular structure using internal coordinates**. *IEEE Transactions in Computational Biology and Bioinformatics, submitted* 2010.
6. Flores S, Altman, RB: **Turning limited experimental information into 3D models of RNA** *RNA, submitted* 2010.
7. Flores S, Wan, Y, Russell, R, Altman, RB: **Predicting RNA structure by multiple template homology modeling**. *Proceedings of the Pacific Symposium on Biocomputing* 2010:216-227.
8. Zhang W, Dunkle JA, Cate JH: **Structures of the ribosome in intermediate states of ratcheting**. *Science (New York, NY* 2009, **325**(5943):1014-1017.
9. Selmer M, Dunham CM, Murphy FVt, Weixlbaumer A, Petry S, Kelley AC, Weir JR, Ramakrishnan V: **Structure of the 70S ribosome complexed with mRNA and tRNA**. *Science (New York, NY* 2006, **313**(5795):1935-1942.
10. Shoji S, Abdi NM, Bundschuh R, Fredrick K: **Contribution of ribosomal residues to P-site tRNA binding**. *Nucleic acids research* 2009, **37**(12):4033-4042.
11. Trabuco LG, Villa E, Mitra K, Frank J, Schulten K: **Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics**. *Structure* 2008, **16**(5):673-683.
12. Birmanns WWaS: **Using Situs for Flexible and Rigid-Body Fitting of Multiresolution Single-Molecule Data**. *Journal of Structural Biology* 2001, **133**:193–202.
13. Chacon P, Wriggers W: **Multi-resolution contour-based fitting of macromolecular structures**. *Journal of molecular biology* 2002, **317**(3):375-384.
14. Fabiola F, Chapman MS: **Fitting of high-resolution structures into electron microscopy reconstruction images**. *Structure* 2005, **13**(3):389-400.
15. Korostelev A, Noller HF: **Analysis of structural dynamics in the ribosome by TLS crystallographic refinement**. *Journal of molecular biology* 2007, **373**(4):1058-1070.
16. Eric F. Pettersen TDG, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, Thomas E. Ferrin: **UCSF Chimera—A Visualization System for Exploratory Research and Analysis**. *Journal of Computational Chemistry* 2004, **25**:1605–1612.
17. Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM, Cate JH: **Structures of the bacterial ribosome at 3.5 A resolution**. *Science (New York, NY* 2005, **310**(5749):827-834.
18. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF: **Crystal structure of the ribosome at 5.5 A resolution**. *Science (New York, NY* 2001, **292**(5518):883-896.

NEW CONFORMATIONAL SEARCH METHOD USING GENETIC ALGORITHM AND KNOT THEORY FOR PROTEINS

Y. SAKAE

Department of Physics, Nagoya University,
Nagoya, Aichi 464-8602, Japan
E-mail: sakae@tb.phys.nagoya-u.ac.jp

T. HIROYASU

Department of Biomedical Information, Doshisha University,
Kyotanabe, Kyoto 610-0394, Japan
E-mail: tomo@is.doshisha.ac.jp

M. MIKI

Department of Intelligent Information Engineering and Sciences, Doshisha University,
Kyotanabe, Kyoto 610-0394, Japan
E-mail: mmiki@mail.doshisha.ac.jp

Y. OKAMOTO

Department of Physics, Nagoya University,
Nagoya, Aichi 464-8602, Japan
and
Structural Biology Research Center, Nagoya University,
Nagoya, Aichi 464-8602, Japan
E-mail: okamoto@phys.nagoya-u.ac.jp

We have proposed a parallel simulated annealing using genetic crossover as one of powerful conformational search methods, in order to find the global minimum energy structures for protein systems. The simulated annealing using genetic crossover method, which incorporates the attractive features of the simulated annealing and the genetic algorithm, is useful for finding a minimum potential energy conformation of protein systems. However, when we perform simulations by using this method, we often find obviously unnatural stable conformations, which have "knots" of a string of an amino-acid sequence. Therefore, we combined knot theory with our simulated annealing using genetic crossover method in order to avoid the knot conformations from the conformational search space. We applied this improved method to protein G, which has 56 amino acids. As the result, we could perform the simulations, which avoid knot conformations.

Keywords: Molecular Simulation; Simulated Annealing; Protein Folding; Genetic Algorithm; Knot Theory

1. Introduction

Computational simulations of biomolecular systems such as proteins and DNA are performed using molecular simulation techniques such as Monte Carlo (MC) and molecular dynamics (MD) methods. However, as the biomolecular system has a large number of degrees of freedom associated with a lot of atoms and is characterized by many local minima separated by high energy barriers, it is not yet possible to perform enough conformational searches in this extremely high dimensional space, and efficient sampling techniques are required.

In order to solve this problem, various sampling and optimization methods for conformations of biomolecules have been proposed such as generalized-ensemble algorithms.[1] Simulated annealing[2] and genetic algorithm[3,4] have been recognized by researchers as powerful tools for difficult optimization problems. Simulated annealing mimics an annealing process in which the temperature of a system is lowered very slowly from a sufficiently high initial temperature to "freezing" temperature. This method has been applied to molecular simulations of biomolecules[5–11] (and also to various other research fields). The genetic algorithm mimics the process of natural evolution and has been applied to various research fields and is one of well-known techniques. The genetic algorithm uses the optimization procedures of natural gene-based evolution, that is, mutation, crossover, and replication. For a certain optimization problems, this algorithm has been found to be an excellent strategy to find global minima. The conformational search or optimization approaches for biomolecules using the genetic algorithm have also been performed.[12–16]

We proposed a new conformational search method, in which a simulated annealing simulation is combined with genetic algorithm, namely, *parallel simulated annealing using genetic crossover* (PSA/GAc),[17] and applied it to be the search of the global-minimum-energy structures for protein systems.[18,19] Here, the genetic crossover is one of the operations of genetic algorithm. The conformational search using simulated annealing is based on local conformational updates. On the other hand, the genetic algorithm is based on global conformational updates. Our method incorporates these two attractive features of the simulated annealing and the genetic crossover. In our previous work, in order to examine the effectiveness of our method, we compared our method with those of the conventional simulated annealing molecular dynamics simulations using an α-helical miniprotein, namely, Trp-cage.[20]

However, in the case of the conformational search of a protein constructed by a certain length of amino acids, we often found the lower energy conformation in spite of the completely different structure in comparison with the native structure. The structure has a very compact fold as if there is a knot in the string of the amino-acid sequence. For example, in Fig. 1, two conformations, namely, a native structure and a stable conformation obtained from the simulation by using our method, of a protein, which is the B1 domain in the immunoglobulin G (IgG) binding domains of protein G[21] and has 56 amino residues are shown. For the stable conformation obtained from the simulation, there is one knot in the string of the protein. Knot conformations of some proteins are already found by experiments of X-ray crystallography.[22,23] However, the knotted chains in the knot conformations obtained from the simulations by using our method have obviously different length from those of the experimental results. Although the length of the knotted chains known by experiments is at least 35–45 amino residues, the length of the knotted chains of the conformations obtained from the simulations is about 15 amino residues. Namely, the knot conformations obtained from the simulations are unnatural. As the reasons of getting the unnatural knot conformations, it is thought to be causally related to using the inaccurate force field and/or the unusual simulation technique in comparison with the conventional MC or MD. As far as we know, knotted conformations were never found with other conformational sampling methods, which suggests the powerfulness of our conformational search method.

Therefore, we propose the improved conformational search method, which can avoid the unnatural knot conformations. In order to check whether a knot conformation or not, we use the Kauffman polynomial[24] in the mathematical theory of knots. If a trial conformation generated by the crossover operation in the simulation has knots, the conformation is rejected regardless of the value of the potential energy. In this paper, we performed the conformational search of protein G by using the improved PSA/GAc and the conventional one in order to examine the simulation results of the improved method.

In section 2 the details of our conformational search method and its improved one are given. In section 3 the results of applications of the folding simulations of protein G are presented. Section 4 is devoted to conclusions.

2. Method

2.1. *Parallel simulated annealing molecular dynamics using genetic crossover*

Let M be the total number of *individuals*. In parallel simulated annealing using genetic crossover (PSA/GAc), a crossover operation is carried out in a fixed interval of a certain time steps of the M parallel conventional simulated annealing simulations. The entire process of the general formalism of parallel simulated annealing using genetic crossover[17–19] is illustrated in Fig 2 (in the schematic illustration there, we have $M = 6$). In parallel simulated annealing molecular dynamics using genetic crossover (PSAMD/GAc), M conventional simulated annealing molecular dynamics simulations (instead of Monte Carlo simulations) are performed in parallel. Although we employed a genetic one-point crossover in our previous study,[17–19] we can employ various kinds of genetic crossover operations such as one-point crossover, two-point crossover, etc. In this study, we employed the genetic two-point crossover, and we refer to the entire method as PSAMD/GAc2. The crossover operation in this method exchanges a part of dihedral angles between two conformations of a protein.

In the two-point crossover operation, the following procedure is carried out (see Fig. 3) :

(1) $M/2$ pairs of conformations are selected from "parental" group randomly.

(2) Consecutive amino acids of length n residues in the amino-acid sequence of the conformation are selected randomly for each pair of selected conformations.

(3) All dihedral angles (in backbone and side chains) in the selected n amino acids are exchanged between the selected pairs of conformations.

Note that the length n of consecutive amino-acid residues is in general different for each pair of selected conformations. Motivated by the fragment assembly method,[25] we take n to be an integer ranging from 2 to 10. In this procedure, we obtain two new "child" conformations. After that, we have to select two superior "chromosomes" (conformations) from the total of four conformations (two parental conformations and two new child conformations). We perform the energy minimizations for these four conformations by a standard method such as Newton-Raphson method and conjugate gradient method. We then select two lower-energy conformations based on the four minimized energy values. Finally, using the selected two energy-minimized conformations, the parallel simulated annealing simulations continue.

In our previous works, we did not perform the energy minimization after the genetic crossover operation. However, the conformations generated by the genetic crossover operation often have unusually high potential energy, because the genetic crossover operation brings about a large global change of conformations. This leads to very low acceptance ratio of child conformations. Therefore, in this study, we perform the energy minimization after the genetic crossover operation in order to avoid this difficulty of low acceptance ratio. Because the conformational change by the energy minimization is very small (in the example of a mini-protein presented below, the root-mean-square deviations of C$_\alpha$ atoms between before and after energy minimizations was only about 0.45 Å on the average), we believe that this energy minimization does not affect the nature of the new conformational generation of the crossover operations.

2.2. ***PSAMD/GAc with knot theory***

In this paper, in order to check whether a trial conformation, which generated by the genetic crossover operation, has knots or not, we use the Kauffman polynomial[24] in knot theory.

2.2.1. *Calculation of knot invariants*

To characterize the topological properties of knots and links of strings algebraically, polynomials can be used. These polynomials are knot invariants, which have been discovered and constructed, and have been proposed several polynomials. The Kauffman polynomial $F(L; a, x)$ is one of them and is a two-variable (a and x) invariant.

$$F(L; a, x) = a^{-t(\tilde{L})}\Lambda(|\tilde{L}|; a, x). \tag{1}$$

Here, L is a link. A knot is an embedding of a single circle into three-dimensional space, while a link is an embedding of a collection of circles. The sign of $|\ |$ means unoriented knots, and the tilde $\tilde{}$ means a link represented by a link diagram. $\Lambda(|\tilde{L}|; a, x)$ is defined by the conditions and the Skein relation, which is recursion relations relating the invariants of knots, in Fig. 4(b). Four knots $|L_+|$, $|L_-|$, $|L_\infty|$, and $|L_{-\infty}|$ in Fig. 4(b) correspond to the line configurations $+$, $-$, ∞, and $-\infty$ in Fig. 4(a), respectively. $t(\tilde{L})$ is the sum of the signs of all the crossings. If a knot is unknot (trivial knot), the knot invariant estimated by the Kauffman polynomial is equal to 1 ($F = 1$), if it is other knots, the knot invariant is a polynomial except 1 ($F \neq 1$). Namely, by estimating the knot invariant, we can determine whether a conformation has knots or not.

2.2.2. *Estimation of knotting properties of a protein*

We need to construct a knot diagram from a protein conformation in order to obtain the knot invariant. At first, the coordinate points of C$_\alpha$ atoms in a protein conformation are projected on X-Y Cartesian coordinate space. These points are connected from N-terminal to C-terminal by lines. If two lines intersect, the crossing point is defined as a crossing on the knot diagram, and the sign of the crossing is determined by the relation of Z Cartesian coordinates of the crossing point on the two lines. After that, the two points of the first and last C_α atoms are connected by as few crossings as possible. We use a collection of these lines connected by the points of C_α atoms as a knot diagram.

2.2.3. *Flow of PSAMD/GAc with knot theory*

A chart of the PSAMD/GAc simulation process with knot theory is shown in Fig. 5. In our improved simulation, the calculation of the knot invariant for a conformation is performed after the process of crossover operations. In the conventional PSAMD/GAc simulation process, we select two lower-energy conformations based on the four minimized energy values of four conformations (two parental conformations and two new child conformations). On the other hand, in the improved process, if both two new child conformations generated by genetic crossover operations do not have knots, the simulation process is the same as the conventional one. If one of two new child conformations has knots, we select two lower-energy conformations based on the three minimized energy values of three conformations (two parental and one child conformations) except one child knot conformation. If both two conformations have knots, we do not perform the procedure of the selection, namely, two parental conformations are selected, and after that, the simulation continues.

3. Results and Discussion

We applied our improved method to the protein G (PDB code: 1PGA).[21] Protein G from *Streptococcus* also binds human immunoglobulin G (IgG). This protein consists of a series of small binding domains separated by linkers and a cell-wall anchor near the C-terminus. Two (in some strains, three) of the domains bind IgG. The IgG-binding domains of protein G are identified as B1, B2, etc., numbering from the N-terminus of the native protein G molecule. We used the B1 domain which consists of a four-stranded β-sheet and an α-helix, and was engineered for production as a 56 residue protein with N-terminal methionine (this position was threonine in the wild type) (see Fig. 1(a)).

We incorporated PSAMD/GAc2 by modifying the TINKER program package[26] modified by us. The unit time step was set to 2.0 fs, and all bonds to hydrogen atoms at ideal bond lengths were constrained by RATTLE method.[27] Each simulation was carried out for 2.0 nsec (hence, it consisted of 1,000,000 MD steps) with 32 individuals ($M = 32$) and repeated 5 times. The temperature during MD simulations was controlled by Berendsen method.[28] For each run the temperature was decreased exponentially from 1000 K to 200 K. As for the conformational energy calculations, we used the AMBER ff96 force field.[29] As for solvent effects, we used the GB/SA model[30,31] included in the TINKER program package.[26] These folding simulations were started from a fully extended conformation and different sets of randomly generated initial velocities (for repetition of 5 times). The genetic crossover operations in PSAMD/GAc2 simulation were performed 1000 times at the fixed interval of 1000 MD steps. Moreover, we incorporated the calculation program of knot invariants by the Kauffman polynomial to the PSAMD/GAc2 program based on a program of Ochiai *et al.*[32] After the genetic crossover operation, the energy minimization by the quasi-newton method (L-BFGS)[33] included in TINKER was performed. Additionally, we performed the conventional simulated annealing molecular dynamics simulations for comparison. In order to balance the computational cost, we performed 160 simulation runs of 2 nsec in length ($32 \times 5 = 160$). The other simulation conditions were the same (except for with or without crossover operations).

We remark on the dependence of the frequency of knotted conformation creation on the force fields. We found that three out of five simulations with OPLS-AA/L and one out of five simulations with CHARMM22 created knotted conformations, while five out of five simulations with AMBER ff96 found knotted conformations. Because AMBER ff96 gave the most number of knotted conformations, we present the results of our knot-avoiding method with AMBER ff96 below.

In Fig. 6, the lowest-energy final minimized conformations obtained from the normal PSAMD/GAc2, and the improved PSAMD/GAc2 with knot theory are shown. As these results, all the conformations obtained from the normal PSAMD/GAc2 have unnaturally knot conformations. On the other hand, the conformations obtained from the improved PSAMD/GAc2 with knot theory have the stable conformations without knots.

In Fig. 7, the minimized potential energy of the final 160 conformations obtained from the normal PSAMD/GAc2, the improved PSAMD/GAc2 with knot theory, and the conventional simulated annealing is shown. As a reference, the value for the native conformation is also shown. Here (and in Fig. 8 below), the "native conformation" means the conformation that was obtained as follows. A canonical MD simulation of 100 psec at a low temperature (200 K) with the initial conformation being the native PDB conformation was first performed. The final conformation was then energy-minimized. The heavy-atom RMSD of this "native conformation" from the PDB coordinates was 1.4 Å. In comparison with the conventional simulated annealing method, the potential energy is obviously lower in both the normal PSAMD/GAc2 and the improved PSAMD/GAc2 with knot theory as a whole. The lowest energy and the average energy obtained from the normal PSAMD/GAc2 are −2322.7 kcal/mol and −2306.6 kcal/mol, respectively. Those obtained from the improved PSAMD/GAc2 with knot theory are −2310.4 kcal/mol and −2297.6 kcal/mol, respectively. On the other hand, those obtained from the conventional simulated annealing method are −2277.3 kcal/mol and −2237.9 kcal/mol. The differences of the energy values between the normal and improved PSAMD/GAc2 are 12.3 kcal/mol and 9.0 kcal/mol. The differences of the energy values between the normal PSAMD/GAc2 and the conventional simulated annealing are 45.4 kcal/mol and 68.7 kcal/mol. As these results, the conformations obtained from both the normal and improved PSAMD/GAc2 are more stable than those of the conventional simulated annealing. Namely, by incorporating the crossover operation into the simulated annealing method, we can obtain more stable structures than the conventional simulated annealing method. Additionally, the conformations obtained from the normal PSAMD/GAc2 are slightly more stable than the improved PSAMD/GAc2 with knot theory. This result shows that the more unnatural conformations with knots are more stable than the conformations without knots. We suppose that one of the reasons is the inaccuracy of the force field for the simulations.

In Fig. 8, the radius of gyration of the final minimized conformations obtained from the normal PSAMD/GAc2 and the improved PSAMD/GAc2 with knot theory, and the conventional simulated annealing are shown. These results obviously illustrate that the final conformations obtained from both the normal PSAMD/GAc2 and the improved PSAMD/GAc2 with knot theory become more compact conformations in comparison of those of the conventional simulated annealing on the whole. Namely, the improved method as well as the normal method can search compact conformations.

4. Conclusions

In this article, for the parallel simulated annealing using genetic crossover (PSA/GAc), we proposed the improved method, which can avoid the unnatural knot conformations. In order to check whether a conformation has knots or not, we used the Kauffman polynomial in the mathematical theory of knots and links and incorporated the check function to PSA/GAc.

As a test simulation, we applied this improved conformational search method to the protein G. We succeeded in performing the simulations which avoided unnatural knot conformations and could obtain stable conformations as well as the normal PSAMD/GAc2, in comparison with the conventional simulated annealing. Additionally, the knot conformations obtained from the normal PSAMD/GAc2 were slightly more stable than the unknoted conformations obtained from the improved PSAMD/GAc2 with knot theory. One of the supposable reasons is inaccuracy of the force field for the simulations. Therefore, in a future work we are going to perform the conformational search by PSAMD/GAc2 with the force field optimized by our optimization methods.[34,35]

Once all these preparations are successfully made, we will be ready to apply the present method to multi-scale modelling of biosystems.

Acknowledgments

The computations were performed on SuperNOVA of Miki Laboratory at Doshisha University and the computers at the Research Center for Computational Science, Institute for Molecular Science. This work was supported, in part, by the Grants-in-Aid for the Academic Frontier Project, "Intelligent Information Science", for Scientific Research on Innovative Areas ("Fluctuations and Biological Functions"), and for the Next Generation Super Computing Project, Nanoscience Program from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

1. A. Mitsutake, Y. Sugita and Y. Okamoto, *Biopolymers (Pept. Sci.)* **60**, 96 (2001).
2. S. Kirkpatrick, C. D. Gelatt Jr. and M. P. Vecchi, *Science* **220**, 671 (1983).
3. J. H. Holland, *Adaptation in Natural and Artificial Systems* (The University of Michigan Press, Ann Arbor, 1975).
4. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, 1989).
5. M. Nilges, G. M. Clore and A. M. Gronenborn, *FEBS Lett.* **229**, 317 (1988).

6. A. T. Brünger, *J. Mol. Biol.* **203**, 803 (1988).
7. A. T. Brünger, M. Karplus and G. A. Petsko, *Acta Cryst.* **A45**, 50 (1989).
8. S. R. Wilson, W. Cui, J. W. Moskowitz and K. E. Schmidt, *Tetrahedron Lett.* **29**, 4373 (1988).
9. H. Kawai, T. Kikuchi and Y. Okamoto, *Protein Eng.* **3**, 85 (1989).
10. C. Wilson and S. Doniach, *Proteins* **6**, 193 (1989).
11. Y. Okamoto, M. Fukugita, T. Nakazawa and H. Kawai, *Protein Eng.* **4**, 639 (1991).
12. T. Dandekar and P. Argos, *Protein Eng.* **5**, 637 (1992).
13. S. Sun, *Protein Science* **2**, 762 (1993).
14. R. Unger and J. Moult, *J. Mol. Biol.* **231**, 75 (1993).
15. A. A. Rabow and H. A. Scheraga, *Protein Science* **5**, 1800 (1996).
16. J. Lee, H. A. Scheraga and S. Rackovsky, *J. Comput. Chem.* **18**, 1222 (1997).
17. T. Hiroyasu, M. Miki and M. Ogura, *Proceedings of the 44th Institute of Systems* , 113 (2000).
18. T. Hiroyasu, M. Miki, M. Ogura and Y. Okamoto, *J. IPS Japan* **43**, 70 (2002), in Japanese.
19. T. Hiroyasu, M. Miki, M. Ogura, K. Aoi, T. Yoshida and Y. Okamoto, *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003)* , 117 (2003).
20. J. W. Neidigh, R. M. Fesinmeyer and N. H. Andersen, *Nature Struct. Biol.* **9**, 425 (2002).
21. T. Gallagher, P. Alexander, P. Bryan and G. L. Gilliland, *Biochemistry* **33**, 4721 (1994).
22. K. Lim, H. Zhang, A. Tempczyk, W. Krajewski, N. Bonander, J. Toedt, A. Howard, E. Eisenstein and O. Herzberg, *Proteins* **51**, 56 (2003).
23. O. Nureki, M. Shirouzu, K. Hashimoto, R. Ishitani, T. Terada, M. Tamakoshi, T. Oshima, M. Chijimatsu, K. Takio, D. G. Vassylyev, T. Shibata, Y. Inoue, S. Kuramitsu and S. Yokoyama, *Acta Cryst.* **D58**, 1129 (2002).
24. L. H. Kauffman, *Trans. Am. Math. Soc.* **318**, 417 (1990).
25. K. T. Simons, C. Kooperberg, E. Huang and D. Baker, *J. Mol. Biol.* **268**, 209 (1997).
26. Tinker program package software available at http://dasher.wustl.edu/tinker/.
27. H. C. Andersen, *J. Comput. Phys.* **52**, 24 (1983).
28. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.* **81**, 3684 (1984).
29. P. A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot and A. Pohorille, *Computer simulations of biological systems* (Escom, Netherlands, 1997), ch. The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data, pp. 83–96.
30. W. C. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson, *J. Am. Chem. Soc.* **112**, 6127 (1990).
31. D. Qiu, P. S. Shenkin, F. P. Hollinger and W. C. Still, *J. Phys. Chem. A* **101**, 3005 (1990).
32. M. Ochiai, S. Yamada and E. Toyoda, *Computer Aided Knot Theory* (Makino Shoten, 1996). in Japanese.
33. J. Nocedal, *Math. Comp.* **35**, 773 (1980).
34. Y. Sakae and Y. Okamoto, *Chem. Phys. Lett.* **382**, 626 (2003).
35. Y. Sakae and Y. Okamoto, *Mol. Sim.* **36**, 159 (2010).

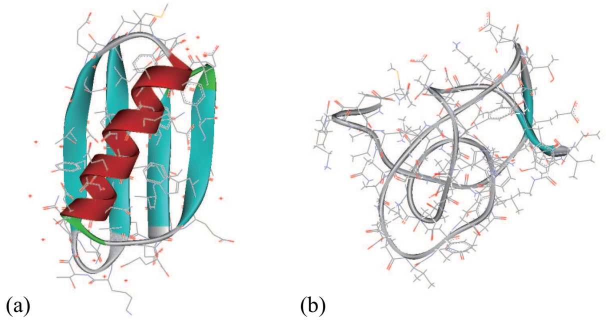


Fig. 1. The structure of Protein G. (a) is the native structure (PDB ID: 1PGA). (b) is the final conformation obtained from PSAMD/GAc2.

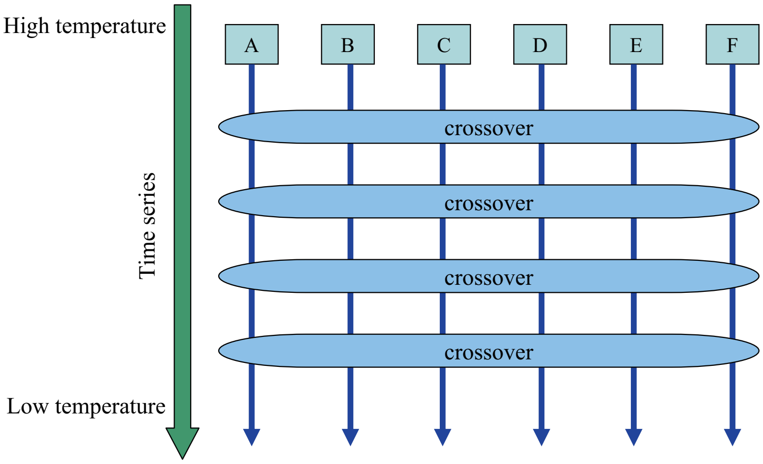


Fig. 2. Schematic process of the parallel simulated annealing using genetic crossover. In this method, the crossover operation, which is shown in Fig. 3, is performed during parallel simulated annealing simulations.

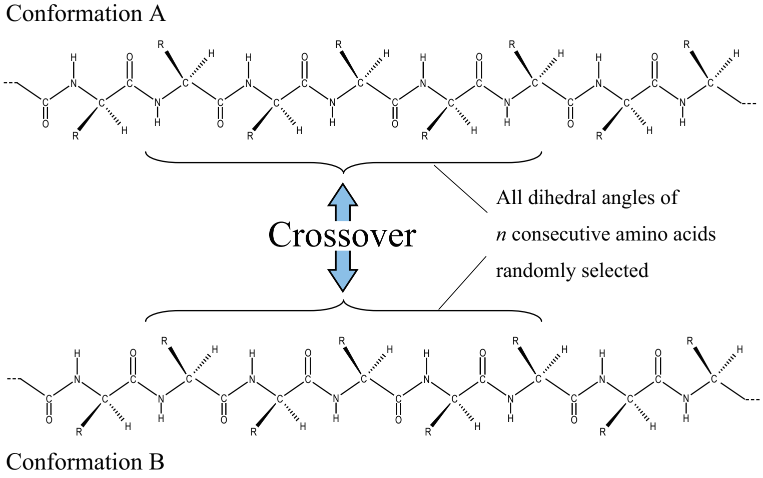


Fig. 3. Schematic process of the two-point crossover operation. In this process, all dihedral angles (in backbone and side chains) within the randomly selected n consecutive amino acids are exchanged between a pair of conformations. Motivated by the fragment assembly method,[25] we take n to be an integer ranging from 2 to 10.

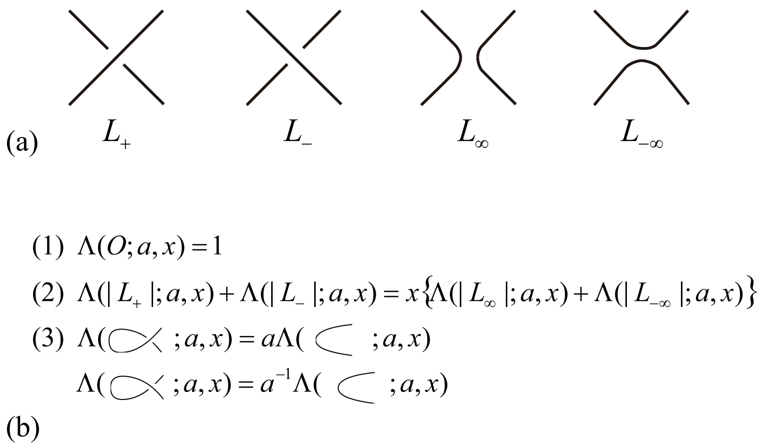

(a)
L_+
L_-
L_∞
$L_{-\infty}$
(1) $\Lambda(O;a,x)=1$
(2) $\Lambda(|L_+|;a,x)+\Lambda(|L_-|;a,x)=x\{\Lambda(|L_\infty|;a,x)+\Lambda(|L_{-\infty}|;a,x)\}$
(3) $\Lambda(\;;a,x)=a\Lambda(\;;a,x)$
$\Lambda(\;;a,x)=a^{-1}\Lambda(\;;a,x)$
(b)

Fig. 4. The four line configurations (a), the Skein relation (2) of (b), and the conditions (1,3) of (b) defined by the Kauffman polynomial.

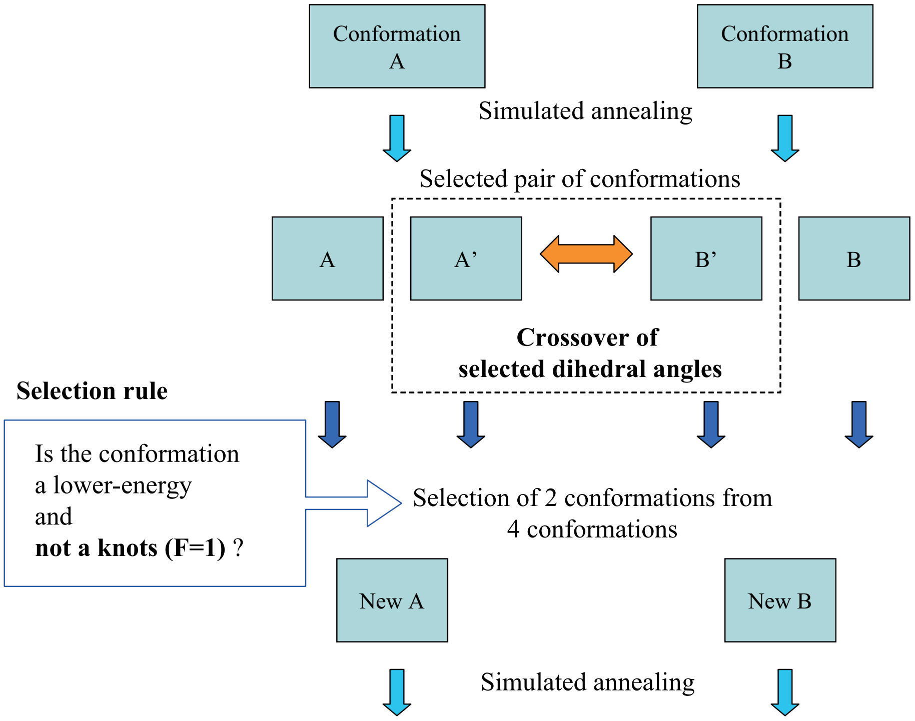

Conformation A
Conformation B
Simulated annealing
Selected pair of conformations
A
A'
B'
B
Crossover of selected dihedral angles
Selection rule
Is the conformation a lower-energy and not a knots (F=1) ?
Selection of 2 conformations from 4 conformations
New A
New B
Simulated annealing

Fig. 5. Schematic process of the simulated annealing using genetic crossover with knot theory. In this simulation, the two unknotted conformations are selected by using the Kauffman polynomial after genetic crossover operations.

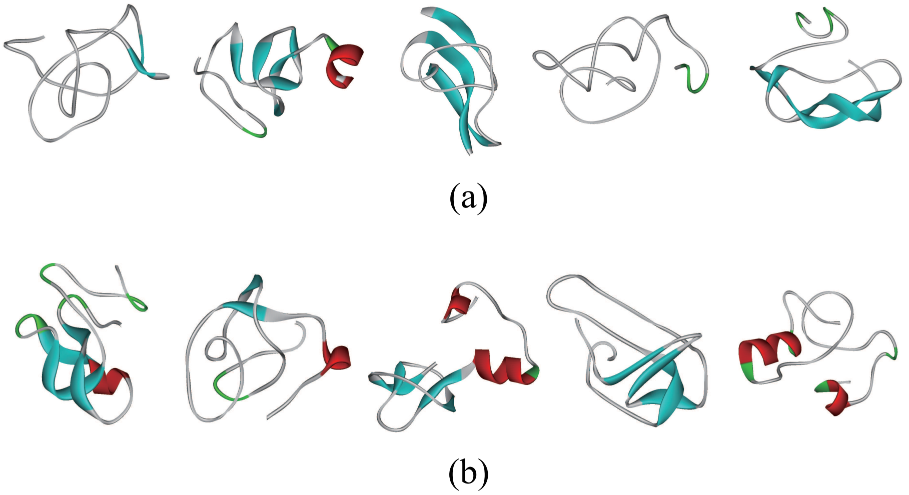


Fig. 6. The final conformations obtained from PSAMD/GAc2. (a) shows the conformations obtained from the normal method, and (b) shows the conformations obtained from the improved method with knot theory. The simulations were performed five times for both cases.

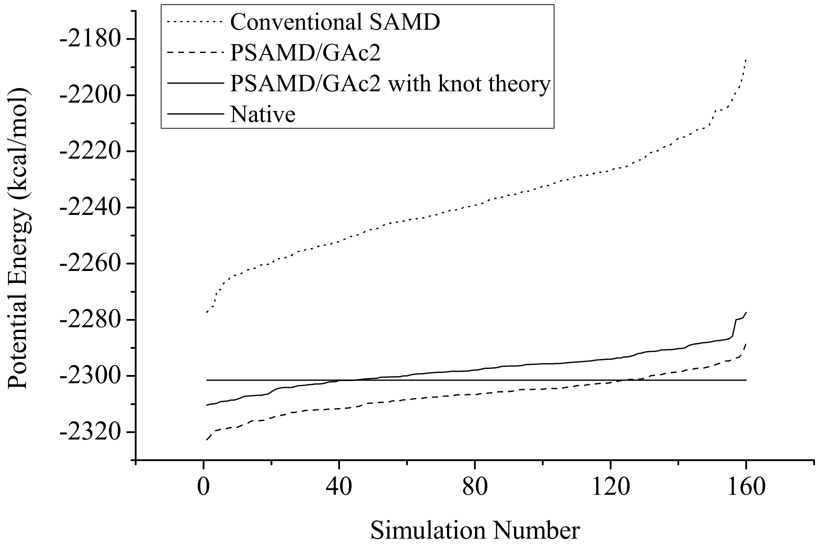


Fig. 7. Comparison of the minimized potential energy of the final conformations obtained from the conventional simulated annealing MD simulation (dotted line), the normal PSAMD/GAc2 (broken line), the improved PSAMD/GAc2 with knot theory (normal line). The value for the native structure is also shown (normal horizontal line).

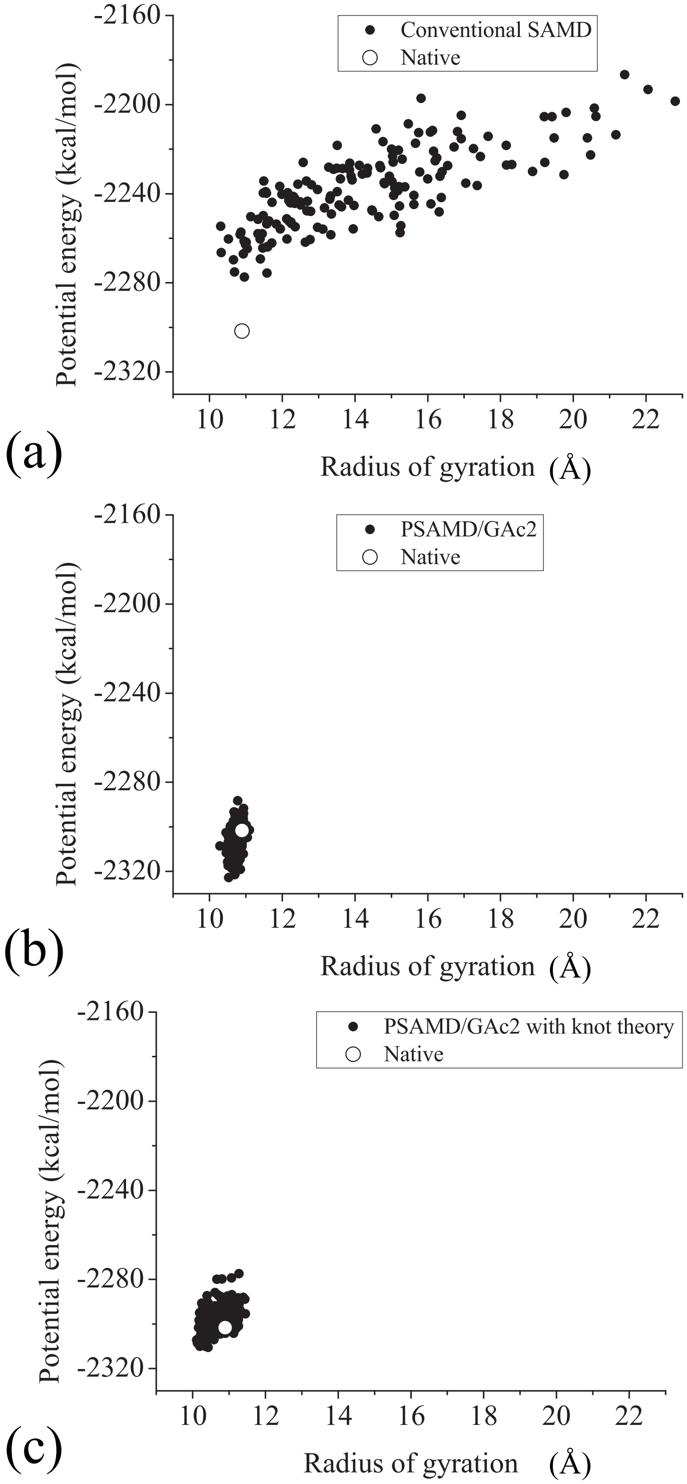


Fig. 8. Radius of gyration of the final minimized conformations obtained from the conventional simulated annealing (a), the normal PSAMD/GAc2 (b), the improved PSAMD/GAc2 with knot theory (c). The radius of gyration was caluclated with respect to all atoms. The value for the native structure is also shown (open circle).

PERSONAL GENOMICS

Can Alkan

Department of Genome Sciences, University of Washington.
and the Howard Hughes Medical Institute.
Seattle, WA, 98195, USA

Emidio Capriotti

Department of Bioengineering, Stanford University.
Stanford, CA, 94305, USA

Eleazar Eskin

Department of Computer Science, University of California Los Angeles.
Los Angeles, CA, 90095, USA

Fereydoun Hormozdiari

School of Computing Science, Simon Fraser University.
Burnaby, BC, V5A 1S6, Canada

Maricel G. Kann

Department of Biological Sciences, University of Maryland Baltimore County.
Baltimore, MD, 21250, USA

1. Introduction

Our genetic identity not only determines our physical differences, but it also defines our susceptibility against diseases. Several groups are working on various methods to exploit the power of cost efficient sequencing technologies as well as more traditional genome analysis approaches (SNP microarrays, arrayCGH, etc.) to better perform genotype-phenotype associations, in particular to identify susceptibility to disease, and eventually diagnose disease at its early stages. The ultimate goal is to vastly improve the field of pharmacogenomics, which can broadly be defined as the study of the relationship between genotype and drug response and how the drugs affect our metabolism. The abundance of new sequence data gives many opportunities to advancing our understanding of how to optimize drug combinations for each individual's genetic makeup. The underlying computational tools for such studies analyze available sequence data to identify differences between a reference genome and high-throughput sequenced genomes and perform sequence oriented clustering and classification to obtain both normal and disease-related phenotype associations.

This session focuses on the development of novel computational methods in all aspects of Personal Genomics including genetic and epigenetic variation discovery, genotype-phenotype associations, indexing and cataloguing both normal and disease-related variation, exome capture and resequencing, and personalized medicine. This session has a broad target audience that includes algorithm developers working on sequence analysis, genomics researchers, pharmacogeneticists, and medical geneticists.

2. Session Summary

This session includes an invited talk, six reviewed oral presentations, and a tutorial. The studies presented in this session focus on the development of computational methods to analyze genomic data generated with various types of methods.

2.1. *Oral Presentations*

The following six talks will be presented at the Personal Genomics session:

- "Haplotype Inference from Single Short Sequence Reads Using a Population Genealogical History Model" by Jim Zhang and Yufeng Wu,
- "Multivariate Analysis of Regulatory SNPs: Empowering Personal Genomics by Considering Cis-Epistasis and Heterogeneity" by Stephen D. Turner and William S. Bush,
- "Visual Integration of Results from a large DNA Biobank (BIOVU) using Synthesis-View" by Sarah Pendergrass, Scott M. Dudek, Dan M. Roden, Dana C. Crawford, and Marylyun D. Ritchie,
- "Use of Biological Knowledge to Inform the Analysis of Gene-Gene Interactions Involved in Modulating Virologic Failure with Efavirenz-Containing Treatment Regimens in Art-Naïve ACTG Clinical Trials Participants" by Benjamin J. Grady, Eric C. Torstenson, Paul J. Mclaren, Paul W. De Bakker, David W. Haas, Gregory K. Robbins, Roy M. Gulick, Richard Haubrich, Heather Ribaudo and Marylyn D. Ritchie,
- "The Reference Human Genome Demonstrates High Risk of Type 1 Diabetes and Other Disorder" by Rong Chen and Atul J. Butte
- "Matching Cancer Genomes to Established Cell Lines for Personalized Oncology" by Joel T. Dudley, Rong Chen, and Atul J. Butte

We are excited by the breadth of research in the field of Personal Genomics, and are hopeful that our session will help bring together researchers in these areas. The six papers presented at our session were selected with the help of several reviewers, whose help we gratefully acknowledge.

3. Acknowledgments

We would like to thank all the authors who submitted their work to the Personal Genomics Session. We are also indebted to the anonymous reviewers who contributed their time and expertise to evaluate the submitted papers.

THE REFERENCE HUMAN GENOME DEMONSTRATES HIGH RISK OF TYPE 1 DIABETES AND OTHER DISORDERS

RONG CHEN

Department of Pediatrics, Stanford University School of Medicine
Stanford, CA 94305-5479, USA

ATUL J. BUTTE

Department of Pediatrics, Stanford University School of Medicine
Stanford, CA 94305-5479, USA
Email: abutte@stanford.edu

Personal genome resequencing has provided promising lead to personalized medicine. However, due to the limited samples and the lack of case/control design, current interpretation of personal genome sequences has been mainly focused on the identification and functional annotation of the DNA variants that are different from the reference genome. The reference genome was deduced from a collection of DNAs from anonymous individuals, some of whom might be carriers of disease risk alleles. We queried the reference genome against a large high-quality disease-SNP association database and found 3,556 disease-susceptible variants, including 15 rare variants. We assessed the likelihood ratio for risk for the reference genome on 104 diseases and found high risk for type 1 diabetes (T1D) and hypertension. We further demonstrated that the risk of T1D was significantly higher in the reference genome than those in a healthy patient with a whole human genome sequence. We found that the high T1D risk was mainly driven by a R260W mutation in PTPN22 in the reference genome. Therefore, we recommend that the disease-susceptible variants in the reference genome should be taken into consideration and future genome sequences should be interpreted with curated and predicted disease-susceptible loci to assess personal disease risk.

1. Introduction

With the advance of sequencing technology and assembling tools, whole genome sequencing has become a commodity with 10,000 personal genomes being sequenced in the next two years. An urgent question is how to interpret personal genome sequences to comprehensively assess disease risk and optimize personalized treatment. Sixteen personal genomes (*1-13*) have been fully sequenced and described in the literature, while companies state they are sequencing as many as 500 individuals per month. However, due to the limited samples and lack of case/control design, the current interpretation of these genomes had been mainly focused on the identification and functional annotation of the DNA variants that are different from the reference genome sequence, with an aim to find interesting genomic features. The reference genome was not from a single normal individual; instead, the reference was deduced from a collection of DNAs from anonymous individuals with primarily European origins and assembled into a mosaic haploid genome (*14, 15*). To our knowledge, the clinical and phenotypic information of the participants had never been published. Although they were very likely to be healthy at the time of study, some of them might be carriers of disease risk alleles. The identification of biologically and clinically important rare and common disease variants in the reference genome and a comprehensive disease risk assessment will improve our understanding of the reference to better assemble and interpret future genome sequences.

We have previously developed a method to assess the risk of a patient for 55 diseases using a quantitative human disease-SNP association database, and showed that we could suggest useful and clinical relevant information using his personal genome sequence (*16*). Here, we queried the reference genome sequence against our database and identified 3,556 disease-susceptibility variants, including 15 rare variants. We comprehensively assessed the risk of the reference genome for 104 diseases and found high risk for type 1 diabetes (T1D) and hypertension. We further demonstrated that the risk of T1D was also significantly higher in the reference genome than in the genome of the healthy male we previously described (*16*). Comparing all contributing alleles, we found that the high T1D risk was mainly driven by a R260W mutation in the intracellular tyrosine phosphatase (*PTPN22*) in the reference genome.

2. Methods

2.1 Identifying the disease susceptible/protective alleles in the reference genome

We downloaded the alleles at 24.5 million SNPs (dbSNP 131 on hg19) of the reference genome from the UCSC genome browser (*17, 18*), and removed all SNPs that were mapped to multiple locations.

As described previously (*16*), we manually curated quantitative human disease-SNP associations from the full text, figures, tables, and supplemental materials of 3,333 human genetics papers, and recorded more than 100 features from each paper, including the disease name (e.g. coronary artery disease), specific phenotype (e.g. acute coronary syndrome in coronary artery disease), study population (e.g. Finnish individuals), case and control population (e.g. 2,508 patients with coronary artery disease proven by angiography), gender distribution, genotyping technology, major/minor risk alleles, odds ratio, 95% confidence interval of the odds ratio, published p-value, and genetic model. Studies on similar diseases were categorized and mapped to the Concept Unique Identifiers (CUI) in the Unified Medical Language System (UMLS) (*19*). For each study, the frequency of each genotype and allele in the case and control populations was recorded.

We queried the reference genome against this disease-SNP database using dbSNP identifiers (*17*), and identified all disease susceptible or protective alleles in the reference. We then retrieved the Minor Allele Frequency (MAF) from the HapMap II and III projects (*20*) and identified rare disease-susceptible alleles in the reference that had an MAF<1% in the CEU population.

2.2 Assessing the risk of the reference genome on 104 diseases

We had previously reported the medical assessment of a personal genome sequence from a healthy 40-year-old male by calculating his pre-test probability, likelihood ratio (LR), and post-test probability across 55 diseases (*16*) using a curated high-quality quantitative human disease-SNP association database. Similarly, for each of 104 diseases, we queried the reference genome sequence against our database, identified all independent disease-associated loci, treated the genotype at each locus as an independent genetic test, and calculated the LR as the increased disease odds from all tests.

For each disease, we identified all SNPs that had been significantly associated with the disease with a p value of $\leq 10^{-6}$ in Genome-Wide Association Studies on more than 5000 individuals, or with a p value of ≤ 0.01 in candidate gene studies on more than 1000 individuals. We estimated genetic risk using a likelihood ratio for each SNP defined by the relative frequency of the individual's genotype in the diseased vs. healthy control populations (e.g., given an allele "A", LR = Pr(A|diseased)/Pr(A|control)). The LR incorporates both the sensitivity and specificity of the test and provides a direct estimate of how much a test result will change the odds of having a disease (*21*). In addition, the likelihood ratio is taught to medical students and physicians in training(*22*).

We excluded studies with diseased patients in the control group, and included studies across all ethnicities and genders, because the reference genome was deduced from a mixture of people with different ethnicities and genders. For each allele, we averaged the LRs from multiple studies with a weight of the square root of the sample size to give higher confidence to studies with larger sample size. After removing SNPs in high linkage disequilibrium ($R^2 \geq 0.8$ in HapMap CEU populations), we assumed each locus as an independent genetic test and multiplied LRs to report the combined LR or risk.

2.3 Comparing the disease risk between the reference genome and a healthy patient

We plotted the log(LR) of a 40-year-old healthy male (*16*) against the log(LR) of the reference genome across 62 shared diseases to identify the diseases where the reference genome had significantly higher risk. All contributing SNPs were plotted for the disease to identify SNPs that drove the observed risk difference between the two genomes. For each SNP, its associated gene was identified using the NCBI Entrez dbSNP (*17*), and annotated using the UCSC genome browser (*18*) for its functional type and chromosome location.

3. Results:

3.1 Disease susceptible and protective alleles in the reference genome

The reference genome (hg19) contains 21.8 million SNPs, with 17,429 of them known to associate with human disease and other phenotypes, and 12,190 of them known to associate with human diseases (Table 1). It contains slightly more diseases-protective alleles and genotypes (4,052 SNPs for 381 diseases) than disease-susceptible alleles and genotypes (3,556 SNPs for 349 diseases).

Table 1: Number of disease susceptible and protective alleles in the reference genome

	SNPs	Phenotypes	PubMed count
Disease/traits#	17,429	1,026	3,333
Associated with disease	12,190	561	2,695
Susceptibility to disease	3,556	349	1,416
Protection from disease	4,052	381	1,600

Non-disease phenotypes included drug response and clinical measurements

3.2 Rare disease-susceptible variants in the reference genome

The reference genome carries minor alleles at 0.93 million SNPs in the CEU population, and 0.15 million of them were rare variants with MAF<1% in the HapMap II and III projects (*20*). We found that 15 rare alleles in the reference genome are known to increase the risk of a variety of diseases (Table 2). For example, rs10849033 is close to the 5' end of *C12orf5*, a TP53-induced glycolysis and apoptosis regulator. The reference genome has a rare G allele at rs10849033 with an MAF of 0.8%. The G allele had been found to significantly increase the risk of acute lymphoblastic leukemia (ALL) by 2.55 fold, with a p value of 8.5×10^{-6} in a study on 317 children with ALL and 17,958 non-ALL individuals in a control group (*23*). This rare ALL-susceptibility variant would likely be missed by recent personal genome resequencing efforts focusing on reporting and studying only those variants different from the reference genome.

Table 2: Rare disease-susceptible variants (MAF<1%# in Caucasian) in the reference genome

Disease	Gene	SNP	Allele	Type	PubMed
Acute lymphoblastic leukemia	C12orf5	rs10849033	G	near 5'	19684603
Asthma		rs10837012	G	unknown	19187332

		rs1335159	C	unknown	19187332
Breast cancer	RRP1B	rs9306160	T	missense	19825179
Coronary artery disease	PON2	rs7493	G	missense	12588779
Focal segmental glomerulosclerosis	WT1	rs2234591	T	intron	15687485
Juvenile idiopathic arthritis	SLC26A2	rs30832	T	missense	17393463
Malaria	FAM53B	rs7076268	C	intron	19465909
Obesity		rs7173766	A	unknown	19584900
Parkinson's disease	ADH1C	rs283413	A	nonsense	15642852
	NUCKS1	rs823128	G	intron	19915575
Placental abruption	F5	rs6025	T	coding-synon	18277167
Prostate cancer	GDF15	rs1058587	C	missense	16775185
Schizophrenia		rs4568102	A	unknown	18347602
Type 2 diabetes	ARHGEF11	rs861086	G	near 5'	17369523
Venous thrombosis	F5	rs6025	T	coding-synon	17284699

[#] MAF (minor allele frequency) was retrieved from the HapMap II and III projects

We further found two rare variants in the reference genome increasing the risk of Parkinson’s disease (Table 2). One of them is rs283413, containing an A allele in the reference genome, which leads to the early truncation of *ADH1C* protein, and has been known to increase the risk of Parkinson’s disease by 3.25 fold (p=0.007) in multiple Swedish and Caucasian studies (*24*).

A large survey across 17,429 disease SNPs in our database showed that the effect sizes or the odds ratio of disease SNP associations were consistently and negatively associated with the MAF in Caucasian, African, Chinese, and Japanese. This indicated that rare disease-associated SNPs conveyed significantly larger effect size to the observed genetic association across human diseases. With the discovery of several rare alleles known to be associated with disease in the reference genome, we suggest that whole genome resequencing would very likely identify other causal SNPs, possibly explaining some of the currently missing genetic heritability of complex diseases (*25*). As such, some of the other 0.15 million rare variants in the reference genome could also potentially be associated with disease. Comparing genome sequences against curated disease and rare variants would likely discover many causal variants.

3.3 Risk likelihood ratio of the reference genome on 104 diseases

We analyzed the risk likelihood ratio (LR) of the reference genome on 104 diseases using the independent test likelihood ratio model. We found that the reference genome had an increased risk on 48 diseases (LR>1) and a decreased risk on 56 diseases (LR<1). The LR ranged from 0.14 to 5.14 with a mean LR close to 1.0 (p=0.39, t-test). Strikingly, T1D demonstrated the highest risk with a product LR of 5.14. This LR was calculated from 31 T1D-susceptible alleles and 14 T1D-protective alleles in the reference genome.

The reference genome also had a high likelihood ratio of risk for hypertension with 11 risk and 3 protective alleles. The high risk of hypertension was mainly driven by a G allele at rs3741691 in *THAP2* with a LR of 1.26 (*26*), an A allele at rs2106809 in *ACE2* with a LR of 1.26 (*27*), and an A risk allele at rs3761987 with a LR of 1.21 (*26*). Table 3 lists the LR and the number of susceptible and protective SNPs on just the 44 diseases with 10 or more SNPs.

Table 3: Disease risk profile of the reference genome on 44 diseases with ≥10 SNPs

Disease	LR	Susceptible SNPs	Protective SNPs
Type 1 diabetes	5.14	31	14
Hypertension	2.58	10	3
Ankylosing spondylitis	1.90	9	6
Myocardial infarction	1.78	10	3
Prostate cancer	1.56	22	19
Breast cancer	1.28	17	17
Multiple sclerosis	1.25	10	4
Inflammatory bowel disease	1.21	7	8
Colorectal cancer	1.20	9	12
Lung cancer	1.03	6	5
Parkinson's disease	1.01	14	7
Alzheimer's disease	0.89	10	8
Coronary artery disease	0.86	8	9
Celiac disease	0.83	9	10
Rheumatoid arthritis	0.76	12	11
Bipolar disorder	0.75	5	5
Schizophrenia	0.71	5	10
Ulcerative colitis	0.70	6	12
Systemic lupus erythematosus	0.66	26	29
Type 2 diabetes	0.61	34	37
Crohn's disease	0.55	12	17
Glioma	0.53	4	9
Psoriasis	0.47	11	10
Obesity	0.43	6	14
Basal cell carcinoma	0.33	3	8
Melanoma	0.14	4	11

We then plotted the histogram of log(LR) across all 198 diseases, and observed a symmetric distribution with no significant difference from the mean of zero (p=0.07, t-test). This suggests that our method is unbiased towards overcalling susceptibility or protection across all diseases.

3.4 Disease risk comparison between the reference and a personal genome

We plotted the log(LR) of a 40-year-old healthy Caucasian male against the log(LR) of the reference genome across 104 shared diseases (Figure 2). Interestingly, the reference genome showed a strikingly increased risk on T1D than the healthy male, and a decreased risk on Melanoma. This indicats that the high T1D risk was likely a result of T1D-susceptible alleles in the reference genome instead of biased T1D-susceptible alleles in the database. Although the reference genome was deduced from a group of healthy persons, some of them might be carriers of T1D-sueceptible alleles. Therefore, the reference genome is not free of predicted disease-risk and these disease-susceptible alleles in the reference genome need to be taken into consideration in interpreting future genome sequences.

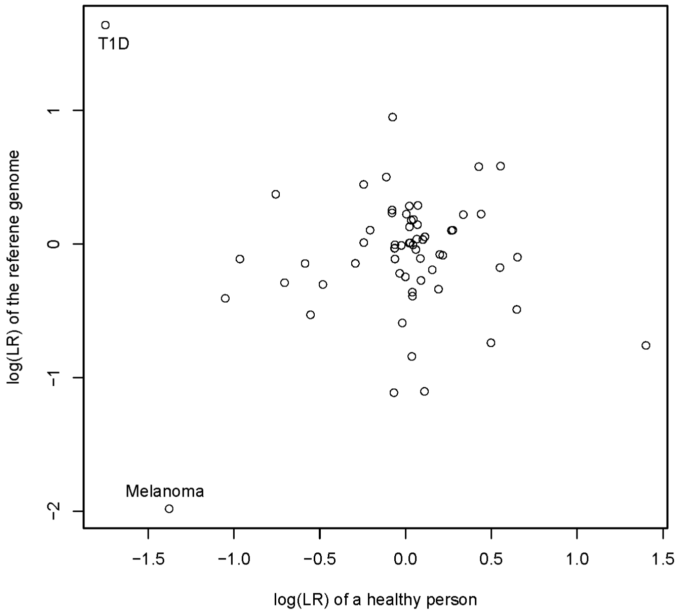
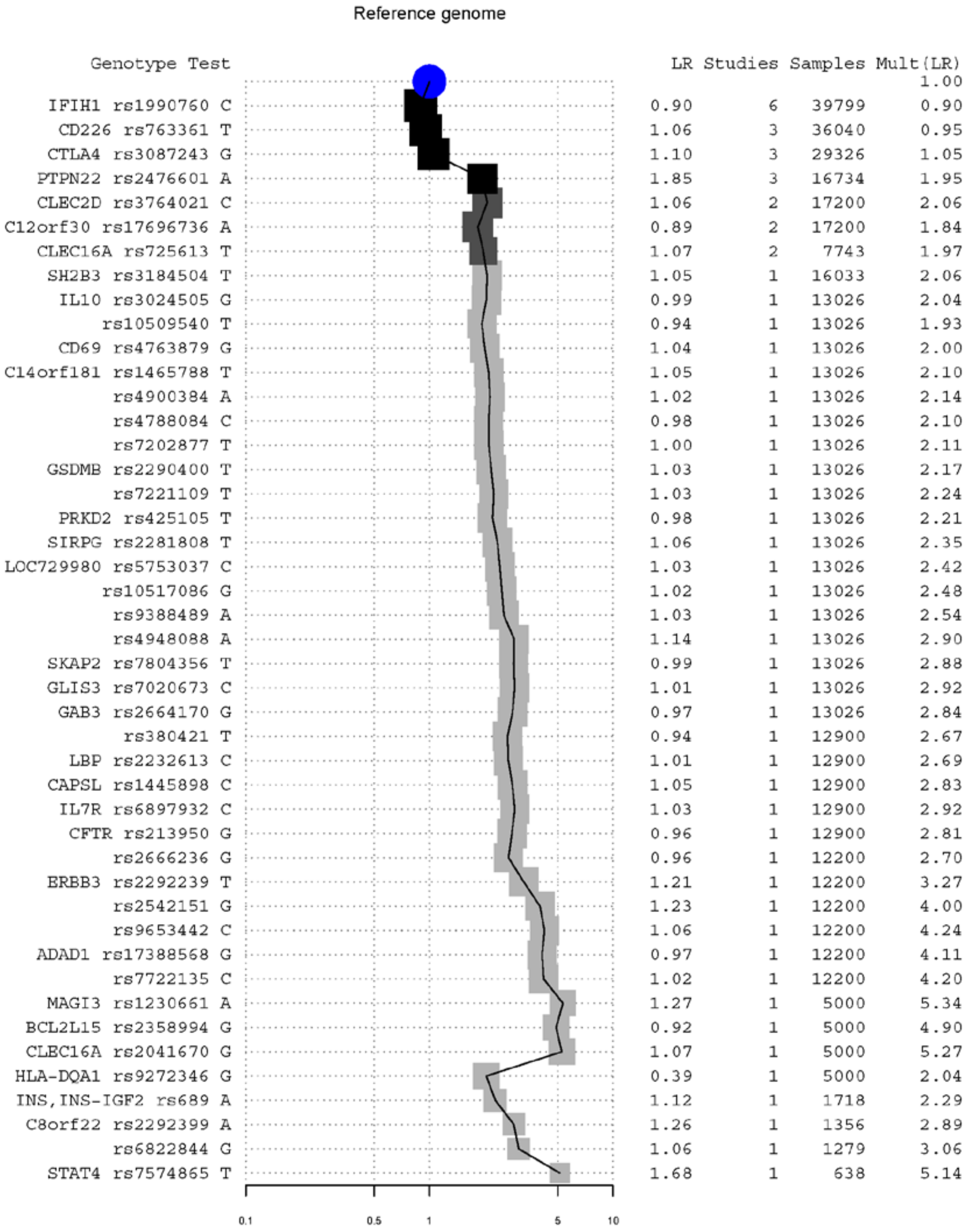


Fig. 1: The disease risk comparison between the personal genome of a healthy male and the reference genome. Each circle represents the genetic risk of a disease for the patient and the reference genome.

