Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence

Julian Echave[1,2] and Claus O. Wilke[3]

[1] Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, 1650 San Martín, Buenos Aires, Argentina; email: julian.echave@unsam.edu.ar

[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

[3] Department of Integrative Biology, The University of Texas at Austin, Texas 78712; email: wilke@austin.utexas.edu

Annu. Rev. Biophys. 2017. 46:85–103

First published online as a Review in Advance on March 15, 2017

The *Annual Review of Biophysics* is online at biophys.annualreviews.org

https://doi.org/10.1146/annurev-biophys-070816-033819

Keywords

fitness landscape, protein folding, protein misfolding, protein–protein interaction, evolutionary rate

Abstract

For decades, rates of protein evolution have been interpreted in terms of the vague concept of functional importance. Slowly evolving proteins or sites within proteins were assumed to be more functionally important and thus subject to stronger selection pressure. More recently, biophysical models of protein evolution, which combine evolutionary theory with protein biophysics, have completely revolutionized our view of the forces that shape sequence divergence. Slowly evolving proteins have been found to evolve slowly because of selection against toxic misfolding and misinteractions, linking their rate of evolution primarily to their abundance. Similarly, most slowly evolving sites in proteins are not directly involved in function, but mutating these sites has a large impact on protein structure and stability. In this article, we review the studies in the emerging field of biophysical protein evolution that have shaped our current understanding of sequence divergence patterns. We also propose future research directions to develop this nascent field.

Contents

1. INTRODUCTION

The success of theoretical physics in the nineteenth and twentieth centuries was driven primarily by the development of fundamental and elegant mathematical theories that accurately describe and predict physical processes from first principles. In biology, by contrast, development of fundamental theory has been lacking. Biology is, to a much greater extent, a science of individual special cases and historical contingencies, and it has shown surprising resilience against physicists' desire to describe it with theories that are both simple and general. However, the one area in which biology may be most amenable to the development of fundamental theory is the field of protein evolution. Evolutionary biology is probably the most mathematical branch of biology, and sophisticated mathematical machinery exists to describe how populations accumulate mutations and evolve over time. Similarly, protein biophysics is governed by equilibrium and nonequilibrium thermodynamics, both of which are well-developed areas of theoretical physics. In recent years, several groups have made efforts to combine evolutionary theory and protein biophysics, with the ultimate goal of developing a general, mechanistic theory of protein evolution.

In this article, we review both fundamental concepts and recent advances in the field of protein evolution. Because this field is so extensive, no single review can cover it exhaustively. Recent reviews have covered empirical approaches to studying protein evolution, specifically reviewing the causes of rate variation among proteins (90) and among sites within proteins (21). Other reviews have emphasized how protein biophysics shapes fitness landscapes (14, 70, 74). To complement these works, we describe how theoretical population genetics and protein biophysics have been combined into predictive models of evolutionary rate variation among and within proteins. We emphasize the development of theory and attempt to provide a comprehensive and self-contained description of the key concepts, starting from first principles.

2. PROTEIN EVOLUTION ON FITNESS LANDSCAPES

The evolution of a protein's amino acid sequence (its genotype) can be represented as a trajectory in sequence space. Because evolutionary dynamics depends on fitness, protein evolution can be pictured as following paths on a fitness landscape (84). This section provides a basic introduction to the theory of evolution on fitness landscapes.

2.1. Population Dynamics on Fitness Landscapes

Finite populations of genotypes evolve on the fitness landscape under genetic drift, selection, and mutation. The population's evolution can be modeled by randomly drawing the genes of the offspring population from the pool of parent genotypes. Such stochastic sampling is called genetic drift. To model selection, one draws individuals with weights proportional to their fitness. The drawn individuals may be mutated to introduce new genotypes.

A frequently used sampling model is the Moran process, in which, for a population of haploid individuals of effective population size N_{e}, at each time step a random individual is removed and another individual, drawn randomly with a probability proportional to its fitness, is duplicated. Another popular model is the Wright-Fisher process of discrete nonoverlapping generations, in which the offspring generation is obtained by N_{e} random draws with replacement from the parent generation, with probability proportional to fitness.

In finite populations, stochastic fluctuations of allele frequencies result in genes becoming lost from the gene pool. If no new alleles are introduced, eventually all alleles but one will be lost and one allele will become fixed. If the time between the introduction of new mutations is much larger than the timescale of fixation, then, in most cases, all individuals of the population will share the same allele (the population is monophyletic). Given a parent allele i, when a new mutant allele j is introduced, it eventually either is lost or becomes fixed. For a haploid population evolving under the Moran process, the fixation probability (p_{fix}) is (44, 66)

$$p_{\mathrm{fix}}(s_{ji}, N_{\mathrm{e}}) = \frac{1 - e^{-s_{ji}}}{1 - e^{-N_{\mathrm{e}} s_{ji}}}, \qquad 1.$$

where s_{ji} is the selection coefficient (66). The selection coefficient is defined as

$$s_{ji} = \ln(f_j/f_j) = \ln(f_j) - \ln(f_i), \qquad 2.$$

where f_j and f_i are the fitness values of alleles j and i, respectively. The selection coefficient measures the selective advantage of the mutant relative to the parent: s_{ji} is zero for neutral mutations and positive (negative) for advantageous (deleterious) mutations. For the haploid Wright-Fisher process, p_{fix} can be obtained by replacing s_{ji} with $2s_{ji}$ in Equation 1. In certain conditions, fixation probabilities for diploid populations may be obtained by replacing N_{e} with $2N_{\mathrm{e}}$ (44).

