Amino acid coevolution induces an evolutionary Stokes shift

David D. Pollock[a], Grant Thiltgen[b], and Richard A. Goldstein[b,1]

[a]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045; and [b]Division of Mathematical Biology, National Institute for Medical Research, London NW7 1AA, United Kingdom

Edited by Andrew J. Roger, Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, Canada, and accepted by the Editorial Board April 3, 2012 (received for review December 7, 2011)

The process of amino acid replacement in proteins is context-dependent, with substitution rates influenced by local structure, functional role, and amino acids at other locations. Predicting how these differences affect replacement processes is difficult. To make such inference easier, it is often assumed that the acceptabilities of different amino acids at a position are constant. However, evolutionary interactions among residue positions will tend to invalidate this assumption. Here, we use simulations of purple acid phosphatase evolution to show that amino acid propensities at a position undergo predictable change after an amino acid replacement at that position. After a replacement, the new amino acid and similar amino acids tend to become gradually more acceptable over time at that position. In other words, proteins tend to equilibrate to the presence of an amino acid at a position through replacements at other positions. Such a shift is reminiscent of the spectroscopy effect known as the Stokes shift, where molecules receiving a quantum of energy and moving to a higher electronic state will adjust to the new state and emit a smaller quantum of energy whenever they shift back down to the original ground state. Predictions of changes in stability in real proteins show that mutation reversals become less favorable over time, and thus, broadly support our results. The observation of an evolutionary Stokes shift has profound implications for the study of protein evolution and the modeling of evolutionary processes.

A major focus of modern evolutionary studies is to understand how structural and functional contexts determine the patterns of evolutionary change at different positions in a biological macromolecule. Such an understanding is important to phylogenetics, partially because the position-specific processes of evolution are known to determine our ability to reconstruct deep nodes in the tree of life (1) but also because features of the evolutionary process such as convergence can deterministically mislead phylogenetic reconstruction (2). Understanding patterns of evolution and how they respond to details of structure and function can also potentially help us to better decode the evolutionary record, allowing us to distinguish between structural and functional constraints and identify the signatures of positive selection. This finding can lead to improved understanding of a biomolecule's structure, dynamics, thermodynamics, functionality, and physiological context. Key questions concern how evolutionary processes vary among sites and over time and particularly, how the evolution at different locations influences each other. For example, coevolution between different locations in a protein can slow the amino acid replacement process, allowing phylogenetic inference at deep nodes that would otherwise have been swamped by recurrent neutral changes. Furthermore, coevolution tends to depend on proximity in the 3D structure of proteins, leading to the hope that, if properly understood, it could improve our ability to predict important features of protein structure (3–11).

To understand patterns of molecular evolution, it is necessary to model the process of evolution. Models of protein evolution must, out of necessity, make assumptions or simplifications concerning the underlying replacement process. In the past, most model assumptions have been driven by the lack of data to determine parameters and the overwhelming computational demands of more realistic but complex models. Recent advances in high-throughput sequencing, high-performance computing, and phylogenetic model building have improved the situation, but it is still necessary to make simplifications and assumptions. The question is which simplifications are most useful for addressing a particular problem or issue? The model including the simplifications should be correct enough to decipher key features of how protein structure and function interact with protein evolution, allowing us to interpret the terms of the model in terms of the basic biology and biochemistry. To answer this question, we need to consider the mechanistic theory underlying different evolutionary models.

The most direct approach to formulating a mechanistic model of protein evolution is to make the replacement rates dependent on the resulting change in various protein properties, which is calculated as a function of the entire protein sequence. This approach has been the rationale for thermodynamic energy-based models, which allow for direct calculation of contextual interactions. Although this approach might seem a priori preferable because of its directness, it has dual disadvantages: it is slow, and it does not seem to explain the data as well as site-specific empirical models (12, 13). Thermodynamic models are compromised by the unlikely assumption that the fitness of a protein is a simple function of its thermodynamic stability and the assumptions necessary to make the thermodynamic calculations computationally feasible. For example, the energy potentials used are almost certainly incorrect. We still do not have adequate ways to include the effect of side chain and backbone flexibility in these calculations. Thermodynamic properties generally involve small differences between large numbers, meaning that these larger quantities must be computed to excruciating accuracy. Small errors in energy potentials can accumulate across the many atomic interactions in a protein, compounding error in even the best energy models. Most of the time in evolutionary studies, however, many variants need to be evaluated, and therefore, simple (and even more error-prone) pairwise contact potentials are used for computational reasons. Furthermore, to correctly calculate the free energy of the native fold, it is necessary to consider the energy of all thermodynamically relevant alternative folds. Because it is currently impossible to include the vast multitude of such folds or even know what the relevant folds are, decoy datasets are generally used, consisting of, for example, a subset of the known folds in protein databases. These decoy datasets are inadequate representations of

Author contributions: D.D.P. and R.A.G. designed research; G.T. and R.A.G. performed research; D.D.P. and R.A.G. analyzed data; and D.D.P. and R.A.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.J.R. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: richard.goldstein@nimr.mrc.ac.uk.

See Author Summary on page 7961 (volume 109, number 21).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1120084109/-/DCSupplemental.

the energetically relevant competing folds, leading to large amounts of error in energy analyses.

Because of these limitations, phylogenetic analyses have been dominated by phenomenological models. Such models attempt to capture the results of evolutionary change rather than model the selective constraints acting on the evolutionary process per se. Historically, empirical models of amino acid evolution were obtained from observed differences among sequences (14–16). Phenomenological substitution models were initially constructed based on the assumption that evolution at all locations was identical and independent or that different locations might evolve at different rates, with the rate acting as a simple scaling factor for the rate matrix. More elaborate models have been constructed that allow substitution rates to depend on local structure (17–19), whereas other models have allowed substitution rates to vary between branches and locations in ways that are determined by the sequence data (20–22). In these models, it was observed that replacements were more frequent among amino acids with similar physicochemical properties (for example, among small hydrophobic amino acids) than more disparate amino acids (for example, between amino acids with positive and negative charge). This idea of exchangeability can be rationalized using the graphical models in the work by Fisher (23), which noted that, if a trait is approximately optimized, it is far more likely for smaller, more conservative changes to be acceptable and not eliminated by purifying natural selection. More radical changes, in contrast, are more likely to be deleterious.

