Novartis Foundation Symposium 254

IMMUNOINFORMATICS: BIOINFORMATIC STRATEGIES FOR BETTER UNDERSTANDING OF IMMUNE FUNCTION

2003

John Wiley & Sons, Ltd

IMMUNOINFORMATICS:
BIOINFORMATIC STRATEGIES FOR BETTER UNDERSTANDING OF IMMUNE FUNCTION

The Novartis Foundation is an international scientific and educational charity (UK Registered Charity No. 313574). Known until September 1997 as the Ciba Foundation, it was established in 1947 by the CIBA company of Basle, which merged with Sandoz in 1996, to form Novartis. The Foundation operates independently in London under English trust law. It was formally opened on 22 June 1949.

The Foundation promotes the study and general knowledge of science and in particular encourages international co-operation in scientific research. To this end, it organizes internationally acclaimed meetings (typically eight symposia and allied open meetings and 15–20 discussion meetings each year) and publishes eight books per year featuring the presented papers and discussions from the symposia. Although primarily an operational rather than a grant-making foundation, it awards bursaries to young scientists to attend the symposia and afterwards work with one of the other participants.

The Foundation's headquarters at 41 Portland Place, London W1B 1BN, provide library facilities, open to graduates in science and allied disciplines. Media relations are fostered by regular press conferences and by articles prepared by the Foundation's Science Writer in Residence. The Foundation offers accommodation and meeting facilities to visiting scientists and their societies.

Information on all Foundation activities can be found at
http://www.novartisfound.org.uk

Novartis Foundation Symposium 254

IMMUNOINFORMATICS: BIOINFORMATIC STRATEGIES FOR BETTER UNDERSTANDING OF IMMUNE FUNCTION

2003

John Wiley & Sons, Ltd

Published in 2003 by John Wiley & Sons Ltd,
The Atrium, Southern Gate,
Chichester PO19 8SQ, UK

National 01243 779777
International (+44) 1243 779777
e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on http://www.wileyeurope.com
or http://www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Novartis Foundation Symposium 254
viii+263 pages, 32 figures, 11 tables

Library of Congress Cataloging-in-Publication Data

Immunoformatics : bioinformatic strategies for better understanding of immune function / [editors, Gregory Bock and Jamie Goode].
p. cm. – (Novartis Foundation symposium ; 254)
Includes bibliographical references and index.
ISBN 0-470-85356-5 (alk. paper)
1. Immunoinformatics. I. Bock, Gregory. II. Goode, Jamie, III. Series.
QR182.2.I46I46 2003
571.9′6–dc22 2003057599

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 470 85356 5

Typeset in 10½ on 12½ pt Garamond by Dobbie Typesetting Limited, Tavistock, Devon.
Printed and bound in Great Britain by T. J. International Ltd, Padstow, Cornwall.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry, in which at least two trees are planted for each one used for paper production.

Contents

Symposium on Immunoinformatics: bioinformatic strategies for better understanding of immune function, held at the Novartis Foundation, London, 8–10 October 2002

Editors: Gregory Bock (Organizer) and Jamie Goode

This symposium is based on a proposal made by Nikolai Petrovsky and Vladimir Brusic

Participants

Stephan Beck Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Massimo Bernaschi IAC 'Mauro Picone' (C.N.R.), Viale del Policlinico 137, I-00161 Rome, Italy

Francisco Borras-Cuesta Department of Internal Medicine, School of Medicine, University of Navarra, Irunlarrea 1, 31008 Pamplona, Spain

Vladimir Brusic Knowledge Discovery Department, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, 119613, Singapore

Annie De Groot Brown University, TB/HIV Research Laboratory, Brown University, Box G, Providence, RI 02912 , USA

Charles DeLisi Center for Advanced Genomic Technology, Boston University, 1st Floor, Room 102, 48 Cummington Street, Boston, MA 02215, USA

Darren R. Flower Bioinformatics Group, The Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berkshire RG20 7NN, UK

Kamalakar Gulukota gvk bioSciences Private Limited, #210, 'My Home Tycoon', 6-3-1192, Begumpet, Hyderabad 500 016, India

Paul Kellam Virus Genomics and Bioinformatics Group, Department of Immunology & Molecular Pathology and Department of Virology, Windeyer Institute of Medical Sciences, Windeyer Building, 46 Cleveland Street, London W1T 4JF, UK

Can Kesmir Department of Theoretical Biology, Utrecht University, Padualaan 8, 3584 CH, Utrecht, Netherlands

Marie-Paule Lefranc IMGT, the international ImMunoGeneTics information system®, Université Montpellier II, Laboratoire d'ImmunoGénétique

Moléculaire, LIGM, UPR CNRS 1142, Institut de Génétique Humaine, 141 rue de la Cardonille, F-34396 Montpellier Cedex 5, France

Tim Littlejohn Biolateral, PO Box A51, Enfield South, NSW, 2133, Australia

Terry Lybrand Department of Chemistry, Vanderbilt University, Center for Structural Biology, 5142 Biosci/MRB III, Nashville, TN 37232-8725, USA

Hanah Margalit Department of Molecular Genetics & Biotechnology, Hebrew University Hadassah Medical School, PO Box 12272, Ein Kerem, Jerusalem 91120, Israel

Steven G. E. Marsh Anthony Nolan Research Institute, Royal Free Hospital, Pond Street, Hampstead, London NW3 2QG, UK

Alan S. Perelson MS K710, T-10, Theoretical Division, Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM 87545, USA

Nikolai Petrovsky Canberra Hospital, Autoimmunity Research Unit, PO Box 11, 2606 Woden, ACT, Australia

Hans-Georg Rammensee (*Chair*) Interfakultäres Institut für Zellbiologie, Abteilung Immunologie, Universität Tübingen, Auf der Morgenstelle 15, D-72076 Tübingen, Germany