2.2. Origin-Fixation Models

For long timescales, evolution is often studied using origin-fixation models (46). Evolution is modeled using a Markov process that is completely specified by a rate matrix **Q** with off-diagonal

Protein genotype: the amino acid sequence of a given protein

Sequence space: possible amino acid sequences connected by single amino acid point mutations

Fitness: expected number of offspring produced by an individual

Fitness landscape: map from sequence space to fitness

Effective population size (N_{e}): number of individuals that may contribute offspring to the next generation

Allele: variant of a gene sequence generated through mutation

Fixed: describes an allele that is present in all members of the population

Fixation probability (p_{fix}): probability of an allele becoming fixed

Selection coefficient (s_{ji}): fitness of allele j relative to allele i

Origin-fixation model: model of evolution that considers only the most abundant genotype in the population; this genotype is assumed to change over time through the origination of individual mutations and their subsequent fixation in the population

$(j \neq i)$ elements:

$$Q_{ji} = N_e \mu_{ji} p_{\text{fix}}(s_{ji}, N_e), \tag{3.}$$

Substitution rate: the rate at which mutations fix in the population per unit of time

where i is the parent genotype, j is the mutant allele, μ_{ji} is the $i \to j$ mutation rate per individual, and p_{fix} is the fixation probability of j. $Q_{j \neq i}$ represents the rate of $i \to j$ transitions; the diagonal elements satisfy $Q_{ii} = -\sum_{j \neq i} Q_{ji}$.

In general, the time-dependent state of the system can be described by a (column) probability vector $\boldsymbol{\pi}(t) = (\pi_1(t), \pi_2(t), \ldots)^T$, where $\boldsymbol{\pi}_i(t)$ is the probability that at time t the fixed genotype is i. The evolutionary dynamics obeys

$$\frac{\mathrm{d}\boldsymbol{\pi}(t)}{\mathrm{d}t} = \mathbf{Q}\boldsymbol{\pi}(t). \tag{4.}$$

Eventually, $\boldsymbol{\pi}(t)$ approaches an equilibrium distribution of genotypes $\boldsymbol{\pi}^*$. For origin-fixation models given by Equation 3, with fixation probability given by Equation 1 and a mutational process that satisfies $\mu_{ji} m_i^* = \mu_{ij} m_j^*$ (following 66), it can be found that

$$\pi_i^* = \frac{m_i^* e^{-\nu h_i}}{Z}, \tag{5.}$$

where m_i^* is the equilibrium probability of i under mutation only, $h_i = -\ln(f_i)$, $Z = \sum_k m_k^* e^{-\nu h_k}$, and $\nu = N_e - 1$ for the haploid Moran process. [For the Wright-Fisher process, $\nu = 2(N_e - 1)$, and the diploid cases can be obtained by replacing N_e with $2N_e$.]

Notably, Equation 5 is mathematically similar to the Boltzmann distribution of statistical physics, with m_i^* playing the role of the degeneracy of state i, $h_i = -\ln(f_i)$ playing the role of energy, and ν playing the role of the inverse temperature β. This mathematical analogy is useful to import methods and concepts from statistical physics into molecular evolution, but it can also lead to confusion.

2.3. Substitution Rate, Entropy, and Number of States

For further reference, we define in this section some quantities that reflect the degree of evolutionary constraint on protein sequences or on individual protein sites. Consider a distribution of genotypes $\boldsymbol{\pi}$. One measure of constraint is the substitution rate K, which is the mean number of amino acid substitutions per unit of time:

$$K = \sum_i \sum_{j \neq i} Q_{ji} \pi_i = -\sum_i Q_{ii} \pi_i. \tag{6.}$$

Another measure of constraint is sequence entropy, which quantifies how dispersed the distribution of genotypes is:

$$S = -\sum_i \pi_i \ln \pi_i. \tag{7.}$$

Often, a more convenient measure is

$$\Omega = e^S, \tag{8.}$$

which is a number between 1 and the total number of states; it represents the effective number of different states allowed.

3. MOLECULAR, BIOPHYSICS-BASED FITNESS LANDSCAPES

The key challenge in developing a biophysical description of protein evolution is determining how fitness f depends on the amino acid sequence. A mutation will impact the fitness of the organism in relationship to how the mutation (*a*) perturbs the activity of the protein and (*b*) creates undesired side effects on the organism. For example, to be functional, the protein may need, for example, to be able to fold rapidly into a given specific native structure or bind some substrate. Similarly, to avoid creating side effects, a protein must not interact with off-target binding partners or create toxic metabolic products. These traits (structure, folding, stability, binding affinity, enzymatic activity, etc.) constitute the molecular phenotype of the protein, i.e., those protein properties that affect organism fitness.

