Received: 17 March 2021 | Revised: 25 June 2021 | Accepted: 29 June 2021

DOI: 10.1002/pro.4155

FULL-LENGTH PAPER

THE PROTEIN SOCIETY WILEY

A thermodynamic model of protein structure evolution explains empirical amino acid substitution matrices

Christoffer Norn[1] | Ingemar André[1] | Douglas L. Theobald[2]

[1]Biochemistry and Structural Biology, Lund University, Lund, Sweden

[2]Biochemistry Department, Brandeis University, Waltham, Massachusetts, USA

Correspondence
Ingemar André, Biochemistry and Structural Biology, Lund University, PO BOX 124, Lund, Sweden.
Email: ingemar.andre@biochemistry.lu.se

Douglas L. Theobald, Biochemistry Department, Brandeis University, Waltham, MA 02453, USA.
Email: dtheobald@brandeis.edu

Funding information
National Institute of General Medical Sciences, Grant/Award Numbers: R01GM096053, R01GM132499; Vetenskapsrådet, Grant/Award Number: 2015-04203

Abstract

Proteins evolve under a myriad of biophysical selection pressures that collectively control the patterns of amino acid substitutions. These evolutionary pressures are sufficiently consistent over time and across protein families to produce substitution patterns, summarized in global amino acid substitution matrices such as BLOSUM, JTT, WAG, and LG, which can be used to successfully detect homologs, infer phylogenies, and reconstruct ancestral sequences. Although the factors that govern the variation of amino acid substitution rates have received much attention, the influence of thermodynamic stability constraints remains unresolved. Here we develop a simple model to calculate amino acid substitution matrices from evolutionary dynamics controlled by a fitness function that reports on the thermodynamic effects of amino acid mutations in protein structures. This hybrid biophysical and evolutionary model accounts for nucleotide transition/transversion rate bias, multi-nucleotide codon changes, the number of codons per amino acid, and thermodynamic protein stability. We find that our theoretical model accurately recapitulates the complex yet universal pattern observed in common global amino acid substitution matrices used in phylogenetics. These results suggest that selection for thermodynamically stable proteins, coupled with nucleotide mutation bias filtered by the structure of the genetic code, is the primary driver behind the global amino acid substitution patterns observed in proteins throughout the tree of life.

KEYWORDS

amino acid substitution, exchangeabilities, protein evolution, protein stability, replacement matrices

1 | INTRODUCTION

Protein amino acid sequences change due to spontaneous mutations at the DNA level. Amino acid exchange rates depend not only on the background mutation rate, but also on how the mutation at the protein level impacts overall organismal fitness. Empirical models that incorporate global amino acid substitution matrices have wide-ranging applications in bioinformatics and evolutionary science, including homology search, phylogenetics, ancestral sequence reconstruction, and the prediction of functional residues. While such empirical models have tremendous practical utility, the importance of various fitness pressures in shaping the patterns of amino acid substitutions is still unknown. Amino acid mutations can impact fitness by modifying protein properties such as catalysis, binding, expression, aggregation, non-specific interactions, and protein stability, the latter of which is known to be a major selection pressure at most sites for all proteins.[1–4] Here we develop from first principles a biophysical and genetic model that predicts amino acid substitution rates by assuming that the fitness effects of amino acid mutations arise from changes in protein stability alone. We then evaluate the ability of thermodynamic stability to explain empirical amino acid substitution patterns.

The prevailing phylogenetic models of protein evolution explicitly describe amino acid substitution processes using instantaneous rate matrices whose parameters are inferred from observed sequences using statistical methods. Substitutions are modeled as an aggregated Markov process in which sites evolve independently with a substitution rate that depends on the identity of the current amino acid. All widely used phylogenetic programs (e.g., PhyML,[5] IQ-TREE,[6] RAxML,[7] MrBayes,[8] BAli-Phy,[9] and PhyloBayes[10]) incorporate global substitution (or exchangeability) matrices as a key component of the default evolutionary model. Some of the most common and successful global matrices include the JTT,[11] WAG,[12] and LG[13] exchangeability matrices. Similarly, NCBI blastp[14] uses the BLOSUM62 matrix[15] by default for remote homology detection. Global substitution matrices are empirically constructed by averaging rate processes over numerous sites in many proteins. These global matrices contain amino acid rate information that has been incredibly useful and effective in practice, and they are essential to modern phylogenetics and computational biology. All these global amino acid substitution matrices have highly similar, consistent patterns that call out for an explanation.

The pairwise amino acid rate constants (known as the "exchangeabilities") in global substitution matrices are a function of the inherent physical properties of amino acids and folded proteins, and hence they should be predictable from first principles. To understand the origin of the variation in global amino acid exchangeabilities, several studies have established correlations with biophysical amino acid descriptors such as hydrophobicity, secondary structure propensity, charge, and codon table structure.[16–19] While these phenomenological correlations help identify factors that influence amino acid substitutions, biophysical and genetic models are required for a mechanistic understanding of the underlying fitness constraints that give rise to the empirical substitution patterns.

Over the past two decades, mechanistic models that combine evolutionary dynamics and protein biophysics have led to significant advances in our understanding of the causes of evolutionary rate variation in proteins.[20–23] Biophysical models of rate variation typically treat molecular fitness as dominated by fold stability or the stability of an active structure. Evolutionary trajectories simulated with biophysical fitness models have shown how fluctuations in the structural environment within proteins influence substitution rates[24] and produce epistatic effects between sites.[25] However, such structure-based evolutionary models have yet to be applied to understand the origin of global amino acid substitution patterns.

Building on this previous work, here we develop a biophysical model of amino acid substitution rate variation to investigate the evolutionary and thermodynamic basis for global substitution patterns in proteins. Since global substitution matrices are constructed by averaging over protein sites, our theoretical analysis must necessarily also average over sites to offer a plausible explanation of these ubiquitous global substitution patterns. We first calculate the thermodynamic effects of amino acid mutations in native protein structures and combine a fitness function solely based on protein stability[26] with a position-specific codon-level model of sequence evolution[27] to predict global amino acid substitution rates. Using this model, we find that our calculated amino acid substitution rates are strongly correlated with experimental values described by global empirical substitution matrices such as the widely used LG matrix. Most of the empirical amino acid substitution pattern can be explained wholly by mutation combined with selection pressure to maintain thermodynamically stable protein structures.

