A Model of Substitution Trajectories in Sequence Space and Long-Term Protein Evolution

Dinara R. Usmanova,1,2,3 Luca Ferretti,4,5 Inna S. Povolotskaya,2,3 Peter K. Vlasov,2,3 and Fyodor A. Kondrashov*,2,3,6
1Moscow Institute of Physics and Technology, Institutskiy Pereulok 9, g.Dolgoprudny, Russia
2Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain
3Universitat Pompeu Fabra (UPF), Barcelona, Spain
4Systématique, Adaptation et Evolution (UMR 7138), UPMC University Paris 06, CNRS, MNHN, IRD, Paris, France
5CIRB, Collège de France, Paris, France
6Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
*Corresponding author: E-mail: fyodor.kondrashov@crg.eu.
Associate editor: Hideki Innan

Abstract

The nature of factors governing the tempo and mode of protein evolution is a fundamental issue in evolutionary biology. Specifically, whether or not interactions between different sites, or epistasis, are important in directing the course of evolution became one of the central questions. Several recent reports have scrutinized patterns of long-term protein evolution claiming them to be compatible only with an epistatic fitness landscape. However, these claims have not yet been substantiated with a formal model of protein evolution. Here, we formulate a simple covarion-like model of protein evolution focusing on the rate at which the fitness impact of amino acids at a site changes with time. We then apply the model to the data on convergent and divergent protein evolution to test whether or not the incorporation of epistatic interactions is necessary to explain the data. We find that convergent evolution cannot be explained without the incorporation of epistasis and the rate at which an amino acid state switches from being acceptable at a site to being deleterious is faster than the rate of amino acid substitution. Specifically, for proteins that have persisted in modern prokaryotic organisms since the last universal common ancestor for one amino acid substitution approximately ten amino acid states switch from being accessible to being deleterious, or vice versa. Thus, molecular evolution can only be perceived in the context of rapid turnover of which amino acids are available for evolution.

Key words: molecular evolution, fitness landscape, epistasis.

Introduction

Whether or not epistasis, a situation when fitness is dependent on the interaction of alleles, plays a major role in molecular evolution is the subject of scrutiny and debate (de Visser et al. 2011; Lehner 2011; Breen et al. 2012; Hansen 2013; McCandlish et al. 2013; de Visser and Krug 2014). Two types of approaches are being used to reveal the type and amount of epistasis in protein evolution: First, studies aiming to reconstruct recent evolutionary trajectories revealing potential epistatic interactions among substitutions that occurred recently (Weinreich et al. 2006; Bridgham et al. 2009; Lozovsky et al. 2009; Romero and Arnold 2009; Lunzer et al. 2010; Khan et al. 2011; Zhang et al. 2012; Covert et al. 2013); second, studies quantifying the degree of epistasis among substitutions that occurred across long evolutionary periods (Miyamoto and Fitch 1995; Huelsenbeck 2002; Kondrashov et al. 2002; Choi et al. 2005; Bazykin et al. 2007; Wang et al. 2007; Rogozin et al. 2008; Rokas and Carroll 2008; Bollback and Huelsenbeck 2009; Povolotskaya and Kondrashov 2010; Gloo et al. 2010; Soylemez and Kondrashov 2012; Naumenko et al. 2012; de Juan et al. 2013; Wellner et al. 2013).

Such studies are often statistical in nature and usually cannot identify specific interactions, yet they provide a broader outlook on the nature of the fitness landscape across different areas of the sequence space. Many of these studies claimed that different aspects of molecular evolution are not compatible with evolutionary models devoid of epistatic interactions. However, the dynamics of protein sequence divergence has not been subject to modeling with explicit parameters of epistasis.

Modeling protein (or DNA) sequence evolution is often concerned with estimating the rate at which two sequences diverge typically describing sequence divergence as a Markov chain process. Initially, such models considered the neutral divergence of DNA sequences. The most general neutral, site-independent, general time-reversible (GTR) model (Tavaré 1986), which was created following more restricted models (Jukes and Cantor 1969; Kimura 1980; Felsenstein 1981), allows different substitution rates for each nucleotide pair (see O’Meara 2012 for review). Within the existing Markov chain models, the probabilities of each site being occupied by each of the four nucleotides are estimated across time using a
4 × 4 matrix of nucleotide substitution rates (fig. 1). Amino acid substitution models are analogous in a sense that a 20 × 20 matrix of amino acid substitution rates can be used to estimate the probabilities of a site being occupied by each of the 20 amino acids.

The first level of complication of these models arises from the impact of selection on sequence divergence. When the matrix of substitution rates reflects only the rate of mutation the Markov chain models reflect neutral sequence divergence. The typical way to introduce selection to these models is to assume a multiplier of the mutation rate, \( \omega \), which models the neutral rate of change of nucleotides across sites. As selection influences the rate of substitution in sites it is introduced as a parameter \( \omega \) with \( Z'[t + \tau] = Z[t] \cdot e^{\omega R \tau} \) models. In that case \( \omega > 1 \) reflects the action of positive selection and accelerates the rate of evolution and \( \omega < 1 \) reflects negative selection slowing down the rate of change of \( Z \). As the action of selection may be different in different sites, some models attempt to capture the resulting rate variation across sites by assigning a different \( \omega \) to different sites. The covarion models reflect the possibility that the rate of evolution of a site is itself subject to change with time. They introduce extra parameters allowing for sites to switch among the different \( \omega \) categories.

Sequence divergence models often use a precomputed matrix of amino acid substitution rates, such as BLOSUM or PAM (Dayhoff et al. 1978; Henikoff S and Henikoff JG 1992; Whelan and Goldman 2001).

The second level of complication of these models comes from the realization that the strength of selection may be different at different sites. A class of Markov chain models with variable rate of evolution across sites has been created (Nei and Gojobori 1986; Yang 1994) in which \( \omega \) follows a specific distribution, typically a Gamma distribution. Practically, a discrete Gamma distribution is used and sites are, therefore, classified into a number of categories with a different rate multiplier \( \omega_i \). Within this approach some sites may be completely invariable, in that case for those sites \( \omega_i = 0 \).