An alternative approach is to implement site-specific mutation selection models (24–27) that attempt to capture the mechanistic aspects of the evolutionary process while avoiding the conceptual and practical limitations of thermodynamic models. Such models can be related to biophysical models of protein structure and thermodynamics by considering that each location (or site) l in a protein has a propensity for each of the 20 amino acids, depending on its local structural and functional environment. This propensity is then related to fitness by assuming that organisms with a particular amino acid a at location l in the protein will have an average fitness ω_l^a compared with organisms with other amino acids at that site. The propensities and fitnesses are generally assumed to be constant over the course of evolution. The general structure of this model is, thus, that the probability of substitution at location l is equal to $\mu_{ij}^l \times f_{ij}^l$, where μ_{ij}^l is the mutation rate from nucleotide i to nucleotide j at location l and f_{ij}^l is the probability of fixation at that location. If the change is synonymous, then the mutation is usually considered neutral (although codon biases can be included); however, if the mutation of a codon encodes a replacement of an amino acid, the probability of fixation is usually calculated using the relative fitnesses of the two amino acids and the equations by Crow and Kimura (28) and Kimura (29, 30) (see below). In contrast to the thermodynamic approach, the parameters in such models can be estimated from the sequence data rather than calculated from first principles. The amino acid substitution process in such cases is time reversible as long as the mutation process is time reversible.

The site-specific constant fitness mutation selection models suffer from a number of limitations. The important properties of proteins (e.g., structure, function, and stability) are holistic and depend on interactions between amino acids. As different locations undergo substitutions, the context of other sites will change. These changing interactions between locations with substitutions in the protein have motivated work on correlated substitutions or coevolution (3, 4, 8, 10, 11), and most such work has found strong evidence that coevolution is extremely common. The model also implies that a nonpreferred amino acid will always be nonpreferred, and therefore, although occasionally observed, it will be unstable over evolutionary time.

Modifications of exchange rates do occur, however, and are mediated through coevolutionary interactions with exchanges of amino acids at other locations. A fundamental difference between expectations under what we will call a coevolutionary fitness model and the site-specific constant fitness (independent evolution) model described above can be seen by considering forward and reverse substitutions. In the constant model, if a forward substitution (for example, from amino acid a_k to a_l) is advantageous, then the reverse substitution (from a_l to a_k) will always be disadvantageous or deleterious and will be unlikely to occur. In other words, the change in fitness of the reverse substitution, $\Delta\omega(a_l \Rightarrow a_k)$, is the same magnitude and the opposite direction as the change in fitness of the forward substitution of $\Delta\omega(a_l \Rightarrow a_k) = -\Delta\omega(a_k \Rightarrow a_l)$. In contrast, the coevolutionary fitness model would allow for the possibility that $\Delta\omega(a_k \Rightarrow a_l)$ and $\Delta\omega(a_l \Rightarrow a_k)$ might change because of substitutions at other sites.

Despite their limitations and the lack of a plausible underlying theoretical rationale, the overall comparative success of the phenomenological models suggests that we should consider how they can be further developed. We should consider what underlying processes make sense to model phenomenologically and whether we can improve the successful application of such models. The development of better models requires an improved understanding of the substitution process and how it is affected by epistasis. In particular, we want to know when the evolutionary process at a site changes, the magnitude and timescale of these changes, and whether the changes occur in a predictable fashion. We are interested in knowing the extent to which a coevolutionary scenario is justified when considering protein evolution and how coevolution might affect substitutions through fluctuating fitness differences among amino acids at each site. We explore these ideas with a simple model of protein evolution in which the fitness of a protein is calculated based on the probability that it folds to its native structure (the purple acid phosphatase). We investigate how substitution propensities change with substitutions at individual locations because of coevolutionary substitutions at other locations in the protein. Finally, we consider the structure of a general model that would take our observations into account.

Results

Scale and Rate of Fluctuations in Selective Constraints. We first performed 18 evolutionary simulations on the structure of purple acid phosphatase for a total of over 2.4 million substitutions. We considered six focal locations: exposed sites 111 and 147, partially exposed sites 168 and 273, and buried sites 7 and 135. We calculated the instantaneous propensities Π_l^X at these sites for each of the 20 amino acids X after every substitution in the evolutionary pathway, defined as the equilibrium frequency of the amino acid given fixed amino acids at all other locations and ignoring the effect of base composition and the degeneracy of the genetic code (Eq. **4**). If there were no coevolutionary interactions (the standard assumption of most phylogenetic models), these propensities would not change over the course of the simulations. Instead, there is considerable variation in propensities in all simulations, which can be seen in the example segment of evolution at site 168 shown in Fig. 1. It is also notable in Fig. 1 that the amino acid that is actually at that location is almost always the amino acid with the highest propensity.

We can characterize the range of the fluctuations by calculating the distribution of the propensities at a given location, such as in the distributions of the various amino acids at location 168 shown in Fig. 2. (The corresponding distribution for locations 111, 147, 273, 135, and 7 is shown in Fig. S1.) The most frequent propensities for all amino acids range from 0 to 0.2 and are centered around 1/20th, but the most common propensity for any amino acid at the location is that it is highly disfavored. However, all amino acids also have some probability of reaching propensities near 1.0, in which case all other amino acids are essentially disallowed.

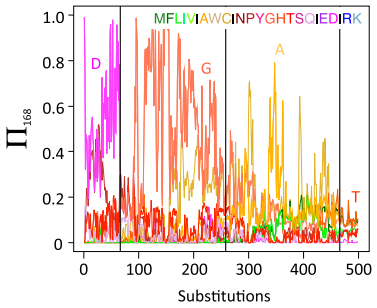

Fig. 1. Propensities shift caused by coevolution. Results are shown for propensities of various amino acids at site 168 during an evolutionary period encompassing 500 substitutions. Black lines represent changes in the amino acid resident at this location, with the current occupant during each time period noted. During this time, site 168 underwent substitutions from aspartic acid (D) to glycine (G) to alanine (A) and then, to threonine (T). The propensities of the 20 amino acids are indicated by different colored lines, which are indicated by the single-letter International Union of Pure and Applied Chemistry (IUPAC) amino acid codes in the legend. For clarity, amino acids with low propensities during this time period were omitted.

We can quantify the degree to which a site is constrained in its amino acid composition by considering A_l, the effective size of the alphabet of amino acids possible at a given location l, calculated with Eq. **5**. The average values of A_l for the six locations $\langle A_l \rangle$ averaged over each point in the simulation using the instantaneous propensities Π_l^X are 10.48 and 10.90 (exposed sites 111 and 147, respectively), 8.44 and 8.65 (partially buried sites 168 and 273, respectively), and 2.94 and 3.69 (buried sites 7 and 135, respectively). In contrast, the values of A_l calculated using the average values of the propensities $\langle \Pi_l^X \rangle$ are 18.71, 19.10, 14.52, 17.80, 6.83, and 8.03, respectively. This finding indicates that the average amino acid position is considerably more constrained at any instant than one would predict based on considering the average frequencies observed over long periods of time. Notably, this observation holds true even with the varying levels of constraint at the three sites (the surface sites are least constrained and the buried site is most constrained).

