Sequence Entropy and the Absolute Rate of Amino Acid Substitutions

Richard A. Goldstein¹ and David D. Pollock²

¹Division of Infection & Immunity, University College London, London, WC1E 6BT, UK.
²Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045 USA.

The evolution of model proteins under selection for thermodynamic stability suggests parallels between evolutionary behavior and chemical reaction kinetics. We developed a statistical mechanics theory of protein evolution by dividing amino acid interactions into site-specific and ‘bath’ components, and show that substitutions between two amino acids occur when their site-specific contributions to stability are nearly identical. Fluctuating epistatic interactions drive stabilities into and out of these regions of near neutrality, with the time spent in the neutral region and thus the rate of substitution governed by physicochemical similarities between the amino acids. We derive a theoretical framework for how site-specific stabilities are determined, and demonstrate that substitution rates and the magnitude of the evolutionary Stokes shift can be predicted from biophysics and the effect of sequence entropy alone. Population genetics underlays our analysis, but population size does not determine the absolute rate of amino acid substitutions.

Introduction

Modeling the rate at which protein sequences change is central to understanding how proteins adapt to their structural, functional, and thermodynamic requirements. It is also key to deciphering the patterns of conservation and variation that reflect evolutionary processes and the properties of specific proteins. An important step was Kimura’s calculation of the probability of fixation of a single mutation given constant relative fitnesses of the wild type and mutant (1-3). Fixation probabilities alone, however, do not address how or why fitness differences come to be, and therefore cannot explain observed substitution rates. Empirically derived substitution rates have long been obtained by analyzing differences between related protein sequences (4-6), providing estimates of average rates but not explaining them. Although this approach has been extremely useful, its successes were achieved by ignoring the underlying biophysics, molecular biology, and population dynamics, as well as how these rates vary amongst sites and time.

In recent years, the number of protein sequences, computational speeds, and our knowledge of protein biophysics have increased substantially. This has led to an expansion in the potential scope of evolutionary analyzes and a growing awareness of the limitations of standard empirical models. Proteins are under selection for traits – function, foldability, stability, solubility – that depend on a complex network of interacting amino acids. These forms of selection induce epistatic interactions (or coevolution) among sites in the protein, resulting in substantial effects on the evolutionary process (7-12). Models that ignore this epistasis can seriously compromise evolutionary analyzes by misrepresenting the frequency and time
dependence of convergence and homoplasy (13). Empirical models can be modified to allow the substitution process to vary among sites (14, 15) and over time (6, 16-20), but information available from sequences to obtain accurate site- and time-dependent substitution rates is fundamentally limited. Efforts to use protein structure to predict substitution rates (21, 22) are compromised by our lack of understanding of the relationship between protein sequence, protein properties, and organismal fitness, and our inability to predict the effect of mutations as differences accumulate.

The development of more accurate and powerful models of protein evolution depends on our ability to represent the process of molecular evolution at a mechanistic level, ideally enabling us to calculate substitution rates based on the protein’s sequence and biophysical properties. Our purpose here is to develop from first principles a theory of how proteins evolve and how substitution rates are determined. We approach the problem by building a conceptual framework to translate protein evolution into the formalisms of statistical mechanics, demonstrating the primacy of sequence entropy. Using evolutionary simulations of model proteins, with fitness determined by thermodynamic stability, we demonstrate that substitution rates depend on how amino acid energy contributions fluctuate as the rest of the protein sequence evolves. We show that substitution rates can be predicted based on 1) the stability distributions at a site in the absence of selection on that site; and 2) the relative numbers of sequences with different protein stabilities; no other adjustable parameters, such as expected population size, are needed. This forms a mechanistic framework for constructing improved models of amino acid substitution rates.

Results

Site-specific stabilities and relative substitution rates

To develop a mechanical theory of the evolutionary process, we consider the relationship between protein stability and substitution rates at a site. The stability \( \Xi(X) \) of a protein sequence \( X = \{x_1, x_2, x_3 \ldots x_n\} \) was defined as the negative of the free energy difference between the sequence in the native structure and in the ensemble of possible alternative structures, so that more positive values indicate greater stability. The Malthusian fitness \( m(X) \) was set equal to the fraction of such sequences that would be folded in a pre-specified native conformation at thermodynamic equilibrium (Equation (2, Methods) (11, 23, 24). Thus, increases in stability lead to increases in fitness.

To understand how the rest of the protein influences the substitution rate at individual sites, we focus our attention on an amino acid \( \alpha \) at a specific focal site \( k \), and partition the stability into \( \Xi(X) = \xi_{k,\alpha}(X_{\beta \neq k}) + \xi_{k,Bath}(X_{\beta \neq k}) \). The first term, \( \xi_{k,\alpha}(X_{\beta \neq k}) \), is the site-specific stability contribution due to interactions (in both the folded and unfolded states) between \( \alpha \) at site \( k \) and the amino acids at all other sites excluding \( k \). The second term, \( \xi_{k,Bath}(X_{\beta \neq k}) \), is the ‘background’ contribution resulting from interactions among amino acids at sites excluding the focal site. Because only a small fraction of contacts involve site \( k \), we assume that the site-
specific stability contribution is small relative to the background contribution, so that this second term fulfills the role of the ‘thermal bath’ in statistical physics.