2 | RESULTS

2.1 | A parametric all-atom thermodynamic model of protein evolution

We selected a non-redundant, curated set of 52 high-quality protein structures for the basis of our analysis.[28,29] These 52 protein structures were subjected to a computational analog of saturated mutagenesis. For each site in every protein, we calculated the thermodynamic effect ($\Delta\Delta G$) of all possible amino acid mutations with the Rosetta macromolecular modeling suite.[30] We assume that fitness is proportional to the fraction of folded protein,[26] which is determined by the $\Delta\Delta G$ value, and then calculate the probability of fixation in a finite population using the Kimura equation[31,32] (Figure 1). A global amino acid substitution matrix can then be constructed from these amino acid fixation probabilities and a codon-level model of nucleotide mutation. There are only four free parameters in this mechanistic evolutionary model: (a) the free energy of the native protein, ΔG_{nat}, (b) the effective population size, N_e, (c) the nucleotide transition rate, κ, and (d) a whole-codon rate parameter, ρ. Values of these four parameters are required to calculate a substitution matrix, because we do not know the values of these parameters a priori, we optimize them with the method of maximum likelihood. The end-result of this method is an amino acid exchangeability matrix (189 free exchangeability rate constants) and an amino acid equilibrium frequency vector (19 free probabilities) that has been fit using only three or four free parameters, each with a clear physical and evolutionary interpretation.

2.2 | Thermodynamic effects of amino acid mutations

A full atomistic model of protein structure and energetics is necessary to realistically model site-specific behavior. However, computing the folding stability of proteins at this level of detail is currently intractable for simulations of protein evolution, as it requires extensive sampling of alternative conformations for every sequence evaluated. Instead, we treat the folding stability of native sequences (ΔG_{nat}) as a global free parameter in the model. In contrast to folding stability, the free energy change upon amino acid mutation ($\Delta\Delta G$) is both reasonably fast to compute and fairly accurate ($r^2 = .56$ between prediction and experiment[33]). We therefore calculate the folding stability of a sequence variant at site L as:

$$\Delta G_j^L = \Delta G_{\text{nat}} + \Delta\Delta G_{ij}^L \tag{1}$$

where i indicates the index of the native amino acid at site L and j indicates the index of the proposed amino acid mutation at that site. The equilibrium properties of each position are conditioned only on the fitness pressure exerted by the native structural environment in the selected proteins. In contrast to using a mutation-after-mutation simulation method, this eliminates the need for sequence pre-equilibration, avoids compounding errors of model and energy function, and allows computation of a site-specific Q-matrix from just 20 $\Delta\Delta G$ calculations.[27]

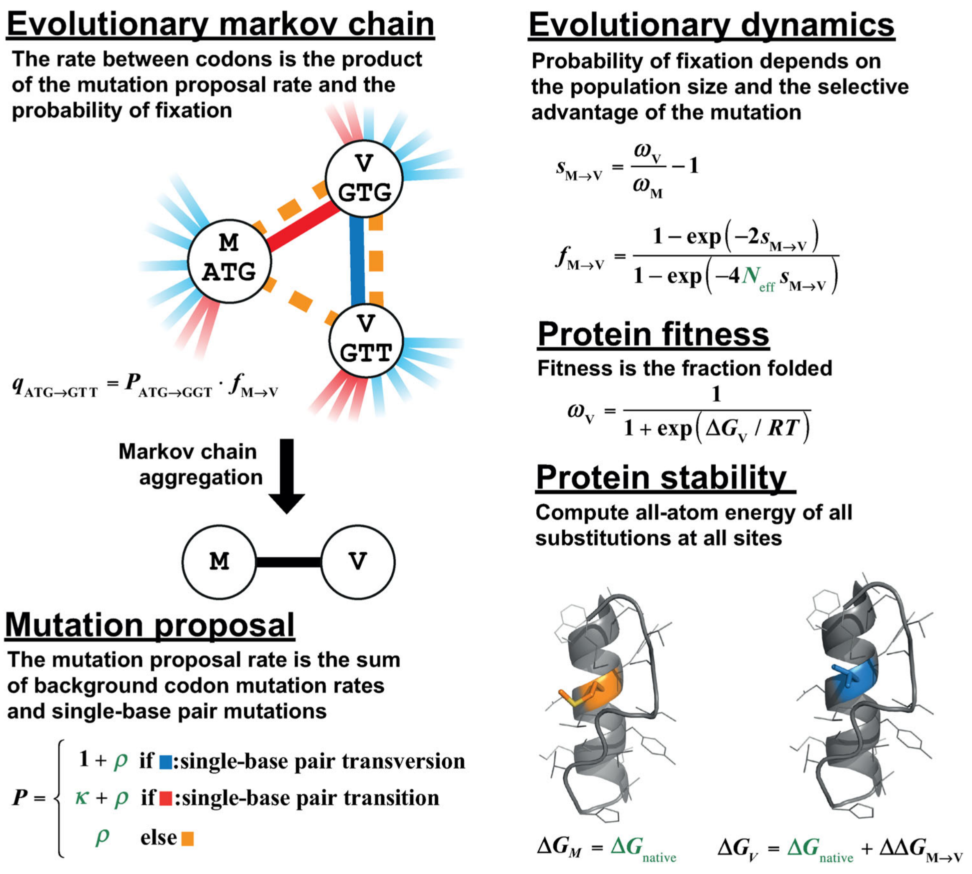
Evolutionary markov chain
The rate between codons is the product of the mutation proposal rate and the probability of fixation
V GTG
M ATG
V GTT
$q_{\text{ATG}\to\text{GTT}} = P_{\text{ATG}\to\text{GGT}} \cdot f_{\text{M}\to\text{V}}$
Markov chain aggregation
M
V
Mutation proposal
The mutation proposal rate is the sum of background codon mutation rates and single-base pair mutations
$P = \begin{cases} 1+\rho & \text{if ■:single-base pair transversion} \\ \kappa+\rho & \text{if ■:single-base pair transition} \\ \rho & \text{else ■} \end{cases}$
Evolutionary dynamics
Probability of fixation depends on the population size and the selective advantage of the mutation
$s_{\text{M}\to\text{V}} = \frac{\omega_\text{V}}{\omega_\text{M}} - 1$
$f_{\text{M}\to\text{V}} = \frac{1-\exp(-2s_{\text{M}\to\text{V}})}{1-\exp(-4N_{\text{eff}}\, s_{\text{M}\to\text{V}})}$
Protein fitness
Fitness is the fraction folded
$\omega_\text{V} = \frac{1}{1+\exp(\Delta G_\text{V}/RT)}$
Protein stability
Compute all-atom energy of all substitutions at all sites
$\Delta G_M = \Delta G_{\text{native}}$
$\Delta G_V = \Delta G_{\text{native}} + \Delta\Delta G_{\text{M}\to\text{V}}$

FIGURE 1 Overview of method illustrated for a M to V mutation. The rate of substitution between any two amino acids in a protein is evaluated based on an underlying codon mutation Markov process. Mutations arise according to the mutation proposal model and fix depending on the fitness difference according to an evolutionary dynamics model. Fitness is proportional to the fraction folded (i.e., the probability of the native state) based on an all-atom energy function. The model contains four free parameters (ΔG_{nat}, N_e, κ, ρ) that are simultaneously optimized