Lukas Roth Novartis Pharma, Transplantation Research, WSJ-386.9.26, CH-4002 Basel, Switzerland

Diego Silva (*Novartis Foundation Bursar*) Autoimmunity Research Unit, The Canberra Hospital, Canberra, ACT 2065, Australia

Christian Schönbach RIKEN Genomic Sciences Center, Biomedical Knowledge Discovery Team, E-209, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa, 230-0045, Japan

Stefan Stevanović Interfakultäres Institut für Zellbiologie, Abteilung Immunologie, Universität Tübingen, Auf der Morgenstelle 15, D-72076 Tübingen, Germany

Edgar Wingender GBF-Braunschweig, Genome Analysis, Mascheroder Weg 1, D-38124 Braunschweig, Germany

Dominik Wodarz Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP-655, Seattle, WA 98109-1024, USA

Chair's introduction

Hans-Georg Rammensee

Interfakultäres Institut für Zellbiologie, Abteilung Immunologie, Universität Tübingen, Auf der Morgenstelle 15, D-72076 Tübingen, Germany

This is a timely meeting. Although Vladimir Brusic's opening paper is titled 'Immunoinformatics — the new kid in town', this is actually a field that has been around for a while, although under a different name. At least part of what we know of as immunoinformatics was previously known as 'theoretical immunology'. There was an important meeting on this subject in New Mexico in 1988, which resulted in a two-volume book (Perelson 1988).

The subject of immunoinformatics as we see it today can roughly be divided into three areas: the hard, the soft and the semi-soft. A challenge for this group is to decide by the end of the meeting whether I am correct with this classification! Let me start with a description of hard immunoinformatics. This contains what I will call 'hard facts': DNA, RNA and peptide sequences that we can write down. This part of immunoinformatics can be used for a growing number of applications that will have a direct impact on biomedicine. One example is peptides for T cell recognition, working out which peptides are recognized by the T cell receptor during an infection. Hard immunoinformatics is one of the newest parts of the field and is only a few years old. The amount of information in this realm is growing exponentially. 15 years ago all we had were a few DNA sequences, but now we have a tremendous amount of data stored in various databases.

Semi-soft immunoinformatics comprises algorithms and parameters which we use to create the 'hard' part. It includes all the prediction algorithms we use in DNA or peptide sequences: we say that a particular DNA sequence will interact with some regulatory protein or this piece of protein sequence will interact with the MHC. The one hallmark of this semi-soft area is that all the predictions can be tested accurately. You can predict the peptide sequence to bind to HLA, and then go on and test whether this is true. Some of the predictions will be correct and others won't. At one point, though, we may get to a stage where we can omit the verification of the prediction by experiment. I personally think this will never be the case, and we will always have to verify our predictions, but others may disagree.

Then we come to the soft part of immunoinformatics. This is I would to define as something that can never be tested with hard facts. This may raise some

controversy. I would classify this part of immunoinformatics as what has previously been known as 'theoretical immunology'. This includes mathematical descriptions of the behaviour of populations, whether this is at the level of the individual, or at cellular or antibody levels. It involves interactions between antibodies, infectious agents and T cells. I would like to propose that these kinds of models will stay soft because it is not possible to verify the predictions experimentally. If you predict that you need 30 T cells in a human to start an efficient immune response against a viral infection using mathematical modelling, you will never be able to prove this. On the other hand, while these predictions cannot be tested accurately, they can certainly be of help. For example, if one can calculate in a mathematical model the percentage of people that need to be immunized against measles to avoid an epidemic, this will be of great use.

So I propose that it is useful to break down immunoinformatics into these three categories of hard, semi-soft and soft. At the end of the meeting we can discuss whether or not my proposal is correct. Two important questions related to this are whether soft immunoinformatics can ever be tested accurately, and whether the predictions from semi-soft immunoinformatics can stand alone without experimental verification. Let's now move to the first presentation.

Reference

Perelson AS (ed) 1988 Theoretical immunology. Proceedings of the Theoretical Immunology Workshop, June 1987, Santa Fe, New Mexico. Addison-Wesley, Reading, MA

Immunoinformatics — the new kid in town

Vladimir Brusic*† and Nikolai Petrovsky†‡

**Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, †Centre for Medical Informatics, Division of Science and Design, University of Canberra, Bruce ACT 2617 and ‡National Health Sciences Centre, Canberra Clinical School, Woden ACT 2606, Australia*

Abstract. The astounding diversity of immune system components (e.g. immunoglobulins, lymphocyte receptors, or cytokines) together with the complexity of the regulatory pathways and network-type interactions makes immunology a combinatorial science. Currently available data represent only a tiny fraction of possible situations and data continues to accrue at an exponential rate. Computational analysis has therefore become an essential element of immunology research with a main role of immunoinformatics being the management and analysis of immunological data. More advanced analyses of the immune system using computational models typically involve conversion of an immunological question to a computational problem, followed by solving of the computational problem and translation of these results into biologically meaningful answers. Major immunoinformatics developments include immunological databases, sequence analysis, structure modelling, mathematical modelling of the immune system, simulation of laboratory experiments, statistical support for immunological experimentation and immunogenomics. In this paper we describe the status and challenges within these sub-fields. We foresee the emergence of immunomics not only as a collective endeavour by researchers to decipher the sequences of T cell receptors, immunoglobulins, and other immune receptors, but also to functionally annotate the capacity of the immune system to interact with the whole array of self and non-self entities, including genome-to-genome interactions.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 3–22

Biotechnology has provided methods and instrumentation for analysis and manipulation of biological systems on a massive scale. Information technology has provided hardware and software that enable data processing at an unprecedented speed and efficiency. Bioinformatics, defined as the storage, manipulation and interpretation of biological data (MacLean & Miles 1999), has emerged at the interface of life and information sciences. Bioinformatics has evolved as a crucial methodology in genomics, proteomics, and structural

biology. Immunoinformatics (also known as computational immunology) is a subset of bioinformatics focusing on the field of immunology. Immunoinformatics applications are increasingly becoming important to immunological research. The major findings of structural, functional and regulatory aspects of molecular immunology, coupled with the rapid accumulation of immunological data have been complemented by the development of more sophisticated computational solutions for immunology research.