This statistical mechanics formalism can now be applied to modeling the amino acid substitution rate. Consider $Q_{k,\alpha \to \beta}(X_{\beta k})$, the instantaneous rate of an $\alpha$ to $\beta$ substitution at site $k$, equal to the mutation rate times the fixation probability. The fixation probability depends on the difference in fitnesses $\Delta m_{k,\alpha \to \beta}$, which is a function of the initial stability $\Xi(X)$ and the stability of the mutant $\Xi(X') = \Xi(X) + \Delta \Xi_{k,\alpha \to \beta}(X_{\beta k})$. We can simplify the situation by noting that sequences from real proteins, as well as proteins from evolutionary simulations under selection for thermostability, tend to have a narrow range of stability values (23, 25-28). This stability range occurs where the decreasing effectiveness of selection for greater stability is balanced by destabilizing mutations fixed by genetic drift. The precise value depends on a variety of factors such as temperature, effective population size, sequence length and protein function. As long as these factors are approximately constant, we can assume that a given protein will evolve to the mean of this narrow range $\Xi(X) = \bar{\Xi}$. If $\bar{\Xi}$ is a known constant, calculating the fixation probability requires only the difference in site-specific stabilities $\Delta \Xi_{k,\alpha \to \beta}(X_{\beta k}) = \xi_{k,\beta}(X_{\beta k}) - \xi_{k,\alpha}(X_{\beta k})$; the bath component, $\xi_{k,Bath}(X_{\beta k})$ is independent of the amino acid at focal site $k$, and is therefore unchanged by the substitution.

From this perspective, the key distribution determining the substitution rate from $\alpha$ to $\beta$ is $\rho_{k,\alpha,\beta}(\xi_{k,\alpha}, \xi_{k,\beta})$, the joint probability density of $\xi_{k,\alpha}(X_{\beta k})$ and $\xi_{k,\beta}(X_{\beta k})$ given that amino acid $\alpha$ is resident at site $k$, integrating over the distributions of amino acids at other locations. The distribution depends on which amino acid occupies position $k$ because that amino acid will have affected the evolution in the rest of the protein; in this case, $\xi_{k,\beta}$ is the local stability contribution that would result if $\beta$ were to replace $\alpha$ at that site with no other changes in the sequence. For simplicity, we will consider that the rest of the protein sequence has evolved sufficiently that $\rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta})$ has reached a stationary distribution; the effect of a breakdown in this assumption will be considered below.

To help visualize these distributions, and evaluate our theoretical results, we modeled the evolution of real proteins using the simulated evolution of a 300-residue protein under selection for thermodynamic stability. This model is not meant to make quantitative predictions in particular cases. Instead, it is meant to predict general characteristics of evolutionary behavior for proteins that require the native confirmation to carry out some critical biological function, and has demonstrated its ability to reproduce fundamental aspects of the evolutionary process (11, 23, 24). By using a simple pair-contact model of protein thermodynamics, we were able to perform replicate simulations over long periods of evolutionary time, corresponding to approximately 5 billion years given typical substitution rates.

We grouped sites with similar substitution patterns into four different site classes, where class 1 is the most exposed and 4 is the most buried. Figures 1A-D shows the observed joint probability distributions of these site classes for glutamic acid and lysine, as well as the stability
Figure 1: A-H) Relative local contributions to stability for pairs of amino acids in different site classes (A-D) or different pairs of amino acids in the same site class (E-H). Points were sampled either when the amino acid in the abscissa is resident (blue), when the amino acid in the ordinate is resident (green), or during transitions between the two (yellow). I-L: Distributions of local contributions to stability in reference state when the non-interacting null amino acid was present ($\rho_{k,\alpha}(\xi_{k,\alpha},\xi_{k,\beta})$, magenta), when the amino acid in the abscissa was present as predicted using Equation 1 ($\bar{\rho}_{k,\alpha}(\xi_{k,\alpha},\xi_{k,\beta})$, cyan), or as observed ($\rho_{k,\alpha}(\xi_{k,\alpha},\xi_{k,\beta})$, blue).

Distributions when substitutions between these two amino acids occurred. Figures 1E-H shows these distributions for four different pairs of amino acids in site class 3. There are wide ranges of values for $\xi_{k,\alpha}$ and $\xi_{k,\beta}$, consistent with earlier results demonstrating fluctuating selective pressures at sites due to substitutions elsewhere in the protein (11). The distributions of $\xi_{k,\alpha}$ and $\bar{\xi}_{k,\beta}$ strongly depend on the resident amino acid. In particular, the potential contribution of an amino acid to the protein stability tends to be greater when that amino acid is resident at a site, a phenomenon we previously named the ‘evolutionary Stokes shift’ (11). The amount of this increase appears to be correlated with the observed variance in $\xi_{k,\alpha}$.