2.3 | Protein fitness as a function solely of thermodynamic stability

To understand to what degree thermodynamic stability explains the process of protein evolution, we base our fitness function on folding stability alone. Following previous work,[26] we assume that a protein's contribution to fitness is proportional to the fraction of the protein folded into its functional conformation. The fraction of protein folded ω for a two-state folding model with stability ΔG_i^L is given by

$$\omega_i^L = \frac{1}{1+\exp\left(\Delta G_i^L/RT\right)} \tag{2}$$

2.4 | Fixation probability in a finite population

Given enough time, a proposed mutation will either spread throughout the population and fix or be purged due to negative selection or genetic drift. We assume throughout a monomorphic population evolving under weak-mutation.[34] The probability of fixation when amino acid i mutates to amino acid j at site L depends on the relative change in fitness due to that mutation (the selection coefficient $s_{ij}^L = \omega_j^L/\omega_i^L - 1$) and the effective population size N_e:[32]

$$f_{ij}^L = \frac{1-\exp\left(-2s_{ij}^L\right)}{1-\exp\left(-4N_e s_{ij}^L\right)} \tag{3}$$

The effective population size N_e is another global free parameter in the model. The fixation probability we use above is for diploid populations with non-overlapping generations in which a new mutation arises as a single copy. Similar equations, with exponential terms differing by a factor of two, apply to haploids and overlapping generations,[35] but the effects on our analysis are negligible for large population sizes (e.g., $N_e > 100$).

2.5 | Nucleotide and codon mutation proposals

Most spontaneous mutations involve changes to single-base pairs. Due to the intrinsic chemical properties of nucleotides,[36,37] point mutations occur with greater rates for transitions (pyrimidine to pyrimidine or purine to purine) than transversions. To capture this bias in our model, the transversion rate is fixed at 1.0 and the transition rate κ is treated as a free parameter. A smaller subset of mutations involve changes of multi-nucleotides or whole codons, for instance as may result from insertions, deletions, or tandem mutations due to error-prone replication[38] or UV damage.[39] These types of mutations have not been modeled previously in structure-based simulations of evolution, but they play an important role in the evolution of natural sequences.[40] We model the rate of whole-codon mutations, scaled relative to transversions, with a single parameter ρ. The mutation proposal probability P at a given site is:

$$P = \begin{cases} 1+\rho & \text{if transversion} \\ \kappa+\rho & \text{if transition} \\ \rho & \text{else} \end{cases} \tag{4}$$

2.6 | Construction of amino acid rate matrices from codon mutations and protein fixation

We model the site-specific evolutionary process as a continuous-time Markov process, where the unit of change is the codon and the relative instantaneous substitution rate q_{uv}^L from codon u with amino acid i to codon v with amino acid j is the product of the mutation proposal rate, P_{uv}, and the fixation probability, f_{ij}^L:[27]

$$q_{uv}^L = P_{uv} \cdot f_{ij}^L \tag{5}$$

Codons representing the same amino acids are then aggregated, allowing determination of site-specific amino acid flux matrices. The flux matrices are averaged across sites and proteins to construct a global 20 × 20 amino acid instantaneous rate matrix Q. For comparison with empirical global substitution matrices such as JTT, WAG, and LG, and for use in phylogenetic analyses, the Q-matrix was decomposed into a diagonal amino acid equilibrium probability matrix, π, and an independent symmetric 20 by 20 exchangeability matrix R:[13,41]

$$Q = \pi R \tag{6}$$

The rate constants in the symmetric exchangeability matrix R can be considered normalized equilibrium fluxes between amino acids, where the fluxes have been normalized by the appropriate equilibrium amino acid probabilities (see Methods). The equilibrium probabilities, in turn, describe the expected frequency distribution of amino acids when the substitution process has progressed long enough that it reaches equilibrium.

2.7 | Optimized model parameters provide typical values

Using the computed mutation energies ($\Delta\Delta G$), an optimal exchangeability matrix R was found by optimizing values of the four parameters of the model $\theta = \{\Delta G_{\text{nat}}, N_e, \kappa, \rho\}$ by maximizing the phylogenetic likelihood for the 52 non-redundant protein families. With grid-search optimization, we found an optimal parameter set of $\Delta G_{\text{nat}} = -6.0 \pm 0.1$ kcal/mol, $\log(N_e) = 3.8 \pm 0.1$, $\kappa = 2.1 \pm 0.1$, and $\rho = 0.11 \pm 0.01$ (standard error estimated by bootstrap,[42] see supporting methods). We refer to this optimized matrix as the Thermodynamic Mutation-Selection (TMS) matrix. Alternatively, parameter values can also be determined by direct optimization against, say, the LG matrix, a procedure that gives nearly identical results as the maximum likelihood method (Figure S1). Because the TMS model parameters are based on fundamental quantities in population genetics (N_e), spontaneous mutation processes (κ, ρ) and protein thermodynamics (ΔG_{nat}), the optimal parameter values can be compared with independent empirical measurements from protein biochemistry and population genetics. As discussed further below, these optimal values from our thermodynamic, evolutionary model correspond surprisingly well to representative physical and biological values.

The stability of natural proteins is typically between -5 and -10 kcal/mol,[43] while effective population sizes vary over many orders of magnitude. For instance, the effective population size for humans is on the order of 10^{13},[44] while *Escherichia coli* have an population size of approximately 10^3.[45] The optimal ΔG_{nat} and N_e TMS values compare well with these ranges. By inspecting the likelihood surface (see Figure S2), the stability ΔG_{nat} and population N_e parameters are seen to be highly correlated, with optimal parameters found along a line given by

$$\Delta G_{\text{opt}} \approx -1.5 \cdot \log\left(N_{\text{opt}}\right) - 0.2 \qquad (7)$$

Along this line, substitution parameters do not vary substantially (Figure S3). A similar dependence of protein stability ΔG on N_e at equilibrium has been noted previously.[4,46] The strong correlation between ΔG and N_e suggests that they represent a single latent parameter, reducing the effective parameter space of our model from four to three.

In contrast to stability and population size, the parameters describing the mutation process (κ and ρ) are largely independent (Figure S2). The optimal transition/transversion bias $\kappa = 2.1$ is consistent with experimental data for spontaneous mutation rates from *E. coli*,[47] where $\kappa \simeq 2.6$ (assuming a K80-type Markov model,[48] see Appendix S1). For the whole codon mutation parameter ρ, the MLE is 0.1 relative to the transversion rate. This corresponds to that 29% of all proposed mutations are multi-nucleotide mutations. Obtaining direct empirical estimates of ρ is difficult, but for indels of 2–4 nucleotides in *E. coli*, the spontaneous mutation rate is approximately 10% of the single-nucleotide mutation rate.[47] In eukaryotes, multi-nucleotide mutations comprise approximately 3% of all mutations[49] and have been shown to be important in phylogenetic tests.[50] Given that processes other than indels can result in multi-nucleotide mutations, and that indels typically result in multiple codon mutations, our ρ_{opt} value appears reasonable.