The final level of complication considers the possibility that the strength of selection at a site changes with time.

![General Time Reversal Model](image)

**Fig. 1.** Three categories of Markov chain models of protein evolution. The general time reversal models estimate the probability that a site is occupied by a specific nucleotide, \( Z \). The probability of finding specific nucleotides at a site changes with time and the rate of change is described by a 4 × 4 matrix, \( R \), because each of the four nucleotide can change into the other three nucleotides with a certain rate \( r_{i,j} \). The \( r_{i,j} \) rates typically reflect the rate of mutation and, therefore, \( Z'[t + \tau] = Z[t] \cdot e^{R \tau} \) models the neutral rate of change of nucleotides across sites. As selection influences the rate of substitution in sites it is introduced as a parameter \( \omega \) with \( Z'[t + \tau] = Z[t] \cdot e^{\omega R \tau} \) models. In that case \( \omega > 1 \) reflects the action of positive selection and accelerates the rate of evolution and \( \omega < 1 \) reflects negative selection slowing down the rate of change of \( Z \). As the action of selection may be different in different sites, some models attempt to capture the resulting rate variation across sites by assigning a different \( \omega \) to different sites. The covarion models reflect the possibility that the rate of evolution of a site is itself subject to change with time. They introduce extra parameters allowing for sites to switch among the different \( \omega \) categories.
Such covarion models (Fitch and Markowitz 1970) introduce parameters, which describe the rate of change between different substitution rate multipliers. The first such models allowed a site to switch between invariable and variable states (Tuffley and Steel 1998; Huelsenbeck 2002) the second ones to switch between several \( \omega_i \) categories (Galtier 2001). Wang et al. (2007) combined both approaches into a more general covarion model.

The models that incorporate selection do so by varying the rate of evolution across different sites without varying the probability of different substitution across sites. Biologically speaking we know that the fitness impact of a \( \text{L} \rightarrow \text{C} \) substitution in one site may be radically different from that of the \( \text{L} \rightarrow \text{C} \) substitution at a different site. However, assuming a different substitution matrix for different sites is impractical as it overparametrizes the model. Therefore, for the time being the assumption that sites differ from each other only in their rate of evolution, which in a covarion model is time-dependent, remains widely accepted. This assumption may not interfere with the models ability to reliably estimate the overall rate of evolution. However, such models may not be appropriate when differences in the rate of the same type of substitutions across different sites contribute substantially to the sequence divergence process.

A widely utilized verbal model of molecular evolution has been formulated by Maynard Smith (1970), which compares protein evolution to a word game, whereby two words (proteins) with meaning (function) are connected by a series of one letter (amino acid) changes (substitutions) such that a continuous pathway between these two words is created. The example used by Maynard Smith is that of the word “GENE” through meaningful intermediates within the sequence space. For the trajectory considered by Maynard Smith (1970) \( \text{WORD} \rightarrow \text{WORE} \rightarrow \text{GORE} \rightarrow \text{GONE} \rightarrow \text{GENE} \) the entire sequence space, in English, is \( 26^4 = 456,976 \). However, to understand the local restrictions imposed by the interactions of letters in different places across the word it is sufficient to describe the fitness landscape one substitution away from each of the five words in the trajectory (fig. 2). The current sequence as well as the potential fitness impact of all single letter substitutions can be shown in a matrix, which we call the fitness matrix. In the case of a four letter word in English, a cell of the fitness matrix may have three different states, A, B, or C. First, C, or the “current” letter represents the current state of the letter in a specific position. For example, the second letter of \( \text{WORD} \) is O and, therefore, the \((2,0)\) cell of the matrix has the state C. Second, A stands for a letter that is currently “available” for evolution. In the second letter of the word \( \text{WORD} \), the \((2,A)\) cell has the state A because O \( \rightarrow \) A substitution in the second letter creates the word \( \text{WARD} \), which has meaning in English. Finally, a substitution may be “blocked” (B), whereby if the substitution were to occur it would not create a meaningful word, but this letter at the same site may be present in another word. All substitutions in \( \text{WORD} \) in the second site other than \( O \rightarrow A \) are of such type. For example, \( O \rightarrow A \) substitution creates a meaningless \( \text{WJRD} \); however, the letter J at the second site can be part of an actual word, AJAR and, therefore, the \((2,J)\) cell is occupied by state B (see table 1 for a list of all abbreviations).

Given a trajectory of substitutions in sequence space, it is possible to track not only the substitutions that have occurred but also the associated changes in the fitness impact of other substitutions. The first substitution in the given trajectory is the \( \text{D} \rightarrow \text{E} \) substitution in the fourth letter. This is reflected in the \( \text{A} \rightarrow \text{C} \) switch of the cell \((4,E)\) and the reciprocal \( \text{C} \rightarrow \text{A} \) switch of the \((4,D)\) cell. Thus, a \( \text{C} \rightarrow \text{A} \) switch represents one substitution. Furthermore, each substitution may cause a number of \( \text{A} \rightarrow \text{B} \) and \( \text{B} \rightarrow \text{A} \) switches. In this case, the \( \text{D} \rightarrow \text{E} \) substitution at site 4 causes two \( \text{B} \rightarrow \text{A} \) switches in cells \((2,E)\) and \((2,I)\). This occurs because these substitutions, \( O \rightarrow \text{E} \) and \( O \rightarrow \text{I} \), now lead to meaningful words (WERE and WIRE, respectively), whereas the same substitutions prior to the \( \text{D} \rightarrow \text{E} \) substitution did not lead to a meaningful combination of letters (WERD and WIRD). If we consider the second substitution in the sequence \( \text{W} \rightarrow \text{G} \) in the first letter (leading to \( \text{WOE} \rightarrow \text{GORE} \)), another \( \text{B} \rightarrow \text{A} \) switch occurs in the cell \((2,Y)\), which leads to the word GYRE, and three \( A \rightarrow B \) switches at the same site in the cells \((2,A)\), \((2,E)\), and \((2,I)\).
which lead to the meaningless combinations of GARE, GERE, and GIRF, respectively.