Rapidly evolving sites with few selective constraints tend to have compact distributions with smaller variances in $\xi_{k,\alpha}$ and $\xi_{k,\beta}$ than slowly evolving sites (Figures 1A-D). Distributions for physicochemically similar amino acids (e.g., aspartic acid versus glutamic acid, Figure 1E) appear highly correlated, while those for dissimilar amino acids (e.g., arginine versus leucine, Figure 1H) seem anti-correlated. This is because background sequences that confer a high site-specific stability on aspartic acid tend to do the same for the highly similar glutamic acid, while background sequences that stabilize arginine tend to destabilize the dissimilar leucine (29). A
non-resident amino acid is generally stabilized if the distributions are correlated (e.g. glutamic acid when aspartic acid is present, Figure 1E), but destabilized if the distributions are anti-correlated (e.g. glutamic acid when alanine is present, Figure 1F).

To determine whether substitution rates can be predicted from $\rho_{k,\alpha}(\bar{\xi}_{k,\alpha}, \bar{\xi}_{k,\beta})$, class-specific stability distributions were modeled with the best fitting bivariate normal distribution for each pair of amino acids

$$
\rho_{k,\alpha}(\bar{\xi}_{k,\alpha}, \bar{\xi}_{k,\beta}) = \mathcal{N}\{\bar{\xi}_{k,\alpha|\alpha}, \bar{\xi}_{k,\beta|\alpha}, \sigma^2_{k,\alpha|\alpha}, \sigma^2_{k,\beta|\alpha}, \rho_{k,\alpha|\alpha}\}. $$

The expected substitution rates between each pair were then estimated by numerical integration over these distributions using Kimura’s formula for the probability of fixation (30-32) (see Equation (3, Methods). There is extremely good agreement between expected substitution rates derived from this approximation and substitution rates obtained by counting substitutions that occurred during the simulations (Figure 2). This validates the utility of the bivariate normal approximation and the assumption that variation in $\Xi(X)$ has little effect on substitution rates.

A striking feature of Figure 1 is the strong tendency for substitutions to occur in the overlap region between $\rho_{k,\alpha}(\bar{\xi}_{k,\alpha}, \bar{\xi}_{k,\beta})$ and $\rho_{k,\beta}(\bar{\xi}_{k,\alpha}, \bar{\xi}_{k,\beta})$, centred on the diagonal $\Delta\Xi_{k,\alpha \rightarrow \beta} = \bar{\xi}_{k,\beta} - \bar{\xi}_{k,\alpha} = 0$ where substitutions are neutral. This suggests the possible applicability of transition state theory (TST), a method for predicting the rate of chemical reactions (33). In TST, the reaction rate is given by the fraction of reactants in a ‘transition state’ in which the energies of reactant and product are approximately equal, times the rate of conversion from transition state to products. Adapting this theory, we model the substitution rate as equal to the fraction of joint stabilities for which the fitness of wild type and mutant are approximately equal, times the rate of substitution under neutral conditions.

The probability that the background sequence results in nearly equal fitnesses between $\alpha$ and $\beta$ at site $k$ was estimated as the density $\rho_{k,\alpha}(\bar{\xi}_{k,\alpha}, \bar{\xi}_{k,\beta})$ integrated along the neutral line $\bar{\xi}_{k,\alpha} = \bar{\xi}_{k,\beta}$, multiplied by the width of the neutral zone on both sides of the neutral line, $2\varepsilon$, the region in which the effect of selection is small. The neutral substitution rate is equal to the mutation rate $v_{\alpha \rightarrow \beta}$, allowing us to write a closed-form expression for the average substitution rate (Equation (7, Methods).

As described in the Methods section, the extent of the neutral zone, $\varepsilon$, can be naturally defined by the falloff in the number of sequences with greater stabilities. Because the stability values $\Xi$ for folded proteins represent the far tail of a distribution dominated by unstable sequences, we
modeled $\Omega(\Xi)$, the number of sequences with stability $\Xi$, as an exponential $\Omega(\Xi) \propto \exp(-\gamma \Xi)$, where $\gamma$ characterizes the decrease in number of sequence with increasing stability; thus, the bias of the drift effect does not depend on $\Xi$. The scale of the neutral zone is given by $1/\gamma$, which is equal to the range of stabilities at which the fitness changes by less than $\frac{1}{4N_e}$ regardless of population size (see Methods). To calculate substitution rates, we estimated $\gamma = 1.26$ (kcal mol$^{-1}$)-1 based on the relative numbers of destablizing and stabilizing mutations, yielding $\varepsilon = 0.79$ kcal mol$^{-1}$. Notably, because this calculation considers only neutral substitutions, it produces strikingly accurate predictions (Figure 2) without the need for Kimura’s formula.