Immunology is essentially a combinatorial science. The diversity in the human immune system is enormous — the total number of combinatorial arrangements of immunoglobulins (Ig) in an individual is greater than 10^9 (Jerne 1993). The T cell receptor (TCR) diversity in humans has been estimated (Arstila et al 1999) at between 10^7 and 10^{15} different clonotypes. There are approximately 10^{12} B cell clonotypes in an individual human (Jerne 1993). More than 500 allelic variants of class I human histocompatibility complex (MHC) molecules characterized to date allow theoretically more than 10^{13} class I haplotypes. The theoretical number of linear epitopes composed of nine amino acids, common targets in cellular immunity, is of the order 10^{11}. The number of conformational epitopes is far higher. These crude numbers, reflecting the complexity of the immune system in a very simplistic manner, indicate its enormous diversity. This diversity underpins our ability to discriminate between friend (self) and foe (non-self) and mount appropriate immune responses. Additional information includes multi-step processing pathways, network-type interactions, complex signalling and mechanisms for modulation of immune responses. Currently available data represent only a tiny fraction of possible situations and the amount of information will keep growing. With the steadily increasing amount of immunological information our ability to decipher the specific mechanisms of immune responses or correct undesirable immune responses is increasingly dependent on exploiting immunoinformatics strategies.

A major role of immunoinformatics is the management and analysis of immunological data with the basic infrastructure comprising numerous immunology database systems (Brusic et al 2000). Immunology databases provide access to, data extraction from, and analysis of immunological data. Standard bioinformatics methods, e.g. sequence analysis (Foster & Chanock 2000) and structural methods, e.g. structure modelling (immunoglobulin, Martin et al 1989; MHC, Schueler-Furman et al 1998, Rognan et al 1999; or TCR, Garcia et al 1998) are routinely applied to immunology studies. More advanced analyses of the immune system using computational models typically involve conversion of an immunological problem to a computational one, solving the computational problem, and translating the results into biologically meaningful interpretations. Examples include data-driven modelling of peptide binding to MHC molecules (Brusic et al 2001), theoretical modelling and complex analysis of the immune

system (Perelson 1989, Kepler & Perelson 1993), and statistical support for immunological experimentation (Merrill 1998). Virtually every aspect of immunology research uses some form of immunoinformatics. The appropriate use of informatics techniques has potential, as supported by examples of practical applications, to vastly improve the efficiency of immunology research. Complete genomes of more than 900 viruses and more than 80 microbes have been sequenced to date (Wheeler et al 2002). High-throughput approaches such as microarray technology (Glynne & Watson 2001), proteomics (Marshall & Williams 2002) and large-scale T cell epitope screening (Schönbach et al 2002) provide for genomic-scale screening and study of the immune system, and its role in beneficial and pathological immune responses. Practical immunoinformatics applications include screening of genomes for vaccine components (De Groot et al 2002), disease-specific gene expression (Saito 2001), studies of cell differentiation pathways, tolerance/immunity decision process and B cell transformation (Glynne & Watson 2001), antibody recognition site identification (Yoshimori & Del Carpio 2001), and integration of data into high level models of the immune system (Yates et al 2001). In the following sections we describe the status and challenges within the subfields of immunoinformatics and discuss the prospects for future developments.

Immunoinformatics

The immune system is intertwined with all other body systems. Bioinformatics applications are relatively well developed for some immunological areas, such as databases (Brusic et al 2000), genomic applications (Glynne & Watson 2001), study of T cell epitopes (Brusic & Zeleznikow 1999), or modelling immune responses (Bernaschi & Castiglione 2002). In other fields of immunology bioinformatics applications are still in their infancy, such as analysis of allergenicity of proteins (Gendel 2002) or proteomics (Klade 2002). Because of the combinatorial nature of immunological data, the importance of efficient, accurate and comprehensive use of immunoinformatic tools will continue to grow in importance for support of immunology research.

Immunological databases

Both molecular biology and immunology produce large amounts of data that have to be stored in general-purpose and specialist immunological databases. General-purpose biological databases contain annotated entries of biological sequences. These entries typically contain the sequence, a short description, the source organism, a list of structural or functional features and literature references. The major public databases include the nucleotide or protein

sequence databases GenBank/GenPept (*www.ncbi.nlm.nih.gov/Genbank/index.html*), EMBL/TrEMBL (*www.ebi.ac.uk/embl*), DDBJ/DAD (*www.ddbj.nig.ac.jp*), PIR (*www-nbrf.georgetown.edu*), SWISS-PROT (*www.expasy.ch/sprot*), PDB (*www.rcsb.org/pdb*), PROSITE (*www.expasy.ch/prosite*) and KEGG (*www.genome.ad.jp/kegg/kegg2.html*). The nucleotide databases — Genbank, EMBL and DDBJ — focus on collecting, annotating, and providing access to the entries of DNA sequences and the related information. GenPept, TrEMBL and DAD are protein databases derived from the translations of coding sequences of the three main nucleotides databases. SWISS-PROT and PIR are protein databases that are manually annotated. Their content is of higher quality than GenPept, TrEMBL and DAD, but they contain fewer entries. PDB is a database of 3D molecular structures. The PROSITE database contains biologically significant patterns and motifs. The KEGG databases comprise repositories on molecular interaction networks, chemical compounds and reactions relevant to cellular processes, and genomics data.

