An Information Theoretic Quantitative Approach to Natural Selection and Evolutionary Processes at the Genomic Level

(Extended Abstract)

Gill Bejerano * Hanah Margalit † Naftali Tishby ‡

The Hebrew University

October 9, 2003

Abstract

We propose a novel *quantitative* description of the genomic-proteomic cycle based on the tradeoff between the genomic stochasticity and proteomic fitness. Considering a model mapping $Protein \rightarrow DNA \rightarrow DNA \rightarrow Protein$, from the genome to the proteome, and vice versa, as a composition of the codon usage, all forms of mutations, and translation using the genetic code, enables us to quantify this stochastic mapping via the *mutual information* between the decoded and phenotypic amino acids. On the other hand, the proteomic fitness is quantified by the *fitness loss* or *distortion* of the protein sequences, which on the single amino acid level is captured by commonly used amino acid substitution matrices.

The relationship between the information loss through the stochastic mapping of the proteins through the DNA and their fitness loss has a theoretical optimal tradeoff, described by the celebrated "Rate-Distortion theory" of Shannon and Kolmogorov. Using this optimal tradeoff we quantify the selection of evolution at the single codon level and compare random mutations with those selected for by evolution, using codon to codon substitution matrices based on human-mouse orthologs.

We find that nature's mutation matrices are within 10% of the optimal tradeoff, and show that longer evolution exhibits a downward shift along the optimal theoretical curve. We also depict the evolution of this curve over growing time scales, and show, quantitatively, how the genetic code is set to adopt mutation bias in favour of transitions over transversions in its favour.

We believe that this framework opens new venues for the combined analysis of the co-evolution of genes and proteins.

*School of Computer Science and Engineering, The Hebrew University, Jerusalem 91904, Israel. e-mail: jill@cs.huji.ac.il , **corresponding author**.

†Department of Molecular Genetics and Biotechnology, Hadassah Medical School, The Hebrew University, POB 12272 Jerusalem 91120, Israel. e-mail: hanah@md2.huji.ac.il

‡School of Computer Science and Engineering, and Center for Neural Computation, The Hebrew University, Jerusalem 91904, Israel. e-mail: tishby@cs.huji.ac.il

1 Introduction

Consider the well known central dogma of Molecular Biology - that information on the molecular level flows in one direction only. From DNA, the blue-print repository of life, through transcription into RNA, and translation of the majority of RNA sequences, into proteins. This chain of events is not foolproof. Errors, or mutations, can occur at any stage, causing the resulting protein to diverge from its DNA mold.

When adding evolution into this framework, the chain of events just described turns into a loop, a cycle (the cycle of life, if you like). Through the means of natural selection, resulting phenotypes struggle for survival and multiplication, giving rise to the next generation undergoing the very same process. In this context hereditary mutations affecting the DNA itself are potentially most harmful.

The main objective of this paper is to capture this view of molecular evolution in a mathematical framework that will allow us to ask *quantitative* questions about it, while making use of the multitude of data pouring in at the genomic level.

As the model we define is far too rich to be analyzed completely in one work we have chosen in this paper to begin by concentrating on one interpretation of it. We will therefore currently accept the genetic code as given, and ask whether this "information processing channel" from DNA to protein can be expected to preserve a successful phenotype, through many generations, while fighting against dispersive mutational forces. In this respect we will consider the large codon redundancy in the genetic code as a means of resisting phenotypical change while allowing for inevitable hereditary mutations.

The tradeoff between the quality of the mapping between protein sequences and the DNA coding regions on the one hand, and at the protein phenotypical fitness level on the other hand, suggests a powerful analogy with the problem of lossy compression in information theory. When formulating protein-DNA encoding and decoding as a probabilistic, or stochastic, mapping (a Markov chain), a natural fidelity measure between the proteins at the two ends of this map emerges - *mutual information*. Mutual information, an already well known quantity in other areas of computational biology (see, e.g. [11]), is the natural measure of the statistical dependence between two variables. In our case it tells us how faithfully any given phenotype is preserved through the 'noise' of mutation, henceforth we call it the "fidelity" of the map (or code-rate). Just as in lossy compression, however, the quality of this map is "tested" at a very different level - that of potential fitness loss resulting from mutant proteins. This measure, captured indirectly by comparing protein sequences that survived mutations, is formally identical to the distortion measure in lossy compression. While it is intuitively clear that the lower the fidelity of the mapping – the higher its expected distortion or fitness loss, the details of this tradeoff are quantitatively described by a fundamental result in information theory: *Rate-Distortion theory* (see Ch. 13 in [5]). This celebrated theorem, proved independently by Shannon and Kolmogorov, states that the minimal possible fidelity at a given fitness loss level, or equivalently the minimal distortion (fitness loss) that can be achieved at a given fidelity (mutation) level are given by one function, the rate-distortion function - which in our case can be considered as *the optimal fidelity-fitness tradeoff curve*. We believe that the introduction of this relation in our context is one of the main contributions of this paper.