3.1. Thermodynamics of Protein Folding

According to statistical thermodynamics, at constant pressure and temperature, the probability P_i to observe a protein in a given conformation i is given by the Boltzmann distribution,

$$P_i = \frac{e^{-\beta E_i}}{Z}, \qquad 9.$$

where E_i is the potential energy of conformation i and $\beta = (k_B T)^{-1}$, with k_B as the Boltzmann constant and T as the physical temperature of the system. The quantity $Z = \sum_j e^{-\beta E_i}$ is called the conformational partition function. The energy E_i itself is calculated from some potential energy function, e.g., a pairwise potential summed over all pairs of atoms in the conformation i.

In practice, we are usually interested in the ground state of the protein. The ground state will typically correspond to a set of conformations with low energy; we also refer to it as the folded state. The probability of observing the folded state, P_{folded}, relative to the probability of observing any other state, P_{unfolded}, also follows from a Boltzmann distribution:

$$\frac{P_{\text{folded}}}{P_{\text{unfolded}}} = \frac{e^{-\beta G_{\text{folded}}}}{e^{-\beta G_{\text{unfolded}}}} = e^{-\beta \Delta G}. \qquad 10.$$

The quantity $\Delta G = G_{\text{folded}} - G_{\text{unfolded}}$ is the free energy of folding, and it measures the stability of the folded protein. The more negative is ΔG, the more stable is the protein. The stability ΔG of a protein will typically change if the protein is mutated. In many modeling scenarios, we are interested in the magnitude of this change, commonly denoted by $\Delta\Delta G$.

Equation 9, which describes the probability of observing a specific protein in a particular conformation, looks similar to Equation 5, which describes the probability of observing the evolutionary trajectory in a given genotype state. However, these two equations describe two entirely distinct processes, and one does not follow from the other. To relate Equation 9 to biological evolution, we must introduce an explicit mapping to fitness, as is discussed in the next section.

3.2. Fitness Landscapes Based on Protein Fold Stability

Most biophysical models of protein evolution map ΔG to fitness. Several examples are illustrated in **Figure 1**. For example, if fitness is assumed to be proportional to the probability of folding, then we obtain the soft-threshold model shown in **Figure 1*a*** (82):

$$f_{\text{folding}} \propto P_{\text{folded}} = \frac{1}{1 + e^{\beta \Delta G}}. \qquad 11.$$

A simplified version of the soft-threshold model uses a hard stability threshold (**Figure 1*b***) (6, 7, 20). This model can be considered the limit of the soft-threshold model for very large β (low physical temperature). Alternatively, if protein stability and protein activity trade off, then we expect fitness to be maximal at some intermediate optimum stability (**Figure 1*c***) (14, 71). Finally, if fitness declines due to protein toxicity with increasing amounts of misfolded protein in the cell,