The fitness matrix approach to investigating the trajectory of substitutions in the WORD→GENE analogy can be further dissected to reveal various parallels with protein evolution. Some letters (sites) tend to have more substitutions available at any given moment (the first letter compared with the second in the example), some substitutions can transition rapidly between allowed and blocked, whereas others switch between these states at a much slower rate, etc. However, investigating many or long trajectories of substitutions using the full fitness matrix may still be troublesome, although this has been done experimentally for a small number of proteins with very short trajectories (McLaughlin et al. 2012; Gong et al. 2013; Roscoe et al. 2013; Thyagarajan and Bloom 2014). Therefore, for ease of mathematical treatment we follow site-independent models and assume that all sites across the sequence have the same properties. Our model aims to study the probability that a cell in the fitness matrix is one of these three states (C, A, or B) and the rate of A→C switches with every sequence substitution (A→C switch). The GTR models investigate the rate of nucleotide substitutions on average across the entire sequence without tracking individual sites yet they can reconstruct the expected sequence divergence between the evolving and the ancestral sequence. Similarly, our model investigates the rate of switches between different states of the fitness matrix cells across the entire matrix without tracking individual sites and retains the ability to determine the expected divergences between the evolving and the ancestral sequence.

**Fitness Matrix**

The protein fitness matrix has $L$ columns and 20 rows, where $L$ is the length of the protein and 20 is the number of all

---

**Fig. 2.** The fitness matrices of words encountered in the trajectory of substitutions WORD→GENE described by Maynard Smith (1970). The fitness matrix of a specific sequence reflects both the current (C) sequence, with the state C in the corresponding cell of the matrix, as well as the fitness impact of all possible single letter substitutions. For example, in the first word in the trajectory, “WORD” there are 16 available (A) substitutions, out of 100 total possible ones, that would lead to another word in English (having high fitness). All other 84 states are blocked (B), meaning that if such a substitution were to occur would not lead to a meaningful sequence of letters. A substitution that actually occurred in the trajectory is reflected by a bidirectional C→A switch in two cells of the matrix. With every substitution the potential impact of other substitution also changes (changes between the current and the previous fitness matrix are shown in orange).
amino acids. In addition to the three aforementioned states A, B, and C we introduce another state $F$, or a “forbidden” amino acid, which confers low fitness in all possible genetic backgrounds. In the WORD<--GENE analogy, forbidden states do not appear because for each letter of the alphabet at each of the four sites there exists at least one word in English in which that letter is used at the site in question. The fitness matrix as we define it assumes a binary distribution of fitnesses, such that a genotype can have a high, 1, or low, 0, fitness without intermediate values (C and A correspond to high fitness whereas B and C to low fitness).

Because we consider trajectories of amino acid substitutions we must also take into account properties of the genetic table. Some amino acid substitutions may be available from a fitness perspective; however, they cannot occur at the present moment because more than one nucleotide substitution is required on the DNA level. We, therefore, segregate the available (A) and blocked (B) cells into mutationally “neighboring” ($n$) states, $A_n$ and $B_n$ or nonneighboring states labeled with the $f$ subscript (for “far”), $A_f$ and $B_f$. Thus, we define six distinct states: $C, A_n, A_f, B_n, B_f, F$ with $A_n$ and $A_f$ forming set A whereas $B_n$ and $B_f$ forming set B. Conversely, $A_n$ and $B_n$ form set $n$, all amino acid states that can be reached by a single substitution and $A_f$ and $B_f$ form set $f$, those amino acid states that cannot. Each cell of the fitness matrix has one of these six states and the model considers the rate of switches among them (fig. 3).

**Evolution and Switches Between States in the Fitness Matrix**

Three types of switches of the cell state are possible in the fitness matrix. First, an amino acid substitution causes one cell to switch from state $C$ to state $A_n$ and another cell to have the opposite switch—from state $A_n$ to state $C$. Second, in the same site the sets of $n$-labeled and $f$-labeled cells also change, with some amino acids that were previously more than one nucleotide substitution away become neighboring amino acids and vice versa. Third, with every substitution some previously blocked amino acids became available and some available amino acids became blocked. Thus, the dynamics of accumulating substitutions in a sequence can be described with the change of the state of cells of the fitness matrix with a total of ten different switches: $C$,$\rightarrow$,$A_n$,$\rightarrow$,$C$, $A_n$,$\rightarrow$,$A_f$, $A_f$,$\rightarrow$,$A_n$, $B_n$,$\rightarrow$,$B_f$, $B_f$,$\rightarrow$,$B_n$, $A_n$,$\rightarrow$,$B_n$, $B_n$,$\rightarrow$,$A_n$, $A_f$,$\rightarrow$,$B_f$, and $B_f$,$\rightarrow$,$A_f$ (fig. 3).

**Markov Process for States of a Single Cell in the Fitness Matrix**

We introduce an approach to analyze the state of a single cell of the fitness matrix. We describe a non-$F$ cell with a vector $Z$ that contains probabilities that the cell is in one of the five possible states, $Z = (Z_C, Z_{A_n}, Z_{A_f}, Z_{B_n}, Z_{B_f})$. The probability a state $F$ of a cell is not included because non-$F$ cells never become $F$ and vice versa. Vector $Z$ changes with time because states switch between each other. We measure the rate of the switches $s_{i \rightarrow j}$ as the expected number of switches occurring with every substitution in a site. Therefore, $s_{C \rightarrow A_n} = 1$.  