The equilibrium distributions of site-specific stabilities: The mechanism behind the evolutionary ‘Stokes Shift’

It appears that the rate of amino acid substitutions is substantially determined by $\rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta})$ in the regions where $\xi_{k,\alpha} = \xi_{k,\beta}$. Our goal for the rest of the paper is to show the degree to which these distributions, and therefore substitution rates, can be explained using the principles of statistical mechanics.

As above, we assume that proteins evolve to a specific stability value $\Xi(X) = \Xi$. All sequences with stability $\Xi$ have, in our model, identical fitnesses, so none are preferred over another by selection. If evolution has had sufficient time to sample from the stationary distribution, the fraction of sequences with any property $\theta$ is proportional to $\Omega(\theta, \Xi)$, the number of sequences with property $\theta$ and stability $\Xi = \Xi$. The log of this quantity, $S(\theta, \Xi) = \ln[\Omega(\theta, \Xi)]$, is the ‘sequence entropy’ of such sequences, analogous to thermodynamic entropy. Under these conditions, the probability of property $\theta$ is given by $p_{\text{Eq}}(\theta) = \frac{\Omega(\theta, \Xi)}{\Omega(\Xi)} = e^{S(\theta, \Xi) - S(\Xi)}$, where $\Omega(\Xi)$ and $S(\Xi)$ are the number of sequences with stability $\Xi = \Xi$ and the log of this quantity.

To calculate $\rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta})$, an estimate of $\rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta})$, we first considered $\Omega_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta}, \Xi)$, the number of sequences with stability $\Xi = \Xi$, $\alpha$ resident at site $k$, and site-specific stability contributions $\xi_{k,\alpha}$ and $\xi_{k,\beta}$ at that location. We approximated this number as the product of $\Omega_{k}^{\text{Loc}}(\xi_{k,\alpha}, \xi_{k,\beta})$, the number of amino acid arrangements resulting in the site-specific $\xi_{k,\alpha}$ and $\xi_{k,\beta}$, times $\Omega_{k}^{\text{Bath}}(\xi_{k,\beta} = \Xi - \xi_{k,\alpha})$. The latter term is the number of sequences furnishing the background stability required to complement the site-specific contribution furnished by $\xi_{k,\alpha}$, for a total stability equal to $\Xi$. This calculation assumes independence of the bath and local contributions to total stability; although not strictly accurate (the relevant sites in the protein overlap), it is likely to be approximately true because the interactions involved are different.

We note that $\Omega_{k}^{\text{Loc}}(\xi_{k,\alpha}, \xi_{k,\beta})$ does not depend on selection, so to characterize it requires removing selection at site $k$. To do this we performed simulations with the focal site permanently occupied by a non-interacting amino acid, $\emptyset$, and with all other sites evolving freely. The resulting distributions $\rho_{k,\beta}(\xi_{k,\alpha}, \xi_{k,\beta})$ are proportional to $\Omega_{k}^{\text{Loc}}(\xi_{k,\alpha}, \xi_{k,\beta})$, and
represent the null distributions of local stability contributions that would occur if interactions between amino acids at site \( k \) and the rest of the protein did not affect the evolutionary dynamics. Because the number of possible sequences is immense, and because \( \xi_{k,\alpha} \) and \( \xi_{k,\beta} \) are the result of many interactions, the central limit theorem suggests that \( \rho_{k,\theta}(\xi_{k,\alpha}, \xi_{k,\beta}) \) can be approximated by a bivariate normal distribution

\[
\rho_{k,\theta}(\xi_{k,\alpha}, \xi_{k,\beta}) \propto N\{\bar{\xi}_{k,\alpha|\theta}, \bar{\xi}_{k,\beta|\theta}, \sigma_{k,\alpha|\theta}^2, \sigma_{k,\beta|\theta}^2, \varphi_{k,\alpha|\theta}, \varphi_{k,\beta|\theta}\}.
\]

Interactions involving the focal amino acid represent a small fraction of total interactions, allowing us to approximate \( \Omega_k^{\text{Bath}}(\xi_{k,\text{Bath}}) \propto \Omega(\xi = \xi_{\text{Bath}}) \). The normalized product of \( \rho_{k,\theta}(\xi_{k,\alpha}, \xi_{k,\beta}) \) and the exponential \( \Omega(\xi = \xi_{k,\text{Bath}} = \bar{\xi} - \xi_{k,\alpha}) \) results in a shifted bivariate normal distribution

\[
\tilde{\rho}_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta}) = N\{\bar{\xi}_{k,\alpha|\alpha}, \bar{\xi}_{k,\beta|\alpha}, \sigma_{k,\alpha|\alpha}^2, \sigma_{k,\beta|\alpha}^2, \varphi_{k,\alpha|\alpha}\}
\]

(1)

Selection in the presence of amino acid \( \alpha \) at site \( k \) shifts the average local contribution to stability by an amount \( \xi_{k,\alpha|\alpha} = \gamma \sigma_{k,\alpha|\theta}^2 \) compared to its contribution to stability in the absence of interactions; this stabilization can be viewed as the basis for the evolutionary Stokes shift. The mechanism for the shift is the large increase in sequence entropy gained from a decrease in \( \xi_{k,\text{Bath}} \), combined with the trade-off between \( \xi_{k,\alpha} \) and \( \xi_{k,\text{Bath}} = \bar{\xi} - \xi_{k,\alpha} \).