The optimal fidelity-fitness tradeoff curve can be completely calculated when given the prior probability over the protein sequences (or amino acids) and the distance (distortion) measure between the sequences, via a well known iterative algorithm in information theory - the Blahut-Arimoto algorithm [5]. This allows us to vary and compare different components of the map, in particular mutation magnitude and codon usage, and examine the fidelity and fitness loss of real genomes for different (empirical) fitness measures. This gives an interesting insight into evolutionary selection at the molecular genomic level.

Several groups attempted to describe the consistency of the genomic-proteomic cycle by examining the optimality of the canonical genetic code for accurate information transmission from DNA (via mRNA) to proteins. Woese [25] proposed that the genetic code has evolved to its current structure of amino acid/codon assignments to minimize the effect of errors on the resulting proteins. Hurst [16, 14, 15] and colleagues have continued this line of reasoning by comparing the current code to alternative, theoretical codes, and demonstrating that the natural code performs exceedingly well in minimizing translational errors. Di Giulio and colleagues investigated the possibility that evolution of the genetic code was driven by the optimization of the physicochemical distances between amino acids [8, 9]. Wong proposed the co-evolution theory of genetic code and its protein products [26], which was further investigated and supported by Di Giulio and Medugno [10]. Another aspect is provided by studies that aim at the codon usage. These have ranged from studies that tabulate and record the codon usage of various genomes [24] to studies that attempt to identify

relationships between codon usage and various gene features. In particular, a correlation has been observed between codon usage and gene expression (e.g. [12, 19]). Other studies have attempted to correlate codon usage and gene expression with some properties of the amino acids, such as amino acid conservation and hydrophobicity [1, 7]. Thus, previous studies addressed one or two components of the genomic-proteomic cycle. Here we provide, for the first time, a simple formal description of all components in one system, and demonstrate in a quantitative manner one facet of how they all interact to shape the process of molecular evolution.

The rest of the paper is built as follows: In sec. 2 we define the model, and in sec. 3 we analyze it. We then move on to describe the data we have used in sec. 4, and our experimental results in sec. 5. Note that discussion, as a whole, is differed to sec. 6.

2 Model Definition

Consider the following abstracted view of evolution at the molecular level:

We consider two quantities in our formulation. A molecular genotype, represented by an organism's genome, more specifically in this setting by its coding portion. And a molecular phenotype, comprised by the full repertoire of proteins that an organism may synthesize at will.

The mapping between a DNA coding region and the respective protein segment encoded for within it is governed by that organism's genetic code. The vast majority of known organisms share the single, aptly termed standard code, but our discussion is by no means limited to that particular choice.

A further simplification, that can later be waved (see sec. 6), is made by observing both sequence collections at the level of a single meaningful symbol. The amino acid basic building block of proteins, later denoted A, and the codon, the complementing elementary meaningful unit in a coding region of DNA, denoted C. Different codon, or amino acid compositions will therefore represent different genomes, or (proteomic) phenotypes, respectively.

We now turn to define the evolutionary mapping. The mapping we are interested in $Protein \rightarrow DNA \rightarrow DNA \rightarrow Protein$ (see fig. 1) is mostly to be expected. The last two transitions depict the hereditary flow of information. A genome may undergo mutation, resulting in a phenotypic change. The first link and transition in our mapping, are best explained through the question we aim to probe - must it not be so that the above described process is able to propagate across many generations any given (successful) phenotype in a relatively robust manner? The intuitive answer to this question seems an obvious yes. Current species seem well adapted to their environments, and mutationally defunct organisms do not overly burden any given species.

This quality can be captured quantitatively by placing the $Protein$ link *infront* of the obvious mapping, and asking how well will any given 'prototypical' phenotype be preserved at the 'output' of this mapping. Codon redundancy allows for a multitude of mappings between this single phenotype whose preservation we wish to measure and the many genotypical forms it may assume. An organism's choice of codon usage can now be probed in relation to the above question.

Two complementary measurements can now be taken. One measuring the dependence between 'input' and 'output' of this process, or the *fidelity* of the process itself, as a lossy transfer of information (due to hereditary mutations). The other is interested in the actual *distortion*, or fitness gap, between the two.

Note that a single iteration over the chain can be set to represent not only a single generation but rather an evolutionary time scale of our desire, simply by placing the correct values for each of the stages in the process, in that time scale.

Another attractive feature of this model is that all quantities involved are measurable in the post genomic era. The proteins comprising a plethora of organisms are being discovered continuously, along with the organisms' choice of codon usage. Several genetic codes are well known. And our estimates for time scaled mutation magnitudes, and amino acid potential interchangeability, or distance, are constantly improved.