We next considered the timescale of the fluctuations in amino acid propensities. One way this question can be framed is by asking about the rate at which the system loses memory of its previous amino acid preferences. We measured this rate for the six focal locations using the decay of the autocorrelation function over different periods of time (Fig. 3). The autocorrelation drops rapidly to 0.75–0.90 after 50 substitutions, after which it drops off only very slowly and comes close to equilibrium after 10^4–10^5 substitutions or more (note the log scale for substitutions in Fig. 3). The drop off can be reasonably well-fit by a stretched exponential, characteristic of the dynamics of systems such as spin-glasses with large numbers of degrees of freedom and a rough energy landscape (Table S1). The rate of loss of predictability is highly dependent on the exposure of a site, with the

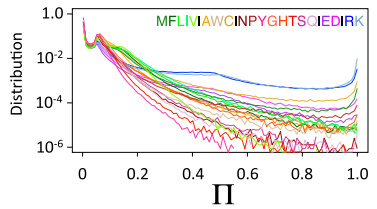

Fig. 2. Long-term distribution of propensities. Distributions of marginal propensities for all 20 amino acids across 18 simulations totaling over 2.4 million substitutions are shown for partially buried site 168. Line colors are as in Fig. 1. Similar plots for all six focal locations are shown in Fig. S1.

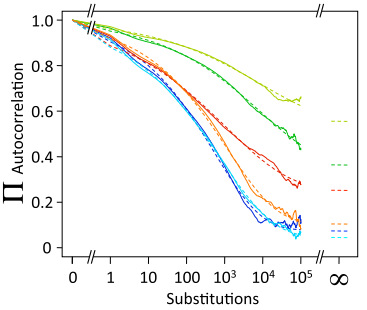

Fig. 3. Decay of the autocorrelation function of amino acid propensities. The dynamics of the decay of the autocorrelation function are shown for exposed locations (blue, site 111; cyan, site 147), partially exposed locations (red, site 168; orange, site 273, and buried location (green, site 135; lime, site 7). Dashed lines fit to a stretched exponential plus baseline as described in the text. Value of the fit at infinity represents the baseline value. Note the modified logarithmic scale on the abscissa.

buried site retaining predictability the longest and coming to a higher asymptotic value and the exposed site losing predictability the quickest and dropping to the lowest value in the end. Notably, the predictability for each site does not drop to zero (although the exposed site comes close), indicating that information is retained about the nature of the site from the long-term process, even if there is a great deal of fluctuation over the short term. This finding is not surprising, because it reflects the biases of different amino acids to different local structures as well as interactions of the focal location with the biased distributions at other sites. (The presence of long-term information about the amino acid distribution is also shown by the effective size of the alphabet of amino acids calculated using the average values of the propensities, which is less than 20.) Still, the equilibrium correlation is only 0.36 for buried sites, which indicates that recent propensities at a site (over the last 50–1,000 substitutions in the protein overall) are a better predictor of future propensities than the location in the protein.

Coadaptation Between Protein Locations. As noted in Fig. 1, despite considerable fluctuations, the amino acid that is at a location is almost always the amino acid with the highest propensity. This finding comes about, first, because as might be expected, amino acids tend to be replaced by residues that have relatively high propensities at the time of the substitution (that is, the protein is preadapted to accept the new residue). This finding is clear in Fig. 4, where the average propensity for a new amino acid just after it was substituted into position 273 was 0.13. In comparison, the overall average propensity of the same amino acid was 0.07. This degree of preequilibrium is especially high for the charged amino acids. Second, however, amino acid propensities tend to evolve so that the new amino acid becomes more preferred at that position. The average propensity of amino acids in the period after a substitution, while they are still resident, is 0.17, significantly higher than the average propensities at both the time of the substitution and overall. This finding means that, after a replacement, coevolution tends to equilibrate the protein to the presence of the new amino acid. This equilibration tends to increase over time, slowly approaching the average propensity for amino acids that are fixed at a location (Fig. 4) (see below). It can also be seen that these effects are much more pronounced for the buried amino acids compared with the partially buried and especially, exposed amino acids.

To separate the tendency of an amino acid to occur where it is preferred from the tendency of the amino acid to become preferred where it occurs, we performed additional simulations in which the amino acid at the focal location was not allowed to

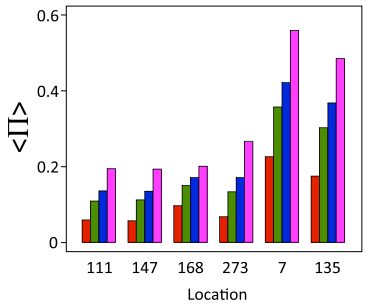

Fig. 4. Increase in propensity for current amino acid caused by coevolution. For each location, the average propensity of all amino acids during free simulations (red) was compared with the average propensity of new amino acids after a substitution (green), the average propensity of resident amino acids during their residency (blue), and the average propensity of fixed amino acids (magenta). The six locations examined were exposed locations 111 and 147, partially exposed locations 168 and 273, and buried locations 7 and 135.

change. In this way, the amino acid found at a location is independent of its propensity for the location. Fig. 5 shows the average propensities of selected amino acids at position 168 when this location was constrained to have different fixed amino acids. (More complete propensities for locations 111, 168, and 135 are shown in Fig. S2; average propensity for fixed residues, averaged over the equilibrium distribution, are shown in Fig. 4.) As can be seen, the protein in each case adjusts so that the fixed amino acid becomes the preferred amino acid: the average propensity of leucine (L) is 0.21 when L is fixed at location 168 but only 0.01 when aspartic acid (D) is fixed at this location. Furthermore, the adjustment to the fixed amino acid also increases the propensity for amino acids that are similar to the resident amino acid: the propensity of L is increased when the similarly hydrophobic valine (V) is fixed at that location, adjustment of the protein to serine (S) increases the propensity for threonine (T), and the presence of the negatively and positively charged glutamic acid (E) and lysine (K) increase the propensity of the similarly charged D and arginine (R), respectively. Note that these different propensities occurred in a single location in the protein and therefore, cannot represent the constraints caused by local structure.