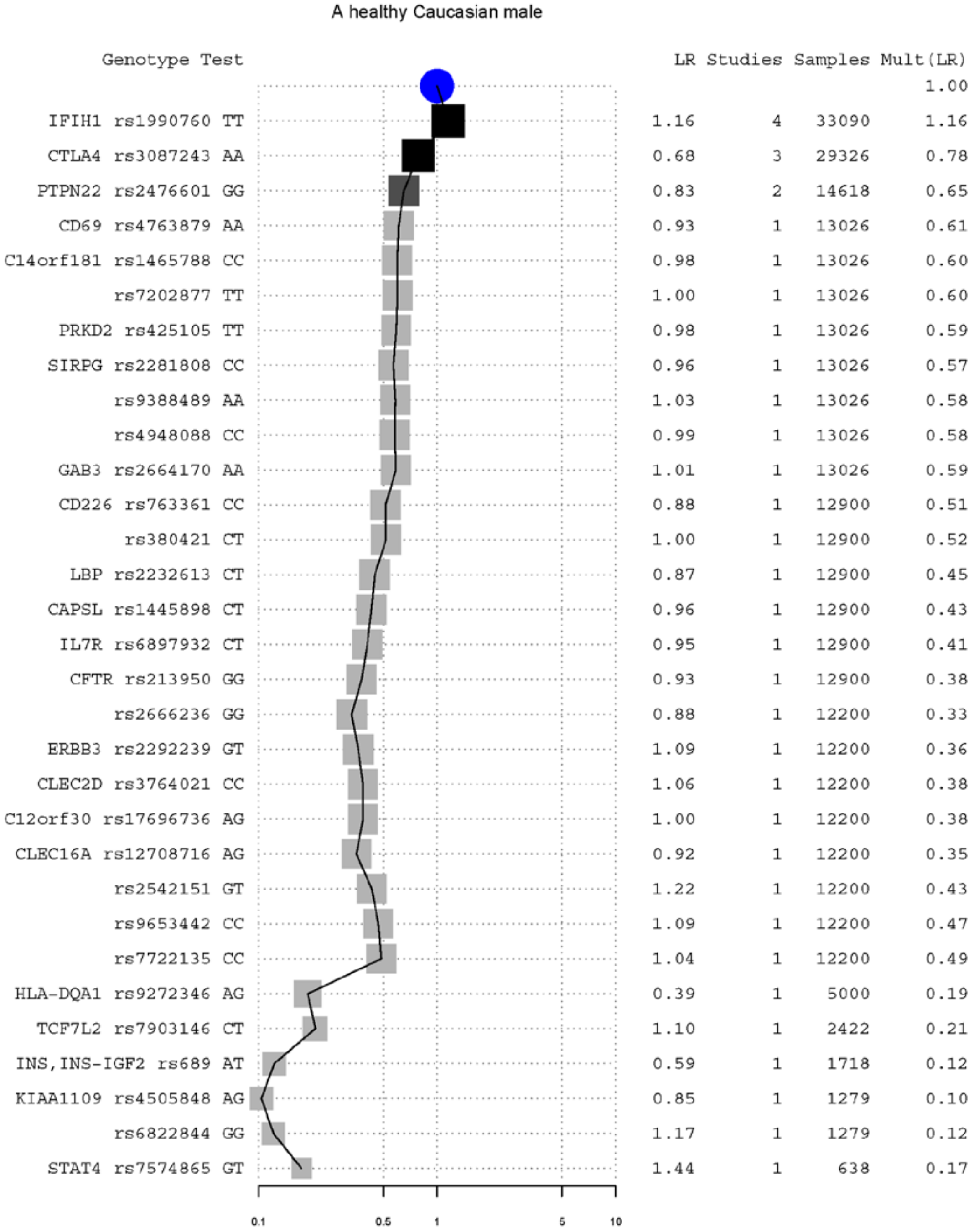
3.5 T1D-susceptible alleles in the reference genome

To identity the specific alleles that led to the striking difference on predicted T1D risk between the reference genome and the healthy male, we plotted all contributing T1D susceptible and protective alleles in both the reference genome (Figure 2) and the previously studied 40 year old patient (Figure 3).



Genotype Test	LR	Studies	Samples	Mult(LR)
				1.00
IFIH1 rs1990760 C	0.90	6	39799	0.90
CD226 rs763361 T	1.06	3	36040	0.95
CTLA4 rs3087243 G	1.10	3	29326	1.05
PTPN22 rs2476601 A	1.85	3	16734	1.95
CLEC2D rs3764021 C	1.06	2	17200	2.06
C12orf30 rs17696736 A	0.89	2	17200	1.84
CLEC16A rs725613 T	1.07	2	7743	1.97
SH2B3 rs3184504 T	1.05	1	16033	2.06
IL10 rs3024505 G	0.99	1	13026	2.04
rs10509540 T	0.94	1	13026	1.93
CD69 rs4763879 G	1.04	1	13026	2.00
C14orf181 rs1465788 T	1.05	1	13026	2.10
rs4900384 A	1.02	1	13026	2.14
rs4788084 C	0.98	1	13026	2.10
rs7202877 T	1.00	1	13026	2.11
GSDMB rs2290400 T	1.03	1	13026	2.17
rs7221109 T	1.03	1	13026	2.24
PRKD2 rs425105 T	0.98	1	13026	2.21
SIRPG rs2281808 T	1.06	1	13026	2.35
LOC729980 rs5753037 C	1.03	1	13026	2.42
rs10517086 G	1.02	1	13026	2.48
rs9388489 A	1.03	1	13026	2.54
rs4948088 A	1.14	1	13026	2.90
SKAP2 rs7804356 T	0.99	1	13026	2.88
GLIS3 rs7020673 C	1.01	1	13026	2.92
GAB3 rs2664170 G	0.97	1	13026	2.84
rs380421 T	0.94	1	12900	2.67
LBP rs2232613 C	1.01	1	12900	2.69
CAPSL rs1445898 C	1.05	1	12900	2.83
IL7R rs6897932 C	1.03	1	12900	2.92
CFTR rs213950 G	0.96	1	12900	2.81
rs2666236 G	0.96	1	12200	2.70
ERBB3 rs2292239 T	1.21	1	12200	3.27
rs2542151 G	1.23	1	12200	4.00
rs9653442 C	1.06	1	12200	4.24
ADAD1 rs17388568 G	0.97	1	12200	4.11
rs7722135 C	1.02	1	12200	4.20
MAGI3 rs1230661 A	1.27	1	5000	5.34
BCL2L15 rs2358994 G	0.92	1	5000	4.90
CLEC16A rs2041670 G	1.07	1	5000	5.27
HLA-DQA1 rs9272346 G	0.39	1	5000	2.04
INS,INS-IGF2 rs689 A	1.12	1	1718	2.29
C8orf22 rs2292399 A	1.26	1	1356	2.89
rs6822844 G	1.06	1	1279	3.06
STAT4 rs7574865 T	1.68	1	638	5.14

Fig. 2: Contribution of individual alleles to overall risk LR of T1D of the reference genome. Alleles and their associated genes are listed on the left, ordered from top to bottom by the number of studies in which each was published and the total sum of cohort sizes across those papers. The LR of each independent SNP/allele is listed. A user of this figure could draw a horizontal line at a given threshold of belief, include and exclude alleles, and retrieve the accumulated LR at the right column and shown graphically in the middle. The central graph displays the change in accumulated LR, with darker squares representing more publications and larger squares representing larger sample size.



Genotype Test	LR	Studies	Samples	Mult(LR)
				1.00
IFIH1 rs1990760 TT	1.16	4	33090	1.16
CTLA4 rs3087243 AA	0.68	3	29326	0.78
PTPN22 rs2476601 GG	0.83	2	14618	0.65
CD69 rs4763879 AA	0.93	1	13026	0.61
C14orf181 rs1465788 CC	0.98	1	13026	0.60
rs7202877 TT	1.00	1	13026	0.60
PRKD2 rs425105 TT	0.98	1	13026	0.59
SIRPG rs2281808 CC	0.96	1	13026	0.57
rs9388489 AA	1.03	1	13026	0.58
rs4948088 CC	0.99	1	13026	0.58
GAB3 rs2664170 AA	1.01	1	13026	0.59
CD226 rs763361 CC	0.88	1	12900	0.51
rs380421 CT	1.00	1	12900	0.52
LBP rs2232613 CT	0.87	1	12900	0.45
CAPSL rs1445898 CT	0.96	1	12900	0.43
IL7R rs6897932 CT	0.95	1	12900	0.41
CFTR rs213950 GG	0.93	1	12900	0.38
rs2666236 GG	0.88	1	12200	0.33
ERBB3 rs2292239 GT	1.09	1	12200	0.36
CLEC2D rs3764021 CC	1.06	1	12200	0.38
C12orf30 rs17696736 AG	1.00	1	12200	0.38
CLEC16A rs12708716 AG	0.92	1	12200	0.35
rs2542151 GT	1.22	1	12200	0.43
rs9653442 CC	1.09	1	12200	0.47
rs7722135 CC	1.04	1	12200	0.49
HLA-DQA1 rs9272346 AG	0.39	1	5000	0.19
TCF7L2 rs7903146 CT	1.10	1	2422	0.21
INS,INS-IGF2 rs689 AT	0.59	1	1718	0.12
KIAA1109 rs4505848 AG	0.85	1	1279	0.10
rs6822844 GG	1.17	1	1279	0.12
STAT4 rs7574865 GT	1.44	1	638	0.17

Fig. 3: Contribution of individual genotypes to the overall risk LR of T1D for a previously published 40-year-old healthy Caucasian male. See Figure 3 for details on the graphical elements.

Comparing Figure 2 and 3, we found that the increased T1D risk in the reference genome was mainly due to a highly T1D-susceptible allele A at rs2476601, causing a R260W mutation in the intracellular tyrosine phosphatase (*PTPN22*). This SNP had been reported to increase the risk of T1D by 2 fold in more than nine studies (*28*-*31*). Comparing with the patient, the reference genome also has increased risk of T1D due to the lack of two T1D-protective alleles at rs3087243 in cytotoxic T-lymphocyte-associated protein 4 (*CTLA4*) (*32*) and at rs689 in the insulin (*INS*) (*28*). These three alleles increased the T1D risk for the reference genome by 6.8 fold comparing with our previously published patient. Interestingly, for rs2476601 in *PTPN22*, the T1D susceptible allele in the reference genome is the minor allele in most population. The 3,556 known disease-susceptible variants and many unknown ones especially rare variants could be potentially missed if only variants different from the reference were analyzed.

3.6 Disease-susceptible alleles deleted in the reference genome

The reference genome also contains a deletion at 2.7M SNPs with a dbSNP identifier in the dbSNP build 131 (*17*). We found that 16 SNPs that are known to associate with human diseases at these points of deletion. The clinical relevance of these missing base pairs is not clear.

4. Discussion

We identified 3,556 disease-susceptible variants including 15 rare variants (MAF<1%) in the reference human genome, which provides a useful tool for the annotation of personal genome sequences. Using a curated high-quality quantitative human disease-SNP association database, we assessed the likelihood ratio of increased risk over healthy population on 104 diseases for the reference genome and found the high predictive T1D risk with a R260W mutation in the intracellular tyrosine phosphatase (*PTPN22*). It reminded us that the reference genome was not from a regular person and was certainly not disease free. Although it had dramatically accelerated personal genome sequencing efforts, focusing on variants different from the reference will likely miss many disease causal variants including rare variants.

With the likely incoming deluge of 10,000 personal genome sequences arriving within the next two years, a method to estimate personal disease risk is urgently needed. Here, we described a method to estimate personal genetic risk using a likelihood ratio for each SNP as the relative frequency of the individual’s genotype in the diseased vs. healthy control populations. We further described a very simple method to treat multiple disease loci outside the linkage disequilibrium as independent genetic test, and estimated their combined effect. We acknowledge that assuming independence of tests is actually a different assumption than assuming that each variant contributes independently to risk. If each measured variant is viewed as an independent test probing disease state, this is arguably closer to our understanding of their use as markers associated with disease instead of actual causal variants (*22*). We admit that it is likely to be too simple to accurately model the risk of many common diseases, especially those like T1D, which are also influenced by unknown environmental and gene-environmental factors, and we are currently investigating different models to estimate combined effects.

The accurate assessment on personal disease risk is also dependent on the quality and coverage of the genotype/allele frequency in the disease and control population in the literature. We found that many studies, including genome-wide association studies (GWAS) only reported the odds ratio of disease risk between genotypes/alleles, and not their frequencies in the case and control population, which were required for the calculation of the likelihood ratio. For studies reporting both the odds ratio and the minor allele frequency in the control group, we recalculated their allele frequencies. We excluded studies reporting only the odds ratio, and we are investigating the possibility of estimating the genotype/allele frequencies in the control group using the data in the HapMap III project (*33*). There have been many debates on whether the aggregated genotype frequency data should be published in GWASs (*34*). Analyses showing association of a single biomarker with disease typically report very detailed characteristic of the populations studied; this is radically different from typical genetic association studies, which often report almost nothing about the subjects (*22*). Therefore, we strongly recommend the release of the genotype frequency in future GWAS studies as it is critical for us to quantitatively evaluate the disease-SNP association, enabling an accurate personal risk assessment.

We further found that many disease SNPs had been reported as the genotypes in the negative strand without indicating their strand directions. We had identified the strand direction by comparing the major/minor alleles in the study with the major/minor alleles in similar population in the HapMap projects. However, the identification process became difficult when the C/G or A/T alleles share similar frequencies. Therefore, we strongly recommend investigators to report the genotype frequencies in the case and control population and their strand direction in the future GWAS publications. With exponentially increasing personal genome sequences with phenotype information, we will likely to discover more rare causal variants and comprehensively predict personal risk on a variety of diseases.

5. Acknowledgements

This work was supported by Lucile Packard Foundation for Children's Health, National Institute of General Medical Sciences (R01 GM079719), National Library of Medicine (R01 LM009719), National Cancer Institute (R01 CA138256), and Howard Hughes Medical Institute. We thank Alex Morgan for R scripts for graphics and suggestions on the maximum likelihood ratio model. We thank Alex Skrenchuk and Boris Oskotsky for computer support, and thank Prajkta Bhide, Priyanka Korde, Anuj Kharnal, Harshal Darokar, and Priyanka Khandelwal from Optra Systems for curating disease-SNP association from the literature under contract.

References

1. J. R. Lupski *et al.*, *N Engl J Med* **362**, 1181 (Apr 1, 2010).
2. E. D. Pleasance *et al.*, *Nature* **463**, 191 (Jan 14, 2010).
3. R. Drmanac *et al.*, *Science* **327**, 78 (Jan 1, 2010).
4. D. Pushkarev, N. F. Neff, S. R. Quake, *Nat Biotechnol* **27**, 847 (Sep, 2009).
5. E. R. Mardis *et al.*, *N Engl J Med* **361**, 1058 (Sep 10, 2009).
6. J. I. Kim *et al.*, *Nature* **460**, 1011 (Aug 20, 2009).
7. K. J. McKernan *et al.*, *Genome Res* **19**, 1527 (Sep, 2009).
8. S. M. Ahn *et al.*, *Genome Res* **19**, 1622 (Sep, 2009).
9. T. J. Ley *et al.*, *Nature* **456**, 66 (Nov 6, 2008).
10. J. Wang *et al.*, *Nature* **456**, 60 (Nov 6, 2008).

11. D. R. Bentley *et al.*, *Nature* **456**, 53 (Nov 6, 2008).
12. D. A. Wheeler *et al.*, *Nature* **452**, 872 (Apr 17, 2008).
13. S. Levy *et al.*, *PLoS Biol* **5**, e254 (Sep 4, 2007).
14. M. Snyder, J. Du, M. Gerstein, *Genes Dev* **24**, 423 (Mar 1, 2010).
15. E. S. Lander *et al.*, *Nature* **409**, 860 (Feb 15, 2001).
16. E. A. Ashley *et al.*, *Lancet* **375**, 1525 (May 1, 2010).
17. S. T. Sherry *et al.*, *Nucleic acids research* **29**, 308 (Jan 1, 2001).
18. W. J. Kent *et al.*, *Genome Res* **12**, 996 (Jun, 2002).
19. O. Bodenreider, *Nucleic acids research* **32**, D267 (Jan 1, 2004).
20. *Nature* **426**, 789 (Dec 18, 2003).
21. http://www.childrens-mercy.org/stats/definitions/likelihood.htm.
22. A. A. Morgan, R. Chen, A. J. Butte, *Genome Med* **2**, 30 (2010).
23. L. R. Trevino *et al.*, *Nat Genet* **41**, 1001 (Sep, 2009).
24. S. Buervenich *et al.*, *Arch Neurol* **62**, 74 (Jan, 2005).
25. T. A. Manolio *et al.*, *Nature* **461**, 747 (Oct 8, 2009).
26. N. Kato *et al.*, *Hum Mol Genet* **17**, 617 (Feb 15, 2008).
27. X. Fan *et al.*, *Clin Pharmacol Ther* **82**, 187 (Aug, 2007).
28. C. Cervin *et al.*, *Diabetes* **57**, 1433 (May, 2008).
29. D. J. Smyth *et al.*, *Diabetes* **57**, 1730 (Jun, 2008).
30. E. Kawasaki *et al.*, *Am J Med Genet A* **140**, 586 (Mar 15, 2006).
31. L. A. Criswell *et al.*, *Am J Hum Genet* **76**, 561 (Apr, 2005).
32. J. M. Howson *et al.*, *Diabetologia* **50**, 741 (Apr, 2007).
33. D. M. Altshuler *et al.*, *Nature* **467**, 52 (Sep 2, 2010).
34. G. Church *et al.*, *PLoS Genet* **5**, e1000665 (Oct, 2009).

MATCHING CANCER GENOMES TO ESTABLISHED CELL LINES FOR PERSONALIZED ONCOLOGY

JOEL T. DUDLEY[1,2,3], RONG CHEN[2,3] AND ATUL J. BUTTE[2,3,*]

[1]Training Program in Biomedical Informatics and [2]Department of Pediatrics, Stanford University School of Medicine
Stanford, CA 94305, USA; [3]Luclie Packard Children's Hospital, Palo Alto, CA 94304, USA

The diagnosis and treatment of cancers, which rank among the leading causes of mortality in developed nations, presents substantial clinical challenges. The genetic and epigenetic heterogeneity of tumors can lead to differential response to therapy and gross disparities in patient outcomes, even for tumors originating from similar tissues. High-throughput DNA sequencing technologies hold promise to improve the diagnosis and treatment of cancers through efficient and economical profiling of complete tumor genomes, paving the way for approaches to personalized oncology that consider the unique genetic composition of the patient's tumor. Here we present a novel method to leverage the information provided by cancer genome sequencing to match an individual tumor genome with commercial cell lines, which might be leveraged as clinical surrogates to inform prognosis or therapeutic strategy. We evaluate the method using a published lung cancer genome and genetic profiles of commercial cancer cell lines. The results support the general plausibility of this matching approach, thereby offering a first step in translational bioinformatics approaches to personalized oncology using established cancer cell lines.

1. Introduction

Despite innovations in relevant diagnostics and therapeutics over the past decades, cancers remain among the leading causes of mortality in developed nations. Although many common molecular drivers of oncogenesis are known to exist, the majority of cancers are heterogeneous in their molecular characteristics, leading to disparities in response to standard cancer therapies. High-throughput sequencing technologies, with promise to offer complete DNA sequence profiling of cancer genomes, present novel opportunities understanding the unique molecular characteristics of tumors profiled in clinical populations. Knowledge of the unique molecular characteristics of a tumor, as detailed by its genomic sequence, could inform diagnosis, prognosis and treatment, thereby establishing a basis for personalized oncology.

In order to gain clinical utility from personal cancer genomes, the molecular characteristics latent in the cancer genomic sequence must be related to a broader biological context. Aberrations in a cancer genome, such as somatic variations in single nucleotides, copy number or novel gene fusions can serve as informative biomarkers that inform diagnosis, prognosis or treatment. For example, mutations in the epidermal growth factor receptor (*EGFR*) have been associated with response to gefitinib in non-small cell lung cancer (NSCLC)[1], and mutations in *KRAS* are known to be predictive of response to cetuximab in colon cancers[2]. Such markers have great clinical value when they are well characterized, however a complete genomics sequence of a cancer is likely to present many novel molecular aberrations that have minimal to no precedence in the

[*] Corresponding author: abutte@stanford.edu

literature. Furthermore, consideration for only a subset of the markers available in a fully sequenced cancer genome might miss molecular and biological features important for individualized treatment.

In order to assess functional correlates of disease progression or therapeutic susceptibility, approaches to personalized oncology need to consider molecular phenotypes salient in individual tumor biology along with the tumor's genotype. For example, expression levels of human epidermal growth factor receptor 2 (*HER2*) are predictive of response to trastuzumab[3], and various cellular metabolic features have been associated with tumor progression[4]. Ideally, it would be possible to functionally investigate these molecular phenotypes towards a personalized course of clinical care (e.g. test the response of several different chemotherapies to determine the best course of treatment), however it is not possible to conduct such clinical experimentation *in vivo* without placing the patient in danger of serious harm. One solution is to create autologous tumor cell lines from tumor tissue excised from the patient. However, the technical capacity to establish, maintain, and functionally test autologous cell lines is not at all common in most clinical settings, and therefore may not be as viable as a therapeutic option during the course of clinical care for cancer patients.

Here we describe a method to match a personal cancer genome with commonly studied commercially available cancer cell lines based on shared genetic profiles. Commercial cell lines serve as an attractive option for personalized oncology, because they are readily and economically available through commercial suppliers, and the pharmacological and biochemical characteristics of many of the available cancer cell lines are well reported in the literature. Furthermore, it has been shown that large collections of cancer cell lines can serve as "systems" to functionally characterize the pathophysiological properties of individual tumors[5]. Once a personal cancer genome is matched to a commercial cell line, it is possible that the cell line and the prior knowledge around that cell line could serve as an *in vitro* surrogate for clinical functional assessment of tumor biology. We offer a profile similarity approach that matches a cancer genome with commercial cell lines based on profiles of shared somatic variability at multiple loci. The method is assessed using data from a recently published genomic sequence of a lung cancer tumor, which was matched to genotyped cell lines found in the GlaxoSmithKline cancer cell line genomic profiling data.

2. Methods

2.1. *Data*

A set of somatic single nucleotide variants discovered in a NSCLC genome through paired genome sequencing in a lung cancer patient was obtained from the supplementary information provided by Lee et al[6]. Variant positions were mapped to dbSNP rsId's by genomic location. SNP genotype profiles for commercial cancer cell lines were downloaded from the Cancer Biomedical Informatics Grid (caBIG) website (https://cabig.nci.nih.gov/caArray_GSKdata/) via FTP. Allele frequency information was downloaded from data provided by the International HapMap Project Phase IIa[7]. We aggregated *in vivo* tumor xenograft screening data made available through the National Cancer Institute (NCI) Developmental Therapeutics Program (DTP) website (http://dtp.nci.nih.gov/webdata.html). The DTP screening data provides assessments of the anti-tumor efficacy of a wide range of chemical compounds evaluated across various clinical endpoints in human tumor xenograft models[8].

2.2. *Profile similarity*

A profile similarity metric was computed by comparing common variant loci between the cancer genome and the cancer cell line SNP profiles. The SNP profiles for the commercial cell lines only represent the genotype of various primary cancer cells, and therefore offer no means to distinguish somatic variants from neutral variation. We used allele frequency data from the HapMap project as a proxy for the normal baseline genotype. In this way, a locus was said to be a cancer-associated variation if it was not found to harbor the associated major allele for that locus found in the HapMap data. We then derived a multi-locus identity metric to compute a similarity score between to genomic profiles based on shared genotypes at somatically variant positions. For each locus an identity-by-similarity (IBS) score was computed based on the number of alleles shared between the profiles at that locus. The IBS score = 0 if no alleles are share, 1 if one allele is shared, or 2 if both profiles are homozygous for the same allele. The multi-locus profile identity score (*mIS*) was computed by summing the IBS scores across all shared loci and dividing by twice the number of common loci:

$$mIS_{i\,j} = \frac{\sum_{I=1}^{L} IBS_{ij}(g_i^I, g_j^I)}{2L}$$

Where L is the number of common variant loci between two genomic profiles i and j, and g_i^I is the genotype of the I^{th} locus in profile i, and g_j^I is the genotype of the I^{th} locus in profile j.

2.3. *Matching the lung cancer genome to cell lines*

To match the NSCLC genome to cell lines we computed the mIS score between the somatic variants and the SNP profiles for all cell lines found in the GSK data set. To estimate a p-value for mIS scores we computed a random distribution of mIS scores by constructing random genotype profiles by sampling randomly from the GSK data, and computing the mIS score between the NSCLC profile and the random genotype for one thousand iterations. The empirical p-value for an mIS score was computed as the proportion of mIS scores from the random distribution greater than the given mIS score.

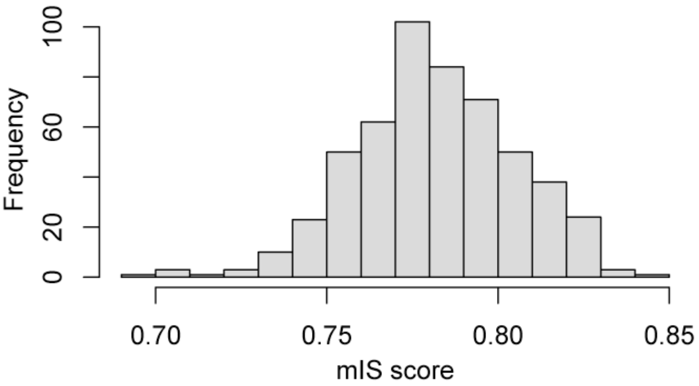


Figure 1. Distribution of genetic profile similarity scores between the lung cancer genome and GSK cancer cell lines.

2.4. *Clustering tumors by therapeutic profiles*

The DTP inhibition data was averaged by tumor type and compound. For each tumor type defined in the DTP data set, a chemotherapeutic profile was defined as the average inhibition for each compound against which the tumor was evaluated. A distance matrix was computed between tumors using the Pearson's correlation of compound inhibition response values. Only statistically significant correlations were retained. Hierarchical clustering was performed on the correlation distance matrix (1 - correlation) using the average agglomeration method. The significance of the compound inhibition clustering was assessed by multiscale bootstrap resampling across 1,000 bootstrap replicates using the *pvclust* package (http://www.is.titech.ac.jp/~shimo/prog/pvclust/). All computations were performed using the R language for statistical computing (http:// www.r-project.org).

3. Results

Using genomic location information we mapped 9,754 somatic single nucleotide variants and their genotypes to dbSNP rsId identifiers. Among these loci we found 391 that overlapped with the SNPs measured on the SNP array used to profile the cancer cell lines in the GSK data set. This common set of loci was used to compute the profile similarity between the NSCLC genome and the cancer cell lines. After computing mIS profile similarity scores (see methods) between the NSCLC genome and all cell lines profiled in the GSK data set, we find 16 cell lines to be significantly associated with the personal cancer genome by genetic profile (Table 1). The distribution of mIS scores across the GSK data set is shown in Figure 1. The top match among the GSK cancer cell lines is bladder carcinoma line J82. While other lung carcinomas are found among the top results, we also find non-obvious associations between various leukemias and lymphomas.

To explore the plausibility of these cell line associations, we obtained chemotherapeutic screening data from the NCI Developmental Therapeutics Program (DTP) and clustered tumors based on their response to various chemotherapies (Figure 2). Based on chemotherapy response profiles, we find that Lewis lung carcinomas, a model for non-small cell lung cancer, generally cluster with several leukemias and reticular (lymphoid) sarcoma, which is reflective of our cell line match results.

Table 1. Cancer cell lines from the GSK genomic profiling data set with genetic profiles significantly similar to the individual NSCLC genome based on mIS scores.

Cancer Type	Cell Line	mIS score	P-value
Carcinoma of Bladder	J82	0.84	$2.3x10^{-2}$
Acute T Cell Lymphoblastic Leukemia of Hematopoietic and lymphatic system	CCRFCEM	0.83	$3.3 x10^{-2}$
Lymphoma of Hematopoietic and lymphatic system	SR	0.83	$3.3 x10^{-2}$
Hodgkin Lymphoma of Hematopoietic and lymphatic system	RPMI6666	0.83	$3.3 x10^{-2}$
Lung Adenocarcinoma	NCIH1975	0.82	$4.8x10^{-2}$
Lung Adenocarcinoma	NCIH2228	0.82	$4.8x10^{-2}$
Atypical Carcinoid Tumor of Lung	NCIH720	0.82	$4.8x10^{-2}$
Small Cell Lung Carcinoma of Lung	NCIH524	0.82	$4.8x10^{-2}$
Burkitt Lymphoma of Hematopoietic and lymphatic system	MC116	0.82	$4.8x10^{-2}$
Burkitt Lymphoma of Hematopoietic and lymphatic system	1A2	0.82	$4.8x10^{-2}$
Carcinoma of Uterus	KLE	0.82	$4.8x10^{-2}$
Sarcoma of Bone	SW1353	0.82	$4.8x10^{-2}$
Carcinoma of Uterus	RL952	0.82	$4.8x10^{-2}$
Myeloma of Hematopoietic and lymphatic system	HuNS1	0.82	$4.8x10^{-2}$
Carcinoma of Breast	MT3	0.82	$4.8x10^{-2}$
Acute T Cell Lymphoblastic Leukemia of	CEMC1	0.82	$4.8x10^{-2}$

4. Discussion

In effort to relate a personal cancer genome to cancer cell lines for personalized oncology, we developed a profile similarity method that computes a similarity score between two genetic profiles based on shared alleles at somatically variant sites. We applied this method to a published non-small cell lung cancer genome and a set of SNP profiles from the GSK cancer genomic profiling data set. We found that the personal cancer genome could be significantly matched with 16 cell lines from the GSK data set by genetic profile (Table 1). While we find a number of lung cancer cell lines among these significant matches, we also find equally significant matches for non-lung cancers, including various Hodgkin lymphomas, leukemias and bladder cancer.

It is not immediately apparent why the lung cancer genome would be associated with these seemingly unassociated cancers. One possible explanation is that there are many passenger mutations after the cancer initiation event has started[9], and that the similarities are being driven by these mutations. Since passenger mutations are not necessarily causal, and could therefore confound variation based similarity metrics like the one used in this study. In this case, future work might involve inclusion of prior knowledge of cancer causal variants to reduce false positives, or look across multiple cancer genomes to understand patterns of earlier versus later mutations from a data-driven perspective.

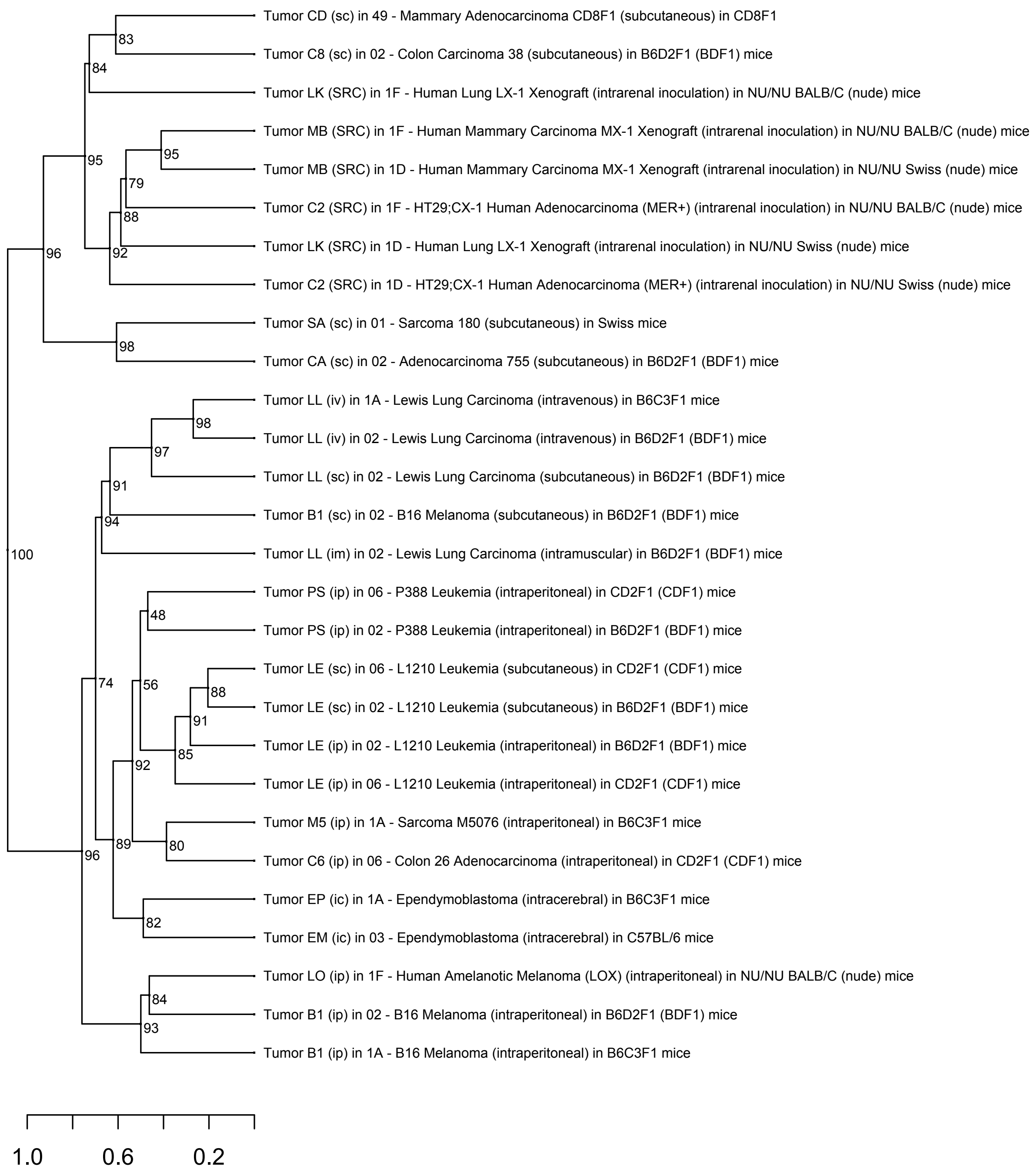

Tumor CD (sc) in 49 - Mammary Adenocarcinoma CD8F1 (subcutaneous) in CD8F1
Tumor C8 (sc) in 02 - Colon Carcinoma 38 (subcutaneous) in B6D2F1 (BDF1) mice
Tumor LK (SRC) in 1F - Human Lung LX-1 Xenograft (intrarenal inoculation) in NU/NU BALB/C (nude) mice
Tumor MB (SRC) in 1F - Human Mammary Carcinoma MX-1 Xenograft (intrarenal inoculation) in NU/NU BALB/C (nude) mice
Tumor MB (SRC) in 1D - Human Mammary Carcinoma MX-1 Xenograft (intrarenal inoculation) in NU/NU Swiss (nude) mice
Tumor C2 (SRC) in 1F - HT29;CX-1 Human Adenocarcinoma (MER+) (intrarenal inoculation) in NU/NU BALB/C (nude) mice
Tumor LK (SRC) in 1D - Human Lung LX-1 Xenograft (intrarenal inoculation) in NU/NU Swiss (nude) mice
Tumor C2 (SRC) in 1D - HT29;CX-1 Human Adenocarcinoma (MER+) (intrarenal inoculation) in NU/NU Swiss (nude) mice
Tumor SA (sc) in 01 - Sarcoma 180 (subcutaneous) in Swiss mice
Tumor CA (sc) in 02 - Adenocarcinoma 755 (subcutaneous) in B6D2F1 (BDF1) mice
Tumor LL (iv) in 1A - Lewis Lung Carcinoma (intravenous) in B6C3F1 mice
Tumor LL (iv) in 02 - Lewis Lung Carcinoma (intravenous) in B6D2F1 (BDF1) mice
Tumor LL (sc) in 02 - Lewis Lung Carcinoma (subcutaneous) in B6D2F1 (BDF1) mice
Tumor B1 (sc) in 02 - B16 Melanoma (subcutaneous) in B6D2F1 (BDF1) mice
Tumor LL (im) in 02 - Lewis Lung Carcinoma (intramuscular) in B6D2F1 (BDF1) mice
Tumor PS (ip) in 06 - P388 Leukemia (intraperitoneal) in CD2F1 (CDF1) mice
Tumor PS (ip) in 02 - P388 Leukemia (intraperitoneal) in B6D2F1 (BDF1) mice
Tumor LE (sc) in 06 - L1210 Leukemia (subcutaneous) in CD2F1 (CDF1) mice
Tumor LE (sc) in 02 - L1210 Leukemia (subcutaneous) in B6D2F1 (BDF1) mice
Tumor LE (ip) in 02 - L1210 Leukemia (intraperitoneal) in B6D2F1 (BDF1) mice
Tumor LE (ip) in 06 - L1210 Leukemia (intraperitoneal) in CD2F1 (CDF1) mice
Tumor M5 (ip) in 1A - Sarcoma M5076 (intraperitoneal) in B6C3F1 mice
Tumor C6 (ip) in 06 - Colon 26 Adenocarcinoma (intraperitoneal) in CD2F1 (CDF1) mice
Tumor EP (ic) in 1A - Ependymoblastoma (intracerebral) in B6C3F1 mice
Tumor EM (ic) in 03 - Ependymoblastoma (intracerebral) in C57BL/6 mice
Tumor LO (ip) in 1F - Human Amelanotic Melanoma (LOX) (intraperitoneal) in NU/NU BALB/C (nude) mice
Tumor B1 (ip) in 02 - B16 Melanoma (intraperitoneal) in B6D2F1 (BDF1) mice
Tumor B1 (ip) in 1A - B16 Melanoma (intraperitoneal) in B6C3F1 mice
83
84
95
95
79
88
92
96
98
98
97
91
94
100
48
74
56
88
91
85
92
89
80
96
82
84
93
1.0
0.6
0.2

Figure 2. Hierarchical clustering of tumors profiled by the National Cancer Institute Developmental Therapeutics Program based on their chemotherapeutic inhibition response profiles. Values at the inner nodes represent bootstrap p-values estimated by multiscale bootstrap resampling using 1,000 boostrap replicates.

Table 2. Gene-associated variants driving the similarity score between the personal lung cancer genome profile and the top cell-line match bladder carcinoma (J82). Both the lung cancer genome and J82 exhibit somatic variation at these positions and share at least one variant allele.

dbSNP rsID	Gene region	Gene symbol	Gene description
rs169124	intronic	BMP6	bone morphogenetic protein 6
rs13378247	intronic	ENOX1	ecto-NOX disulfide-thiol exchanger 1
rs11182675	intronic	NELL2	NEL-like 2 (chicken)
rs7824149	intronic	NECAB1	N-terminal EF-hand calcium binding protein 1
rs938726	intronic	EIF2C2	eukaryotic translation initiation factor 2C, 2
rs10983337	intronic	ASTN2	astrotactin 2
rs639839	intronic	NRG3	neuregulin 3
rs16907794	intronic	NELL1	NEL-like 1 (chicken)
rs2425562	intronic	PTPRT	protein tyrosine phosphatase, receptor type, T
rs2837583	intronic	DSCAM	Down syndrome cell adhesion molecule
rs10852799	intronic	DNAH9	dynein, axonemal, heavy chain 9
rs8024401	intronic	GABRG3	gamma-aminobutyric acid (GABA) A receptor, gamma 3
rs9555507	intronic	MYO16	myosin XVI
rs10483422	intronic	NPAS3	neuronal PAS domain protein 3
rs11158839	intronic	SLC8A3	solute carrier family 8 (sodium/calcium exchanger), member 3
rs9620769	intronic	TTC28	tetratricopeptide repeat domain 28
rs13112477	intronic	C4orf22	chromosome 4 open reading frame 22
rs6720773	intronic	COL6A3	collagen, type VI, alpha 3
rs10932540	intronic	VWC2L	von Willebrand factor C domain-containing protein 2-like
rs7550703	intronic	HHAT	hedgehog acyltransferase
rs1881410	intronic	LOC730124	similar to hCG2041586
rs4730038	intronic	LHFPL3	lipoma HMGIC fusion partner-like 3
rs2642484	intronic	CNTNAP2	contactin associated protein-like 2
rs7819262	intronic	TUSC3	tumor suppressor candidate 3
rs2910639	intronic	ADAMTS12	ADAM metallopeptidase with thrombospondin type 1 motif, 12
rs16870537	intronic	C7	complement component 7

Another explanation is that these associations might point towards some shared etiological or pathophysiological characteristics. Smoking is a well-known risk factor for lung cancers, leading to consistent genetic lesions observable in the genomes of lung cancer tumors. Smoking is also a substantial risk factor for bladder cancer[10], which is the top match in our results, and is also known to be associated with increased risk of various leukemia's and lymphomas[11]. Therefore the computed similarity between the lung cancer genome and these cell lines might have a basis in shared common genetic lesions due to smoking. It is also known that individuals affected by Hodgkin's lymphoma have an increased risk of lung cancer and non-Hodgkin lymphomas[12], suggesting a possible shared molecular pathophysiology among the various forms of cancer. Therefore, despite the fact that many of the matches are not of the same tumor type as the lung cancer genome, it is possible that they still might serve as functional surrogates for personalized clinical investigation.

To gain functional support for the plausibility of these cell line associations, we clustered tumors based on their response to various chemotherapies (Figure 2). Based on chemotherapy response profiles, we find that non-small cell lung cancer model tumors (Lewis lung) cluster significantly with both each other and other non-lung tumor types. A scatterplot of the chemotherapeutic profile similarity between a NSCLC tumor and leukemia is shown in Figure 3. Although the cell lines used in the DTP screening data set are not precise matches for the cell lines in the GSK data set, we can draw support for the notion that unrelated cancers such as lymphomas or leukemias could serve as functionally relevant clinical surrogates for lung cancer tumors.

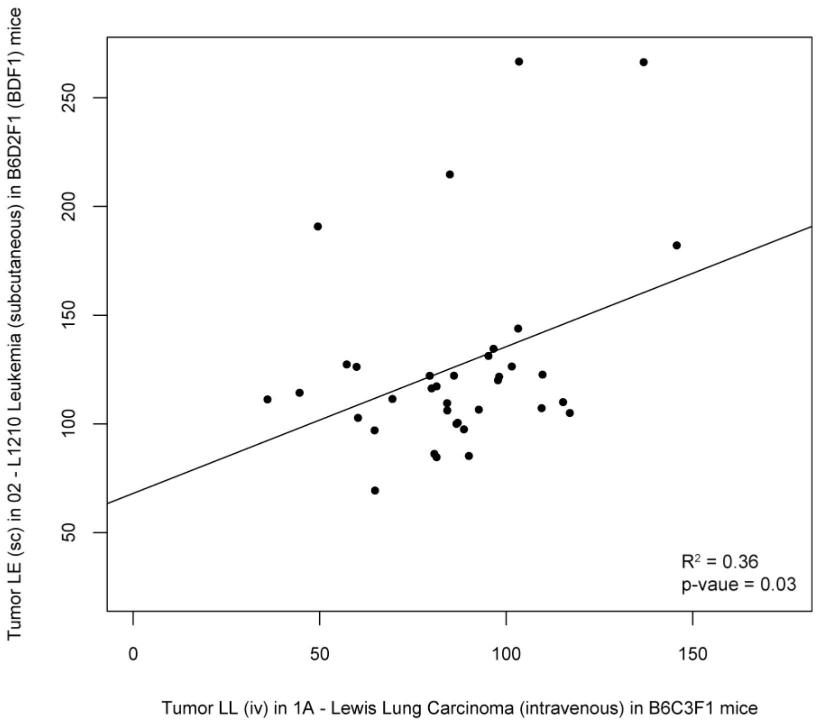


Figure 3. Comparison of the chemotherapeutic response profiles between a model of non-small cell lung cancer tumor model and a leukemia characterized in the NCI DTP data. The points represent the inhibition proportion (treatment/control) for a compound.

We find additional support for a plausible functional relationship through investigation of the variants driving the similarity between the lung cancer genome and cell lines. The best match in our data set was a bladder carcinoma cell line (J82). The gene associated variants shared between the lung cancer genome and the J82 cell line are shown in Table 2. Although all of these shared loci are intronic, it's still possible that they could be disrupting gene function through an effect on alternative splicing, or might serve as surrogate markers for mutational disruption of other loci in the same gene through linkage disequilibrium. Among these genes we find several known to be associated with cancers. *PTPRT*, a protein tyrosine phosphatase receptor, is a signaling molecule known to be implicated in oncogenic transformation in several different cancers[13], including colon cancer[14,15], glioma[16], and melanoma[17]. *NELL1* and *NELL2*, growth factor like protein thought to be involved in regulation of cell growth, has also been associated with multiple cancer types, including esophageal adenocarcinoma[18], colon cancer and Burkitt's lymphoma[19]. *TUSC3*, a putative tumor suppressor gene, has been associated with pancreatic cancer[20], prostate cancer[21] and ovarian cancer[22]. It's possible that these pleiotropic oncogenes are driving the similarity relationship between the lung cancer genome and J82 based on common patterns of oncogenic mutation. Several other genes underlying this similarity are not known to be oncogenic, however variants in *BMP6*, *COL6A3*, *C7*, *GABRG3* and *NRG3* are known to be associated with various complex and Mendelian diseases.

We acknowledge several limitations in our approach. Foremost, we recognize that since the GSK cell lines were profiled by SNP microarray, that the analysis was appreciably constrained to only the loci measured on the array platform. Future work might employ sophisticated imputations algorithms to expand the genotype profiles in the GSK data set, but ideally full genome sequencing data for these cell lines would likely be necessary for clinical application of this approach. We also acknowledge that the DTP chemotherapeutic profiling data can only offer indirect support for functional associations between these cell lines, as many of the cell lines profiled in the GSK data set are not represented in the NCI DTP screening data set. Efforts are needed to comprehensively characterize the chemotherapetuic response profiles of these cell lines and to provide a machine-readable representation of these data in the public domain.

Future work in this area will incorporate improved similarity metrics that give added importance to somatic variations more likely to play a causal role in tumorigenesis or metastasis, such as mutations in evolutionary conserved regions, or in loci known to act as expression quantitative trait loci (eQTLs) for genes associated with oncogenesis. More importantly, future work should incorporate experimental validation of predicted cell line matches to test whether or not the predicted cell line match exhibits clinical characteristics (e.g. chemotherapeutic response) similar to the individual tumor genome to which it was matched. Developments in this area will provide novel directions in personalized oncology that leverage the clinical, economic, and scientific benefits of well studied and characterized commercial cancer cell lines.

Acknowledgements

JTD is supported by the Graduate Training in Biomedical Informatics grant (R01 LM009719) from the National Library of Medicine. AJB is supported by the National Cancer Institute (R01 CA138256), the Lucile Packard Foundation for Children's Health and the Hewlett Packard Foundation. We thank Alex Skrenchuk and Boris Oskotsky from Stanford University for computer support.

References

1. Kobayashi, S., *et al.* EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N Engl J Med* **352**, 786-792 (2005).

2. Lievre, A.*, et al.* KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res* **66**, 3992-3995 (2006).
3. Hudis, C.A. Trastuzumab--mechanism of action and use in clinical practice. *N Engl J Med* **357**, 39-51 (2007).
4. Tennant, D.A., Duran, R.V. & Gottlieb, E. Targeting metabolic transformation for cancer therapy. *Nat Rev Cancer* **10**, 267-277 (2010).
5. Neve, R.M.*, et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515-527 (2006).
6. Lee, W.*, et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).
7. International HapMap, C. The International HapMap Project. *Nature* **426**, 789-796 (2003).
8. Teicher, B.A. & Andrews, P.A. *Anticancer drug development guide : preclinical screening, clinical trials, and approval*, (Humana Press, Totowa, N.J., 2004).
9. Greenman, C.*, et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).
10. Boffetta, P. Tobacco smoking and risk of bladder cancer. *Scand J Urol Nephrol Suppl*, 45-54 (2008).
11. Willett, E.V., O'Connor, S., Smith, A.G. & Roman, E. Does smoking or alcohol modify the risk of Epstein-Barr virus-positive or -negative Hodgkin lymphoma? *Epidemiology* **18**, 130-136 (2007).
12. van Leeuwen, F.E.*, et al.* Increased risk of lung cancer, non-Hodgkin's lymphoma, and leukemia following Hodgkin's disease. *J Clin Oncol* **7**, 1046-1058 (1989).
13. Lee, J.W.*, et al.* Mutational analysis of PTPRT phosphatase domains in common human cancers. *APMIS* **115**, 47-51 (2007).
14. Zhao, Y.*, et al.* Identification and functional characterization of paxillin as a target of protein tyrosine phosphatase receptor T. *Proc Natl Acad Sci U S A* **107**, 2592-2597 (2010).
15. Ruivenkamp, C.A.*, et al.* Ptprj is a candidate for the mouse colon-cancer susceptibility locus Scc1 and is frequently deleted in human cancers. *Nat Genet* **31**, 295-300 (2002).
16. Norman, S.A., Golfinos, J.G. & Scheck, A.C. Expression of a receptor protein tyrosine phosphatase in human glial tumors. *J Neurooncol* **36**, 209-217 (1998).
17. Yu, J.*, et al.* Tumor-derived extracellular mutations of PTPRT /PTPrho are defective in cell adhesion. *Mol Cancer Res* **6**, 1106-1113 (2008).
18. Jin, Z.*, et al.* Hypermethylation of the nel-like 1 gene is a common and early event and is associated with poor prognosis in early-stage esophageal adenocarcinoma. *Oncogene* **26**, 6332-6340 (2007).
19. Kuroda, S.*, et al.* Biochemical characterization and expression analysis of neural thrombospondin-1-like proteins NELL1 and NELL2. *Biochem Biophys Res Commun* **265**, 79-86 (1999).
20. Bashyam, M.D.*, et al.* Array-based comparative genomic hybridization identifies localized DNA amplifications and homozygous deletions in pancreatic cancer. *Neoplasia* **7**, 556-562 (2005).
21. Bova, G.S.*, et al.* Physical mapping of chromosome 8p22 markers and their homozygous deletion in a metastatic prostate cancer. *Genomics* **35**, 46-54 (1996).
22. Pils, D.*, et al.* Five genes from chromosomal band 8p22 are significantly down-regulated in ovarian carcinoma: N33 and EFA6R have a potential impact on overall survival. *Cancer* **104**, 2417-2429 (2005).

USE OF BIOLOGICAL KNOWLEDGE TO INFORM THE ANALYSIS OF GENE-GENE INTERACTIONS INVOLVED IN MODULATING VIROLOGIC FAILURE WITH EFAVIRENZ-CONTAINING TREATMENT REGIMENS IN ART-NAÏVE ACTG CLINICAL TRIALS PARTICIPANTS

BENJAMIN J. GRADY[1], ERIC S. TORSTENSON[1], PAUL J. MCLAREN[2], PAUL I.W. DE BAKKER[2], DAVID W. HAAS[3], GREGORY K. ROBBINS[4], ROY M. GULICK[5], RICHARD HAUBRICH[6], HEATHER RIBAUDO[7] AND MARYLYN D. RITCHIE[1*]

[1]Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232, USA
[2]Broad Institute, Harvard University, Cambridge, Massachusetts 02138, USA
[3]Departments of Medicine, Microbiology & Immunology, Vanderbilt University, Nashville, TN 37232, USA
[4]Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[5]Department of Medicine, Cornell University, New York, NY 10065, USA
[6]Department of Medicine, University of California, San Diego, San Diego, CA 92103, USA
[7]Department of Biostatistics, Harvard University, Boston, MA 02115, USA
**Corresponding Author Email: ritchie@chgr.mc.vanderbilt.edu*

Personalized medicine is a high priority for the future of health care. The idea of tailoring an individual's wellness plan to their unique genetic code is one which we hope to realize through the use of pharmacogenomics. There have been examples of tremendous success in pharmacogenomic associations however there are many such examples in which only a small proportion of trait variance has been explained by the genetic variation. Although the increased use of GWAS could help explain more of this variation, it is likely that a significant proportion of the genetic architecture of these pharmacogenomic traits are due to complex genetic effects such as epistasis, also known as gene-gene interactions, as well as gene-drug interactions. In this study, we utilize the Biofilter software package to look for candidate epistasis contributing to risk for virologic failure with efavirenz-containing antiretroviral therapy (ART) regimens in treatment-naïve participants of AIDS Clinical Trials Group (ACTG) randomized clinical trials. A total of 904 individuals from three ACTG trials with data on efavirenz treatment are analyzed after race-stratification into white, black, and Hispanic ethnic groups. Biofilter was run considering 245 candidate ADME (absorption, distribution, metabolism, and excretion) genes and using database knowledge of gene and protein interaction networks to produce approximately 2 million SNP-SNP interaction models within each ethnic group. These models were evaluated within the PLATO software package using pair wise logistic regression models. Although no interaction model remained significant after correction for multiple comparisons, an interaction between SNPs in the *TAP1* and *ABCC9* genes was one of the top models before correction. The *TAP1* protein is responsible for intracellular transport of antigen to MHC class I molecules, while *ABCC9* codes for a transporter which is part of the subfamily of ABC transporters associated with multi-drug resistance. This study demonstrates the utility of the Biofilter method to prioritize the search for gene-gene interactions in large-scale genomic datasets, although replication in a larger cohort is required to confirm the validity of this particular *TAP1-ABCC9* finding.

1. Introduction

1.1. *The HIV pandemic*

Human Immunodeficiency Virus (HIV) Type 1 infection has been in a state of pandemic for several years. The 2008 UNAID report estimates that 33 million people are currently infected, with approximately 3 million new infections during the year of 2008 [1]. Within regions in Sub-Saharan Africa, the prevalence of HIV-1 infection rises as high as 25-30% [1]. Because there is no cure for HIV-1 infection, one of the best tools available to combat the epidemic currently is antiretroviral therapy (ART). ART helps with treating those individuals already infected and helps to reduce the chance of spreading the disease[2]. ART consists of a regimen of two or three antiretroviral drugs and is successful in drastically increasing the lifespan of HIV-1 infected individuals and improving their quality of life[3]. By reducing the amount of virus circulating freely in the blood of an infected person, ART also greatly decreases the probability of transmitting the virus through sexual contact[4], and child birth [5]. Despite the benefits of ART for its use in fighting HIV, there are unfortunately several issues that accompany the use of the therapy. Arguably the most significant issue among these is the prevalence of adverse drug reactions (ADR) and the failure of the drug to suppress viral load. Adverse reactions to antiretroviral drugs range from skin rash and nausea to neurologic impairment and fatal hypersensitivity, as is sometimes seen in response to the drug abacavir [6]. ADRs contribute to ineffectiveness of ART by reducing adherence to drug regimens and requiring temporary discontinuation of treatment [7]. The failure of a drug to suppress viral load in a patient is known as virologic failure [8]. Virologic failure refers either to initial inefficacious response to the drug and a failure to ever reach a controlled viral load or to the phenomenon whereby viral load rebounds subsequent to reaching a controlled level.

1.2. *Pharmacogenomics and HIV treatment*

The way in which people respond to drug treatment has been shown, in many cases, to be influenced by their genetics. The field of pharmacogenomics attempts to discover the exact genetic variants which predict success, failure or ADR in response to treatment. There have been successes in identifying genetic polymorphisms which explain large proportions of variance in drug response. Approximately 20-30% of the variance in initial dosing of the anti-coagulant warfarin, for example, can be explained by variation in the gene *VKORC1*[9], which codes for vitamin K epoxide reductase complex subunit 1. Vitamin K epoxide reductase creates the enzymatically active form of vitamin K [10] which is in turn extremely important in modulating the function of proteins involved in blood clotting. For this reason, it makes biological sense that a polymorphism which affects the expression of VKORC1 would also affect how much warfarin is necessary to prevent over-clotting. Arguably the most significant pharmacogenomic discovery has been made in the field of HIV ART. Hypersensitivity reaction (HSR) in response to the nucleoside reverse-transcriptase inhibitor (NRTI) abacavir, a commonly-used drug in ART regimens, has been shown to be strongly tied to *HLA* genotype. *HLA-B*5701* genotype has a 100% negative predictive value (NPV) for predicting HSR from abacavir [11, 12]. As a result of this relationship, *HLA-B*5701* has become one of the first genetic tests approved by the FDA for use in determining risk prior to prescription of a drug. Although the abacavir story represents the pinnacle of pharmacogenomic discovery and there may not be another single genetic polymorphism with 100% NPV for ART in the future, there are still many possibilities for utilizing genetic prediction models in determination of the optimal ART drug regimen to prescribe in order to control HIV. It might be that a combination of genetic variants in concert would best predict antiretroviral drug response.