General-purpose databases contain large numbers of immunologically relevant entries and are invaluable resources, therefore, for immunology research. They do not, however, provide sufficient detail on immunological function. Specialist immunology databases provide more detailed information on immunologically relevant molecules, systems and processes. They are typically annotated by experts and contain immunology-specific annotations. Kabat database (*kabatdatabase.com*) contains entries of proteins of immunological interest: Ig, T cell receptors (TCR), major histocompatibility complex (MHC) molecules and other immunological proteins. The IMGT databases (*imgt.cines.fr*) contain high-quality annotations of DNA and protein sequences of Ig, TCR and MHC. They also contain IMGT-related genomic and structural data. The FIMM database (*sdmc.lit.org.sg/fimm*) focuses on protein antigens, MHC molecules and structures, MHC-associated peptides and relevant disease associations. The SYFPEITHI database (*syfpeithi.bmi-heidelberg.com*) contains entries of MHC ligands and peptide motifs. The HIV molecular immunology database (*hiv-web.lanl.gov/immunology*) is an annotated searchable repository of HIV1 T cell and B cell epitopes. More detailed reviews of important immunological databases and related issues can be found in (Brusic et al 2000, Petrovsky & Brusic 2002). The important database issues relate to data standardisation, data quality, interpretation of database entries, and the quality of computational tools for data extraction and analysis (Petrovsky & Brusic 2002), which will be discussed later in this text.

Bioinformatics applications to the study of T cell epitopes

The identification of T cell epitopes relies heavily on bioinformatics for initial screening followed by experimental validation. MHC molecules bind short peptides produced mainly by intracellular (MHC class I) and extracellular (MHC

class II) degradation of proteins and display them on the cell surface for recognition by the T cells (using TCRs) of the immune system. Binding of peptides to the MHC molecule is a prerequisite for immune recognition, but the number of peptides that can bind to a specific MHC molecule is limited. Peptides that bind specific MHC molecules are involved in initiation and regulation of immune responses. Determining peptides that bind specific MHC molecules is important for understanding immunity and has applications to vaccine discovery and design of immunotherapies. The combinatorial nature of this problem makes computational approaches necessary for systematic mapping of T cell epitopes.

Prediction methods are based on binding motifs (Rammensee et al 1999), quantitative matrices (Parker et al 1994) or higher complexity prediction models such as artificial neural networks (ANN) (Brusic et al 2001), hidden Markov models (HMM) (Brusic et al 2002) or molecular modelling (Schueler-Furman et al 1998, Rognan et al 1999). The binding motif describes amino acids commonly occurring at particular positions within peptides that bind to a specific MHC molecule. Quantitative matrices provide coefficients for each amino acid and each position within the peptide that can be used with appropriate formulae to calculate scores that predict peptide binding. The artificial intelligence methods of ANNs and HMMs are based on higher order models that can capture non-linear dependencies in the data sets. The data-driven models (binding motifs, quantitative matrices, ANNs and HMMs) are derived from experimental data sets and can be used for large-scale screening of potential vaccine components (Schönbach et al 2002, De Groot et al 2002). The important property of these models is that each binding motif can be encoded as a quantitative matrix, and each quantitative matrix can, in turn, be encoded as an ANN or a HMM. The accuracy of data-driven methods depends on the complexity of the model relative to the complexity of the peptide–MHC interaction, and on the quantity and representativeness of the data available for building a particular model. Molecular modelling methods utilise comparative modelling where known crystal structures and protein-peptide interactions are used as templates for building 3D models of molecular structures. If initial structural data are not available, *ab initio* modelling based on atomic simulations and residue statistics can be used. Molecular modelling is useful for detailed analysis of specific 3D structures and interactions, but being computationally intensive it is less useful for large-scale screening. Molecular modelling can be used for building complex data-driven methods, such as those for prediction of promiscuous MHC-binding peptides (Brusic et al 2002), or quantitative structure–activity relationships (QSAR) for vaccine discovery (Doytchinova & Flower 2002). The main issues for prediction of MHC-binding peptides are the quality, quantity, and representativeness of data available for model development, the complexity of

the selected predictive model relative to the natural complexity of the peptide–MHC interaction and the training and testing of the predictive model.

Mathematical modelling of the immune system

Observations of immune responses and cellular interactions at the organism level produce definite measurements, but are difficult to interpret at the molecular level. An example is the idiotypic network theory (Jerne 1993) which can be translated into speculative explanations at the molecular level. Mathematical modelling implemented as computational programs can easily translate speculative hypotheses into quantitative descriptions (Perelson 1989). The parameters of the mathematical models can easily be tuned to represent real behaviour of the immune system. These models can then be used for determining the framework for study of the kinetics of immune responses and practical applications such as prediction of immune interventions. Mathematical models of the immune system can model interactions of a large number of elements (10^6 or higher) thereby approaching the complexity of the human immune system. Remarkably accurate simulations using mathematical models have been developed for study of B cell (Kepler & Perelson 1993) and T cell responses (Coussens & Nobis 2002). More specific examples (Yates et al 2001) include modelling of tumour necrosis factor oscillations in allografts, differentiation of T helper cells (Th1/2), modelling T cell memory and cross-talk between TCRs.

Systemic level mathematical models provide a framework for understanding of the immune system as whole. We foresee the convergence of mathematical models at the systemic and molecular level in the future. Huge experimental data sets produced by genomics, proteomics and molecular biology efforts will ultimately be integrated with mathematical models of the immune system at the organism level to produce models of whole organism.