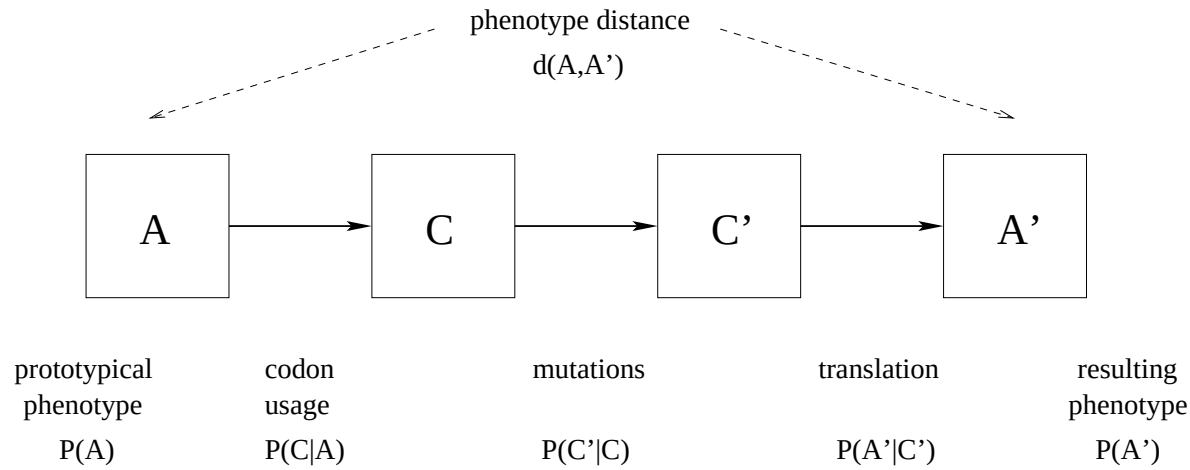

Figure 1: **An abstract model for molecular evolution.** See text for details.

3 Information Theoretic Analysis

3.1 Quantifying stochasticity: Protein coding fidelity

The mapping $Protein \to DNA \to DNA \to Protein$ described in the previous section is characterized in terms of several probability distributions (see fig. 1). The first is the *protein phenotype* distribution. In principle this should capture the probability over all observed sequences. At this point we examine the chain at the single amino acid level, hence this is taken as the prototypical (observed) amino-acid distribution, $P(A)$. The second distribution captures the *codon usage* or the probability that a certain amino acid A is mapped into one of the 64 possible codons C, $P(C|A)$. The next stage of this chain describes the map between codons to codons, which is a *stochastic* representation of all hereditary mutations, $P(C'|C)$. This is the only source of noise, or information loss, in our current model chain. The final stage is the translation from the abstracted DNA (or RNA) back into proteins, prescribed by the genetic code, $P(A'|C')$. The amino acid distribution at the end, $P(A')$ is, in general, different from $P(A)$. The relationship between the observed amino acid distribution and the stationary distribution of this process is an intriguing question which we discuss elsewhere.

The composition of all 3 stages of the chain gives a stochastic mapping from amino acid to amino acid, given by:

$$P(A'|A) = \sum_{C,C'} P(A'|C')P(C'|C)P(C|A)$$

which describes a Markov process that captures the protein encoding-decoding path via the genome. Our fidelity measure for this process is chosen as the mutual information between A' and A, namely:

$$I(A;A') = \sum_{A} P(A) \sum_{A'} P(A'|A) \log \frac{P(A'|A)}{P(A')} \ .$$

$I(A;A')$ is a symmetric non-negative quantity which vanishes if, and only if, A and A' are statistically independent. The mutual information is maximal and equal to the entropy of A, $H(A)$, when there is no loss of information, or no mutations. In general, we have defined a lossy process.

3.2 Protein phenotype fitness loss: the distortion function

The biological quality of the above (stochastic) mapping between the protein sequences is not determined by the information loss, or by mutual information, but rather by the fitness of the whole organism composed of the mutated proteins. This fitness loss can in principle be captured by the distance, or distortion between

the sequences, a measure at the heart of sequence analysis. At the amino acid level it is sufficient to consider amino acid interchangeability, as captured by the examination of families of homologous proteins. Defining a distance matrix $d(A, A')$ by this procedure captures the *average* fitness loss in the substitution $A \to A'$ in the organism. This is just the complementary measure of distortion, or fitness, required for the application of rate distortion theory.

Given the distance matrix $d(A, A')$, the survivability (fitness) of the organism is determined (within our first order discussion) by the average distortion,

$$\langle d \rangle = \sum_{A,A'} P(A)P(A'|A)d(A, A') = E_{A,A'}(d(A, A')) \ .$$

Notice that this expected distortion depends on both the (marginal) joint distribution $P(A, A')$ and the choice of the matrix $d(A, A')$. We thus expect some relationship between the amount of information preserved in the mapping of A to A', and the potential fitness loss captured by $\langle d \rangle$. This intuition is formalized and quantified in the next section.

3.3 Optimal tradeoff between fidelity and fitness: Evolutionary Rate Distortion Theory

The mutual information $I(A; A')$ and the average distortion $\langle d \rangle$ must be related. A lossless map has no distortion and maximally preserved information ($H(A)$), while independent A and A' have very high distortion (fitness loss). We thus expect a monotonic decrease of $I(A; A')$ with $\langle d \rangle$. A celebrated theorem of Shannon and Kolmogorov, known as Rate-Distortion theory provides the optimal tradeoff between $\langle d \rangle$ and I [1] given by,

$$I(D) = min_{\ P(A'|A):\langle d \rangle \leq D}\ I(A; A') \ ,$$

where the minimum is taken over all possible (stochastic) mappings $P(A'|A)$. This is a constrained minimization problem which can be solved by introducing a Lagrange multiplier for the distortion constraint [5]. Denoting the Lagrange multiplier by β, the *optimal* mapping is given by:

$$P(A'|A) = \frac{P(A')}{Z(A, \beta)} \exp[-\beta d(A, A')] \ ,$$

with $Z(A, \beta) = \sum_{A'} P(A') \exp[-\beta d(A, A')]$ a normalization factor. $P(A')$ must satisfy, self-consistently, the marginal condition,

$$P(A') = \sum_{A} P(A'|A)P(A) \ .$$

The repeated iterations of these two equations, for any given value of β, are guaranteed to converge to the unique solution [6] and are known as the Blahut-Arimoto algorithm for the rate-distortion function (see fig. 2).