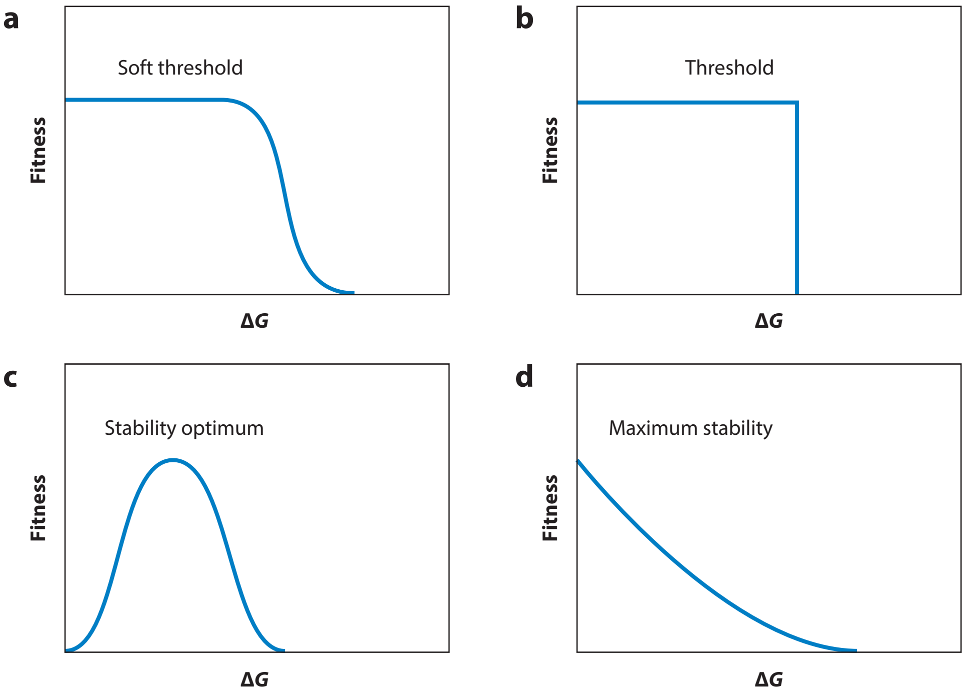


Figure 1

Common stability-based fitness models. (*a*) The soft-threshold model assumes that sufficiently stable proteins (low ΔG) all have the same high fitness, whereas unstable proteins (high ΔG) have fitness of zero. Between these two extremes, there is a smooth, sigmoidal transition region. (*b*) The threshold model can be considered as a limiting case of the soft-threshold model. It assumes that all proteins with sufficient stability ($\Delta G <$ some threshold) have the same fitness, whereas all other proteins are completely inviable. (*c*) The stability-optimum model assumes that there is an ideal stability for maximum fitness, and both more and less stable proteins will be less fit. (*d*) The maximum-stability model assumes that fitness is higher the more stable is the protein.

then we obtain the maximum-stability model of **Figure 1*d*** (18, 68, 86):

$$f_{\text{toxicity}} = \exp[-c\,(\text{number of misfolded proteins})] = \exp\left[-c\sum_i \frac{A_i}{1+e^{-\beta\Delta G_i}}\right], \qquad 12.$$

where A_i is the abundance of protein i, c is a measure of toxicity per misfolded protein, and the sum runs over all genes i in the organism's genome.

There are several modeling approaches that are commonly used to calculate ΔG for a given protein sequence. First, we can evaluate Equation 9 directly for models in which we can exhaustively enumerate all possible energy states E_j (74). Second, we can estimate Z from random energy models (22, 73) or ignore it entirely (42). These approaches assume that individual mutations have a negligible effect on Z. Alternatively, a more realistic approach is to estimate Z from a limited number of decoy structures (82). Some authors also use detailed, all-atom off-lattice models (2, 15, 20, 37, 71), but to date this approach has seen limited use in evolutionary modeling due to its high computational cost. Finally, some models operate entirely in ΔG space, keeping track of how ΔG changes due to the stability effects $\Delta\Delta G$ of successive mutations (68).

3.3. Fitness Landscapes Based on Protein Function

Stability-based models, such as the soft-threshold or stability-optimum models, implicitly assume that stability determines function; either all folded proteins or only the proteins with the right optimum stability are functional. However, the exact way in which a protein's function contributes to organism fitness is highly specific, and thus, in principle, we would have to derive function-based fitness landscapes from the biochemical properties and interactions of each individual protein. However, there are at least two broad classes of models that can be used to describe a generic, functional protein.

First, many authors have modeled function through protein–protein or protein–ligand binding (9, 31, 37, 49, 59, 83, 85). For example, fitness can be assumed to be proportional to the binding energy of the protein–ligand complex or the probability of finding the protein in the bound state. In all cases, the probability of binding is measured through a change in free energy $\Delta G_{\text{binding}}$, which is calculated according to the same principles of thermodynamics outlined in Section 3.1 (for further details, see, e.g., 49, 83).

Second, for enzymes, we can consider the flux through metabolic pathways (69). The fitness contribution due to flux through a linear pathway can be written as (69)

$$f_{\text{flux}} = a \Big/ \sum_i \frac{\epsilon_i}{C_i}, \qquad 13.$$

where a is the abundance of the input metabolite, ϵ_i is the enzymatic activity of protein i in the pathway, and C_i is the number of functional copies of protein i. C_i can, in turn, be written as the protein expression level A_i times the probability that protein i is folded (Equation 11), yielding the expression

$$f_{\text{flux}} = a \Big/ \sum_i \frac{\epsilon_i(1 + e^{\beta \Delta G_i})}{A_i}. \qquad 14.$$

This equation describes how f_{flux} depends on protein stability. It does not, however, explicitly model how the proteins' enzymatic activities ϵ_i depend on mutations.

Models of ϵ_i for actual proteins are still lacking. However, enzyme-like constraints have been explicitly considered in studies with simplified protein-like polymers. A model by Sasai and coworkers (62, 63, 88) assumed that fitness increases monotonically as the conformation of a few active site residues converges to a predefined active conformation. Similarly, a recent study (51) used a threshold-like fitness function that constrains the structure of a ligand and a few sites that bind it.

4. MODELING AMONG-PROTEIN RATE VARIATION

In the following sections, we discuss models that combine molecular fitness landscapes with models of sequence evolution to calculate the rates at which proteins accumulate substitutions. These models attempt to explain why some proteins evolve faster or slower than others. The models typically consider only the average rate of evolution across the entire protein and neglect variation among sites. Variation among sites may be implicitly present in the model, but this variation is not the focus of the analysis and is averaged out in the final modeling predictions.