<table>
<thead>
<tr>
<th>Table 1. Abbreviations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possible states of cells in the fitness matrix</td>
</tr>
<tr>
<td>$C$</td>
</tr>
<tr>
<td>$A_n$</td>
</tr>
<tr>
<td>$A_f$</td>
</tr>
<tr>
<td>$B_n$</td>
</tr>
<tr>
<td>$B_f$</td>
</tr>
<tr>
<td>$F$</td>
</tr>
<tr>
<td>Constants</td>
</tr>
<tr>
<td>$L$</td>
</tr>
<tr>
<td>$m$</td>
</tr>
<tr>
<td>$\pi$</td>
</tr>
<tr>
<td>$\gamma$</td>
</tr>
<tr>
<td>$\delta$</td>
</tr>
<tr>
<td>$\varphi$</td>
</tr>
<tr>
<td>Rates of switches</td>
</tr>
<tr>
<td>$s_{C \rightarrow A_n}$</td>
</tr>
<tr>
<td>$s_{A_n \rightarrow C}$</td>
</tr>
<tr>
<td>$s_{A_n \rightarrow B_n}$</td>
</tr>
<tr>
<td>$s_{B_n \rightarrow A_n}$</td>
</tr>
<tr>
<td>$s_{B_n \rightarrow B_f}$</td>
</tr>
<tr>
<td>$s_{A_f \rightarrow A_n}$</td>
</tr>
<tr>
<td>$s_{B_f \rightarrow B_n}$</td>
</tr>
<tr>
<td>Markov process variables</td>
</tr>
<tr>
<td>$Z = (Z_C, Z_{A_n}, Z_{A_f}, Z_{B_n}, Z_{B_f})$</td>
</tr>
<tr>
<td>$R$</td>
</tr>
<tr>
<td>Empirical observations</td>
</tr>
<tr>
<td>$D$</td>
</tr>
<tr>
<td>$U$</td>
</tr>
<tr>
<td>$N_e$</td>
</tr>
<tr>
<td>$N_a$</td>
</tr>
<tr>
<td>$K_c$</td>
</tr>
<tr>
<td>$K_d$</td>
</tr>
<tr>
<td>$K_s$</td>
</tr>
</tbody>
</table>
Moreover, we consider the fitness impact of an amino acid state to be an independent parameter from its mutational neighborhood, that is, whether a site is in state A or B is independent of whether or not it is n or f. Therefore, $s_{A_n \rightarrow A_B} = s_{A_f \rightarrow A_B}$, $s_{B_n \rightarrow A_B} = s_{B_f \rightarrow A_B}$, $s_{A_n \rightarrow A_f}$, $s_{A_f \rightarrow A_n}$, $s_{B_n \rightarrow B_f}$, $s_{B_f \rightarrow B_n}$. The switches form a Markov process with the rate matrix, $R$:

$$R = \begin{pmatrix}
[C] & * & 1 & 0 & 0 & 0 \\
[A_n] & s_{A_n \rightarrow C} & * & s_{n \rightarrow f} & s_{A \rightarrow B} & 0 \\
[A_f] & 0 & s_{f \rightarrow n} & * & 0 & s_{A \rightarrow B} \\
[B_n] & 0 & s_{B \rightarrow A} & 0 & * & s_{n \rightarrow f} \\
[B_f] & 0 & 0 & s_{B \rightarrow A} & s_{f \rightarrow n} & *
\end{pmatrix}.$$ 

where diagonal entries are determined by the constraint that rows sum to 0.

The transition probabilities between different states can then be obtained by taking the exponent of the product of the rate matrix and time measured as the number of substitutions per site, or $e^{Rt}$. The distribution of probabilities to obtain a cell of a specific particular state at time $t$ equals:

$$Z[t] = Z[0] \cdot e^{Rt}. \quad (1)$$

Figure 4 shows an example of evaluation $Z[t]$ for $Z[0] = (1,0,0,0,0)$ and for switching rates in the same value range as that obtained for real proteins families (see next section).

**Constants**

To estimate $s_{i \rightarrow j}$ rates in the rate matrix, several constants are necessary. First, we introduce $m$, the average number of non-$F$ cells per column. Biologically, $m$ is the number of amino acids per site that can confer nonzero fitness at least one genetic background. Second, we introduce $\alpha$, the fraction of cells with state $A$ among all cells with states $A$ and $B$. Third, we introduce $\gamma$, the number of switches $A \leftrightarrow B$ that happen with one amino acid substitution. The $\alpha$ and $\gamma$ constants have an affinity in that $\alpha$ describes the relative amount of $A$ and $B$ states and $\gamma$ describes the rate of change between them. Two more constants are estimated from the genetic code. The fraction of cells with state $n$ among all cells with state $n$ and $f$, $\delta = 0.4$ calculated considering two amino acids as neighbors if any of their codons are separated by one nucleotide substitution. The number of switches $n \leftrightarrow f$, $\varphi_0 = 3.3$, also calculated as an average across all codons for all 20 amino acids. The $\delta$ and $\varphi$ constants are related in that $\delta$ stands for the number of neighboring amino acids whereas $\varphi$ stands for how rapidly this set of neighboring amino acids changes. However, we are interested only in switches between non-$F$ states. Thus, we define the number of possible $n \leftrightarrow f$ switches per site as

$$\varphi = \frac{\varphi_0 m - 1}{20 - 1}. \quad (2)$$

The constant $-1$ appears in the numerator and denominator because at each site one amino acid is $C$, which is neither $n$ nor $f$.