1.3. *Genetic interactions*

Decades of research into the pharmacokinetics of drug metabolism have shown that the enzymes which process and transport pharmaceuticals function as part of highly-interconnected networks [13]. For example, studies have shown that many drugs, including phenytoin [14] and irinotecan [15], can be metabolized, activated, or deactivated by more than one enzyme[16]. It is as a result of this complementation that it is reasonable to expect the necessity of multiple genetic polymorphisms to experience a large change in the resulting phenotype. The phenomenon of gene-gene interaction, or epistasis as it is often referred to in the field of genetic epidemiology, has been a subject of much discussion over the past decade [17-19]. Although the term epistasis was coined separately by Bateson [20] and Fisher [21] in the early 20th century to refer to the effect of one gene “masking” another’s effect or a non-additive effect of multiple elements observed simultaneously, respectively, the necessary technology to explore its presence has only recently been developed. The HapMap project, the sequencing of the human genome, and the steady increase in computational power have been the driving factors in the ability to analyze genetic data for gene-gene interaction effects. Despite the rising computational power available, genotyping technology has far out-paced the ability to exhaustively analyze multi-locus genetic effects for genome-wide association study (GWAS) data. To search exhaustively for epistasis between two single nucleotide polymorphisms (SNPs) in a current GWAS containing 1 million SNPs would require 5×10^{11} tests. Although it is still possible to perform this pair wise exploration by utilizing parallel computation, it is clear that with the advent of whole-exome and whole-genome sequencing technology as a primary source for genetic information in association studies in the near future, an alternative to exhaustive searches must be found. One such solution is that of biasing the search using prior knowledge to search for combinations of genes that are likely to interact biologically. The Biofilter tool [22] was developed to use databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Protein Family database (PFAM) in order to build SNP-SNP models based on known interactions between genes and proteins in curated pathways and networks. Especially in a field such as pharmacogenomics, in which the knowledge of the drug metabolism networks is extensive, enriching the search for epistasis with knowledge from known biological interactions could prove valuable. Not only does this alleviate the issues of computational complexity, but it also substantially reduces the number of tests and associated multiple-comparison issues. As opposed to considering 10's or 100's of billions of two-way interaction models, one would search a more reasonable subset of a few million models with a solid biological basis. The time necessary to perform the subsequent statistical analysis declines from days to hours, using a single processor. One concern of pursuing only a biologically-informed subset of interaction models is the possible loss of novel significant interactions during filtering. Due to long-standing knowledge in pharmacology, the potential reduction in noise outweighs the concern.

2. Methods

2.1. *Study population*

DNA samples in the current study come from individuals who were randomized to receive efavirenz (in multidrug ART regimens) in the AIDS Clinical Trials Group (ACTG) randomized clinical trials (RCT) ACTG 384, A5095 and A5142 and were collected under protocol A5128 [23]; study designs are described in depth elsewhere [24-30]. ACTG 384 [29, 30] and A5095 [24-26] were double-blind, multicenter RCTs designed to test the effectiveness of differing ART drug regimens. Of the 980 individuals enrolled into ACTG 384, 526 were consented for DNA extraction and 347 of those with DNA available were on efavirenz-containing regimens. A5095 enrolled 1147 subjects for comparison of protease inhibitor-sparing regimens. Of the enrollment in A5095, a total of 600 individuals were available who both consented for DNA and had ART containing efavirenz. The final study used in this multi-study analysis was A5142 [27, 28]. This ACTG study was a Phase III comparison of 3 ART regimens. Of the 757 participants of A5142, 411 were randomized to receive ART containing efavirenz and provided DNA samples.

Table 1. Sample size broken down by study and race ethnicity grouping.

			Study			
			ACTG 384	**A5095**	**A5142**	**Total**
Pre-QC	**Total**	Cases	45	124	97	266
		Controls	302	476	314	1092
Post-QC	**Total**	Cases	38	100	59	197
		Controls	228	319	160	707
	White	Cases	16	34	24	74
		Controls	116	163	78	357
	Black	Cases	18	47	30	95
		Controls	69	97	49	215
	Hispanic	Cases	4	19	5	28
		Controls	43	59	33	135

The total number of individuals available with DNA samples and GWAS genotyping across these three studies was 1358. Self-described ethnicity of the combined study population reveals the study population to consist of 45% white (N = 606), 34% non-Hispanic black (N = 459), 19% Hispanic (N = 265) and 2% other (N = 28). After quality control (QC) and exclusions were applied, 904 participants remained available for analysis. Of this 904, 48% (N = 431) were non-Hispanic white, 34% were non-Hispanic black (N = 310), and 18% (N = 163) were Hispanic. The endpoint used in this study was virologic failure as defined by a spike in viral load above 200 copies/mL after achieving viral load less than 200 copies/mL on ART. Individuals who experienced virologic failure on efavirenz are categorized as cases while those who did not are categorized as controls (**Table 1**).

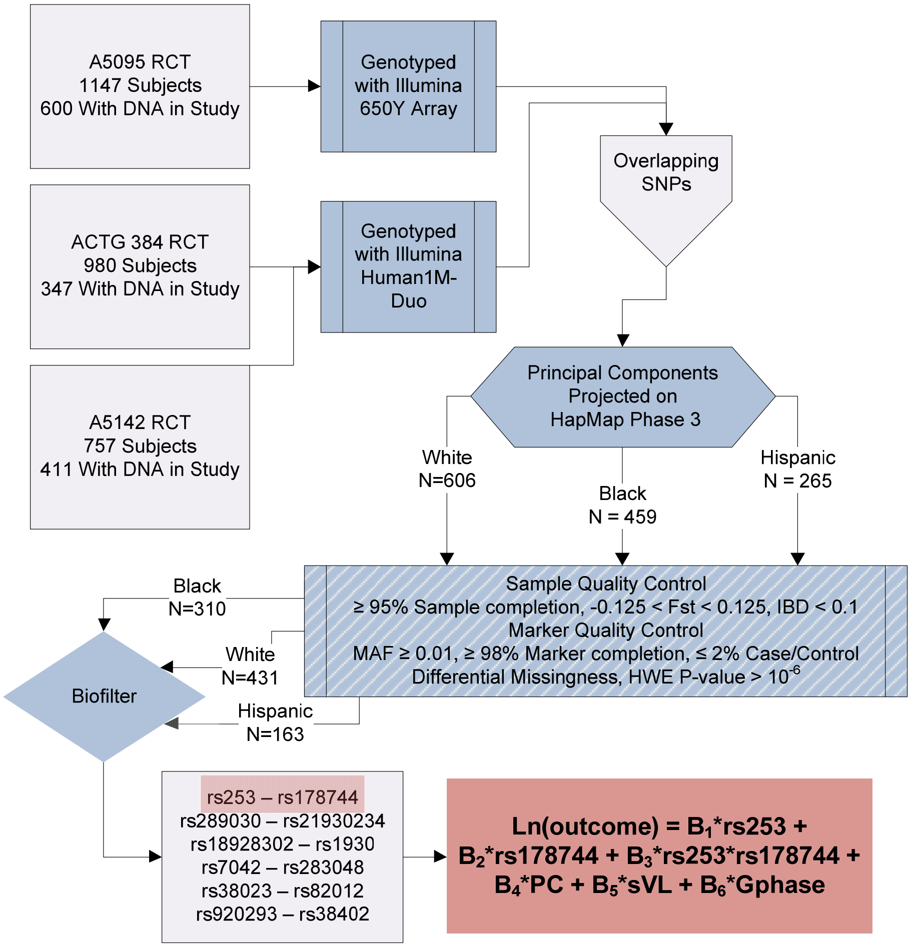


Figure 1. An outline of the analysis plan used in this study. A more detailed description of the Biofilter step is available in Figure 2.

2.2. *Genotyping and quality control exclusions*

Individuals from the ACTG 384 study were genotyped on the Illumina 650Y array while those from A5142 and A5095 were genotyped with the Illumina Human1M-Duo platform (Illumina, Inc. San Diego, CA). In combining data from these two platforms, only the SNPs which overlapped were used for this analysis. Principal components analysis was performed referencing the HapMap phase 3 sample data to map each individual back to one of three major ethnic groups: white, black, and Hispanic. There was greater than 95% concordance between self-reported ethnicity and that found through principal components analysis. Within each race stratum, quality control was performed to filter out samples and SNPs of low quality (Figure 1). Samples with low genotyping rate (<95%), high or low heterozygosity (inbreeding coefficient > 0.125 or < -0.125) and related individuals (IBD estimate > 0.1) were removed. SNPs with missingness > 2%, large deviations from Hardy-Weinberg equilibrium ($p < 10^{-6}$) and those with differential missingness between cases and controls > 2% were removed from analysis.

2.3. *Biofilter*

The Biofilter[22] was developed to provide prior biological knowledge to influence the search for gene-gene interactions in large-scale data. Given a set of variants, Biofilter first maps the SNPs back to genes based on gene definitions in Ensembl and then builds models using disease-dependent (i.e. those biological associations previously known with respect to the trait under investigation) or disease-independent relationships (i.e. known biological interactions with no particular association to the trait under consideration). A unique option for the Biofilter is to provide a personally curated list of genes based on expert knowledge of the phenotype under study as a starting point, using that list to search both disease-dependent and disease-independent data sources to map all other genes that are related to the genes in the curated list. Based on user options, the Biofilter can query the set of databases, which currently includes KEGG, PFAM, Reactome, DIP, PFAM, GO, and NetPath, to establish groups of interacting genes. Once these groups are established, SNP-SNP interaction models are created by exhaustively pairing two SNPs from two genes in the group. Biofilter allows flexibility in choosing how restrictive the creation of interaction models will be. For example, when inputting a list of self-curated genes, the user has the option to ensure that at least one of the SNPs comes from a gene in the list. Alternatively, restrictions can be relaxed to allow models with SNPs from other genes in the same group or even pathway as those genes in the list. As shown in Figure 2, we provided a list of 245 absorption, distribution, metabolism and elimination (ADME) genes which were curated by the authors and allowed for the inclusion of SNPs which were within 10kb of the gene boundaries. Interactions were restricted to allow only those models for which at least one of the SNPs in the model belonged to a gene in the list – although the search was conducted in the disease-independent databases. Two-SNP interaction models were generated separately for non-Hispanic white (henceforth referred to as white), non-Hispanic black (henceforth referred to as black), and

Hispanic ethnic groups, while those participants self-describing as other races were excluded. We used all six databases currently integrated in the Biofilter to generate our SNP-SNP models.

2.4. *Statistical analysis*

All statistical analyses were performed using the Platform for the Analysis, Translation and Organization of large scale data (PLATO) software package (http://chgr.mc.vanderbilt.edu/ritchielab/subscriptions) [31]. PLATO is a scaffold which allows for recoding, quality control, and analysis of data as part of a pipeline. The Biofilter models were used as input for PLATO. The statistical analyses performed used logistic regression to assess the risk of each pair of interaction models. Logistic regression models included terms for each SNP separately and a term for multiplicative interaction. In addition, variables deemed important with respect to the outcome of virologic failure were included as covariates in the model. Principal components vectors were utilized to adjust for population substructure within each racial group, as might exist between northern and southern European white individuals or African Americans. Indicator variables for genotyping phase and baseline viral load (≥ or < 100,000 copies/mL) were also incorporated. Regression analysis was performed separately within each ethnic group as defined by principal components analysis.

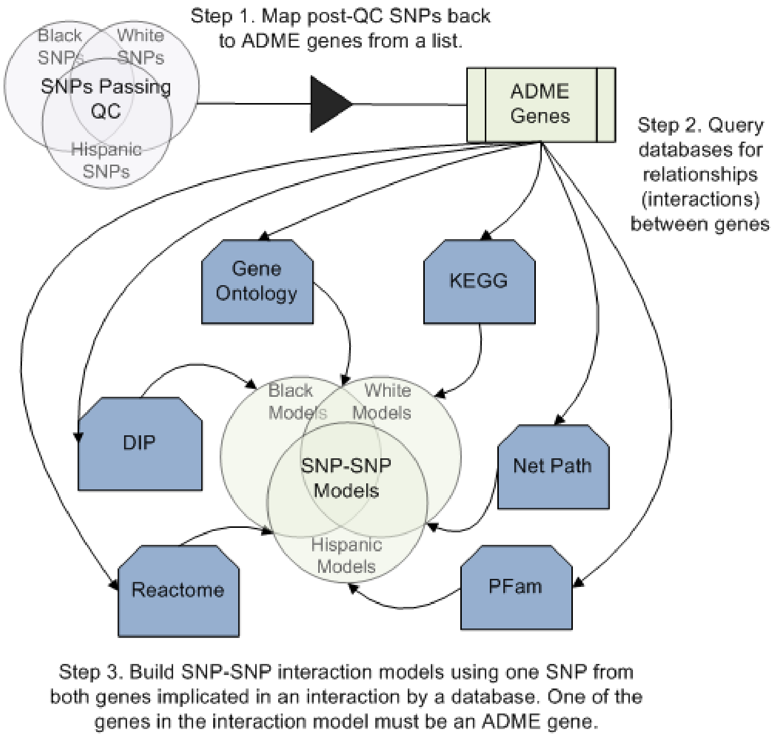

Step 1. Map post-QC SNPs back to ADME genes from a list.
Black SNPs
White SNPs
SNPs Passing QC
Hispanic SNPs
ADME Genes
Step 2. Query databases for relationships (interactions) between genes
Gene Ontology
KEGG
DIP
Black Models
White Models
SNP-SNP Models
Hispanic Models
Net Path
Reactome
PFam
Step 3. Build SNP-SNP interaction models using one SNP from both genes implicated in an interaction by a database. One of the genes in the interaction model must be an ADME gene.

Figure 2. A schematic of the procedure involved in a Biofilter run.

3. Results

Genome-wide genotyping of 1358 AIDS Clinical Trials Group (ACTG) participants with exposure to the NNRTI efavirenz was conducted to elucidate the genetic basis of virologic failure. Race-stratification was performed using principal components analysis based on HapMap phase 3 samples. After quality control processes, 904 individuals remained. The Biofilter software tool was used to take a list of 245 ADME genes and build putative gene-gene interactions based on biological knowledge provided by KEGG, DIP, Pfam, Net Path, Reactome, and Gene Ontology. The SNPs from each ethnic group were then mapped back to these ADME genes and SNP-SNP models were created by taking one SNP from each gene in a proposed gene-gene interaction. Running Biofilter resulted in 2,144,157 models to evaluate in whites, 2,471,201 models in blacks, and 2,099,614 models in Hispanics. These models were derived from a total of 33067, 35764 and 32698 SNPs for white, black and Hispanic groups respectively. If all two-way interactions between these SNPs were exhaustively tested, it would result in the evaluation of 546 million models for the white group and 638 million and 534 million models for the black and Hispanic groups respectively. The differences in model number between ethnic groups are due to race-stratified quality control. SNP-SNP models from Biofilter were passed to PLATO[31] to perform logistic regression analysis. Due to the highly correlated nature of many of the interaction models, a Bonferroni correction would too conservative for correcting for multiple testing. Instead, a false-discovery rate (FDR) correction was applied using the qvalue package available in R. No interaction models were found to be significant at an FDR level of 0.10, although the most significant interactions were significant at an FDR level of 0.45. The interaction models with lowest p-values are shown in Table 2.

Table 2. Most significant interaction models resulting from gene-gene interaction analysis.

SNP1	SNP2	Model P-value	Interaction P-value	SNP1 Odds Ratio	SNP2 Odds Ratio	Interaction Odds Ratio
rs2318785 (*NME2*)	rs2157597 (*NME7*)	1.69E-06	5.98E-07	0.257	0.253	4.732
rs2318785 (*NME2*)	rs12118611 (*NME7*)	1.69E-06	5.98E-07	0.253	0.257	4.732
rs2318785 (*NME2*)	rs12121994 (*NME7*)	1.68E-06	6.15E-07	0.247	0.263	4.769
rs2318785 (*NME2*)	rs17349439 (*NME7*)	3.30E-06	1.20E-06	0.265	0.277	4.487
rs2318785 (*NME2*)	rs6703463 (*NME7*)	1.48E-05	3.93E-06	0.258	0.289	3.884
rs2318785 (*NME2*)	rs12744184 (*NME7*)	7.46E-06	6.46E-06	0.253	0.306	3.821
rs735883 (*TAP1*)	rs1283807 (*ABCC9*)	4.38E-06	9.05E-06	2.245	3.236	0.154
rs735883 (*TAP1*)	rs1352909 (*ABCC9*)	4.61E-06	9.76E-06	2.231	3.205	0.155
rs735883 (*TAP1*)	rs4148665 (*ABCC9*)	4.83E-06	9.89E-06	2.230	2.850	0.172
rs735883 (*TAP1*)	rs1283798 (*ABCC9*)	1.04E-05	1.00E-05	2.395	2.482	0.240

4. Discussion

As genotyping technologies progress and we move into the era of whole-genome sequencing, the need to improve analysis schemes is ever-present. This is especially true when gene-gene, gene-environment, and gene-drug interactions are concerned. Allowing our biological knowledge of gene and protein network dynamics to guide the search for the genetic basis of disease is a promising solution to this dilemma. While our current state of biological knowledge is limited, and that knowledge-base will continue to grow and develop over time, if we develop techniques that use the information we have, while still exploring novel interactions, we have a greater chance for success. By narrowing the dimensions of the search space, the computational complexity of the problem becomes much more amenable to current analytical techniques. In addition, interpretation of results is more straightforward. We utilized a list of 245 genes involved in absorption, distribution, metabolism and elimination of drugs and their metabolites to focus the search for gene-gene interactions associated with virologic failure during HIV treatment with efavirenz. Although there were no gene-gene interactions which remained significant after correction for multiple testing, this could be related to the small sample size present in this study. Due to race-stratification, the largest group in the analysis had 74 cases and 357 controls. But the development of this analytic pipeline and software tools will be immensely useful for future analyses.

The interactions which appeared most significant in the results of the logistic regression analysis occur between a SNP - rs2318785 - in the *NME2* gene and multiple SNPs in the *NME7* gene. Both *NME2* and *NME7* are part of the NDK family, coding for nucleoside diphosphate kinase enzymes involved in the synthesis of non-ATP nucleoside triphosphates. Although it is not readily apparent as to why purine and pyrimidine metabolism would be involved in the predisposition towards virologic failure, it is possible that this could represent novel biological knowledge in this field. Currently known reasons for virologic failure include lack of adherence to drug regimen, presence of drug resistance mutations in the HIV strain, and drug interactions which might limit efficacy. In the absence of environmental heterogeneity, little is known about the etiology of virologic failure. Small sample size precludes our ability to draw conclusions about the role of nucleoside triphosphate metabolism on risk for virologic failure. Other SNP interaction models which were among the most significant results involve a SNP in the *TAP1* gene - rs735883 - and multiple SNPs in the *ABCC9* gene. *TAP1* encodes a transporter responsible for the shuttling of antigen into the endoplasmic reticulum for association with MHC class I while *ABCC9* is part of the MRP subfamily of ABC transporters associated with multi-drug resistance and codes for a protein thought to be a subunit of a pancreatic potassium channel responsible for drug-binding modulation of the channel. It could be that down-regulation of *TAP1* through mutation prevents proper immune response to the virus even after it has been affected by NNRTI action and this allows it to rebound during treatment. The results of the current study require validation with larger sample size before any firm conclusion can be drawn. The current results are meant to demonstrate the pipeline for analysis and the general approach rather than attempting to draw general statements regarding true biological associations with HIV therapy.

Despite the lack of statistical power to elucidate a significant genetic interaction, this study shows the promise of the use of Biofilter for focusing the search for gene-gene interactions during large-scale genetic association studies. The number of polymorphisms typed in association studies is nearing our limits to perform exhaustive explorations of two-way interactions during analysis. Reducing the set of interesting models to evaluate presents itself as a capable alternative. Utilizing Biofilter to provide the set of interesting models and PLATO to perform analysis has at least three advantages over traditional exhaustive gene-gene interaction analysis. First, it partially alleviates issues of multiple comparisons. Second, interpretation of results is significantly eased due to models construction. Third, the use of regression framework allows for the adjustment of the analysis taking into account important covariates. Although the use of Biofilter might not be as promising an option in cases where very little biological knowledge exists on the phenotype being analyzed, in the case of pharmacogenomics, where extensive drug metabolism networks have been elucidated, utilizing this knowledge to direct the analysis is a superior alternative, particularly when epistasis is concerned. As the search for the genetic architecture underlying complex traits such as drug pharmacokinetics continues, utilities such as the Biofilter can play an important role. Drug response is a nuanced trait and, as such, is likely to have genetic components which are monogenic as well as those that are multi-locus. Now that whole-genome sequencing technology is almost ready for wide-spread implementation, rare genetic variation is likely also to become an important component to consider for pharmacogenomic traits. Due to the nature of rare variants, the same pathway knowledge which is exploited by Biofilter to search for epistasis should be useful to group these rare variants to look for patterns predicting drug response. In summary, Biofilter is a tool which is likely to prove invaluable for the analysis of large-scale genetic association data for complex disease, especially in pharmacogenomic data where the biological knowledge is extensive.

5. Acknowledgements

This work was supported in part by the AIDS Clinical Trials Group funded by the National Institute of Allergy and Infectious Diseases (AI68636, AI38858, AI68634, AI38855), and by virology laboratory contracts (201VC001 with Vanderbilt University, 200VC001 with the University of Alabama at Birmingham, and 200VC011 with Stanford University). Grant support included AI069439, RR000095, AI54999 (DWH), AI51966, RR024996 (RMG), AI069472, AI062435 (GKR), AI64086, AI36214 (RH), GM080178 (BJG), LM010040, HG004798, AI077505, HL065962 (MDR). Clinical Research Sites that participated in ACTG protocols ACTG 384, A5095, and 5142 and collected DNA under protocol A5128, were supported by the following grants from NIH: AI69472, AI69465, AI69532, AI69556, AI27666, AI69424, AI27660, AI69432, AI69502, AI27663, AI69477, AI69494, AI46383, AI69511, AI27658, AI69428, AI69434, AI27664, AI27661, AI69484, AI38844, AI69495, AI25903, AI69474, AI69513, AI25897, AI69501, AI25859, AI69471, AI25915, AI46370, AI69472, AI46381, AI69423, AI25868, AI69439, AI46339, AI46376, AI38858-09S1, AI69447, AI34853, AI34835, AI69415, AI69452, AI69418, AI69450, AI69467, AI32783, AI32782, AI69419, AI46386, AI69426, AI69470, AI69471, AI69503, and AI69470.

References

[1] UNAIDS. 2008 Report on the Global AIDS epidemic. 129-158. 2008.

[2] Granich R, Crowley S, Vitoria M, Smyth C, Kahn JG, Bennett R *et al.* Highly active antiretroviral treatment as prevention of HIV transmission: review of scientific evidence and update. *Curr Opin HIV AIDS* 2010; **5:**298-304.

[3] Tsibris AM, Hirsch MS. Antiretroviral therapy in the clinic. *J Virol* 2010; **84:**5458-5464.

[4] Attia S, Egger M, Muller M, Zwahlen M, Low N. Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis. *AIDS* 2009; **23:**1397-1404.

[5] Dorenbaum A, Cunningham CK, Gelber RD, Culnane M, Mofenson L, Britto P *et al.* Two-dose intrapartum/newborn nevirapine and standard antiretroviral therapy to reduce perinatal HIV transmission: a randomized trial. *JAMA* 2002; **288:**189-198.

[6] Hetherington S, McGuirk S, Powell G, Cutrell A, Naderer O, Spreen B *et al.* Hypersensitivity reactions during therapy with the nucleoside reverse transcriptase inhibitor abacavir. *Clin Ther* 2001; **23:**1603-1614.

[7] Hawkins T. Understanding and managing the adverse effects of antiretroviral therapy. *Antiviral Res* 2010; **85:**201-209.

[8] d'Ettorre G, Zaffiri L, Ceccarelli G, Mastroianni CM, Vullo V. The role of HIV-DNA testing in clinical practice. *New Microbiol* 2010; **33:**1-11.

[9] Wadelius M, Chen LY, Downes K, Ghori J, Hunt S, Eriksson N *et al.* Common VKORC1 and GGCX polymorphisms associated with warfarin dose. *Pharmacogenomics J* 2005; **5:**262-270.

[10] Oldenburg J, Bevans CG, Muller CR, Watzka M. Vitamin K epoxide reductase complex subunit 1 (VKORC1): the key protein of the vitamin K cycle. *Antioxid Redox Signal* 2006; **8:**347-353.

[11] GlaxoSmithKline. Ziagen Medication Guide. 2009.

[12] Mallal S, Phillips E, Carosi G, Molina JM, Workman C, Tomazic J *et al.* HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med* 2008; **358:**568-579.

[13] Zanella F, Lorens JB, Link W. High content screening: seeing is believing. *Trends Biotechnol* 2010; **28:**237-245.

[14] Anderson GD. Pharmacogenetics and enzyme induction/inhibition properties of antiepileptic drugs. *Neurology* 2004; **63:**S3-S8.

[15] Iyer L, King CD, Whitington PF, Green MD, Roy SK, Tephly TR *et al.* Genetic predisposition to the metabolism of irinotecan (CPT-11). Role of uridine diphosphate glucuronosyltransferase isoform 1A1 in the glucuronidation of its active metabolite (SN-38) in human liver microsomes. *J Clin Invest* 1998; **101:**847-854.

[16] Wilke RA, Reif DM, Moore JH. Combinatorial pharmacogenetics. *Nat Rev Drug Discov* 2005; **4:**911-918.

[17] Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 2002; **11:**2463-2468.

[18] Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* 2004; **27:**141-152.

[19] Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* 2005; **27:**637-646.

[20] Bateson,W, Saunders,ER, Punnett,RC, Hurst,CC. Reports to the Evolution Committee of the Royal Society, Report II. London: Harrison and Sons; 1905.

[21] Fisher RA. The correlations between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* 1918; **52:**399-433.

[22] Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput* 2009;368-379.

[23] Haas DW, Wilkinson GR, Kuritzkes DR, Richman DD, Nicotera J, Mahon LF *et al.* A multi-investigator/institutional DNA bank for AIDS-related human genetic studies: AACTG Protocol A5128. *HIV Clin Trials* 2003; **4:**287-300.

[24] Clifford DB, Evans S, Yang Y, Acosta EP, Goodkin K, Tashima K *et al.* Impact of efavirenz on neuropsychological performance and symptoms in HIV-infected individuals. *Ann Intern Med* 2005; **143:**714-721.

[25] Gulick RM, Ribaudo HJ, Shikuma CM, Lalama C, Schackman BR, Meyer WA, III *et al.* Three- vs four-drug antiretroviral regimens for the initial treatment of HIV-1 infection: a randomized controlled trial. *JAMA* 2006; **296:**769-781.

[26] Gulick RM, Ribaudo HJ, Shikuma CM, Lustgarten S, Squires KE, Meyer WA, III *et al.* Triple-nucleoside regimens versus efavirenz-containing regimens for the initial treatment of HIV-1 infection. *N Engl J Med* 2004; **350:**1850-1861.

[27] Haubrich RH, Riddler SA, DiRienzo AG, Komarow L, Powderly WG, Klingman K *et al.* Metabolic outcomes in a randomized trial of nucleoside, nonnucleoside and protease inhibitor-sparing regimens for initial HIV treatment. *AIDS* 2009; **23:**1109-1118.

[28] Riddler SA, Jiang H, Tenorio A, Huang H, Kuritzkes DR, Acosta EP *et al.* A randomized study of antiviral medication switch at lower- versus higher-switch thresholds: AIDS Clinical Trials Group Study A5115. *Antivir Ther* 2007; **12:**531-541.

[29] Robbins GK, De G, V, Shafer RW, Smeaton LM, Snyder SW, Pettinelli C *et al.* Comparison of sequential three-drug regimens as initial therapy for HIV-1 infection. *N Engl J Med* 2003; **349:**2293-2303.

[30] Shafer RW, Smeaton LM, Robbins GK, De G, V, Snyder SW, D'Aquila RT *et al.* Comparison of four-drug regimens and pairs of sequential three-drug regimens as initial therapy for HIV-1 infection. *N Engl J Med* 2003; **349:**2304-2315.

[31] Grady BJ, Torstenson ES, Dudek SM, Giles P, Ritchie MD. Finding Unique Filter Sets in PLATO: A Precursor to Efficient Interaction Analysis in GWAS Data. *Proceedings of the Pacific Symposium in Biocomputing* 2009.

VISUAL INTEGRATION OF RESULTS FROM A LARGE DNA BIOBANK (BIOVU) USING SYNTHESIS-VIEW [*]

SARAH PENDERGRASS

Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University
507D Light Hall, Nashville, TN 37205, USA
Email: sarah.a.pendergrass@vanderbilt.edu

SCOTT M. DUDEK

Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University
509 Light Hall, Nashville, TN 37205, USA
Email: dudek@chgr.mc.vanderbilt.edu

DAN M. RODEN

Department of Medicine, Department of Pharmacology, Office of Personalized Medicine, Vanderbilt University
536 Robertson Research Building, Nashville, TN 37205, USA
Email: dan.roden@vanderbilt.edu

DANA C. CRAWFORD

Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University
505 Light Hall, Nashville, TN 37205, USA
Email: dana.crawford@chgr.mc.vanderbilt.edu

MARYLYN D. RITCHIE

Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University
509 Light Hall, Nashville, TN 37205, USA
Email: ritchie@chgr.mc.vanderbilt.edu

In this paper, we describe using Synthesis-View, a new method of presenting complex genetic data, to revisit results of a study from the BioVU Vanderbilt DNA databank. BioVU is a biorepository of DNA samples coupled with de-identified electronic medical records (EMR). In the Ritchie et al. study [1] ~10,000 BioVU samples were genotyped for 21 SNPs that were previously associated with 5 diseases: atrial fibrillation, Crohn Disease, multiple sclerosis, rheumatoid arthritis, and type 2 diabetes. In the proof-of-concept study, the 21 tests of association replicated previous findings where sample size provided adequate power. The majority of the BioVU results were originally presented in tabular form. Herein we have revisited the results of this study using Synthesis-View. The Synthesis-View software tool visually synthesizes the results of complex, multi-layered studies that aim to characterize associations between small numbers of single-nucleotide polymorphisms (SNPs) and diseases and/or phenotypes, such as the results of replication and meta-analysis studies. Using Synthesis-View with the data of the Ritchie et al. study and presenting these data in this integrated visual format demonstrates new ways to investigate and interpret these kinds of data. Synthesis-View is freely available for non-commercial research institutions, for full details see https://chgr.mc.vanderbilt.edu/synthesisview.

[*] This work was supported in part by LM010040 (SAP, SMD, MDR), HG004798 (SAP, DCC, MDR), HL065962 (MDR, DCC)

1. Introduction

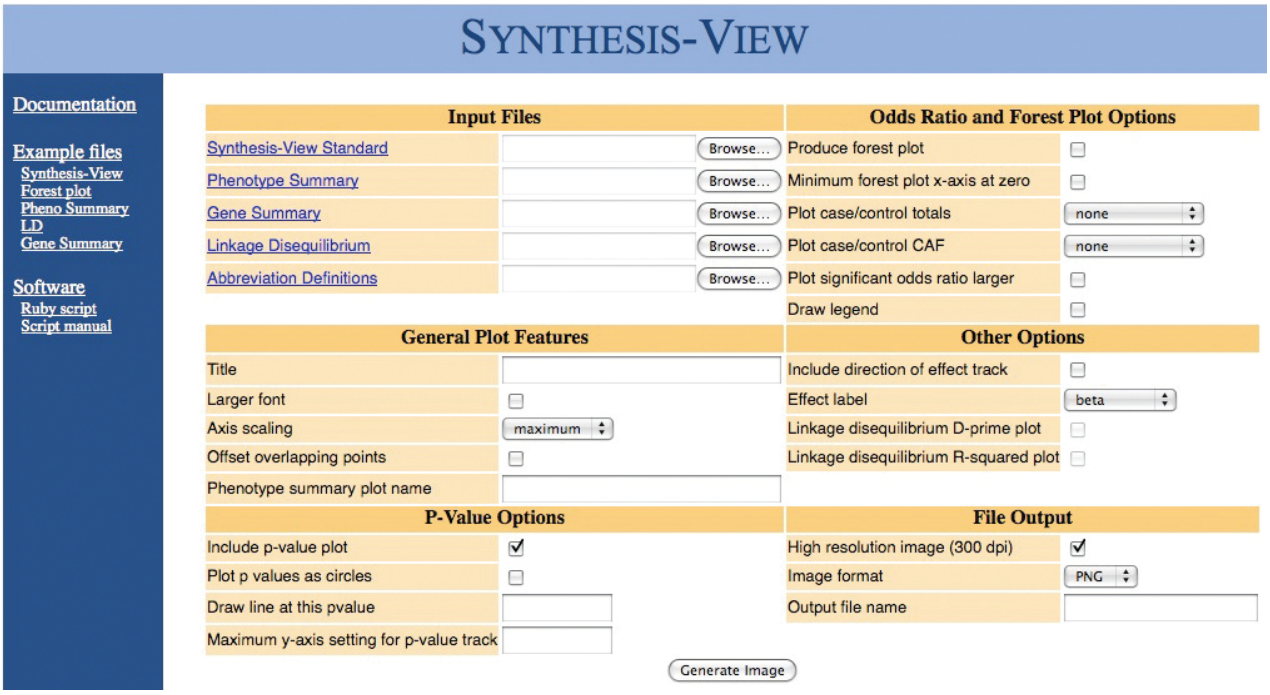
The use of results from genome-wide association studies (GWAS) in the emerging field of personal genomics requires the further investigation and characterization of potentially functional single nucleotide polymorphisms (SNPs) originally identified in GWAS. The additional studies required usually characterize less than 100 SNPs, often include multiple and correlated phenotypic measurements, and can include data from multiple-sites, multiple-studies, as well as multiple race/ethnicities. The Vanderbilt University biobank (BioVU)[2] aims to both characterize previously detected SNPs, as well as discover new associations between genetic variation and diseases and phenotypes. BioVU has an "opt-out" system, whereby DNA samples are collected from blood remaining after routine clinical testing at Vanderbilt Medical Center. De-identified electronic-medical record (EMR) data, called the "synthetic-derivative" (SD) is coupled to DNA of the biorepository. Cases and controls for phenotype-genotype association are identified using the synthetic-derivative through the use of electronic phenotyping algorithms developed in by EMR content experts along with biomedical informaticists.

In the Ritchie et al. study [1], the first approximately 10,000 DNA samples collected in BioVU were genotyped for a series of SNPs that each had a previously known and robust association with of one of five common diseases. The goal of this proof-of-concept study was to demonstrate that EMR data can successfully be used to accurately define phenotypes that enable the investigation of genotype-phenotype correlations. In this study the electronic phenotyping algorithms were deployed in the SD to determine cases and controls for atrial fibrillation, Crohn disease, multiple sclerosis, rheumatoid arthritis, and type 2 diabetes in a sample of largely European American descent. A total of 9483 DNA samples were successfully genotyped, and 21 tests of association were performed. Significant associations ($p < 0.05$) were found for 8/14 tests where SNPs had a previously reported odds ratio (OR_{PR}) > 1.25, and 0/7 where SNPs had a lower OR_{PR}. In the initial presentation of the results of this study, the majority of the results were provided in a tabular form. While tabular data provides a record of the exact results of a study, it can be challenging to identify and convey the trends and patterns within a set of results using a tabular data alone.

Visualizing data results such as those of the BioVU study as well as other candidate-gene replication studies that move beyond initial GWAS findings, provides a way to interpret the complex and multi-layered results of these studies in a more integrated way, and allows for rapid comparisons of multiple forms of information not easily achievable through reviewing large tables of numbers. To visualize the results of these forms of studies, we developed the software tool "Synthesis-View" to visually synthesize the results of candidate gene and GWAS replication studies in stacked data-tracks, providing a single image where p-values (or other measures of significance), odds-ratios, allele frequencies, sample sizes, effect size, and direction of effect are all incorporated. While Manhattan plots already exist for the effective visualization of GWAS data, the results of candidate gene studies, studies investigating genetic variation in specific regions in detail, or even isolated GWAS results, are not often presented in visual form. Our tool provides a unique and direct way to generate accessible visual information from these kinds of data.

2. Methods

The Synthesis-View software tool used herein was developed in Ruby and utilizes the RMagick graphics library. Synthesis-View is available for use through a web interface, and can alternately be used at the command line. Figure 1 shows a screen-capture of the web interface, which allows for the flexible choice of various options for Synthesis-View plots. The required and optional tab-delimited text input file format to produce a Synthesis-View plot are briefly described here, and are also described in greater detail at the Synthesis-View website along with example input files. One file is necessary to produce a standard Synthesis-View plot, a file containing a column for SNP identification (such as RS number), a column for which chromosomes the SNPs map to, and a column for SNP genomic location information. The rest of the standard input file can optionally contain information on p-values, odds-ratios, allele frequencies, and sample size, with tracks plotted if data are present. Other files can be provided for Synthesis-View to plot additional tracks of data. If a phenotype summary file is supplied, summary information about continuous phenotypes will be plotted. If a gene summary file is included, information on gene name and location in relation to SNPs plotted will be in a track at the top of the plot. If a linkage disequilibrium file is provided that contains D′ or r^2 correlation data, the data will be plotted in Haploview style format [3]. Finally, if abbreviation definitions are provided, an additional legend



Input Files		
Synthesis-View Standard		Browse...
Phenotype Summary		Browse...
Gene Summary		Browse...
Linkage Disequilibrium		Browse...
Abbreviation Definitions		Browse...

Odds Ratio and Forest Plot Options	
Produce forest plot	□
Minimum forest plot x-axis at zero	□
Plot case/control totals	none
Plot case/control CAF	none
Plot significant odds ratio larger	□
Draw legend	□

General Plot Features	
Title	
Larger font	□
Axis scaling	maximum
Offset overlapping points	□
Phenotype summary plot name	

Other Options	
Include direction of effect track	□
Effect label	beta
Linkage disequilibrium D-prime plot	□
Linkage disequilibrium R-squared plot	□

P-Value Options	
Include p-value plot	☑
Plot p values as circles	□
Draw line at this pvalue	
Maximum y-axis setting for p-value track	

File Output	
High resolution image (300 dpi)	☑
Image format	PNG
Output file name	

Figure 1 – Synthesis-View web interface screen capture.

describing plot abbreviations will appear below OR/forest plots when "Draw Legend" is selected. Table 1 describes the various possible settings available in the web interface.

3. Visualization of Results

The focus of the proof-of-concept BioVU study was to both show and characterize the utility of using electronic phenotype algorithms deployed in an EMR linked to a DNA biobank. As described in Ritchie et al. [1], blood samples that showed poor-quality or that yielded insufficient DNA, blood samples from individuals < 18 years of age, a lack of consent-to-treatment form, any indication of opt-out, or discovery of a duplicate sample, resulted in exclusion from the study. In addition, 2% of samples in BioVU are randomly dropped out, further randomizing individuals not included in the biobank and consequent studies. After filtering for exclusions, definite cases of European Ancestry (EA) and probable EA were defined using the administrative information recorded in the EMR. Almost a tenth of the records (9.2%) did not include ancestry information, or recorded the ancestry as "unknown". The data were thus analyzed with cases and controls that indicated EA specifically as the race/ethnicity, and also separately analyzed with cases and controls defined as both EA and individuals characterized as unknown.

To define disease state for case/control status, for one set of association tests, identification of case/control status was solely determined using an electronic phenotyping algorithm (see Ritchie et al. appendix for algorithm details). Content experts were used to develop the algorithm that used disease-specific billing codes and patient encounter information, including records such as medication information, electrocardiogram data, and past medical history from the SD. "Definite" cases were defined by the algorithm as disease present, excluding those with indications of overlapping disease or symptoms, or lack of a clear diagnosis. Controls were defined as those with clear absence of the specific disease used in the case/control association. In the case of multiple sclerosis, algorithm classified cases were also manually reviewed because of the small sample size. In addition to the algorithm defined Definite cases, for rheumatoid arthritis and multiple sclerosis, a set of association tests were separately performed with both Definite cases as well as cases showing indications of overlapping autoimmune diseases and/or symptoms. These cases were described as "Probable".

After defining cases/controls, association tests for the 21 genotyped SNPs were performed. For SNPs associated with atrial fibrillation, Crohn's disease, or Type 2 diabetes, tests of association were performed for both EA with cases Definite cases and EA + Unknown with Definite cases. For SNPs known to be associated with rheumatoid arthritis and multiple sclerosis, tests of association were performed for EA with Definite cases, EA with Definite and Probable cases, EA + Unknown with Definite cases, EA + Unknown with Definite and Probable cases.

3.1 *Synthesis-View Forest Plot*

The results of the association tests of the BioVU study were presented in Table 1 of the Ritchie et al. manuscript [1]. The results for EA alone with Definite cases were presented in a forest plot along with OR_{PR} from previous studies in Figure 1 of the Ritchie et al. manuscript [1,4-10]. In the current paper, Figure 2 is a modified forest plot using Synthesis-View to visualize the results of the BioVU study. From left to right in Figure 2 are tracks with various pieces of data:

1. The first track is a *physical genome track*, displaying the chromosome and relative location of each SNP used in the 21 association tests. Having the SNP data presented in this way visually shows the location of SNPs in reference to other SNPs within the same study. Lines lead from the relative location of each SNP to the SNP identifier.
2. The next track is the *significance track*, showing the p-values of both the original OR_{PR} as well as the results of the Ritchie et al. paper. A single color consistently represents results for the original OR_{PR} (in blue), as well as for the new associations: EA_D (European American, Definite disease classification, in red); EA+U (European American and Unknown, Definite disease, in orange); EA_P (European American, Definite as well as

Table 1. Synthesis-View plotting options

Synthesis-View Option	Description
Title	Title for Synthesis-View plot
Larger font	Produce a plot with larger sized text than the default
Axis scaling	If set to "maximum", axes limits will start and end utilizing the range of the data with tick-marks at regular intervals in-between. If set to "cleaner" the axes will still encompass the range of the data, however the range will begin and end with a multiple of five or ten, and the plot tick-marks will also be a multiple of five or ten.
Offset overlapping points	When points overlap, this setting will include "jitter", whereby overlapping points are offset horizontally to make them more distinguishable.
Phenotype summary plot name	If phenotypic summary data will be incorporated into the Synthesis-View plot, the title for the phenotype summary plot should be specified here.
Include p-value plot	Include plot of p-values
Plot p-values as circles	To plot p-values as circles, instead of triangles that include direction of effect, even if direction of effect information is supplied in the Synthesis-View standard input file.
Draw line at this p-value	Specification of a horizontal red line at a specific p-value of interest.
Maximum y-axis setting for p-value track	Specify the maximum y-axis value for the p-value track in order to limit the range of the y-axis. Any p-value result more significant than this y-axis cutoff value will be plotted at the cutoff value in larger size.
Produce forest plot	To produce a forest plot in Synthesis-View from odds-ratio results
Minimum forest plot x-axis at zero	To set the minimum value of the forest plot x-axis to zero
Plot case/control totals	The total numbers of cases/controls can be plotted either in two separate tracks ("split plot"), or in one track where the total numbers of cases/controls are indicated using open/closed circles ("combined plot").
Plot case/control CAF	The respective coded allele frequency (CAF) for cases/controls can be plotted either as two separate tracks ("split plot"), or in one track where cases/controls are indicated using open/closed circles ("combined plot").
Plot significant odds ratio larger	Plot significant odds-ratio results in larger size
Draw Legend	When an "Abbreviation Defintions" file is provided, and Draw Legend is selected, an additional legend describing plot abbreviations will appear below OR/forest plots
Include direction of effect track	Even if direction of effect information is supplied, this setting allows for inclusion/exclusion of a direction of effect track.
Effect label	Choice of effect size label
Linkage disequilibrium D-prime plot	If linkage disequilibrium information is included as an input file, select this to include a d-prime correlation track.
Linkage disequilibrium R-squared plot	If linkage disequilibrium information is included as an input file, select this to include an R-squared correlation track.
High resolution image (300 dpi)	Select to produce a 300 dpi image, otherwise the image is 72 dpi
Image format	Choices of image format include PNG, JPEG, and TIFF
Output file name	Choice of file name for output Synthesis-View plot

Probable disease, in purple); and EA+U_P (European American and Unknown, Definite as well as Probable disease, in green). Applying a red line at a p-value cutoff of choice is one of the options of Synthesis-View, in this case the vertical red line was applied at a p-value of 0.05, allowing for a more quick detection of values above and below the chosen p-value. For two SNPs, rs6457620 and rs3135388, in studies prior to Ritchie et al. the results were extremely significant at 4E-186[10] and 9E-81[6] respectively. When these two SNPs were originally plotted on the same track as the rest of the p-values, there was compression of other p-values along the bottom of the plot due to the wide spread of the data points. Synthesis-View allows for the choice of a p-value cutoff, whereby any points more significant than that cutoff are plotted at that cutoff value with a larger sized point. Thus, on this plot, after choosing a p-value cutoff of 1E-50, the two points for SNPs, rs6457620 and rs3135388 are plotted at p-value 1E-50 but are larger in size. Also of note, the various BioVU p-value results for each SNP were very similar, thus when initially plotted, the points had considerable overlap. Synthesis-View allows the application of "jitter", where points that overlap are spread out vertically along the "abacus" line leading down from the SNP identification information. Thus the jitter option was applied, providing more visual discrimination between multiple overlapping points.

3. The next four tracks are *odds-ratio/forest-plot tracks.* Each track shows the individual odds ratio (OR) and confidence intervals for each of the separate sets of associations, such as those for EA_D or EA+U. Each OR result is plotted as a square, with a line indicating the upper and lower 95% confidence interval. In this case a specific option in Synthesis-View was used, whereby if the result is significant (the upper or lower boundary of the confidence intervals do not cross 1.0), the square is plotted in larger size. This allows for quick visual identification of significant results in forest plots that may show many results. In the case of the results for the previous studies, the confidence intervals were small enough they were overplotted by the OR square. As the eye moves from left to right, there are visible trends. Results that were not significant in the BioVU study were in the same direction as OR_{PR}. Also, it is easy to determine how similar the results were in the BioVU study, even with inclusion or exclusion of data from Unknown individuals and Probable case data. With Synthesis-View both an overview of the data as well as individual results are available, and a table can be used to look up exact numerical results of interest.
4. The second to last track is the *coded allele frequency track*. Synthesis-View provides the option of either the coded allele frequencies (CAF) of both cases and controls plotted on the same track, with closed circles indicating cases and open circles indicating controls, or the allele frequencies of cases and controls can be plotted in two separate tracks. In either case, colors match those of the groups of the previous tracks, allowing the user to look at the allele frequencies between groups by eye for trends. This can aid in interpreting the potential lack of replication of results.

5. The last track is the *sample size track*. Like the CAF track, case/control sample size can either be plotted with closed circles indicating cases, and open circles indicating controls. The colors match those of the groups of the previous tracks, allowing the user again by eye to look at sample size across groups.

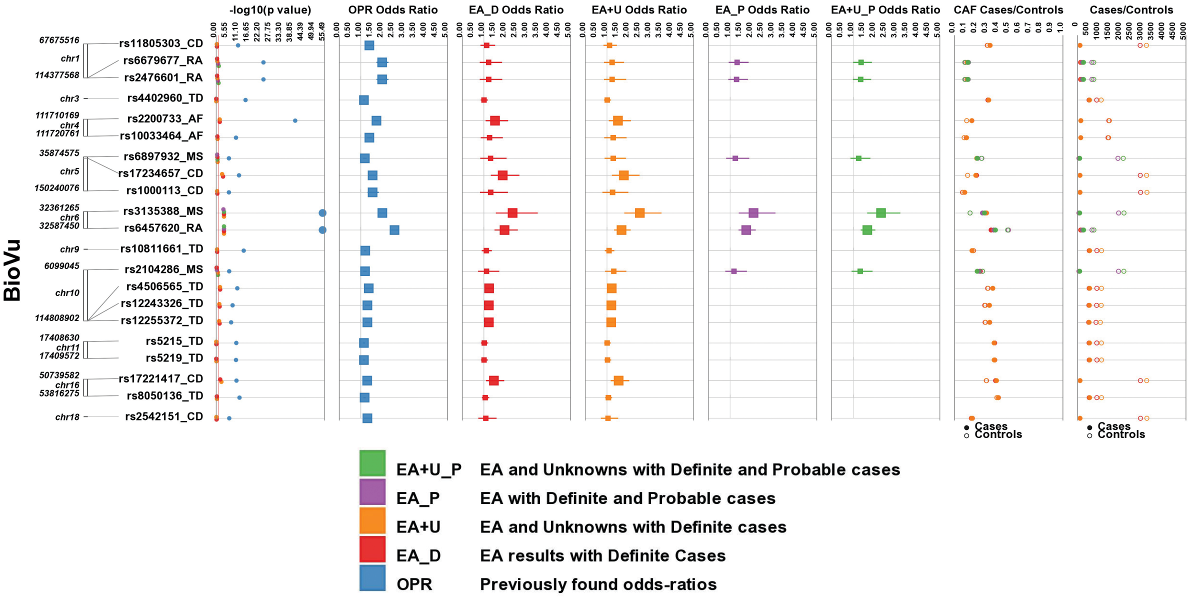


Figure 2 - The results of using the forest-plot option in Synthesis-View and the data of the Ritchie et al. BioVU paper. Moving from left to right, the first track shows SNP location, the next track shows –log$_{10}$(p-value). Each SNP identifier also has the following abbreviations for associated disease: atrial fibrillation (AF), Crohn disease (CD), multiple sclerosis (MS), rheumatoid arthritis (RA), and type 2 diabetes (TD). The next five tracks are odds-ratio/forest-plots. The abbreviations for these tracks are described in greater detail in the figure legend. The coded allele frequency (CAF) track with allele frequencies for both cases/controls is the second-to-last track. Sample sizes for the cases/controls are plotted in the last track.

3.2 *Synthesis-View Standard Plot*

An alternative way to look at the results of the BioVU study is through stacked tracks where the eye moves from top to bottom (Figure 3). If the "forest-plot" option is not chosen in Synthesis-View, the default data plot is in this format. Again the first track is the *physical genome track*, with chromosome number and the relative location of each SNP with lines leading from the chromosome location track to identification of each of the respective SNPs. The next track is the *significance track*, showing p-value results across groups with an optional horizontal red line at a p-value of 0.05 applied. In this case, again to reduce compression of the p-value results when plotted, a p-value cutoff was chosen (1E-30), with larger points plotted directly at the p-value cutoff. SNPs rs6457620 and rs3135388 and rs2200733 had p-values of 4E-186[10], 9E-81[6], and 3.3E-41[4] respectively. The track below the significance track is an *odds-ratio track*. Unlike the forest-plots of Figure 2, here the ORs are plotted as closed circles. If the OR results are significant, the OR closed circle is plotted in a larger size. So while the confidence intervals are not plotted, it is still easy to discriminate OR results that are significant. For studies where OR data are omitted, the OR track will not appear. Below the OR track, there is a CAF track. Again, Synthesis-View provides the option of either viewing the allele frequencies of both cases and controls plotted on the same track, with closed circles indicating cases, and open circles indicating controls. The last track is a sample size track plotted in a similar fashion as the CAF track.

There are available Synthesis-View options that were not used in this presentation of the BioVU results. When summary data regarding a continuous phenotype of interest exists, there is an option to add on a summary data plot, which consists of the mean and standard deviation of the continuous phenotype for each group. Future versions of Synthesis-View will incorporate ways to characterize categorical/case-control phenotype summary data. Also, when linkage disequilibrium (LD) data is provided, a D′ or r^2 correlation plot in Haploview style format [3] is plotted.

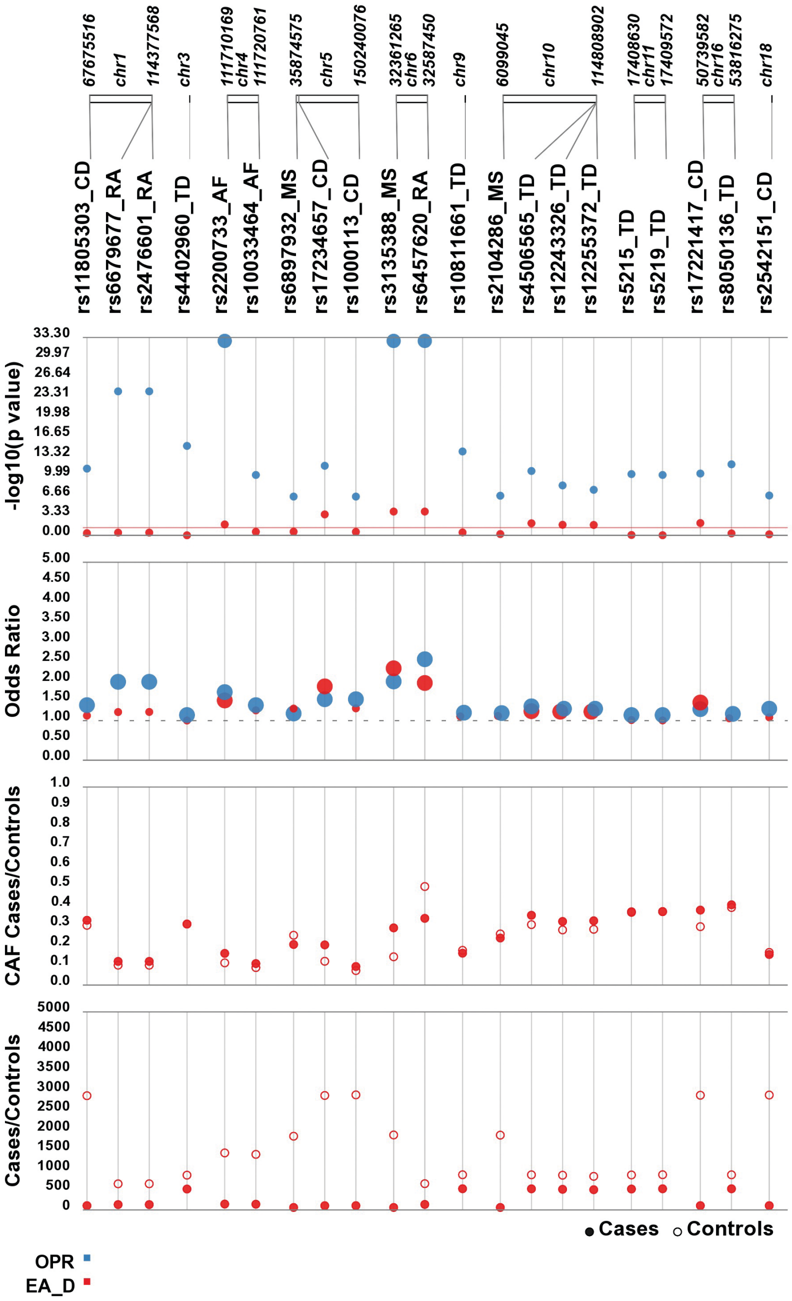


Figure 3 - Default format of the Synthesis-View plot with horizontal data tracks. In red are the results of OR_{PR} (OPR), in blue are results of the BioVU Ritchie et al. study. From top to bottom, data tracks include the physical genome track, odds-ratio track (significant odds-ratio results are plotted larger in size), coded-allele-frequency (CAF) track for cases and controls, and a sample-size track for cases and controls.

4. Conclusions

Synthesis-View was extended from the previous software "LD-Plus". The LD-Plus feature carried through to Synthesis-View is the use of multiple tracks for showing data results, as LD-Plus also uses a flexible data display format of multiple data "tracks" that can be viewed [11]. However, Synthesis-View allows for visualization of data that is not possible with LD-Plus. In Synthesis-View, through the use of stacked data-tracks, SNP genomic location, presence of the SNP in a specific study or analysis, as well as related data such as genetic effect size and summary phenotype data, are plotted according to user preference. With Synthesis View, trends from many different kinds of information can be visualized in a more integrated way than by using tabular data alone. These multi-faceted views are important to understanding in greater depth the relationships between SNPs, strata, sample size, and phenotypic differences expected with the increasing complexity of emerging datasets.

It is important to note here that we present one set of scenarios where Synthesis-View can be used; however, the software is very flexible and that there are no restrictions to how the data are grouped. The Ritchie et al. paper was able to show proof-of-concept, such that the use of a biobank coupled with EMR data can effectively replicate previously well characterized results. The original results of this paper were largely presented in tabular format, and here we show the utility of Synthesis-View in visualizing these kinds of results. Through using Synthesis-View the larger picture of the data as a whole can be seen, with trends and patterns visually evident, while also allowing a user to determine details about individual results. Tables can then be used as a reference for determining specific numerical results in greater detail after areas of interest are located in the plotted data.

5. Acknowledgments

We would like to acknowledge the following individuals for their suggestions and ideas in designing Synthesis-View: Matthew Thomas Oetjens, Fredrick Schumacher, Janina Jeff, Logan Dumitrescu, and Chris Haiman. This work was supported in part by LM010040 (SAP, MDR), HG004798 (SAP, DCC, MDR), and HL065962 (MDR, DCC).

6. References

1. Ritchie, M.D., *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 86, 560-572 (2010).
2. Roden, D.M., *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 84, 362-369 (2008).
3. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263-265 (2005).
4. Gudbjartsson, D.F., *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 448, 353-357 (2007).
5. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-678 (2007).
6. Hafler, D.A., *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med* 357, 851-862 (2007).
7. Groves, C.J., *et al.* Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* 55, 2640-2644 (2006).

8. Zeggini, E., *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316, 1336-1341 (2007).
9. Saxena, R., *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316, 1331-1336 (2007).
10. Raychaudhuri, S., *et al.* Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet* 40, 1216-1223 (2008).
11. Bush, W.S., Dudek, S.M. & Ritchie, M.D. Visualizing SNP statistics in the context of linkage disequilibrium using LD-Plus. *Bioinformatics* 26, 578-579 (2010).

MULTIVARIATE ANALYSIS OF REGULATORY SNPS: EMPOWERING PERSONAL GENOMICS BY CONSIDERING CIS-EPISTASIS AND HETEROGENEITY

STEPHEN D. TURNER[†]

Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University
Nashville, TN 37232, United States
Email: stephen.turner@vanderbilt.edu

WILLIAM S. BUSH

Center for Human Genetics Research, Department of Biomedical Informatics, Vanderbilt University
Nashville, TN 37232, United States
Email: william.s.bush@vanderbilt.edu

Understanding how genetic variants impact the regulation and expression of genes is important for forging mechanistic links between variants and phenotypes in personal genomics studies. In this work, we investigate statistical interactions among variants that alter gene expression and identify 79 genes showing highly significant interaction effects consistent with genetic heterogeneity. Of the 79 genes, 28 have been linked to phenotypes through previous genomic studies. We characterize the structural and statistical nature of these 79 *cis*-epistasis models, and show that interacting regulatory SNPs often lie far apart from each other and can be quite distant from the gene they regulate. By using *cis*-epistasis models that account for more variance in gene expression, investigators may improve the power and replicability of their genomics studies, and more accurately estimate an individual's gene expression level, improving phenotype prediction.

1. Introduction

Epistasis, or gene-gene interaction, is thought to be an important component of complex, multifactorial diseases due to the monumental complexity of biological systems [1]. Over the past 10 years, a wealth of data from model organisms has supported a role for epistasis [2, 3]. Furthermore, epistasis is one way to account for the problem of “missing heritability”, where the analysis of single SNPs (single nucleotide polymorphisms) has explained very little of the heritability estimated from twin and adoption studies for complex traits [4, 5]. Accounting for interactions among SNPs may explain a larger portion of this heritability [6], expanding our understanding of the genomics of human disease and personalized medicine.

One often cited potentially causal mechanism of gene-gene interaction is due to variation in multiple genes in similar pathways, protein families, or genes with similar or redundant biological function [7, 8]. This generally implies that interaction occurs between genes scattered throughout the genome due to a *trans*-epistasis effect. Several approaches have been applied to investigate these effects in genome-wide association studies [9-12].