Emerging applications of immunoinformatics

Genomics focuses on the study and characterization of the complete set of DNA sequences (genome) from an organism. Similarly, proteomics focuses on study and characterization of the full protein complement of the genome. Following successful integration of bioinformatics in various fields of molecular biology, notably genomics and proteomics, immunoinformatics is the next frontier, namely the integration of bioinformatics with immunology. A major function of the immune system is to help the organism maintain homeostasis while interacting with self and foreign entities. Beneficial immune responses are targeted towards maintaining homeostasis, while pathological immune responses result in disease states, such as allergies or autoimmunity. The emerging field of immunomics

encompasses the genomics and proteomics of the immune system (Glynne & Watson 2001, Marshall & Williams 2002, Saito et al 2001, Coussens & Nobis 2002, Zagursky & Russell 2001). Immunomics focuses not only on deciphering the sequences of immunoglobulins and various cellular receptors, but is also instrumental for functional annotation of the immune system interactions with the whole array of self and foreign entities, including complete genome-to-genome interactions. Examples of fields that are expected to show rapid growth are immunoinformatics of disease (allergies, cancer, autoimmunity, infectious diseases), host–pathogen interactions, animal immunology, improved predictions of organ rejections, cytokine signalling and other regulatory network analysis, among others. In respect of development of immunoinformatics tools, we expect to see the integration of immunological databases with generic interfaces and ultimately the integration of system level mathematical models with molecular level models leading to applications in the development of novel therapeutic regimens and disease management.

Unifying concepts

The main issues that need to be resolved are those of common data standards, data quality and the accuracy of computational methods. These issues are critical for establishing a common immunoinformatics platform and enabling efficient and adequate use of immunoinformatics resources.

Standardization

Biochemical and molecular biology terms have been standardized by nomenclature committees, such as IUPAC/IUBMB (*www.chem.qmw.ac.uk/iubmb/nomenclature*). The gene ontology consortium (*www.geneontology.org*) has produced a dynamic controlled vocabulary of genes and proteins that can be applied to all organisms in rapidly changing environments. The immunogenetics ontologies and nomenclature for immunoglobulins have been defined recently (Ruiz & Lefranc 2002) and are accessible at the IMGT database. The HLA nomenclature system has been well-defined and accepted (*www.anthonynolan.org.uk/HIG/nomen/nomen_index.html*). Although the MHC nomenclature for other organisms has been under development (e.g. swine and bovine leukocyte antigens) a unifying system for the MHC nomenclature is lacking. Cytokine and cytokine related gene nomenclature is also not well defined — a comprehensive list of cytokine names can be found at the COPE web site (*www.copewithcytokines.de*).

In addition, each immunological database has its own unique structure, data models, and interfaces. Common interfaces, such as SRS (*srs6.ebi.ac.uk*) can integrate multiple databases and search tools, but are general tools. A common

interface for multiple immunological tools and databases would provide long-term benefits for immunology research. This common interface would provide seamless access to data and easy integration of both general and specialist bioinformatics tools.

Data issues

The interpretation of data extracted from the databases is highly dependent on the skills and knowledge of the user. In many cases the complicating factors are lack of standards, ad hoc nomenclature, variable quality annotations of database entries, incomplete data and biases embedded in the data. The optimal database searching tools for addressing a particular problem may require careful selection as well as setting of search parameters. Although the situation is slowly improving, the lack of bioinformatics education represents a serious obstacle to extracting the best value from data and unfortunately this problem often goes unnoticed by users. Data residing in databases are not of uniform quality, and even well-curated databases contain numerous errors (for a case study of errors in databases, see Srinivasan et al 2002).

Accuracy of computational methods

Hundreds of bioinformatics tools are available for analysing biological data. Many of these, such as sequence comparison and sequence alignment tools (such as standard bioinformatics tools BLAST or FASTA) calculate the distance between the query sequence and the database entries. This distance is based on user-selected parameters of the search and statistical assessment of the data and method. Therefore, search results may differ and assessing the accuracy of these tools is not informative. On the other hand, assessment of accuracy of predictive methods (such as prediction of peptide binding to MHC molecules) is of critical importance. In the past, most of the predictive models were generated and provided to the research community without careful assessment of their predictive performance. This resulted in some predictions of poor accuracy and a low level of acceptance of predictive bioinformatics models by the majority of researchers. More recently, assessment of predictive performance has become standard and vastly improved and refined predictive methods are appearing. A comparative study of the predictive performance of various methods has been recently published (Yu et al 2002). In addition, it was shown that predictive methods, when combined with experimental research in a cyclical fashion (Fig. 1.) can significantly improve the efficiency of research (Brusic et al 2001).

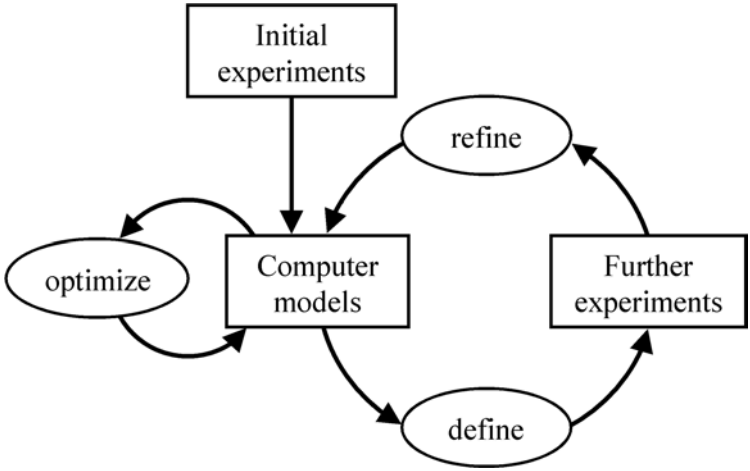

FIG. 1. Cyclical refinement of computer models used to define further experiments, including the optimization step. The optimization uses computer science methods, while refinement uses new experimental data.

Conclusion

Immunoinformatics is an enabling technology that will increasingly dominate immunology research, following the pattern set by genomics and proteomics. The scope of immunoinformatics is huge — it comprises databases, molecular-level and organism-level models, genomics and proteomics of the immune system, as well as genome-to-genome studies. Immunomics is thus the natural extension of genomics and proteomics and includes the study of organism-to-self and organism-to-organism interactions.