4.1. More Robust Protein Structures Evolve More Rapidly

First, we can ask how rapidly a protein evolves if it is only required to fold stably into its native conformation (the threshold or soft-threshold models of **Figure 1**). Under this model, we would expect that proteins whose structures are, on average, more tolerant to substitution (i.e., more mutationally robust) should evolve more rapidly (5). However, in practice, it is not entirely clear how the robustness of a protein structure should be measured. A random-energy model predicts that a structure's robustness is determined by its contact matrix (22). Based on this prediction, the average number of residue–residue contacts, the principal eigenvalue of the contact matrix, and the size of the protein core (as measured, e.g., by the fraction of buried residues) have all been proposed as simple measures of mutational robustness (5). Indeed, there is evidence that proteins with higher contact density or with a larger core evolve more rapidly (5, 23, 26, 27, 64). Alternatively, it should be possible to calculate a protein's robustness to mutations from its $\Delta\Delta G$ distribution (8, 80). However, this approach has been less successful at explaining observed rate variation among proteins (23).

Protein neutral network: the network of protein sequences that fold stably into the native conformation of the protein

Importantly, the mutational robustness of a protein changes as the protein mutates, and the evolutionary rate is determined by the average robustness of all sequences fixed during the evolutionary trajectory. This average robustness will depend on intrinsic properties of the structure (some structures are, on average, inherently more robust to mutations than others), as well as extrinsic factors such as the mutation rate μ and population size N_e. If the product of the mutation rate and population size is large, i.e., $\mu N_e \gg 1$, then populations concentrate in the most connected (i.e., most mutationally robust) regions of the protein neutral network (10, 78). Therefore, as we increase μ, N_e, or both, the mean substitution rate increases as well (7, 79).

4.2. Proteins Evolve to Be Robust Against Misfolding

In many different systems, one of the best predictors of among-protein evolutionary rates is the cellular abundance of proteins. The more highly expressed a protein is, the slower it evolves. Because of both the strength and the universality of this effect, it requires an explanation that applies in any organism, is consistent with protein biophysics, and is capable of reproducing the key patterns of sequence variation observed in natural proteins. The first explanation that met these criteria was the idea that organisms experience strong selection pressure against protein misfolding, specifically against misfolding due to errors introduced during translation (17).

Models of selection against protein misfolding generally assume that fitness declines exponentially with an increasing number of misfolded proteins, as in Equation 12 (18, 81). This number can be calculated, for example, via a lattice-protein model that simulates error-prone translation (18). The lattice-protein model produces a power-law relationship between evolutionary rate and abundance, as observed in natural sequences, and it also reproduces observed codon-level selection pressures, such as an increase in the fraction of preferred codons for more highly expressed genes. Although this model only considers misfolding induced by translation errors, proteins may also misfold spontaneously without a proximate cause (19). A variant of the lattice-protein simulation allowing for spontaneous misfolding (86) has demonstrated that the source of protein misfolding, whether spontaneous or induced by translation errors, is not critical to the presence of the key evolutionary adaptations that protect against misfolding.

However, the lattice-protein simulations used to date have considered only a single protein fold; variation among folds was not considered (18, 86). Alternatively, Lobkovsky et al. (42) estimated protein misfolding in a coarse-grained off-lattice model and calculated the resulting rate distribution across folds. They found that their calculated rate distribution matched the observed distribution of rates in both bacteria and mammals. This study was limited, however, by the fact that it did not consider variation in protein abundance.

As an alternative to using an explicit protein folding model, either on or off lattice, one can also describe the protein misfolding probability in terms of the $\Delta\Delta G$ distribution (67, 68). This approach has the benefit of yielding analytic insight into the relationship among substitution rate, protein abundance, and population size. For the simplest such model, which assumes that the $\Delta\Delta G$ distribution is constant and independent of ΔG (2, 39, 76), the mean substitution rate is independent of protein abundance (68). To recover a rate–abundance anticorrelation, mutations must become increasingly destabilizing as protein stability increases (67, 68). This relationship between protein stability and the mutational effect of mutations is realized in all explicit folding models.

4.3. Proteins Evolve to Avoid Misinteraction

Protein misfolding is not the only mechanism that creates a fitness cost proportional to protein abundance. Even if a protein is folded correctly, it may interact with off-target partners, and such interactions will likely be deleterious. Because the probability of off-target interactions increases in proportion to the abundance of a given protein, it is reasonable to assume that the fitness cost due to such protein misinteractions similarly increases with abundance.

The importance of protein misinteractions was first pointed out by Zhang et al. (89), who estimated the amount of protein misinteraction in a cell and proposed an upper limit to gene expression and proteome size. Subsequently, Yang et al. (85) used a lattice-protein simulation similar to the simulation models employed to study misfolding (18, 86) but using a fitness function based on on-target and off-target interactions. In their model, fitness increases linearly with the proportion of correct interactions and decreases exponentially in proportion to the number of nonspecific interactions. This model produces an evolutionary rate–expression level anticorrelation, as do the models based on misfolding. More importantly, it provides an explanation for the observation that there is a larger decrease in the evolutionary rate at solvent-exposed residues than at buried residues with increasing expression level (27, 64, 85).