2.8 | The thermodynamic TMS model reproduces empirical substitution patterns

Most of the variation in empirical amino acid substitution matrices can be explained by our thermodynamic evolutionary model, as judged by the logarithmic correlation coefficient with the ML TMS matrix (Figure 2). The TMS matrix appears to be a rather typical exchangeability matrix, as the average correlation of TMS with many widely used empirical matrices is $r^2 = .54$, whereas the correlation of those empirical matrices with each other

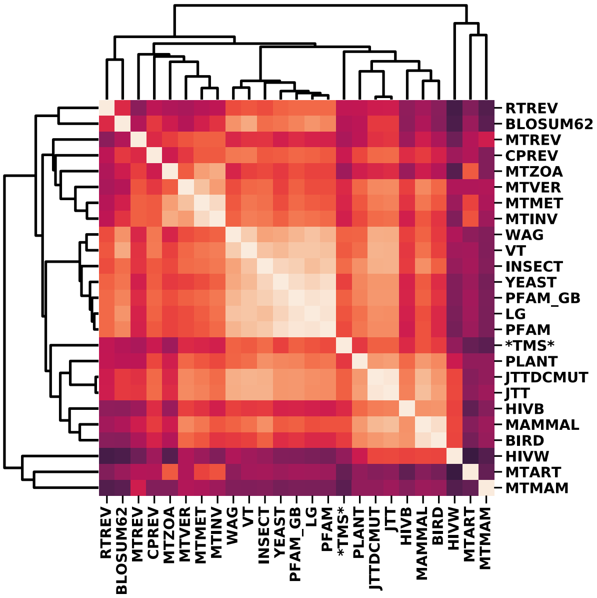

FIGURE 2 Heatmap of correlations between popular exchangeability matrices. Squared Pearson correlations (r^2) were computed in log-space over the 190 exchangeability parameters. Matrices were clustered based on r^2 using hierarchically clustering.[51] The TMS matrix is indicated by asterisks

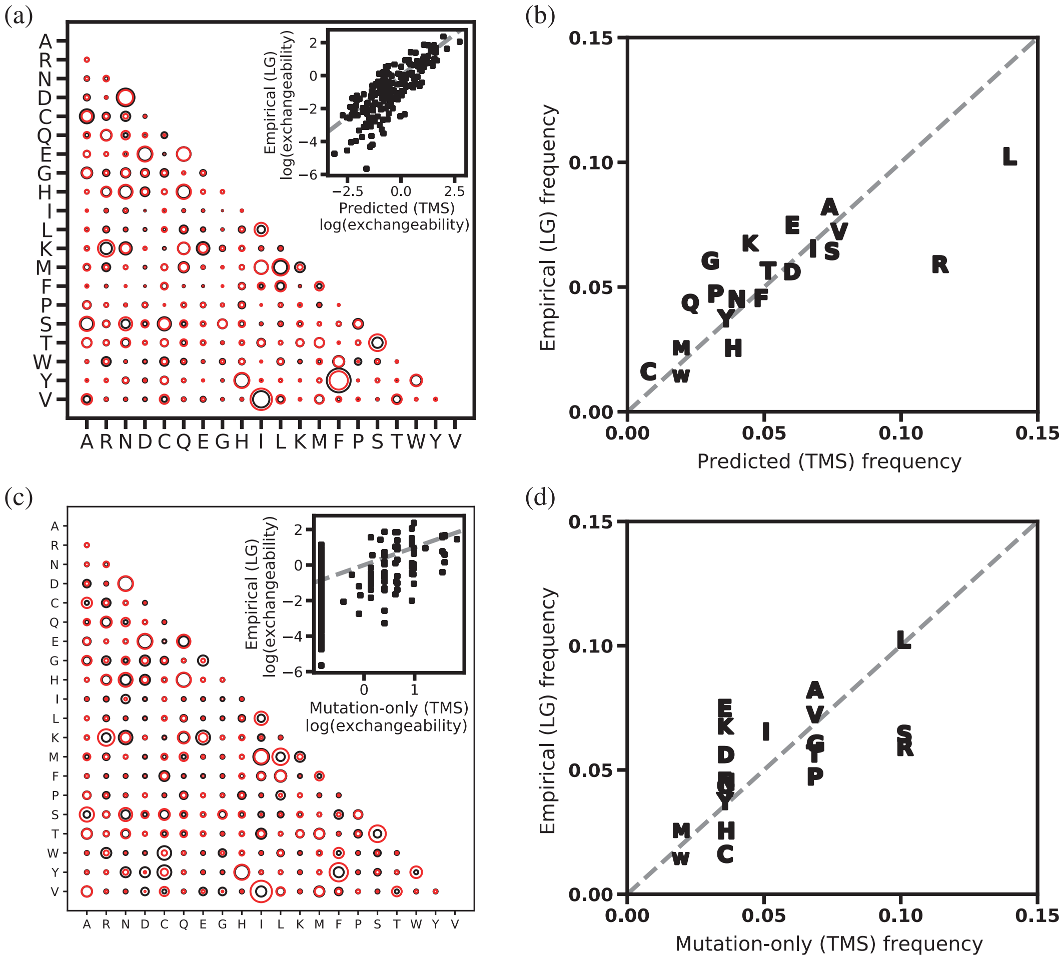

FIGURE 3 The TMS model recapitulates the mean substitution behavior between amino acids. (a) Comparison of the amino acid exchangeability matrix predicted by TMS (black circles) and the LG phylogenetic empirical exchangeability matrix (red circles). Inset, correlation between exchangeabilities from TMS (*x*-axis) versus LG (*y*-axis). (b) Correlation between amino acid equilibrium probabilities predicted by TMS (*x*-axis) and values from LG (*y*-axis). Identity shown as dashed line. (c) The "mutation-only" exchangeability matrix, without selection for stability, vs LG. (d) Correlation between mutation-only equilibrium probabilities and those from LG. Identity shown as dashed line

(excluding TMS) is $r^2 = .52$ (Figure 2). The average correlation with globular matrices is higher ($r^2 = .59$), whereas the correlation with mitochondrial matrices is substantially lower ($r^2 = .42$). For example, the TMS exchangeability matrix has an r^2 of .67 and .64 with the widely used WAG (Figure S4) and LG substitution matrices (Figure 3), respectively. The lower correlation with mitochondrial matrices is likely due to the preponderance of transmembrane proteins in the mitochondrial datasets,[52,53] as the Rosetta energy function used in our thermodynamic free energy calculations is intended for soluble proteins. Breaking exchangeabilities down by individual amino acids, we see the lowest correlation for cysteine and proline ($r^2 = .50$), likely reflecting shortcomings of the energy function and limited modeling of backbone flexibility. The other amino acids have higher correlations, up to $r^2 = .85$ (Figure 4).