Response to Deleterious Substitutions. Substantially destabilizing substitutions may occur for a variety of reasons, including selection at active site locations for improved or new functionality or selective sweeps in nearby genes. How does the protein react to such perturbations? How fast does the shift in amino acid propensities approach equilibrium? To address these questions, we chose a pair of residues X and Y and found three sequences where (i) there was an X at the focal location and (ii) an $X \rightarrow$ Y substitution would be moderately destabilizing, with a value of $\Delta\Delta G_{X \rightarrow Y} = 1$ kcal/mol. We then made this substitution in this sequence, fixed the Y at this location, and monitored the energetics of the $Y \rightarrow X$ back substitution. Each of these three simulations was repeated 1,000 times. An example is shown in Fig. 6 for substitutions from X = R to Y = D at site 111, which has broadly similar characteristics to other types of substitutions. Looking at these runs, it seems that there are rapid short-term fluctuations on the order of 50 substitutions as well as consistent long-term directional trends on longer timescales.

By averaging together multiple runs, a more general trend emerges. Initially, before any other substitutions have occurred in the rest of the protein, the back mutation $R \rightarrow D$ would recover the original sequence, and therefore, $\Delta\Delta G_{R \rightarrow D} = -1$ kcal/mol. As additional substitutions occur, the arginine at this location becomes increasingly accommodated relative to the aspartic acid. After ~50 substitutions, the reverse mutation is, on average, deleterious, becoming increasingly deleterious with additional substitutions.

Fig. 7 presents the evolution of the propensities of the 20 amino acids after a fixed R111D substitution. Within a small number of substitutions (fewer than 10), the propensity for the resident amino acid rises about 1/20th (0.05). The propensity for the resident amino acid eventually increases to about 0.25, on average, after 1,000 substitutions, but it does not seem to have yet equilibrated. This finding indicates that there is a rapid response adjusting to such a destabilizing amino acid replacement (50–100 substitutions to rise from low marginal propensity to around 12%) followed by a gradually decelerating rise over a much longer period. Perhaps the most interesting aspect of Fig. 7 is what happens to the propensities of the other amino acids. With this negatively charged amino acid at the focal site (111), the propensities for the hydrophobic MFLIV of group amino acids and the positively charged (RK) amino acids plummet, whereas amino acids in the large NPYGHTSQ group remain at moderate propensities (0.04–0.07). Most interestingly, the propensity for the other negatively charged amino acid, glutamic acid, slowly increases its propensity in step with the increase in aspartic acid propensity, rising to about one-half the propensity of aspartic acid. This finding is in agreement with the results in Fig. 5.

Evolution of Mutational Stability and the Evolutionary Stokes Shift. A prediction of our model that can be measured is the effect of subsequent substitutions on the effect of mutations and their inverse over time. Recall that the standard model in the work by

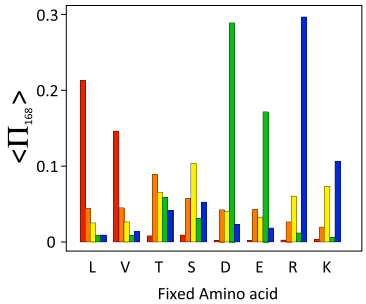

Fig. 5. Effect of current amino acid on amino acid propensities. Average propensity of selected amino acids at location 168 ($\langle \Pi^X_{168} \rangle$) depending on the amino acid fixed at that location. Results are shown for X equal to lysine (L; red), threonine (T; orange), serine (S; yellow), aspartic acid (D; green), and arginine (R; blue).

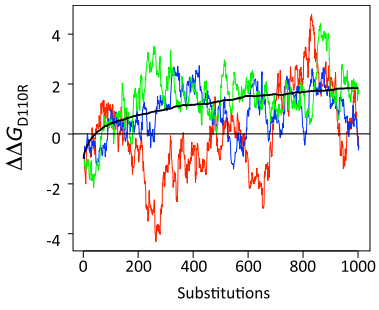

Fig. 6. Evolution of the change in stability for the back mutation ($\Delta\Delta G_{D111R}$) after an R110D mutation. Three sequences were chosen for which an R111D mutation was slightly deleterious ($\Delta\Delta G_{R111D} \approx 1$ kcal/mol). This mutation was made, and the D was fixed at this location. Red, blue, and green traces represent individual simulations. Black curve represents the average of 1,000 simulations for each of the three initial sequences.

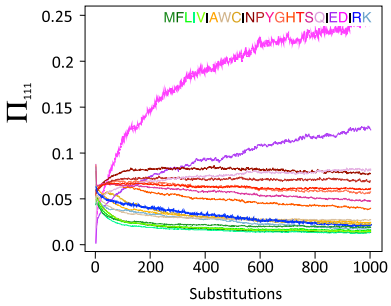

Fig. 7. Evolution of changes in propensities after mutation. Evolution of amino acid propensities at site 110 over 1,000 substitutions. As in Fig. 5, three sequences were chosen for which an R110D mutation was slightly deleterious ($\Delta\Delta G_{R111D} \approx 1$ kcal/mol). This mutation was made, and the D was fixed at this location. Curves represent the average of 1,000 simulations for each of the three initial sequences. Amino acid color codes are as in Figs. 1 and 2.

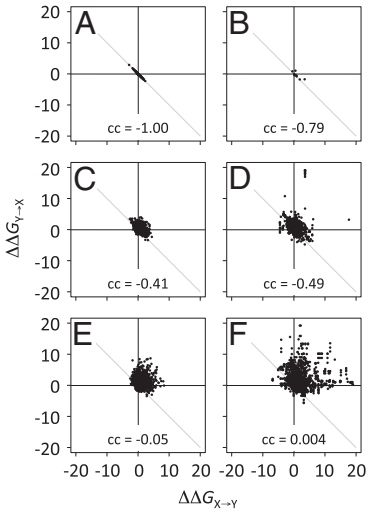

Fig. 8. Correlation between energetics of forward and backward mutations. (*A*, *C*, and *E*) Values of $\Delta\Delta G_{X\to Y}$ compared with $\Delta\Delta G_{Y\to X}$ for mutations where one sequence is changed to match the amino acid in the other sequence at that location as a function of the pairwise identity at other locations: (*A*) 100%, (*C*) 75%, and (*E*) 20%. If all locations are independent, $\Delta\Delta G_{Y\to X} = -\Delta\Delta G_{X\to Y}$, which is the case for *A*. (*B*, *D*, and *F*) Similar calculations for different homologs of ferrodoxin, where the values of $\Delta\Delta G_{X\to Y}$ and $\Delta\Delta G_{Y\to X}$ are computed using Rosetta (31). Calculations were based on the crystal structures of the two proteins for sets of proteins where the pairwise identity at other locations was (*B*) 100%, (*D*) 70–80%, or (*E*) <25%. Correlation coefficients (cc), calculated after excluding outliers ($|\Delta\Delta G| > 15$), are included in the plots.