The efficient development and use of immunoinformatics will require the coordinated efforts of immunologists and bioinformaticians to establish common standards and protocols as well as standardized tools and interfaces. While coordinating efforts may be a challenge in this fast developing field, it is essential if we are to make sense out of the mountains of immunological data that will be produced in coming decades.

References

Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P 1999 A direct estimate of the human $\alpha\beta$ T cell receptor diversity. Science 286:958–961

Bernaschi M, Castiglione F 2002 Selection of escape mutants from immune recognition during HIV infection. Immunol Cell Biol 80:307–313

Brusic V, Zeleznikow J 1999 Computational binding assays of antigenic peptides. Lett Pept Sci 6:313–324

Brusic V, Zeleznikow J, Petrovsky N 2000 Molecular immunology databases and data repositories. J Immunol Methods 238:17–28

Brusic V, Bucci K, Schönbach C, Petrovsky N, Zeleznikow J, Kazura JW 2001 Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding. J Mol Graph Model 19:405–411, 467

Brusic V, Petrovsky N, Zhang G, Bajic VB 2002 Prediction of promiscuous peptides that bind HLA class I molecules. Immunol Cell Biol 80:280–285

Coussens PM, Nobis W 2002 Bioinformatics and high throughput approach to create genomic resources for the study of bovine immunobiology. Vet Immunol Immunopathol 86:229–244

De Groot AS, Sbai H, Aubin CS, McMurry J, Martin W 2002 Immuno-informatics: mining genomes for vaccine components. Immunol Cell Biol 80:255–269

Doytchinova IA, Flower DR 2002 Quantitative approaches to computational vaccinology. Immunol Cell Biol 80:270–279

Foster CB, Chanock SJ 2000 Mining variations in genes of innate and phagocytic immunity: current status and future prospects. Curr Opin Hematol 7:9–15

Garcia KC, Degano M, Pease LR et al 1998 Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. Science 279:1166–1172

Gendel SM 2002 Sequence analysis for assessing potential allergenicity. Ann NY Acad Sci 964:87–98

Glynne RJ, Watson SR 2001 The immune system and gene expression microarrays—new answers to old questions. J Pathol 195:20–30

Jerne NK 1993 The Nobel Lectures in Immunology. The Nobel Prize for Physiology or Medicine, 1984. The generative grammar of the immune system. Scand J Immunol 38:1–9

Kepler TB, Perelson AS 1993 Cyclic re-entry of germinal center B cells and the efficiency of affinity maturation. Immunol Today 14:412–415

Klade CS 2002 Proteomics approaches towards antigen discovery and vaccine development. Curr Opin Mol Ther 4:216–223

MacLean M, Miles C 1999 Swift action needed to close the skills gap in bioinformatics. Nature 401:10

Marshall T, Williams KM 2002 Proteomics and its impact upon biomedical science. Br J Biomed Sci 59:47–64

Martin AC, Cheetham JC, Rees AR 1989 Modeling antibody hypervariable loops: a combined algorithm. Proc Natl Acad Sci USA 86:9268–9272

Merrill SJ 1998 Computational models in immunological methods: an historical review. J Immunol Methods 216:69–92

Parker KC, Bednarek MA, Coligan JE 1994 Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152:163–175

Perelson AS 1989 Immune network theory. Immunol Rev 110:5–36

Petrovsky N, Brusic V 2002 Computational immunology: the coming of age. Immunol Cell Biol 80:248–254

Rammensee HG, Bachmann J, Emmerich NP, Bachor OA, Stevanović S 1999 SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V 1999 Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. J Med Chem 42:4650–4658

Ruiz M, Lefranc MP 2002 IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53:857–883

Saito H, Nakajima T, Matsumoto K 2001 Human mast cell transcriptome project. Int Arch Allergy Immunol 125:1–8

Schönbach C, Kun Y, Brusic V 2002 Large-scale computational identification of HIV T-cell epitopes. Immunol Cell Biol 80:300–306
Schueler-Furman O, Elber R, Margalit H 1998 Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. Fold Des 3:549–564
Srinivasan KN, Gopalakrishnakone P, Tan PT et al 2002 SCORPION, a molecular database of scorpion toxins. Toxicon 40:23–31
Wheeler DL, Church DM, Lash AE et al 2002 Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res 30:13–16
Yates A, Chan CCW, Callard RE, George AJT, Stark J 2001 An approach to modelling in immunology. Brief Bioinform 2:245–257
Yoshimori A, Del Carpio CA 2001 Automatic epitope recognition in proteins oriented to the system for macromolecular interaction assessment MIAX. Genome Inform Ser Workshop Genome Inform 12:113–122
Yu K, Petrovsky N, Schönbach C, Koh JYL, Brusic V 2002 Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol Med 8:137–148
Zagursky RJ, Russell D 2001 Bioinformatics: use in bacterial vaccine discovery. Biotechniques 31:636–640

DISCUSSION

Petrovsky: I would like to start in a slightly argumentative mode, by questioning the idea that Hans Georg Rammensee brought up in this introduction that peptide binding data constitute hard evidence and immunoinformatic predictions constitute semi-soft or soft evidence. I would argue that the quality of data is dependent on the level of its validation rather than whether it is derived from laboratory studies or computer models. Hence, couldn't well validated computer algorithms be considered hard and poorly validated experimental assays be soft?

Rammensee: It is a matter of quality control.

Petrovsky: Exactly. The quality of the data is a reflection of their statistical validation rather than their source. As an example, consider how MHC restriction was originally described: when did this evidence go from being soft to being hard? We initially started with Zinkernagel and Doherty's original description of MHC restriction of viruses, but the nature of the molecules involved and the manner in which they interacted was pure conjecture. Over time experimental details led to the proposal that a complex of MHC, antigen and a TCR underpinned this phenomenon. At that stage, however, given that no-one had actually seen an MHC molecule or a TCR, was this hard or soft evidence of the existence of these molecules. Later there was argument about how MHC was binding antigens with some people believing the peptide was bound in the cleft whereas others thought it was bound outside the cleft. More recently crystal structures have begun to appear and for the first time we can actually visualize what, up to that point, people had been hypothesizing about. Hence substance is a question of validation. Sometimes we fool ourselves into thinking that because something was measured in a lab it must be hard, whereas

if it is derived from a computer model it must be soft. I do not think this reasoning is correct.