The high correlation between TMS and other experimental matrices like LG is a result of both the thermodynamic model, in which more chemically similar amino acids have higher fixation probabilities, and the mutation model, which biases amino acid replacements in the genetic code due to preferred nucleotide mutations, codon number, and codon connectivity. To understand the impact of the genetic component on the correlations, we computed a "mutation-only" *R*-matrix, assuming only the connectivity of the codon table and the transition-transversion bias by setting all fixation probabilities in Equation 5 to 1 and setting $\rho = 0.10$ and $\kappa; = 2.1$ to their optimal values. This mutation-only *R*-matrix accounts for 32% of the variation in the LG substitution pattern (Figure 3c), suggesting that genetic and thermodynamic factors contribute roughly equally to the patterns of empirical amino acid substitutions.

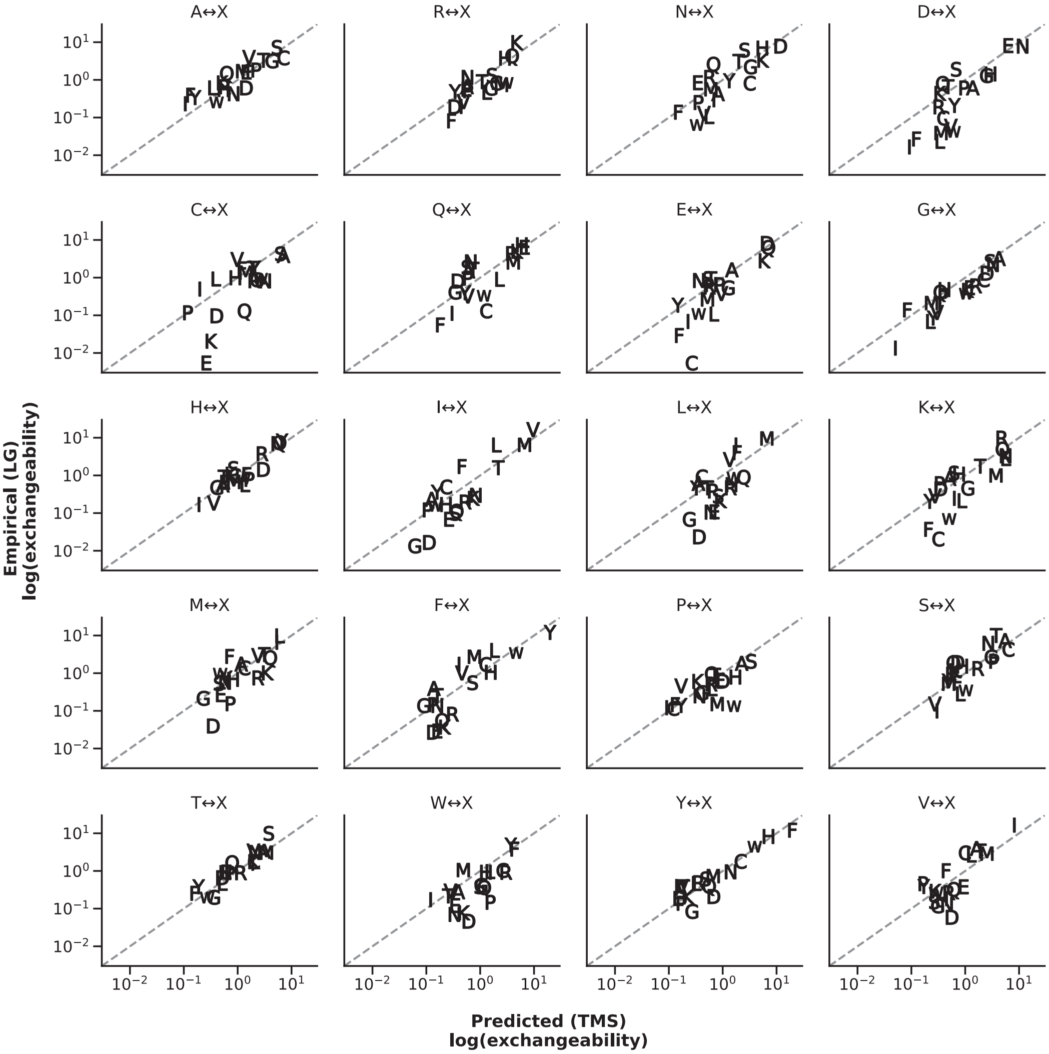

FIGURE 4 The TMS model recapitulates amino acid-specific exchangeabilities. Identity is shown as a dashed line

Correct recapitulation of amino acid exchangeabilities is strongly dependent on the thermodynamic component of the model. For example, using LG as the standard for comparison, the empirical exchangeability for tryptophan to tyrosine is underestimated by 7.4-fold by the mutation-only *R*-matrix but is underestimated by 1.1-fold by TMS. Similarly, the mutation-only model overestimates the empirical exchangeability of asparagine to phenylalanine by 4.8-fold, while TMS overestimates it by 1.3-fold.

Our ML TMS matrix was optimized over the exchangeabilities and not the amino acid equilibrium frequencies. Nevertheless, from the decomposition of the Q-matrix (given in Equation 6), we can provide the implied TMS equilibrium frequencies and compare them with those provided with common exchangeability matrices (Figure S5). For example, the calculated TMS equilibrium frequencies correlate well with those from LG (linear $r^2 = .65$), but suggest unmodelled fitness effects. For instance, we do not model disulfide formation, and unsurprisingly the stationary frequency of cysteine calculated from TMS is 2.8-fold too low. Likewise, we ignore codon usage biases, which could explain the overestimation of the background frequency of arginine and leucine, both of which have exceptionally skewed codon usage.[54]

Although our model correlates strongly with empirical exchangeability matrices like LG, some of the variation in the rate constants remains unexplained. One possible contribution to this discrepancy stems from the fact that we selected a set of proteins that are different from those used to infer the LG-matrix. Our set of proteins will likely have a slightly different substitution process than the LG set. To investigate this possibility, we used a maximum likelihood phylogenetic inference method,[55] similar to that used for the construction of the LG matrix, to infer exchangeabilities from the sequence alignments for our set of benchmark proteins (Figure S6). The inferred phylogenetic exchangeability matrix for our proteins is highly correlated with LG ($r^2 = .97$), indicating that our benchmark proteins are similar to those that LG was derived from, and that stochastic errors originating from the rate-inference itself are minimal.