1. Input distribution $P(A)$, distortion matrix $d(A, A')$ and curve parameter β.
2. Initialize $P(A')$ to some distribution.
3. Repeat
   (a) $P(A'|A) = \frac{P(A')e^{-\beta d(A,A')}}{\sum_{A'} P(A')e^{-\beta d(A,A')}}$
   (b) $P(A') = \sum_A P(A)P(A'|A)$
   Until convergence of $P(A'|A)$
4. Calculate and return $D = \sum_{A,A'} P(A)P(A'|A)d(A, A')$
   and $I = \sum_{A,A'} P(A)P(A'|A) \log \frac{P(A'|A)}{P(A')}$.

Figure 2: Evaluation of $I(D)$ using the Blahut-Arimoto Algorithm (adapted from [5]).

[1] Strictly speaking, we only use the "easy" part of the theorem, as we do not use the existence of the lossy code.

This optimal tradeoff between fidelity and fitness defines a concave function, $I(D)$, whose slope is given by $-1/\beta$. By varying β one can obtain the full $I(D)$ curve. *Only* points above this curve are achievable by a stochastic mapping $P(A'|A)$. For a typical curve, together with several calculated points see fig. 5.

4 Data Sources and Manipulation

We review the data used in the experiments below, following the flow of our model (see fig. 1). A discussion of the data can be found in sec. 6.

Prototypical phenotypes of representative organisms and their complementing codon usage tables were obtained from the CUTG database [24] version based on the NCBI GenBank [4] flat file Release 117.0 (April 15, 2000). The organisms chosen were human, mouse, *E.coli*, *S.Cerevisiae*, *D.Melanogaster*, *C.Elegans* (all having data collected from many coding sequences), *Methanoccocus J.* (representative Archaea), *Pimelodella Ch.* mitochondria (using the non standard vertebrate mitochondrial genetic code) and *S.Cerevisiae* mitochondria (using another variant of the code). Technically speaking the acquired $P(A, C)$ data was split into $P(A)$ and $P(C|A)$.

Two kinds of mutation matrices were used. The first, courtesy of Compugen (used in [2]), were collected in the following manner: mouse and human protein sequences were compared using the Smith-Waterman algorithm, and the defaultive Blosum62 matrix setting. Pairs within a varying range of similarity were declared homologous and their DNA sequences aligned according to the best protein alignment. Frequency counts were derived from these. Available matrices were also slightly noised artificially, at Compugen, but at very low noise levels. Four matrices were used, with protein homology ranges of 50–80%, 50–90%, 50–100% and 60–100%.

Theoretical mutation matrices, $P(C'|C)$, were also derived for various values of three parameters p, w and m. The mutation magnitude was governed by an i.i.d. probability of a single base substitution. All three bases of a codon were treated equally. p denotes the transition probability, the transversion[2] probability is $\frac{p}{w}$ (making w a skew factor), and m was defined to be the total mutation probability $P(C' \neq c|C = c)$. Writing the combinatorial relationship between the three parameters, mutation matrices were generated numerically for all combinations of $m = 0.01, 0.1, 0.2, 0.3, 0.4$ and $w = 1, 2, 3, 4, 5, 6$.

The genetic code tables were obtained from the NCBI taxonomy repository [13] (March 31, 2000 update). As mentioned above, we used three tables - the standard, the vertebrate mitochondrial and the yeast mitochondrial.

Well known amino acid distance matrices were collected from the AAindex database [20], Release 4.0 (September 1999). Chosen were Mclachlan's [21, 22], Grantham's [17], and Miyata's [23] distance matrices based on the physicochemical properties of the amino acids. Also chosen was the PAM 74-100 matrix [3], augmented by all Blosum matrices numbered between 30 and 100, acquired from the Blosum tar file version 5.0 (October 19, 1992) to comprise the subgroup of matrices where distance was assessed based on alignments of protein sequences of varying degrees of similarity.

In order to derive $d(A, A')$ from these matrices two issues had to be addressed - scaling and slight expansion. Scaling is needed in order to compare the different matrices with each other. On the face of it this is a difficult task - different matrices have different units and typical orders of magnitude. Luckily our formulation of the problem gives an elegant solution to this problem. All matrices were transformed through a linear transformation to hold values exactly between 0 and 1. The justification can be verified by observing the nature of the Blahut-Arimoto algorithm for computing the optimal curve (see fig. 2, clause 3a). Substituting $d(A, A') \mapsto K_1 d(A, A') + K_2$ in the $P(A'|A)$ equation causes K_2 to cancel out, while K_1 is absorbed by β (and we are interested in all values of β). A final adjustment is to shift all curves from their respective minimal distortion value to zero, along the D axis.