Halpern and Bruno (24) assumes that the propensities do not change; thus, mutations and their inverse should have opposite effects on stability (in our model, stability reflects propensity), and this effect should not change over time, because evolution is independent among sites in this model. We can examine this effect by considering two divergent sequences. At every location that differs, we can evaluate $\Delta\Delta G$ for the mutation that converts the residue found in one sequence to the residue found at that location in the other sequence and compare this finding with the corresponding inverted situation. We perform this analysis for sequences generated by the evolutionary simulation, where the two sequences are separated by different degrees of identity at the other locations in the protein (100%, 75%, and 20% in Fig. 8 *A*, *C*, and *E*, respectively). [Identity is used to represent distance between the sequences to allow comparisons with sequences from nature (Fig. 8 *B*, *D*, and *F*), for which the number of substitutions separating them is unknown.] Because our calculations for simulated sequences are exact, the $\Delta\Delta G$ for mutations and their inverses when no other changes between the two sequences exist (100% identity) is exactly on the diagonal, indicating that the $\Delta\Delta G$ for inverse mutations is always of the same magnitude and opposite sign (Fig. 8*A*). As sequences diverge, the relationship moves off the diagonal (Fig. 8*C*) until, with low sequence identity (20%) (Fig. 8*E*), the inverse mutations are almost always up and to the right, indicating that mutations away from the resident amino acid are almost always worse (i.e., have positive values of $\Delta\Delta G$). To test this finding on sequences from nature, we sampled ferrodoxin sequences with high-resolution crystal structures in the protein database and predicted single-mutation $\Delta\Delta G$ using Rosetta (31) (Fig. 8 *B*, *D*, and *F*). The pattern is strikingly similar to the pattern from our simulated proteins, strongly supporting the idea that there is a shift in real sequences as well.

A comparison can be made with spectroscopy. If a simple atom in vacuum is excited by absorption of a photon of light, there is an increase in the energy of the system $\Delta E_{\text{Absorption}} = E_{\text{Excited state}} - E_{\text{Ground state}}$. Subsequent emission of a photon through fluorescence changes the energy of the system by the negative of this amount: $\Delta E_{\text{Emission}} = E_{\text{Excited state}} - E_{\text{Ground state}} = -\Delta E_{\text{Absorption}}$. As a result, a plot of $\Delta E_{\text{Absorption}}$ vs. $\Delta E_{\text{Emission}}$ would consist of points on the line $\Delta E_{\text{Emission}} = -\Delta E_{\text{Absorption}}$, similar to the plot $\Delta\Delta G_{X\to Y}$ vs. $\Delta\Delta G_{Y\to X}$ shown in Fig. 8*A*. In more complicated molecules, each electronic state corresponds to a series of vibrational states, which add vibrational energy to its total energy. At equilibrium at room temperature, the molecule is generally predominantly at the ground vibrational state, but changes in the electronic state often leave the molecule in an excited vibrational state; after the excitation, the vibrational state relaxes back to the ground vibrational state, generally through radiationless transitions. This tendency for electronic transitions (of either direction) to be accompanied by vibrational excitation results in an increase in both $\Delta E_{\text{Absorption}}$ and $\Delta E_{\text{Emission}}$, increasing the energy gain and reducing the energy loss, respectively, which is an effect called the Stokes shift. As a result, plots of $\Delta E_{\text{Absorption}}$ vs. $\Delta E_{\text{Emission}}$ would consist of points above and to the right of the line $\Delta E_{\text{Emission}} = -\Delta E_{\text{Absorption}}$, similar to the effect of the evolutionary dynamics shown in Fig. 8 *C*–*F*. We believe that the spectroscopic Stokes shift is an apt analogy for the change in amino acid preference that occurs when an amino acid is substituted at a position, which is characterized by the tendency for substitutions in both directions to be deleterious with larger values of $\Delta\Delta G$ as observed in Fig. 8. We, therefore, refer to this effect in proteins as an evolutionary Stokes shift.

Discussion

We have presented evidence here that strongly suggests that our understanding of how proteins evolve and coevolve needs to be fundamentally revised. We have long understood that evolution at different positions in proteins is context-dependent (even if our models have not often incorporated this knowledge), but we cannot make strong claims to understand what determines this context dependence. We should understand more about general trends, whether the structural context or sequence context dominates evolutionary trends, and what amounts of sequence change (what timescales) affect sequence context. Here, our most notable finding is the existence of what we call an evolutionary Stokes shift, by which we mean that, on substitution of an amino acid at a position in a protein, the protein will tend to adjust through coevolutionary processes to having that amino acid at that position; therefore, the inherent propensity for that amino acid at that position will be, on average, higher than it was when the substitution occurred. As a result, for most of the time that the amino acid is resident, the probability that a reverse

mutation or any other mutation at that position will be accepted will be substantially smaller than when the substitution occurred. In our model, the evolutionary Stokes shift is affected through folding stability, but it is reasonable to suppose that other fitness-inducing complex landscapes may induce this effect as well as long as there is a significant degree of epistasis.

The presence of an evolutionary Stokes shift does not necessarily violate the reversibility of evolutionary dynamics. A new amino acid resident at a given location will bias substitutions at other locations to increase the propensity of the new amino acid at that location. Other substitutions may arise, however, that decrease the propensity, enabling another substitution at this location to occur. The dynamics can be imagined as movement on a fitness landscape between successive fitness peaks, each representing an amino acid at that location. Differences in the heights of the peaks represent the inherent suitability of any amino acid for that location. Initially, when a substitution occurs, the sequence is at a lower point on the peak, but its height on the peak increases as coevolutionary substitutions occur at other locations in the protein. Stochastic movement on the peak can decrease the fitness, enabling a substitution at this location and the jump to another peak. In contrast to this situation of mutation selection balance, violations of reversibility would be expected after the fixation of a substitution because of positive selection, and such nonreversibility can affect the reversibility at other sites that are linked through coevolution. In this case, there would be irreversible adjustment of the rest of the protein, improving the fitness of the newly substituted amino acid, which is seen in Figs. 6 and 7.

Even without the Stokes shift, we expect substitutions to similar amino acids to be favored compared with dissimilar amino acids (the Fisherian model) simply because of site heterogeneity; amino acids would predominantly be found where they are preferred, and a location with a preference for a given amino acid would likely accommodate similar amino acids. Our model, however, would greatly enhance the Fisherian nature of protein evolution; as the protein equilibrates to the presence of a resident amino acid at a position, it will similarly increase the propensity for amino acids that share physicochemical properties. In the classical point accepted mutation (PAM) matrix style models (14–16), lower rates of substitution are modeled by a phenomenological rate factor. Here, however, the explanation is that the propensities for similar amino acids increase along with the increase in propensity for the resident amino acid because of coevolutionary adjustments in the rest of the protein sequence.