A second source of the unexplained variation could be errors in the energy function used for $\Delta\Delta G$ prediction, as Rosetta energies have an imperfect correlation with empirical values ($r^2 = .56$). We explored this possibility by simulating noisy data using empirically determined $\Delta\Delta G$ Rosetta prediction errors. This analysis suggested at least 21% of the unexplained variance could be caused by energy function errors (see Appendix S1).

Finally, our mutation-selection model operates at the codon level, producing rates that are subsequently aggregated to the amino acid level—a reduction in dimensionality that can affect the assumed Markov property of codon and amino acid evolution.[56] It is possible that our codon-level model could explain codon exchangeability matrices better than amino acid matrices. Additionally, when determining an amino acid exchangeability matrix like LG by phylogenetic methods, all 189 exchangeability rate constants are free parameters that are estimated independently from the sequence data. In contrast, our codon model is highly constrained by only two parameters (κ and ρ) and the K80 assumption of equal nucleotide frequencies. More complex mutational models could provide higher correlations at the expense of model simplicity and interpretability. To explore this possibility, we extended our codon model with the T92 nucleotide mutation proposal model.[57] Compared to the K80 model, the T92 model adds a single additional parameter that addresses G+C content bias (%-GC), which captures most of the variation in nucleotide frequencies per Chargaff's second rule.[58] The optimal %-GC bias for the exchangeabilities R is 0.51 and for the equilibrium frequencies π is 0.52, which barely deviates from the K80 value fixed at 0.5 (see Figure S7). At the optimal %-GC bias values, the T92 model does not improve the correlation of the TMS matrix with empirical substitution matrices (such as LG), suggesting that equal nucleotide frequencies in the K80 model is a relatively weak assumption that does not significantly limit the fit of the TMS matrix.

2.9 | TMS explains some phylogenies better than empirical substitution matrices

The purpose of our study is to see how far a simple thermodynamic and mutation-selection model could go in explaining empirical amino acid substitutions; it was not intended to provide a better substitution matrix for practical use. Nevertheless, it is interesting to see how well our TMS matrix fares in phylogenetic analyses. For each of the 52 individual phylogenetic trees in our benchmark set, we compared the TMS maximum likelihood values to the LG maximum likelihood values for the same alignments. The mean TMS likelihood was −13,664 while the mean LG likelihood was −13,492, a difference of 172 on average in favor of LG. For most trees, the LG matrix is better, but for three of the 52 trees TMS has a higher likelihood than LG. Using a larger set of 500 Pfam alignments,[13,59] independent of both our model parameters and LG, we similarly found that in eight alignments TMS had higher likelihood than LG (see Table S1). Thus, for a certain subset of protein families, TMS does appear to provide a better substitution model than LG.

3 | DISCUSSION

We present an approach to predict relative rates of global amino acid substitutions in evolution from first principles by combining population genetics and protein biophysics. We find that most of the global substitution behavior of amino acids in proteins can be explained by a simple evolutionary fitness model that captures (a) the thermodynamic effects of mutations, (b) the biases in spontaneous mutations, and (c) the structure of the genetic code. Our model is based on an extremely simple assumption: the fitness of a gene variant is controlled only by constraints on protein stability. Nevertheless, this naive model can remarkably recapitulate the complex amino acid substitution patterns seen in empirically derived substitution matrices. Our model incorporates transition/transversion mutation biases, multi-nucleotide codon changes, variation in codon counts due to the genetic code, and free energy changes calculated from the detailed atomic interactions at specific sites in proteins. By calculating the fitness effects of mutations in the structural environment of native sequences and averaging over many sites in many proteins, we are able to reproduce the majority of the amino acid substitution patterns quantified in common global exchangeability matrices.

Previous work by others has used structure-based models of evolution to predict amino acid substitution rates and equilibrium frequencies at specific sites in proteins.[46,60–64] Such methods have been applied to predict site-specific substitution matrices, rather than global matrices, but the results are difficult to validate against empirical site-specific matrices due to the lack-of-data problem at specific sites.[65] Arenas and Bastolla have emphasized that there are currently two incomplete kinds of structure-based evolutionary models described in the literature: stability-constrained fitness models and structurally-constrained fitness models.[66,67] Stability-constrained models calculate fitness from the effect of a mutation on protein stability, but ignore the structural changes due to the mutation.[46,60,61,67,68] On the other hand, structurally constrained models calculate fitness from the effect of a mutation on the structure, but ignore the change in stability.[62,64,69] Notably, our method bridges both categories of structure-based fitness model, as we simultaneously estimate the change in structure due to a mutation (allowing local backbone rearrangements and side chain repacking) and calculate the free energy change of the resulting perturbed structure relative to the native state using a state-of-the-art energy function.

Like the previous structure-based evolutionary methods described above, we also calculate site-specific rates. However, in our analysis these rates are then averaged over sites and proteins to arrive at the predicted global TMS exchangeability matrix. What then is the physical interpretation of a global substitution matrix? The exchangeabilities in amino acid substitution matrices are rate constants that can be considered to be instantaneous probabilities. The probability of a substitution at a given site is a function that is conditional on the particular physical environment at that protein site and depends on the chemical and physical properties of the current and mutated amino acids. These amino acid properties are largely constant regardless of the different chemical environments at different sites in the different proteins found throughout life. Amino acid mutations with similar chemical and physical properties are expected to have little effect on fitness. Highly dissimilar mutations will likely decrease stability and be eliminated quickly, while mutations that increase stability due to favorable interactions in the site-specific environment will fix quickly. Since the exchangeabilities in global matrices are calculated by averaging over many different sites, exchangeabilities are seen to be marginal probabilities in which all site-specific variation has been integrated out. In most common applications of substitution matrices, like phylogenetic analyses and homology detection, we lack detailed physical information at specific protein sites. Hence, in the absence of atomic-level site-specific information, global (i.e., marginal) exchangeability matrices should provide the maximum information possible about amino acid substitution behavior. After site-specific variation has been removed, the remaining information in a global exchangeability matrix quantifies the degree of similarity between amino acids in their inherent biological, chemical, and physical properties, as seen by selection in the context of folded proteins.[13] The fact that we can use site-specific physics and genetics to accurately predict empirical global substitution matrices helps explain why global matrices have been so successful in practice in phylogenetics, bioinformatics, homology detection, and ancestral sequence reconstruction.

4 | MATERIALS AND METHODS

4.1 | Protein dataset

We selected a non-redundant subset of 52 proteins structures from a curated set of high-quality protein structures.[28,29] All structural modeling was done with the Rosetta macromolecular modeling suite.[30] To ensure structural diversity, sequence redundancy was decreased so that no sequence shared more than 60% identical positions with any other sequence. Before analysis, each structure was adapted to the energy function using the *FastRelax* protocols as described by Nivon et al.[70] allowing cartesian space minimization.