The occurrence of epistatic interactions, however, is not restricted to variation between distant genes. Epistatic interactions could also occur between genetic variants in close proximity which may impact transcriptional regulation. Recent work investigating the transcriptome of HapMap-

[†] Work partially supported by National Research Service Award F31-NS066638.

based cell lines has led to the identification of expression quantitative trait loci (eQTLs) - genetic variants that influence the expression of a gene [13, 14]. Veyrieras et al. published an analysis of gene expression for 11,446 genes from HapMap-based lymphoblastoid cell lines leveraging genotypes for roughly 3 million single nucleotide polymorphisms (SNPs) to identify eQTL SNPs in a 500 kilobase (kb) window both upstream of the transcription start site and downstream of the transcription end site [15]. This work discovered 744 genes containing at least one significant eQTL SNP ($p<7\times10^{-6}$). The single-SNP analysis, however, does not assess the variance in gene expression that can be explained by the interaction of multiple SNPs in regulatory regions of the gene. It has been shown that the underlying mechanisms of gene expression are incredibly complex, involving the binding of multiple factors to DNA to facilitate transcription and mRNA stability [16]. Furthermore, polymorphisms within the binding sites of multiple factors may alter binding affinities to various degrees, exerting a non-linear influence on gene expression due to synergistic effects [17, 18]. This principle has been demonstrated with multiple sclerosis where severity is impacted by functional effects of two alleles in close proximity in the MHC region [19]. Despite the known complexity of gene regulation, multi-SNP interaction analysis has been previously examined only for genes having highly heritable expression but lacking single SNP associations [20]. As a secondary analysis of eQTLs using lymphoblastoid lines isolated from children with asthma, the authors successfully explain some of the missing heritability from single SNP analysis using interactions. From this limited assessment, the authors conclude that genetic interactions may have an important role in the regulation of gene expression. From these points, we hypothesize that combinations of SNPs within the 500 kb window of potential transcriptional influence will alter gene expression in humans in a non-linear fashion, here dubbed *cis*-epistasis.

An analysis of gene expression phenotypes provides a unique opportunity to systematically assess the degree to which epistasis, or nonlinear interactions between genetic variants, might influence human traits. Linking the HapMap cell line expression data from [15] with publicly available genotype data on the same cell lines gives us a dense collection of genetic variants in regions with strong biological plausibility for non-linear multi-SNP interaction within 11,466 quantitative expression outcomes with established main effects. Here we leverage this data to investigate the nature and degree to which *cis*-epistasis affects gene expression in humans. Furthermore, if epistasis plays an important role in influencing gene regulation, then it logically follows that epistasis is an important part of more complex downstream human disease phenotypes, as these traits are often associated to SNPs that alter gene expression [21]. Finally, investigators could prioritize established combinations of eQTL SNPs to inform a SNP-SNP interaction analysis in complex human traits to reduce both the computational and multiple testing burdens that plague epistasis analysis in high-throughput genetic analysis. This would also motivate reanalysis of existing datasets for multi-SNP interactions that influence complex disease, many of which are publicly available at the database of genotypes and phenotypes (dbGaP) [22]. Put simply, if a study design which considers *cis*-epistasis can explain more heritability in gene expression, then personal genomics studies that account for *cis*-epistasis should be more fruitful.

2. Methods

2.1. *Genotype and Gene Expression Data*

As a starting point for these analyses, we retrieved the full eQTL results database and normalized gene expression data from the Veryrieras et al. analysis (available online: http://eqtnminer.sourceforge.net/), containing 11,966,533 results (significant and non-significant) from 2,437,821 distinct SNPs and 11,466 distinct microarray probes [15]. These results establish a mapping between eQTL SNPs and the genes they regulate using a 500kb window both upstream and downstream of the regulated gene. We limited all analyses to these SNPs and microarray probes. Genotype data for these SNPs was retrieved from release #23 of the International HapMap project for 210 unrelated individuals, including 60 Yoruba (YRI) and 60 CEPH (CEU) parents, and 90 unrelated Chinese (CHB) and Japanese (JPT) samples [23]. Processed gene expression data was retrieved from (http://eqtnminer.sourceforge.net/) that had been normalized first by quantile normalization within replicates and then median normalized across all HapMap individuals. We then applied the normalization procedure from [15], which is a Gaussian quantile normalization for each gene within each population separately to avoid results confounded by population stratification (the distribution of expression values within each population is now the same).

2.2. *Statistical Analysis*

From the Veryrieras et al. analysis results database, we extracted all SNPs with eQTL p-values <0.05 and their associated microarray probe - that is, all nominally significant SNPs falling within 500 kb upstream of the transcription start site and 500 kb downstream of the transcription end site. Based on this data we generated all possible pair-wise combinations of associated SNPs for each microarray probe, constructing 12,107,627 two-SNP models in total. For each model, we performed a multiple linear regression analysis fitting a model with additive main effect terms (AA = 0, Aa = 1, aa = 2) for the two individual SNPs and a multiplicative interaction term. We tested for significance of interaction via a student's T-test of the interaction term coefficient. All regression analyses were conducted using the 'rms' package for the R statistical computing environment [24]. Statistical significance was determined by controlling the false discovery rate (FDR) at 0.20, using the 'qvalue' package available for R [25]. Linkage disequilibrium was computed using PLINK software, analyzing the combined set of 210 HapMap samples without phasing using the '`--r2`' option [26].

2.3. *Annotation of Results Using GWAS Catalog*

The National Human Genome Research Institute (NHGRI) actively maintains a catalog of all significant ($p<10^{-5}$) findings from published Genome-Wide Association Studies (GWAS) [27](accessed March, 2010). The National Heart, Lung, and Blood Institute (NHLBI) also recently released comprehensive open access database of 118 GWAS studies containing 56,411 significant SNP-phenotype associations [28]. Illumina expression probe IDs were matched to transcripts within the Ensembl database (Release 49). Transcripts were matched to Ensembl Genes which have associated gene symbols within the Ensembl database. These symbols were matched to the "gene" fields in the GWAS catalogs to assess the number of matches. We also referenced the SNPs from our most significant results against these catalogs to determine if any single SNPs in the regions around our findings were known to influence any complex human phenotypes.

3. Results

3.1. *Gene Expression in Humans is Influenced by Cis-Epistasis*

After exhaustively fitting two-SNP models between known eQTL SNPs surrounding each microarray probe (12,107,627 two-SNP models in total), we examined the distribution of the p-values from the interaction term. The full results catalog from this analysis is available online at http://chgr.mc.vanderbilt.edu/bushlab/. Figure 1 is a quantile-quantile plot showing that the distribution of interaction term p-values deviates highly from the expected uniform distribution under the null hypothesis of no epistasis (diagonal line). This indicates that multi-SNP interaction may be common among eQTL SNPs that influence gene expression in humans.

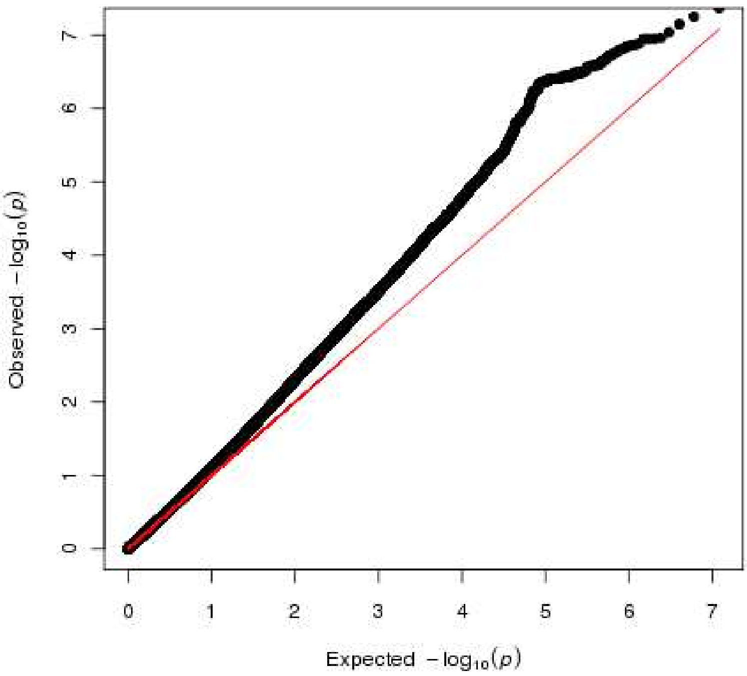


Fig. 1. Quantile-quantile plot showing the distribution of observed –$\log_{10}$(p-values) against the expected –$\log_{10}$(p-values) for the interaction term among 12,107,627 *cis*-epistasis models. Deviation from the expected uniform distribution of p-values under the null hypothesis (indicated by the red line) indicates an abundance of significant *cis*-epistatic interactions.

Because a large number of statistical tests were performed, we corrected for multiple testing using the false discovery rate (FDR) method described in the methods section. Of the ~12 million two-SNP interaction models tested with multiple linear regression, 706 were still significant after correcting for multiple testing. It is of note that our multiple testing correction is extremely conservative because our tests of interaction are not independent of each other. The deviation from the null hypothesis of no interaction shown in figure 1 suggests that there may be many more than 706 SNP-SNP interactions truly influencing gene expression that we are insufficiently powered to detect when applying our FDR correction. These 706 significant SNP-SNP interaction models influenced the expression of 79 unique probes, representative of 79 unique genes. 706 SNP-SNP interactions reduce to 79 genes because multiple SNP-SNP pairs are associated with the same gene. This redundancy is due to LD between SNPs across models, for example when SNP 1 of model 1 and SNP 1 of model 2 show strong correlation. However, there was relatively weak LD between the two SNPs participating within the interaction; i.e. SNP 1 and SNP 2 of model 1. The distribution of LD statistics (measured by r^2) between the SNPs in each interacting pair is shown in Figure 2. The median r^2 was 0.043, with a median distance between each pair of 108 kb. Taken together, this suggests that the majority of the most significant results are indeed epistatic effects between independent SNPs, not simple haplotype effects.

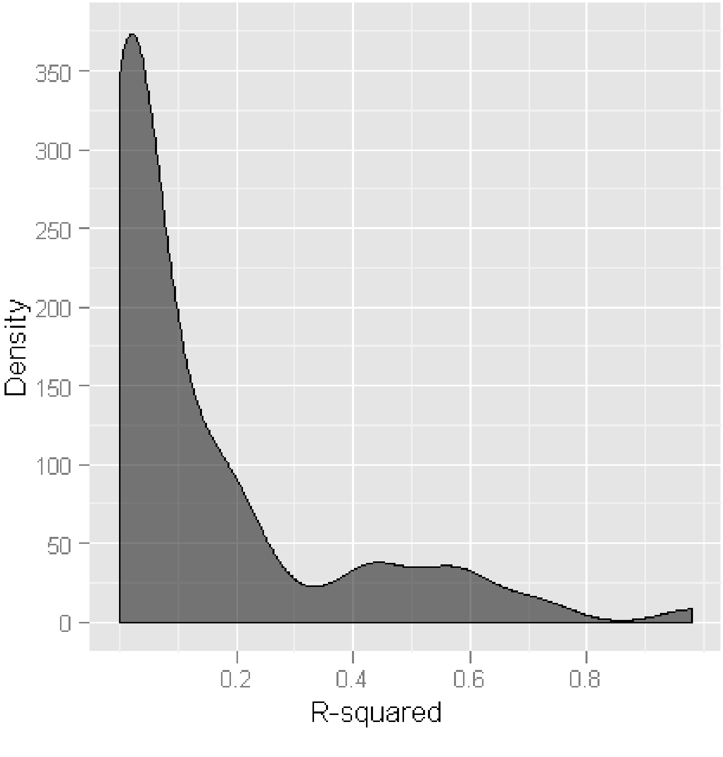


Fig. 2. Density histogram showing distribution of linkage disequilibrium (LD) values (r^2) between the most significant interacting SNP pair influencing expression of 79 genes after correcting for multiple testing. r^2 was calculated using genotype data from the combined set of 210 HapMap samples.

Table 1a. Significant two-SNP interactions where the regulated gene has been previously associated to one or more complex human disease or morphological phenotypes. The specific SNPs which interact to regulate the gene were not necessarily reported as associated to the phenotype.

Assoc. Gene	eQTL1	eQTL2	R^2_{full}	R^2_{redu}	R^2_{diff}	β_1	β_2	β_{int}	LD (r^2)	eQTL1 Pvalue	eQTL2 Pvalue	INT Pvalue	Model Pvalue	GWAS Associated Phenotype
ABCA13	rs17132158	rs6945363	0.104	0.017	0.087	1.18	1.19	-0.77	0.00	0.0308	0.0012	1.3E-05	4.8E-05	Height,Triglycerides, Systolic Blood Pressure, Fasting glucose
AEBP2	rs7135885	rs11044945	0.100	0.013	0.087	-1.36	-0.80	0.58	0.74	0.0254	0.0222	1.4E-05	7.4E-05	Insulinogenic index
BLK	rs11986748	rs2572430	0.095	0.012	0.082	1.23	0.83	-0.77	0.02	0.0346	0.0461	2.4E-05	0.00013	Lupus*, Rheumatoid arthritis*
C10orf97	rs2883029	rs4748176	0.086	0.001	0.085	0.51	0.39	-0.47	0.02	0.0205	0.0188	1.9E-05	0.00032	Crohn's disease
CDKL1	rs1955926	rs7151406	0.115	0.037	0.078	-0.09	1.52	-0.58	0.56	1E-06	7E-09	3.3E-05	1.4E-05	Cognitive Performance
CPNE8	rs2387836	rs12818797	0.087	0.006	0.081	-1.19	-1.65	0.67	0.39	0.0114	0.0235	3.2E-05	0.00032	Waist circumference
DTNB	rs1369704	rs7607198	0.101	0.015	0.085	-0.02	2.94	-1.21	0.13	7E-15	0.0249	1.7E-05	7.2E-05	Type II Diabetes (T2D)
EEFSEC	rs2811484	rs2713590	0.099	0.013	0.086	-0.58	-0.28	1.24	0.00	0.0245	0.0307	1.4E-05	8E-05	Alzheimer's disease (AD)
FRMD3	rs10868025	rs11792634	0.124	0.042	0.082	-2.03	-1.30	0.70	0.06	0.0165	0.0307	2E-05	5.5E-06	HDL cholesterol
GNG2	rs1272117	rs3742536	0.087	0.005	0.082	-1.09	-1.23	0.90	0.01	0.0064	0.0171	2.6E-05	0.00031	AD, T2D, Crohn's disease
GRIP2	rs2607765	rs2607737	0.093	0.009	0.084	-0.19	-0.84	0.78	0.18	0.0262	0.0355	2E-05	0.00016	Cognitive performance
KIF7	rs17807856	rs3803530	0.081	0.003	0.078	-1.01	-0.88	0.58	0.01	0.0103	0.0012	4.4E-05	0.00057	LDL Cholesterol
MCOLN2	rs657309	rs6690583	0.095	0.003	0.092	0.42	1.02	-0.54	0.04	3E-13	0.0223	8E-06	0.00012	Fasting glucose
NMNAT3	rs10935317	rs7648532	0.121	0.033	0.088	1.54	1.40	-0.64	0.08	0.0273	2E-24	9.7E-06	7.1E-06	BMI, Fasting glucose
NPY	rs198723	rs16189	0.085	0.006	0.080	-0.63	-0.80	0.61	0.01	0.0461	0.0006	3.4E-05	0.00036	Early onset extreme obesity
NRN1	rs3763180	rs7763755	0.114	0.034	0.079	-1.10	-1.05	0.59	0.00	0.0169	0.0012	2.7E-05	1.6E-05	Waist/height ratio squared
OBFC1	rs2986059	rs3124	0.123	0.036	0.088	1.17	0.81	-0.70	0.01	0.002	0.0372	9.8E-06	5.5E-06	Parkinson's disease,brachial artery flow velocity, Height,Endothelial traits
PCM1	rs385139	rs7816561	0.101	0.019	0.082	0.76	1.17	-0.52	0.61	0.0377	0.0462	2.3E-05	6.9E-05	Triglyceride/HDL ratio
TYK2	rs10403787	rs4804480	0.166	0.095	0.071	1.48	0.43	-0.64	0.07	0.0004	0.0006	4.1E-05	3.7E-08	Type 1 Diabetes*,Lupus
ZBTB38	rs6802753	rs7626871	0.084	0.002	0.082	0.97	1.52	-0.76	0.08	0.0053	7E-10	2.8E-05	0.00042	Height*

* indicates significant association of a gene to a complex human phenotype with $p < 5E-8$ (genome-wide significance)

Table 1b. Significant two-SNP interactions where one of the SNPs regulating a gene was previously associated to one or more human disease or morphological phenotypes. The involvement of the regulated gene in disease pathogenesis has not been investigated.

Gene	eQTL1	eQTL2	R^2_{full}	R^2_{redu}	R^2_{diff}	β_1	β_2	β_{int}	LD (r^2)	eQTL1 Pvalue	eQTL2 Pvalue	INT Pvalue	Model Pvalue	GWAS Associated Phenotype
SLIC1	rs2160683	rs9635531	0.087	0.009	0.079	1.17	1.17	-0.66	0.61	0.0461	0.0433	3.7E-05	0.00029	Crohn's disease
TMBIM1	rs7605980	rs12471773	0.087	0.011	0.076	0.37	0.64	-0.73	0.09	0.0002	0.0408	5E-05	0.00032	Type I Diabetes
PCM1	rs396462	rs2955427	0.091	0.003	0.088	0.95	0.92	-0.50	0.47	0.0326	0.0039	1.3E-05	0.00019	Crohn's disease
OTUB2	rs6575354	rs12433627	0.079	0.000	0.079	-1.00	-0.90	0.58	0.00	0.0346	0.0203	4E-05	0.00071	Type I Diabetes
DDX19A	rs929840	rs2303791	0.092	0.009	0.082	1.45	0.41	-0.80	0.32	0.042	0.0135	2.4E-05	0.00019	Alzheimer's disease
DDX19A	rs929840	rs4985534	0.087	0.008	0.079	1.43	0.39	-0.79	0.33	0.042	0.0104	3.7E-05	0.00032	Alzheimer's disease
ORMDL1	rs7568054	rs7568449	0.123	0.044	0.079	-0.24	-0.44	0.49	0.01	0.0263	3E-22	2.6E-05	5.5E-06	Amyotrophic Lateral Sclerosis
C3orf31	rs7615782	rs440746	0.115	0.020	0.095	1.56	0.16	-0.82	0.11	0.0031	4E-17	4.9E-06	1.4E-05	Waist circumference, Hypertension
XKR9	rs268625	rs7828552	0.137	0.051	0.085	-1.84	-1.29	0.66	0.23	0.0009	0.0002	1.1E-05	1.2E-06	Systolic blood pressure post-exercise
C17orf53	rs228769	rs2526021	0.086	0.007	0.079	0.95	1.02	-0.58	0.61	0.0103	0.0001	3.7E-05	0.00035	Bone mineral density

Of the 706 interactions significant after FDR correction, we examined one interaction with the most significant model fit statistic for each of these 79 genes, referencing each regulated gene to the GWAS results catalog described in the methods section. The GWAS results catalog contains SNPs that have been previously associated to a human phenotype, and the associated gene reported by the original GWAS publication. We matched the significant *cis*-epistatic interactions to the GWAS results catalog in two ways: matching the 79 genes being regulated to the gene reported in the GWAS study, and matching SNPs participating in the 706 interactions to a SNP associated in a GWAS study. When matching by gene, we found that 20 of the 79 genes regulated by *cis*-epistasis have been previously reported in studies of approximately 20 human disease and morphological phenotypes (Table 1a). When matching by SNP, we found 10 additional *cis*-interactions where one of the specific SNPs has been associated to one or more disease or morphological phenotypes in humans (Table 1b). These data indicate that genes regulated by *cis*-epistasis are implicated in human phenotypes.

For the majority the genes in Table 1, examining single SNP effects on expression only resulted in a nominal level of statistical significance (Table 1, columns "eQTL[1/2] P-value"). Examining the *cis*-epistasis interaction between the two SNPs allowed us to achieve a much greater degree of statistical significance (Table 1, columns "INT P-value" and "Model P-value"). Furthermore, accounting for *cis*-epistasis allows us to explain a much larger proportion of the heritability (variance) in gene expression (Table 1, column "R^2_{diff}", which is the difference in variance explained by the full model accounting for the interaction, "R^2_{full}", and the reduced model with main effects only, "R^2_{redu}").

3.2. *Structural Characterization of Significant Two-SNP Interactions*

3.2.1. *Genomic Structure*

Next we examined the genomic structural characteristics of the single most significant two-SNP epistatic interaction that impact the expression for each of these 79 genes. Specifically, we examined the location of the two eQTL SNPs relative to each other and relative to the transcription start site (TSS) and transcription end site (TES) of the regulated gene. Based on structural characteristics, we defined four distinct classes of regulatory epistatic interactions: *upstream*, where both eQTL SNPs lie upstream of the TSS of the gene; *downstream*, where both eQTL SNPs lie downstream of the TES; *spanning*, where one eQTL SNP is upstream of the TSS and one eQTL SNP is downstream of the TES; and *intragenic*, where at least one eQTL SNP lies within the genic region, and the other may be either upstream, downstream, or also in the genic region.

We observed 25 upstream interactions (32%), 18 downstream interactions (23%), 17 spanning interactions (21%), and 19 intragenic interactions (24%). Interestingly, all our significant results were evenly distributed among the four structural classes, as a *z*-test for population proportions revealed no significant difference from 25%. However, this test does not account for gene size or SNP density in the surrounding region. Small genes are less likely to harbor spanning or intragenic interactions, and perhaps the fact that we observe an even distribution of genomic structural classes is meaningful. Figure 3 shows that the four structural classes are distributed evenly among these most significant 79 *cis*-epistatic interactions. Figure 3 also reveals that the distribution of structural class does not correlate with gene size, organized vertically along the figure.

3.2.2. *Structure of the Statistical Model*

Statistical epistasis is classically defined as the deviation from additivity in a linear model [29]. We have shown that there are significant nonlinear effects impacting gene expression throughout the genome. Next we examined the structure of the statistical models of the most significant interactions impacting the 79 unique genes discussed above. Specifically, we examined the direction of the coefficients of both main effect terms and the interaction term in each statistical model.

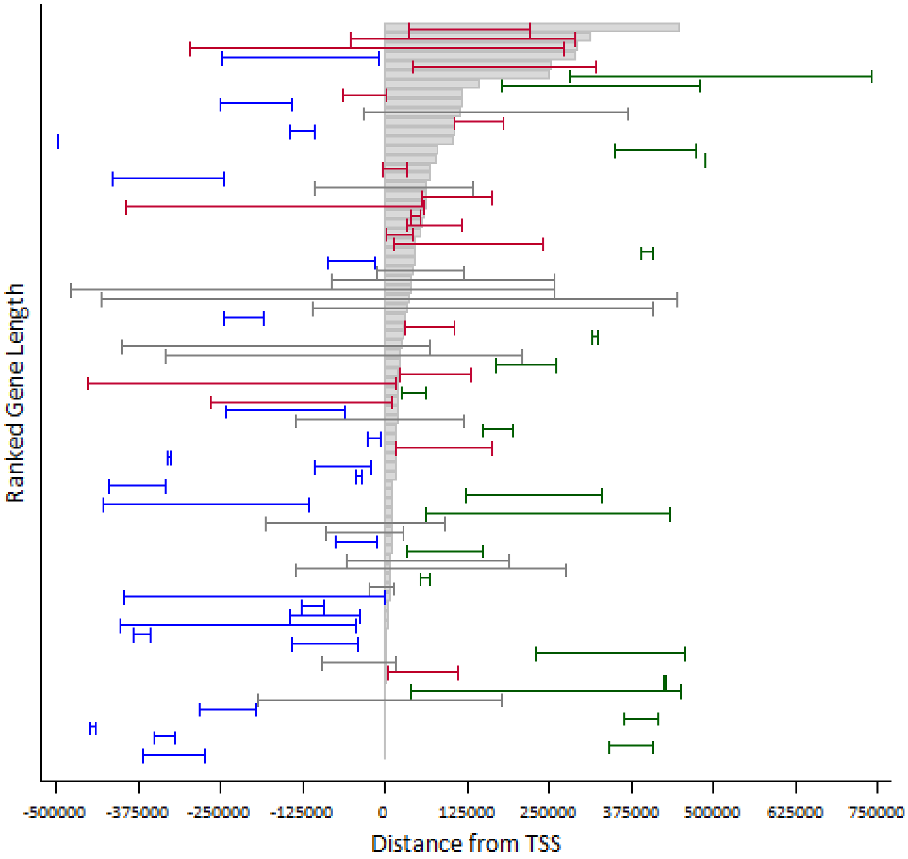


Fig. 3. Transcribed regions of these 79 genes (gray boxes) are aligned by transcription start site, ordered by gene size. Epistatically interacting SNPs that influence the gene's expression are shown as connected hash marks, color coded by class: upstream (blue), downstream (green), spanning (gray) and intragenic (red). Analysis of the genomic structure of *cis*-epistatic interactions reveals that all four structural classes are evenly represented among the most significant *cis*-epistatic interactions, and that structural class does not correlate with gene size.

We found that of these 79 significant *cis*-epistasis interactions, the main effect coefficients in 75 of these models were in the same direction. That is, if inheriting one copy of the minor allele of a single variant caused an increase in expression, the main effect of the other SNP also resulted in an increase in expression. Recall that we only tested for SNP-SNP interactions among eQTL SNPs that had an established main effect. Interestingly, of these 75 *cis*-epistasis models where both main effects were in the same direction, the statistically significant interaction term coefficient was in the *opposite* direction. That is, if the main effect of each variant alone caused an increase in expression by *x* units, inheriting both variants resulted in an expression level that is significantly lower than the expected 2*x* increase. Of the remaining four significant *cis*-epistasis interactions, the main effects were in opposing directions. For three of these four, the main effect coefficient of one SNP in the model approached zero after accounting for the interaction. This suggests a classical modifier effect, where one variant only exerts an effect in the presence of another. In all three of these models, the presence of the "modifier SNP" ($\beta \approx 0$) results in a mitigation of the main effect of the other SNP.

The pattern of coefficients can be seen by examining β_1, β_2, and β_{int} for the models presented in Table 1 (showing only models related to a human phenotype from a GWAS). These results indicate that the overwhelming majority of significant non-additive two-SNP interactions influencing gene expression represent epistatic genetic heterogeneity rather than multiplicative effects. We consider this in greater detail in the discussion section below.

We also investigated the possibility that aspects of the genomic structure of the model might impact the statistical nature of the interaction. However these analyses revealed no significant relationships between genomic structure characteristics (such as class or the physical distance between the two SNPs) to the variance explained (R^2) or magnitude of the interaction coefficient.

4. Discussion

In this work we examined eQTL SNPs known to impact gene expression in humans for non-additive epistatic effects by combining transcriptome-wide expression data from HapMap lymphoblastoid cell lines with genome-wide SNP data from the same cell lines. Specifically, we analyzed over 12 million potential two-SNP interactions for *cis*-epistasis among SNPs known to regulate transcription of a nearby gene, and found that multiple independent eQTL SNPs may often interact to influence gene expression non-additively. After correcting for multiple testing, we found 706 highly significant *cis*-epistasis interactions that influence the expression of 79 unique genes.

We characterized the genomic and statistical structure of the most significant *cis*-epistasis model corresponding to each of these 79 genes. Here we discovered that in the vast majority of *cis*-epistasis interactions (1) the main effects are in the *same* direction, and (2) the interaction was in the *opposite* direction. While still considered a nonlinear epistatic interaction, the structure of this type of model is referred to as a *heterogeneity model* [30, 31] rather than a multiplicative model. While we observe primarily heterogeneity-type models, our particular approach using linear regression may be underpowered to detect models of other statistical structures. Genetic heterogeneity is a serious concern with large-scale genetic studies, and is often cited as a reason for the widespread lack of replication in GWAS studies [32, 33]. Because epistatic genetic heterogeneity may commonly impact regulation of gene expression, and since SNPs associated to complex human phenotypes often result in changes of gene expression [21], it follows that *cis*-epistatic genetic heterogeneity could exert a significant influence over complex human traits and should be investigated as such. Others have recently argued that epistatic genetic heterogeneity should be considered when analyzing genomic data for association to disease [34]. Despite the fact that statistical tools have been available for some time now to accomplish this [35, 36], analyses of genome-wide datasets accounting for the possibility of *cis*-epistasis is a task rarely undertaken. Accounting for genetic heterogeneity in gene expression may improve the replicability of existing personal genomics studies.

By matching *cis*-epistasis interactions to the GWAS results catalogs by SNP, we discovered that of the 79 significant *cis*-epistasis interactions, 10 contained one SNP previously associated to a human phenotype via GWAS studies. Nearly all of these associations fall short of "genome-wide" statistical significance [37] and thus would not be reported in the literature as a relevant gene for the phenotype. Furthermore, the statistical significance of each single SNP on the expression of a gene is weak. However, when we consider the joint effect of both SNPs involved in the *cis*-epistasis interaction, we see a dramatic improvement in the variance of gene expression explained. As such, we hypothesize that some of these reported associations from the GWAS catalog would show stronger associations to the phenotype if modeled with their *cis*-epistasis partner SNP. In light of the prevalence of *cis*-epistatic interactions, these examples provide motivation to re-examine existing datasets for *cis*-epistatic effects on human phenotypes. Our models provide a compelling set of specific regulatory hypotheses to examine in existing data.

Many new approaches have been recently used to examine epistasis in GWAS data [9-12]. All of these approaches focus on interactions among SNPs within genes related to a common biological mechanism, such as pathways, and structural or functional similarity. With these approaches, interaction models consist of SNPs from each of two distant genes – a *trans*-epistasis effect. In most cases, this precludes the possibility of capturing *cis*-epstasis effects. While *trans*-epistasis effects are likely to be important for complex disease etiology, we argue that *cis*-epistasis may be of equal or greater importance, and coupling *cis*- and *trans*-epistasis analysis methods may be more successful.

Furthermore, the collection of available tools for the analysis of multi-locus interactions in personal genomics studies is not likely to discover the *cis*-epstasis effects we describe here. Knowledge-based approaches generally test models of *trans*-epistasis (as discussed above). Sliding window-based haplotype association approaches typically use window sizes based on a fixed physical distance or number of SNPs [38]. These approaches would likely not discover *cis*-epstasis effects due to the variable and often large distances between the pairs of regulatory SNPs within the model (see Figure 3).

Moreover, any gene-based analysis approach that uses SNP data requires mapping SNPs to genes. This is exclusively done using either physical distance (base-pair proximity) or genetic distance (linkage disequilibrium). The genomic window generated using these approaches is typically conservative, including a small region upstream and downstream of the gene region. Others have shown in model organisms that regulatory elements exert effects from extremely long distances [39]. Likewise, the many of the single SNP eQTLs used in the examination of this study illustrate long range regulatory effects [15]. From our analysis, we provide additional evidence that SNPs can influence the regulation of a gene at great distances from the transcription start site, and existing SNP-to-gene mapping approaches should take this into account.

There are several possible molecular phenomena that may underlie these statistical observations. SNPs upstream or downstream of the gene may alter transcription factor binding sites or otherwise affect the efficiency of the transcriptional machinery. SNPs may also alter the binding of micro RNA molecules known to regulate gene transcripts. SNPs in untranslated regions may affect the stability of mRNA molecules. The impact of common variation on these processes is, however, still largely unknown.

We therefore suggest that the re-analysis of existing datasets and the development of new analysis approaches take into account the possibility that long range regulatory interactions could alter gene expression and thus influence human phenotypes. By accounting for more variance in gene expression (thus increasing statistical power), this will improve performance of analytical methods and potentially improve the replicability of GWAS findings. One basic approach would be to use the models we have generated as templates for the analysis of *cis*-epistasis in existing and future personal genomics studies. The 79 genes we identified after multiple testing correction suggest the most compelling cases of *cis*-epistasis. However, interaction models with less significant p-values may explain sufficient variance in a gene's expression to resolve an association with a phenotype.

One limitation of this study is that whole-transcriptome data was available for only 210 HapMap samples. However dense genome-wide SNP data is available for 1397 individuals in 11 diverse human sub-populations through the HapMap project [23], so if additional gene expression data were collected we could improve the statistical power of this analysis to detect *cis*-epistasis effects. Also, we only considered interactions among eQTL SNPs with a known regulatory effect ($p < 0.05$). A reanalysis of this data including all SNPs, (even those without a known regulatory effect) would be straightforward, perhaps revealing additional *cis*-epistasis effects; however this would cause a power loss from the increased burden of multiple testing correction.

In summary, we have shown that *cis*-epistasis is an important phenomenon regulating gene expression in humans. Using this information, we suggest ways in which the performance of existing and future analysis approaches can be improved, and how additional insights into human biology and disease pathogenesis could be gained from personal genomics studies.

References

1. A. L. Tyler, F. W. Asselbergs, S. M. Williams, J. H. Moore, *Bioessays* **31**, 220 (2009).
2. H. Shao *et al.*, *Proc. Natl. Acad. Sci. U. S. A* **105**, 19910 (2008).
3. X. He, W. Qian, Z. Wang, Y. Li, J. Zhang, *Nat. Genet.* **42**, 272 (2010).
4. E. E. Eichler *et al., Nat. Rev. Genet.* **11**, 446 (2010).
5. T. A. Manolio *et al., Nature* **461**, 747 (2009).
6. B. Maher, *Nature* **456**, 18 (2008).
7. P. S. Aguilar *et al., Nat. Struct. Mol. Biol* **17**, 901 (2010).

8. M. Costanzo *et al.*, *Science* **327**, 425 (2010).
9. S. E. Baranzini *et al.*, *Hum. Mol. Genet.* **18**, 2078 (2009).
10. W. S. Bush, S. M. Dudek, M. D. Ritchie, *Pac Symp Biocomput* **14**, 368 (2009).
11. G. Peng *et al.*, *Eur. J Hum. Genet.* **18**, 111 (2010).
12. D. Ruano *et al.*, *Am. J Hum. Genet.* **86**, 113 (2010).
13. J. K. Pickrell *et al.*, *Nature* **464**, 768 (2010).
14. B. E. Stranger *et al.*, *Science* **315**, 848 (2007).
15. J. B. Veyrieras *et al.*, *PLoS. Genet.* **4**, e1000214 (2008).
16. T. Ravasi *et al.*, *Cell* **140**, 744 (2010).
17. S. F. Boj, D. Petrov, J. Ferrer, *PLoS. Genet.* **6**, e1000970 (2010).
18. W. Du, D. Thanos, T. Maniatis, *Cell* **74**, 887 (1993).
19. J. W. Gregersen *et al.*, *Nature* **443**, 574 (2006).
20. A. L. Dixon *et al.*, *Nat. Genet.* **39**, 1202 (2007).
21. E. R. Gamazon *et al.*, *Bioinformatics* **26**, 259 (2010).
22. M. D. Mailman *et al.*, *Nat. Genet.* **39**, 1181 (2007).
23. International hapmap consortium, *Nature* **449**, 851 (2007).
24. R Development Core Team, *R: A language and environment for statistical computing. ISBN 3900051070, URL http://www.R-project.org.*
25. J. D. Storey, J. E. Taylor, D. Siegmund, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **66**, 187 (2004).
26. S. Purcell *et al.*, *Am. J. Hum. Genet.* **81**, 559 (2007).
27. L. A. Hindorff *et al.*, *Proc. Natl. Acad. Sci. U. S. A* **106**, 9362 (2009).
28. A. D. Johnson, C. J. O'Donnell, *BMC Med. Genet.* **10**, 6 (2009).
29. R. A. Fisher, *Trans. R. Soc. Edinb.* **52**, 399 (1918).
30. H. J. Cordell, *Hum. Mol. Genet.* **11**, 2463 (2002).
31. R. J. Neuman, J. P. Rice, *Genet. Epidemiol.* **9**, 347 (1992).
32. J. McClellan, M. C. King, *Cell* **141**, 210 (2010).
33. M. J. Sillanpaa, K. Auranen, *Ann. Hum. Genet.* **68**, 646 (2004).
34. J. H. Moore, F. W. Asselbergs, S. M. Williams, *Bioinformatics* **26**, 445 (2010).
35. K. L. Lunetta, L. B. Hayward, J. Segal, E. P. Van, *BMC Genet.* **5**, 32 (2004).
36. T. A. Thornton-Wells, J. H. Moore, J. L. Haines, *Trends Genet.* **20**, 640 (2004).
37. I. Pe'er, R. Yelensky, D. Altshuler, M. J. Daly, *Genet. Epidemiol.* **32**, 381 (2008).
38. S. Lin, A. Chakravarti, D. J. Cutler, *Nat. Genet.* **36**, 1181 (2004).
39. R. L. Chandler, K. J. Chandler, K. A. McFarland, D. P. Mortlock, *Mol. Cell Biol.* **27**, 2934 (2007).

HAPLOTYPE INFERENCE FROM SHORT SEQUENCE READS USING A POPULATION GENEALOGICAL HISTORY MODEL

JIN ZHANG and YUFENG WU*

Department of Computer Science and Engineering
University of Connecticut
Storrs, CT 06269, U.S.A.
E-mail: {jinzhang,ywu}@engr.uconn.edu

High-throughput sequencing is currently a major transforming technology in biology. In this paper, we study a population genomics problem motivated by the newly available short reads data from high-throughput sequencing. In this problem, we are given *short* reads collected from individuals in a population. The objective is to infer haplotypes with the given reads. We first formulate the computational problem of haplotype inference with short reads. Based on a simple probabilistic model on short reads, we present a new approach of inferring haplotypes directly from given reads (i.e. without first calling genotypes). Our method is finding the most likely haplotypes whose local genealogical history can be approximately modeled as a perfect phylogeny. We show that the optimal haplotypes under this objective can be found for many data using integer linear programming for modest sized data when there is no recombination. We then develop a related heuristic method which can work with larger data, and also allows recombination. Simulation shows that the performance of our method is competitive against alternative approaches.

Keywords: High-throughput sequencing; haplotype inference; bioinformatics algorithms; population genomics.

1. Introduction

High throughput DNA sequencing is increasingly recognized as a major transforming technology in biology. During the last decade, several novel high throughput sequencing (HTS) technologies have been developed and commercialized (such as the Roche 454 FLX, Illumina Genome Analyzer, and ABI SOLiD), and several more are under development. These high throughput technologies dramatically bring down the sequencing cost and are generating huge amount of data. Several individual genomes have been sequenced,[1,2] and an effort is underway to sequence one thousand individuals.[3] Sequencing may give entire *diploid* genomes of individuals in a population and potentially reveal *all* the common (and many of the *rare*) variations in the sequenced region. Thus, increasingly complete sequencing using HTS technologies will become the preferred approach to attack population genomics problems.

On the other hand, the current HTS technologies have some technical *limitations.* First, the reads generated by HTS technologies are often *short.* Although longer sequence reads may become available in the near future,[4] it is expected that short sequence reads are likely to be still useful in the coming years. Thus, we focus on short reads in this paper. Second, many HTS technologies have higher error rates than the traditional Sanger sequencing. Some technologies have error rates of 1% or even higher, which can make it difficult to distinguish between error and population-scale variation. Additional error sources include inaccurate sequence

*Corresponding author.

reads mapping (i.e. locating the reads within a reference genome). Sometimes high coverage sequencing may reduce the noise and uncertainty, but with increased cost. Therefore, robust data analysis methods are needed to process the (somewhat noisy) HTS data.

In this paper, we focus on a population genomics problem: inference of a pair of haplotypes for each individual in the population from the given HTS reads for diploid organisms (such as human). Here, a haplotype refers to the DNA sequence collected from the same chromosome, which describes the alleles at polymorphic sites on this chromosome. Collecting haplotypes from populations is an important population genomics problem, which is evident in the HapMap project.[5,6] See Section 2 for more description on haplotypes. To formulate a concrete computational problem, we make several assumptions:

(1) In this paper, we only consider *short* reads. That is, our problem is different from the haplotype reconstruction problem based on *long* sequence reads (e.g. Bansal and Bafna,[7] He, et al.[8]). Since the sequence reads are short and often the variations in a population are relatively sparsely located along the genome, we assume that a short (single or paired-end) read covers no more than *one* SNP site. When there is a read covering more than one SNP sites, our current implementation treats this read as multiple reads, each covering one SNP, although our implementation can be modified to use the haplotype phase information contained in such reads.
(2) We do *not* consider pooling here: we know the individual a sequence read originates.
(3) In this paper, we only concern single nucleotide polymorphisms (SNPs), which can be stated as a binary value: 0 or 1.
(4) A standard analysis step in analyzing short reads is *mapping* the short reads against a reference genome (which we assume is available). We assume that reads mapping is performed properly so that reads covering one polymorphic site are properly mapped. We only consider reads that are uniquely mapped and remove reads that are ambiguous in mapping. Once the reads are mapped, we can identify polymorphic genomic positions by comparing the mapped reads with the reference genome. Thus, we assume that the SNP sites can be determined from the mapped reads.

We are now ready to define the precise problem formulation.

Haplotype Inference with Single Short Reads. We are given a set of mapped single short reads $\mathcal{R}$, each covering a specific SNP site. That is, a sequence read reports an allele at a polymorphic site for an individual, but we do not know which homologous chromosome it comes from and also there is some chance the allele reported is incorrect. The goal is inferring two haplotypes for each individuals from the reads $\mathcal{R}$.

Note that our method also calls genotypes: once haplotypes are inferred, we can obtain genotypes from the haplotypes. This problem formulation may be useful for (1) sequencing a new population, where no previously sampled population haplotypes (such as those provided by the HapMap project) are available, and (2) whole-genome sequencing, where we want to infer haplotypes for *all* SNPs (not only common SNPs but also *rare* SNPs). We note that rare variants are becoming more important in understanding genotype-phenotype association.[9]

2. Background

2.1. *Haplotypes and Genealogical History*

An important genetic variation is the single nucleotide polymorphism (SNP). A SNP site in the genome can generally take only two states (alleles) among the individuals in a population. Thus, we use binary alleles (0 and 1) to represent the state at any SNP site. In this paper, we focus on SNPs and do not consider other variations such as copy number variation (CNV) or polymorphism (CNP). Often, we collect genetic variations data at multiple genomic sites. We call a sequence of genetic variations at these sites a *haplotype*. A haplotype based on SNPs can be represented as a *binary* vector. A *diploid* organism (such as human) has two haplotypes per chromosome, and although these are often called 'copies', they are not identical. A description of the conflated (mixed) data from the two haplotypes is called a *genotype*. When both haplotypes have state 0 (resp. 1) at a site, the genotype has state 0 (resp. 2), and is called a *homozygote*. Otherwise, the genotype has state 1 at that site and is called a *heterozygote*. We let n be the number of individuals sampled in the population, and m be the number of SNP sites. The genotypes of these individuals are represented by an n by m matrix with entries 0/1/2, while their haplotypes are represented by a $2n$ by m binary matrix. We call the two ordered alleles from the two haplotypes at a single site of a diploid individual *diploid type*. Diploid type can be 0/0, 0/1, 1/0 and 1/1. Note that there are two diploid type (0/1 and 1/0) for the same genotype 1.

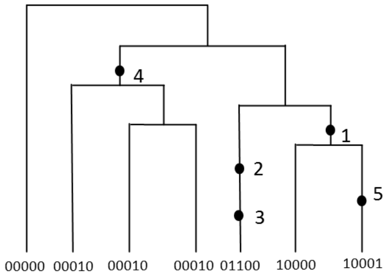


Fig. 1. *A genealogical tree.*

Genealogical history of sequences in a population explicitly shows the origin and derivation of extant sequences, the locations of all the genomic alterations (both in the genome and in time), and how the variants are transmitted from parents to descendants. The simplest genealogical model is the tree model, when recombination is ignored. See Figure 1 for an illustration. A common assumption is that at most one mutation occurs at any site, which is supported by the *infinite sites model*[10] from population genetics. We assume infinite sites model throughout this paper. Therefore, the genealogical tree is a perfect phylogeny (see, e.g. Gusfield[11]). A perfect phylogeny implies that at any two SNP sites, the four ordered pairs of alleles 00, 01, 10 and 11 (called gametes) can *not* be all present (called four-gamete test in population genetics). Two sites satisfying this property are said to be *compatible*. If all pairs of sites are compatible, the sequences allow a perfect phylogeny (see e.g. Gusfield[11]). Note that although gamete and diploid type use similar values, conceptually they are different: gamete means the setting of the two alleles at two sites of the same haplotype, while diploid type is for the two alleles at the same site of an individual.

When meiotic recombination is considered, a more complex model is needed. Recombination takes two homologous chromosomes (haplotypes) and produces a third chromosome consisting of alternating segments (usually a small number) of the two chromosomes. With recombination, genealogical history can no longer be modeled as a single tree. Nonetheless, sometimes we can use local trees to represent local genealogical history for a short region, within which recombination does not affect the genealogy of the sampled sequences.

2.2. *High Throughput Sequencing (HTS)*

A main application of the HTS is on *resequencing*. In resequencing, we want to find genetic variations (e.g. SNPs) in a sample of individuals by sequencing the genomes of those individuals, when an existing, fully-sequenced, reference genome is already known. The general procedure for many resequencing applications is to first find where a new sequence read originates by comparing the sequence read with the reference genome (called reads mapping). Once the originating positions of sequence reads are found, we can then examine the mapped reads to find variations such as SNPs.

The current HTS data does not contain information on which of the two haplotypes (from a diploid organism) a read is from. This often adds complexity to data analysis. For example, suppose we have two mapped sequence reads that give the same alleles as the reference genome. We can not assert that the individual is a homozygote because the two reads may come from the *same* haplotype, and yet the sequenced individual is a heterozygote at the site. Moreover, suppose we have two mapped reads that give allele 0 and 1 at a SNP site. The individual can still be a homozygote 0 if the read with 1 allele is caused by a sequencing error. See Figure 2 for an illustration of sequencing diploid samples.

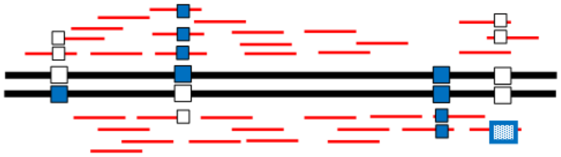


Fig. 2. *Illustration of the HTS technologies. Two thick lines are the two haplotypes of a diploid individual. Boxes are the genetic variations (e.g. SNPs), where colors indicate different allele. The short, red lines are the short sequence reads from this diploid individual, which are mapped to the proper location. The read with a dotted box (on lower right) has a sequencing error.*

3. Haplotyping with Short Reads

Haplotype inference from given *genotypes* has been actively studied recently.[12–15] Thus, a straightforward approach of inferring haplotypes with short reads is a two-stage one: first call the genotypes from the given reads (say taking the genotypes with the highest posterior probability as described in Section 3.1) and then run a population haplotype inference program (e.g. fastPHASE[15]) on the called genotypes. The main problem with this two-stage approach is that inaccurately called genotypes may lead to haplotypes of low quality. This is especially a concern when the sequencing coverage is low, which may lead to more noise in the called genotypes. In this paper, we present a new method based the *one-stage* approach, which infers haplotypes directly from the reads (i.e. without calling genotypes first). We note that few published haplotype inference approaches work directly on sequence reads, with the exception of program Beagle.[16] In Section 4, we compare out method with program Beagle.

3.1. *Posterior Probability of Genotypes at a Single Site*

Given the reads at a SNP site, it is easy to compute the posterior probability of a genotype. For ease of exposition, we assume each read has probability ϵ of reporting an incorrect allele at the site. Note that it is straightforward to allow reads having reads-specific error probability. Consider an individual i with genotype g at site s_j. We let $\mathcal{R}_{i,j}$ be the reads covering s_j for individual i, which report $r_{i,j,0}$ 0-allele and $r_{i,j,1}$ 1-allele for s_j. The single SNP genotypic posterior probability is the probability of observing a particular genotype $g \in \{0, 1, 2\}$ at a site s_j given all the reads for all individuals at this site (denoted as $\mathcal{R}_{-,j}$). We define $f_j(g)$ as the genotype frequency for genotype g at site s_j. We assume that the read of interest was obtained with equal prior probability from either haplotype. Now the posterior probability of genotype g can be calculated [b]:

$$P(g = 0|\mathcal{R}_{-,j}) \propto P(\mathcal{R}_{-,j}|g = 0)P(g = 0) = (1 - \epsilon)^{r_{i,j,0}} \epsilon^{r_{i,j,1}} f_j(0) \tag{1}$$

$$P(g = 1|\mathcal{R}_{-,j}) \propto P(\mathcal{R}_{-,j}|g = 1)P(g = 1) = 0.5^{r_{i,j,0}+r_{i,j,1}} f_j(1) \tag{2}$$

$$P(g = 2|\mathcal{R}_{-,j}) \propto P(\mathcal{R}_{-,j}|g = 2)P(g = 2) = (1 - \epsilon)^{r_{i,j,1}} \epsilon^{r_{i,j,0}} f_j(2) \tag{3}$$

We use the Hardy-Weinberg equilibrium to estimate genotype frequency $f_j(g)$ at site s_j, from the frequency of alleles 0 and 1 in the population. Allele frequency can be estimated from the reads $\mathcal{R}_{-,j}$ from the observed alleles at site j. Once posterior probability is computed, a simple two-stage approach calls genotypes at each locus by picking the genotypes with maximum posterior probability, and then infer haplotypes for the called genotypes using some population haplotype inference method. As shown in Section 4, this approach is generally not as accurate as the one-stage approach we now present.

3.2. *The Special Case: No Recombination with Small Number of SNPs*

We now present an *one-stage* approach, which infers haplotypes from short reads directly. Our method rely on the shared *genealogical history* of the sampled sequences to infer haplotypes. To get started, we first consider the case when there is *no* recombination. Later, we will extend our method to allow recombination.

When there is no recombination, the underlying genealogy is a perfect phylogeny. Gusfield[13] first exploited the approach of inferring haplotypes with the perfect phylogeny model. Here, we develop a perfect phylogeny based method for inferring haplotypes with short reads. That is, we want to infer haplotypes that allow a perfect phylogeny. Note that perfect phylogeny alone can not determine the haplotypes since there are many possible haplotypes allowing perfect phylogeny. Since some haplotypes fit the given short reads better than others, a natural objective is to find the haplotypes that allow a perfect phylogeny *and* the probability of short reads given these haplotypes is maximized.

We now give the technical details. The short reads based perfect phylogeny haplotyping is, given short reads $\mathcal{R}$, finding a set of haplotypes H s.t. $P(\mathcal{R}|H)$ is maximized *and* H allows a perfect phylogeny. We let H_i denote the i-th haplotype, where $1 \leq i \leq 2n$. We let $H_{i,j}$ denote the allele (0 or 1) at the j-th site of H_i. As before, we let $\mathcal{R}_{i,j}$ be the set of reads that are taken from individual i, and cover site s_j. Consider a read $R_{i,j,k} \in \mathcal{R}_{i,j}$, which reports allele $k \in \{0, 1\}$ for site s_j. Now, $P(R_{i,j,k}|H)$ depends on $H_{2i-1,j}$ and $H_{2i,j}$. The following is related to equations 1 to 3 in Section 3.1.

[b] Similar equations have been used in Duitama, et al.,[17] and also in other statistical genetics papers

$$P(R_{i,j,k}|H) = 0.5 \times (P(R_{i,j,k}|H_{2i-1,j}) + P(R_{i,j,k}|H_{2i,j}))$$

Here, $P(R_{i,j,k}|h) = \epsilon$ if $k \neq h$, and $1-\epsilon$ otherwise. We consider all the reads covering site s_j in individual i, where there are $r_{i,j,0}$ 0-reads and $r_{i,j,1}$ 1-reads. Then, based on the assumption that all reads are independent, we have:

$$logP(\mathcal{R}_{i,j}|H) = r_{i,j,0}log(P(R_{i,j,0}|H)) + r_{i,j,1}log(P(R_{i,j,1}|H))$$

When $H_{2i-1,j}$ and $H_{2i,j}$ are known, these two alleles determine the diploid type $d(H) \in \{0/0, 0/1, 1/0, 1/1\}$. To simplify notations, we simply use d for diploid type. When d is given, $H_{2i-1,j}$ and $H_{2i,j}$ are also known. We let $w_{i,j,d} = logP(\mathcal{R}_{i,j}|H)$, where d is the diploid type at site s_j of individual i. We assume the reads are independent, since reads are short and thus can be treated as independent given the haplotypes. Note, however, that in practice there may exist other factors such as mapping bias that can make this assumption less accurate. Then,

$$logP(\mathcal{R}|H) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j,d}$$

Our goal is finding haplotypes H, s.t. H allows a perfect phylogeny and $logP(\mathcal{R}|H)$ is maximized. Since $logP(\mathcal{R}|H)$ can be computed easily for fixed H, naively we can enumerate all possible haplotypes H to find the ones that allow a perfect phylogeny and maximize $logP(\mathcal{R}|H)$. But this is infeasible even for data of moderate size. We do not currently know an efficient algorithm for finding the optimal solution. To develop a practical method, we use integer linear programming (ILP) to solve the optimization problem *exactly.*

In our ILP formulation, we have a binary variable $D_{i,j,d}$ for individual i, site s_j and diploid type $d \in \{0/0, 0/1, 1/0, 1/1\}$, where $D_{i,j,d} = 1$ if the diploid type formed by $H_{2i-1,j}$ and $H_{2i,j}$ is d. That is, $D_{i,j,d}$ specifies which diploid type individual i carries at site s_j. For any two sites s_{j_1} and s_{j_2}, we define a binary variable $G_{j_1,j_2,g}$. $G_{j_1,j_2,g} = 1$ if sites s_{j_1} and s_{j_2} have gamete $g \in \{00, 01, 10, 11\}$. Now we give the sketch of the ILP formulation.

Objective: maximize $\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{d\in\{0/0,0/1,1/0,1/1\}} w_{i,j,d} \times D_{i,j,d}$.
Subject to

1 $D_{i,j,0/0} + D_{i,j,0/1} + D_{i,j,1/0} + D_{i,j,1/1} = 1$, for each $1 \le i \le n$ and $1 \le j \le m$.

[We now impose constraints on $G_{j_1,j_2,d}$. We only give the constraints for $G_{j_1,j_2,00}$. The rest are similar and thus omitted.]

2 $G_{j_1,j_2,00} + D_{i,j_1,1/1} \ge D_{i,j_2,0/0}$, for all $1 \le j_1 < j_2 \le m$ and $1 \le i \le n$.

3 $G_{j_1,j_2,00} + D_{i,j_2,1/1} \ge D_{i,j_1,0/0}$, for all $1 \le j_1 < j_2 \le m$ and $1 \le i \le n$.

4 $G_{j_1,j_2,00} + 1 \ge D_{i,j_1,0/1} + D_{i,j_2,0/1}$, for all $1 \le j_1 < j_2 \le m$ and $1 \le i \le n$.

5 $G_{j_1,j_2,00} + 1 \ge D_{i,j_1,0/1} + D_{i,j_2,0/0}$, for all $1 \le j_1 < j_2 \le m$ and $1 \le i \le n$.

6 $G_{j_1,j_2,00} + 1 \ge D_{i,j_1,1/0} + D_{i,j_2,1/0}$, for all $1 \le j_1 < j_2 \le m$ and $1 \le i \le n$.

7 $G_{j_1,j_2,00} + 1 \ge D_{i,j_1,1/0} + D_{i,j_2,0/0}$, for all $1 \le j_1 < j_2 \le m$ and $1 \le i \le n$.

[We now ensure no four gametes exists at any pair of sites]

8 $G_{j_1,j_2,00} + G_{j_1,j_2,01} + G_{j_1,j_2,10} + G_{j_1,j_2,11} \le 3$ for all $1 \le j_1 < j_2 \le m$.

For each $1 \le i \le n$, $1 \le j \le m$ and $d \in \{0/0, 0/1, 1/0, 1/1\}$, there is a binary variable $D_{i,j,d}$.

For each $1 \leq j_1 < j_2 \leq m$ and $g \in \{00, 01, 10, 11\}$, there is a binary variable $G_{j_1,j_2,g}$.

Briefly, constraint (1) states that each individual must take exactly one of the four diploid type at a site. Constraints (2) to (7) relate diploid variables $D_{i,j,d}$ with the gamete variables $G_{j_1,j_2,00}$. For example, constraint (2) states that if the diploid type at site s_{j_1} is not 1/1 (i.e. $D_{i,j_1,1/1} = 0$) and the diploid type at site s_{j_2} is 0/0 (i.e. $D_{i,j_2,0/0} = 1$), then there exists gamete 00 at sites s_{j_1} and s_{j_2}. Constraint (8) states that there are at most three gametes for any two sites, which is required by the perfect phylogeny model. The constraints for the other diploid type variables are similar. Finally, the objective function uses the diploid type variables times the weights, which means that only the selected diploid types (i.e. $D_{i,j,d} = 1$) contribute to the objective. Once the ILP formulation is solved, the haplotypes are readily retrieved from the values of the $D_{i,j,d}$ variables.

Simulation in Section 4 shows that this ILP formulation can be practically solved for many data, especially when the number of sites (i.e. m) is relatively small (say less than 20).

3.3. *The General Case: with Recombination and Larger Number of SNPs*

When data size grows or recombination occurs, we can no longer directly use the ILP-based approach in Section 3.2. We now extend our approach to handle data with recombination and/or larger number of sites. Our strategy is similar in high-level to the approach in Halperin and Eskin:[14] we first infer haplotypes using the ILP based approach in Section 3.2 on small number of consecutive (and overlapping) SNPs (called *windows*); then we *concatenate* these overlapping haplotypes to create complete haplotypes for the entire data. This approach may work well when recombination rate is relatively low: in this case, there are relatively long genomic regions with no recombination. Also, even when there is a small number of recombinations within a region, perfect phylogeny may still be a good approximation of the genealogical history of the region.

Specifically, we let the size of the sliding window (i.e. number of sites) be W, which starts from the first site. Each time, we move the window to the right by $\frac{W}{2}$ sites to obtain haplotypes in a list of overlapping windows by the ILP approach. Then, we concatenate the haplotypes of the overlapped windows from the left to the right. Let h_{2i-1} and h_{2i} be the haplotypes of individual i in a window, and h'_{2i-1} and h'_{2i} be the haplotypes in an overlapping window. Note that the haplotypes of an individuals within two overlapped windows are obtained from different ILP solutions, and thus the two pairs of haplotypes need to be paired up properly. Moreover, sometimes concatenation may require changes to these haplotypes for consistency.

Here are the main steps of haplotype concatenation.

(1) First concatenate obvious haplotypes. Sometimes only one pairing between the two pairs of haplotypes is perfect (e.g. the overlapped portions of h_{2i-1} and h'_{2i-1} match perfectly, so do those of h_{2i} and h'_{2i}, and the other pairing of the haplotypes is not perfect). In this case, we simply greedily choose the obvious pairing to obtain two concatenated haplotypes.
(2) The previous step often generates a set of inferred haplotypes. Now we use these already inferred haplotypes to help to resolve the other undecided haplotype pairs. If two haplotypes (say h_{2i-1} and h'_{2i-1}) can be merged perfectly (i.e. with no mismatches within the

overlapped region) to generate one of the existing haplotypes, we just take this particular pairing if the other haplotype pair is approximately consistent.