To develop a new framework for empirical model building in evolutionary analysis, it is useful to build on the old framework. In the models in the work by Halpern and Bruno (24) as well as PAM-like models, sites evolve independently and reversibly. As mentioned in the Introduction, in the models by Halpern and Bruno (24), the rate of substitution is $Q_{ij}^l = \mu_{ij}^l f_{ij}^l$, where μ_{ij}^l is the mutation rate and f_{ij}^l is the probability of fixation of this change at that location. At equilibrium, substitutions require a simultaneously high propensity for both the original amino acid (or it would not be resident and available for substitution) and the new amino acid (or the substitution would not be accepted). Locations with a high propensity for one amino acid will generally have a high propensity for similar amino acids; this tendency for the propensities of similar amino acids (Π_i^l and Π_j^l or alternatively, the equilibrium frequencies π_i^l and π_j^l) to covary between locations provides a mechanism for the observation in the work by Fisher (23) of the predominance of conservative evolutionary changes. This mechanism relies on site dependence of the propensities; the more uneven the distribution of propensities among different locations, the more conservative the observed substitutions.

In contrast, PAM-like general reversible models applied to a single site, Q_{ij}^l, can be expressed as $Q_{ij}^l = \mu_{ij}^l \lambda_{ij}^l \pi_{ij}^l$, where λ_{ij}^l is the parameter of mathematical convenience that results in a reversible model as long as $\frac{\lambda_{ij}^l}{\lambda_{ji}^l} = \frac{\mu_{ji}^l}{\mu_{ij}^l}$. We note that the rate parameters of such models are not generally fit to individual sites because of the large number of rate parameters (190). In general, not even the equilibrium frequencies are fit to individual locations, which means that the value of λ_{ij}^l is a parameter adjusted to the average observed frequencies among sites. Because site-dependent propensities are not included in PAM-like models, the mechanism for favoring conservative changes that is incorporated into the model by Halpern and Bruno (24), which depends on correlations between site-dependent amino acid propensities, is disallowed. Instead, the conservative nature of substitutions can only be contained in λ_{ij}^l, incorrectly suggesting that the effect is one of rates rather than covarying propensities.

In our framework, which we will designate the Stokes–Fisher framework, the substitution rate depends on the timescale. On a very short timescale, when no other substitutions have occurred in the protein, the substitution rate at each location is based on the instantaneous amino acid propensities. The location-specific rate of substitution over longer periods of time is, however, dependent on coevolutionary changes at other positions. Substitution from amino acid i to j will occur only when propensities have drifted such that j is sufficiently fit relative to i in that it has a reasonable probability of substitution. The substitution probabilities will depend on the distribution of relative fitnesses over time, and it seems unlikely that probability of substitution given arrival into a substitutable state can be separated from the probability of arriving into a substitutable state. The most reasonable approach for future model-building may, therefore, be to combine the processes of drift to substitutability and fixation into a single set of parameters. These parameters would represent the probability of a given shift in propensities times the rate of substitution given this shift integrated over all possible propensity shifts. The shifts would reflect random fluctuations in propensities as well as the systematic tendency of proteins to adjust, on average, to the current amino acid at a location and prefer similar amino acids.

Unfortunately, in addition to the large number of parameters inherent in such an approach, our results show that the propensities and therefore, the substitution probabilities are likely to change to varying degrees over different timescales. Propensities may change rapidly, especially immediately after a substitution at that site and especially after a selected substitution that would have been deleterious with regard to the structure alone. Conversely, the results presented in Figs. 3 and 7 show that the longest timescales are on the order of thousands of substitutions in the protein, corresponding to branch lengths on the order of 10 substitutions per site. There is good reason to question whether such a process will ever be knowable. However, there is hope that the process may be moderately stable over moderate periods of time as long as the amino acid at the focal site does not change. In this manner, modifications of schemes that use a set of Markov-modulated substitution models (32, 33) might be promising. In contrast to currently-implemented Markovian schemes, however, there would need to be interplay between the changes in the amino acid at a given location and the length of time that the amino acid has been at that location and the appropriate substitution model. We can characterize the types of changes in the substitution models through concepts such as coevolutionary latency, the timescale during which the protein adjusts to the new amino acid at the site through coevolution at other sites. After this latency period, we can consider longer periods of evolutionary time where there is a temporarily stable process and for which any small fluctuations are sufficiently fast that they may be averaged. We might also consider that, when the process does change, the concept of coevolutionary latency will again apply. These coherent coevolutionary propensity processes may be expected to change in an a priori unpredictable

way, but as long as there are coherent and measurable processes in the tree on both sides of the change, then the coevolutionary latency transition might be presumed to be fairly regular. This finding might allow its characterization by a single posthoc inferable parameter governing the rate of change from one coherent process to the other.

There have been other substitution models including context dependency (34, 35), but most of these include the effect of context on the mutation process. The mechanism that we discuss here, conversely, involves the impact of the context (the protein sequence) on the fixation probability. Because of this difference in mechanism, the consequences are very different. First, the local propensities are a function of all of the amino acids in the protein; even locations not in contact with the focal location can affect these propensities through either contacts made in unfolded structures or changes in the protein stability and thus, the degree of selective pressure. The magnitude of the epistasis between different locations, however, has a wide distribution. The result is amino acid propensities that fluctuate over a wide range of timescales from near instantaneous to changes that occur over thousands of substitutions throughout the protein. Second, the impact of a given amino acid on the fixation process at other locations in the protein results in a tendency of the protein to adapt to this amino acid, a process that we have termed the evolutionary Stokes shift. One consequence of this Stokes shift is that the amino acid at a given location will generally have a higher propensity, not only because substitutions will favor favorable amino acids but also because resident amino acids will become favorable because of their impact on other locations. This type of effect is difficult to imagine occurring through changes in the mutation rates.

This Stokes shift will also affect similar amino acids, increasing the rate of conservative substitutions. This finding can be understood by considering the original argument in the work by Fisher (23) that conservative changes would be more likely to be accepted than more radical mutations. This effect requires the organism to be near a fitness optimum. This optimum can occur because of the adaptive process of evolution. Alternatively, if we consider the fitness landscape for a given location in the protein, the location of the fitness optimum will move as the other locations in the protein change. The evolutionary Stokes shift will result in the fitness optimum shifting to better match the amino acid at that location. In this way, the criterion for the Fisher construction would be satisfied, not only because the amino acid adapts to the fitness landscape but also because the fitness landscape adapts to the current amino acid (23).