4.2 | Prediction of the energetic effects of amino acid mutations

The $\Delta\Delta G$ prediction method is based on a modified version of the method presented by Park et al.,[33] but with a cutoff in the Lennard-Jones potential set to 6.0 Å. This $\Delta\Delta G$ method samples backbone degrees of freedom for the mutated and neighboring residues in the sequence and allows repacking of all side chains in energetic contact (>0.1 kcal/mol) with the mutated residue.

4.3 | Markov chain aggregation and averaging

To calculate the amino acid substitution rate matrix Q_{20}^{L} for site L from the codon substitution rate matrix Q_{64}^{L}, we first determine the frequency of each codon:[71,72]

$$\pi_i^L = \frac{\exp\left(4N_e\omega_i^L\right)}{\sum_i \exp\left(4N_e\omega_i^L\right)} \tag{8}$$

Next, we determine the rate between two amino acids (i,j) with codons $u\in i$ and $v\in j$ using the aggregation approach presented by Yang et al.:[73]

$$q_{ij}^L = \sum_{u\in i}\sum_{v\in j}\frac{\pi_v^L}{\pi_i^L}q_{vu}^L \tag{9}$$

where q_{vu}^L is the rate between codon v to u at site L. The flux between a pair of amino acids at a site L is:

$$\Phi_{ij}^{L} = \pi_i^L q_{ij}^L = \sum_{u \in i} \sum_{u \in j} \pi_v^L q_{vu}^L \tag{10}$$

The site-specific rate μ^L (the total flux) is:

$$\mu^L = \sum_i \sum_{j \neq i} \Phi_{ij}^L \tag{11}$$

We normalize the site-specific flux matrix, Φ_{20}^L, so that a unit of time corresponds to one expected amino acid substitution per site.

$$\Phi_{ij,\text{norm}}^L = \Phi_{ij}^L / \mu^L \tag{12}$$

To calculate the mean instantaneous rate for amino acid *i* to *j* we average over sites by normalizing by the equilibrium frequency of an amino acid *i* at that site:

$$q_{ij} = \frac{\langle \Phi_{ij,\text{norm}}^L \rangle_L}{\langle \pi_i \rangle_L} = \frac{\sum_L \Phi_{ij,\text{norm}}^L}{\sum_L \pi_i^L} \tag{13}$$

Sites with rates with $\mu < 10^{-10}$ were excluded from the analysis to avoid numerical errors. Next, we determine the exchangeability matrix $R_{20} = (r_{ij})$ as

$$r_{ij} = q_{ij} / \pi_j \tag{14}$$

From this, it can be seen that the exchangeability matrix is equivalent to the equilibrium flux matrix normalized by the appropriate amino acid equilibrium frequencies:[41]

$$r_{ij} = \frac{\Phi_{ij}}{\pi_i \pi_j} \tag{15}$$

4.4 | Model parameterization based on phylogenetic trees

Optimal values for the four free parameters of the model $\hat{\theta} = \{\Delta G_{\text{nat}}, N_e, \kappa, \rho\}$ were determined by maximizing the sum log-likelihood of phylogenetic trees for the 52 nonredundant protein alignments with a total of 8,907 sites:

$$\hat{\theta} = \text{argmax} \sum_{i=1}^{n} \log L(\text{MSA}_i | Q(\theta, \Delta\Delta G)) \tag{16}$$

The maximum likelihood trees (optimizing branch lengths, topology, equilibrium frequencies, and site rate variation parameter α) were determined with IQ-TREE[6] (model TMS + FO + G4, where TMS is an exchangeability matrix calculated from specific values of $\Delta G_{\text{nat}} = -5.91$, $N_e = 10^{3.8}, \kappa = 2.1, \rho = 0.1$). Note that because we use the IQ-TREE "FO" model option, rather than specifying the equilibrium frequencies predicted by our TMS model, this method optimizes over the TMS exchangeabilities alone. To speed the calculation the tree search was seeded with a tree inferred using LG. To find the parameterization that maximizes the sum log-likelihood, we performed a grid-search over the parameter space. Grid search optimization was performed over linearly spaced steps in the ranges $\Delta G_{\text{nat}} \in \left[-8.0 : -4.6 \frac{\text{kcal}}{\text{mol}}\right], \kappa \in [1.4 : 2.8]$ and over log-linearly spaced steps in the ranges $\rho \in [0.05 : 0.25]$, $N_e \in [10^3 : 10^5]$. The maximum likelihood TMS matrix is available in supplementary materials.

4.5 | Correlations between substitution matrices

The agreement between exchangeability matrices was quantified using a Pearson correlation coefficient calculated in log-space for the 190 exchangeability parameters.

ACKNOWLEDGMENTS

CN and IA was supported by a grant from the Swedish Research Council (2015-04203). DLT was supported by NIH grants R01GM096053 and R01GM132499.

CONFLICT OF INTEREST

The authors declare no competing financial interest.

DATA AVAILABILITY STATEMENT

Data, substitution matrices, and scripts for analysis are available here: https://github.com/Andre-lab/TMS

AUTHOR CONTRIBUTIONS

Christoffer Norn: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; software; visualization; writing-original draft; writing-review & editing. **Ingemar Andre:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; software; supervision; visualization; writing-original draft; writing-review & editing. **Douglas L. Theobald:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; software; supervision; visualization; writing-original draft; writing-review & editing.

ORCID

Christoffer Norn https://orcid.org/0000-0002-1450-4651
Ingemar André https://orcid.org/0000-0002-4753-8233
Douglas L. Theobald https://orcid.org/0000-0002-2695-8343