Gulukota: I would add that when you have an interface between computational biology and bench biology, often the computational side believes that 10 computations are not as good as a single experiment, but this could just be because they don't know that much about experiments. On the other side, however many experiments the experimentalists do, they don't quite believe it until a computer prediction says something similar. There needs to be a cultural shift. Hard and soft is very much in the eye of the beholder. When we talk about biology, it is pretty much all soft!

Rammensee: I was restricting the use of 'hard' to just DNA, RNA and peptides sequences. The hard facts about MHC restriction are the sequence of the MHC, the sequence of the peptide and the sequence of the TCR.

DeLisi: In effect, you are distinguishing data from concepts.

Gulukota: Even in data there are gradations of softness. If you consider data such as MHC peptide binding, there are three or four different ways of measuring this. I'm sure we all have our personal preferences about whether IC_{50} is better than half-life, for example. Until we have a consensus, we can't even call experimental data hard.

Brusic: I have experience with assessing which method is best for measuring peptide binding. I started from the computational end and interviewed people who measure peptide binding and asked them which method they considered the best. I got a unified answer, 'mine'! Then I took a fuzzy approach to interpreting measurement data by converting all the values to approximate measures of values.

Rammensee: I don't think the peptide binding is the most important component of the quality of a certain peptide. The most important part is whether this peptide is recognized under physiological conditions. If you have a virus-infected cell and a T cell, does the T cell that is specific for a particular peptide recognize the virus-infected cell? This is the acid test. We again come to the point about what the right test is and what the best criteria are for calling something solid or soft.

Stevanović: It is still difficult to judge the properties of peptides. You say that sequences are hard data, and I agree with this. But the properties of peptides in terms of binding to MHC molecules or recognition by TCRs vary with the experimental setting. In particular, in cancer immunology, we know very well that there are so-called T cell epitopes that do not function in many labs. Even T cell recognition can't be called 'hard' data.

Rammensee: We are talking about immunoinformatics, but sequences in databases are hard data.

Littlejohn: I think your concept of hard data is a useful one. Hard data should be seen as discrete information, observations that can be digitized, and that are qualitative and not continuously variable. 'Hard data' are the foundation stones

in molecular observations. Then, on top of hard data, we can superimpose 'noise' and biological variation, and the contextually dependent observations that we have discussed here. If this is what you mean by 'hard immunoinformatics data' then I think this is an extremely useful concept that constitutes a good reference point against which we can compute (i.e. carry out rigorous immunoinformatics).

DeLisi: Of course, there is noise in the hard data also. Sequencing has an error rate of about one base in a thousand.

Rammensee: This brings us back to the issue of quality control.

Marsh: I like this idea of hard data. The HLA database that we run is a 'hard database': it is a database of sequences. The difficulty we have is knowing how to link our hard database to other databases. For example, there are many databases doing peptide prediction for MHC binding peptides. Which one should we link up with? We don't want to link our hard database with a semi-soft database that gives poor predictions.

DeLisi: There needs to be more benchmarking. We have done this with peptide MHC. Zhiping Weng and her colleagues have an algorithm that is about 90% reliable in terms of both specificity and sensitivity. This has been benchmarked in terms of all the standard algorithms on the web. Parker comes close to that. If we have more benchmarking like this, then this will go some way to alleviating this problem.

Littlejohn: I think the problem is elsewhere: it lies with evidence. Many of the databases do not ascribe evidence as to how the information was derived. Was it experimental? What experiment? Was it computational? What method was used? Who did it, when, and in what context? This is the big problem. The Gene Ontology consortium is battling with this issue of 'evidence', and this consortium has only just begun to think about how to ascribe evidence codes to the methods used to assign function to genes. I'd argue that this is one of the great problems in bioinformatics in general, and it needs to be tracked in the database as well as the derived information.

Wingender: That was exactly the point I was going to make. When we start differentiating between hard, soft and semi-soft data, we have to assign where the 'facts' come from. What is the source of the experimental or computational evidence? Whenever we model these data and provide them through a database, we simply have to provide the evidence, along with these data and facts. We then need to try to make a quality assignment to the data on the basis of this information. I would like to add a caution here against databases that have been made using data collected in an automated manner. There are some terrible mistakes in these. The data must be extracted manually from the literature, but this is also an error-prone process. The original data in the paper can even contain errors. At some point we have to rely on the quality control step of peer-review, though.

Rammensee: With regard to the problem of interconnectivity of databases, I would say that if I had a database which is quality controlled and contains good data, I don't want to have it connected with a bad database — for instance, one made automatically without adequate curation. I would like to protect my database from being corrupted by poor data. Thus we need to discuss the two important issues of interconnectivity and quality control together.

Margalit: We all agree about the need for quality control and good documentation. Who can do this? Most of the databases are assembled by research groups and are not commercial. I know from other fields that I am involved in, such as transcription factor binding and protein–protein interactions, that these databases may start in the academy, but at some point they decide they can't handle the scale of the database and they make a consortium or go commercial. Perhaps this meeting represents an opportunity to think how we can best develop a single, quality controlled immunoinformatic database that isn't spoiled by bad data.