The expansion is much simpler. None of the matrices includes a distance measure between any amino acid and the stop codon, denoted '$*$'. This is solved by setting (normalized) $d(*, *) = 0$ and using a parameter $s = d(A, *)$ for all A, with tested values of $s = 1, 2, ..., 9$.

[2] A transition is a substitution between $A \leftrightarrow G$ or $C \leftrightarrow T$. All the rest are transversions. The distinction stems from structural similarity between A and G (the purines), and between C and T (the pyramidines).

5 Experimental Results

We briefly recount the experiments conducted and their results. We use the nomenclature developed in preceding sections. A discussion of the results and their implications can be found in sec. 6.

5.1 The Boundary Curve

As explained earlier on, the theoretical tradeoff boundary curve for the fitness-fidelity plane depends only on our choice of $P(A)$, and $d(A, A')$.

We begin by exploring the relationship between the distance matrix expansion parameter s and the resulting curve. Various pairs of $P(A)$ and $d(A, A')$ were taken. For each pair several boundary curves were drawn - one for each value of s in the range we have defined. In all cases the set of graphs for a single pair was found to almost overlap for all choices of s (data not shown)[3].

We proceed to examine the curve's dependence on $P(A)$. Several distance matrices were chosen, for each we have plotted the boundary curve for every one of the nine model organisms chosen earlier on. A representative result is given in fig. 3. It is seen that the shape of all graphs is very similar, and in fact most of them are very close to each other (most noticeable was an almost complete overlap between human and mouse).

We conclude this part by an examination of the curve's dependence on $d(A, A')$ itself. Several model organisms were chosen, for each all distance matrices were drawn. A typical result is shown in fig. 4. A representative subset of the Blosum matrices is given, while all 16 of them show the following clear relationship - the lower the number of the matrix, the further to the left it resides. PAM 74-100 is shown to reside to the left of all these. The two McLachlan chemical properties matrices 'off-shoot' to the right, while two others (Grantham's and Miyata's) lie somewhere in the Blosum 35-40 region.

5.2 Measuring points on the achievable plane

Having described the boundary curve we move over to measure specific points on the achievable plane. Each such point requires the specifications of all relevant quantities defined in our model.

We currently concentrate on measuring the effect of different mutation matrices on our location on the fitness-fidelity curve. We probe the difference between theoretical mutation matrices, and those measured in practice. Since for the later resource we now have at our disposal only the Compugen matrices, acquired through a comparison of human and mouse sequences, we will confine ourselves to these organisms. A typical result is given in fig. 5. The graph reveals three different phenomena:

First, observing all skewless ($w = 1$) theoretical mutation matrices one sees a clear down sloping curve that advances as the overall codon mutation magnitude (m) grows.

Second, for each such skewless point several points are drawn sharing the same mutation magnitude, at increasing skewness. It is shown that the larger the skew in favour of transition over transversion, the better the resulting points (i.e. they climb back up the curve).

Third, and most interesting, the points using actual statistics collected from human - mouse similarities are much closer to the boundary curve. In fact all points seem to maintain a similar gap of roughly 10% from the optimal curve, in both respects. It is also clear that the more similar the sequences from which the mutation matrix is built, the better the process (again, climbing up the curve).

To conclude the section we note that the same behaviour has been observed under all examined distance matrices, including the Grantham and Miyata chemical properties based matrices, with distance from the optimal curve falling to some 5%.

5.3 Some Analysis of the Empirical Points

Having observed the great efficiency achieved by the human - mouse mutation matrices, it is interesting to note that on first look these matrices seems very un-conserved. An exemplary matrix, the 50–90%, shown in fig. 6, has a minimal codon conservation $(1 - m)$ value of 0.07 and an uninspiring average of 0.30 ± 0.16. However when one utilizes our knowledge of the genetic code in use and collates all codons for a particular

[3] All other experiments will therefore be presented with a canonical choice of $s = 5$.

amino acid into a single group, a far more coherent picture arises, where the minimal conservation is now 0.58 and the average has risen to 0.71 ± 0.08. Note that these values also apply to the process as a whole, as both the codon usage and genetic code transformations are both lossless.

6 Discussion

It is difficult, within the space limitations of an extended abstract, to raise and discuss properly every single issue involved in a model like this. We shall thus try to raise several important issues, and discuss them as succinctly as possible.

One strong criticism of the model, as presented, is the simplifying assumption of representing meaningful sequences merely by their composition. The best answer to this claim, is perhaps in the results themselves. Clearly the essence of what we qualitatively expect of the model is already here in this simplest order approximation. Naturally a more realistic treatment of sequences should sharpen our quantitative results and bring them closer to reality itself.

Note that the claim at the heart of the analysis we chose here, that the evolutionary channel, so to speak, can preserve any phenotypic 'idea', is justified, at least for the standard code, by its proliferation in different organisms carrying very different phenotypes. Existing phenotypes are by themselves natural candidates for 'inputs' into this channel, both since they do represent the wild type nowadays, and because it is the most reasonable substitute for a wild type phenotype of some vintage which we cannot currently deduce.