(3) Since haplotypes within a window are usually closely related through mutation and recombination, this provides more hints on how to concatenate the haplotypes. Suppose we are evaluating two choices of pairing, which generate two sets of candidate haplotypes. We compare the two sets of haplotypes and choose the ones that are closely related to the already inferred haplotypes. A haplotype h is closely related to a set of haplotypes H if (a) the Hamming distance between h and a haplotype $h' \in H$ is small, or (b) h can be broken into a small number of segments, s.t. each segment appears in H. The later can be easily evaluated by either a dynamic programming algorithm or a greedy algorithm.[18]

(4) Here is one more rule in deciding how to concatenate the haplotypes, which is applied if the previous step leads to multiple equally good choices. When recombination occurs, some pairs of sites become incompatible. However, a site is still likely to be compatible with its neighboring sites.[19] For a site s, the compatible region of s is a continuous set of sites, each of which is compatible with s (but there may exist two incompatible sites among these sites other than s). Based on this observation, we select the haplotype pairings that give longer compatible regions.

4. Results

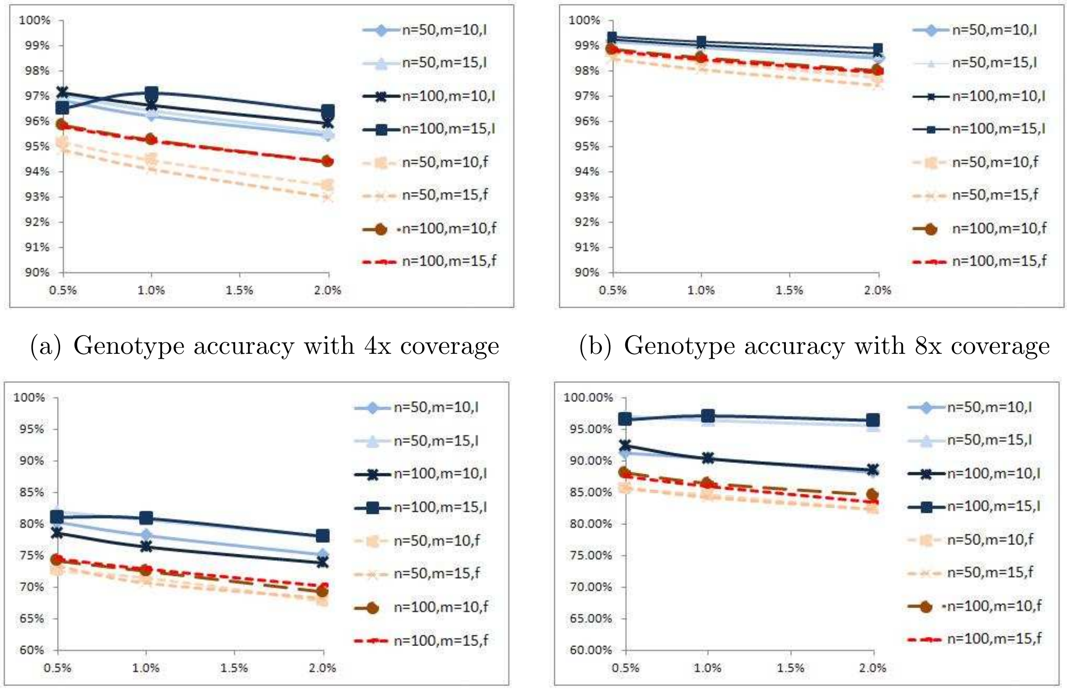
We have implemented our method in a program (called HapReads) written with C++, which uses either CPLEX (a commercial and faster ILP solver) or GNU GLPK ILP solver. HapReads can be downloaded from: http://www.engr.uconn.edu/˜jiz08001. Our simulation results are from the CPLEX version. We test our method on simulated data on a 3192 MHz Intel Xeon workstation. We use Hudson's program ms[20] to generate haplotypes for different settings on the number of diploid individuals, the number of sites and recombination rate. Then, for each set of haplotypes, we simulate the sequence reads by (1) deciding the number of reads to generate based on the sequencing coverage, and (2) randomly picking the sites for the reads and one of the two haplotypes when reporting the alleles in the reads. To simulate the sequencing errors and other noise, we generate sequence reads with some error probability ϵ (the probability of reporting a wrong allele). We generate 100 datasets for each setting.

To evaluate the accuracy of our method (and the two-stage approach using fastPHASE), we compare the inferred haplotypes with the true simulated haplotypes. We run program fastPHASE by letting the program to choose the number of clusters itself. We assume error probability ϵ is known to *both* our method and fastPHASE. Different from haplotyping from given genotypes (where there is only phasing errors), there are two types of errors: (a) genotype errors, and (b) haplotype phase errors. Genotype errors refer to the genotypes (implied by the inferred diploid types) that are different from the true genotypes. We define genotype accuracy A_g as the percentage of correctly called genotypes. We define phase accuracy A_p as the switching accuracy[21] that is related to the incorrectly phased neighboring heterozygotes. Note that calculating phase accuracy needs first *correcting* the genotype errors (i.e. changing the diploid types in some ways so that the corresponding genotypes match the true genotypes). There is a subtle issue in computing phase accuracy A_p when there are genotype errors.

Suppose the true diploid type is 0/1 (i.e. heterozygote), while the inferred diploid type is 0/0 (i.e. homozygote). We can use either 0/1 or 1/0 to correct the genotype that may lead to different phase accuracy. To get over this issue, we use the *average* phase accuracy over all possible choices for these corrected diploid types.

4.1. *Accuracy of the ILP formulation*

We first evaluate the accuracy of the ILP-based approach in Section 3.2. Recall that the ILP approach is practical when there is no recombination and the number of sites is relatively small. In Figure 3, we show the average (genotype and phase) accuracy for various number of individuals and sites, sequence read error rates and coverage.



(a) Genotype accuracy with 4x coverage

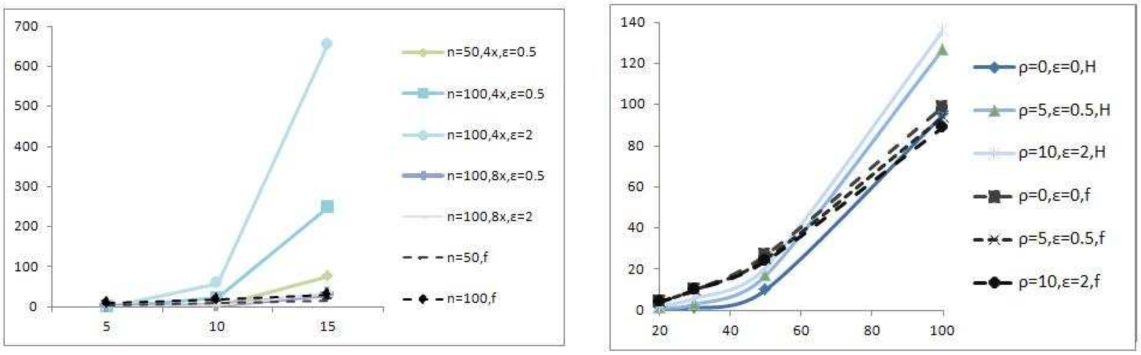
(b) Genotype accuracy with 8x coverage

(c) Phase accuracy with 4x coverage

(d) Phase accuracy with 8x coverage

Fig. 3. Accuracy of ILP-based method and fastPHASE under different reads error rates and coverage. n: the number of individuals, m: the number of sites. I: ILP-based method (solid line). f: fastPHASE (dashed line).

Figure 3 shows that our ILP approach outperforms the two-stage approach using fastPHASE (or simply fastPHASE) in both genotype accuracy and phase accuracy in most datasets of the simulations. For example, for 50 individuals, 15 sites, error rate 1% and coverage 4x, the phase accuracy of our method is roughly 10% more than that of the two-stage approach, even when the difference between genotype accuracy is about 2.5%. This suggests that our method works well in inferring haplotypes when there is no recombination. As expected, when read error rate is higher and coverage is lower, phase accuracy tends to be lower. One downside is that the ILP solving gets slower when the number of sites increases, which is shown in Figure 4(a).



(a) Our ILP-based method and fastPHASE

(b) Our heuristic approach and fastPHASE

Fig. 4. Running time of ILP-based method, heuristic approach and fastPHASE. n: the number of individuals. x: reads coverage. ϵ: reds error rates. ρ: recombination rate. H: our heuristic approach. f: fastPHASE.

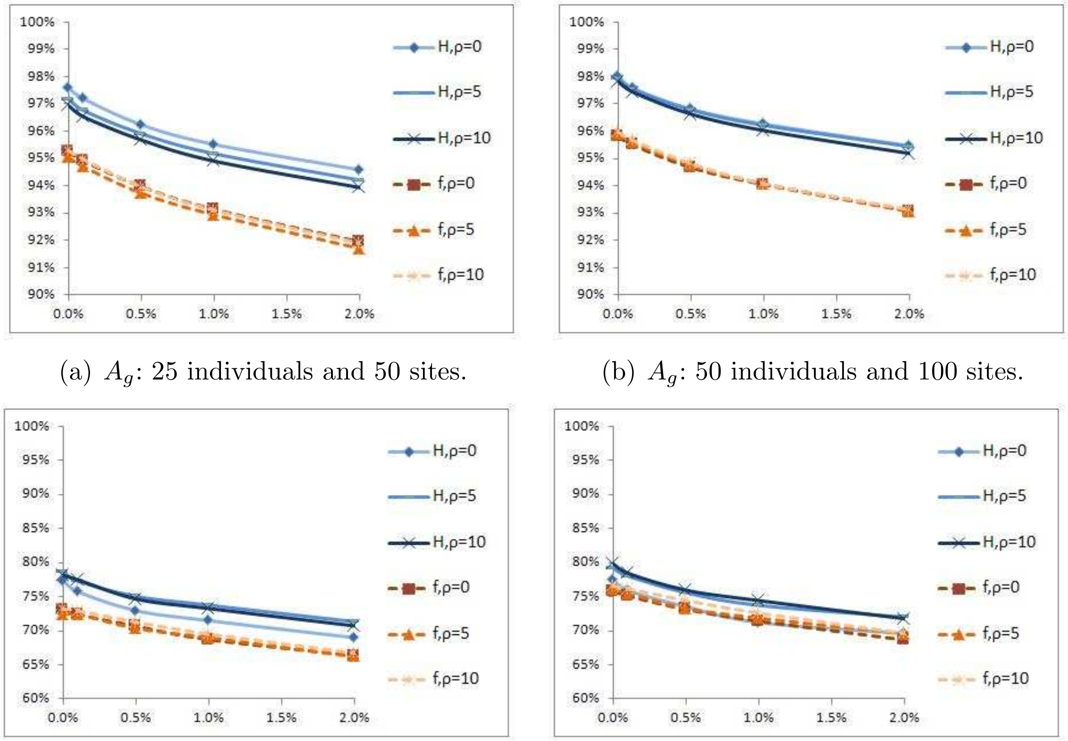
4.2. *Accuracy of the case with larger data*

We now evaluate the performance of the heuristic approach in Section 3.3, which allows us to handle problem instances that are larger or with recombination. We use 4x coverage in this simulation. The results are obtained by inferring haplotypes from a sliding window of 10 sites, and then concatenating the overlapped haplotypes.

Figure 5 shows that in terms of genotype accuracy and phase accuracy, our one-stage approach is consistently more accurate. Thus, the simulation results show that our one-stage approach outperforms the two-stage approach in general. Also, our method remains reasonably accurate with higher sequence reads error (up to 2%) or when recombination rate increases (up to 10). We note that genotype accuracy in our simulation is often fairly accurate. Phase accuracy, on the other hand, is in general not very high for both methods. One reason may be the low sequencing coverage: we use 4x coverage here and increasingly coverage may improve the phase accuracy. Moreover, as shown in Figure 4(b), the running time of our method is similar to the two-stage approach for the data we simulate.

4.3. *Comparing with program Beagle with simulated and biological data*

Program Beagle[16] allows uncertain genotypes which are specified by genotype probabilities. Thus, Beagle can be used as a one-stage approach so we compare program Beagle and our approach. We run program Beagle with the same data sets generated by program ms in Section 4.2. The result is given in Figure 6. For data sets with 25 individuals and 50 sites, our method and Beagle have similar genotype accuracy and our method has slightly higher phase accuracy, but from data sets with 50 individuals and 100 sites, our method is less accurate than Beagle. We also test the two approaches on simulated reads for HapMap haplotypes. We generated 100 data sets of 25 individuals by 50 sites from 100 regions on chromosome 1 of CEU population. The results are similar (results omitted, with Beagle being slightly more accurate). One possible reason is that HapMap haplotypes are only for common SNPs, where haplotypes within a window are less likely to allow a perfect phylogeny. More simulations are needed to further compare the two methods. Overall, one-stage approaches appear to perform



(a) A_g: 25 individuals and 50 sites.

(b) A_g: 50 individuals and 100 sites.

(c) A_p: 25 individuals and 50 sites.

(d) A_p: 50 individuals and 100 sites.

Fig. 5. Accuracy of heuristic approach and fastPHASE with different reads error rates. H refers to heuristic approach (solid lines) and f refers to fastPHASE (dashed lines). ρ is recombination rates.

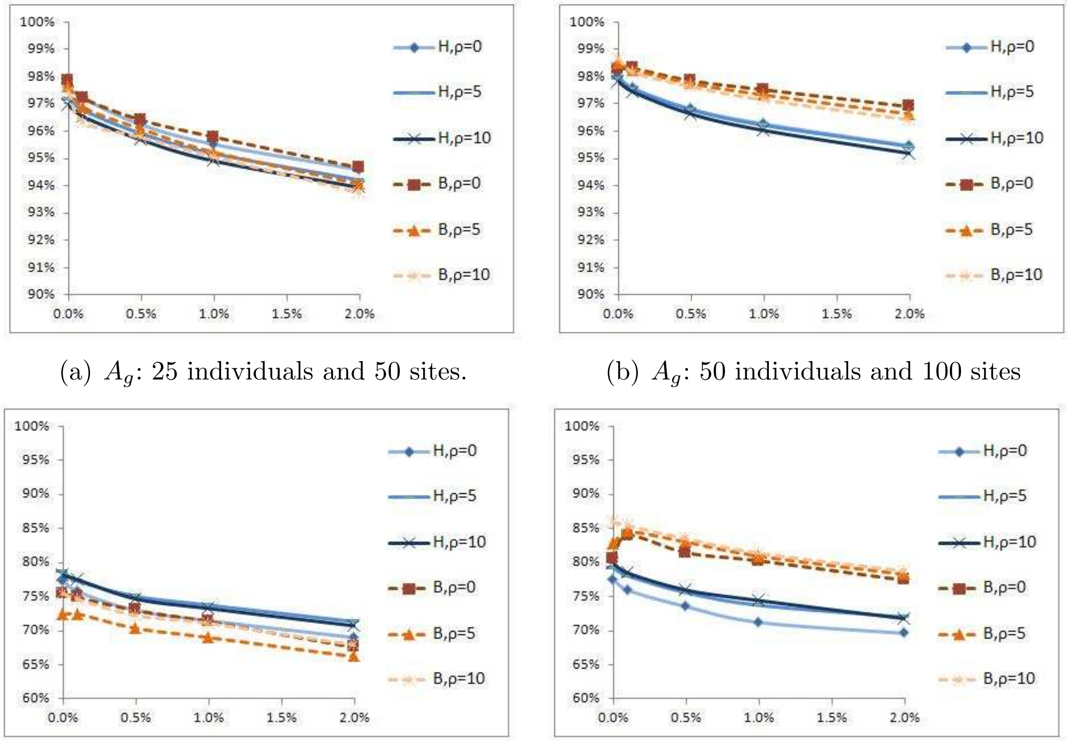
better than two-stage approaches.

Funding and Acknowledgment

The research is supported by grants from U.S. National Science Foundation (IIS-0803440, IIS-0916948 and IIS-0953563).

References

1. S. Levy *et al.*, *PLoS Biology* **5**, e254+ (2007).
2. D. Wheeler *et al.*, *Nature* **452**, 872 (2008).
3. The 1000 genomes project consortium http://www.1000genomes.org/.
4. Pacific Biosciences http://www.pacificbiosciences.com/index.php?q=home.
5. International HapMap Consortium, *Nature* **426**, 789 (2003).
6. International HapMap Consortium, *Nature* **449**, p. 851861 (2007).
7. V. Bansal and V. Bafna, *Bioinformatics* **24**, 153 (2008).
8. D. He, A. Choi, K. Pipatsrisawat, A. Darwiche and E. Eskin, *Bioinformatics* **26**, i183 (2010).
9. T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll and P. M. Visscher, *Nature* **461**, 747 (2009).



(a) A_g: 25 individuals and 50 sites.

(b) A_g: 50 individuals and 100 sites

(c) A_p: 25 individuals and 50 sites.

(d) A_p: 50 individuals and 100 sites

Fig. 6. Accuracy of heuristic approach (solid lines) and Beagle (dashed lines) under different reads error rates. H refers to Heuristic approach and B refers to Beagle. ρ is recombination rates.

10. G. A. Watterson, *Theoretical Population Biology* **7**, 256 (1975).
11. D. Gusfield, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology* (Cambridge University Press, Cambridge, UK, 1997).
12. M. Stephens, N. Smith and P. Donnelly, *Am. J. Human Genetics* **68**, 978 (2001).
13. D. Gusfield, Haplotyping as Perfect Phylogeny: Conceptual Framework and Efficient Solutions (Extended Abstract), in *Proceedings of RECOMB 2002: The Sixth Annual International Conference on Computational Biology*, 2002.
14. E. Halperin and E. Eskin, *Bioinformatics* **20**, 1842 (2004).
15. P. Scheet and M. Stephens, *Am. J. Human Genetics* **78**, 629 (2006).
16. B. L. Browning and Z. Yu, *American Journal of Human Genetics* **85**, 847 (2009).
17. J. Duitama, J. Kennedy, S. Dinakar, Y. Hernández, Y. Wu, and I. Măndoiu, Linkage Disequilibrium Based Genotype Calling from Low-Coverage Shotgun Sequencing Reads, manuscript.
18. Y. S. Song, Y. Wu and D. Gusfield, *Bioinformatics* **21**, i413 (2005), Bioinformatics Suppl. 1, Proceedings of ISMB 2005.
19. Y. Wu, New methods for inference of local tree topologies with recombinant snp sequences in populations (2010), IEEE/ACM Trans. of Comput. Biol. and Bioinfo., in press.
20. R. Hudson, *Bioinformatics* **18**, 337 (2002).
21. S. Lin, D. Cutler, M. Zwick and A. Chakravarti, *Am. J. of Hum. Genet.* **71**, 1129 (2002).

REVERSE ENGINEERING AND SYNTHESIS OF BIOMOLECULAR SYSTEMS

GIL ALTEROVITZ

Division of Health Sciences and Technology, Harvard University/Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Children's Hospital Informatics Program, Boston, MA 02115, USA. Department of Electrical Engineering and Computer Science, Cambridge, MA 02139, USA. Partners Healthcare Center for Personalized Genetic Medicine, Harvard Medical School, Boston, MA 02115, USA. gil@mit.edu

SILVIO CAVALCANTI

Department of Bioengineering, University of Bologna, Bologna, Italy.

TARO M. MUSO

Partners Healthcare Center for Personalized Genetic Medicine, Harvard Medical School, Boston, MA 02115, USA.

MARCO F. RAMONI*

Children's Hospital Informatics Program, Boston, MA 02115, USA. Division of Health Sciences and Technology, Harvard University/Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Partners Healthcare Center for Personalized Genetic Medicine, Harvard Medical School, Boston, MA 02115, USA.

MAY WANG

Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

1. Introduction

Synthetic biology is the new frontier of biological engineering. Instead of incrementally altering living organisms, synthetic biologists propose to use biological knowledge, modular biological parts, and computer-aided design to quickly develop systems capable of unprecedented biochemical feats. Synthetic biology therefore promises dramatic improvements in green chemistry [1], alternative energy [2], drug manufacture [3,4], and therapeutirs [5].

There have been numerous recent advancements in synthetic biology. The need for accuracy at the design and simulation stage have inspired dialogue on how to add functional characterizations to parts documentation in the Registry of Standard Biological parts [6,7]. In addition, a design strategy -- constructing networks from quantitatively characterized libraries of diversified components -- has been proposed [8]. A synthetic network must be integrated into an engineering chassis. To this end the development of evolved ribosome-mRNA pairs may be the first step towards an orthogonal cellular network [9] [10] [11] [12].

Although scientists have made significant progress in synthetic biology, the field must still overcome a number of challenges. To this end, this session offers novel methodologies in three general areas: namely, in designing synthetic systems, in developing novel biological parts, and in analyzing complex networks.

* Deceased

2. Session Papers

Design principles and development strategies from other engineering disciplines must be adjusted to the peculiarities of biological systems. **Kharam *et al*.** propose a rate independent scheme to implement binary counting using chemical reactions. **McDermott *et al*.** develop enhanced network models to determine biological dependencies that help predict behavior of a system. **Senum *et al*.** present a collection of computational modules implemented with chemical reactions, independent of exact reaction rates. **Uhlendorf *et al*.** have proposed and developed a system towards in vivo control of gene expression using an experimental platform combining micro-fluidic device, an epi-flouresence microscope and software approaches. **Verdicchio *et al*.** have demonstrated how logic minimization of the collections of state in Boolean network basins of attraction can help identify targets for intervention.

Acknowledgments

Thank you to all the authors who submitted their work to this session and to the reviewers who graciously contributed their time.

References

1. Marguet, P., Balagadde, F., Tan, C. & You, L., *J R Soc Interface* **4**, 607-23 (2007).
2. Lee, S.K., Chou, H., Ham, T.S., Lee, T.S. & Keasling, J.D., *Current Opinion in Biotechnology* **19**, 556-563 (2008).
3. Chang, M.C.Y. & Keasling, J.D., *Nat Chem Biol* **2**, 674-681 (2006).
4. Weber, W., Schoenmakers, R., Keller, B., Gitzinger, M., Grau, T. *et al.*, *Proceedings of the National Academy of Sciences* **105**, 9994-9998 (2008).
5. Lu, T.K. & Collins, J.J., *Proceedings of the National Academy of Sciences* **106**, 4629-4634 (2009).
6. Canton, B., Labno, A. & Endy, D., *Nat Biotech* **26**, 787-793 (2008).
7. Purnick, P.E.M. & Weiss, R., *Nat Rev Mol Cell Biol* **10**, 410-422 (2009).
8. Ellis, T., Wang, X. & Collins, J.J., *Nat Biotech* **27**, 465-471 (2009).
9. Rackham, O. & Chin, J.W., *Biochem Soc Trans* **34**, 328-9 (2006).
10. Rackham, O. & Chin, J.W., *Nat Chem Biol* **1**, 159-166 (2005).
11. An, W. & Chin, J.W., *Proceedings of the National Academy of Sciences* **106**, 8477-8482 (2009).
12. Filipovska, A. & Rackham, O., *ACS Chemical Biology* **3**, 51-63 (2008).

BINARY COUNTING WITH CHEMICAL REACTIONS*

ALEKSANDRA KHARAM, HUA JIANG, MARC RIEDEL, and KESHAB PARHI

Electrical and Computer Engineering, University of Minnesota
Minneapolis, MN 55455
`http://cctbio.ece.umn.edu`
E-mail: {`veden002, hua, mriedel, parhi`}`@umn.edu`

This paper describes a scheme for implementing a binary counter with chemical reactions. The value of the counter is encoded by logical values of "0" and "1" that correspond to the absence and presence of specific molecular types, respectively. It is incremented when molecules of a trigger type are injected. Synchronization is achieved with reactions that produce a sustained three-phase oscillation. This oscillation plays a role analogous to a clock signal in digital electronics. Quantities are transferred between molecular types in different phases of the oscillation. Unlike all previous schemes for chemical computation, this scheme is dependent only on coarse rate categories for the reactions ("fast" and "slow"). Given such categories, the computation is exact and independent of the specific reaction rates. Although conceptual for the time being, the methodology has potential applications in domains of synthetic biology such as biochemical sensing and drug delivery. We are exploring DNA-based computation via strand displacement as a possible experimental chassis.

1. Introduction

In the nascent field of synthetic biology, researchers are striving to create biological systems with functionality not seen in nature. The field aims to apply engineering methods to biology in a deliberate way. Beyond engineering ends, such methods also provide a constructive means to validating new science. Understanding is achieved by constructing and testing simplified systems from the bottom up, teasing out and nailing down fundamental principles in the process.[1]

We bring a particular mindset to tackle the problem of synthesizing new biological functions. We tackle synthesis at a *conceptual* level, working with abstract molecular types. Working at this level, we implement *computational* constructs, that is to say, chemical reaction networks that compute specific outputs as a function of inputs. Then we map the conceptual designs onto specific chemical substrates.

We model the chemical dynamics in terms of mass-action kinetics:[2,3] reaction rates are proportional to (1) the quantities of the participating molecular types; and (2) reaction constants. We aim for robust constructs: systems that compute exact results independently of specific reaction constants. All of our designs are formulated in terms of two coarse rate categories (e.g., "fast" and "slow"). Given such categories, the computation is exact and independent of the specific reaction rates.

The analogy for this approach is the design flow for digital electronics, where different designs are systematically explored at a *technology-independent* level, in terms of Boolean functions. Once the best design is found, it is mapped to specific technology libraries in silicon.[4] An overarching goal of the digital paradigm is robustness: digital electronics delivers voltage values that correspond to zero and one reliably, in spite of fluctuations in the signals.

*This work is supported by an NSF EAGER Grant, #CCF-0946601, and by an NSF CAREER Award, #0845650.

In our prior and related work, we have described a variety of computational constructs for chemical reaction networks: logical operations such as copying, comparing and incrementing/decrementing;[5] programming constructs such as "for" and "while" loops;[6] arithmetic operations such as multiplication, exponentiation and logarithms;[5,6] and signal processing operations such as filtering.[7]

In this paper, we describe a scheme for implementing a binary counter with chemical reactions. The value of the counter is encoded by logical values of "0" and "1" that correspond to the absence and presence of specific molecular types, respectively. It is incremented by one every time molecules of a trigger type are injected. Synchronization is achieved with reactions that produce a sustained three-phase oscillation. This oscillation plays a role analogous to a clock signal in digital electronics. Quantities are transferred between molecular types in different phases of the oscillation.

This paper is organized as follows. In Section 2, we summarize the main principles and the basic algorithm for our implementation of the binary counter. In Section 3, we introduce some specific concepts that we use, namely the concepts of "prereactants" and "absence indicators." We also introduce the essential synchronization mechanism that we use, a three-phase oscillation that we call "red-green-blue" (RGB). Then we present the design of the molecular counter. In Section 4, we present simulation results obtained with an ordinary differential equations (ODE) solver. Finally, in Section 5, we discuss DNA strand-displacement reactions as a possible experimental chassis for our method.[8]

2. Counting in Binary

We first review some of the algorithmic principles of counting in binary. Then we present an intuitive description of our approach to implementing a molecular binary counter.

Z	Y	X
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1
⋮	⋮	⋮

Fig. 1. Sequence of values in a three-bit binary counter.

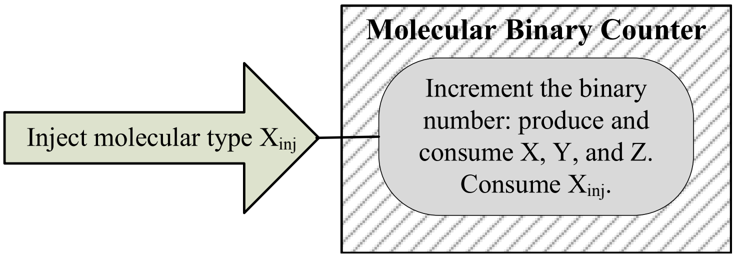

Molecular Binary Counter
Inject molecular type X_{inj}
Increment the binary number: produce and consume X, Y, and Z. Consume X_{inj}.

Fig. 2. Basic functionality of the molecular counter.

2.1. *General Principles*

Figure 1 lists the binary numbers that a 3-bit binary counter cycles through, starting at “000” and ending at “111.”

(1) *Every time the binary count is incremented, the least significant (i.e., right-most) bit is flipped.* For instance, in the sequence $00\underline{0} \rightarrow 00\underline{1} \rightarrow 01\underline{0} \rightarrow 01\underline{1} \rightarrow 10\underline{0} \rightarrow 10\underline{1}$, note that the least significant bit (underlined) alternates: $0 \rightarrow 1 \rightarrow 0 \rightarrow 1 \rightarrow 0 \rightarrow 1$.

(2) *Every time the binary count is incremented, exactly one bit changes from “0” to “1”. (However, several bits may change from “1” to “0.”)* For instance, in the sequence $00\underline{0} \rightarrow 0\underline{0}1 \rightarrow 01\underline{0} \rightarrow \underline{0}11 \rightarrow 10\underline{0} \rightarrow 1\underline{0}1$, the bits that change from “0” to “1” are underlined. Note that there is exactly one such bit each time. (As will be discussed in Section 3, this principle is important for synchronizing our molecular counter.)

(3) *When the binary count is incremented, a given bit changes from “0” to “1” only if all bits of lesser significance (i.e., all bits to the right of it) are “1.”* For instance, in the sequence $000 \rightarrow 00\underline{1} \rightarrow 01\underline{0}$, the second bit changes from “0” to “1” when the first bit is “1.” In the sequence $0\underline{11} \rightarrow 1\underline{00} \rightarrow 101$ the third bit changes from “0” to “1” when the first and second bits are “1.”

2.2. *Towards a Molecular Binary Counter*

Throughout this paper, the exposition will be in terms of a three-bit binary counter. The ideas can readily be generalized to an n-bit counter. We encode the binary values of “0” and “1” by the presence or absence of specific molecular types, respectively. For the binary sequence in Figure 1, we use the types X, Y and Z. (We will call these “bit types.”) For instance, if types X and Z are present, while type Y is absent, the corresponding binary number is “101”.

Figure 2 shows the basic functionality of our molecular counter. Every time we want to increment it, we inject some amount of a “trigger” type X_{inj}. The system consumes X_{inj} and increments the binary value specified by the quantities of X, Y and Z. Once all the molecules of X_{inj} have been consumed, the counter can be incremented again.

Tables 2 and 3 specify the set of chemical reactions for our three-bit counter. In order to elucidate the final design, we will provide a succession of design refinements:

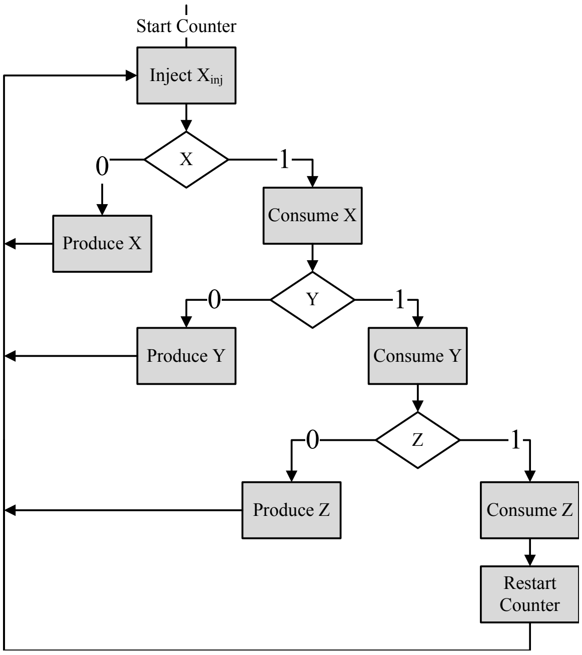


Fig. 3. Basic algorithm for the molecular counter.

(1) We start with a simple intuitive set of reactions, ignoring issues such as synchronization (Section 2.3).

(2) We introduce two specific concepts that we use to implement the counter: the concept of "pre-reactants" and that of "absence indicators" (Section 3.1).

(3) We introduce our synchronization mechanism: a three-phase chemical oscillation that we call "red-green-blue" (RGB) (Section 3.2).

(4) Finally, we provide the full design of the counter, consisting of 24 chemical reactions (Sections 3.3).

2.3. *Intuitive Model*

A molecular counter cannot directly set bits to "0" or to "1"; rather the functionality must be achieved by reactions that produce and consume the molecular types corresponding to these bits. Call the three bits of the counter, the *high*, *middle* and *low* bits, encoded by the presence/absence of types Z, Y and X, respectively. The low bit is set to "1" by producing molecules of X whenever the type X_{inj} is injected into the system:

$$X_{\text{inj}} \rightarrow X. \tag{1}$$

The middle bit is set to "1" by producing molecules of Y whenever the type X is present:

$$X \rightarrow Y. \tag{2}$$

The high bit is set to "1" by producing molecules of Z whenever both types X and Y present:

$$X + Y \rightarrow Z. \tag{3}$$

Note that, in each of these reactions, the system consume molecules of X, Y and Z, reseting the corresponding bits to "0." When molecules of all three types X, Y and Z are present, the corresponding binary number is "111". The counter is reset:

$$X + Y + Z \rightarrow \varnothing. \tag{4}$$

(The symbol $\varnothing$ as a product indicates "nothing", meaning that the type degrades into products that are no longer tracked or used.)

A flowchart for the algorithm that we use is given in Figure 3. In the figure, decisions to produce and consume molecular types are made according to the presence and absence of types. (As we refine the design, we will have to implement these "decisions" through chemical reactions.) Let us assume the current binary number is set to "101". This number corresponds to the absence of Y and the presence of X and Z. Suppose that we inject the trigger type X_{inj}; we move to the first decision box. Since X is present, we do not produce more of it. We consume molecules of X and move to the next decision box. Here we check for the presence of type Y. Since Y is absent, we move to the left and produce molecules of Y. With the absence of X and the presence of Y and Z, the binary number has changed to "110". Next we return to "idle state," waiting for the next injection.

3. Synchronization

The challenge in setting up the molecular counter is that all the chemical reactions fire asynchronously. Each reaction starts producing its products as soon as its reactants are available. If these products participate as reactants in other reactions, then they immediately start getting consumed.

Accordingly, with Reactions 1–4, we will not get a binary counter, encoded by the presence and absence of X, Y and Z. Rather, we will get a jumble of all of these. In particular, note that with Reaction 2, Y is produced from X. As soon as molecules of Y are available, Reaction 3 starts consuming molecules of X and Y to produce molecules of Z. This contradicts the second principle described in Section 2.1: we should only change one bit from "0" to "1" in each increment operation. To mitigate against this issue, we introduce additional molecular types called "prereactants." We also introduce "absence indicator" types to coordinate the transfer between prereactants and reactants.

3.1. *Prereactants and Absence Indicators*

We use the following notation to describe these concepts. For each bit i of the counter,

(1) Q_i is a **bit type** corresponding to i^{th} bit. (For our three-bit molecular counter, we have $Q_1 = X$, $Q_2 = Y$ and $Q_3 = Z$.)
(2) a_{qi} is an **absence indicator** type for type Q_i. (For our three-bit molecular counter, we have $a_{q1} = a_x$, $a_{q2} = a_y$ and $a_{q3} = a_z$.)
(3) Q_{pi} is a **prereactant** type for Q_i. (For our three-bit molecular counter, we have $Q_{p1} = X_p$, $Q_{p2} = Y_p$ and $Q_{p3} = Z_p$.)
(4) We set $X_p = X_{\text{inj}}$: the trigger type is the first prereactant.

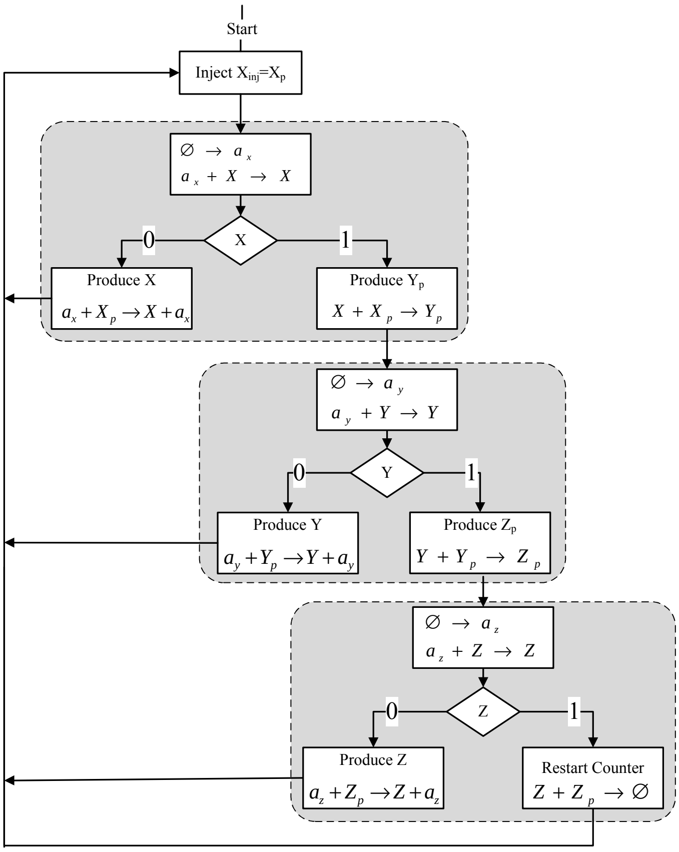


Fig. 4. Modified algorithm for the molecular counter, with prereactants and absence indicators.

All the absence indicators a_{qi} are produced continuously at the slow rate:

$$\varnothing \xrightarrow{\text{slow}} a_{qi}; \tag{5}$$

Here the symbol $\varnothing$ as a reactant indicates that the reaction does not alter the quantity of the reactant types, perhaps because the quantity of these is large or replenishable. If Q_i is present, then its absence indicator a_{qi} is destroyed at the fast rate:

$$a_{qi} + Q_i \xrightarrow{\text{fast}} Q_i. \tag{6}$$

However, if Q_i is absent, then a_{qi} persists, so its presence indicates the absence of Q_i, as required. If both a prereactant Q_{pi} and the absence indicator a_{qi} for the i-th bit are present, we produce type Q_i

at the fast rate:

$$a_{qi} + Q_{pi} \xrightarrow{\text{fast}} Q_i + a_{qi}. \tag{7}$$

Finally, the prereactant $Q_{p(i+1)}$ for the $(i+1)$-st bit is produced at the fast rate if both the prereactant Q_{pi} and the type Q_i for the i-th bit are present:

$$Q_i + Q_{pi} \xrightarrow{\text{fast}} Q_{p(i+1)}. \tag{8}$$

Table 1 lists the corresponding reactions for our three-bit counter in terms of the bit types X, Y and Z instead of generic Q_i's. Figure 4 shows a modified version of the flowchart in Figure 3, this time with prereactants and absence indicators.

Table 1. Reactions for the molecular counter, with prereactants and absence indicators.

#	Q_i	Z	Y	X
1	$\varnothing \xrightarrow{\text{slow}} a_{qi}$ $a_{qi} + Q_i \xrightarrow{\text{fast}} Q_i$	$\varnothing \xrightarrow{\text{slow}} a_z$ $a_z + Z \xrightarrow{\text{fast}} Z$	$\varnothing \xrightarrow{\text{slow}} a_y$ $a_y + Y \xrightarrow{\text{fast}} Y$	$\varnothing \xrightarrow{\text{slow}} a_x$ $a_x + X \xrightarrow{\text{fast}} X$
2	$a_{qi} + Q_{pi} \xrightarrow{\text{fast}} Q_i + a_{qi}$	$a_z + Z_p \xrightarrow{\text{fast}} Z + a_z$	$a_y + Y_p \xrightarrow{\text{fast}} Y + a_y$	$a_x + X_p \xrightarrow{\text{fast}} X + a_x$
3	$Q_i + Q_{pi} \xrightarrow{\text{fast}} Q_{p(i+1)}$	$Z + Z_p \xrightarrow{\text{fast}} \varnothing$	$Y + Y_p \xrightarrow{\text{fast}} Z_p$	$X + X_p \xrightarrow{\text{fast}} Y_p$

3.2. *Three-Phase Synchronization*

Including absence indicators and prereactants establishes an order for the transfers of molecular quantities in the counter, but we need a mechanism to ensure that each transfer completes before the next one begins. As indicated on the left-hand side of Figure 5, we must ensure that the accumulation or destruction of the absence indicator completes before the production of the bit type begins; in turn, we must ensure that the production of the bit type completes before the production of the next prereactant begins, and so on. Similarly, as indicated on the right-hand side of Figure 5, we must ensure that the bit types X, Y and Z are not produced simultaneously. We must turn the two "dials" shown in Figure 5 simultaneously. To do so, we introduced a synchronization mechanism based on sustained *chemical oscillation*.

Chemical oscillations, such as those produced by Belousov–Zhabotinsky (BZ) system, have been widely studied by the chemical engineering community.[9] For our purposes, we require an oscillator with a specific property: it must have three symmetric phases for synchronizing both of the "dials" in Figure 5. To this end, we have developed a scheme for chemical oscillation that we call "Red-Green-Blue" (RGB). A detailed analysis of the scheme is given in related work.[7] Here we give only a cursory description of it.

Like the BZ system, our scheme is a perfect oscillator, producing sustained oscillations for a wide range of reaction rates. The scheme is illustrated in Figure 6. Reactions are "color coded" – that is to say assigned to one of the three categories. Quantities are transferred between color categories based

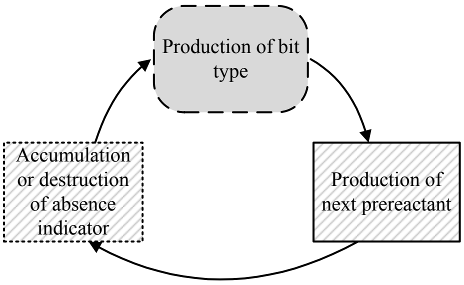
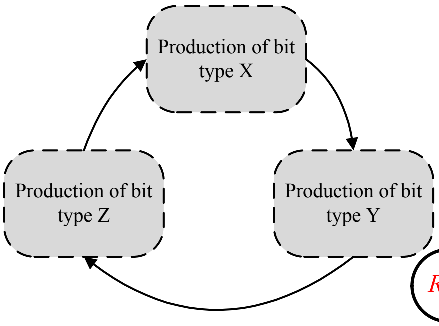

Production of bit type
Production of next prereactant
Accumulation or destruction of absence indicator

Fig. 5. Sequence of reactions for the molecular counter.



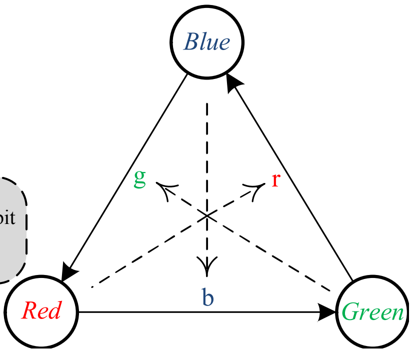


Fig. 6. The three-phase transfer scheme.

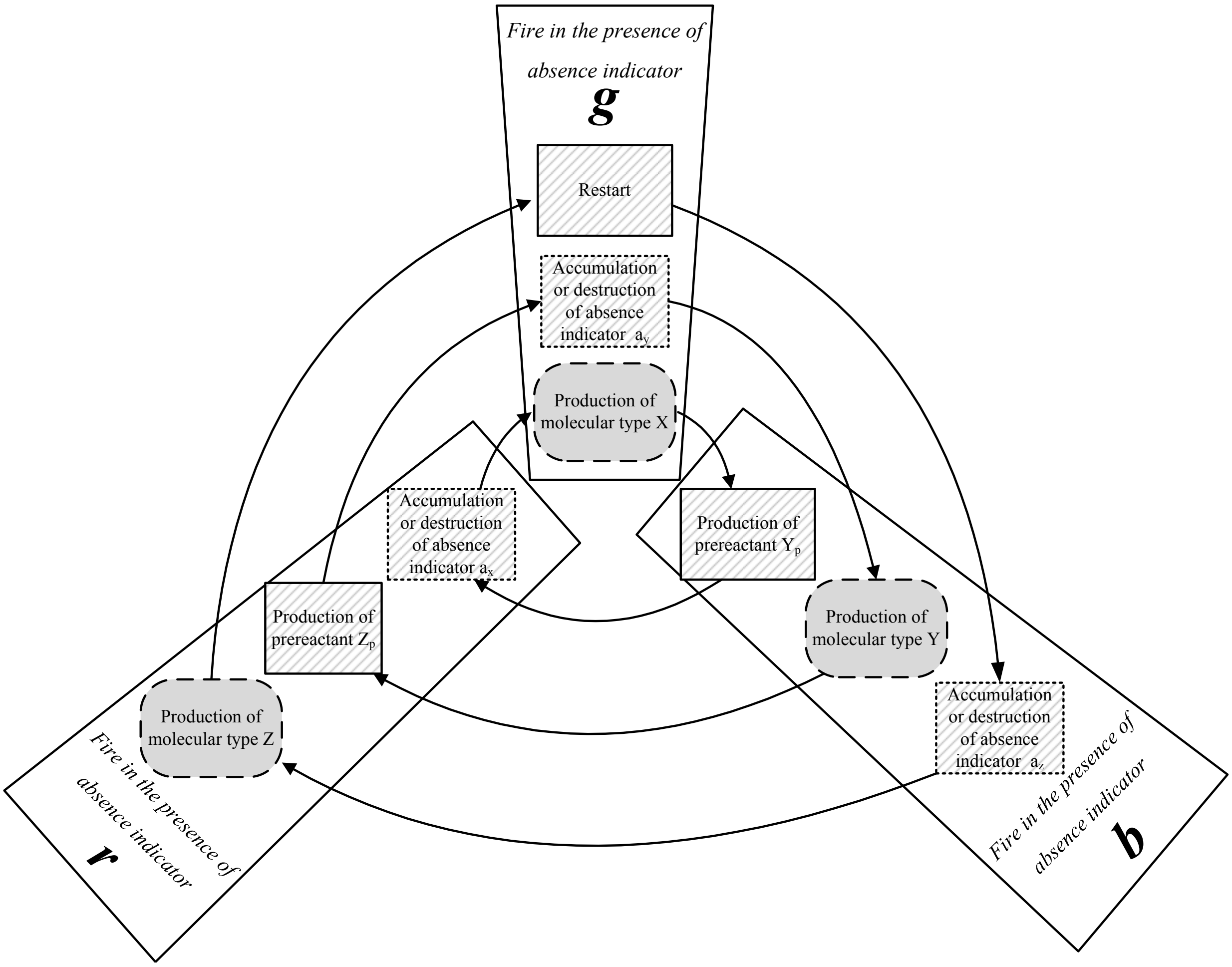

Fire in the presence of absence indicator
g
Restart
Accumulation or destruction of absence indicator a_y
Production of molecular type X
Accumulation or destruction of absence indicator a_x
Production of prereactant Y_p
Production of prereactant Z_p
Production of molecular type Y
Production of molecular type Z
Accumulation or destruction of absence indicator a_z
Fire in the presence of absence indicator
r
Fire in the presence of absence indicator
b

Fig. 7. Combined diagrams for synchronization of reactions.

on the absence of types in the third category: red goes to green in the absence of blue; green goes to blue in the absence of red; and blue goes to red in the absence of green. We introduce molecular types R, G and B. Computation cycles are implemented by transferring quantities among three types R, G and B, with following reactions:

$$\mathbf{b} + R \xrightarrow{\text{slow}} G + \mathbf{b} \quad (9) \qquad \mathbf{r} + G \xrightarrow{\text{slow}} B + \mathbf{r} \quad (10) \qquad \mathbf{g} + B \xrightarrow{\text{slow}} R + \mathbf{g} \quad (11)$$

We generate "absence indicators" types **r**, **g** and **b** corresponding to R, G and B:

$$\begin{array}{r}\varnothing \xrightarrow{\text{slow}} \mathbf{r}\\ R + \mathbf{r} \xrightarrow{\text{fast}} R\end{array} \quad (12) \qquad \begin{array}{r}\varnothing \xrightarrow{\text{slow}} \mathbf{g}\\ G + \mathbf{g} \xrightarrow{\text{fast}} G\end{array} \quad (13) \qquad \begin{array}{r}\varnothing \xrightarrow{\text{slow}} \mathbf{b}\\ B + \mathbf{b} \xrightarrow{\text{fast}} B\end{array} \quad (14)$$

The absence indicators are continually generated. However, they only persist in the absence of the corresponding color-coded signals, since they are quickly consumed by signal molecules in their corresponding color categories. This feature assures that as long as any reaction in a given phase has not fired to completion, the succeeding phase cannot begin. We also include reactions that accelerate and isolate the transfers in each phase. For instance, in Reaction 15 two molecules of G combine with one molecule of R to produce three molecules of G. The transfer will occur at a higher rate. Simulation results illustrating the RGB oscillation are shown in Figure 8. In the next section, we incorporate this scheme to synchronize the molecular counter, using RGB in a way analogous to a clock signal in digital electronics.

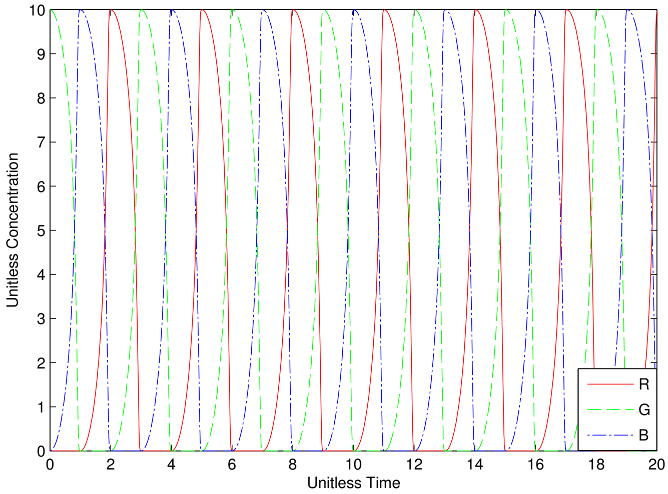


Fig. 8. Simulation results for RGB oscillation.

$$R + 2G \xrightarrow{\text{slow}} 3G \quad (15) \qquad G + 2B \xrightarrow{\text{slow}} 3B \quad (16) \qquad B + 2R \xrightarrow{\text{slow}} 3R \quad (17)$$

3.3. *The Molecular Binary Counter with RGB scheme*

Figure 7 shows the assignment operations to phases of the computation. Absence indicators **r**, **g** and **b** are used to initiate reactions in each phase. In lieu of the generic transfer reactions 9– 11, we use transfer reactions that produce the absence indicators a_x, a_y and a_z for X, Y and Z, respectively:

$$\mathbf{r} + G \xrightarrow{\text{slow}} B + a_x + \mathbf{r} \quad (18) \qquad \mathbf{g} + B \xrightarrow{\text{slow}} R + a_y + \mathbf{g} \quad (19) \qquad \mathbf{b} + R \xrightarrow{\text{slow}} G + a_z + \mathbf{b} \quad (20)$$

This obviates the need for reactions of the form of Reaction 5 to generate a_x, a_y and a_z.

A set of reactions for the counter that incorporates the RGB transfer reactions is described in Figure 9. This is nearly the final design. However, we need a few more reactions to deal with accumulation of unused absence indicators. The transfer reactions 18–20 supply absence indicators a_x, a_y and a_z in every RGB cycle. The scheme cycles continuously, irrespective of injections of X_{inj}. Accordingly, unused absence indicators a_x, a_y and a_z will accumulate. To mitigate against this, we include "**clean-up**" reactions initiated in the presence of corresponding absence indicators **r**, **g** and **b**:

$$\mathbf{b} + a_x \xrightarrow{\text{fast}} \mathbf{b} \quad (21) \qquad \mathbf{r} + a_y \xrightarrow{\text{fast}} \mathbf{r} \quad (22) \qquad \mathbf{g} + a_z \xrightarrow{\text{fast}} \mathbf{g} \quad (23)$$

As shown in Figure 9 the corresponding clean-up reactions always complete before the production of a_x, a_y and a_z begins. For instance, the absence indicator a_z is pushed into the system by Reaction 20 whenever the absence indicator **b** is present. Therefore, the clean-up Reaction 23 for a_z fires in the preceding RGB phase that was initiated in the presence of absence indicator **g**.

Table 2 shows the final set of RGB reactions and Table 3 shows the final set of reactions for X, Y and Z. Together, these comprise our complete design of the molecular counter.

Table 2. Final version of RGB reactions for the molecular counter.

Production of r, g, b	Destruction of r, g, b	Transfer reactions	Speed-up reactions	Clean-up reactions
$\varnothing \xrightarrow{\text{slow}} \mathbf{r}$	$R + \mathbf{r} \xrightarrow{\text{fast}} R$	$\mathbf{b} + R \xrightarrow{\text{slow}} G + a_z + \mathbf{b}$	$R + 2G \xrightarrow{\text{slow}} 3G$	$\mathbf{b} + a_x \xrightarrow{\text{fast}} \mathbf{b}$
$\varnothing \xrightarrow{\text{slow}} \mathbf{g}$	$G + \mathbf{g} \xrightarrow{\text{fast}} G$	$\mathbf{r} + G \xrightarrow{\text{slow}} B + a_x + \mathbf{r}$	$G + 2B \xrightarrow{\text{slow}} 3B$	$\mathbf{r} + a_y \xrightarrow{\text{fast}} \mathbf{r}$
$\varnothing \xrightarrow{\text{slow}} \mathbf{b}$	$B + \mathbf{b} \xrightarrow{\text{fast}} B$	$\mathbf{g} + B \xrightarrow{\text{slow}} R + a_y + \mathbf{g}$	$B + 2R \xrightarrow{\text{slow}} 3R$	$\mathbf{g} + a_z \xrightarrow{\text{fast}} \mathbf{g}$

Table 3. Final version of reactions for molecular types X, Y and Z.

Accumulation or destruction of absence indicators	Production of molecules	Production of prereactant
$\mathbf{r} + a_x + X \xrightarrow{\text{fast}} X + \mathbf{r}$	$\mathbf{g} + a_x + X_p \xrightarrow{\text{fast}} X + a_x + \mathbf{g}$	$\mathbf{b} + X + X_p \xrightarrow{\text{fast}} Y_p + \mathbf{b}$
$\mathbf{g} + a_y + Y \xrightarrow{\text{fast}} Y + \mathbf{g}$	$\mathbf{b} + a_y + Y_p \xrightarrow{\text{fast}} Y + a_y + \mathbf{b}$	$\mathbf{r} + Y + Y_p \xrightarrow{\text{fast}} Z_p + \mathbf{r}$
$\mathbf{b} + a_z + Z \xrightarrow{\text{fast}} Z + \mathbf{b}$	$\mathbf{r} + a_z + Z_p \xrightarrow{\text{fast}} Z + a_z + \mathbf{r}$	$\mathbf{g} + Z + Z_p \xrightarrow{\text{fast}} \mathbf{g}$

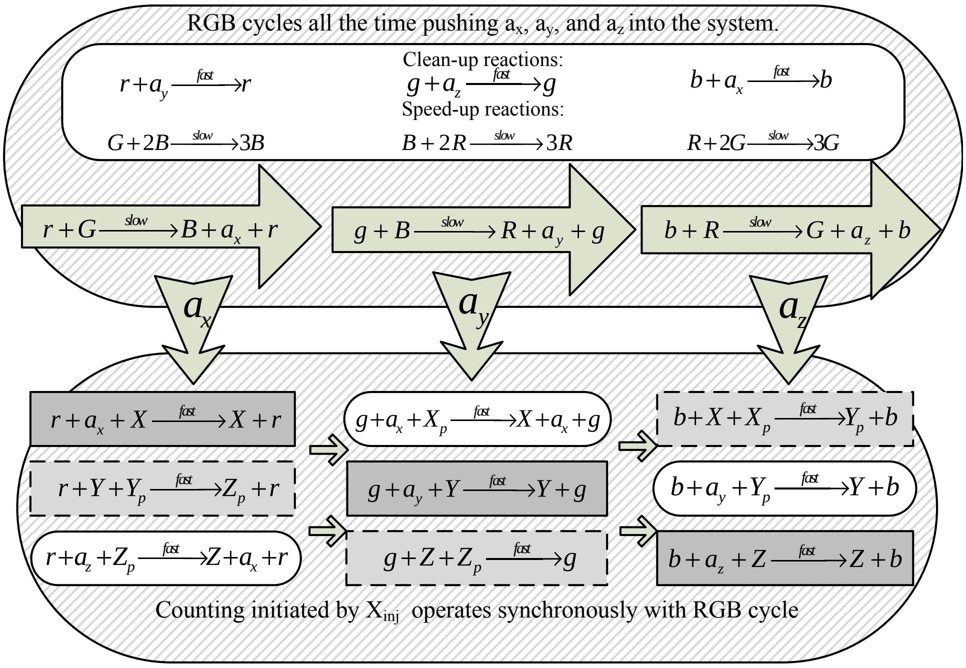

RGB cycles all the time pushing a$_x$, a$_y$, and a$_z$ into the system.
Clean-up reactions:
$r+a_y \xrightarrow{fast} r$
$g+a_z \xrightarrow{fast} g$
$b+a_x \xrightarrow{fast} b$
Speed-up reactions:
$G+2B \xrightarrow{slow} 3B$
$B+2R \xrightarrow{slow} 3R$
$R+2G \xrightarrow{slow} 3G$
$r+G \xrightarrow{slow} B+a_x+r$
$g+B \xrightarrow{slow} R+a_y+g$
$b+R \xrightarrow{slow} G+a_z+b$
a_x
a_y
a_z
$r+a_x+X \xrightarrow{fast} X+r$
$g+a_x+X_p \xrightarrow{fast} X+a_x+g$
$b+X+X_p \xrightarrow{fast} Y_p+b$
$r+Y+Y_p \xrightarrow{fast} Z_p+r$
$g+a_y+Y \xrightarrow{fast} Y+g$
$b+a_y+Y_p \xrightarrow{fast} Y+b$
$r+a_z+Z_p \xrightarrow{fast} Z+a_x+r$
$g+Z+Z_p \xrightarrow{fast} g$
$b+a_z+Z \xrightarrow{fast} Z+b$
Counting initiated by X$_{inj}$ operates synchronously with RGB cycle

Fig. 9. Diagram for the molecular counter with RGB synchronization.

4. Simulation results

As we discuss in Section 5, we are targeting DNA strand displacement as a potential experimental chassis for our molecular counter.[8] Accordingly, the constituent chemical reactions must all be either uni- or bimolecular reactions. Thus, we split all trimoleclar reactions of the form

$$R_1 + R_2 + R_3 \xrightarrow{\mathrm{k}_1} R_4 + \cdots \tag{24}$$

in Table 2 and Table 3 into the sequence of bimolecular reactions

$$\begin{aligned} R_1 + R_2 &\underset{\mathrm{k}_2}{\overset{\mathrm{k}_1}{\rightleftharpoons}} I \\ I + R_3 &\xrightarrow{\mathrm{k}_2} R_4 + \cdots . \end{aligned} \tag{25}$$

The first step in this process is reversible: two molecules R_1 and R_2 can combine at a rate k_1, but in the absence of any molecules R_3, the combined form will dissociate back into molecules R_1 and R_2 at a rate k_2 which is greater than k_1. In the presence of R_3, the sequence of reactions will proceed, producing $R_4 + \cdots$. The overall rate of reactions is determined by the slowest reaction and therefore set by k_1.

With such transformations into uni- and bimolecular reactions, we simulate the chemical kinetics of our molecular counter with an ordinary differential equation (ODE) solver. We chose the parameters, the concentration values and reaction rates as follows. The concentrations are unitless; for an experimental setup, these would be scaled appropriately. The initial concentration of our trigger type X_p was set to 0.05. (Recall that we use X_p as the trigger type, so $X_{\text{inj}} = X_p$.) The initial concentration of G was set to a value much greater than that of X_p, namely 10. The rates of all the "slow" reactions were set to unity. The rates of all the "fast" reactions were set five orders of magnitude higher.