Using these constants we calculate the number of cells with different states in one column of the fitness matrix and the stationary probabilities of non-$F$ states to which components of vector $Z$ converge independently of the initial state (table 2). We then derive rates of switches as the number of switches divided by the number of cells from which a switch could have occurred:

$$s_{A_n \rightarrow C} = \frac{1}{N_{A_n}} = \frac{1}{\alpha \delta (m - 1)},$$

$$s_{A \rightarrow B} = \frac{\gamma}{N_{A_n} + N_{A_f}} = \frac{\gamma}{\alpha (m - 1)},$$

$$s_{B \rightarrow A} = \frac{\gamma}{N_{B_n} + N_{B_f}} = \frac{\gamma}{(1 - \alpha) \delta (m - 1)},$$

$$s_{n \rightarrow f} = \frac{\varphi}{N_{A_n} + N_{B_n}} = \frac{\varphi}{\delta (m - 1)},$$

$$s_{f \rightarrow n} = \frac{\varphi}{N_{A_f} + N_{B_f}} = \frac{\varphi}{(1 - \delta) (m - 1)}.$$ 

Thus, our model has five constants, with two of them, $\delta$ and $\varphi$, characterizing the genetic code whereas the other three, $m$, $\alpha$ and $\gamma$, characterize the fitness landscape. To estimate realistic ranges of $m$, $\alpha$, and $\gamma$, we use empirical data on evolution of protein sequences.

**Fitting Observations to Model**

Our model is aimed at testing recent statements regarding the epistatic nature of long-term protein evolution. Specifically, the model is used to fit previously obtained data on protein sequence divergence from Povolotskaya and Kondrashov (2010) and Breen et al. (2012). We thus
provide a brief description of the previously published data that we investigate here with the present model.

Polarization of amino acid substitutions with one or more outgroup sequences reveals the directionality of evolution. For example, if one sequence has an Alanine (A) and another sequence in the orthologous site contains a Threonine (T) and closely related outgroup sequences also contain a Threonine then an T→A substitution is inferred. Such polarized substitutions can then be related to a fourth, reference, sequence when the reference sequence is an outgroup to all three sequences involved in the polarization. Following our example, if the orthologous site in the reference sequence is occupied by a different amino acid in the ancestor of the sister sequences and the reference sequence and are separated by a different amino acid in the ancestor of the sister sequences, the relative rate of divergence of the sister sequences and the reference sequence. The observation that sister and reference sequences that have already diverged considerably continue to do so, that is, $N_s/N_a < 1$ for large values of sequence divergence (protein distance, $D$), has been claimed to be a consequence of epistatic interactions between sites in the evolving sequences (Povolotskaya and Kondrashov 2010).

The rate of divergence of sequences from each other can be estimated in an analogous manner to the $N_s/N_a$ measure by deconstructing $N_s/N_a$ into two independent rates of evolution—the rate of divergent evolution ($K_a$) and the rate of convergent evolution ($K_c$). The ratio of divergent evolution is estimated as $K_a/N_a$ divided by the number of sites in which an away substitution could have occurred. The estimate of the number of such divergent sites is equal to the number of sites in which the ancestral state of the two sister sequences matches the reference sequence. Similarly, $K_c$ is $N_s$ divided by the number of convergent sites, in which a toward substitution could have occurred. The number of sites that are occupied by a different amino acid in the ancestor of the sister sequences and the reference sequence are separated with only a single nucleotide substitution is the target number of convergent sites. $K_d$ was estimated to be substantially slower than the rate of synonymous divergence in 4-fold sites ($K_s$), independently of $D$ (Povolotskaya and Kondrashov 2010). Alternatively, $K_c$ is comparable to $K_s$ when $D$ is near 0 and rapidly declines as $D$ increases reaching a plateau slightly above $K_s$ (fig. 5). The dependence of $K_a/K_s$ but not $K_d/K_s$ on $D$, and similar observations (Kondrashov et al. 2002; Choi et al. 2005; Rogozin et al. 2008; Rokas and Carroll 2008; Bollback and Huelsenbeck 2009; Gloo et al. 2010; Naumenko et al. 2012; Soylemez and Kondrashov 2012), has been claimed to show support for epistasis in protein evolution but has not been modeled.

The final observation that at present requires more formal modeling deals with amino acid usage ($U$), the number of different amino acids in an orthologous site. In a large multiple sequence alignment, $U \sim 9$ (Breen et al. 2012), or approximately half of all possible amino acids. However, the same proteins exhibit a short-term rate of nonsynonymous evolution ($K_a$) up to an order of magnitude lower than $K_s \sim 0.5$ that is expected if an amino acid state has the same effect in all species. The reason why amino acids are accepted in the long-term yet rejected in the short term may be due to epistatic interactions being common (Fitch and Markowitz 1970; Maynard Smith 1970; Povolotskaya and Kondrashov 2010; Breen et al. 2012); however, this parameter has also not been put into a formal context of a macroevolutionary model.

We obtained quadruplet alignments from Povolotskaya and Kondrashov (2010) where each quadruplet alignment consisted of two sister sequences, one outgroup sequence and one reference sequence. Such quadruplet alignments were available for 572 clusters of orthologous groups (COGs), functional gene families that were predicted to have been present in the last universal common ancestor, LUCA (Mirkin et al. 2003). Reference sequence from the alignments corresponds to the sequence at $t = 0$ in our model.
Distance
As a sequence diverges from the sequence it was at \( t = 0 \), we can calculate the protein distance between them as a function of time. The protein distance (\( D \)) between two sequences is defined as 1 minus the sequence identity. The sequence identity between the sequence at time \( t = 0 \) and at \( t = 0 \) equals the probability that a cell that was in state \( C \) at \( t = 0 \) is at state \( C \) at time \( t \), that is, \( Z_C[t] \) when \( Z[0] = (1,0,0,0,0) \). This gives:

\[
D[t] = 1 - Z_C[t] \mid Z[0] = (1,0,0,0,0).
\]

We then consider the value of \( D \) at equilibrium, at \( t = \infty \) as \( D^\infty = 1 - Z_C^\infty = 1 - 1/m \). In Povolotskaya and Kondrashov (2010), the distance limit has been estimated as the time when \( N_t/N_a = 1 \). In principle, we can use \( D^\infty \) to estimate one of the constants, the number of non-\( F \) amino acids \( m \), as

\[
m = \frac{1}{1 - D^\infty}.
\]

From Povolotskaya and Kondrashov (2010) the divergence equilibrium (\( N_t/N_a = 1 \)) was estimated as \( D^\infty \approx 0.90 - 0.95 \). However, from equation (4) it is clear that when \( D^\infty \) is large a difference of just 5% leads to a large scatter of \( m \) with \( m = 10 - 20 \). Thus, although it appears that \( m \) is likely to be high, consistent with a large \( U \) observed by Breen et al. (2012), this approach is not suitable for an accurate estimation of \( m \).