The fit between estimated equilibrium values of \( \bar{\xi}_{k,\alpha|\alpha} \) for each amino acid and values of \( \bar{\xi}_{k,\alpha|\alpha} \) calculated directly from the simulations is surprisingly good given the approximations made (Figure 3A). The entropic stabilization as a function of \( \sigma_{k,\alpha|\theta}^2 \) is linear (correlation coefficient 0.857; Figure 3B) as predicted by Equation (1). The slope of 1.00 (kcal mol\(^{-1}\))\(^{-1}\) (95% CI: 0.87 – 1.14) is close to the expected value of \( \gamma = 1.26 \) (kcal mol\(^{-1}\))\(^{-1}\), confirming the trends evident in Figure 1. The observed entropic stabilization is smaller than predicted for the two largest shifts in the slowest rate class, involving the negatively charged aspartic acid and glutamic acid. Earlier work demonstrated that equilibration for the most buried states can be extremely slow(11), and these outliers may represent cases where the protein has had insufficient time to adjust to the presence of the new amino acid.

The key result here is that the magnitude of the entropic stabilization that drives the evolutionary Stokes shift depends only on the number of protein sequences with given protein stabilities and on the underlying distributions of interactions in the absence of selection: the effect can be understood purely in terms of biophysics and sequence entropy.
The predicted and observed distributions of $\rho_i, \xi_i$ are shown in Figures 1L. The values of $\xi_i$ are shifted by an amount $\zeta_i = \gamma \phi_i \sigma_k,\alpha|\emptyset \sigma_k,\beta|\emptyset$ to either higher or lower values depending on the physicochemical similarities between the amino acids. From Equation (1) we can see that the realized evolutionary ‘Stokes shift’ after a substitution, the expected average difference in stability before and after the protein adjusts to the new resident amino acid, is equal to $\zeta_i - \zeta_i = \gamma (\sigma_k,\beta|\emptyset - \phi_k,\alpha|\emptyset \sigma_k,\alpha|\emptyset \sigma_k,\beta|\emptyset)$. The full entropic stabilization is reduced by $\zeta_i$, which can be viewed as the average amount of preadaptation (or lack thereof) to amino acid $\beta$ caused by the residency of amino acid $\alpha$. As with the average entropic stabilization, the realized evolutionary Stokes shifts depend deterministically on the site-specific stability distributions in the absence of selection with no adjustable parameters. Substitution rates estimated with the TST approximation (Equation (7) using the site-specific stabilities calculated from Equation (1) are remarkably accurate for all four site classes and over four orders of magnitude of rate variation (Figure 4).

**Discussion**

The understanding of evolutionary mechanics developed here represents a fundamental shift in
how we conceptualize the process of amino acid substitution. Although stability is approximately constant, the way this stability is partitioned among various interactions fluctuates as a protein evolves. In particular, the contribution that a resident amino acid at a site makes to the stability of the protein, as well as the contribution a non-resident amino acid would make if substituted in, will fluctuate. Occasionally, these fluctuations lead to approximately equal stabilities for a pair of amino acids so that substitutions from one to the other are nearly neutral. The frequencies of these nearly neutral states then determine the relative substitution rates. The fluctuations in stabilizing contributions of different amino acids at a site are not superfluous or unwanted complications in the construction of substitution models, but rather are central to the substitution process. In evolutionary theory it is common to evoke the idea of a fixed ‘adaptive landscape’, but for a single amino acid position a more appropriate analogy may be a fluctuating adaptive seascape; the site explores the space of possible amino acids by moving along fluctuating local contours in the context of approximately constant overall fitness.

By developing a statistical mechanics view of protein evolution, the evolutionary Stokes shift can be seen as a direct consequence of sequence entropy. Increases in the stabilizing contributions of an amino acid occupying a given site reduce the amount of stabilization required by the rest of the sequence, increasing the number of sequences that can contribute this reduced stability. Our theoretical analysis of the balance between the number of states available to the system (the amino acid at the focal site and its interactions) and the ‘bath’ (the rest of the sequence) yields an expectation that the relative magnitude of entropic stabilization of an amino acid at a site is proportional to the variance of the underlying null site-specific stability distribution. Furthermore, the stabilization of all amino acids at all sites are scaled by a protein-wide proportionality constant determined by the decline in the number of available sequences as protein stability increases. Thus, surprisingly, the strength of selection and the effective population size do not affect the evolutionary Stokes shift or substitution rates if the protein is in a steady state(34, 35). Thus, although our evolutionary mechanics theory fully incorporates population genetics theory and Kimura’s equation for the probability of a substitution, if the system is near equilibrium we do not need Kimura’s formula to predict and explain substitution rates among amino acids.