REFERENCES

1. Jacquier H, Birgy A, Le Nagard H, et al. Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci U S A. 2013;110:13067–13072.
2. Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc Natl Acad Sci U S A. 2012;109:16858–16863.
3. Bloom JD, Arnold FH. In the light of directed evolution: Pathways of adaptive protein evolution. Proc Natl Acad Sci U S A. 2009;106:9995–10000.
4. Wylie CS, Shakhnovich EI. A biophysical protein folding model accounts for most mutational fitness in viruses. Proc Natl Acad Sci U S A. 2011;108:9916–9921.
5. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–321.
6. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–1534.
7. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313.
8. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17:754–755.
9. Suchard MA, Redelings BD. BAli-Phy: Simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics. 2006;22: 2047–2048.
10. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004;21:1095–1109.
11. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 1992;8:275–282.
12. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 2001;18: 691–699.
13. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008;25:1307–1320.
14. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402.
15. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89:10915–10919.
16. Tomii K, Kanehisa M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 1996;9:27–36.
17. Venkatarajan M, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. J Mol Model. 2001;7: 445–453.
18. Atchley WR, Zhao J, Fernandes AD, Druke T. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A. 2005; 102:6395–6400.
19. Creixell P, Schoof EM, Tan CSH, Linding R. Mutational properties of amino acid residues: Implications for evolvability of phosphorylatable residues. Phil Trans R Soc B Biol Sci. 2012; 367:2584–2593.
20. Liberles DA, Teichmann SA, Bahar I, et al. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 2012;21:769–785.
21. Echave J, Wilke CO. Biophysical models of protein evolution: Understanding the patterns of evolutionary sequence divergence. Annu Rev Biophys. 2017;46:85–103.
22. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface. 2014;11:20140419.
23. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 2016;17:109–121.
24. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary stokes shift. Proc Natl Acad Sci U S A. 2012;109:E1352–E1359.
25. Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. Proc Natl Acad Sci U S A. 2015;112:E3226–E3235.
26. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol. 2006;2:598–605.
27. Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: Modeling site- specific residue frequencies. Mol Biol Evol. 1998;15:910–917.
28. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011;79:830–838.
29. Kumar MDS. ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006;34:D204–D206.
30. Leaver-fay A, Tyka M, Lewis SM, et al. ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–574.
31. Kimura M. Some problems of stochastic processes in genetics. Ann Math Stat. 1957;28:882–901.
32. Kimura M. On the probability of fixation of mutant gennes in a population. Genetics. 1962;47:713–719.
33. Park H, Bradley P, Greisen P, et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J Chem Theory Comput. 2016;12:6201–6212.
34. Golding B, Felsenstein J. A maximum likelihood approach to the detection of selection from a phylogeny. J Mol Evol. 1990;31: 511–523.
35. Ewens WJ. Mathematical population genetics. New York: Springer, 2004.
36. Topal MD, Fresco JR. Base pairing and fidelity in codon-anticodon interaction. Nature. 1976;264:289–293.
37. Topal MD, Fresco JR. Complementary base pairing and the origin of substitution mutations. Nature. 1976;263:285–289.
38. Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res. 2014;24: 1445–1454.
39. Reid TM, Loebt LA, Gottstein J. Tandem double CC-> TT mutations are produced by reactive oxygen species. Proc Natl Acad Sci U S A. 1993;90:3904–3907.
40. Kosiol C, Holmes I, Goldman N. An empirical codon model for protein sequence evolution. Mol Biol Evol. 2007;24:1464–1479.
41. Yang Z. Molecular evolution: A statistical approach. Oxford: Oxford University Press, 2014.

42. Efron B. Bootstrap methods: Another look at the jackknife. Ann Stat. 1979;7:1–26.
43. Ghosh K, Dill K. Cellular proteomes have broad distributions of protein stability. Biophys J. 2010;99:3996–4002.
44. Yu N, Jensen-Seaman MI, Chemnick L, Ryder O, Li WH. Nucleotide diversity in gorillas. Genetics. 2004;166:1375–1383.
45. Charlesworth J, Eyre-Walker A. The rate of adaptive evolution in enteric bacteria. Mol Biol Evol. 2006;23:1348–1356.
46. Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins. 2011;79:1396–1407.
47. Lee H, Popodi E, Tang H, Foster PL. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc Natl Acad Sci U S A. 2012;109:2774–2783.
48. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–120.
49. Schrider DR, Hourmozdi JN, Hahn MW. Pervasive multinucleotide mutational events in eukaryotes. Curr Biol. 2011;21:1051–1054.
50. Venkat A, Hahn MW, Thornton JW. Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nat Ecol Evol. 2018;2:1280–1288.
51. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental algorithms for scientific computing in python. Nat Methods. 2020;17:261–272.
52. Le VS, Dang CC, Le QS. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. BMC Evol Biol. 2017;17:1–13.
53. Jimenez-Morales D, Liang J. Pattern of amino acid substitutions in transmembrane domains of β-barrel membrane proteins for detecting remote homologs in bacteria and mitochondria. PLoS One. 2011;6:e26400.
54. Plotkin JB, Kudla G. Synonymous but not the same: The causes and consequences of codon bias. Nat Rev Genet. 2011;12:32–42.
55. Dang CC, Lefort V, Le VS, Le QS, Gascuel O. ReplacementMatrix: A web server for maximum-likelihood estimation of amino acid replacement rate matrices. Bioinformatics. 2011;27:2758–2760.
56. Spielman SJ, Shapiro B. Relative model fit does not predict topological accuracy in single-gene protein phylogenetics. Mol Biol Evol. 2020;37:2110–2123.
57. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol. 1992;9:678–687.
58. Rudner R, Karkas JD, Chargaff E. Separation of *B. subtilis* DNA into complementary strands, III. Direct analysis. Proc Natl Acad Sci U S A. 1968;60:921–922.
59. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol. 1994;39:306–314.
60. Serohijos AWR, Shakhnovich EI. Merging molecular mechanism and evolution: Theory and computation at the interface of biophysics and evolutionary population genetics. Curr Opin Struct Biol. 2014;26:84–91.
61. Bastolla U, Dehouck Y, Echave J. What evolution tells us about protein physics, and protein physics tells us about evolution. Curr Opin Struct Biol. 2017;42:59–66.
62. Parisi G, Echave J. Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol. 2001;18:750–756.
63. Parisi G, Echave J. The structurally constrained protein evolution model accounts for sequence patterns of the LBH superfamily. BMC Evol Biol. 2004;4:41.
64. Echave J. Evolutionary divergence of protein structure: The linearly forced elastic network model. Chem Phys Lett. 2008;457:413–416.
65. Fornasari MS, Gustavo P, Echave J. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. Mol Biol Evol. 2002;19:352–356.
66. Jimenez MJ, Arenas M, Bastolla U. Substitution rates predicted by stability-constrained models of protein evolution are not consistent with empirical data. Mol Biol Evol. 2018;35:743–755.
67. Santos MJJ, Arenas M, Bastolla U. Influence of mutation bias and hydrophobicity on the substitution rates and sequence entropies of protein evolution. PeerJ. 2018;6:e5549.
68. Arenas M, Sánchez-Cobos A, Bastolla U. Maximum-likelihood phylogenetic inference with selection on protein folding stability. Mol Biol Evol. 2015;32:2195–2207.
69. Parisi G, Echave J. Generality of the structurally constrained protein evolution model: Assessment on representatives of the four main fold classes. Gene. 2005;345:45–53.
70. Nivón L, Moretti R, Baker D. A pareto-optimal refinement method for protein design scaffolds. Curr Sci. 2011;101:1435–1439.
71. Iwasa Y. Free fitness that always increases in evolution. J Theor Biol. 1988;135:265–281.
72. Sella G, Hirsh A. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A. 2005;102:12690–12693.
73. Yang Z, Nielsen R, Hasegawa M. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol. 1998;15:1600–1611.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Norn C, André I, Theobald DL. A thermodynamic model of protein structure evolution explains empirical amino acid substitution matrices. Protein Science. 2021;30:2057–68. https://doi.org/10.1002/pro.4155