A final general comment regarding the model concerns the nature of mutations. We claim that it is valid in our model to consider all mutations as harmful, as our goal is in observing phenotypic 'scattering' throughout the ages, and not (at least here) the advancement of a phenotype interacting with its environment.

Turning over to available data, there are two seeming tautologies that need refuting. The first is the use of amino acid distance matrices (like Blosum and PAM) based themselves on a comparison of divergent protein sequences. Is our measure of closeness then not derived from the products of the channel itself? To this we offer two answers - the first, specific to our model, is a reminder that the figures we show remain the same (qualitatively) when using the Grantham or Miyata physicochemical properties derived matrices. For another way of making away with this problem we refer the reader to [15] which, put shortly, claims that divergent enough similarity matrices are exactly that - enough time has passed for them to reflect real amino acid interchangeability relationships and not the channel.

A second, similar argument calls our attention to the empirical human - mouse homology matrices. These matrices, used for estimating the magnitude of hereditary DNA mutation, were again themselves based on protein sequence similarities. Here we offer an argument closer in spirit to the second one above. We claim that the lower threshold for these matrices, at 50% or 60% similarity, is high enough for us to both conclude that all observed pairs are truly homologous, and furthermore that any other scoring scheme we'd pick would claim the same over again.

Note that selection as a driving force enters our mapping in the form of these empirically derived hereditary mutational matrices. This is what sets these matrices well apart from the oblivious selectionless theoretical matrices we use in comparison.

Turning to our results, the graph of fig. 3 is indeed somewhat surprising. It is possible that a part of the evolutionary processes we have observed here has pushed existing phenotypes into a subset sharing certain common features. In this respect it is rather easy to verify that the major variation in the observed amino acid distribution is due to the structure of the genetic code itself and is approximately given by $\sum_C P(A|C)P(C)$.

However, our analysis shows that the rather significant deviations of the observed phenotypical amino acid distribution from this genetic code prediction are almost entirely explained by the stationary distribution of our Markov chain model. This suggests taking our "cycle" even more seriously and it quantitatively justifies the idea of co-evolution of proteins and codons. We are currently investigating this intriguing observation and its implications.

The interesting observation in fig. 4, that the more diverged the protein pairs at the base of the Blosum or PAM[4] matrix the further it resides to the left, can explained as follows: First it is natural, and expected, that as the magnitude of divergence increases, so will the fidelity of any mapping, as well as the optimal, drop.

[4]Note that this newer PAM matrix, in contrast to the Dayhoff ones, is derived from highly diverged sequences.

This explains the I axis shift with evolution. Since the distortion D puts a direct limit on the survivability of the mutations the only way we can observe lower fidelity in real evolution is when it is accompanied by lower average distortion. This is why the (observed) curves of $I(D)$ are shifted to the left with evolution.

Moving over to our central results of fig. 5, all of which we claim are quantitative affirmations of common sense qualitative arguments. First the down-sloping curve for increasing skewless hereditary mutation magnitude can be taken to signify where life would have been without the driving force of natural selection, and that is very far away from the optimal tradeoff curve.

The observation that DNA structure-dependent skewness is actually favoured by the genetic code, in the sense that it causes less translational errors, can be affirmed by observing the code itself. This tendency is most obvious when considering how amino acids encoded by two codons *always* use one pair or the other.

And the upward motion of our empirical points is clear, again in light of the fact that less divergent sequences, by definition conserve more of the mutual information between them. It is however intriguing to observe, both the roughly constant distance this curve keeps from the optimal tradeoff curve, and the dramatic improvement it has over the selectionless theoretical one.

And finally, adding fig. 6 into the picture, one again sees something which is expected qualitatively, that although the dispersion at the codon level is seemingly large, in fact natural selection keeps the real conservation in these matrices in check quite strictly.

To conclude, we believe we have presented here a novel formulation of molecular evolution at the genomic-proteomic level. Interesting results have been shown with respect to specific aspects of this process. We also believe that this model can be used to ask many more intriguing questions, such as those relating to the origin of the code itself, in a quantitative manner, and we aim to further pursue this exciting line of work.

Acknowledgements

The authors wish to thank Mor Amitai, Vera Asodi and Alex Diber from Compugen for providing the mutation matrices, and Jorja Henikoff from FHCRC for help with the Blosum database. GB is supported by a grant from the Ministry of Science, Israel.

References

[1] Alff-Steinberger, C., A comparative study of mutations in escherichia coli and salmonella typhimurium shows that codon conservation is strongly correlated with codon usage, *J. Theor. Biol.*, **206(2)** , 307–311 (2000).

[2] Asodi, V., A. Diber, R. Gill-More, H. Safer and M. Amitai, EstSearch: A Tool for More Informative EST Alignments, in *Currents in Comp. Mol. Bio.*, 72–73, Universal Academy, Tokyo (2000).

[3] Benner, S.A., M.A. Cohen and G.H. Gonnet, Amino Acid Substitution During Functionally Constrained Divergent Evolution of Protein Sequences, *prot. eng.*, **7**, 1323–1332 (1994).