Importantly, no model has combined both misfolding and misinteractions. Experiments support the existence of a fitness cost from both protein misfolding (28) and widespread off-target interactions among proteins (13). Thus, the challenge for future work will be to determine the relative importance of these and other mechanisms.

4.4. Proteins Evolve to Carry Out Their Function

Although models based on protein stability, protein misfolding, and/or protein misinteraction reproduce major patterns of evolutionary rate variation in natural proteins, they neglect a key component that must have some effect on rate variation: protein function (90). For example, a negative correlation between rate and abundance can also be explained by the expression-cost hypothesis, which argues that the fitness reduction due to a deleterious mutation affecting protein function should scale with protein abundance (11, 30; although see also discussion in 90).

In general, although there are well-developed models of how fitness may depend on protein function (Section 3.3), few authors have used these models to calculate among-protein rate variation. One exception is the misinteraction model of Yang et al. (85), which also contains a fitness term for correct, functional interactions. However, both this fitness term and the one for misinteractions are ad hoc, not derived from first principles.

An alternative, promising direction was recently taken by Serohijos & Shakhnovich (69), who modeled the metabolic flux through a pathway and wrote fitness as the difference between the fitness components f_{flux}, due to flux (Equation 14), and f_{toxicity}, due to toxicity caused by protein misfolding (Equation 12). However, the ways in which these different fitness components interact in their effect on evolutionary rate have not yet been studied in detail.

5. MODELING AMONG-SITE RATE VARIATION

Mean evolutionary rates vary not only among proteins but also among sites within proteins. More generally, not only rates but also the whole pattern of sequence divergence are site-dependent. In this section, we describe modeling efforts aimed at explaining among-site variation of sequence divergence patterns.

5.1. Evolution at the Site Level

Sequence divergence patterns at the site level can be obtained from simulations. For instance, suppose that we set up a model combining some evolutionary model from Section 2 with some molecular fitness landscape from Section 3 and simulate divergent evolution by running several independent trajectories that start at the same ancestral protein sequence. Then, for any given site l, we may estimate its amino acid probability distribution $\boldsymbol{\pi}^l$ as the fraction of trajectories that have each amino acid at the focus site. Given $\boldsymbol{\pi}^l$, we may calculate the site-specific sequence entropy S^l and number of observed amino acids Ω^l (using Equations 7 and 8, respectively). Counting amino acid substitutions along the trajectories, we can calculate the site-specific substitution rate K^l and even the site-specific 20 × 20 matrix of rates of amino acid substitutions $\mathbf{Q}^l$.

5.2. Site-Specific and Pooled Analyses

A simple but often overlooked issue that arises when analyzing site-specific variation is that sequence patterns of individual sites are generally different from those of sites pooled together into groups. For instance, the variability of individual sites increases monotonically with relative solvent accessibility (RSA) (75), as shown in **Figure 2*a***. In contrast, the variability of groups of sites of

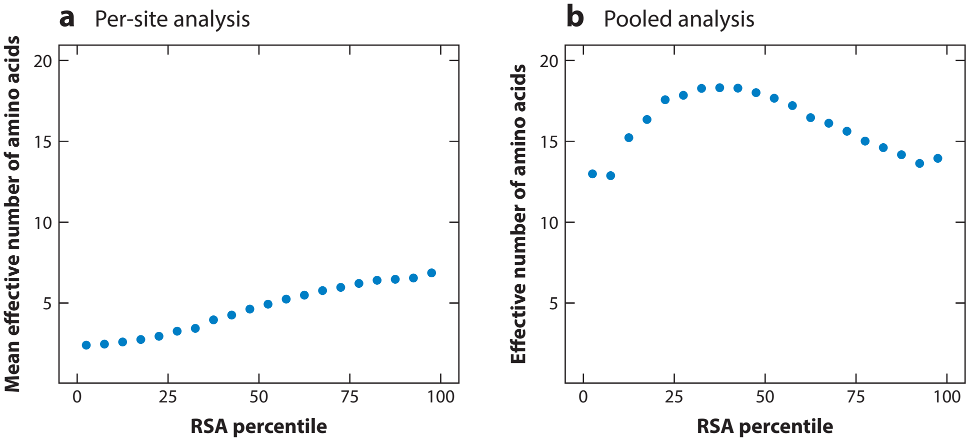


Figure 2

The relationship between the effective number of amino acids Ω and the relative solvent accessibility (RSA) of sites depends on whether amino acid distributions are pooled across similar sites. (*a*) Without pooling, an Ω value is calculated at each site, and Ω is then averaged over all sites within the same RSA percentile. This procedure yields relatively low Ω values, between 2 and 10, and a nearly linear increase of Ω with increasing RSA. (*b*) Alternatively, amino acid distributions are pooled among all sites within the same RSA percentile, and a single Ω value is then calculated for each pooled distribution. This procedure yields much higher Ω values (>10) and shows a maximum at intermediate RSA values. Protein sequences and structural data for this figure were taken from Reference 34. We selected 266 distinct enzymes for which each available alignment consisted of at least 400 sequences.