Sequence Divergence and Convergence
The ratio \( K_d/K_4 \) relates the rate of divergence of nonsynonymous substitutions to the rate of 4-fold synonymous evolution. In terms of our model, a nucleotide substitution leading to amino acid with state \( A \) fixes with the same probability as a synonymous substitution. Thus, \( K_d/K_4 \) equals to the proportion of substitutions that lead to an amino acid with state \( A \), in other words it is the ratio of the number of cells with state \( A \) and the number of cells with all non-\( C \) states. Therefore,

\[
\frac{K_d}{K_4} = \frac{N_A}{20 - N_C} = \alpha \cdot m - 1 \quad 19.
\]

The ratio \( K_c/K_4 \) relates the rate of convergent amino acid substitutions to the 4-fold synonymous rate of evolution. \( K_c \) measures the rate of substitution toward the reference sequence, or toward cells in the fitness matrix which were \( C \) at \( t = 0 \). Amino acids with state \( A \) fix with the same probability as synonymous substitutions. But when calculating \( K_c \), only amino acids in the mutational neighborhood are taken into account. Thus, \( K_c/K_4 \) equals the probability that a cell that was in state \( C \) at time \( t = 0 \) is in state \( A \), at time \( t \) divided by the probability that a cell that is in state \( n \):

\[
\frac{K_c}{K_4} [t] = \frac{Z_{An}[t]}{Z_{An}[t] + Z_{Bn}[t]} \mid Z[0] = (1,0,0,0,0).
\]

We then explore equation (6) as \( t \rightarrow \infty \). At \( t = 0 \) \( Z_{An} = 0 \) and \( Z_{Bn} = 0 \) whereas \( K_c/K_4 \) is not defined, because there are no convergent sites. Applying L'Hôpital's rule, we find

\[
\frac{K_c}{K_4} \mid t \rightarrow 0 = \frac{Z_{An}^\infty}{(Z_{An}^\infty + Z_{Bn}^\infty)} = 1. \quad \text{That is in a good agreement with \( K_c/K_4 \) from Povolotskaya and Kondrashov (2010).}
\]

Second, we explore \( K_c/K_4 \) when \( t \rightarrow \infty \). From table 2:

\[
K_c \mid K_4 \mid D^\infty = \frac{Z_{An}^\infty}{Z_{An}^\infty + Z_{Bn}^\infty} = \alpha.
\]

From data we estimate \( K_c/K_4 \rightarrow 0.06 \), which implies \( \alpha = 0.06 \). \( K_d/K_4 \) is approximately constant and equals 0.02, substituting it and \( \alpha \) into equation (5) we get \( m = 7.3 \). From equation (2), we calculate \( \varphi = 1.1 \).

Estimating the Degree of Epistasis in Protein Evolution
In the previous sections, we estimated values for all constants used in the model except for \( \gamma \), which reflects the amount of epistasis in evolution. To obtain it, we fit the empirically obtained \( K_c/K_4 \) from Povolotskaya and Kondrashov (2010) to the estimated \( K_c/K_4 \mid D \) that we obtain from our model. We evaluate equation (1) numerically varying \( \gamma \) as a parameter. For a given \( \gamma \), we use the functions \( Z_C(t), Z_{An}(t), Z_{Bn}(t), Z_{Ab}(t) \), and \( Z_{Bb}(t) \) obtained from equation (1) to calculate \( K_c/K_4 \) in equation (7) and \( D[t] \) in equation (3). We repeat this process varying the \( \gamma \) parameter until we obtain a good fit between the observed and predicted \( K_c/K_4 \) across \( D \) (fig. 5). Mathematically speaking we minimize relative errors of fit:

\[
\min \gamma \frac{1}{N_D} \sum_{i=1}^{N_D} \left( \frac{K_c/K_4 \mid \text{emp}(D_i) - K_c/K_4 \mid \text{theor}(D_i)}{K_c/K_4 \mid \text{theor}(D_i)} \right)^2 ,
\]

where \( N_D \) is the number of data points for \( K_c/K_4 \mid D \). \( \gamma = 5 \pm 1 \) provides the best fit (fig. 5).

Evolution Relative to a Reference Sequence, \( N_t/N_a \)
As we have estimated all parameters of the model, we can now estimate \( N_t/N_a \):

\[
\frac{N_t}{N_a} = \frac{N_{\text{sites}}}{N_{\text{sub}}},
\]

where \( N_{\text{sites}} \) is amount of convergent (divergent) sites in protein, and \( N_{\text{sub}} \) defines how many substitutions from one convergent (divergent) site toward (away) the reference are possible. Convergent sites are those from which a single nucleotide substitution can lead to reference amino acids that is currently available. Thus, \( N_{\text{sites}} \sim Z_{An} \). For every convergent site, only one substitution can lead toward the reference amino acid: \( N_{\text{sub}} = 1 \). Sites with the amino acid which matches reference are divergent, so \( N_{\text{sub}} = Z_{Bn} \). \( Z_{An} \) equals the average number of available and neighboring amino acids per column in the fitness matrix: \( N_{\text{sub}} = N_{An} = (m - 1) \cdot \alpha \cdot \delta \). Therefore, we get:

\[
\frac{N_t}{N_a} [t] = Z_{An}[t] \cdot (m - 1) \cdot \alpha \cdot \delta = \frac{K_c}{K_d} \mid Z_{An}[t] + Z_{Bn}[t] \quad 19 \cdot \delta \cdot Z_C[t] \mid Z[0] = (1,0,0,0,0).
\]

Equation (10) allows us to convert \( K_c/K_4 \) into \( N_t/N_a \) and vice versa. We calculate \( N_t/N_a \) using rates of divergent and
convergent evolution from Povolotskaya and Kondrashov (2010) and estimate Z from equation (1) when $Z[0] = (1,0,0,0)$ and with the five estimated constants. Figure 6 shows the comparison of data (fig. 3 from Povolotskaya and Kondrashov 2010) and the estimate obtained with our model.