Figure 10 shows the change in concentration X, Y and Z for 20 injections. We observed a stable behavior of the molecular binary counter for 40 injections. The data from the simulation for the first 8 injections is shown in Table 4. We see exactly the behavior that we expect for a binary counter.

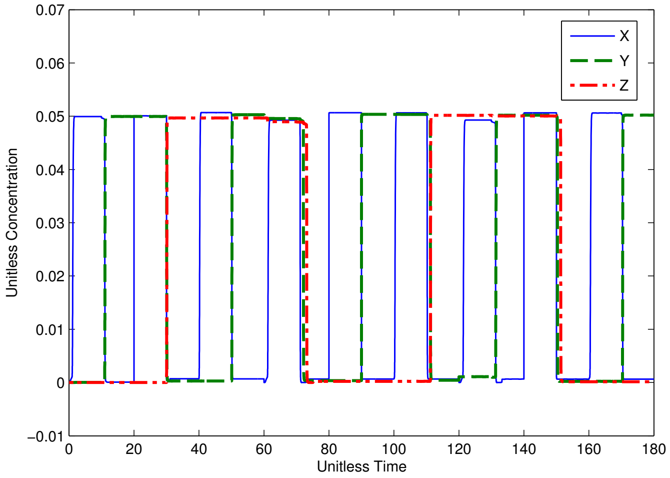


Fig. 10. Simulation results for molecular types X, Y and Z for 20 injections.

- The threshold for logical "1" for the bit types X, Y and Z can be set at 97% of the injected concentration of X_p.
- The threshold for logical "0" for the bit types X, Y and Z can be set at 3% of the injected concentration of X_p.

Table 4. Data from the ODE simulation of the molecular counter for 8 successive increment operations.

# Injection	Binary number	Concentration Z	Concentration Y	Concentration X
0	000	0.0000	0.0000	0.0000
1	00**1**	0.0000	0.0000	**0.0499**
2	0**1**0	0.0000	**0.0500**	0.0001
3	0**11**	0.0000	**0.0500**	**0.0501**
4	**1**00	**0.0497**	0.0003	0.0007
5	**1**0**1**	**0.0497**	0.0003	**0.0507**
6	**11**0	**0.0497**	**0.0503**	0.0007
7	**111**	**0.0490**	**0.0496**	**0.0493**
8	000	0.0002	0.0004	0.0007

5. Discussion

We have demonstrated the design of a molecular counter that is robust and accurate. Given only rate categories of "slow" and "fast, our counter computes exact binary values. It does not matter how fast any "fast" reaction is relative to another, or how slow any "slow" reaction is relative to another – only that "fast" reactions are fast relative to "slow" reactions. Throughout the paper, the exposition was in terms of a three-bit counter. In future work, we will generalize the construction to n bits.

Our contribution is to tackle the problem of synthesizing computation at a conceptual level, working not with actual molecular types but rather with abstract types. In future work, we will demonstrate our binary counter through *in vitro* experiments with DNA. It has been shown that DNA strand displacement reactions can emulate the chemical kinetics of nearly any chemical reaction network. Indeed, in recent work, researchers at Caltech have developed a compiler that translates abstract chemical reactions of the sort that we design into specific DNA reactions.[8]

Recent work has demonstrated both the scale of computation that is possible with DNA-based computing,[10] as well as exciting applications.[11] We comment that our design of a molecular counter could be applied for the task of counting cell divisions. This task is important for the analysis of aging and, perhaps, for the detection of cancer, where cell divisions run rampant. Also, our design might find applications in biochemical sensing and drug delivery.

References

1. D. Endy, "Foundations for engineering biology," *Nature*, vol. 438, pp. 449–453, 2005.
2. F. Horn and R. Jackson, "General mass action kinetics," *Archive for Rational Mechanics and Analysis*, vol. 47, pp. 81–116, 1972.
3. P. Érdi and J. Tóth, *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*. Manchester University Press, 1989.
4. L. Lavagno, G. Martin, and L. Scheffer, *Electronic Design Automation for Integrated Circuits Handbook*. CRC Press, 2006.
5. P. Senum and M. D. Riedel, "Rate-independent biochemical computational modules," in *Proceedings of the Pacific Symposium on Biocomputing*, 2011.
6. A. Shea, B. Fett, M. D. Riedel, and K. Parhi, "Writing and compiling code into biochemistry," in *Proceedings of the Pacific Symposium on Biocomputing*, 2010, pp. 456–464.
7. H. Jiang, A. P. Kharam, M. D. Riedel, and K. K. Parhi, "A synthesis flow for digital signal processing with biomolecular reactions," in *IEEE International Conference on Computer-Aided Design*, 2010.
8. D. Soloveichik, G. Seelig, and E. Winfree, "DNA as a universal substrate for chemical kinetics," *Proceedings of the National Academy of Sciences*, vol. 107, no. 12, pp. 5393–5398, 2010.
9. I. R. Epstein and J. A. Pojman, *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos*. Oxford Univ Press, 1998.
10. L. Qian and E. Winfree, "A simple DNA gate motif for synthesizing large-scale circuits," in *DNA Computing*, 2009, pp. 70–89.
11. S. Venkataramana, R. M. Dirks, C. T. Ueda, and N. A. Pierce, "Selective cell death mediated by small conditional RNAs," *Proceedings of the National Academy of Sciences*, 2010 (in press).

DEFINING THE PLAYERS IN HIGHER-ORDER NETWORKS: PREDICTIVE MODELING FOR REVERSE ENGINEERING FUNCTIONAL INFLUENCE NETWORKS

JASON E. MCDERMOTT[1], MICHELLE ARCHULETA[1], SUSAN L. STEVENS[2], MARY P. STENZEL-POORE[2], AND ANTONIO SANFILIPPO[1]

[1]Pacific Northwest National Laboratory, Richland, WA, USA

[2]Oregon Health & Science University, Portland, OR, USA

Determining biological network dependencies that can help predict the behavior of a system given prior observations from high-throughput data is a very valuable but difficult task, especially in the light of the ever-increasing volume of experimental data. Such an endeavor can be greatly enhanced by considering regulatory influences on co-expressed groups of genes representing functional modules, thus constraining the number of parameters in the system. This allows development of network models that are predictive of system dynamics. We first develop a predictive network model of the transcriptomics of whole blood from a mouse model of neuroprotection in ischemic stroke, and show that it can accurately predict system behavior under novel conditions. We then use a network topology approach to expand the set of regulators considered and show that addition of topological bottlenecks improves the performance of the predictive model. Finally, we explore how improvements in definition of functional modules may be achieved through an integration of inferred network relationships and functional relationships defined using Gene Ontology similarity. We show that appropriate integration of these two types of relationships can result in models with improved performance.

1. Introduction

Stroke is currently the second leading cause of death in the Western world [1] and is estimated to cause 10% of deaths worldwide. Patients who do not die from a stroke suffer from neurological impairment that is significantly disabling in a large percentage of survivors. Preconditioning by induction of a small stroke or treatment with Toll-like receptor (TLR) agonists prior to induction of a large stroke provides a significant degree of neuroprotection in animal models [2]. To provide molecular level understanding of the dynamics of stroke processes we have previously used high-throughput transcriptomic profiling using microarrays to follow the dynamics of stroke and neuroprotection in a mouse model [2, 3]. Predictive models of regulatory and functional processes occurring during neuroprotection and stroke would offer a very powerful tool to investigate novel methods for prevention and treatment of this important disease.

Models that can predict aspects of system behavior from the observation of a small number of system inputs or components have been largely limited to very general models [4], focused models that can be fully parameterized, or models for which there is a large body of existing data about the molecular interactions between components [5]. Inference of specific interactions between large numbers of system components is limited by the number of observations of the system being examined. Specific interactions of interest include protein-protein interactions, interactions between signal transduction pathway members (e.g. phosphorylation events), and transcription factor mediated regulatory events (activation or repression of a gene or set of genes). Even with high-throughput experimental techniques most experimental designs are limited in their ability to produce detailed molecular networks of regulatory influences on a system-wide level. An alternative to determination of mechanistic networks between individual components is to constrain the parameter space by considering networks that describe the most important regulatory influences between groups of genes that represent important functions [6], here called functional influence networks. Functional influence networks involve regulatory processes that govern a specific set of system responses. The networks can be represented as causal influences between regulators, which mediate transitions between system states, and functional modules that provide the mechanism of action for the system [7]. For example, immune cells such as macrophages respond to certain kinds of stimuli (e.g. pathogen detection) by activating an inflammatory program that includes the transcriptional activation and subsequent release of inflammatory cytokines, pro-inflammatory effectors, and other components of the inflammatory program [8, 9]. These responses are regulated by a set of transcription factors (e.g. AP1, NFκB, and IRFs) that respond to pathogen associated molecular patterns (PAMPs) detected by TLRs. In a functional influence network the inflammatory response genes responding with similar dynamics would be considered to be functional modules and the genes that regulate their activation would be considered their regulatory influences. In this way the dynamic behavior of the network is simplified to facilitate modeling and represented only as expression patterns that represent collections of similarly behaving genes.

Modeling the dynamic behavior of functional influence networks makes it possible to chart the development of a biological network through time, with reference to experimental evidence from gene expression data. For example, Tegner et al. (2003) have created a method that models the change in each gene's expression as a linear process [10]. Another algorithm created recently for such dynamic modeling uses an ODE model for regulatory dynamics and L1 shrinkage as a means of selecting parsimonious models [11, 12]. The result is a coupled set of ODEs, each ODE describing the expression of a set of co-regulated genes as a function of the expression of genes identified as being regulators. A separate model is learned for each functional module, with each model defining the network edge connections between that cluster and its regulators and assigning strengths (coefficients) to each such regulatory interaction. Thus, the approach infers the regulatory network structure as it builds individual dynamic models for each regulated functional module.

Despite the significance of dynamic regulatory models, the performance of many inference methods is highly dependent on the initial clustering techniques. Inference methods require determination of subtle differences in patterns of gene expression profiles to best identify co-regulated functional modules. Unfortunately gene expression data has inherent noise and standard clustering techniques applied to limited sets of observations will inappropriately identify clusters. Existing knowledge, for example functional information about genes represented in the Gene Ontology (GO), can be used to augment clustering approaches. Methods to incorporate knowledge-driven techniques into predictive models of pathways have been recently proposed in which the GO is used to filter [13], enrich [14] or restructure [15] gene associations inferred from gene expression data through reverse engineering methods. These approaches have been shown to improve the biological plausibility of the network inferences drawn and the accuracy of the predictive models built. However, they still treat data- and knowledge-based inferences as incommensurable inputs, and the impact of each approach on the inferred network is factored in separately.

The goal of the current study is to show how dynamic modeling using functional influence networks can be used to infer the important regulatory influences that drive neuroprotection or injury during stroke in a mammalian model system and how incorporation of data from other sources can be used to improve model performance. We accomplish this using clustering, network topology and functional associations to refine components of functional influence networks (regulators and functional modules). We then use a machine-learning approach to learn relationships between components that can be used to robustly predict system dynamics. Our results show that predictive modeling in complex eukaryotic systems can be a useful way of generating hypotheses about the high-level functional regulation of the system, even with relatively few observations of the system. This approach provides valuable information about the processes of neuroprotection and injury during stroke in a whole animal model system, and generates a number of interesting hypotheses that are being experimentally validated.

2. Methods

2.1. Data sets

Briefly, we used a dataset of microarray results from blood of mice in a neuroprotection study, and data processing was performed as previously described [3]. The dataset comprises five treatments; ischemic preconditioning, lipopolysaccharide (LPS) or CpG injection, or control treatments, saline injection and sham surgery. The samples were taken 3, 24 and 72 hours post-preconditioning treatment, a stroke was induced at 72 hours then two more samples, 3 and 24 hours post-stroke induction, were taken for each preconditioning treatment.

2.2. Co-expression networks

We filtered this data to exclude probes that do not change significantly (p value > 0.05, fold-change expression < 2.0) resulting in 7352 transcripts. The expression levels of these transcripts (fold change relative to control untreated animals) were used as input to the CLR method [16] and the resulting relationships were filtered to a Z score of 5.0, yielding a network with 1880 nodes and 14205 relationship edges. We inverted the adjacency matrix for this network and treated it as a distance matrix for hierarchical clustering using complete linkage agglomeration and cut the dendrogram to generate 46 clusters to serve as initial targets for modeling.

The igraph library in the R statistical language was used to calculate the topology of the inferred networks. Bottlenecks are considered to be those genes with high betweenness centrality measures in the network [17, 18]; the highest 2.5% in this study.

2.3. Predictive modeling cross-validation approach

To infer a predictive regulatory model of neuroprotection during stroke we expanded on an algorithm that was previously applied to transcriptomics from prokaryotes and yeast [11, 12, 19]. We first applied the multivariate regression method, the Inferelator, to the targets defined from network analysis (above) using sets of potential regulators as described in the text. This method infers parsimonious sets of regulatory influences between regulators and targets (functional modules). In the learned model the relation between the expression of a target (y) and the expression levels of regulators with non-null influences on y (X) is expressed as:

$$\tau \frac{dy}{dt} = -y + \sum \beta_j X_j \tag{1}$$

Here, τ is the time step used in model construction and β is the weight for relationship X on y as determined by L_1 shrinkage using least angle regression [20]. To make predictions using a learned model eq. 1 can be solved for y, the expression of the target cluster. Assuming equilibrium conditions the derivative dy/dt is 0 and so equation (1) can be represented simply as a linear weighted sum:

$$y = \sum \beta_j X_j \tag{2}$$

and the dynamic version (for time series) is expressed for each time point (m) as:

$$y_m = \frac{-y_{m-1} + \sum \beta_j X_{m-1j}}{\tau} - y_{m-1} \tag{3}$$

In our modeling we used a τ of 30 minutes, which is appropriate for mRNA dynamics in a eukaryote [21].

Given the limited amount of transcriptomic data available for training we wanted to ensure that the models being inferred were robust, that is, that they were predictive of target expression under novel conditions not included in the training data. To accomplish this we employed a cross-validation approach to evaluate the performance of inferred models using different starting components (sets of regulators or target clusters, as described in the text). In the cross-validation the transcriptomic data is divided into five sets based on the treatment (i.e., LPS, CpG or ischemic preconditioning pretreatment, or saline or sham control treatments; see Figure 2B), five models are trained on the data excluding each treatment set in turn, and the performance of each model is evaluated on the left out set. Performance is evaluated as the average correlation of observed versus predicted expression values for each target weighted by the number of genes in each target, to produce a weighted gene-normalized overall performance score for the model, as:

$$P = \frac{\sum_{i=1}^{T} corr(pred_i, obs_i) n_i}{\sum_{i=1}^{T} n_i} \tag{4}$$

where *P* is the overall performance score, *T* is the number of targets in the model, *pred* and *obs* are the predicted and observed expression patterns, respectively, and *n* is the number genes in the target *i*. This cross-validation approach allows relatively unbiased assessment of model performance because the data used to evaluate the model is not included in the training data.

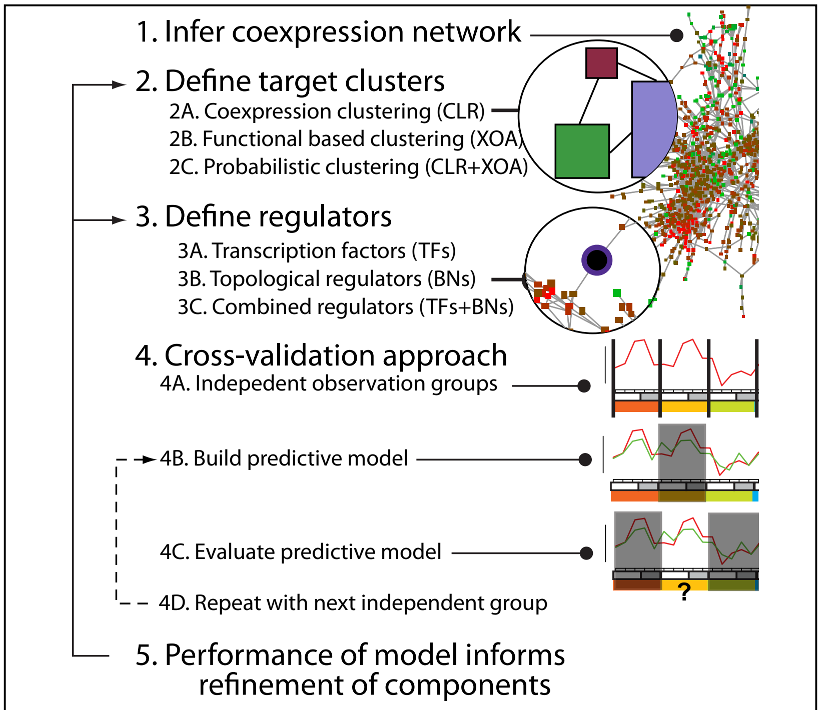

1. Infer coexpression network
2. Define target clusters
2A. Coexpression clustering (CLR)
2B. Functional based clustering (XOA)
2C. Probabilistic clustering (CLR+XOA)
3. Define regulators
3A. Transcription factors (TFs)
3B. Topological regulators (BNs)
3C. Combined regulators (TFs+BNs)
4. Cross-validation approach
4A. Indepedent observation groups
4B. Build predictive model
4C. Evaluate predictive model
?
4D. Repeat with next independent group
5. Performance of model informs refinement of components

Figure 1. Overview of iterative cross-validation predictive modeling approach. 1). Network inference from transcriptomic data using CLR. **2)** Definition of target clusters for modeling using several partitioning methods. **3)** Definition of potential regulators from existing knowledge or topological analysis. **4)** Cross-validation of predictive model: **A**. Divide expression data into related independent groups of observations (i.e. different treatments); **B**. Build a predictive model using all but one group with the Inferelator; **C.** Evaluate the performance of the model using the left out group; **D**. Repeat with next independent group. **5)** Use the overall predictive performance to evaluate and refine methods used to determine the network components (targets [2] and regulators [3]).

2.4. Probabilistic integration of relationships

Our previous results showed that partitioning co-regulated clusters of genes using either CLR or XOA associations could improve the performance of our cross-validated model. We were interested in combining the networks generated by both methods. Our approach was to treat the score for association between two genes as a p value for each method (see below), then partition the parent target cluster into subclusters using hierarchical clustering. We then used the predictive model generated for the genes in the subclustered target to assess which approach provides the best performance. We tested several approaches for integrating p values: maximum p value, minimum p value, mean p value, and the product of p values. Associations unique to either approach were transferred into the final similarity matrix directly, thus creating a union set of associations. Though the product of p values is the appropriate probabilistic combination of p values, the other methods were used because they may be more appropriate for specific instances. Additionally, the p values from each method do not have exactly the same meaning due to the differences in assumptions used in generating them. For CLR p values we converted the output of CLR (Z scores for the edge relative to the all other edges for each interaction partner) to p values using the normal distribution in R.

The p value for an XOA relationship is obtained by comparing the observed XOA score against the distribution of (a sample of) all possible scores obtained by computing the XOA similarity between all pairs of GO terms from the three subontologies. For example, the p value 0.14 associated with the XOA score of 3.76 assessing the similarity of GO:0007179 (BP: TGF-beta receptor signaling pathway) and GO:0016301 (MF: kinase activity) indicates that fewer than 14% of all XOA scores have higher semantic similarity than 3.76. Higher XOA scores are regularly found in association with lower p values. For example, statistically relevant values (< 0.05) typically correspond to XOA scores above 4.73. The p value across gene expresses the same idea, since the semantic similarity between two genes is the highest XOA score found pairing GO categories across the two genes:

$$XOA(GP1, GP2) = max\ XOA(c1_i, c2_j) \qquad (5)$$

where i=1,…,n and j=1, …,m, GP1 and GP2 are genes, $c1_i$ is one of the GO categories associated with $GP1$, and $c2_j$ one of the GO categories associated with GP2.

3. Results and Discussion

3.1. Reverse-engineering by predictive modeling of transcriptomic data

We are interested in developing a predictive model of neuroprotection in stroke at a systems level. There are significant gaps in knowledge about the regulation, functional mechanisms, and components that are involved in neuroprotection and stroke. These gaps prevent the development of molecular-level representations of the stroke process. We therefore have chosen to use a reverse-engineering approach that considers the regulatory influences and functional processes that these influences induce at a more abstract level. The resulting models will still provide useful and interpretable predictions that can be used for further experimental or computational investigation.

Our approach is to develop a predictive model of transcriptomic data using a machine-learning approach and cross-validation, and use the ability of this model to predict behavior under novel conditions as a way to refine the reverse engineering process (Figure 1). The reverse-engineering algorithm [11, 12] uses multivariate regression to learn ordinary differential equations (ODEs) that describe the relationship between the expression levels of a parsimonious set of regulators and the target functional module. Here, we apply this approach to a higher eukaryotic system with observations that are focused specifically on stroke response and neuroprotection.

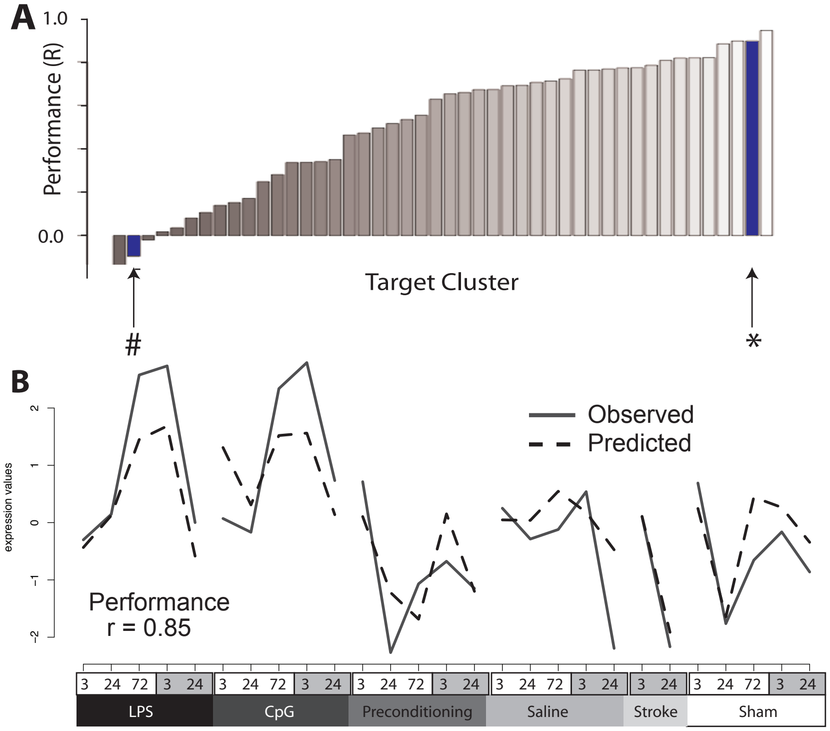


Figure 2. Performance of a predictive model of neuroprotection and injury during stroke in a mouse model system. A. Target cluster performance. The coexpressed clusters used as targets for modeling are shown (X axis) with bar height (Y axis) indicating the performance (correlation of predicted versus observed expression) for that target in the cross-validation approach. # indicates the poorly performing cluster used in further partitioning and * indicates the accurately predicted cluster shown in panel B. **B. Expression of an accurately predicted target.** The observed (red line) versus predicted (green line) expression levels (Y axis) for one cluster representing 180 genes is shown over the treatments/time points (X axis). The independent groups used in the cross-validation are indicated in colored boxes, and time points post-treatment (white boxes) and post-stroke induction (grey boxes) are also shown.

To define functional modules that are the targets in the model we used a transcriptomic data set from a mouse stroke model to infer functional relationships between genes using the context likelihood of relatedness (CLR) method [16] and used hierarchical clustering to define targets (see Methods). We initially treated all genes annotated as transcription factors (85 genes in the network) as potential regulators for reverse engineering.

To evaluate the performance of the model in a relatively unbiased manner we used a cross-validation approach (see Figure 1) that allows all the observations of the system to be treated as ‘independent’ data sets. We obtained an overall model performance of 0.52 (mean correlation per gene) observed versus predicted expression. In Figure 2A we show the performance (Y axis) of each cluster in the model (bars) ordered by performance. We mark the performance bar corresponding to the poorly performing cluster used for further analysis (see below) with a number sign and mark the bar corresponding to a well-predicted cluster with an asterisk. In Figure 2B we show the predicted (green line) and observed (red line) expression of the well-predicted cluster marked in panel A, over all the conditions examined (Y axis). The shaded bars below the X axis in Fig. 2B show the independent groups used for cross-validation. This correlation between observed and predicted expression shows that the model is robustly predictive of the behavior of the majority of the genes considered. This is an important result as it shows that regulatory influences that act as predictors can be learned from a relatively limited set of expression data. We note that the model itself provides a large number of interesting predictions about regulatory influences and expression of particular functional groups that are the focus of future studies. In this study we use this output of the model (predicted target behavior) to refine the components and relationships that are used for model generation.

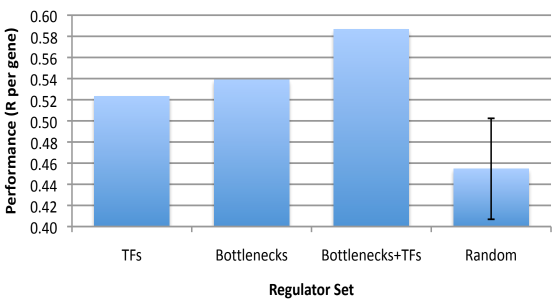


Figure 3. Bottlenecks are complementary to transcription factors as candidate regulatory influences. Predictive models were constructed using annotated transcription factors (TFs), topological bottlenecks, or a combination of the two groups (X axis). The mean and standard deviation (error bar) of ten randomly selected sets of genes is shown as a control. Performance (Y axis) using our cross-validation approach indicates that bottlenecks are robustly predictive of system behavior.

3.2. Network topology identifies important points of regulatory control

Many approaches for reverse-engineering regulatory networks preselect regulators based on sequence-based annotation, and then attempt to identify regulatory relationships between these sets of transcriptional regulators. Functional influence networks may be driven by mediators that are not transcriptional regulators, but could include effectors (e.g. immune effectors), signaling pathway components, metabolic enzymes, or any other component whose change mediates or reflects major changes in the state of the system. Previously our research has suggested the hypothesis that topological bottlenecks identified from transcriptional coexpression networks represent mediators of state transitions in systems [18, 22]. We thus tested the ability of topological bottlenecks to predict system behavior reasoning that true mediators of system transitions should be more predictive of system behavior than randomly chosen differentially regulated genes.

We examined the ability of bottlenecks to serve as regulators in our cross-validated modeling. As a comparison we randomly selected ten sets of differentially expressed genes in the network and evaluated their ability to predict the behavior of the targets in the model. Our results (Figure 3) show the performance of models that include transcription factors only, bottlenecks only, a combination of bottlenecks and transcription factors, or the mean of ten randomly chosen sets of genes. Bottlenecks provide modestly better performance than either the transcription factors set used initially or randomly selected genes. Furthermore, combining the transcription factors with the list of bottlenecks further improved the ability of the resulting model to predict expression behavior under novel conditions. This shows that the expression of bottlenecks is somewhat predictive of system behavior.

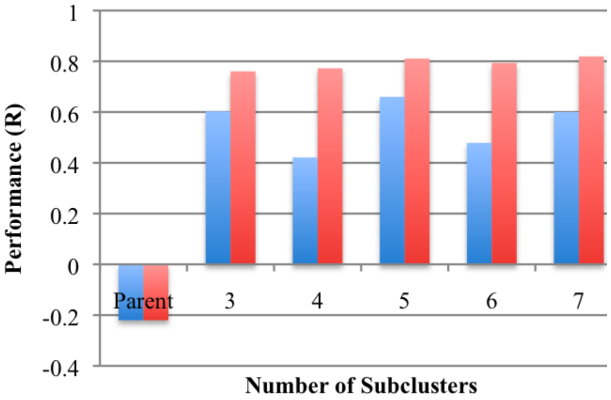


Figure 4. Performance of CLR and XOA defined subclusters for prediction. The parent cluster was subclustered using either the CLR (red)- or XOA (blue)-derived associations between genes into the indicated number of subclusters. Performance (mean correlation of observed versus predicted expression levels) is shown on the Y axis. These results support our previous observations that both methods can improve performance.

A surprising result of this analysis was that the randomly selected gene sets performed significantly worse than any of the selected regulator groups but the performance was still high (R = 0.45). This is likely to be due to the limited number of observations of the system that we are using for this work. Essentially the model is able to identify randomly selected genes which are somewhat predictive of the behavior of the targets because the dynamics of expression over the limited observations are relatively simple. Adding additional observations and/or data gathered for other purposes (TLR agonist treatment of mice, e.g.) should improve performance of our model. Further study is required to determine whether bottlenecks are indeed robustly predictive of system behavior.

3.3. Probabilistic integration of relationships improves delineation of functional modules

We next wanted to examine how the model could be further improved by better determination of target clusters. We examined how best to partition target gene clusters by combining results from the CLR and XOA algorithms to delineate subclusters. As a test case we focused on a problematic cluster with very poor performance (Figure 2A) identified in our previous study [15]. This cluster is made up of 335 genes and has a performance of -0.22 (correlation of predicted versus observed behavior) in the original model.

In Figure 4 we show the cross-validation performance on this cluster subdivided the cluster into 3-7 subclusters using either CLR (red bars) or XOA (blue bars) associations. These results show that using both expression-driven (CLR) or function-driven clustering can improve performance of the predictive model dramatically over that of the parent cluster.

We next examined how combining the two sets of associations could improve results. We chose to consider the strength of the associations as p values in order to directly compare the scores from different algorithms. We used four simple methods for combining p values for XOA and CLR scores when there were overlapping associations within a cluster: the minimum XOA or CLR p value, the maximum XOA or CLR p value, the mean of the XOA and CLR p values, and the product of the XOA and CLR p values. As shown in Figure 5, either the mean p value or maximum p value strategy provides the best performing solution for most cluster sizes, showing significant, but modest, improvement for a model composed of four subclusters. These findings indicate that an appropriate combination of approaches can improve the performance of predictive transcriptomic models

4. Conclusions

We have presented an approach to reverse-engineering from limited, but focused, transcriptional datasets and used it to infer functional influence networks of mouse blood during stroke. This approach uses a machine-learning method to iteratively define and refine the components of the network, both potential regulatory influences and coexpressed functional modules that are the targets of prediction (Figure 1). The approach is applicable to problems in which there are not well-established regulatory pathways already understood, where there are a limited number of observations of the system available, and where there may be complex and multiscale effects that need to be captured by the model, but not necessarily explicitly modeled. Our results demonstrate that the approach can be applied to provide biological insight into a complex and poorly understood pathology, such as neuroprotection and injury during ischemic stroke.

We show that a machine-learning method that employs multivariate regression techniques to learn ODEs describing relationships between regulators and target clusters can be applied to model transcriptomic dynamics from multicellular eukaryotic time course samples (Figure 2). This is an advance in modeling such systems that have traditionally been underrepresented in reverse-engineering applications due to their complexity and lack of ‘gold standard’ networks for validation. The results from cross-validation show that the models we produce can predict transcriptomic behavior of the majority of the genes considered under conditions not used to train the models. This approach is limited by the requirement that the gene-expression level changes of the regulatory influences must be indicative of their activity, an assumption that is clearly not true for many regulators. Additionally, regulatory influences inferred from such a limited set of observations, though predictive of system behavior to a significant degree, are unlikely to be highly accurate. However, this approach provides the foundation for more detailed investigation, both computationally and experimentally. These results represent an important first step toward more detailed and nuanced models of complex systems.

Using network topology we show that highly central bottlenecks are more predictive of system behavior than a similarly sized group of transcription factors (Figure 3). This result is consistent with the notion that bottlenecks from inferred networks represent mediators of transitions between system states [18, 22]. We further show that combining transcription factors and bottlenecks provides even better predictive performance. These gains are modest but statistically significant and we foresee that including more varied observations of the system will improve the results of the modeling, and should improve the definition of important mediators that we identify through network topology. However, the integration of such data will have to be undertaken carefully [23].

In our approach the performance of the predictive models is dependent on definition of the underlying functional modules used as targets for prediction. We initially define functional modules using hierarchical clustering based on expression profiles of genes. This approach gives good performance for a number of resulting clusters (Figure 2A) but does not provide accurate predictions for a number of significantly sized clusters. We show that further subclustering of a poorly performing cluster using either co-expression relationships from CLR or functional relationships from XOA [24] can dramatically improve the gene-wise performance of the parental cluster. Further, we use a probabilistic integration method and show that the combination of the two relationships can provide better performance than either individual method. This relatively simple approach has the advantage of being able to integrate arbitrary kinds of relationships between genes, so long as they can be associated with p values. We are currently examining what other kinds of relationships between genes will improve performance of the predictive models (e.g. protein-protein interactions, phylogenetic relationships).

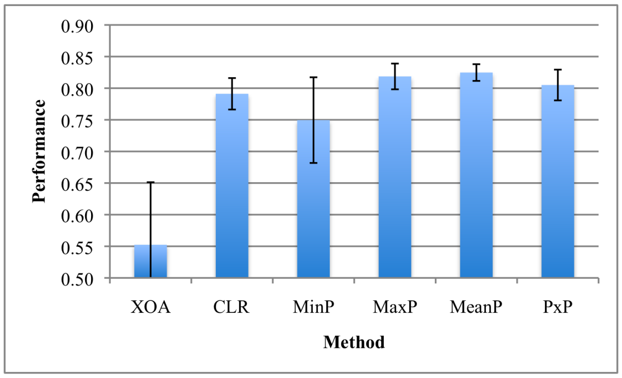


Figure 5. Comparison of subclustering methods. The mean performance of the methods examined (X axis) across different subclustering levels (3-7 clusters, as in Figure 1) is shown (Y axis). The error bars represent one standard deviation. The methods used are XOA and CLR alone, minimum p value (MinP), maximum p value (MaxP), mean of p values (MeanP) and product of p values (PxP). These results show that combining the CLR and XOA associations using probabilities can improve performance over the individual methods alone, but that only when non-standard methods (maximum p value or mean of p values) are employed to do so.

5. Acknowledgements

This work was supported under NIH/NINDS grant R01NS057484-03.

6. References

1. Donnan, G.A., et al., *Stroke.* Lancet, 2008. **371**(9624): p. 1612-23.
2. Stenzel-Poore, M.P., et al., *Effect of ischaemic preconditioning on genomic response to cerebral ischaemia: similarity to neuroprotective strategies in hibernation and hypoxia-tolerant states.* Lancet, 2003. **362**(9389): p. 1028-37.
3. Marsh, B., et al., *Systemic lipopolysaccharide protects the brain from ischemic injury by reprogramming the response of the brain to stroke: a critical role for IRF3.* J Neurosci, 2009. **29**(31): p. 9839-49.
4. Thakar, J., et al., *Modeling systems-level regulation of host immune responses.* PLoS Comput Biol, 2007. **3**(6): p. e109.
5. Oberhardt, M.A., B.O. Palsson, and J.A. Papin, *Applications of genome-scale metabolic reconstructions.* Mol Syst Biol, 2009. **5**: p. 320.
6. De Smet, R. and K. Marchal, *Advantages and limitations of current network inference methods.* Nat Rev Microbiol, 2010.
7. McDermott, J.E., et al., *Separating the drivers from the driven: Integrative network and pathway approaches aid identification of disease biomarkers from high-throughput data.* Dis Markers, 2010. **28**(4): p. 253-66.
8. Glass, C.K. and K. Saijo, *Nuclear receptor transrepression pathways that regulate inflammation in macrophages and T cells.* Nat Rev Immunol, 2010. **10**(5): p. 365-76.
9. Jenner, R.G. and R.A. Young, *Insights into host responses against pathogens from transcriptional profiling.* Nat Rev Microbiol, 2005. **3**(4): p. 281-94.
10. Tegner, J., et al., *Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5944-9.
11. Bonneau, R., et al., *A predictive model for transcriptional control of physiology in a free living cell.* Cell, 2007. **131**(7): p. 1354-65.
12. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.* Genome Biol, 2006. **7**(5): p. R36.
13. Gamalielsson, J., P. Nilsson, and B. Olsson, *A GO-Based Method for Assessing the Biological Plausibility of Regulatory Hypotheses*, in *ICCS 2006, Part II, LNCS 3992*, V.N.A.e. al., Editor. 2006. p. 879–886.
14. Sanfilippo, A., et al., *Using the gene ontology to enrich biological pathways.* Int J Comput Biol Drug Des, 2009. **2**(3): p. 221-35.
15. McDermott, J., et al., *An Integrated Approach to Predictive Genomic Analytics.*, in *ACM International Conference on Bioinformatics and Computational Biology,*. 2010: Niagra Falls, New York, USA.
16. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.
17. Yu, H., et al., *The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.* PLoS Comput Biol, 2007. **3**(4): p. e59.
18. McDermott, J.E., et al., *Bottlenecks and hubs in inferred networks are important for virulence in Salmonella typhimurium.* J Comput Biol, 2009. **16**(2): p. 169-80.
19. Madar, A., et al., *DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator.* PLoS ONE. **5**(3): p. e9803.
20. Efron, B., et al., *Least angle regression.* Annals of Statistics, 2003. **32**: p. 407-499.
21. Ross, J., *mRNA stability in mammalian cells.* Microbiol Rev, 1995. **59**(3): p. 423-50.
22. Diamond, D.L., et al., *Temporal proteome and lipidome profiles reveal hepatitis C virus-associated reprogramming of hepatocellular metabolism and bioenergetics.* PLoS Pathog, 2010. **6**(1): p. e1000719.
23. Cosgrove, E.J., T.S. Gardner, and E.D. Kolaczyk, *On the Choice and Number of Microarrays for Transcriptional Regulatory Network Inference.* BMC Bioinformatics, 2010. **11**(1): p. 454.
24. Posse, C., et al., *Cross-ontological analytics: Combining associative and hierarchical relations in the gene ontologies to assess gene product similarity.* Lecture Notes in Computer Science, 2006. **3992**: p. 871-878.

RATE-INDEPENDENT CONSTRUCTS FOR CHEMICAL COMPUTATION*

PHILLIP SENUM and MARC RIEDEL
Electrical and Computer Engineering, University of Minnesota
Minneapolis, MN 55455
http://cctbio.ece.umn.edu
E-mail: {senu0004, mriedel}@umn.edu

This paper presents a collection of computational modules implemented with chemical reactions: an inverter, an incrementer, a decrementer, a copier, a comparator, and a multiplier. Unlike previous schemes for chemical computation, ours produces designs that are dependent only on coarse rate categories for the reactions ("fast" vs. "slow"). Given such categories, the computation is exact and independent of the specific reaction rates. We validate our designs through stochastic simulations of the chemical kinetics. Although conceptual for the time being, our methodology has potential applications in domains of synthetic biology such as biochemical sensing and drug delivery. We are exploring DNA-based computation via strand displacement as a possible experimental chassis.

Keywords: synthetic biology; molecular programming; molecular computing; chemical reaction networks

1. Introduction

The theory of reaction kinetics underpins our understanding of biological and chemical systems.[1] It is a simple and elegant formalism: chemical reactions define *rules* according to which reactants form products; each rule fires at a *rate* that is proportional to the quantities of the corresponding reactants that are present. On the computational front, there has been a wealth of research into efficient methods for simulating chemical reactions, ranging from ordinary differential equations (ODEs)[2] to stochastic simulation.[3] On the mathematical front, entirely new branches of theory have been developed to characterize the dynamics of chemical reaction networks.[4]

Most of this work is from the vantage point of *analysis*: a set of chemical reaction exists, designed by nature and perhaps modified by human engineers; the objective is to understand and characterize its behavior. Comparatively little work has been done at a conceptual level in tackling the inverse problem of *synthesis*: how can one design a set of chemical reactions that implement specific behavior?

Of course, chemical engineers, genetic engineers and other practitioners strive to create novel functionality all the time. Generally, they begin with existing processes and pathways, and modify these experimentally to achieve the desired new functionality.[5,6] In a sense, much of the theoretical work on the dynamics of chemical reactions also addresses the synthesis problem by delineating the range of behaviors that are possible. For instance, theoretical work has shown that fascinating oscillatory and chaotic behaviors can occur in chemical reaction networks.[7,8]

Perhaps the most profound theoretical observation is that chemical reaction networks are, in fact, *computational processes*: regardless of the complexity of the dynamics or the subtlety of

*This research is supported an NSF CAREER Award, #0845650.

the timing, such networks transform *input* quantities of chemical species into *output* quantities through simple primitive operations. The question of the computational power of chemical reactions has been considered.[9] (The answer is interesting and subtle: *stochastic* chemical reactions can compute any function – they are "Turing-universal" in the jargon of computer science. However, *deterministic* chemical reactions are not so powerful – they are not Turing-universal.)

One of the great successes of integrated circuit design has been in abstracting and scaling the design problem. The physical behavior of transistors is understood in terms of differential equations – say, with models found in tools such as SPICE.[10] However, the design of circuits occurs at more abstract levels – in terms of switches, gates, and modules. Many analogous levels of abstraction exist for biological systems. These range from molecular dynamics, to protein networks, to genetic regulatory networks, to signaling pathways, to complete cellular systems, to multicellular organisms. Several authors have made implicit or explicit connections between biochemical reactions and digital electronics.[11–13]

Our contribution is tackle the problem of computation with chemical reactions from a conceptual vantage point focusing on robustness. Unlike previous schemes for chemical computation, ours produces designs that are dependent only on coarse rate categories for the reactions ("fast" and "slow"). Given such categories, the computation is exact and independent of the specific reaction rates. In particular, it does not matter how fast any "fast" reaction is relative to another, or how slow any "slow" reaction is relative to another – only that "fast" reactions are fast relative to "slow" reactions.

In our prior and related work, we have described a variety of computational constructs with chemical reaction networks, including programming constructs such as "for" and "while" loops,[14] signal processing operations such as filtering,[15] and arithmetic operations such as multiplication, exponentiation and logarithms.[14]

In this paper, we present designs of chemical reaction networks that implement specific computational modules: inverters, incrementers, decrementers, copiers, comparators, and multipliers. In contrast to some of our earlier published constructs, all of these constructs depend on only two rate categories. Although conceptual for the time being, our methodology has potential applications in domains of synthetic biology such as biochemical sensing and drug delivery.

2. Chemical Model

We adopt the model of discrete, stochastic chemical kinetics.[3,16] Molecular quantities are whole numbers (i.e., non-negative integers). Reactions fire and alter these quantities by integer amounts. The reaction rates are proportional to (1) the quantities of the reacting molecular types; and (2) rate constants. We aim for robust constructs: systems that compute exact results independently of specific rate constants. All of our designs are formulated in terms of two coarse rate constant categories (e.g., "fast" and "slow"). Given such categories, the computation is exact and independent of the specific reaction rates.

Consider the reaction

$$X_1 \xrightarrow{\text{fast}} X_2 + X_3. \tag{1}$$

When this reaction fires, one molecule of X_1 is consumed, one of X_2 is produced, and one of X_3 is produced. (Accordingly, X_1 is called a *reactant* and X_2 and X_3 the *products.*) Consider what this reaction accomplishes from a computational standpoint. Suppose that it fires until all molecules of X_1 have been consumed. This results in quantities of X_2 and X_3 equal to the original quantity of X_1, and a new quantity of X_1 equal to zero:

```
X2 = X1
X3 = X1
X1 = 0
```

Consider the reaction

$$X_1 + X_2 \xrightarrow{\text{fast}} X_3. \tag{2}$$

Suppose that it fires until either all molecules of X_1 or all molecules of X_2 have been consumed. This results in a quantity of X_3 equal to the lesser of the two original quantities:

```
X3 = min(X1, X2)
X1 = X1 - min(X1, X2)
X2 = X2 - min(X1, X2)
```

We will present constructs different arithmetical and logical operations in this vein. Each sets the final quantity of some molecular type as a function of the initial quantities of other types. The challenge in setting up computation with chemical reactions is that they execute asynchronously and at variable rates, dependent on factors such as temperature. In spite of this, we aim to implement computation that does not depend on the rates. We will only speak of rates in qualitative terms, e.g., "fast" vs. "slow" (in our notation, such qualitative rates are listed above the arrows for reactions.)

We validate our designs through stochastic simulations of the chemical kinetics.[17] First proposed by Gillespie, stochastic simulation has become the workhorse of computational biology – the equivalent, one might say, of SPICE for electrical engineering.[10] Such simulation tracks integer quantities of the molecular species, executing reactions at random based on propensity calculations. Repeated trials are performed and the probability distribution of different outcomes is estimated by averaging the results.

3. Computational Constructs

In this section, we present a collection of constituent constructs for rate-independent computation: an inverter, an incrementer/decrementer, a copier, and a comparator. In the next section, we use some of these constructs to implement a multiplier.

An Inverter

We implement an operation that is analogous to that performed by an inverter (i.e., a NOT gate) in a digital system: given a non-zero quantity (corresponding to logical "1") we produce a zero quantity (corresponding to logical “0”). Conversely, given a zero quantity we produce a non-zero quantity. We accomplish this with a pair of chemical types: the given type, call it a, and a corresponding “**absence indicator**” type, call it a_{ab}. The reactions generating the absence indicator are:

$$\varnothing \xrightarrow{\text{slow}} a_{\mathrm{ab}} \tag{3}$$

$$a + a_{\mathrm{ab}} \xrightarrow{\text{fast}} a \tag{4}$$

$$2\, a_{\mathrm{ab}} \xrightarrow{\text{fast}} a_{\mathrm{ab}} \tag{5}$$

Here the symbol $\varnothing$ as a reactant indicates that the reaction does not alter the quantity of the reactant types, perhaps because the quantity of these is large or replenishable.

The first reaction continuously generates molecules of a_{ab}, so in the absence of molecules of a we will have a non-zero quantity of a_{ab} in the system. If there are molecules of a present, then second reaction quickly clobbers any molecules of a_{ab} that are generated, so the quantity of a_{ab} will be close to zero. The third reaction ensures that the quantity a_{ab} remains small.

We use this simple construct in many of our computational modules.[15,18] In general, it can be used to synchronize steps. Suppose that we want to perform the following:

$$a \rightarrow b \tag{6}$$

$$b \rightarrow [\textit{operate on } b] \tag{7}$$

Here the second step is an operation that depends on the quantity of b. We do not want to start consuming molecules of b until the full quantity of it is generated from a. We can accomplish this with an absence indicator a_{ab}:

$$a \rightarrow b \tag{8}$$

$$a_{\mathrm{ab}} + b \rightarrow [\textit{operate on } b] \tag{9}$$

3.1. *Increment and Decrement Operations*

We describe constructs to implement incrementation and decrementation. These operations form the basis of more complex arithmetical operations, such as multiplication. The inputs consist of two molecular types g, the “start signal,” and x, the quantity to be incremented or decremented. We assume that some external source injects molecules of g. Any quantity can be injected; regardless, the quantity of x is incremented or decremented by exactly one. The system consumes all the molecules g. Once the quantity reaches zero, another increment/decrement operation can be performed. The operations proceed as follows:

1) The system waits for the start signal g to be some non-zero quantity.
2) It transfer the quantity of x to a temporary type x'.
3) It sets g to zero.
4) It transfers all but one molecule of x' back to x.
5a) For a decrement, it removes the last molecule x'.
5b) For an increment, it removes the last molecule of x' and adds to two molecules to x.

The following reactions implement this scheme. Given molecules of g, a reaction transfers molecules of x to molecules of x':

$$x + g \xrightarrow{\text{slow}} x' + g \tag{10}$$

The following reaction sets the quantity of g to zero. Using the absence indicator mechanism described in the preceding section, it does so only once all molecules of x have been transfered to x':

$$g + x_{\text{ab}} \xrightarrow{\text{slow}} \varnothing \tag{11}$$

Reactions of the form of 3– 5 are needed to generated x_{ab}; we omit them here. The following reaction transfers all but one molecule of x' back to x. It does so by repeatedly selecting pairs of x'. In essence, this is a repeated integer division by two. Again, using the absence indicator mechanism, it proceeds only once all molecules of g have been removed:

$$g'_{\text{ab}} + 2\,x' \xrightarrow{\text{fast}} x + x' + x'' \tag{12}$$

This reaction also produces molecules of a supplementary type x''. Note that this reaction is in the "fast" category. The new type x'' is consumed by the reaction:

$$x'' \xrightarrow{\text{slow}} \varnothing. \tag{13}$$

Note that this reaction is in the "slow" category. We introduce x'' because we cannot directly use an absence indicator for x' to detect when Reaction 12 has completed; here x' is not completed consumed. Instead, in reactions below we use an absence indicator for x''. Again, reactions of the form of 3– 5 are needed to generated x''_{ab}; we omit them here.

In Reaction 12, we do not directly use an absence indicator for g_{ab}, since that reaction is in the "fast" rate category. A design restriction for the absence indicator types is that they should never be directly involved in "fast" reactions. They are produced slowly and consumed quickly if the corresponding type is present in the system; if they were involved in a "fast" reaction, there would be competition. We avoid this by transferring the corresponding absence indicator g_{ab} to a secondary type g'_{ab} via a "slow" reaction:

$$\varnothing \xrightarrow{\text{slow}} g_{\text{ab}} \tag{14}$$

$$g_{\text{ab}} \xrightarrow{\text{slow}} g'_{\text{ab}} \tag{15}$$

We setup the absence indicator reactions for both types:

$$g + g_{\text{ab}} \xrightarrow{\text{fast}} g \tag{16}$$

$$g + g'_{\text{ab}} \xrightarrow{\text{fast}} g \tag{17}$$

$$2\,g_{\text{ab}} \xrightarrow{\text{fast}} g_{\text{ab}} \tag{18}$$

$$2\,g'_{\text{ab}} \xrightarrow{\text{fast}} g'_{\text{ab}}. \tag{19}$$

Finally, we include the following reaction to perform a decrement:

$$x''_{\text{ab}} + x' + g'_{\text{ab}} \xrightarrow{\text{slow}} \varnothing \quad \text{[Decrement]} \tag{20}$$

Or we include the following reaction to perform an increment:

$$x''_{\text{ab}} + x' + g'_{\text{ab}} \xrightarrow{\text{slow}} 2\,x \quad \text{[Increment]} \tag{21}$$

3.2. *A Copier*

In digital computation, one of the most basic operations is copying a quantity from one register into another. The programming construct is “set the value of b to be the value of a”:

```
let b = a;
```

To implement an equivalent operation with chemical reactions, we could use a reaction that simply transfers the quantity of a to b:

$$a \rightarrow b \tag{22}$$

However, this is not ideal because this reaction consumes all the molecules of a, setting its quantity to zero. We would like a chemical construct that copies the quantity without altering it. The following reaction does not work either:

$$a \rightarrow a + b \tag{23}$$

It just creates more and more molecules of b in the presence of a. A more sophisticated construct is needed.

In our construct, we have a “request-to-copy” type cr. When an external source injects molecules of cr, the copy operation proceeds. (The quantity of cr that is injected is irrelevant.) It produces an output quantity of b equal to the input quantity of a. It leaves the quantity of a unchanged. The reactions for the copier construct are as follows. Firstly, in the presence of cr, a reaction transfers the quantity of a to a':

$$cr + a \xrightarrow{\text{slow}} cr + a' \tag{24}$$

After all molecules of a have been transferred to a', the system removes all the molecules of cr:

$$cr + a_{\text{ab}} \xrightarrow{\text{slow}} \varnothing \tag{25}$$

Here, again, we are using the concept of an absence indicator. (The symbol $\varnothing$ as a product indicates “nothing”, meaning that the type degrades into products that are no longer tracked or used.) Removing cr ensures that a is copied exactly once. After cr has been removed, a reaction transfers the quantity of a' back to a and also creates this same quantity of b:

$$cr_{\text{ab}} + a' \xrightarrow{\text{slow}} a + b \tag{26}$$

We also generate absence indicators a_{ab} and cr_{ab} by the method described above. We note that, while this construct leaves the quantity of a unchanged after it has finished executing, it temporarily consumes molecules a, transferring the quantity of these to a', before transferring it back. Accordingly, no other constructs should use a in the interim.

3.3. *A Comparator*

Using our copier construct, we can create a construct that compares the quantities of two input types and produces an output type if one is greater than the other. For example, let us assume that we want to compare the quantities of two types a and b:

```
if  (a > b) {
    t = TRUE
} else {
    t = FALSE
}
```

If the quantity of a is greater than the quantity of b, the system should produce molecules of an output type t; otherwise, it should not produce any molecules of t.

Our construct for a comparator is as follows. First, we create temporary copies, c and d, of the types that we wish to compare, a and b, respectively, using the copier construct described in the previous section. (We omit these reactions; they are two verbatim copies of the copier construct, one with a as an input and c as an output, the other with b as an input and d as an output.) We split the copy request so that the two copiers are not competing for it:

$$cr \xrightarrow{\text{fast}} cr_1 + cr_2 \tag{27}$$

Now we compare a and b via their respective copies c and d. To start, we first consume pairs of c and d:

$$c + d \xrightarrow{\text{fast}} \varnothing \tag{28}$$

Note that this is a *fast* reaction; we assume that it fires to completion. The result is that there are only molecules of c left, or only molecules of d left, or no molecules of c nor d left. Molecules of the type that originally had a larger quantity have persisted. If the quantities were equal, then both types were annihilated. We use absence indicators c_{ab} and d_{ab} to determine which type was annihilated:

$$\varnothing \xrightarrow{\text{slow}} c_{\text{ab}} \tag{29}$$

$$c + c_{\text{ab}} \xrightarrow{\text{fast}} c \tag{30}$$

$$2\, c_{\text{ab}} \xrightarrow{\text{fast}} c_{\text{ab}} \tag{31}$$

$$\varnothing \xrightarrow{\text{slow}} d_{\text{ab}} \tag{32}$$

$$d + d_{\text{ab}} \xrightarrow{\text{fast}} d \tag{33}$$

$$2\, d_{\text{ab}} \xrightarrow{\text{fast}} d_{\text{ab}} \tag{34}$$

If a was originally greater than b, there will now be a presence of c and an absence of d. We produce molecules of type t if this condition is met. We preserve the quantities of c and d_{ab}; the amount t that we produce depends on the quantity of a fuel type:

$$\text{fuel} + c + d_{\text{ab}} \xrightarrow{\text{slow}} c + d_{\text{ab}} + t \tag{35}$$

For robustness, we also add a reaction to destroy t in the case that the asserted condition is not true:

$$c_{\text{ab}} + d + t \xrightarrow{\text{slow}} c_{\text{ab}} + d \tag{36}$$

$$c_{\text{ab}} + d_{\text{ab}} + t \xrightarrow{\text{slow}} c_{\text{ab}} + d_{\text{ab}} \tag{37}$$

We can readily generalize the construct to all types of logical comparisons. Table 1 lists these operations and their corresponding reactions.

Table 1. Logical operations via chemical reactions.

Operation	Creation	Destruction	Operation	Creation	Destruction
`a == b`	$a_{\mathrm{ab}} + b_{\mathrm{ab}}$	$a + b_{\mathrm{ab}}$ $a_{\mathrm{ab}} + b$	`a >= b`	$a + b_{\mathrm{ab}}$ $a_{\mathrm{ab}} + b_{\mathrm{ab}}$	$a_{\mathrm{ab}} + b$
`a > b`	$a + b_{\mathrm{ab}}$	$a_{\mathrm{ab}} + b$ $a_{\mathrm{ab}} + b_{\mathrm{ab}}$	`a <= b`	$a_{\mathrm{ab}} + b$ $a_{\mathrm{ab}} + b_{\mathrm{ab}}$	$a + b_{\mathrm{ab}}$
`a < b`	$a_{\mathrm{ab}} + b$	$a + b_{\mathrm{ab}}$ $a_{\mathrm{ab}} + b_{\mathrm{ab}}$	`a != b`	$a_{\mathrm{ab}} + b$ $a + b_{\mathrm{ab}}$	$a_{\mathrm{ab}} + b_{\mathrm{ab}}$

4. A Multiplier

Building upon the constructs in the last section, we show a construct that multiplies the quantities of two input types. Multiplication, of course, consists of iterative addition. Consider the following lines of pseudo-code:

```
while x > 0 {
   z = z + y
   x = x - 1
}
```

The result is that z is equal to x times y. We implement multiplication chemically using the constructs described in the previous sections: the line `z = z + y` is implemented with a copy operation; the line `x = x - 1` is implemented using a decrement operation. Only one additional reaction is needed to handle the `while` statement.

Firstly, we have reactions that copy the quantity of y to z. We use a "copy-request" sa type to synchronize iterations; it is supplied from the controlling reaction 52 below.

$$sa + y \xrightarrow{\text{slow}} sa + y' \tag{38}$$

$$sa + y_{\mathrm{ab}} \xrightarrow{\text{slow}} \varnothing \tag{39}$$

$$sa_{\mathrm{ab}} + y' \xrightarrow{\text{slow}} y + z \tag{40}$$

Secondly, we have reactions that decrement the value of x. We use sb as the signal to begin the decrement.

$$x + sb \xrightarrow{\text{fast}} x' + sb \tag{41}$$

$$sb + x_{\mathrm{ab}} \xrightarrow{\text{slow}} \varnothing \tag{42}$$

$$sb_{\mathrm{ab}} \xrightarrow{\text{slow}} sb'_{\mathrm{ab}} \tag{43}$$

$$2\,x' + sb'_{\mathrm{ab}} \xrightarrow{\text{fast}} x' + x + x'' \tag{44}$$

$$x'' \xrightarrow{\text{slow}} \varnothing \tag{45}$$

$$x' + x''_{\mathrm{ab}} + sb'_{\mathrm{ab}} \xrightarrow{\text{slow}} \varnothing \tag{46}$$

$$2\,sb'_{\mathrm{ab}} \xrightarrow{\text{fast}} sb'_{\mathrm{ab}} \tag{47}$$

$$sb'_{\mathrm{ab}} + sb \xrightarrow{\text{fast}} sb \tag{48}$$

Thirdly, we have a controlling set of reactions to implement the `while` statement. This set generates sa and sb to begin the next iteration, preserving the quantity of x:

$$x + x'_{\text{ab}} + y'_{\text{ab}} \xrightarrow{\text{slow}} x + \text{start} \tag{49}$$

$$\text{start} + x' \xrightarrow{\text{fast}} x' \tag{50}$$

$$\text{start} + y' \xrightarrow{\text{fast}} y' \tag{51}$$

$$\text{start} \xrightarrow{\text{slow}} sa + sb \tag{52}$$

This set initiates the next iteration of the loop if such an iteration is not already in progress and if there are still molecules of x in the system. The types x' and y' are present when we are decrementing x or copying y, respectively; thus, they can be used to decide whether we are currently inside the loop or not. Finally, we generate the four absence indicators according to the template in Reactions 3– 5.

5. Simulation Results

We validated our constructs using stochastic simulation. Specifically, we performed a time homogeneous simulation using Gillespie's "Direct Method"[3] with the software package "Cain" from Caltech.[19] In each case, the simulation was run until the quantities of all types except the absence indicators converged to a steady state. We used a rate constant of 1 for the "slow" reactions. We tried rate constants between two to four orders of magnitude higher for the "fast" reactions. (We refer to the ratio of "fast" to "slow" as the *rate separation.*) For each of the graphs below, the initial quantity of each type is zero, with the exception of the types specified.