We characterized our aim here as considering the consequences of evolution on a complex fitness landscape when they are viewed from the perspective of a single site. If we view this finding as part of a broader program, the goals might be to characterize the effect of thermodynamically generated epistatic interactions on the different types of models and see how the next generation of models can include the impact of collective protein properties on the evolution of individual locations considered independently. Essentially, we must understand the dialectic among amino acids. The broader program would include understanding sign epistasis and genetic constraint on evolutionary trajectories (36), developing Dobzhansky–Muller incompatibilities (37–39), and predicting differences in fitness of different mutations in different organisms caused by compensatory change (40). We have addressed this program through a particularly simple model, where the collectiveness involved the calculation of protein thermodynamic stability. The next stage, necessarily, is to evaluate the existence, magnitude, and nature of such effects through detailed phylogenetic analysis.

Methods

Protein Model. The model used to simulate protein evolution in this study is based on calculating a sequence's free energy of folding to a particular target structure (what we will call the native conformation) as described previously (41, 42). The free energy $G(S, C_k)$ of a protein sequence $S = \{a_1, a_2, a_3 \ldots a_M\}$ in a particular conformation C_k is calculated based on the sum of pairwise energies between amino acids that are in contact in that conformation [that is, $G(S, C_k) = \sum_{i<j} \gamma(a_i, a_j) U^k_{i,j}$, where $\gamma(a_i, a_j)$ is the contact potential between amino acids a_i and a_j and $U^k_{i,j}$ is one if i and j are in contact in structure k and zero otherwise]. We use the contact potentials determined in the work by Miyazawa and Jernigan (43) based on their analysis of protein structures. Amino acids are considered to be in contact if their C_β atoms (C_α in the case of glycine) are closer than 7 Å to each other. After the scaling of the potential in the work by Miyazawa and Jernigan (43), all energies are represented in kilocalorie per mole.

To calculate the free energy of folding $\Delta G_{\text{Fold}}(S)$, we need to calculate the free energy for the native state as well as a large ensemble of alternative folds. For the native state, we use the conformation of the 300-residue purple acid phosphatase (Protein Data Bank ID code 1QHW) (44) to calculate the free energy $G_{\text{NS}}(S)$. We assume that the distribution of the free energies $\rho_{\text{U}}(G)$ of the large ensemble of thermodynamically relevant unfolded and alternative conformations can be represented by a Gaussian distribution with sequence-dependent average $\overline{G}(S)$ and variance $\sigma(S)^2$. Consider a large set (N_{U}) of possible unfolded structures with free energy values drawn from such a distribution. The free energy of folding is equal to (Eq. **1**)

$$\Delta G_{\text{Fold}}(S) = G_{\text{NS}}(S) + \frac{\sigma(S)^2 - 2kT\overline{G}(S)}{2kT} + kT \ln N_{\text{U}}. \quad [1]$$

N_{U} was set equal to 10^{160}. T was set equal to 20 °C.

We estimated the values of $\overline{G}(S)$ and $\sigma(S)^2$ by calculating the average free energy and variance of the free energies of the sequence in the conformation of the first 300 residues of 55 different structurally diverse protein structures. Combined with the value of $G_{\text{NS}}(S)$, we can compute $\Delta G_{\text{Fold}}(S)$ with Eq. **1**. We can then calculate the probability $P_{\text{Fold}}(S)$ that the protein would be folded at equilibrium (Eq. **2**):

$$P_{\text{Fold}}(S) = \frac{\exp(-\Delta G_{\text{Fold}}(S)/kT)}{1 + \exp(-\Delta G_{\text{Fold}}(S)/kT)}. \quad [2]$$

As in previous work, we considered the fitness of a sequence $\omega(S)$ to equal the probability of folding.

Evolutionary Dynamics. We initialized a protein sequence by choosing 300 codons at random (ignoring stop codons) using the standard genetic code to determine the encoded amino acids. At any point in the simulation, a random base underwent a mutation with probabilities based on the K80 model ($\kappa = 2$) (45). The fitness ω' of the resulting sequence was then computed based on the value of $\Delta G_{\text{Fold}}(S')$, the free energy of folding for this sequence, and the corresponding folding probability $P_{\text{Fold}}(S')$. This fitness was then compared with the fitness of the premutated sequence ω; the mutation was then accepted with a probability αf, with f (the fixation probability) calculated using the formula for diploid organisms by Crow and Kimura (28) and Kimura (29, 30) of (Eq. **3**)

$$f = \frac{1 - \exp(-2s)}{1 - \exp(-4N_{\text{Eff}}s)}, \quad [3]$$

where s is the selection coefficient equal to $s = \frac{\omega' - \omega}{\omega}$, N_{e} is the effective population size set equal to 10^6, and α is a number that varied over the course of the simulations but was always chosen so that $f < 1$ for all mutations. The evolutionary dynamics are only sensitive to relative rates of acceptance for the different mutations (as long as the evolutionary time is represented in terms of accepted substitutions; e.g., branch lengths). We note that this finding assumes that mutations have nonoverlapping phases, and because only one mutation is considered at a time, stochastic tunneling is not possible. Because the inclusion of α did not affect these relative acceptance rates, it had no effect on the results.

The simulation proceeded for a sufficient number of generations such that the stability of the protein reached equilibrium (i.e., the average fitness was approximately constant over time and across independent runs). Equilibrium is reached because of mutation selection balance, the point where there stabilizing mutations are relatively uncommon and have smaller relative fitness benefits, whereas destabilizing (but marginally acceptable) mutations are greater in number. The stability at this point was approximately −10 kcal/mol, which was approximately the stability observed in biological proteins of similar size, and was achieved after ~3,000 substitutions. All reported results were obtained after this preequilibration. We note that, because of the

form of Eq. **2** and its use as the fitness function, the relative fitness differences become smaller and smaller, with equivalent decreases in folding energy as the probability of folding approaches 1.0.

Propensities and Constraints. Given a protein sequence S at any point in the simulation, we can calculate the fitness ω_l^X of organisms containing that same sequence but with any amino acid X substituted at a focal location l in that protein. We express the acceptability of any amino acid at this location given fixed amino acids at all other locations by the propensity Π_l^X, which is given by (Eq. **4**)

$$\Pi_l^X = \frac{e^{2N_e\omega_l^X}}{\sum_Y e^{2N_e\omega_l^Y}}, \quad [4]$$

where the sum in the dominator is over the 20 amino acids. This propensity would be equal to equilibrium frequency π_l^X if the frequencies of the nucleotides were equal and there was no redundancy in the genetic code.