Usage

$$N_A = \alpha (m - 1) = 0.06 \cdot (7.3 - 1) = 0.38$$

indicates how many amino acid substitutions are available at a given moment. Thus, on average less than 1 amino acid substitution per site is acceptable at a time. However, when longer time periods are taken into account, up to eight different amino acids may be found at the same site across different species (Breen et al. 2013). These two observations can be reconciled if $\gamma$, the number of $A \leftrightarrow B$ switches associated with one $C \leftrightarrow A_p$ switch, is relatively high. For the data from Povolotskaya and Kondrashov (2010), our model predicts $\gamma \approx 5$. Such a high $\gamma$ implies that per site there are five times more switches between the available and blocked states of amino acids than there are actual amino acid substitutions at the same site. It follows that the distribution of available amino acids changes substantially even for a very small number of substitutions in that site. Therefore, we can model all non-$F$ amino acids ($m$) as having an equal probability to occur at a site after a substitution and the usage expected to be observed after a certain amount of sequence divergence is a simple combinatorial problem.

Let $N$ be the number of substitutions observed per site with every substitution being a random choice from $m$. Reformulating the problem, we have a pool of objects of $m$ different colors, we then take $N$ random objects from the pool and calculate $n$, the number of different colors out of $N$ objects taken from the pool. The value of $n$ varies across trials but we consider amino acid usage to be equal expectation of $n$ distribution. The probability to get $n$ different colors in set of $N$ objects is given by the number of possible sets with $n$ different colors divided by the total number of possible sets, which is $m^N$. The number of ways to partition $N$ objects into $n$ nonempty subsets equals the Stirling number of the second kind, $S(N,n)$. There are $m!/(m-n)!$ ways to associate these subsets with $n$ different colors:

$$p[n] = \frac{S(N,n)}{m^n} \cdot \frac{m!}{(m-n)!}, \text{usage} = E(n)$$

We estimate $m = 9$ (see next section) as the mean value of $m$ distribution among different COGs (fig. 7) and using equation (11) calculate the dependence of usage on $N$ (fig. 8). Due to the high $\gamma$ value, the amino acid usage is considerable even for $N \sim 10$, number of substitutions per site, with usage almost approaching its maximal value with 30 substitutions per site.

Considering the Distribution of Parameters across Gene Families

In the previous sections, we used data which were obtained by taking average values across all 572 COGs (Povolotskaya and Kondrashov 2010). Here, we use data on the same variables that were obtained for each COG separately. For some protein families, there are not enough data to calculate $K_{d}/K_{a}$ and $K_{d}/K_{s}$ with an acceptable degree of precision. Therefore, we selected 119 COGs for which $K_{d}/K_{a}$, and $K_{d}/K_{s}$ were defined across distances 0.1–0.8 without strong outliers. For each COG, we obtained $K_{d}/K_{a}$ as an average $K_{d}/K_{s}$ across different $D$ between the trios and the reference sequence in bins of 0.1 of $D$. We presume that $\alpha$ is a limit of $K_{d}/K_{a}$ when $D \rightarrow 1$. Then, we calculate $m$ using equation (5) and $\varphi$ using equation (2). Finally, we fit $K_{s}/K_{a}$ with numerical evaluation of equation (1) and find $\gamma$. Distributions of these parameters for the 119 COGs are shown on figure 7.

Discussion

The model presented here makes three assumptions. The first assumption is that fitness is either 0 or 1. Binary fitness landscapes, while unrealistic in certain situations, are nevertheless useful when recapitulating the multidimensionality of the sequence space (Gavrilets 1997; Gavrilets and Gravner 1997; van Nimwegen et al. 1999; Aita et al. 2003; Gravner et al. 2007). When fitness values are either 1 or 0 mutations are either neutral or lethal and all permitted substitutions have an equal probability of occurrence. Thus, the impact of slightly deleterious and beneficial substitutions on molecular evolution cannot be taken into account, even though their contribution may be considerable (Ohta 1998; Andolfatto 2005; Popadin et al. 2007; McCandlish et al. 2013). The reason for excluding the effects of such substitutions is 2-fold: First, the practical considerations of modeling a complex and multidimensional fitness landscape in which substitutions are not equal in effect. The nonepistatic impact of slightly deleterious and beneficial substitutions on evolution in the absence of epistasis is well characterized (Crow and Kimura 1970), whereas the effects of epistasis on long-term molecular evolution have not been subject to the same level of scrutiny. Our work develops a null model that may be expanded to incorporate these effects. Second, it appears certain that neither the fixation of slightly deleterious or beneficial alleles can serve as the basis for explaining the measurements considered here, at least without the impact of epistasis. The fixation of slightly deleterious alleles is
compatible with a large $U$ under some circumstances (McCandlish et al. 2013). However, slightly deleterious alleles alone cannot lead to large sequence divergence in fast-evolving proteins and are incompatible with reaching maximum sequence divergence distances substantially slower than the rate of divergence of neutral sequences (Kondrashov et al. 2010). There is a tradeoff between the contribution of slightly deleterious alleles toward amino acid usage and sequence divergence; it is not possible to have a high $U$ and a high $D$ at the same time solely due to the contribution of fixation of deleterious alleles (Breen et al. 2013). Furthermore, periodic fixation of nonepistatic slightly deleterious alleles cannot lead to a decline in the rate of convergent evolution, in contrast to the observed $K_d/K_u$ relationship (Povolotskaya and Kondrashov 2010). Therefore, slightly deleterious alleles may contribute to molecular evolution but a model based on the accumulation of slightly deleterious alleles alone cannot explain all of the available data (also see Kondrashov et al. 2010). A similar argument is applicable to the impact of beneficial alleles on observations of molecular evolution considered here. Large sequence divergence observed between sequences is compatible with most differences between sequences having been fixed by positive selection. However, the high values of $K_d/K_u$, especially for small $D$, are incompatible with sequence divergence being largely driven by the accumulation of beneficial alleles.