[4] Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp and D.L. Wheeler, GenBank, *Nuc. Acids Res.*, **28(1)**, 15–18 (2000).

[5] Cover T.M. and J.A. Thomas, Elements of Information Theory, Wiley, New York (1991).

[6] Csiszár, I. and G. Tusnády, Information geometry and alternating minimization procedures, *Stat. Dec.*, supplement issue **1**, 205–237 (1984).

[7] de Miranda, A.B., F. Alvarez-Valin, K. Jabbari, W.M. Degrave, and G.G. Bernardi, Expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in Mycobacterium tuberculosis and Mycobacterium leprae, *J. Mol. Evol.*, **50(1)**, 45–55 (2000).

[8] Di Guilio, M., On the optimization of the physiochemical distances between amino acids in the evolution of the genetic code, *J. Theor. Biol.*, **168**, 43–51 (1994).

[9] Di Giulio, M. and M. Medugno, Physicochemical optimization in the genetic code origin as the number of codified amino acids increases, *J. Mol. Evol.*, **49(1)**, 1–10 (1999).

[10] Di Giulio, M. and M. Medugno, The robust statistical bases of the coevolution theory of genetic code origin., *J. Mol. Evol.*, **50(3)**, 258–263 (2000).

[11] Durbin, R., S. Eddy, A. Krogh and G. Mitchison, Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge University Press, Cambridge (1998).

[12] Duret, L., tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes, *Trensd. Genet.*, **16(7)**, 287–289 (2000).

[13] Elzanowski A. and J. Ostell, The Genetic Codes, NCBI repository, via `http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html/` (2000).

[14] Freeland, S.J. and L.D. Hurst, The genetic code is one in a million, *J. Mol. Evol.*, **47(3)**, 238–248 (1998).

[15] Freeland, S.J., R.D. Knight, L.F. Landweber and L.D. Hurst, Early Fixation of an Optimal Genetic Code, *Mol. Biol. Evol.*, **17(4)**, 511–518 (2000).

[16] Haig D. and L.D. Hurst, A quantitative measure of error minimization in the genetic code, *J. Mol. Evol.*, **33(5)**, 412–417 (1991).

[17] Grantham R., Amino Acid Difference Formula to Help Explain Protein Evolution, *Science*, **185**, 862–864 (1974).

[18] Henikoff S. and J.G. Henikoff, Amino Acid Substitution Matrices from Protein Blocks, *Proc. Nat. Acad. Sci.*, **89**, 10915–10919 (1992).

[19] Karlin, S. and J. Mrazek, Predicted highly expressed genes of diverse prokaryotic genomes, *J. Bacteriol.*, **182(18)**, 5328–5350 (2000).

[20] Kawashima, S. and M. Kanehisa, AAindex: Amino Acid Index Database, *Nuc. Acids Res.*, **28(1)**, 374 (2000).

[21] McLachlan, A.D., Tests for Comparing Related Amino Acid Sequences Cytochrome C and Cytochrome c551, *J. Mol. Biol.*, **61**, 409–424 (1971).

[22] McLachlan, A.D., Repeating Sequences and Gene Duplication in Proteins, *J. Mol. Biol.*, **64**, 417–437 (1972).

[23] Miyata, T., S. Miyazawa and T. Yasunaga, Two Types of Amino Acid Substitutions in Protein Evolution *J. Mol. Evol.*, **12**, 219–236 (1979).

[24] Nakamura, Y., T. Gojobori and T. Ikemura, Codon Usage Tabulated from International DNA Sequence Databases: Status for the Year 2000, *Nuc. Acids Res.*, **28(1)**, 292 (2000).

[25] Woese, C.R., On the evolution of the genetic code, *Proc. Natl. Acad. Sci.*, **54**, 1546–1552 (1965).

[26] Wong, J.T., A co-evolution theory of the genetic code., *Proc. Natl. Acad. Sci.*, **72**, 1909–1912 (1975).

[27] Wong, J.T., Role of minimization of chemical distances between amino acids in the evolution of the genetic code, *Proc. Natl. Acad. Sci.*, **77**, 1083–1086 (1980).

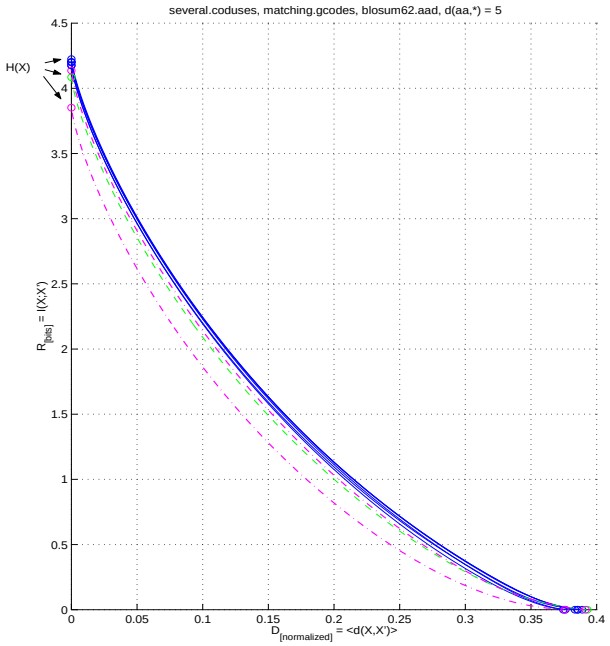

Figure 3: **Boundary curve for different organisms.** Most Prokaryotes/Eukaryotes are in solid (blue), the Archaea is in dashed (green), and the two mitochondria in dash-dot (magenta). Leftmost, apart, is *Pimellodela Ch.* Mitochondria.