The degree to which a site is constrained in its amino acid composition was determined by using A, the effective size of the alphabet of amino acids possible at a given location. This variable is defined as the exponential of the sequence entropy at this location (Eq. **5**):

$$A_l = \exp\left(-\sum_X \Pi_l^X \ln\left(\Pi_l^X\right)\right). \quad [5]$$

The rate at which the system loses memory of its previous amino acid preferences was measured using the decay of the autocorrelation function of the amino acid propensities, which given $K = 20$ amino acids, is defined as (Eq. **6**)

$$R_l(\tau) = \frac{E\left[\left(\Pi_l^X(t) - \frac{1}{K}\right)\left(\Pi_l^X(t+\tau) - \frac{1}{K}\right)\right]}{E\left[\left(\Pi_l^X(t) - \frac{1}{K}\right)^2\right]}, \quad [6]$$

where t and τ are measured in numbers of substitutions. The distribution of the autocorrelation with different values of τ was modeled by fitting to a stretched exponential of the form $\hat{R}(\tau) = (1-b)\exp\left[-\left(\tau/\tau_k\right)^\beta\right] + b$, where b is the equilibrium correlation, τ_k is a scaling parameter, and β is the stretching parameter (46). The stretched exponential is a generalization of the exponential function commonly used to describe relaxation in disordered systems; $\beta = 1$ corresponds to a standard exponential function, whereas values of $\beta < 1$ lead to a stretching effect. The average relaxation time is given by $\langle\tau\rangle = \frac{\tau_k}{\beta}\Gamma\left(\frac{1}{\beta}\right)$.

Thermodynamic Calculations on Biological Proteins. To estimate the change in stability $\Delta\Delta G$ resulting from mutations in biological proteins of known structure, we used the ddg_monomer application from the Rosetta library (31) The structures were first preoptimized to reduce any clashes that may be present in the crystal structure. The optimization process involves running three rounds of energy minimization starting with a lower repulsive value of the van der Waals term and increasing it to the normal value by the third round of minimization. The process also allows for slight backbone movements to compensate for large or small side chain substitutions. The minimization process is done on both the WT and mutated structure. We ran the application using the recommended suggestions by finding the minimum $\Delta\Delta G$ value after 50 iterations of the optimization process.

ACKNOWLEDGMENTS. We thank Jeffrey Thorne and Clemens Lakner for helpful discussions and Michael Sadowski for assistance with Rosetta. We acknowledge the support of National Institutes of Health Grant GM083127 and the Medical Research Council United Kingdom.

1. Pollock DD, Bruno WJ (2000) Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. *Mol Biol Evol* 17:1854–1858.
2. Castoe TA, et al. (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci USA* 106:8986–8991.
3. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* 6:e1000633.
4. Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R (2005) Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics* 21(Suppl 1):i126–i135.
5. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18:309–317.
6. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28:184–190.
7. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6:e28766.
8. Pollock DD, Taylor WR, Goldman N (1999) Coevolving protein residues: Maximum likelihood identification and relationship to structure. *J Mol Biol* 287:187–198.
9. Taylor WR, Jones DT, Sadowski MI (2012) Protein topology from predicted residue contacts. *Protein Sci* 21:299–305.
10. Wang ZO, Pollock DD (2005) Context dependence and coevolution among amino acid residues in proteins. *Methods Enzymol* 395:779–790.
11. Wang ZO, Pollock DD (2007) Coevolutionary patterns in cytochrome c oxidase subunit I depend on structural and functional context. *J Mol Evol* 65:485–495.
12. Rodrigue N, Lartillot N, Bryant D, Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347:207–217.
13. Rodrigue N, Philippe H, Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol* 23:1762–1775.
14. Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, ed Dayhoff M (National Biomedical Research Foundation, Silver Spring, MD), pp 345–352.
15. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282.
16. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699.
17. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL (1992) Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci* 1:216–226.
18. Koshi JM, Goldstein RA (1995) Context-dependent optimal substitution matrices. *Protein Eng* 8:641–645.
19. Thorne JL, Goldman N, Jones DT (1996) Combining protein evolution and secondary structure. *Mol Biol Evol* 13:666–673.
20. Koshi JM, Goldstein RA (1998) Models of natural mutations including site heterogeneity. *Proteins* 32:289–295.
21. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095–1109.
22. Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol* 53:571–581.
23. Fisher R (1930) *The Genetical Theory of Natural Selection* (Clarendon, Oxford).
24. Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Mol Biol Evol* 15:910–917.
25. Rodrigue N, Philippe H, Lartillot N (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci USA* 107:4629–4634.
26. Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* 102:9541–9546.
27. Tamuri AU, Dos Reis M, Goldstein RA (2012) Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* 190:1101–1115.
28. Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory* (Harper & Row, New York).
29. Kimura M (1957) Some problems of stochastic processes in genetics. *Ann Math Stat* 28:882–901.
30. Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics* 47:713–719.
31. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* 79:830–838.
32. Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol* 18:866–873.
33. Penny D, McComish BJ, Charleston MA, Hendy MD (2001) Mathematical elegance with biochemical realism: The covarion model of molecular evolution. *J Mol Evol* 53:711–723.
34. Jensen JL, Pedersen AMK (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv Appl Probab* 32:499–517.
35. Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* 21:468–488.
36. Weinreich DM, Watson RA, Chao L (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59:1165–1174.
37. Orr HA, Turelli M (2001) The evolution of postzygotic isolation: Accumulating Dobzhansky-Muller incompatibilities. *Evolution* 55:1085–1094.
38. Turelli M, Orr HA (2000) Dominance, epistasis and the genetics of postzygotic isolation. *Genetics* 154:1663–1679.
39. Unckless RL, Orr HA (2009) Dobzhansky-Muller incompatibilities and adaptation to a shared environment. *Heredity (Edinb)* 102:214–217.
40. Kulathinal RJ, Bettencourt BR, Hartl DL (2004) Compensated deleterious mutations in insect genomes. *Science* 306:1553–1554.
41. Goldstein RA (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* 79:1396–1407.
42. Williams PD, Pollock DD, Blackburne BP, Goldstein RA (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol* 2:e69.
43. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* 18:534–552.
44. Lindqvist Y, Johansson E, Kaija H, Vihko P, Schneider G (1999) Three-dimensional structure of a mammalian purple acid phosphatase at 2.2 A resolution with a mu-(hydr)oxo bridged di-iron center. *J Mol Biol* 291:135–147.
45. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120.
46. Klafter J, Shlesinger MF (1986) On the relationship among three theories of relaxation in disordered systems. *Proc Natl Acad Sci USA* 83:848–851.