5.1. *Multiplier*

Graph 1 shows the output of a single simulated trajectory for our multiplier. We observe exactly the behavior that we are looking for: the quantity of y cycles exactly 10 times as it exchanges with y' and is copied to z; the quantity of z grows steadily up to 100; the quantity of x decreases once each cycle down to 0. Table 2 presents detailed simulation results, this time tested for accuracy. Errors generally occur if the system executes too many or too few iterations. As can be seen, the larger the quantity of x, the more accurate the result, in relative terms. As expected, the larger the rate separation, the fewer errors we get.

Table 2. Statistical simulation results for "Multiplier" construct

Trial	Rate Separation	Trajectories	x	y	z	Expected z	Error
1	100	100	100	50	4954.35	5000	0.91%
2	100	100	50	100	4893.18	5000	2.14%
3	1000	100	100	50	4991.56	5000	0.17%
4	1000	100	50	100	4995.78	5000	0.08%
5	10000	100	100	50	4998.69	5000	$< 0.01\%$
6	10000	100	50	100	4999.14	5000	$< 0.01\%$
7	10000	100	10	20	200.04	200	$< 0.01\%$
8	10000	100	20	10	200.03	200	$< 0.01\%$

5.2. *Copier*

Graph 2 shows an average simulated trajectory for our copier. Again, we observe exactly the behavior we expect: the quantity of a drops to 0 almost immediately as it turns into a'; this is followed by the removal of cr from the system. When the quantity of cr drops to nearly zero, both a and b rise steadily back to the original quantity of a. Table 3 shows additional simulation results from our copier, this time tested for accuracy. The copier construct appears to be quite robust to errors; however, large rate separations do not help as much as they do for the multiplier. The system seems to prefer a larger injection quantity of cr, but whether it is larger or smaller than the initial quantity of a is irrelevant.

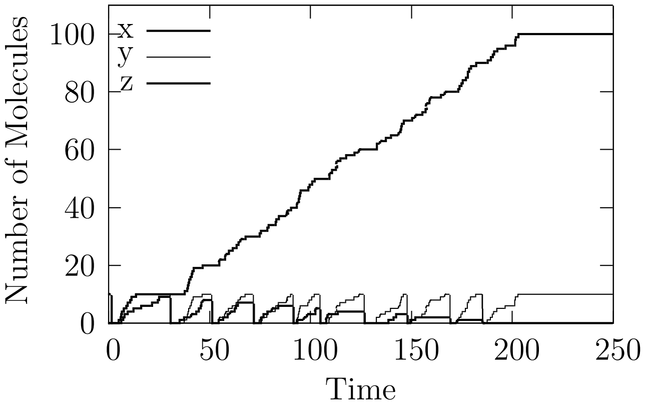
Table 3. Statistical simulation results for "Copier" construct

Trial	Rate Separation	Trajectories	cr	a	b	Expected b	Error
1	100	500	5	100	102.45	100	2.45%
2	100	500	50	100	104.826	100	4.826%
3	1000	500	5	100	100.312	100	0.312%
4	1000	500	50	100	100.516	100	0.516%
5	10000	500	5	100	100.022	100	0.022%
6	10000	500	50	100	100.034	100	0.034%
7	10000	500	5	5000	4938.39	5000	1.232%
8	10000	500	50	5000	4967.26	2	0.655%
9	10000	500	200	5000	4796.38	2	4.072%
10	10000	500	50	2	2	2	4.072%

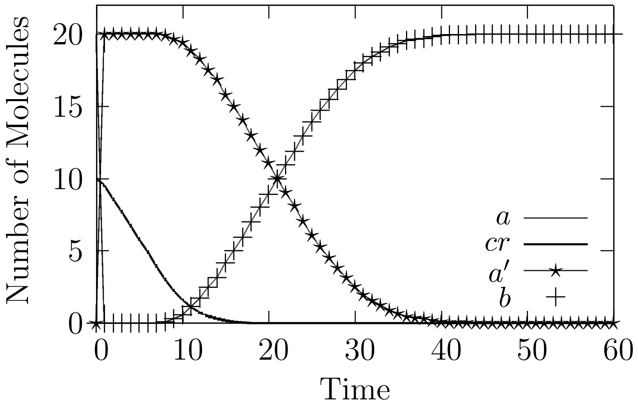
5.3. *Decrementer and Comparator*

Graph 3 shows the output of a single simulated trajectory of our decrementer. Exactly twenty peaks can be seen in the graph, including the initial peak on the far-left margin of the graph. This is exactly the behavior we are looking for – a decrement by exactly one each cycle. Graphs 4 and 5 display simulation results from our comparator. In Graph 4, t is asserted as we would expect; in Graph 5, t is not asserted, also as we would expect.

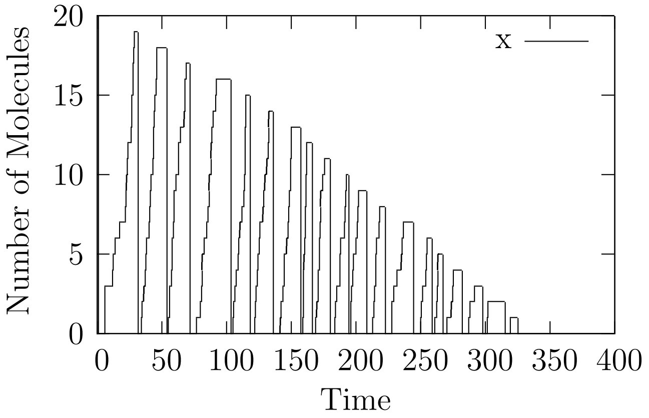
Graph 1: Simulated Multiplier, $x = 10$, $y = 10$



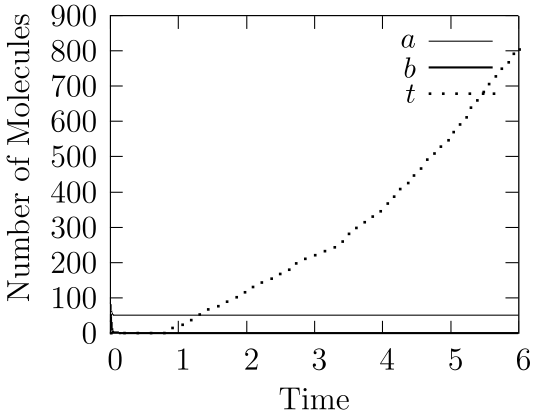
Graph 2: Simulated Copier, $a = 20$, $cr = 10$



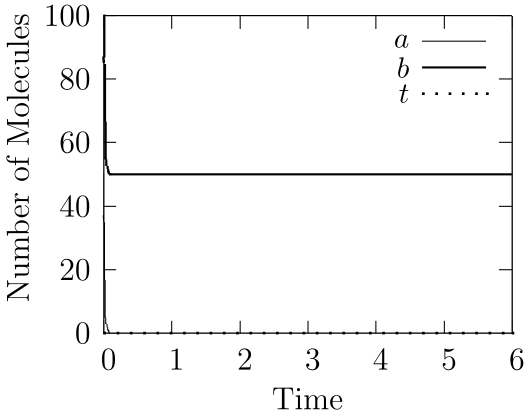
Graph 3: Simulated Decrement, $x = 20$



Graph 4: Comparator ($a > b$), $a = 100$, $b = 50$



Graph 5: Comparator ($a > b$), $a = 50$, $b = 100$



6. Discussion

Our contribution is to tackle the problem of synthesizing computation at a conceptual level, working not with actual molecular types but rather with abstract types. One might question whether actual chemical reactions matching our templates can be found. Certainly, engineering complex new reaction mechanisms through genetic engineering is a formidable task; for *in vivo* systems, there are likely to be many experimental constraints on the choice of reactions.[20] However, we point to recent work on *in vitro* computation as a potential application domain for our ideas.

Through a mechanism called DNA strand-displacement, a group at Caltech has shown that DNA reactions can emulate the chemical kinetics of nearly any chemical reaction network. Indeed, they provide a compiler that translates abstract chemical reactions of the sort that we design into specific DNA reactions.[21] Recent work has demonstrated both the scale of computation that is possible with DNA-based computing,[22] as well as exciting applications.[23] While conceptual, our work suggest a *de novo* approach to the design of biological functions. Potentially this approach is more general in its applicability than methods based on appropriating and reusing existing biological modules.

References

1. F. Horn and R. Jackson, "General mass action kinetics," *Archive for Rational Mechanics and Analysis*, vol. 47, pp. 81–116, 1972.

2. P. Érdi and J. Tóth, *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models.* Manchester University Press, 1989.
3. D. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.
4. S. Strogatz, *Nonlinear Dynamics and Chaos with Applications to Physics, Biology, Chemistry, and Engineering.* Perseus Books, 1994.
5. M. Win, J. Liang, and C. Smolke, "Frameworks for programming biological function through RNA parts and devices," *Chemistry & Biology*, vol. 16, pp. 298–310, 2009.
6. J. Keasling, "Synthetic biology for synthetic chemistry," *ACS Chemical Biology*, vol. 3, pp. 64–76, 2008.
7. I. R. Epstein and J. A. Pojman, *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos.* Oxford Univ Press, 1998.
8. K. D. Willamowski and O. E. Rössler, "Irregular oscillations in a realistic abstract quadratic mass action system," *Zeitschrift fur Naturforschung Section A – A Journal of Physical Sciences*, vol. 35, pp. 317–318, 1980.
9. D. Soloveichik, M. Cook, E. Winfree, and J. Bruck, "Computation with finite stochastic chemical reaction networks," *Natural Computing*, vol. 7, no. 4, 2008.
10. L. Nagel and D. Pederson, "Simulation program with integrated circuit emphasis," in *Midwest Symposium on Circuit Theory*, 1973.
11. R. Weiss, G. E. Homsy, and T. F. Knight, "Toward in vivo digital circuits," in *DIMACS Workshop on Evolution as Computation*, 1999, pp. 1–18.
12. J. C. Anderson, C. A. Voigt, and A. P. Arkin, "A genetic AND gate based on translation control," *Molecular Systems Biology*, vol. 3, no. 133, 2007.
13. Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, and E. Shapiro, "An autonomous molecular computer for logical control of gene expression," *Nature*, vol. 429, no. 6990, pp. 423–429, 2004.
14. A. Shea, B. Fett, M. D. Riedel, and K. Parhi, "Writing and compiling code into biochemistry," in *Proceedings of the Pacific Symposium on Biocomputing*, 2010, pp. 456–464.
15. H. Jiang, A. P. Kharam, M. D. Riedel, and K. K. Parhi, "A synthesis flow for digital signal processing with biomolecular reactions," in *IEEE International Conference on Computer-Aided Design*, 2010.
16. D. T. Gillespie, "Stochastic simulation of chemical kinetics," *Annual Review of Physical Chemistry*, vol. 58, pp. 35–55, 2006.
17. ——, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics*, vol. 22, no. 4, pp. 403–434, 1976.
18. A. Kharam, H. Jiang, M. D. Riedel, and K. Parhi, "Binary counting with chemical reactions," in *Pacific Symposium on Biocomputing*, 2011.
19. S. Mauch and M. Stalzer, "Efficient formulations for exact stochastic simulation of chemical systems," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 99, 2009.
20. R. Weiss, "Cellular computation and communications using engineering genetic regulatory networks," Ph.D. dissertation, MIT, 2003.
21. D. Soloveichik, G. Seelig, and E. Winfree, "DNA as a universal substrate for chemical kinetics," *Proceedings of the National Academy of Sciences*, vol. 107, no. 12, pp. 5393–5398, 2010.
22. L. Qian and E. Winfree, "A simple DNA gate motif for synthesizing large-scale circuits," in *DNA Computing*, 2009, pp. 70–89.
23. S. Venkataramana, R. M. Dirks, C. T. Ueda, and N. A. Pierce, "Selective cell death mediated by small conditional RNAs," *Proceedings of the National Academy of Sciences*, 2010 (in press).

TOWARDS REAL-TIME CONTROL OF GENE EXPRESSION: CONTROLLING THE HOG SIGNALING CASCADE

JANNIS UHLENDORF[1,2], SAMUEL BOTTANI[2], FRANÇOIS FAGES[1],
PASCAL HERSEN[2‖], GREGORY BATT[1♯]

[1] *Contraintes research group, Institut National de Recherche en Informatique et en Automatique,*
INRIA Paris–Rocquencourt, France
[♯]contact: gregory.batt@inria.fr

[2] *Laboratoire Matière et Systèmes Complexes, Université Paris Diderot and Centre National de la Recherche Scientifique, UMR 7057, Paris, France*
[‖]contact: pascal.hersen@univ-paris-diderot.fr

To decipher the dynamical functioning of cellular processes, the method of choice is to observe the time response of cells subjected to well controlled perturbations in time and amplitude. Efficient methods, based on molecular biology, are available to monitor quantitatively and dynamically many cellular processes. In contrast, it is still a challenge to perturb cellular processes - such as gene expression - in a precise and controlled manner. Here, we propose a first step towards *in vivo* control of gene expression: in real-time, we dynamically control the activity of a yeast signaling cascade thanks to an experimental platform combining a micro-fluidic device, an epi-fluorescence microscope and software implementing control approaches. We experimentally demonstrate the feasibility of this approach, and we investigate computationally some possible improvements of our control strategy using a model of the yeast osmo-adaptation response fitted to our data.

1. Introduction

To understand biology at the system level, one has to study both the *structure* and the *dynamics* of cellular processes [18,19,32]. On the one hand, genetic analyses are required to analyze the structure of signaling pathways and genetic networks. On the other hand, to access to the dynamical functioning of cellular processes, one has to observe the time response of cells to well controlled perturbations. Hence, the information level provided by experiments crucially depends on our capacity to observe *and* perturb biological systems at the cellular level. Efficient experimental tools have been developed to *monitor* both quantitatively and dynamically many cellular processes. Gene expression can be measured through micro-arrays or quantitative RT-PCR and conveniently observed at the single cell level through the combination of fluorescent reporter proteins and FACS techniques or microscopy [19,21,26]. In contrast, it is still a challenge to *perturb* cellular processes in a precise and controlled manner. A commonly used strategy resides on using inducible promoters to modulate the expression of a gene of interest by the addition of a diffusible molecule in the external cellular environment [12, 16,28]. However, even if the activity of the inducible promoter can be modulated quantitatively, there is no guarantee that the target gene will reach a desired constant expression level over a long period of time. Indeed, variations may arise because of modifications of the physiological state of the cell due to internal feedback loops and cellular adaptation. The expression of a transcription factor regulating itself is even more problematic. Moreover, both theoretical [1, 15] and recent experimental [14,24] results demonstrate the need for elaborate, time-varying perturbations to decipher quantitatively certain dynamics features of cellular responses. This notably includes the numerous biological processes in which the timing of gene expression plays a central role such as the regulation of the cell cycle. To summarize, existing solutions for the artificial control of gene expression are dissatisfying on two counts since, (i) expressing a gene of interest in a well-controlled, sustained way cannot be conveniently realized at the present time, and (ii) the investigation of certain dynamical properties necessitates dynamical, time-varying perturbations of gene expression for which no solution is currently available.

Here, we propose a first step towards *in vivo* control of gene expression. We have implemented an experimental platform for the *in vivo* control of a signaling pathway in *Saccharomyces cerevisiae*. We chose to control the activity of the HOG cascade which is activated in response to hyper-osmotic perturbations and promotes the transcription of osmo-adaptative response genes. Given a desired temporal profile, the activity of the signaling cascade is monitored in real time and deviations from the desired values are dynamically corrected by varying the osmolarity of the cellular environment (Fig. 1). This can be achieved thanks to a dedicated micro-fluidic device. This experimental platform is driven by software, that notably implements control algorithms, responsible for computing how the cellular environment (osmolarity) should be modified to correct the observed deviations from target values.

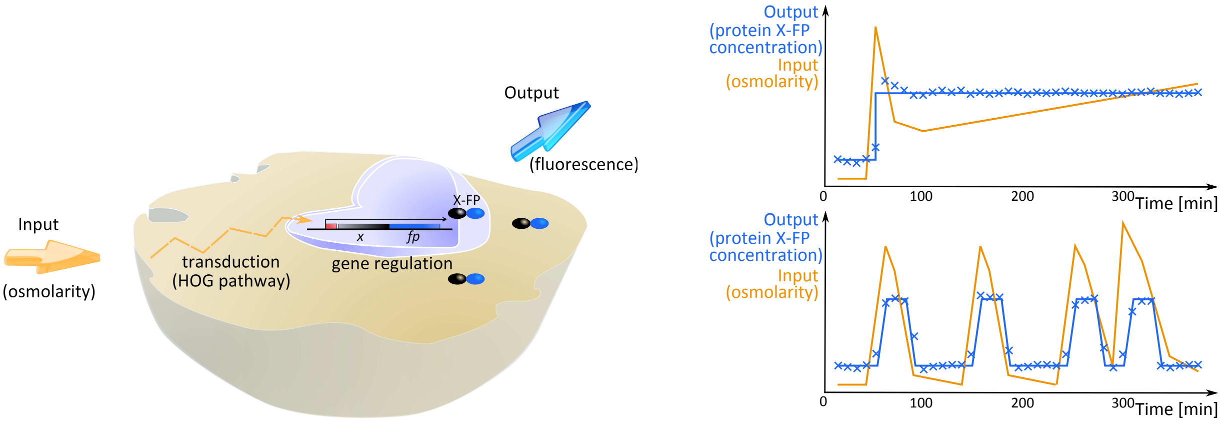


Fig. 1: The control problem. (a) Schematic Input/Output description of the cell. (b) Schematic representation of a desired output (blue), an applied input (orange) and the obtained output (blue crosses), for two different situations. In the first case (top), the goal is to dynamically maintain the concentration of a target protein X, fused to a fluorescent protein (FP), at a constant level. In the second case (bottom), the goal is to create a complex perturbation signal by varying the concentration of the protein X with time.

The work presented here differs significantly from previous applications of control theory in systems and synthetic biology contexts. So far, control theory contributions consisted essentially to shed a new light on biological phenomena, notably by suggesting underlying organization principles in biology [8,9]. An illustrative example is the use of the notion of integral feedback control to explain the robust perfect adaptation observed in bacterial chemotaxis [34]. Other insightful examples are given in a recent textbook [15]. Control theory has also been used in optimal experimental design applications [1,23]. But quite surprizingly, only a few (theoretical) studies focused on the actual control on a biomolecular process, e.g. [2,4,7,17]. Moreover, to the best of our knowledge, control theory has not yet been applied *in vivo* for the *actual feedback control of biological cellular processes at the single cell level*.

The paper is organized as follows. In section 2, we present in details the proposed platform for real-time control of the HOG signaling cascade activity and gene expression. In Section 3, we present preliminary experimental results obtained when controlling the nuclear localization of the Hog1 protein. This represents an essential first step towards controlling gene expression. In Section 4, we discuss possible improvements of our control approach using a simple model of the osmotic stress response. Conclusions are provided in the last section.

2. A platform for real-time control of gene expression

2.1. *An integrated real-time control platform*

Central to control theory is the notion of *feedback control* [30]. The idea is to compute the inputs to apply at the next time instant in function of outputs previously obtained. This way, knowledge of past errors is used to improve the control. In comparison to open loop control where the control strategy is computed beforehand, closed loop control approaches are generally less sensitive to model uncertainties and can compensate for external disturbances. These two features are highly desirable for any biological application. However, performing a control in real-time necessitates a tight integration between measurement device, control software and actuator.

As described in Figure 2, the HOG pathway activity can be monitored at the single cell level using time lapse fluorescent microscopy. The cellular environment can be controlled using the micro-fluidic device developed by Hersen and colleagues [14]. Not only this device allows a fast and well-controlled change of the cellular environment, but also, it guarantees that with the exception of the input signal the cell environment is otherwise held constant. We implemented algorithms for image analysis, state estimation and input computation in a Matlab program that communicates with and drives the microscope via MicroManager [31] and the micro-fluidic pressure controller.

2.2. *Using HOG signaling cascade*

To link external environmental changes to gene expression, we use a natural signaling pathway: the Hyper Osmolar Glycerol (HOG) pathway in the yeast *Saccharomyces cerevisiae*. This MAP kinase pathway is used to sense osmolar pressure changes in the environment and to trigger osmotic stress responses that maintain water homeostasis [29]. More precisely, two osmo-sensor proteins (Sln1 and Sho1) transduce the signal to the Hog1 protein via a phosphorylation cascade. Once phosphorylated, Hog1 promotes the osmo-adaptative response in at least three different ways. Firstly, Hog1 translocates into the nucleus and alters, directly or indirectly, the expression of a large number of genes [27]. Secondly, Hog1 has also a cytoplasmic activity since it regulates negatively glycerol export by inhibiting the activity of the Fps1 glycerol channel [3]. Thirdly, Hog1 activates glycerol producing enzymes, notably Gpd1 [33]. Hence, the osmo-adaptative response involves at least three natural feedback loops.

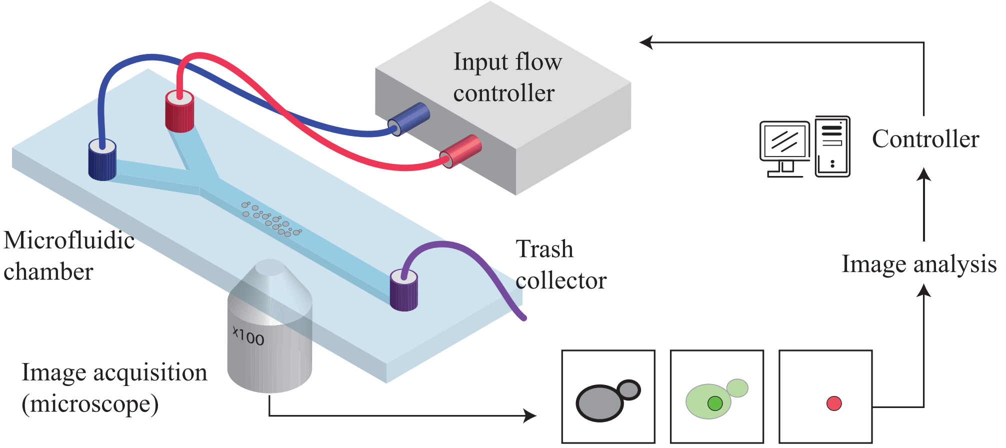


Fig. 2: The integrated control platform. The main elements of the feedback loop are (i) a microfluidic device allowing a rapid control of the cellular environment, (ii) a microscope for phase contrast and fluorescence measurements, (iii) yeast cells with Hog1, a nuclear marker (Htb2) and the protein of interest (X) fused to compatible fluorescent markers, and (iv) Matlab software for image analysis and controller implementation.

Our motivation for using this pathway is triple. Firstly, it has been extensively experimentally studied and quantitative models are available [6,14,20,22,24,25,35]. Second, the output of the signal transduction pathway can be experimentally quantified. Indeed, if Hog1 is fused to a fluorescent protein, its nuclear localization can be quantified and provides a measure of the Hog1 activity [10]. Thirdly, it has been experimentally shown that for fast osmolarity changes, the pathway integrates the signal: the transduction pathway acts as a low-pass filter with a bandwidth approximatively equal to 5×10^{-3} Hz [14]. This property allows us to *emulate* an analog control by rapidly switching (frequencies greater than 0.1 Hz) between two media: the normal growth medium and a sorbitol enriched ($\sim 1\,M$) medium. For example, a two minute osmotic stress corresponding to a $0.4\,M$ sorbitol intensity is obtained by flowing cells 12 times with normal medium during $6\,s$. and with sorbitol-rich medium during $4\,s$.

In this paper, we use a yeast strain with Hog1 fused to GFP and the nuclear protein Htb2 fused to mCherry [14]. The latter is used to conveniently localize the nuclear region. We define the relative Hog1 nuclear localization $h(t)$ as the ratio of the mean fluorescence pixel intensities of Hog1-GFP in the nucleus and in the cytoplasm.

$$h(t) = \frac{\langle Pixel\ intensity \rangle_{nuc}}{\langle Pixel\ intensity \rangle_{cyto}}$$

The normalized Hog1 nuclear localization $h_n(t)$ is then simply $h_n(t) = h(t)/h(t_0)$. These definitions are motivated by the fact that this gives measures that are relatively robust with respect to fluorescent protein photo-bleaching and cell-to-cell variations.

3. Controlling transcription factor nuclear localization using a simple control approach

In this section, we present preliminary results obtained on controlling the Hog1 nuclear localization. The control of Hog1 nuclear activity is a prerequisite for utilizing the Hog pathway

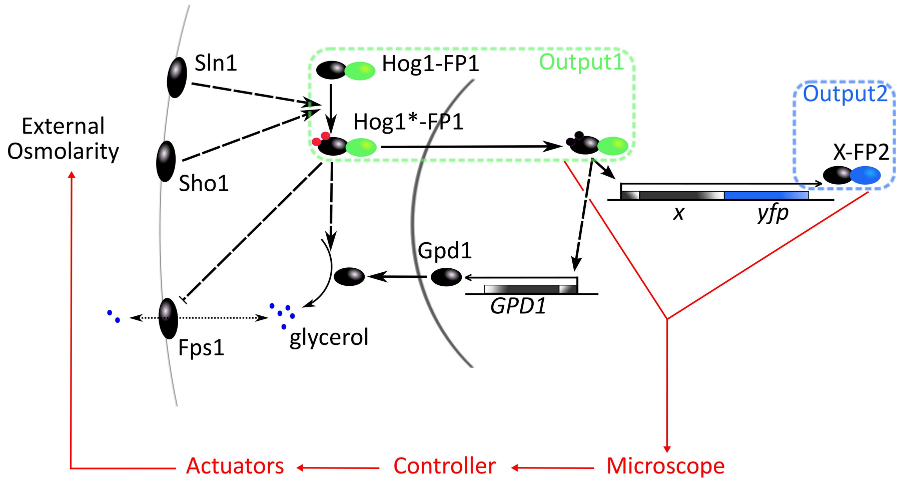


Fig. 3: Schematic representation of the HOG pathway with natural and engineered feedbacks. Solid and dashed arrows indicate direct and indirect effects, respectively. For a detailed description, see the main text. FP1 and FP2 in the figure denote two different fluorescent proteins.

to control gene expression. As a matter of fact, controlling the duration of activated Hog1 residence in the nucleus will lead to bursts of expression for the genes which are placed under Hog1 dependent promoter. There are different options how to encode a certain gene expression profile. We could either work with a constantly high signal and adjust the amount of Hog1 in the nucleus (amplitude modulation), or we could successively activate the Hog pathway for a short duration and control the frequency of these activations (frequency modulation). To test these two alternative strategies, we consider two problems: maintaining a given constant level of Hog1 nuclear localization over a long time period, or obtaining pulses of Hog1 nuclear localization in a repeated manner. These results have been obtained using the simplest control approach: a PID controller.

3.1. *PID control*

A proportional-integral-derivative (PID) controller is a generic closed-loop control algorithm, generic meaning that is does not require any structural knowledge about the controlled system [30]. Due to its simplicity this type of control is very often applied in engineering applications. A PID controller measures the deviations ("errors") of measured states from target states, and uses this information to compute the control. The applied control u at time t is the weighted sum of the error, $e(t)$, its derivative and the integral of past errors $e(\tau)$, $\tau \in [0, t]$:

$$u(t) = k_p \cdot e(t) + k_i \cdot \int_0^t e(\tau) d\tau + k_d \cdot \frac{d}{dt} e(t)$$

where k_p, k_i and k_d are the proportional, integral and derivative gains.

In our case the error $e(t)$ is the difference between the measured normalized Hog1 nuclear localization $h_n(t)$ and its reference value at the corresponding time point. Because we consider tracking problems, only the recent past errors are relevant. Therefore, we integrate the error only on the interval $[t-\Delta, t]$, where Δ is approximatively 2 minutes. We tuned the controller gains manually using a trial and error approach. The derivative term, and to a lesser extend, the proportional term are responsible for implementing a fast system response to target value changes. However large values for these parameters favor oscillations and loss of stability. In practice, we found that setting the derivative gain to zero and using values for k_p and k_i close to 2 and 1.5 leads to a good compromise between response time and stability in our experimental setting.

3.2. *Experimental results*

We designed two control experiments to test the possible strategies discussed above: using amplitude or frequency encoding. The first type of experiment is to try to maintain the system output at a constant target level (Fig 4 left). Quantitatively, the relative Hog1 nuclear localization should remain 20% higher than its nominal value in unstressed cells. The second type of experiment is to try to obtain repeatedly trapezoidal motifs. The amplitude of output variations also corresponds to a 20% increase above nominal value (Fig 4 right).

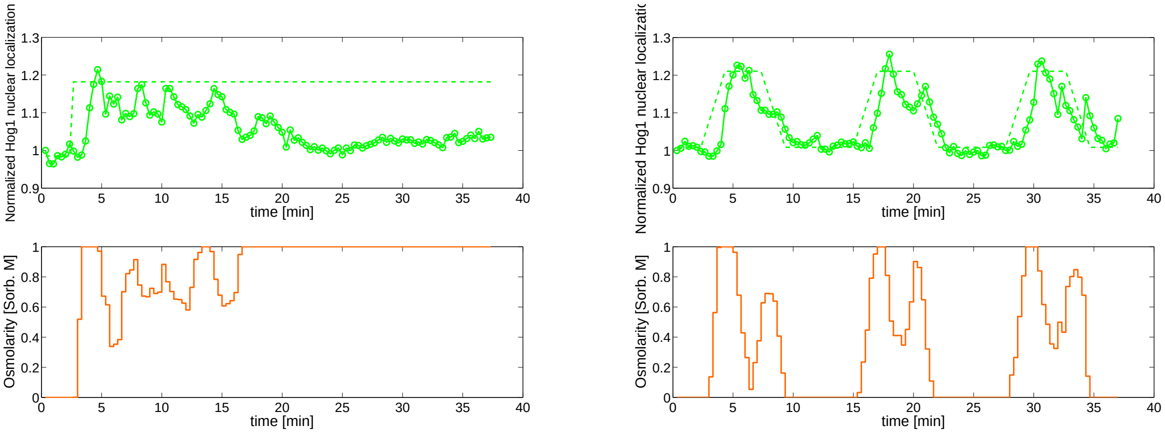


Fig. 4: Experimental results for the control of Hog1 nuclear localization. Left: Controlling the amplitude of Hog1 localization does only work for short durations due to internal feedback and cellular adaptation to sustained hyper osmotic conditions. Right: With a frequency encoded signal, the cell is able to reset between successive shocks and follows the reference values for the whole experiment.

Our experimental results clearly show that the control is effective in yeast cells. Consider for example the step experiment at time 2, when the target value changes. Following this change, the controller applies an osmotic stress, resulting after a 1-2 minute delay in an increase in the Hog1 nuclear localization. Then the system overshoots and the controller decreases the osmolarity of the environment. Oscillations ensue around a level below the target value during approximatively 15 minutes, during which increasing inputs are applied. Finally, even the maximal input is not sufficient to prevent the system from drifting away towards its nominal level.

The interpretation of these control results is simple. Because of internal natural feedbacks, the cells adapt (notably produce glycerol) and become insensitive to high osmolarity environments. Therefore, unless all internal feedback loops are inactivated, the amplitude-based control strategy seems not feasible. The inability of the controller to maintain the output at the target value in osmo-sensitive cells can be explained by the initiation of an osmo-adaptative response causing cells to drift away from the target value, together with the use of a rather narrow integration window in our PID controller.

Concerning the repeated motif experiment, it is fair to say that despite time lags and a relatively noisy behavior, the controller succeeds in producing the desired time varying output (Fig 4 right). As it appears on the plots, the 6 minute time separation between the 8 minute long motifs seem sufficient to fully reset the system to its normal, osmo-sensitive state. Based on these experimental results, the frequency encoding strategy for gene expression seems promising. However, before dealing with the actual control of gene expression, improvements in our control approach are needed. The capacity of the controller to predict rather than just to react -this would help dealing with the lag problem-, and the capacity to filter noise out -this would make the control more robust- are two features of significant interest.

4. Design of an improved control approach

The major advantage of the PID controller is that it does not rely on a model of the system. This makes it particularly easy to deploy. However, performances achieved using model-based control approaches are generally superior. In this section, we use the simple model proposed by Muzzey and colleagues [25], fitted to our data, to compare performances obtained with the PID controller and a model based control approach.

4.1. *Development of a simple linear model*

Numerous models have been developed for the osmotic stress response [6,14,20,22,24,25,35]. Because of its capacity to capture essential aspects of the HOG pathway, including notably the cell adaptation, and of its mathematical simplicity, we reuse the three variables linear model developed by Muzzey and colleagues [25]. In short, the state of the system is described by three variables, s_1, s_2, and s_3, corresponding to the nuclear Hog1 enrichment, its time integral, and glycerol relative concentration, respectively, and one input, u, corresponding to the external osmolarity. Since the osmo-stresses studied in [25] are caused by a different osmolyte (salt versus sorbitol), we introduce a factor σ to rescale the input u, if needed.

$$\dot{s}_1 = k_h\,(\sigma\,u - s_3) - \gamma_h\,s_1 \tag{1}$$

$$\dot{s}_2 = \alpha_d\,s_1 \tag{2}$$

$$\dot{s}_3 = s_2 + \alpha_i\,(\sigma\,u - s_3) - \gamma_g\,s_3 \tag{3}$$

In the above model, $\sigma\,u - s_3$ corresponds to the net osmolarity effectively sensed by the cell. In hyper-osmotic conditions, the production of intracellular glycerol (s_3) and the Hog1 nuclear localization (s_1) are increased. The increased Hog1 nuclear localization increases s_2 and hence s_3. Therefore one distinguishes a *direct* and an *indirect* effect of hyper-osmotic stress on glycerol accumulation [25].

To fit the model parameters to our system we perform two types of experiments in which cells are exposed to hyper-osmotic stresses differing either in magnitude or duration. The first set of experiments is used primarily to estimate the relation between osmotic stress and Hog1 localization, whereas the second set of experiments is used primarily to investigate the cell dynamical adaptation to osmotic stress. One should note that we experimentally measure the normalized Hog1 nuclear *localization* $h_n(t)$, whereas the variable $s_1(t)$ in the Muzzey model corresponds to the Hog1 nuclear *enrichment*. However, it holds that $h_n(t) = s_1(t) + 1$ [25]. In the sequel, to allow for comparison with the experimental results of Section 3, we present all our results -experimental and computational- using $h_n(t)$.

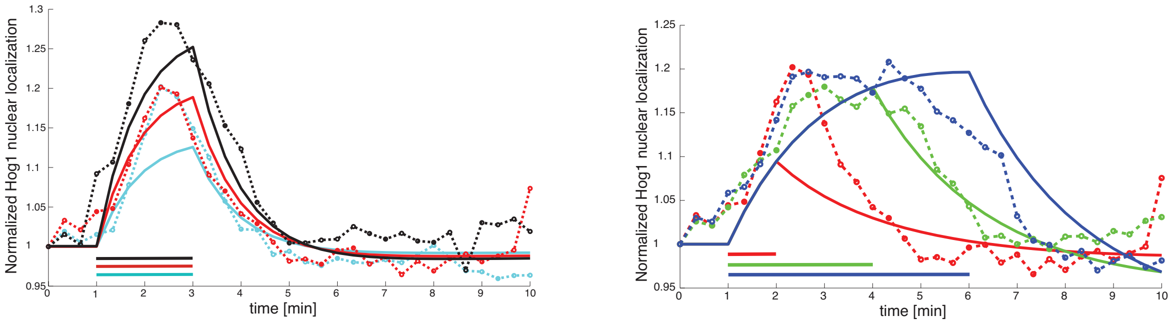


Fig. 5: Cell response to different hyper-osmotic stresses. Left: Stresses of different intensities. Cyan, red, and black plots correspond to a 0.4, 0.6, or 0.8 M stress applied during 2 minutes. Right: Stresses of different durations. Red, green and blue plots correspond to 0.6 M stress applied during 2, 4, or 6 minutes. Dashed and solid lines represent experimental data and model predictions, respectively.

To find parameter values for our model, we use the state-of-the-art non-linear optimization tool CMAES implementing a covariance matrix adaptation evolution strategy [13]. The objective function to minimize is the sum of a mean square error term, where the error is the difference between measured and predicted values for s_1, and a penalty term enforcing the positiveness of parameters and initial conditions. Parameter estimates are then manually fine-tuned (see Table 1).

k_h	γ_h	γ_g	α_h	α_g	σ
1.984	0.9225	0.5950	0.1612	0.0106	0.2

Table 1: Parameter values fitted to the experimental data shown in Fig. 5. All parameter units are min^{-1}, excepted for the dimensionless parameter σ.

As can be seen from the plots shown in Fig. 5, the model is able to capture qualitatively, and up to some degree, quantitatively, the behavior of yeast cells subjected to hyper-osmotic stresses. This is commendable given the extreme simplicity of the model.

4.2. *Comparison of different control approaches*

Equipped with a model of our system, we can computationally simulate the system response and compare various control approaches. Given the time-consuming aspect of experiments, working on simulated but realistic data allows us to rapidly test alternative control approaches. When computationally testing a model-based control approach, one uses the same model in the simulator and in the model based controller. That is, the model based controller knows perfectly the system dynamics. To make fair comparisons, we assume that only the output (and not the full state) is visible by the controller and we add (Gaussian) noise to the system output.

We present here a model predictive control (MPC) approach. The objective of MPC is to minimize the difference between the simulated and the target outputs by using a receding horizon strategy: given an estimate of the current state of the system, a control strategy to be applied during a short time horizon is searched for, and applied for a short period of time. Then, the approach is applied again, with the estimation of the new state, and the computation of a control strategy for a new short time horizon. This receding horizon strategy yields an effective feedback control [11]. Because MPC applies to linear and nonlinear systems, this approach can easily be extended to deal with future improved models. An other motivation for using MPC rather than the conventional control approach for linear system output tracking, based on a linear quadratic gaussian controller [30], is that simple non-linear constraints (e.g. bounded input) can easily be integrated in this framework.

For our application, we implement an MPC approach using Kalman filtering and a simple search strategy. The use of a Kalman filter is a standard approach to estimate the full state of a linear system based on (noisy) observations [30]. Then, at time t, we search for three input values, u_1, u_2, and u_3, that when applied on the time intervals $[t, t+1]$, $]t+1, t+2]$, and $]t+2, t+3]$, respectively, minimize the squared error, again defined as the distance between the target and the simulated outputs, on the time interval $[t, t+3]$. u_1 is applied on $[t, t+1]$ and the procedure is restarted at time $t+1$. At each iteration, we use CMAES, a global optimization, tool to search for the three input osmolarities u_1, u_2, and u_3. Naturally, in our setting, the input (osmolarity) is necessarily positive and bounded. Therefore, we limit the search to the interval of feasible osmolarities. The computational effort remains limited, since less than one second is needed for each iteration. For comparison, the timestep duration of the control loop in our experiments is $20\,s$. So using MPC does not challenges the real-time requirement.

We also consider here the PID controller presented in Section 3, but applied on simulated data as explained in this section. All these computational procedures have been implemented in Matlab.

The results obtained with the two control strategies and the for two different control problems are shown in Figure 6. Regarding the difficulty to maintain pathway activity over a prolonged period and the feasibility of creating repeated short time activity patterns, the results obtained with both control approaches are fully consistent with our experimental findings. The comparison of the results obtained with the PID controller on the experimental (Fig 4) and simulated data (Fig 6) shows that the PID performs better in the second case. This might be explained by a higher complexity of cellular variability (ie the “noise” is not just plain Gaussian). As expected, the lag and incomplete drift compensation are also observed on simulated data, albeit attenuated. In contrast, the model predictive results show neither. This corroborates the fact that they originate -at least partly- from the reactive rather than predictive nature of the PID controller. Moreover, the control is also much more regular in the MPC experiments. Very likely, this comes from the use of Kalman filtering. One should note that this is not due to an improper parametrization of the PID. Indeed the relatively

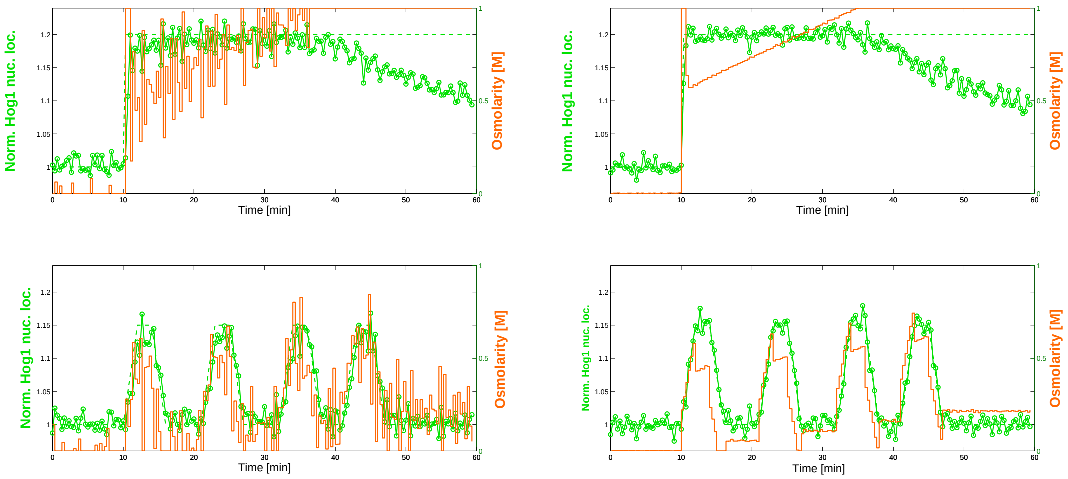


Fig. 6: Comparison of PID (left) and MPC (right) control strategies for two different control problems and on simulated data. *Norm. Hog1 nuc. loc.* stands for normalized Hog1 nuclear localization.

high proportional gain that causes large input changes is needed to ensure a fast response.

To summarize, the model predictive control approach is superior on all counts to the PID controller, at the cost of a very limited computational overhead. However, one should stress that the quality of a model based controller ultimately depends on the quality of the model of the system. So to effectively apply MPC on yeast cells, significant modeling work might be needed. But then one will have the effective proof that the main features of the osmo-adaptative response are captured in sufficient details.

5. Discussion

We presented an integrated experimental platform and demonstrated the feasibility of controlling the nuclear localization of the protein Hog1. Stated differently, we have shown how to create a dynamically controlled inducible promoter. As a matter of fact, it should be possible to place any gene under the control of a Hog1-dependent promoter and then to force its expression by controlling Hog1 nuclear localization. Consequently, this contribution describes a first, crucial step towards real-time control of gene expression.

Using the HOG pathway has several advantages, the most important ones being its quick activation and de-activation which are crucial to ensure efficient dynamics of the control loop, and the established correlation between nuclear localization and activity. It is to be noted though, that contrarily to known inducible promoters such as the Tet system, activating the HOG pathway also affects the cell physiological state, since many genes are transcribed to ensure proper cellular response to an hyper osmotic environment. For real applications, one should achieve a clear separation between controlling gene expression dynamically and altering the physiological state. This might require engineering the HOG cascade, or using other alternative signaling pathways with similar dynamics and nuclear translocation.

Interestingly, our results suggest that for our application, it is preferable to use *frequency* *encoding* to control gene expression. Indeed, because of fast, non-genetic adaptation feedbacks, the output of the signaling cascade can not be held constant over a prolonged period. A Frequency encoding strategy is widely used by neural networks which computing activity relies on action potential pulses. Although it is generally assumed that gene regulation is naturally controlled by amplitude modulation, a recent study by Elowitz's team showed that the expression of some genes in yeast are regulated by the frequency of expression bursts led by the transcription factor Crz1 [5]. The authors proposed that the functional role of frequency modulation is to ease the coordination of the expression of multiple target genes. Based on our results, one can propose an alternative role of regulation by frequency modulation: it allows for both a rapid non-genetic response and a slower transcriptional response leading to a complete adaptation to a given stress.

In a future work, we will use a model-based control approach to improve our results on Hog1 nuclear localization. Moreover, we will progress towards our main goal, that is, gene expression control, by studying a candidate gene fused to a fluorescent tag under the direct control of Hog1. The control platform will be adapted to read as outputs both the localization of Hog1 and the actual expression level of the gene of interest.

We anticipate that this platform to tune in real-time the level of expression of a gene of interest will be a useful tool for the biologist to better understand living processes in single cells. Quoting Feynman saying 'what I cannot built, I cannot understand', synthetic biologists propose that building systems helps to better understand them. Here, we propose that controlling them is an effective way to assess our understanding: what I cannot control, I have not understood.

References

1. J.F. Apgar, J.E. Toettcher, D. Endy, F.M. White, and B. Tidor. Stimulus design for model selection and validation in cell signaling. *PLoS Computational Biology*, 4(2):e30, 2008.
2. S. Azuma, E. Yanagisawa, and J. Imura. Controllability analysis of biosystems based on piecewise affine systems approach. *IEEE Transactions on Circuits and Systems and IEEE Transactions on Automatic Control*, 53, 2008. Joint special issue on Systems Biology.
3. S. E. Beese, T. Negishi, and D.E. Levin. Identification of positive regulators of the yeast Fps1 glycerol channel. *PLoS Genetics*, 5(11):e1000738, 2009.
4. C. Belta, L.C.G.J.M. Habets, and V. Kumar. Control of multi-affine systems on rectangles with applications to hybrid biomolecular networks. In *Proceedings of the 41st IEEE Conference on Decision and Control, CDC'02*, 2002.
5. L. Cai, C.K. Dalal, and M.B. Elowitz. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*, 455(7212):485–490, 2008.
6. A.P. Capaldi, T. Kaplan, Y. Liu, N. Habib, A. Regev, N. Friedman, and E. K O'Shea. Structure and function of a transcriptional network activated by the MAPK Hog1. *Nature Genetics*, 40(11):1300–1306, 2008.
7. M. Chaves and J.-L. Gouzé. Exact control of genetic networks in a qualitative framework: the bistable switch example. Technical Report RR-7359, INRIA Sophia-Antipolis, 2010.
8. M.E. Csete and J.C. Doyle. Reverse engineering of biological complexity. *Science*, 295(5560):1664–1669, 2002.
9. H. El-Samad, H. Kurata, J.C. Doyle, C.A. Gross, and M. Khammash. Surviving heat shock: Control strategies for robustness and performance. *Proceedings of the National Academy of Sciences of the USA*, 102(8):2736–2741, 2005.
10. P. Ferrigno, F. Posas, D. Koepp, H. Saito, and P.A. Silver. Regulated nucleo/cytoplasmic exchange of

HOG1 MAPK requires the importin beta homologs NMD5 and XPO1. *EMBO Journal*, 17(19):5606–5614, 1998.

11. R. Findeisen, F. Allögwer, and L. Biegler. *Assessment and Future Directions of Nonlinear Model Predictive Control*, volume 358 of *LNCIS*. Springer, 2007.
12. M. Gossen, S. Freundlieb, G. Bender, G. Muller, W. Hillen, and H. Bujard. Transcriptional activation by tetracyclines in mammalian cells. *Science*, 268(5218):1766–1769, 1995.
13. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.
14. P. Hersen, M. N. McClean, L. Mahadevan, and S. Ramanathan. Signal processing by the HOG MAP kinase pathway. *Proceedings of the National Academy of Sciences of the USA*, 105(20):7165–7170, 2008.
15. P.A. Iglesias and B.P. Ingalls. *Control Theory and Systems Biology.* MIT Press, 2009.
16. F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3:318–356, 1961.
17. A.A. Julius, A. Halasz, M.S. Sakar, H. Rubin, V. Kumar, and G.J. Pappas. Stochastic modeling and control of biological systems: the lactose regulation system of *Escherichia coli*. *IEEE Transactions on Circuits and Systems and IEEE Transactions on Automatic Control*, 53, 2008. Joint special issue on Systems Biology.
18. H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.
19. E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: Concepts, implementation and application.* Wiley Press, 2005.
20. E. Klipp, B. Nordlander, R. Kruger, P. Gennemark, and S. Hohmann. Integrative model of the response of yeast to osmotic shock. *Nature Biotechnology*, 23(8):975–982, 2005.
21. K.M. Klucher, M.J. Gerlach, and G.Q. Daley. A novel method to isolate cells with conditional gene expression using fluorescence activated cell sorting. *Nucleic Acids Research*, 25(23):4858–4860, 1997.
22. J. Macia, S. Regot, T. Peeters, N. Conde, R. Solé, and F. Posas. Dynamic signaling in the Hog1 MAPK pathway relies on high basal signal transduction. *Science Signaling*, 2(63):ra13, 2009.
23. F. Menolascina, D. Bellomo, T. Maiwald, V. Bevilacqua, C. Ciminelli, A. Paradiso, and S. Tommasi. Developing optimal input design strategies in cancer systems biology with applications to microfluidic device engineering. *BMC Bioinformatics*, 10(Suppl 12):S4, 2009.
24. J.T. Mettetal, D. Muzzey, C. Gomez-Uribe, and A. van Oudenaarden. The frequency dependence of osmo-adaptation in Saccharomyces cerevisiae. *Science*, 319(5862):482–484, 2008.
25. D. Muzzey, C.A. Gomez-Uribe, J.T. Mettetal, and A. van Oudenaarden. A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell*, 138(1):160–171, 2009.
26. D. Muzzey and A. van Oudenaarden. Quantitative time-lapse fluorescence microscopy in single cells. *Annual Review of Cell and Developmental Biology*, 25(1):301–327, 2009.
27. S.M. O'Rourke and I. Herskowitz. Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Molecular Biology of the Cell*, 15(2):532–542, 2004.
28. T.M. Roberts, R. Kacich, and M. Ptashne. A general method for maximizing the expression of a cloned gene. *Proceedings of the National Academy of Sciences of the USA*, 76(2):760–764, 1979.
29. W.H. Mager S. Hohmann. *Yeast stress responses.* Topics in Current Genetics. Springer, 2003.
30. E.D. Sontag. *Mathematical Control Theory. Deterministic Finite-Dimensional Systems.* Springer, 1998.
31. N. Stuurman and A. Edelstein. μmanager: the open source microscopy software v.1.3, 2010. http://www.micro-manager.org.
32. Z. Szallasi, J. Stelling, and V. Periwal, editors. *System Modeling in Cellular Biology: From Concepts to Nuts and Bolts.* MIT Press, 2006.
33. P.J. Westfall, J.C. Patterson, R.E. Chen, and J. Thorner. Stress resistance and signal fidelity independent of nuclear MAPK function. *Proceedings of the National Academy of Sciences of the USA*, 105(34):12212–12217, 2008.
34. T.-M. Yi, Y. Huang, M.I. Simon, and J.C. Doyle. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proceedings of the National Academy of Sciences of the USA*, 97(9):4649–4653, 2000.
35. Z. Zi, W. Liebermeister, and E. Klipp. A quantitative study of the Hog1 MAPK response to fluctuating osmotic stress in Saccharomyces cerevisiae. *PLoS ONE*, 5(3):e9522, 2010.

IDENTIFYING TARGETS FOR INTERVENTION BY ANALYZING BASINS OF ATTRACTION

MICHAEL P. VERDICCHIO[1] AND SEUNGCHAN KIM[1,2*]

[1]*School of Computing, Informatics and Decision Systems Engineering, Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, AZ*

[2]*Computational Biology Division, Translational Genomics Research Institute (TGen), Phoenix, AZ*

mv@asu.edu, dolchan@tgen.org

Motivation: A grand challenge in the modeling of biological systems is the identification of key variables which can act as targets for intervention. Good intervention targets are the "key players" in a system and have significant influence over other variables; in other words, in the context of diseases such as cancer, targeting these variables with treatments and interventions will provide the greatest effects because of their direct and indirect control over other parts of the system. Boolean networks are among the simplest of models, yet they have been shown to adequately model many of the complex dynamics of biological systems. Often ignored in the Boolean network model, however, are the so called basins of attraction. As the attractor states alone have been shown to correspond to cellular phenotypes, it is logical to ask which variables are most responsible for triggering a path through a basin to a particular attractor.

Results: This work claims that logic minimization (i.e. classical circuit design) of the collections of states in Boolean network basins of attraction reveals key players in the network. Furthermore, we claim that the key players identified by this method are often excellent targets for intervention given a network modeling a biological system, and more importantly, that the key players identified are not apparent from the attractor states alone, from existing Boolean network measures, or from other network measurements. We demonstrate these claims with a well-studied yeast cell cycle network and with a WNT5A network for melanoma, computationally predicted from gene expression data.

Keywords: Boolean Networks; Attractors; Logic Minimization; Intervention

1. Introduction

Biological systems are complex in many dimensions as endless transportation and communication networks all function simultaneously. While differential equation models are the most comprehensive at capturing and modeling the true dynamic behaviors of a real biological system,[1] the use of such a framework requires supplying precise model parameters, most of which are not readily measurable with current technologies.

Boolean networks are among the simplest of models, yet they have been shown to adequately model many of the complex dynamics of biological systems. Their popularity is also based on the ease of distilling our knowledge about a particular biological process to positive and negative pair-wise relationships. Since seminal work by Stuart Kauffman in the 1960s relating network attractor states to cell fate,[2] Boolean network dynamics have been studied and related to various biological phenomena. In addition to Boolean networks, many other graphical models have become popular in the modeling of biological interactions, with one interesting property often being the biological significance of network hubs (though this is also a contested view[3]). Specifically, vertices (or nodes) in networks with high degree (also known

[*] Address correspondence to *dolchan@tgen.org*

as network hubs) have often been found to have higher biological significance than those less connected nodes in the same network, especially in scale-free networks. Thus, some simple topological analysis, including network centrality measures, can help to identify interesting variables and possibly even targets for intervention.

Wuensche[4] and others also have studied basins of attraction in Boolean network models of genomic regulation, specifically the relationship of their structures to the stability of attractors (cell types) in the face of perturbations. However, because of the size and transient nature of basins of attraction, they are often neglected in analysis in favor of the attractor states.

As a basin of attraction is a collection of states leading into a corresponding attractor, i.e. phenotype, careful analysis of these basins could reveal interesting biological characteristics that determine cell fate. In this study we employ a logic reduction algorithm to reduce the Boolean states comprising our basins of attraction to their minimal representations, and it is from these minimizations that we identify intervention targets.

2. Background

Despite its simplicity, the Boolean network model has proven to be quite viable at approximating certain aspects of biological processes. For example, it has been used to simulate the yeast cell cycle,[5] which we look at closely in this work. It has also been used to simulate the expression pattern of segment polarity genes in *Drosophila melanogaster*,[6] as well as the vocal communication system of the songbird brain.[7,8] Since we are investigating within a modeling and simulation framework, we employ the often used assumption of *synchronous update*; however, studies on modeling and analysis of *asynchronous update* in the context of random Boolean networks can be found.[9–12]

Since Kauffman's seminal work there have been countless variations and extensions of the use of Boolean networks for modeling biological systems, and various inference procedures have been proposed for them.[13–15] Shmulevich *et al.*[16] pioneered work on a stochastic extension to the model called probabilistic Boolean networks (PBNs), which share the rule-based nature of Boolean networks but also handle uncertainty. Within this extended framework of PBNs, studies focusing on external system control were performed by Datta *et al.*;[17,18] studies by Pal *et al.*[19] and Choudhary *et al.*[20] explored intervention in PBNs to avoid undesirable states.

One major shortcoming of Boolean networks is the exponential growth of the state space with the number of variables, prompting others to work in the Boolean framework itself to achieve some kind of improvement. The approach of Richardson[21] attempted to shrink the size of the state space through the careful removal of "frozen nodes" and network leaf nodes. The smaller state space then lends itself more readily to the discovery of attractors and basins by sampling methods. Dubrova *et al.*[22] explored properties of random Boolean networks, particularly their robustness in the face of topological changes and the removal of "redundant vertices", thus shrinking the state space. While effective in shrinking the space and removing extraneous nodes, neither of these methods is looking for key players in a system or possible intervention targets; in fact both methods have the chance of eliminating such variables.

In an attempt to achieve certain analysis goals, various authors modified or translated the Boolean formalism into another framework. Saez *et al.*[23] as well as Schlatter *et al.*[24] converted their Boolean models of biological systems into hypergraphs, generalizing graphs with edges connecting sets of vertices instead of just pairs or singletons, thus lending themselves to representing Boolean functions. Both papers use analysis techniques to identify important pathways, network motifs and feedback loops. The work of Schlatter *et al.*also mentions the discovery of relevant hubs in the network. Steggles *et al.*[25] employed a classic concept of converting to a different graphical structure, Petri nets. In making this conversion, they used the logic minimization technique we employ (discussed below), albeit in a different way.

Maji and Pradipta[26] did not use a Boolean network but nonetheless work with the notion of state transition using a related discrete model: fuzzy cellular automata. Their work uses multi-valued logic and presents a new way of identifying attractor basins; however it does not focus on the identification of intervention targets in the system. Mar and Quackenbush[27] also employed the notion of a state transition space without the direct use of a Boolean network. Using their regression model they strive to classify core variables (genes in their case) as they decompose state space trajectories. Their method, however, is dependent on time-course data, and furthermore its primary focus is at the pathway level and not the variable level.

In this work we stay with the classical formulation of Boolean networks but concentrate on the basins of attraction themselves to identify the key variables in the system. While limited by the exponential complexity inherent to Boolean network state spaces, we work here with tractible network sizes and describe plans to expand to larger networks in the future. Recently,[28] we successfully used the same yeast network as this study, a human aging network, as well as a version of the WNT5A network for melanoma also presented here in order to study the planning of interventions in biological networks. The intervention targets selected by the Artificial Intelligence planning techniques in that work are in agreement with intervention targets suggested by the methodology presented in this work.

In the coming sections we first formally define our methodology with a sample network and example. Then, we apply our methodology to a well-studied genetic model of the yeast cell cycle. Following this proof of concept we apply our methodology to a WNT5A network computationally predicted from a melanoma gene expression data set. The reader is also referred to our technical report[29] for an additional application to the aforementioned human aging network. We conclude with some comments on our current and future work.

3. Methods

In this section we formally define our methodology. We first briefly summarize the Boolean network formalism and touch upon a basic description of logic reduction. Finally we discuss some measures used in the identification of important variables and intervention targets and then apply all of this to an example network. The reader is referred to our previous technical report[29] for more on the Boolean network formulation, a smaller example, as well as further description of logic reduction; Xiao and Yufei[30] also add to the description of Boolean networks.

3.1. *Boolean Networks*

A Boolean network $\mathbf{B}(V, \mathbf{f})$ is made of a set of binary nodes $V = \{x_1, x_2, \cdots, x_n\}$, where $x_i \in 0, 1$, and a set of functions $\mathbf{f} = \{f_1, f_2, \cdots, f_n\}$ that define a state of $\mathbf{x}$ at time $(t + 1)$

as $x(t+1) = f_i(x_{i1}(t), x_{i2}(t), ..., x_{ik_i}(t))$, where f_i is a Boolean function and k_i is called the *connectivity* of x_i. The state transition diagram $\mathbf{G}(S, E)$ of a Boolean network $\mathbf{B}(V, \mathbf{f})$ with n nodes is a directed graph where $|\mathrm{S}| = |\mathrm{E}| = 2^n$. Each of the vertices represents one possible configuration of the n variables in the network and each of the directed edges represents the transition between two states as Boolean functions are synchronously applied to all variables.