Correlations in the fluctuations between amino acids with similar physicochemical properties increase the probability of near-neutrality, providing a mechanistic explanation for higher rates of conservative change, a general phenomenon rationalized by Fisher with his geometric argument(36). The probability of occupying the neutral zone is lower at interior sites, where the multiplicity of interactions with the focal site increase the distance between $\xi_{k,\alpha}$ and $\xi_{k,\beta}$, and correspondingly higher at surface sites; this is consistent with observed slower internal (buried) than external (surface) substitution rates.

For dissimilar amino acids, the probability of achieving the near-neutrality required for a substitution can be unlikely. However, if such a substitution occurs the protein will subsequently evolve to sequences that partition a larger stability contribution to the newly resident amino acid, causing an increased affinity for this residue. This increased affinity is what
we have called the evolutionary Stokes shift. This evolutionary mechanism can be fully reversible, as is in our evolutionary simulations, with the reversibility coming from the similarity in the processes of moving into and away from the neutral zone (11). These processes, called ‘contingency’ and ‘entrenchment’ by Plotkin and colleagues (12), are mirrors of each other, so that if the substitution were reversed the dissipation process, played backwards, would have the same statistical properties as the pre-adaptation process played forwards. Where previously we might have assumed that the amino acid found at the site had adapted to the requirements of the site, the site may have instead adapted to the resident amino acid.

The fluctuations and the relaxation of the protein are explicitly time-dependent. Here we addressed only the theoretical equilibrium predictions and the result of simulations that were designed to be near equilibrium. This neglect of this time dependence may explain some of the errors in the predicted Stokes shift for charged residues in buried sites. Individual sites at specific time points might be further constrained by conserved neighboring sites in the structure as well as the conserved structural context of their interactions with those sites. Such effects may influence the time-dependent probability of back mutations as well as subsequent substitutions, an important topic for further investigation.

The simulations presented here also consider only the fitness effects of stability, but fitness is also usually determined by other effects such as interactions with substrates, ligands or other proteins. Such alternative fitness components will add additional constraints to the system, and may force non-neutral substitutions if outside selective pressures change. Previous analyzes indicated that when a substitution is compelled by an outside force, an evolutionary Stokes shift occurs in largely the same fashion, except that the process is no longer reversible (11). In this context, evolution can be seen as occurring in a ‘memory foam’ made up by the bath of interactions that occur among all sites other than the selected focal site.

In conclusion, the work described here sets up a theory of evolutionary mechanics, and demonstrates that this theory can be used to predict substitution rates from the basic properties of how amino acids interact. Although the current work is focused on fitness defined by the protein stability, we expect that other kinds of selection will fit well into this framework, either by defining a large nearly neutral landscape in their own right, or by constraining the stability-based nearly neutral network.

**Methods**

**Simulations of protein evolution**

The methods used to simulate protein evolution have been described previously (11, 23, 24). The free energy $G(X, r)$ of a protein sequence $X = \{x_1, x_2, x_3 ... x_n\}$ in conformation $r$ was calculated by summing the pair-wise energies of amino acids in contact in that conformation, using the contact potentials derived by Miyazawa and Jernigan (37). We computed the free energy of folding $\Delta G_{\text{Folding}}(X)$ by first determining the free energy of the sequence in a pre-chosen native state, the conformation of the 300-residue purple acid phosphatase, PDB 1QHW.
(38)). The energies of the unfolded states were assumed to follow a Gaussian distribution with parameters estimated by calculating the free energy of the sequence in an ensemble of 55 different structurally diverse protein structures. The energy of the unfolded state was then calculated by assuming a large set \((10^{160})\) of possible unfolded structures with free energies drawn from that distribution. The free energy of folding \(\Delta G_{\text{Folding}}(X)\) was calculated as the difference between the two, and stability was \(\Xi(X) = -\Delta G_{\text{Folding}}(X)\). The Malthusian fitness of a sequence \(m(X)\) was defined as the fraction of that sequence that would be folded to the native state at equilibrium

\[
m(X) = \frac{\exp\left(\frac{\Xi(X)}{T}\right)}{1 + \exp\left(\frac{\Xi(X)}{T}\right)}
\]

(2)

where \(T\) is the temperature in units of energy, 0.6 kcal mol\(^{-1}\).

Starting from a randomly chosen nucleotide sequence encoding a 300 amino-acid protein, we simulated evolution by considering in each step all possible nucleotide mutations with rates given by the K80 nucleotide model \((\kappa = 2)\)(39). The fixation probability of each mutation was calculated based on the Kimura formula for diploid organisms(30-32),

\[
P_{\text{Fix}}(X, X') = \frac{1 - e^{-2(\Xi(X')-m(\Xi(X)))}}{1 - e^{-4N_e(\Xi(X')-m(\Xi(X)))}}
\]

(3)

where \(X\) and \(X'\) are the sequences before and after the mutation, with the effective population size \(N_e\) set to \(10^6\). One substitution was chosen to be fixed at random with relative probabilities determined by the product of the mutation rates times the acceptance probabilities.