Borras-Cuesta: I have a point about the quality control of databases. One issue is whether a peptide binds or does not bind to MHC: one should control this. The other thing, related to the predictive algorithms, is how these peptides were defined and collected. You could have a database that tells you the truth with respect to binding, but which is skewed with respect to predicting the set of potential binders. This is very important. People who like us work in the induction of immune responses, and have to try to characterize a peptide to induce a response, go through all the steps predicting this with one algorithm and then another. By the end we do not trust any in particular. We use several algorithms, and select the peptides predicted with higher scores from all algorithms. These peptides are synthesized and tested in binding assays, if these are available, or used in immunization experiments. But if algorithms are going to be described which are potent, one should discuss how to build a good database. That is, a database which has no bias for a particular set of peptides because it has been built up using, ideally, several methods (i.e. peptides eluted from MHC molecules, identified with phage display libraries, using peptide libraries, etc.). Peptides from this database could then be used to develop an algorithm for the prediction of binding to MHC molecules.

Rammensee: You raise the important point that predictions can be tested.

Borras-Cuesta: Yes, we predict and then we test in a binding assay. This is not enough, of course, because they could be cryptic peptides. But if we predict and then it binds, then we use it.

DeLisi: The assay has to be quality controlled also. For instance, take the assay used by Parker. He validates it, but when you look at it you find this validity holds only under a certain range of conditions on the rate constants. Something may look valid, you do an analysis of it, and you find there is only some domain of validity.

The first thing that needs to be done, therefore, is to benchmark the assays. Then you benchmark the algorithms on benchmarked assays.

De Groot: I would like to second the idea of having a collective database. I would suggest that we categorize the peptides in the database by peptides that bind MHC and by peptides that are recognized by T cells. I agree that the type of assay is very important in this respect. We all train or benchmark our algorithms on different sets of peptides. We are now finding that the set of epitopes versus the set of binders might be slightly different subsets of HLA binding peptides. I have a second comment: I also wanted to mention that on Vladimir Brusic's time-line, the date that the structure of HLA was published by Don Wiley should be highlighted. When I first met Hannah Margalit and Charles DeLisi in Jay Berzovsky's laboratory, we were talking about how the peptide bound to the groove, and we were discussing about the peptide not being aligned with the sides of the groove. Once the crystal structure was published, this showed everyone the fact that the peptide was aligned parallel to the side of the HLA, and was also tightly constrained within the groove. This was a turning point for the field.

DeLisi: There was no doubt that the peptide was linear; the question was how it was oriented.

Rammensee: In the 1987 crystal structure (Bjorkman et al 1987), it was not clear how the peptide was organized.

Borras-Cuesta: This raises the point of how the peptide is read by the MHC II. In principle, it is possible that the peptide could be read from C-terminus to N-terminus in some circumstances. This is relevant to predictions. Someone should do the following experiment. Synthesize for instance 20 peptides known to be recognised by a given MHC II molecule. These peptides should also be synthesized in the C-terminal to N-terminal sense (that is, with the same amino acid sequence, but read from the C-terminus to the N-terminus, and not in the conventional way N-terminus to C-terminus). If some peptides from this new set were immunogenic in the context of the same MHC II molecule, then predictions should also take into account peptide sequences read from C-terminus to N-terminus.

De Groot: One thing we should add to the databases is information about non-binding peptides. We are all constrained by finances and we don't make the peptides that we predict wouldn't bind, because it is expensive to make them. However, many of us have done assays and found that some of the peptides that we have predicted don't bind. Some of us also make 'negative control' peptides and test these. It will be important to include the negative sets in the databases in order to improve the accuracy of our epitope prediction tools.

Rammensee: The quality of data will be worse if you include non-binding peptides, because the peptide-binding assay might miss some non-binders. What we call 'non-binding' peptides might bind if the assay conditions are altered.

Kellam: What we have been discussing are quality issues. Anyone who has been following the microarray field for the last few years will have seen how people have gone to extreme in describing how to 'quality control' experiments, to the point where you try to document absolutely everything. There is a huge community effort to describe standards and common protocols. In the end, if people start documenting what they are doing experimentally you have a chance of getting to the context of the data in the databases. For example, how many people even know the sex of the cell lines that they are working with? This can become important.

Littlejohn: I would like to comment on that from a standards and sociology point of view. The micoarray MIAME standard is supposed to be a minimum standard, yet it is often referred to by the user-community as a 'maximum irritation' standard, as it requires the biologists to capture more information than they ordinarily might. With regard to the database integration issue, eight years ago I attended the 'Meeting for the Interconnection of Molecular Biology Databases' (see *http://megasun.bch.umontreal.ca/ogmp/abstracts/mimdb.html* on the 'Organelle Genome Megasequencing Program') where many of these issues were discussed. There are a couple of developments in molecular biology databases that would be useful for us to consider by the immunoinformatics community. First, back then Peter Karp proposed that bioinformatics research would benefit from having a unified system of data interchange standards. However, as this idea was discussed, it became clear that each database curator has their own set of objectives, and so was unlikely to redesign their systems to fit a broad-community-developed standard that did not meet their narrower goals. The concept of bioinformatics databank warehouses has been around for a long time and has not made much headway into the community, primarily due to the fact that most databanks have evolved in isolation and have their own schema and specific target audiences, making their absorption in to a warehouse problematic. In spite of this, there has been a vast amount of effort put into systems that allow databank interconnectivity, such as the SRS system (Etzold et al 1996). Databank integration does not come at a quality cost. For example, SWISS-PROT, EMBL and GenBank all have databank cross references and these simply allow cross-databank navigation. Databank integration and 'ontological normalization' (deriving a common set of key-terms for accessing information across databanks) is a vigorous area of research, with many technologies variously called 'wrappers' or 'agents' serving as 'middleware' (software that joins other pieces of software or data) in this area. Interconnectivity is a critical issue, and it isn't in and of itself a problem. The final comment I have is that the debate should continue to focus on biology and not technology, although as Vladimir Brusic points out, at the end of the day this is a technology, a means to an end. Immunoinformatics is all about technologies that underpin the study of immunology, immunology is the