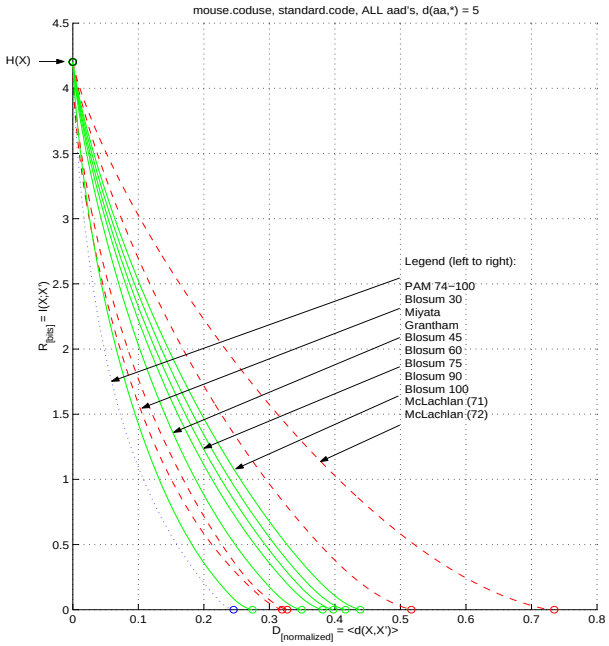

Figure 4: **Boundary curve for different distance matrices.** The four physicochemical properties based matrices are shown in dashed (red). A representative subset of the Blosum matrices is given in solid (green), and the PAM matrix is (blue) dotted.

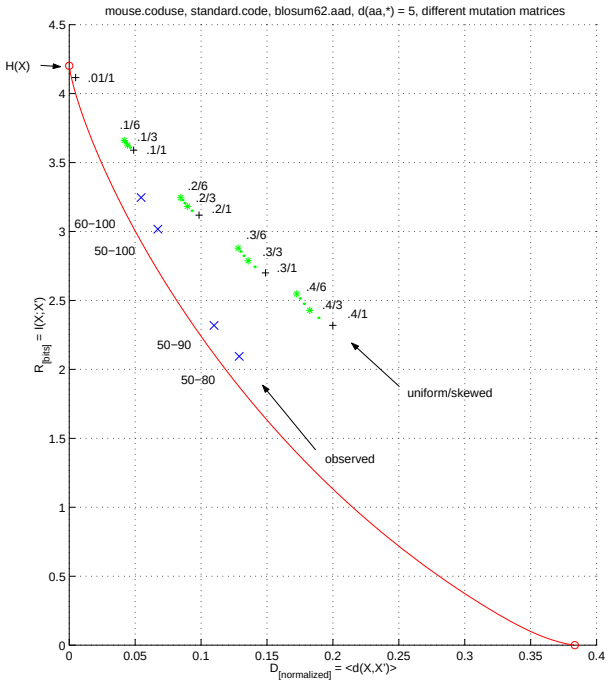

Figure 5: **Different mutation matrices.** Measures taken with theoretical matrices are shown in (black) '+' for skewless mutations and (green) '.' or '*' for matrices with positive skew. The labels on these points correspond to m/w. Finally, (blue) 'x' marks show the four human - mouse matrices, with labels showing the bounding thresholds for sequence similarity used to create them.

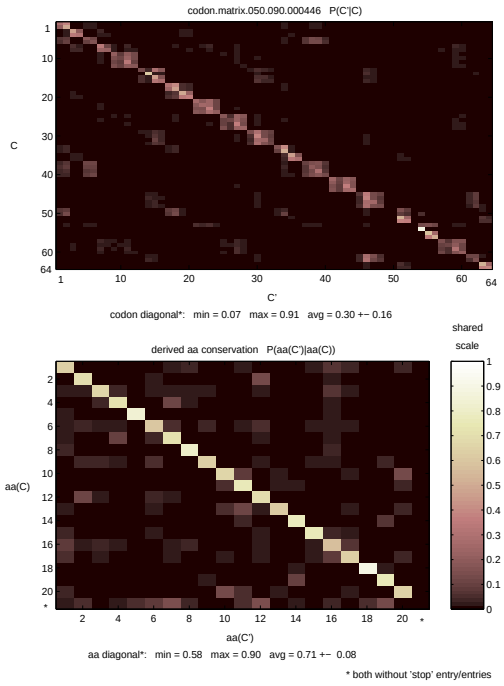

Figure 6: **Hidden conservation.** The upper figure gives a color coded view of the conservation at the codon level, while the lower figure gives the de-facto conservation when considering the genetic code derived groups of codons coding for the same amino acid.