Sequence evolution was simulated for a sufficient number of generations such that the stability of the protein was roughly constant, representing mutation-drift selection balance. 100 such equilibrated proteins were chosen, and three longer simulations were performed using each these equilibrated proteins as initial starting sequences, for a total of 300 simulations. We simulated the evolution of each lineage for an evolutionary distance of approximately seven amino acid replacements per amino acid position.

Grouping of sites

For ease of analysis, we divided the sites in the protein into four classes with similar substitution rates. Substitution matrices were calculated individually for each site; due to the length of the simulations, we had on average over 2000 substitutions at each site. We then clustered the sites based on the off-diagonal elements of the substitution matrices using K-means clustering (40, 41). The resulting clusters were approximately of equal size, and class
membership strongly dependent on how buried or exposed the sites were in the native state (as indicated by number of contacts). We ranked the clusters by surface exposure, where class 1 is the most exposed and 4 is the most buried.

**Calculating the site-specific contribution to protein stability**

The site-specific contribution $\xi_{k,\alpha}(X_{\beta,k})$ of amino acid $\alpha$ at focal site $k$ as a function of the amino acids $X_{\beta,k}$ at all sites excluding $k$ is equal to $\Xi\{x_1, x_2, x_3 \ldots, x_{k-1}, \alpha, x_{k+1} \ldots x_n\}$, the stability when the focal site is occupied by $\alpha$, minus $\Xi\{x_1, x_2, x_3 \ldots, x_{k-1}, \emptyset, x_{k+1} \ldots x_n\}$, the stability of a reference state when by $\alpha$ is replaced by a non-interacting amino acid $\emptyset$, while the rest of the sequence and thus all other interactions, are unchanged. The part of the stability unaffected by this replacement is represented by the ‘bath’ interactions $\xi_{k,\text{Bath}}(X_{\beta,k})$ so that $\Xi(X) = \xi_{k,\alpha}(X_{\beta,k}) + \xi_{k,\text{Bath}}(X_{\beta,k})$.

**Calculating the substitution rate integrating over distributions of local contributions**

The average rate for the substitution $\alpha \rightarrow \beta$ at site $k$, $Q_{k,\alpha \rightarrow \beta}$, is equal to the neutral substitution rate $u_{\alpha \rightarrow \beta}$ times the average probability of fixation, which is a function of the stability of the protein before and after the substitution. The standard deviation of observed values of $\Xi$, 0.71 kcal mol$^{-1}$, was small compared with the range of values of $\xi_{k,\alpha}$ (as shown in Figure 1), allowing us to represent the distribution $\Xi$ by its average, $\Xi = \overline{\Xi} = 9.27$ kcal mol$^{-1}$. We assumed that the stability before the substitution was equal to $\overline{\Xi}$ and after the substitution was $\overline{\Xi} + (\xi_{k,\beta} - \xi_{k,\alpha})$. The average substitution rate was then estimated as

$$Q_{k,\alpha \rightarrow \beta} = u_{\alpha \rightarrow \beta} \int \frac{1 - e^{-2(m(\xi_{k,\beta} - \xi_{k,\alpha}) - m(\overline{\Xi}))}}{1 - e^{-2m(\overline{\Xi})}} \rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta}) \, d\xi_{k,\alpha} \, d\xi_{k,\beta},$$

where $\rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta})$ is the joint distribution of $\xi_{k,\alpha}$ and $\xi_{k,\beta}$ observed when $\alpha$ occupies site $k$.

Based on the observations in Figure 1, we modeled $\rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta})$ as a bivariate normal distribution of the form $\rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta}) = \mathcal{N}(\overline{\xi}_{k,\alpha|\alpha}, \overline{\xi}_{k,\beta|\alpha}, \sigma^2_{k,\alpha|\alpha}, \sigma^2_{k,\beta|\alpha}, \varphi_{k,\alpha\beta|\alpha})$, where the parameters are represented as explicitly depending on the amino acid occupying site $k$. These parameters were calculated directly from the evolutionary simulation, and Equation (4) was integrated numerically. The neutral substitution rate was calculated using the same K80 nucleotide model ($k = 2$) as used in the simulation, with all non-nonsense codons considered equally likely.

**Calculating the substitution rate integrating assuming only neutral substitutions**

As observed in Figure 1, substitutions generally occur in a neutral region in which $\Delta\overline{\Xi}_{k,\alpha \rightarrow \beta} = \xi_{k,\beta} - \overline{\Xi}_{k,\alpha} \approx 0$, so that
\[ \frac{1 - e^{-2(m(\Xi + (\xi_{k,\beta} - \xi_{k,\alpha})) - m(\Xi))}}{1 - e^{-4N_e(m(\Xi + (\xi_{k,\beta} - \xi_{k,\alpha})) - m(\Xi))}} \approx 1. \] (5)

This condition is satisfied in a band of width \(2\varepsilon\) centred on \(\xi_{k,\beta} = \xi_{k,\alpha}\), where \(\varepsilon\) represents the deviation from strict neutrality that is still sufficiently close for Equation (5) to be sufficiently accurate.