The second assumption is that all sites in the protein are expected to have the same properties (described by $\gamma$, $\varphi$, $m$, and $\delta$). This includes genetic properties, such as the number of mutational neighbors at each site, or the properties of the protein, such as the number of available, blocked, and forbidden amino acids at a site. Indeed, it may be possible to alleviate this assumption by introducing a variation of our model in a manner similar to that has been done for models of the rate of protein evolution (Nei and Gojobori 1986; Yang 1994). We hope that the present work will be a stepping stone in this direction. The third assumption is that all of these five parameters are independent of each other. This may not be entirely the case, for example amino acids in the mutational neighborhood of each other are more likely to be available (Freeland et al. 2000).

Our model is not unique in modeling interactions between different alleles to describe the fitness landscape. However, the model presented here differs substantially from models developed previously. Previous efforts were principally concerned with the issue of the shape of the fitness landscape...
(Kauffman and Levin 1987; Macken and Perelson 1989; Gavrilovs 1997; Gavrilovs and Gravner 1997; Ohta 1998; Kondrashov FA and Kondrashov AS 2001; Gravner et al. 2007; Ferrada and Wagner 2010; Crona et al. 2013; Lobkovsky et al. 2013). Our model, on the other hand, considers patterns of long-term divergent and convergent molecular evolution as a factor of the epistatic interactions arising on a multidimensional fitness landscape. Our model shares the same conceptual basis as covarion models (Fitch and Markowitz 1970; Tuffley and Steel 1998; Galtier 2001; Huelsenbeck 2002), which allow the rate of evolution at a site to change over time. Our model approaches the same issue but from a perspective that, in our opinion, more accurately reflects the biological basis for interaction between sites.

Our analysis has revealed several important aspects of protein evolution. The crucial aspect of our model is that we can find a limited parameter space that is consistent with all of the available observations of protein evolution. The three estimated parameters that reveal the nature of the fitness landscape are the number of amino acids that in principle can confer nonzero fitness (m), the fraction of which allowed for substitution in one moment (α), and the number of switches between amino acids in available and blocked states per a single amino acid substitution (γ).

We estimate an m between 5 and 15 for most protein families (fig. 7), which indicates that few states are forbidden, that is, correspond to low fitness regardless of the amino acid composition of other sites. Here, a modest number of universally forbidden amino acids implies that most amino acid sites can accept most of the possible amino acids given the right combination of states at other sites (Fitch and Markowitz 1970; Maynard Smith 1970; Povolotskaya and Kondrashov 2010). Conversely, this implies that a single amino acid site can accept many different amino acids, consistent with the previous observations of a high amino acid usage (Breen et al. 2012).

We estimate α = 0.06, the fraction of available amino acids at a given time as just a small fraction of all amino acid states. Therefore, at any given time many sites cannot accept any substitutions. This estimate is consistent with the observation of slow rate of sequence divergence of sequences in our data set, with Kd/Ks ~ 0.02 and with mutational studies (Guo et al. 2004). Evidently, for faster evolving protein families α is likely to be higher, however, as we considered old gene families, which tend to be conservative, our estimate is unlikely to be representative of the entire diversity of protein rates of evolution found across all taxa.

Within the confines of our model, it is not possible to maintain γ = 0 while maintaining a fit to all of the observed parameters of protein evolution. Specifically, γ appears to be the key parameter that allows the model to fit the observation of the decline in the rate of convergent evolution, Kc/K4 (fig. 5). Because the model is scaled to the number of amino acid substitutions that occur in a protein the γ = 5 estimate cannot be taken as a direct indication that epistatic interactions are intragenic. Individual proteins do not evolve in isolation and cases of intergeneric epistatic interaction have been documented (Lehner 2011). Indeed, our model does not imply a causative interaction between amino acid substitutions and switches between available and blocked states of amino acids in the fitness matrix. Indeed, a general interpretation of γ = 5 is that on average five switches between available and blocked states occur on the same timeframe as a single amino acid substitution. These switches may either be a consequence of changes in the same protein or in the rest of the genome. If most interactions are intragenic then the best fit of γ = 5 implies that the fitness matrix of a protein is changing faster than its sequence implying that the inherently epistatic nature of the fitness landscape is an inseparable and defining factor of molecular evolution.

Our model can be used to analyze the patterns of molecular evolution of specific gene families, as was done here (fig. 7). We observed that all parameters, including γ, vary substantially across different gene families, implying that the nature of the fitness landscape is a feature that may differ across protein families. The application of this model to a broader set of proteins than considered here can lead the way toward a more general characterization of fitness landscapes and evolutionary trajectories in nature.

Acknowledgments

The work was supported by grants from Agence Nationale de la Recherche (ANR-12-JSV7-0007), HHMI International Early Career Scientist Program (55007424), the EMBO Young Investigator Programme, MINECO (BFU2012-31329 and SEV-2012-0208), and an ERC Starting Grant (335980_EinME). All authors participated in the design of the model. D.R.U. and L.F. performed the mathematical analysis. D.R.U and I.S.P. obtained and analyzed data on protein evolution. D.R.U. and F.A.K. wrote the draft.

References