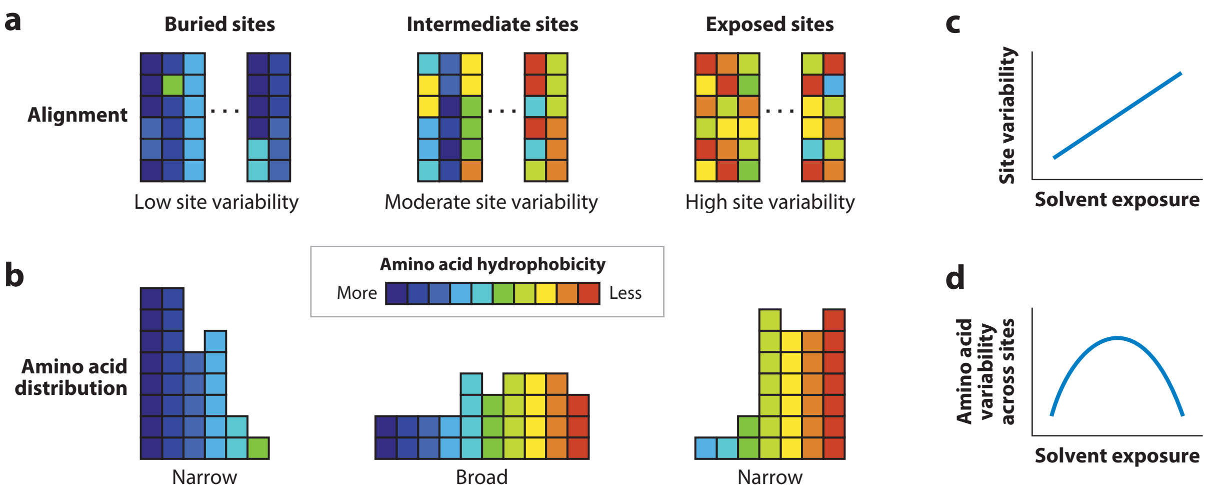


Figure 3

Conceptual explanation for the results shown in **Figure 2**. (*a*) At a per-site level, buried sites are the most conserved, intermediate sites are moderately conserved, and exposed sites are the most variable. (*b*) However, for both buried and exposed sites, the observed types of amino acids are limited to mostly hydrophobic or mostly polar residues, respectively, whereas for intermediate sites, both hydrophobic and polar residues are seen across many sites (58). As a consequence, even though the site variability increases approximately linearly with solvent exposure (*c*), the amino acid variability across sites has a maximum at intermediate solvent exposure (*d*).

similar RSA is much larger and has a peak at some intermediate RSA value, as shown in **Figure 2*b***. The difference between per-site analysis and pooled analysis is explained schematically in **Figure 3**.

Models that are assessed using pooled analysis may fail to reproduce site-level patterns. For instance, a mean-field (MF) model by Bastolla and coworkers (3, 4, 57) predicts that sites with intermediate solvent accessibility are the most variable, with more-exposed sites having conservation levels similar to buried sites. This prediction is satisfied for pooled sites (**Figure 2*b***) but is violated for individual sites (**Figure 2*a***) (58).

5.3. Selection for Rapid Folding

It is reasonable to assume that, because proteins need to fold rapidly, sites critical for folding kinetics should be particularly conserved. Early studies aimed at finding evidence of special conservation of protein sites involved in the folding nucleus (a few sites that are key for rapid folding) (48, 72). Starting at random sequences, sequence trajectories were run selecting for maximization of kinetic stability (the energy gap between the target conformation and the lowest-energy misfolded conformation). By comparing site-specific entropies of simulated and natural sequences, the cited studies found that, whereas folding nucleus sites were more conserved than average, many non-nucleus sites were similarly conserved due to stability constraints alone. Special conservation of the folding nucleus was later disproved by other authors (40, 77). This finding is consistent with kinetic stability correlating strongly with thermodynamic stability (50).

5.4. Selection on Native-State Stability

Because most proteins need to fold stably into their native conformation to be functional and/or to avoid the toxic effects of misfolded and unfolded conformations, it is reasonable to ask whether stability-based models explain among-site evolutionary variation.

Some authors have studied site-specific patterns modeling evolution using sequence design algorithms (15, 35, 52). Sequence design can be seen as a model of evolution with selection for stability (e.g., **Figure 1*d***) because sequences are designed to maximize the stability of a target structure. In general, the agreement between designed and natural sequences is moderate (35, 52). Specifically, predicted site-specific entropies increase too slowly with increasing RSA, and designed proteins have too few hydrophobic residues and too many hydrogen bonds in the core (35). The mismatch between designed and natural sequences may have several causes, such as errors in the energy function or the fitness function. More importantly, a major caveat of sequence design is that it simulates convergent evolution starting from random sequences, rather than divergent evolution starting from a common ancestral sequence. Although the convergent and divergent processes should lead to the same equilibrium distributions at infinite time, natural sequences have diverged for a finite amount of time and finite-time sequence patterns may look different from the infinite-time steady-state picture.