We can obtain a natural scale for \(\varepsilon\) by considering the concept of ‘free fitness’ \(\Phi(\Xi)\) of the protein equal to \(\Phi(\Xi) = m(\Xi) + \frac{S(\Xi)}{4N_e}\) (42, 43). Free fitness, analogous to its thermodynamic equivalent ‘free energy’ where \(T\) is replaced by \(4N_e\), encompasses the contributions of both fitness and sequence entropy in determining the distribution of states; evolutionary dynamics moves towards maximising this quantity. Assuming \(S(\Xi) = \ln(\Omega_0 e^{-\psi})\) where \(\Omega_0\) is a constant, and noting that the system is at equilibrium with \(\frac{\partial \phi(\Xi)}{\partial \Xi} = 0\) when \(\Xi = \Xi\), we can see that

\[ \frac{\partial}{\partial \Xi} 4N_e m(\Xi) \bigg|_{\Xi = \Xi} = \gamma \] (6)

Thus, \(\gamma\) defines the rate of change of the population-weighted fitness \(4N_e m(\Xi)\) with stability. Alternatively, a change in stability of \(\frac{1}{\gamma}\) corresponds to a unit change in the population-weighted fitness. In our calculations, we equated \(\varepsilon = \frac{1}{\gamma}\); the estimation of \(\gamma\) is described below. Note that this calculation demonstrates that \(\varepsilon\) is, surprisingly, independent of effective population size \(N_e\). This is a result of the balance between selection and mutational drift at equilibrium; for fixed effect of mutational drift, the degree of selection (\(\frac{\partial m(\Xi)}{\partial \Xi}\)) adjusts to changes in effective population size so that their product is constant(34, 35).

If we assume that \(\rho_{\alpha}(\xi_{k,\alpha}, \xi_{k,\beta})\) is broader than \(\varepsilon\), and that Equation (5) is satisfied, Equation (4) becomes

\[ Q_{k,\alpha \rightarrow \beta}^{\text{TST}} = 2\varepsilon v_{\alpha \rightarrow \beta} \iint \rho_{\alpha}(\xi_{k,\alpha}, \xi_{k,\beta}) \delta(\xi_{k,\alpha} - \xi_{k,\beta}) d\xi_{k,\alpha} d\xi_{k,\beta} \]

\[ = v_{\alpha \rightarrow \beta} \varepsilon \frac{\exp \left( - \frac{(\bar{\xi}_{k,\alpha} - \bar{\xi}_{k,\beta})^2}{2(\sigma^2_{k,\alpha\beta} + \sigma^2_{k,\beta\alpha} - 2\varphi_k\alpha\beta\sigma_{k,\alpha\beta}\sigma_{k,\beta\alpha})} \right)}{\sqrt{2\pi(\sigma^2_{k,\alpha\beta} + \sigma^2_{k,\beta\alpha} - 2\varphi_k\alpha\beta\sigma_{k,\alpha\beta}\sigma_{k,\beta\alpha})}} \] (7)

where \(\delta(\xi_{k,\alpha} - \xi_{k,\beta})\) is the Dirac delta function.
For highly similar amino acids the entire distribution of $\rho_{k,\alpha}(\xi_{k,\alpha}, \xi_{k,\beta})$ may be contained in a region significantly narrower than the neutral zone, resulting in an overestimation of $Q_{k,\alpha\rightarrow\beta} > \upsilon_{\alpha\rightarrow\beta}$. For this reason, the estimated rate was capped at the neutral rate $\upsilon_{\alpha\rightarrow\beta}$.

**Characterising the bath state distribution**

As described above, we assume that the number of protein sequences with a given value of $\Xi$ in the range of interest around $\Xi = \Xi$ is approximately exponential $\Omega(\Xi) \sim e^{-\gamma \Xi}$. To estimate $\gamma$, we consider the distribution of changes in stability resulting from random mutations, $\rho_{\text{mut}}(\Delta \Xi)$. The average change in stability $\langle \rho_{\text{mut}}(\Delta \Xi) \rangle$ is negative due to the greater number of sequences coding for proteins with lower stability. This suggests that if we correct for the dependence of $\Omega$ on $\Xi$ by multiplying $\rho_{\text{mut}}(\Delta \Xi)$ by $e^{\gamma \Delta \Xi}$, this bias would disappear. We adjusted $\gamma$ so that $\langle \Delta \Xi e^{\gamma \Delta \Xi} \rangle = 0$ where the average was over all possible mutations during the simulations, yielding $\gamma = 1.26$ (kcal mol$^{-1}$)$^{-1}$.

**Literature Cited**