In the absence of interventions or perturbations, beginning in any initial state, repeated application of transition functions will bring the network to a finite set of states, $\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m\} \subseteq S$ and cycle among them forever in fixed sequence. This set of states is known as an *attractor*, denoted $\mathbf{A}$. An attractor with just one state is called a singleton attractor and an attractor with more than one state is called a cyclic attractor. Boolean networks may have anywhere from one cyclic attractor comprised of 2^n states to 2^n point attractors, although most commonly a network will have just a handful of singleton or short-cycle attractors. The complete set of states from which a network will eventually reach $\mathbf{A}$ is known as the *basin of attraction* for $\mathbf{A}$, denoted $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_M\} \subseteq S$. All attractors are subsets of their basins (i.e. $\mathbf{A}_i \subseteq \mathbf{B}_i, \forall_i$), all basins are mutually exclusive (i.e. $\mathbf{B}_i \bigcap \mathbf{B}_j = \emptyset, \forall_{i,j}, i \neq j$), and the complete state space is comprised entirely of all basins (i.e. $\bigcup_i \mathbf{B}_i = S$). In this study we use the BN/PBN Toolbox[31] for Boolean network simulation and processing.

3.2. *Logic Minimization*

Logic minimization (or reduction) is a classic problem from digital circuit design employed to reduce the number of actual logic gates needed to implement a given function.[32] With careful logic minimization one can reduce the number of gates required and thus include more functionality on a single chip. Minimization identifies variables which have no influence on the outcome of a function and marks them appropriately as a *don't-care*. As a simple example, we take the Boolean function: $(A \wedge B) \vee (\neg A \wedge B)$ (2 signals, 4 gates). Since the role of A changes while B remains *ON* with the same output, it is clear to see that the only influencing variable is B, which can be given with just that signal itself (a single gate).

In this study, we use Espresso,[33] which is a heuristic logic minimizer designed to efficiently reduce logic complexity even for large problems. We supply as input the set of states in a particular basin of attraction ($\mathbf{B}_i$); this input comprises the *ON-cover* (or truth table) in disjunctive normal form (DNF) for a Boolean function whose output is *ON* for the states of $\mathbf{B}_i$ ($\{\mathbf{b}_1 \vee \mathbf{b}_2 \vee \cdots \vee \mathbf{b}_{M_i}\} \mapsto ON$) and whose output is *OFF* for the states of $\mathbf{S} \setminus \mathbf{B}_i$. Espresso analyzes this cover and returns a minimal (though not necessarily unique) DNF set comprised of one or more terms, denoted $\mathbf{T}_i = \{\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_{N_i}\}$, where $N_i \leq M_i$. These $\mathbf{t}_i$ have some variables set to *ON*, some set to *OFF*, and some set as *don't-care*. The presence of these don't-care variables in some terms is what allows the reduction.

3.3. *Measures: Popularity, Term Power and Variable Power*

After applying logic minimization to a set of Boolean functions one is left with a minimal DNF representation comprised of a set of terms containing ones, zeros, and don't-cares. We have shown how to spot important variables in a very small example,[29] but a more formalized method is needed to identify key variables and possible targets for intervention from the minimized terms in larger problems. To this end we introduce three simple measures. The first is to measure how frequently a variable (v) is required to be *ON* or *OFF* across different terms, called *Popularity* (p), and is defined as:

$$p(v) = \frac{z(v)}{N_i}, \tag{1}$$

where $z(v) = \sum_{j=1}^{N_i} I\left(v, \mathbf{t}_j\right)$, N_i is the total number of terms in $\mathbf{T}_i$, and $I(v, \mathbf{t}_j)$ is an indicator function: 1 when v is ON or OFF in $\mathbf{t}_j$, 0 otherwise. Next, we define a measure to identify terms where a few variables demonstrate supremacy over many others. These terms are powerful due to the combinatorial effect of their few set variables. If a five-variable term has one variable set and four listed as don't-cares, that one set variable controls 16 configurations covered by the don't-care variables (half of the state space). This term would be more powerful than a term with two variables set and three don't-cares. Formally, *Term Power* (P_T) is defined as:

$$P_T(\mathbf{t}) = 1 - \frac{1}{n} \sum_{j=1}^{n} I\left(v_j, \mathbf{t}\right), \tag{2}$$

where n is the number of variables in the term (and network). Term Power is used in calculating our third measure. Given the notion of term power, one can also consider variables which preside over powerful terms to be potentially important and powerful intervention targets. *Variable Power* (P_V) of a variable v will be defined as the average term power over the terms in which it is explicitly configured, i.e. v is not don't-care:

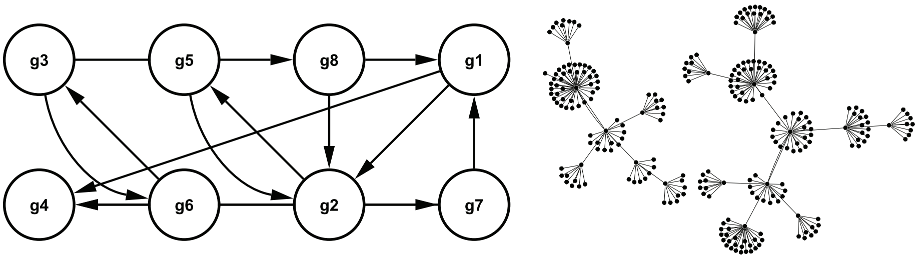
$$P_V(v) = \frac{1}{z(v)} \sum_{j=1}^{N_i} P_T(\mathbf{t}_j) \cdot I\left(v, \mathbf{t}_j\right) \tag{3}$$

3.4. *Other Measures to Identify Key Players*

There are various network centrality measures often used in network studies, particularly concerning biological networks, to identify important variables. We have already touched on the degree of a node, but we also consider the network centrality measures of *betweenness*, *centroid value*, and *eccentricity*. High betweenness indicates that a variable is crucial in maintaining connections between other variables. The centroid value for a variable provides a weighted centrality index. A high eccentricity measure indicates that all other nodes are in proximity. Full definitions as well as biological explanations can be found in the supplementary information of Scardoni *et al.*,[34] but in short, network nodes with high values for these measures can be correlated with biologically significant nodes, possibly even intervention targets.

For Boolean networks, there are also variable-specific measures known as *Influence* and *Sensitivity* for a variable x_i, denoted $r(x_i)$ and $s(x_i)$, respectively. The reader is referred to Shmulevich *et al.*[16,35] for formal definitions. In short, in biological Boolean networks, variables with high influence have the potential to regulate the dynamics of the network, and so they are of interest to this study. Sensitivity represents the degree to which a variable is affected by other variables, and so of the most interest are variables with the highest influence and the lowest sensitivity. Since our measures p and P_V are specific to each basin, this presents an unfair advantage over the network-generality of $r(x_i)$ and $s(x_i)$. Thus, we extend the measures shown in Shmulevich *et al.*[16,35] to be specific to a particular basin of attraction by manipulating the joint probability distribution of the state space; we simply assign a zero probability to any state not in the basin and assign a uniform probability to states within.

Example: An 8-variable Boolean network:



(a) Network (b) State Transition

Fig. 1: Eight-Variable Example Boolean Network

To show these measures and also our claim regarding the utility of our methodology over other measures, we create the 8-variable network shown in Fig. 1(a), in which we assign at most three random inputs and random Boolean functions. Simulation resulted in two basins of attraction, shown in Fig. 1(b). Basin 1 included 160 states converging on a cyclic attractor of length two (`[01011101]` and `[11011100]`), and Basin 2's remaining 96 states converged on another cyclic attractor of length two (`[00011100]` and `[11011101]`). Logic reduction reduced the 160 states in Basin 1 to a set of three terms, and the 96 states of Basin 2 to a set of four terms: $\mathbf{T}_1 = \{$`[0-----00]` $\vee$ `[1-----10]` $\vee$ `[1-----01]`$\}$, $\mathbf{T}_2 = \{$`[1-----00]` $\vee$ `[0-----1-]` $\vee$ `[0------1]` $\vee$ `[------11]`$\}$, where "-" indicates a don't-care.

After analysis with the measures defined in the previous section, we find, based on high p and P_V, $g1$, $g7$ and $g8$ to be of interest. Because each of $g1$, $g7$ and $g8$ are explicitly configured in each of three terms for the larger basin and in 3 out of 4 terms in the smaller basin, their scores for p and P_V are each identical and overshadow the remaining variables. In this example, we again observe that simply identifying vertices in the graph with high degree does not necessarily reveal important variables. With self-loops removed to prevent inflation of degree counts, the variables with the highest degree are $g2$ with six incident edges and $g1$ with four. From our analysis, $g1$ is one of the most important variables. However the variable with the highest degree, $g2$, has been shown to have no influence at all in our analysis. When the network centrality measures of betweenness, centroid value, eccentricity and node degree are calculated for this toy network, we find that $g8$ is frequently reported with high scores, just like our approach. $r(x_i)$, in fact, identifies $g1$, $g7$ and $g8$ as important, which match our three best. However, several of the measures, including $s(x_i)$, incorrectly dismiss $g7$, and many measures also elevate $g2$, which is shown to have no real intervention capabilities. A table of all measures can be found on the supplementary website; an illustrated expansion of this example, along with a simpler one, can be also be found there, and in our previous report.[29]

4. Results

In this section we set out to prove the efficacy of our method on real world examples. For this proof of concept we analyze a Boolean network model of the yeast cell cycle and identify significant variables in the system corroborated by its original manuscript. We then demonstrate our approach on a Boolean network not constructed manually, but rather learned from gene expression data directly. In our technical report[29] we also apply our method to the systems biology of human aging, where we step away from genetic interactions and demonstrate the utility of our method on our Boolean network model for human senescence.

4.1. *Boolean Network Model for Yeast Cell Cycle and Its Analysis*

As a proof of concept on a nonrandom network we will apply our methodology to a well-studied Boolean network model of the yeast cell cycle[5] and show that key variables described in the manuscript are identified by our approach. In their paper, Li *et al.* manually construct a Boolean network modeling the yeast cell cycle using 11 of the most important genes out of the approximately 800 known to play a role in the process. This network is simulated and results in seven basins of attraction, one of which is by far the largest and was studied exclusively in the paper. In this basin of attraction, which included 1,764 states, Li *et al.* were able to trace the trajectory of the yeast cell cycle from one of the fringe, or "Garden of Eden",[4] states down to the eventual point attractor state. The Boolean network adapted from Li *et al.* is shown in on the supplementary website, the original paper,[5] and our technical report.[29]

After applying logic minimization to these 1,764 states we are left with a sum of 39 product terms. An abstraction of these terms can be seen in Table 1. In the table the terms are seen across columns (sorted by P_T), with ones and zeros represented by black and white, respectively, and don't-cares shown in grey. Some variables are set frequently and others are not. Some terms have many requirements, and others have few. The p and P_V measures were calculated for each of the eleven genes in the network. The three most popular variables are Clb5,6, Clb1,2, and Mcm1. The most powerful variable was identified as Cln3.

Table 1: Minimized Yeast Cell Cycle Basin (Black = 1, White = 0, Grey = *don't-care*)

Genes	39 reduced terms across columns																																							p	P_V	s(x)	r(x)
Cln3	□	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	0.05	0.82	0.00	1.00
MBF	▨	▨	■	□	□	□	▨	□	□	■	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	□	□	▨	▨	▨	▨	▨	■	■	0.31	0.63	1.50	0.88
SBF	▨	▨	■	□	□	□	▨	□	□	▨	▨	□	□	□	□	□	□	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	□	□	□	▨	▨	▨	▨	▨	▨	0.38	0.65	1.50	1.50
Cln1,2	▨	▨	▨	▨	▨	▨	▨	□	▨	■	▨	■	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	□	□	▨	■	■	■	▨	■	■	0.28	0.56	1.00	0.56
Cdh1	▨	▨	▨	▨	▨	▨	▨	▨	■	▨	▨	▨	▨	□	▨	▨	▨	□	□	□	□	□	□	□	□	□	□	▨	▨	▨	■	■	▨	▨	▨	▨	▨	▨	▨	0.36	0.62	1.25	0.56
Swi5	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	□	▨	▨	□	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	▨	□	□	▨	▨	□	□	□	0.21	0.55	1.50	0.31
Cdc20	▨	▨	▨	▨	▨	▨	□	■	■	□	□	□	▨	▨	□	□	▨	▨	□	□	□	□	□	▨	▨	▨	▨	□	□	▨	▨	▨	▨	▨	□	▨	□	□	▨	0.46	0.62	1.00	1.75
Clb5,6	▨	▨	▨	■	▨	▨	■	▨	▨	▨	▨	■	■	■	■	■	■	■	■	■	▨	▨	▨	■	■	■	▨	■	▨	■	▨	▨	■	■	■	■	■	▨	▨	0.54	0.62	1.50	1.75
Sic1	▨	▨	▨	▨	▨	▨	□	▨	▨	□	□	▨	□	□	▨	▨	□	□	▨	▨	□	□	▨	□	□	▨	□	▨	□	□	▨	■	□	▨	▨	□	▨	▨	□	0.46	0.62	1.88	1.00
Clb1,2	■	□	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	▨	■	▨	■	▨	■	■	▨	■	■	■	■	■	▨	▨	▨	□	□	□	□	□	□	0.49	0.62	1.88	3.38
Mcm1	▨	▨	▨	▨	▨	■	▨	▨	▨	▨	□	▨	▨	▨	▨	□	▨	▨	▨	■	▨	■	■	▨	■	■	■	■	■	■	▨	▨	□	□	□	□	□	□	□	0.49	0.60	1.00	1.31

Starting with the most popular variable, we find that Clb5,6 is required to be in a particular state 54 percent of the time. Furthermore we find that in each of the 21 terms in which Clb5,6 is in a specific configuration, that configuration is *ON*, or active. Since the Clb5 gene (part of the Clb5,6 variable) is described as being responsible for driving the cell into the S phase (in which the DNA is synthesized and chromosomes are replicated), it seems reasonable to find it strongly represented in the minimized basin. If the role of Clb5 were not known beforehand, analysis of the basin in the manner described could identify it as important (and in the *ON* state) even though it is *OFF* in the eventual attractor state.

Next we look at one of the second-most popular variables in the reduced basin, namely Clb1,2. The Clb2 gene (part of the Clb1,2 variable) is stated as being responsible for the transition in and out of the M phase (in which chromosomes are separated and the cell is divided into two). Thus, like Clb5,6, it is not surprising to find it here among the most frequently specified variables in the basin representing the cell cycle. Unlike Clb5,6, the configuration of Clb1,2 is not consistent—it is found in the *OFF* configuration 7 times and in the *ON* configuration 12 times. However, since it is the activation and subsequent degradation of Clb2 which initiates and terminates the M phase, the split nature of the configurations seems appropriate.

There are other variables with high p which are not explicitly called out in the paper. Given the corroboration of those which are called out in the paper, further investigation of the roles of cyclin inhibitors Cdc20 and Sic1, and of transcription factor Mcm1 is warranted.

Finally we look at the most powerful variable, cyclin Cln3, which was described in the paper as the trigger committing the cell to the division process. Despite its importance, we find it only explicitly configured in 2 of the 39 terms in the reduced basin (once for *OFF* and once for *ON*), which ranks it lowest in the p measure. However, because these two terms are the most powerful, Cln3's P_V score is quickly elevated. It is also interesting to find that in these two terms, only one other variable is specifically configured, namely, Clb1,2. In fact, these two variables are in opposite configurations in these two terms; when Cln3 is *ON*, Clb1,2 is *OFF* and when Cln3 is *OFF*, Clb1,2 is *ON*. This is interesting because Cln3 is described as triggering the G1 phase (the starting phase), and Clb1,2 controls the entry and exit from the M phase (the ending phase). Their opposite configurations in the reduced basin terms seem to agree quite harmoniously with their regulatory control at extreme ends of the cell cycle.

When the network centrality measures of betweenness, centroid value, eccentricity and node degree are calculated for this yeast network, we find that Clb1,2 and Clb5,6 are frequently reported with high scores, just like we find using our approach. This is also the case when $r(x_i)$ is calculated based on the Boolean network properties underlying the topology. However, the centrality measures also report variables such as Clb1,2, SBF and MBF, which are shown mathematically by our method to have little intervention power. Furthermore, these measures give little consideration to other key variables, including Cln3 and Mcm1, which our approach mathematically shows to have some intervention capabilities. Thus, our approach reports the key variables described by Li *et al.* and missed by traditional measures, and avoids reporting mathematically weak variables reported strongly by traditional measures.

4.2. *Application to WNT5A Network for Melanoma*

After applying our approach to a hand-made network, we applied our methodology to a well-studied WNT5A network computationally predicted from a melanoma data set.[36–38] In our previous work,[38] the original data set was narrowed down to the ten most critical variables; these were selected out of 587 total on the basis of their strong interactive connectivity and either their known or likely roles in WNT5A driven induction of an invasive phenotype in melanoma cells, or their close predictive relationship with these genes. For each of the ten variables, we were able to identify the three most ideal predictors out of the remaining nine. Using this connectivity and a binary quantization of the original data set, the best binary logic functions were inferred for each target minimizing the Bayes error.[39,40] From these functions, the Boolean network attractors and basins were identified. The reader is referred to the cited publications for detailed information on the data and connectivity, and to the supplementary website for the functions identified, as well as elucidating figures.

Table 2: WNT5A Basin Attractor States (Black = 1, White = 0) with Basin Measures; $s_i(x)$ and $r_i(x)$ are basin-specific influence and sensitivity, which are discussed in the next subsection

	B1	p	P_V	$s_1(x)$	$r_1(x)$	B2		p	P_V	$s_2(x)$	$r_2(x)$	B3	p	P_V	$s_3(x)$	$r_3(x)$	$s(x)$	$r(x)$
WNT5A	■	0.45	0.57	1.79	2.25	□	■	0.70	0.44	1.72	1.70	□	1.00	0.20	1.00	2.75	1.75	2.00
S100B	□	0.32	0.53	1.03	0.88	□	■	0.40	0.40	0.96	0.59	■	1.00	0.20	1.25	1.25	1.00	0.75
RET1	■	0.23	0.54	0.00	1.22	■	■	0.35	0.46	0.00	1.29	■	0.50	0.20	0.00	1.00	0.00	1.25
MMP-3	■	0.50	0.55	0.00	0.51	■	□	0.60	0.48	0.00	0.47	□	1.00	0.20	0.00	1.00	0.00	0.50
Pho-C	■	0.27	0.53	0.96	0.24	■	■	0.35	0.37	0.52	0.27	■	0.50	0.20	0.50	0.00	0.75	0.25
MLANA	□	0.00	0.00	1.26	0.90	□	□	0.00	0.00	1.24	0.58	□	0.00	0.00	1.00	1.00	1.25	0.75
HADHB	□	0.32	0.50	0.81	0.74	□	□	0.55	0.41	0.64	0.77	□	1.00	0.20	3.00	0.50	0.75	0.75
SNCA	□	0.68	0.54	1.65	0.30	□	□	0.70	0.50	1.31	0.19	□	1.00	0.20	2.50	0.25	1.50	0.25
STC2	■	0.82	0.55	1.31	1.08	■	■	0.75	0.47	1.19	0.92	■	1.00	0.20	1.00	0.25	1.25	1.00
PIR	□	0.86	0.55	1.79	2.48	□	□	0.70	0.49	1.71	2.51	□	1.00	0.20	1.00	3.25	1.75	2.50

The state space (1,024 states) was partitioned into three basins of attraction: Basin 1 had a singleton attractor state with a total basin size of 544 states, Basin 2 has a two-state cyclic attractor with a total basin size of 472 states, and Basin 3 had a singleton attractor with a total basin size of just 8 states. As seen in Table 2, our measures p and P_V reported the intervention capabilities of Pirin, STC2, SNCA, and WNT5A. STC2 is known to interact with MMP-3,[41] another variable in this network, SNCA is known to be aberrantly hypermethylated in human cancer cells,[42] it is known that "cytoplasmic localization of PIR may represent a characteristic of WNT5A network for melanoma progression",[43] and WNT5A has a known role in human melanoma progression.[37] That three of our top four intervention targets are either melanoma-related or cancer-related speaks well for their true intervention capabilities.

When compared to the network centrality measures, as well as $r(x_i)$ and $s(x_i)$, Pirin and WNT5A were identified by most of them. However, also among the high scoring results for these measures was MLANA, which was shown mathematically by our results to have zero influence on the network dynamics. This is not totally surprising, considering this network is derived from melanoma data in which all melanocytes should be present, and that p and P_V are basin-specific (see below). While all variables in such a small, carefully selected set will bear some significance, even MLANA, our approach simply reveals those with true intervention capabilities given the topology. Furthermore, some measures dismissed STC2 and SNCA by including it among the lowest scoring variables despite its influence potential.

4.3. *Usefulness of p and P_V over Other Measures*

We have seen the ability of p and P_V to identify variables with great combinatorial control over the state space of a Boolean network. We have further demonstrated how those variables identified are often known to be suitable targets for intervention. In demonstrating this we have compared p and P_V to $r(x_i)$ and $s(x_i)$, as well as network centrality measures, and here we discuss some differences in these measures.

While $r(x_i)$ and $s(x_i)$ are based on Boolean functions, p and P_V are based on Boolean states. Influence[16] is computed by variable pairs in a matrix and summed by rows and columns to get $r(x_i)$ and $s(x_i)$, where p and P_V are independent measurements on variables and do not depend on pairs. $r(x_i)$ and $s(x_i)$ are general measures, where p and P_V are specific to each basin of attraction. To level the field of comparison, we created a basin-specific version of $r(x_i)$ and $s(x_i)$ ($r_k(x_i)$ and $s_k(x_i)$ for basin k), but they were not able to offer any new insight that $r(x_i)$ and $s(x_i)$ were not already able to. To see this, observe the closeness and value and symmetry in dynamics (based on basin size) between the measurements in Table 2 and in the table on the supplementary website for the human aging network.

There are additional advantages over $r(x_i)$ and $s(x_i)$. p and P_V are not only basin-specific, but they are also value-specific. While we can adapt an influence matrix to be basin-specific, it still cannot be made value-specific. Thus, with p and P_V, because of the minimized terms, we not only know where to intervene, but precisely how to do so. These values, or how we should intervene, can be and often are different than the values in the attractor state (if we're lucky enough to not have a cyclic attractor where values toggle), and furthermore the same target may be viable for more than one basin, but with different values. This kind of information is not available with an influence matrix or the derived measures $r(x_i)$ and $s(x_i)$.

Furthermore, p and P_V allow us to find the minimal effective intervention. Any computational aid to intervention studies will always be human-reviewed in the end, so it need not give one definitive answer. We can say with mathematical certainty that setting certain variables together will force a basin (and thus attractor) to be selected. With a set of minimized terms we can find the smallest interventions (highest P_T) using the most effective targets (high p and/or P_V) which are suitable for intervention with current medical abilities (human evaluation of mathematical possibilities).

5. Conclusion and Future Work

In this paper, we showed the importance of analyzing Boolean network basins of attraction in identifying targets for intervention. Furthermore, we demonstrated that these targets are not always evident in attractor states themselves, in the network topology, or even from various existing measures, both graph-theoretic and Boolean-network-specific. Our use of logic minimization significantly reduces the representation of basins of attraction, and the proposed measures stratify the terms, revealing both the key players and how to manipulate them.

The analysis of the yeast cell cycle network demonstrated that our methodology can identify key variables in the system. We were able to systematically identify three important variables described specifically by the original study and propose others for further study. Our application to the WNT5A network for melanoma demonstrated the applicability of our approach beyond hand-created networks to networks inferred from biological data; furthermore our targets identified for intervention had been previously validated by laboratory studies.

This approach is most appropriate to smaller hand-made or high-confidence networks due to the size complexity issues in Boolean networks. Current efforts involve overcoming the scalability issues inherent in enumerating complete state spaces, which quickly becomes intractable. We are investigating approximation approaches to identify attractor states and enumerate most of their basins. We intend to take full advantage of high performance computing clusters, both in terms of memory and parallelization. We also are working on expanding our implementations and measures to handle multi-valued logic, taking us beyond the Boolean constraint and allowing even more levels of abstraction.

Supplementary Material

http://biocomputing.asu.edu/basinreduction/psb2011/

Acknowledgement

This study funded partly by NIH R21LM009706 (MV, SK) and SFAZ CAA 0243-08 (SK).

References

1. J. Goutsias and S. Kim, *Biophys J* **86**, 1922 (April 2004).
2. S. Kauffman, *Journal of Theoretical Biology* **22**, 437 (March 1969).
3. G. Lima-Mendez and J. Helden, *Mol. BioSyst.* **5**, 1482 (December 2009).
4. A. Wuensche, Genomic regulation modeled as a network with basins of attraction., in *Pacific Symposium on Biocomputing*, 1998.
5. F. Li, T. Long, Y. Lu, Q. Ouyang and C. Tang, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4781 (April 2004).
6. R. Albert and H. G. Othmer, *Journal of Theoretical Biology* **223**, 1 (July 2003).
7. J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink and E. D. Jarvis, *Bioinformatics* **20**, bth448 (July 2004).
8. V. A. Smith, E. D. Jarvis and A. J. Hartemink, *Bioinformatics* **18**, S216 (July 2002).
9. C. Gershenson, Classification of random boolean networks, in *ICAL 2003: Proceedings of the eighth international conference on Artificial life*, (MIT Press, Cambridge, MA, USA, 2003).
10. K. Klemm and S. Bornholdt, *Physical Review E* **72**, 055101+ (Nov 2005).
11. F. Greil and B. Drossel, *Physical Review Letters* **95**, 048701+ (Jul 2005).
12. X. Deng, H. Geng and M. Matache, *Biosystems* **88**, 16 (March 2007).
13. T. Akutsu, S. Miyano and S. Kuhara, *Pacific Symposium on Biocomputing* , 17 (1999).
14. I. Shmulevich, A. Saarinen, O. Yli-Harja and J. Astola, Inference of genetic regulatory networks via best-fit extensions, in *Computational and Statistical Approaches to Genomics*, eds. W. Zhang and I. Shmulevich (Kluwer Academic Publishers, Boston, 2003) pp. 197–210.

15. H. Lähdesmäki, I. Shmulevich and O. Yli-Harja, *Machine Learning* **52**, 147 (July 2003).
16. I. Shmulevich, E. R. Dougherty, S. Kim and W. Zhang, *Bioinformatics* **18**, 261 (February 2002).
17. A. Datta, A. Choudhary, M. L. Bittner and E. R. Dougherty, *Machine Learning* **52**, 169 (July 2003).
18. A. Datta, A. Choudhary, M. L. Bittner and E. R. Dougherty, *Bioinformatics* **20**, 924 (April 2004).
19. R. Pal, A. Datta, M. L. Bittner and E. R. Dougherty, *Bioinformatics* **21**, 1211 (April 2005).
20. A. Choudhary, A. Datta, M. L. Bittner and E. R. Dougherty, *Bioinformatics* **22**, 226 (January 2006).
21. K. A. Richardson, *Advances in Complex Systems* **8**, 365 (2005).
22. E. Dubrova, M. Teslenko and H. Tenhunen, A computational scheme based on random boolean networks, in *Transactions on Computational Systems Biology X*, eds. C. Priami, F. Dressler, O. B. Akan and A. Ngom, 2008) pp. 41–58.
23. J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein, R. Samaga, D. A. Lauffenburger, S. Klamt and P. K. Sorger, *Molecular Systems Biology* **5** (December 2009).
24. R. Schlatter, K. Schmich, I. Avalos Vizcarra, P. Scheurich, T. Sauter, C. Borner, M. Ederer, I. Merfort and O. Sawodny, *PLoS Comput Biol* **5**, e1000595+ (December 2009).
25. L. J. Steggles, R. Banks, O. Shaw and A. Wipat, *Bioinformatics* **23**, 336 (February 2007).
26. P. Maji, *Fundam. Inf.* **86**, 143 (2008).
27. J. C. Mar and J. Quackenbush, *PLoS Comput Biol* **5**, e1000626+ (December 2009).
28. D. Bryce, M. P. Verdicchio and S. Kim, *ACM Transactions on Intelligent Systems and Technology* (To Appear).
29. M. P. Verdicchio and S. Kim, *Reduction of Boolean Network Basins of Attraction Reveals Intervention Targets*, tech. rep., Arizona State University (Tempe, AZ, 2010).
30. Y. Xiao, *Current Genomics* **10**, 511 (November 2009).
31. I. Shmulevich, E. R. Dougherty and W. Zhang, *Bioinformatics* **18**, 1319 (October 2002).
32. A. Marcovitz, *Introduction to Logic Design*, first edn. (McGraw-Hill, Feb 2002).
33. R. L. Rudell and A. L. Sangiovanni-Vincentelli, Espresso-mv: Algorithms for multiple valued logic minimization, in *Proc. of the IEEE Custom Integrated Circuits Conference*, 1985.
34. G. Scardoni, M. Petterlini and C. Laudanna, *Bioinformatics* **25**, 2857 (November 2009).
35. I. Shmulevich and S. A. Kauffman, *Physical Review Letters* **93**, 048701+ (Jul 2004).
36. M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts and V. Sondak, *Nature* **406**, 536 (August 2000).
37. A. T. Weeraratna, Y. Jiang, G. Hostetter, K. Rosenblatt, P. Duray, M. Bittner and J. M. Trent, *Cancer Cell* **1**, 279 (April 2002).
38. S. Kim, H. Li, E. R. Dougherty, N. Cao, Y. Chen, M. Bittner and E. B. Suh, *Journal of Biological Systems* **10**, 337 (2002).
39. E. R. Dougherty, S. Kim and Y. Chen, *Signal Processing* **80**, 2219 (2000).
40. S. Kim, E. R. Dougherty, M. L. Bittner, Y. Chen, K. Sivakumar, P. Meltzer and J. M. Trent, *Journal of biomedical optics* **5**, 411 (October 2000).
41. J. Y. Y. Sung, S. M. M. Park, C.-H. H. Lee, J. W. W. Um, H. J. J. Lee, J. Kim, Y. J. Oh, S.-T. T. Lee, S. R. Paik and K. C. C. Chung, *The Journal of biological chemistry* **280**, 25216 (July 2005).
42. A. Y. Law, K. P. Lai, C. K. Ip, A. S. Wong, G. F. Wagner and C. K. Wong, *Experimental cell research* **314**, 1823 (May 2008).
43. S. Licciulli, C. Luise, A. Zanardi, L. Giorgetti, G. Viale, L. Lanfrancone, R. Carbone and M. Alcalay, *BMC cell biology* **11**, 5+ (January 2010).

MINING THE PHARMACOGENOMICS LITERATURE

KEVIN BRETONNEL COHEN
University of Colorado Denver

YAEL GARTEN
Stanford University

UDO HAHN
Friedrich Schiller University of Jena

NIGAM H. SHAH
Stanford University

Abstract:

The promise of pharmacogenomics for individualized medicine is on the crest of realization as a result of advances that allow us to predict beneficial, non-beneficial, and deleterious drugs for specific individuals based on aspects of both the individual and the drug. In spite of these advances, information management in this field relies on fairly traditional means, which do not scale to the available volume of full text publications. The aim of this workshop is to bring together researchers working on the automatic or semi-automatic extraction of relationships between biomedical entities from the pharmacogenomic research literature. The workshop will focus particularly on methods for the extraction of genotype-phenotype, genotype-drug, and phenotype-drug relationships and the use of the relationships for advancing pharmacogenomic research. Efforts aimed at creating benchmark corpora as well as comparative evaluation of existing relationship extraction methods are of special interest.

Goals of the workshop:

The promise that pharmacogenomic holds for individualized medicine may be on the crest of realization due to technical advances such as large genotyping arrays and analytical advances that allow us to predict beneficial, non-beneficial, and deleterious drugs for specific individuals based on salient features of both the individual and the drug.

However, information management in this field relies on fairly traditional means, e.g. curated databases, which do not scale to (1) the rapid expansion of the pharmacogenomics literature in recent years and (2) the increasingly available volume of full text publications, which contain more specific and (potentially) informative facts than Medline abstracts. Hence, although there is a large demand and significant utility of text analytics to the study of pharmacogenomics, its potential is not fully realized; in part because the work to date has failed to bridge the two distinct worlds—that of (bench) molecular biology and that of (clinically oriented) pharmacology—and because the developers of text analytics are not fully aware of this challenging field.

Last year's workshop (Genotype-Phenotype-Drug relationship extraction from text) examined the current state-of–the-art and reported ongoing research of labs already involved in this area of research. The steady stream of work on extracting interactions from text, the increasing attention in the Semantic Web towards capturing facts as "nano-publications" (individual assertions that are attributable to authors and traceable in their publications), and representing scientific discourse in a structured manner, all indicate that the time seems to be ripe for research that goes even beyond the mere extraction of explicitly stated knowledge in documents, to linking text-mined and database data through formal reasoning to uncover implicit and in some sense "new" knowledge. In order to advance this agenda, it is essential that existing relationship extraction methods be compared to one another and that a community-wide sharable benchmark corpus emerges against which such efforts can be compared. The goal of the workshop is to utilize a corpus put forth by PharmGKB to compare different relationship extraction methods and the corresponding "new" knowledge discovery those methods might drive.

This workshop aims to address the gap in coverage of text mining for pharmacogenomics. The technical area of the workshop will particularly focus on extraction of genotype-phenotype-drug relationships. Work on named entity recognition (e.g. gene taggers) would not be considered for inclusion. Approaches that combine text-mining and knowledge-based systems are of special interest. We invite researchers working on text mining and reasoning to submit applications of their research efforts to the area of pharmacogenomics, and particularly genotype-phenotype-drug relationships. Topics solicited include:

- Relation extraction between genotypes, phenotypes, and drugs, and other semantic classes relevant to pharmacogenomics
- Corpus development for pharmacogenomics text mining
- Associating gene variants (mutations, alleles, rs/ss numbers) to the associated gene name
- Work on the corpus of documents linked to by PharmGKB
- Reasoning systems applied over the PharmGKB knowledge base

We anticipate that this year’s workshop will build on the success of last year’s workshop and seed a community around the shared goal of computationally collecting and distributing pharmacogenomics knowledge.

IDENTIFICATION OF ABERRANT PATHWAY AND NETWORK ACTIVITY FROM HIGH-THROUGHPUT DATA

M. F. OCHS

Departments of Oncology and Health Science Informatics, Johns Hopkins University,
Baltimore, MD 19075, USA
**E-mail: mfo@jhu.edu*

R. KARCHIN

Department of Biomedical Engineering and
Institute for Computational Medicine, Johns Hopkins University,
Baltimore, MD 21218, USA
E-mail: karchin@jhu.edu

H. RESSOM

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University
Washington, DC 20057, USA
E-mail: hwr@georgetown.edu

R. GENTLEMAN

Bioinformatics and Computational Biology, Genentech
South San Francisco, CA 94080, USA
E-mail: gentleman.robert@gene.com

The workshop focused on approaches to deduce changes in biological activity in cellular pathways and networks that drive phenotype from high-throughput data. Work in cancer has demonstrated conclusively that cancer etiology is driven not by single gene mutation or expression change, but by coordinated changes in multiple signaling pathways. These pathway changes involve different genes in different individuals, leading to the failure of gene-focused analysis to identify the full range of mutations or expression changes driving cancer development. There is also evidence that metabolic pathways rather than individual genes play the critical role in a number of metabolic diseases. Tools to look at pathways and networks are needed to improve our understanding of disease and to improve our ability to target therapeutics at appropriate points in these pathways.

Keywords: Signal pathways, metabolic pathways, disease, statistics

1. Introduction

Many complex databases are being developed and maintained to house genetic, epigenetic, genomic, and functional genomic data. Centralized resources such as the National Center for Biomedical Informatics (NCBI) are developing databases to integrate reads from next generation sequencing experiments, tumor-derived somatic DNA sequence variation, and single nucleotide polymorphisms (SNPs) or haplotypes significantly associated with disease phenotypes. Functional genomic data and methylation array data are being captured in the Gene Expression Omnibus (GEO) and ArrayExpress data repositories. The cancer genome atlas (TCGA) combines all these types of data together with detailed information about clinical phenotypes. A vast amount of open-access data now allows data analysts and informaticists the opportunity to develop tools and perform initial demonstrations of their validity independently of new bench experiments. These resources now provide a unique opportunity for the development of tools suitable for analyzing data arising from complex biology.

A key focus in this workshop was the emergence of model-based analysis for high-throughput data. As an example of previous work, Chinnaiyan's group utilized prior knowledge on gene expression and TF binding in prostate cancer to identify a change in a key metabolite associated with prostate cancer progression.[1] Sarcosine was one of many metabolites to show substantial changes in levels during prostate cancer progression, however it is produced by GNMT, a methyl transferase with an androgen receptor binding site upstream. As androgen is known to play an important role in prostate cancer aggressiveness, this allowed prediction that sarcosine might serve as a marker of aggressiveness and potentially even be a driver of such aggressiveness, which was validated in cell line studies. The interactions modeled between the molecular components in this work relied on building a simple mechanistic model of the underlying biology, without which the discovery could not have been made. The focus in this workshop was on efforts to integrate data and build models on a much larger scale.

A particularly promising point of integration is the role of pathways in disease. Biological pathways provide a natural approach to the integration of multiple omics data as well as a means to identify the mechanism through which the effects of mutations, epigenetic variation, protein isoforms, and metabolic changes occur.

2. Pathways in Human Disease

Recognition that biological pathways are critical to understanding human disease emerged along with the elucidation of metabolic and cell signaling pathways by molecular biologists and biochemists. For example, the discovery of the role of MAPK kinases in response to external signals[2] and the later elucidation of the proliferation response due to signaling pathways including these kinases[3] demonstrated the role of pathways in the uncontrolled cell growth that is typical of cancer.[4] Later it was realized that many forms of specific signaling proteins (i.e., different related kinases encoded by different genetic loci) existed, and that each member of a family could substitute for another in specific cell types or be aberrantly expressed in some cancers.[5]

In addition, multiple signaling pathways that play important roles in programmed cell death (PI3K-AKT), proliferation (RAS-RAF), cell cycle (Rb-CDK), DNA damage response (P53), and cell adhesion (FAK) were discovered to play roles in cancer etiology.[6] Each pathway, as with the RAS-RAF-MAPK-ERK pathway, contains multiple signaling proteins, with many proteins having known multiple loci encoding related family members. Overall, this creates a situation in which a single aberrant protein (e.g., an oncogene) in a pathway can activate that pathway inappropriately, leading to escape of a cell from a checkpoint on growth. Effectively, each viable cancer therefore has multiple hits (as first proposed by Knudson[7] for the related case of a dominant tumor suppressor), but the hits may be different (i.e., different pathway members) in each cancer, even for cancers of the same apparent phenotype.

Validation of this new view of cancer came with studies of coordinated methylation, mutation, copy number, and expression changes in glioblastoma multiforme and pancreatic cancer.[8–10] In these studies it was demonstrated conclusively that almost all cancers had changes to one protein in each important pathway, but that these proteins were not the same between different individuals. This result suggested that analysis of pathways would be more informative than analysis of genes across a population.

3. High-Throughput Data

Traditional molecular biology and biochemistry involved detailed study of one or a few genes in tightly controlled experimental systems. This approach changed dramatically with the emergence of gene expression microarrays in the mid 1990's.[11,12] These technologies soon allowed researchers to measure levels of mRNA genome-wide and represented the first of many genome-wide measurement technologies. Subsequent advances since the development of microarrays for gene expression have been very rapid. Tiling arrays and array comparative genomic hybridization (aCGH) have allowed increasingly fine-grained measurements of DNA variations. Use of these arrays and custom arrays coupled with immunoprecipitation permit genome-wide measurement of transcription factor and regulatory factor binding. SNPs are now measured genome-wide as well, and SNP-chips also permit estimation of copy number variation (CNV) at increasingly fine resolution. Recently miRNA chips have been developed, so that the abundance of miRNA families can now routinely be measured for all known miRNAs. Coupling microarray technology to methylation-specific precipitation allows measurement of methylation levels in the genome as well. Next-generation sequencing is replacing some of these technologies, now routinely providing genomic-, epigenomic-, and transcript-level measurements. Emerging technologies in nuclear magnetic resonance and mass spectrometry are beginning to provide large-scale measurements of metabolites and proteins, and antibody and reverse-phase protein arrays have the potential to allow genome-wide measurements of protein levels in a microarray format.

As multiple high-throughput measurements representing different molecular entities (e.g., DNA, mRNA, protein) are now routinely made, methods to integrate the data between these different molecular domains are needed. These can be gene-centric, aligning measurements to the genome for instance, or protein-centric, focusing on protein isoforms and including alternative splicing and post-translational modifications.

4. Analysis Approaches and Tools

The simplest approach to account for the heterogeneity introduced by a pathway effect into analysis of high-throughput data is to realize that only a subset of disease samples may harbor a mutation or change in expression and to generate a statistic to address this. In fact, methods to identify these outlier genes have been developed.[13,14] The next step is to generate a pathway or set statistic to replace single gene statistics, which was the focus of methods now known as gene set analysis.[15] However, a model-based analysis that directly utilizes pathway structures to interpret high-throughput data should provide greater power for biological discovery. The modeling methods discussed in the workshop utilized high-throughput measurements of cell lines, model organisms, and tumors to discover novel insights into biological systems.

Cell lines developed from primary tumors have been among the most important tools for discovering the molecular changes underlying cancer and for drug compound screening. A recent study of 30 breast cancer cell lines used expression and proteomics profiles, along with mutational and copy number variation data to build a discrete, rule-based network signaling model for each cell line,[16] based on the Pathway Logic system.[17] Each model has an initial state that represents all expressed proteins in the cell line. Signaling is represented by rule sets, based on experimentally derived protein-protein interactions, which determine a sequence of model states. This approach involves many simplifying assumptions, in particular discretization of data, i.e. each protein component is either present or absent in each state. However, the simplicity makes the model interpretable and it recaptitulates known breast cancer biology and yields useful new hypotheses about aberrant signaling in breast cancer. For example, model analysis elucidates the role of the gene CAV1 in highly aggressive basal B breast cancers and the relationship of PAK1 to MAPK cascade regulation. In particular, the hypothesized importance of PAK1 led to the discovery that PAK1 over-expression may provide a potential clinical marker for the utility of MEK inhibitors in breast cancer treatment.

Genome-scale studies of primary tumors, in increasingly larger patient cohorts, have become widespread. These studies measure multiple biomolecules in tumor tissue and matched normal samples, including gene expression, copy number variation, somatic mutations, SNPs, and methylation level. The volume and complexity of this data requires new analysis methods to reach translational goals, such as improved prognostics and patient-specific therapies. PARADIGM, a probabilistic graphical model that maps multiple patient-specific genome-scale measurements onto curated cancer-related pathways, can be used to infer which components of a pathway (broadly defined as physical entities, gene families, and abstract processes) are activated with respect to a normal cell.[18] This process yields a matrix of integrated pathway activities (IPAs) for each patient. Based on IPA clustering, clinically relevant subgroups of patients were identified, with the potential for improved stratification of patients for targeted therapeutic regimens.

ResponseNet treats genetic library screening results and transcriptional changes measured by microarray experiments within the context of the relationship between signaling protein interactions and transcriptional regulation, integrating multiple types of data (e.g., microarray, genetic library, ChIP-chip) from different experimental sources. It was used to successfully identify pathways involved with α-synuclein toxicity and genes differentially regulated by these pathways.[19] This approach, however, relies on downstream transcriptional changes to drive discovery, and thus can miss important protein interactions changes that do not drive transcriptional change. An alternative approach, an award gathering Steiner tree, was used to identify changes driven by protein interactions in the yeast pheromone response.[20] The Steiner tree was successful in balancing the introduction of false positive interactions from experimental data with the loss of key interactions.

5. Conclusion

Our understanding of biological processes and their control has led to a model of biology in which biological regulatory and metabolic pathways play the dominant role. Evolution has led to multiple genes in many key families in these pathways, complicating the identification of cell-specific drivers of biological processes. When these drivers are mutated, over-expressed, lost, or replaced by aberrant family members, disease may emerge. Understanding these pathways and identifying the specific members causing disease is critical to elucidating the heterogeneous molecular changes driving disease, identifying subgroups of patients with shared molecular changes, and developing individualized therapies.

References

1. A. Sreekumar, L. M. Poisson, T. M. Rajendiran, A. P. Khan, Q. Cao, J. Yu, B. Laxman, R. Mehra, R. J. Lonigro, Y. Li, M. K. Nyati, A. Ahsan, S. Kalyana-Sundaram, B. Han, X. Cao, J. Byun, G. S. Omenn, D. Ghosh, S. Pennathur, D. C. Alexander, A. Berger, J. R. Shuster, J. T. Wei, S. Varambally, C. Beecher and A. M. Chinnaiyan, *Nature* **457**, 910 (2009).
2. M. H. Cobb, D. J. Robbins and T. G. Boulton, *Curr Opin Cell Biol* **3**, 1025 (1991).
3. G. L. Johnson and R. R. Vaillancourt, *Curr Opin Cell Biol* **6**, 230 (1994).
4. R. Khosravi-Far and C. J. Der, *Cancer Metastasis Rev* **13**, 67 (1994).
5. A. D. Cox and C. J. Der, *Cancer Biol Ther* **1**, 599 (2002).
6. D. Hanahan and R. A. Weinberg, *Cell* **100**, 57 (2000).
7. A. G. Knudson, *Proc Natl Acad Sci U S A* **68**, 820 (1971).
8. S. Jones, X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S. M. Hong, B. Fu, M. T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu and K. W. Kinzler, *Science* **321**, 1801 (2008).
9. D. W. Parsons, S. Jones, X. Zhang, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I. M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, J. Diaz, L. A., J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. Marie, S. M. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu and K. W. Kinzler, *Science* **321**, 1807 (2008).
10. TCGA, *Nature* **455**, 1061 (2008).
11. M. Schena, D. Shalon, R. W. Davis and P. O. Brown, *Science* **270**, 467 (1995).
12. D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton and E. L. Brown, *Nat Biotechnol* **14**, 1675 (1996).
13. J. W. MacDonald and D. Ghosh, *Bioinformatics* **22**, 2950 (2006).
14. R. Tibshirani and T. Hastie, *Biostatistics* **8**, 2 (2007).
15. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church, *Nat Genet* **22**, 281 (1999).
16. L. M. Heiser, N. J. Wang, C. L. Talcott, K. R. Laderoute, M. Knapp, Y. Guan, Z. Hu, S. Ziyad, B. L. Weber, S. Laquerre, J. R. Jackson, R. F. Wooster, W. L. Kuo, J. W. Gray and P. T. Spellman, *Genome Biol* **10**, p. R31 (2009).
17. S. Eker, M. Knapp, K. Laderoute, P. Lincoln, J. Meseguer and K. Sonmez, *Pac Symp Biocomput* , 400 (2002).
18. C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler and J. M. Stuart, *Bioinformatics* **26**, i237 (2010).
19. E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, S. Lindquist and E. Fraenkel, *Nat Genet* **41**, 316 (2009).
20. S. S. Huang and E. Fraenkel, *Sci Signal* **2**, p. ra40 (2009).

WORKSHOP ON THE VALIDATION AND MODELING OF ELECTRON CRYO-MICROSCOPY STRUCTURES OF BIOLOGICAL NANOMACHINES

STEVEN J. LUDTKE

Verna & Marrs McLean Dept. of Biochem. & Mol. Biology, Baylor College of Medicine,
1 Baylor Plaza , Houston, TX 77030, USA
sludtke@bcm.edu

CATHERINE L. LAWSON

Rutgers, The State University of New Jersey, Department of Chemistry & Chemical Biology and Research
Collaboratory for Structural Bioinformatics, 610 Taylor Road Piscataway, NJ 08854, USA
cathy.lawson@rutgers.edu

GERARD J. KLEYWEGT

Protein Data Bank in Europe, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
gerard@ebi.ac.uk

HELEN M. BERMAN

Rutgers, The State University of New Jersey, Department of Chemistry & Chemical Biology and Research
Collaboratory for Structural Bioinformatics, 610 Taylor Road Piscataway, NJ 08854, USA
berman@rcsb.rutgers.edu

WAH CHIU

Verna & Marrs McLean Dept. of Biochem. & Mol. Biology, Baylor College of Medicine
1 Baylor Plaza, Houston, TX 77030, USA
wah@bcm.edu

Electron cryo-microscopy (cryoEM) is a rapidly maturing methodology in structural biology, which now enables the determination of 3D structures of molecules, macromolecular complexes and cellular components at resolutions as high as 3.5Å, bridging the gap between light microscopy and X-ray crystallography/NMR. In recent years structures of many complex molecular machines have been visualized using this method. Single particle reconstruction, the most widely used technique in cryoEM, has recently demonstrated the capability of producing structures at resolutions approaching those of X-ray crystallography, with over a dozen structures at better than 5 Å resolution published to date . This method represents a significant new source of experimental data for molecular modeling and simulation studies. CryoEM derived maps and models are archived through EMDataBank.org joint deposition services to the EM Data Bank (EMDB) and Protein Data Bank (PDB), respectively. CryoEM maps are now being routinely produced over the 3 - 30 Å resolution range, and a number of computational groups are developing software for building coordinate models based on this data and developing validation techniques to better assess map and model accuracy. In this workshop we will present the results of the first cryoEM modeling challenge, in which computational groups were asked to apply their tools to a selected set of published cryoEM structures. We will also compare the results of the various applied methods, and discuss the current state of the art and how we can most productively move forward.

1. Electron Cryo-microscopy

Electron Cryo-microscopy is a versatile experimental technique with several sub-specialties, each of which has its own unique strengths and weaknesses, which must be taken into account when modeling molecular structures or validating results. We briefly introduce each technique, highlighting the most important aspects of each from a modeling perspective.

1.1. *Single Particle Reconstruction*

Single particle reconstruction is the most widely used of the cryoEM methodologies for macromolecular structure determination, responsible for over 80% of the entries in the EMDB (http://EMDatabank.org). In this technique, purified macromolecules in aqueous buffer are vitrified and imaged, yielding images of individual particles in largely random orientations in a layer of vitreous ice. These images are extremely noisy due to the need to avoid radiation damage. They represent a snapshot of the solution conformation at the time of vitrification, thus the particle population includes any structural variability present in solution. Images of tens of thousands to millions of particles are selected and processed using a complex series of algorithms which determines the 3D orientation of each particle, corrects for microscope artifacts, and in certain cases separates the particles into multiple classes based on conformation, ligand binding or other attributes. These particles are then used to produce one or more 3D reconstructions at resolutions as high as 3.5 - 4.5 Å, for example[1-3].

1.2. *Electron Cryotomography*

Historically this technique has been used to for lower resolution studies of cellular architecture and subcellular structures, which, while highly interesting, is not particularly relevant to molecular modeling. Recently, however, a hybrid approach between tomography and single particle reconstruction has gained popularity. In this technique, a tomographic reconstruction is performed on a specimen similar to that used in single particle reconstruction. This provides extremely noisy 3D reconstructions of individual macromolecules, which can then be aligned and averaged in 3D[4]. The advantage over traditional single particle reconstruction is the presence of 3D information for each particle, rather than having only a single 2D projection of each. This 3D information can be used to resolve ambiguities between particle orientation and changes in particle conformation. While this technique is very powerful for studying difficult specimens or specimens displaying structural variability, at present its resolution is limited to ~20-30 Å even in the best cases. Thus, from a modeling perspective it is suitable only for docking large X-ray structure fragments.

1.3. *2D Crystallography*

Electron crystallography was used to elucidate some of the earliest membrane protein structures [5], and still remains a powerful technique for systems that are resistant to 3-D crystallization, but may naturally form 2D arrays. While this technique can produce exceptional resolution in the plane of the crystal, exceeding 2 Å in one case[6], the resolution in the orthogonal direction is necessarily much worse due to the experimental geometry. Nonetheless, this remains a powerful technique that can produce maps amenable to standard X-ray structure model building methods.

1.4. *Helical Reconstruction*

This technique determines the structure of macromolecules arranged in helical arrays. These arrays may be either naturally occurring or 2D crystals that have been formed on the surface of lipid tubes. The advantage of the helical experimental geometry is that a single filament provides images of the target protein and affords full 360-degree tomographic coverage. Until recently, this technique was capable of resolutions well beyond those achieved with single particle reconstruction[7]. However, thanks to recent strides in single particle reconstruction, the gap has narrowed considerably. The best structures using this method have sufficient resolution for traditional X-ray model building methods.

2. Challenges of CryoEM Map Interpretation

Each of the above techniques are powerful, but each also has limitations. Here we will focus on interpretation of maps determined using single particle analysis, since more than 80% of deposited structures have been determined using this method. The fundamental challenge in any single particle analysis project is the high noise level present in the data owing to the need to avoid radiation damage. As the resolution improves, this problem becomes worse, as radiation damage tends to destroy high resolution features first. Single particle reconstruction intrinsically relies on averaging together large numbers of particles. This raises the question of how to assess the interpretability of reconstructed maps. Resolution in this field is a measure of the noise-levels present in the final reconstruction, and is quite distinct from resolvability, which can be adjusted without impacting measured resolution.

The standard resolution metric requires one to split the data into even and odd halves, generate two 'independent' reconstructions, then compare them by Fourier shell correlation (FSC). The resolution is then the point at which the FSC value falls below a threshold value. Unfortunately the FSC is susceptible to overestimation due to noise/model bias[8] and a number of other possible artifacts. While there are rarely any issues with the overall accuracy of a single particle reconstruction, there is some uncertainty over what level of detail in any given structure can be safely interpreted. For example, it is possible to filter a 5 Å resolution map so apparent sidechain densities are visible, but it is almost certain that such densities are simply noise.

Because cryoEM is now able to achieve resolutions that enable molecular interpretations at the near-atomic level, there is a critical need for model data validation tools as well as improved methods for map interpretation.

3. The CryoEM Modeling Challenge 2010

The idea to host a cryoEM modeling challenge (ncmi.bcm.edu/challenge) was developed in order to provide the modeling community with a standard set of maps to test their methods against, enabling comparison of results, and to improve awareness within the cryoEM community of the range of available tools. Unlike a true blind test of the various computational methods as provided by CASP (www.predictioncenter.org/casp9), the modeling challenge utilized known structures and challenged any interested groups to apply their methods to one or more of the structures, with the goal of improving existing map interpretations or developing new tools for map/model validation. The provided maps cover a range of different symmetries, particle sizes, resolutions and experimental methods. The challenge will conclude at the beginning of December, 2010, and results will be presented and discussed in this PSB workshop. After the conclusion of the workshop, all submitted results will be made permanently accessible to the public .

4. Available Modeling Techniques

There are many different computational techniques that can be used to interpret cryoEM maps, depending on the resolution range the map falls in. Each of the following sections describes a category of possible submissions in the cryoEM Modeling Challenge.

4.1. *Volume Interpretation*

Volume interpretation represents a class of techniques that can be applied to structures where the resolution is insufficient for molecular modeling approaches. We have divided these techniques into three broad categories. The first category is map segmentation; separating a map into meaningful sub-regions. Segmentation may be accomplished in a variety of ways, depending on the available information. For example, if crystal structures of domains or components of the map are available, they can be docked into the cryoEM reconstruction. De novo segmentation methods may attempt to perform automated segmentation based on, for example, the location of low-density regions combined with the symmetry of the structure. Validation of results remains a major issue for this technique.

The second technique is secondary structure element annotation. At sub-nanometer resolution, α-helices become resolvable, and as the resolution improves further, β-sheets become discernible, eventually showing strand separation. In this intermediate (~5-10 Å) resolution range, tools for automatic identification and localization of secondary structure elements become quite valuable, but again, in marginal cases there are validation issues. In addition, in this resolution range, it becomes possible to dock crystal structures with much higher levels of confidence.

The final technique in this class is Cα protein backbone tracing. In the 3.5-5 Å resolution range, it is often possible to perform unambiguous tracing of the protein backbone directly from the density map. Some methods for achieving this rely on additional information, such as sequence-based secondary structure prediction or the existence of a crystal structure of a homologue, to help resolve ambiguities.

4.2. *Modeling*

These methods yield true atomistic models derived from cryoEM density maps. The first of the three methods in this class is related to rigid-body docking described above. The implementation of this method may take many forms, and some methods are resolution-dependent. In many cases where flexible modeling is considered impractical, larger models will be broken into domains, for example at hinge points, to attempt to elucidate more information about differences between the cryoEM structure and the model. Variations of this method have been used in cryoEM for decades, even on structures at very low resolutions. Once again, the major difficulty lies in establishing the reliability of the final results.

The second class of modeling techniques comprises flexible docking methods. Rather than simply finding the best 3D position and orientation for an atomistic model within a cryoEM map, in this method the atomic positions are locally adjusted to better match the experimental data. This can be used to model structures in various conformational states, or can make corrections to homology models. However, again there are serious questions related to the level of detail at which such flexible docking can be trusted. For example, at ~8 Å resolution, α-helices are clearly resolved, but if the flexible fitting were to try to use the density map to modify sidechain orientations, the results would obviously be invalid. Groups developing these techniques are working to establish how to balance molecular modeling energy functions against the need to match the information content of the experimental data.

The final technique is true ab initio modeling based on cryoEM maps. This includes established methods for model building in X-ray crystallography. Since the typical resolution of cryoEM experiments is still below the levels typical for crystallographic studies, new techniques are being developed that hopefully allow for accurate model building at lower resolution. An important point, however, is that the two techniques are not entirely the same. While cryoEM and X-ray crystallography both produce density maps, the specific artifacts (e.g. image distortion and image alignment errors in the case of cryoEM and model bias in crystallography) present in each are not necessarily the same, and the definitions of resolution used in the two communities are not entirely compatible. Accurate atomistic modeling has been performed on a number of cryoEM maps at ~4 Å resolution, a resolution that is generally regarded as marginal in X-ray work.

5. Conclusions

As of September 2010, there were over 50 registered participants in the modeling challenge. Many of the major modeling groups using physics and statistical based simulation and cryoEM density restraints are actively participating and applying their methods to the six cryoEM targets selected for the challenge. Many other groups are applying their tools for specific aspects of cryoEM map analysis for segmentation, secondary structure element identification and de novo modeling. Representatives from several groups have been invited to present their work at the workshop, and there will also be a panel discussion of the results.

6. Acknowledgements

This workshop and the associated challenge is supported by NIH grants P41RR02250 and R01GM079429. We would like to thank all of the challenge participants for their contributions to this project.

References

1. X. Zhang, L. Jin, *et al.*, *Cell* **141**, 472-82 (2010).
2. Y. Cong, M. L. Baker, *et al.*, *Proc Natl Acad Sci U S A* **107**, 4967-72 (2010).
3. J. Z. Chen, E. C. Settembre, *et al.*, *Proc Natl Acad Sci U S A* **106**, 10644-8 (2009).
4. J. Walz, D. Typke, *et al.*, *J Struct Biol* **120**, 387-95 (1997).
5. R. Henderson, P. N. Unwin, *Nature* **257**, 28-32 (1975).
6. T. Gonen, Y. Cheng, *et al.*, *Nature* **438**, 633-8 (2005).
7. C. Sachse, J. Z. Chen, *et al.*, *Journal of molecular biology* **371**, 812-35 (2007).
8. A. Stewart, N. Grigorieff, *Ultramicroscopy* **102**, 67-84 (2004).
9. C. Lawson, M. Baker, et al., *Nucleic Acids Research*, in press (Jan. 2011)