Divergence from a common ancestor that has a fitness function with a stability optimum (**Figure 1*c***) is the basis of the structurally constrained protein evolution (SCPE) model (53, 55). This model simulates divergent evolution from a protein of known structure and sequence, which is assumed to be at the fitness peak. Mutations are introduced according to a DNA-level process and selected against departure from the optimum. Site-specific amino acid distributions, entropies, rates, and substitution matrices are calculated as in Section 5.1 (24). For a highly regular structure for which sites can be grouped into classes, the model predicts quantitatively class-specific amino acid distributions and entropies (53, 54). More generally, predicted site-specific substitution matrices fit homologous sequence families better than do site-independent models (55). Interestingly, a recent study (36) showed that for many ligand-binding proteins with more than one stable conformation, the SCPE site-specific matrices obtained from the ligand-binding conformation fit sequence data better than matrices obtained from the inactive conformations.

In contrast with the simulation approach described in the previous paragraphs, a recent study (20) derived substitution rates theoretically for a stability-threshold model (**Figure 1*b***). This work built on a model by Bloom & Glassman (6) to derive an analytic relationship between mutational $\Delta\Delta G$ values and site-specific substitution rates. $\Delta\Delta G$ values for all possible single-point mutations were calculated for a large set of monomeric enzymes using FoldX (32, 65), and predicted rates were found to correlate well with empirical rates (20). This result shows that native-state stability constraints can explain part of the observed rate variation among protein sites.

5.5. Selection on Active Structure Stability

Because proteins fluctuate around the minimum energy conformation, the native state consists of an ensemble of conformations. The stability-based models of Section 5.4 are based on the total probability of the native-state ensemble (e.g., Equation 11). By contrast, a recently developed stress model assumes that fitness depends on the probability of specific active conformation (33, 43). Mutations shift the conformational ensemble, modifying the probability of reaching the active conformation. Basic statistical physics was used to derive an analytic relationship between site-specific rates and the impact of mutations on the stability of the active conformation, $\Delta\Delta G^*_{ji} = \Delta G^*_j - \Delta G^*_i$.

In the published studies, $\Delta\Delta G^*$ was calculated using perturbed elastic network models (33, 43). Amino acids were represented by nodes connected by harmonic springs, and mutations were modeled as random perturbations of the spring lengths of the mutated site. For a large set of monomeric enzymes, the stress model explained the observed linear decrease of site-specific substitution rates with increasing local packing density and their nonlinear increase with site flexibility: The average destabilization of the active conformation is proportional to the local packing density, and therefore, tightly packed sites evolve more slowly than loosely packed ones (26, 33, 87). Moreover, the model predicted the side-chain packing density to be the main structural determinant of a site's rate of evolution (43).

5.6. Selection on Protein Function

The models described in Sections 5.1– 5.5 do not explicitly consider specific functional constraints. Therefore, it is often the case that some sites are observed to be much more conserved than predicted by such models. A few biophysical modeling studies found that including explicit functional constraints may affect site-specific sequence divergence.

Among the forms of functional constraints that may impact site-specific sequence divergence are constraints induced by protein–ligand interactions. Only a few modeling studies have considered the effect of ligand-binding constraints on site-specific sequence divergence (31, 41, 59). Selection for binding affects the site-specific patterns of the binding site. Importantly, selection against binding nontarget ligands must be taken into account to reproduce (qualitatively) the conservation patterns observed in natural sequences (31).

Protein–protein interactions may also affect divergence patterns. For multimeric proteins, SCPE model simulations that include monomer–monomer interactions improve the fit of site-specific substitution matrices for the sites directly involved in the interactions (25). However, the overall conservation of binding surfaces is low. A more recent study (37) that included protein–protein interactions in simulations of sequence divergence found that binding constraints have surprisingly little influence on the rate of divergence of the sites involved in the binding surface. The binding surface diverges nearly as quickly when selection for continued binding is imposed as when this selection pressure is removed. This finding is consistent with divergence patterns observed at binding interfaces in enzymes (34).

In contrast to the low conservation observed in protein–protein binding surfaces, active sites in enzymes are highly conserved and impose long-range constraints on the evolution of most protein sites (34). This conservation is probably caused by the requirement that residues involved in catalysis adopt very precise conformations in the substrate binding pocket. We lack realistic models of functional selection for enzyme evolution. However, the effect of constraining only a few sites on protein evolution has been studied using simple protein-like polymers. Sasai and coworkers (62, 63, 88) simulated the evolution of protein-like polymers with a fitness function that constrains the structure of only a few active residues. Remarkably, they found that such functional constraints alone were enough for sequences to evolve into stable, ordered, protein-like molecules (62, 88) and that the rate of evolution of a site increases with its flexibility (63). Recently, Nelson & Grishin (51) used a similar protein-like model to study more thoroughly the effect of enzyme-like functional constraints on sequence divergence. Starting from ordered structures, they simulated divergent evolution by introducing mutations, insertions, and deletions, constraining only the structure of a small binding pocket whose sites were not mutated. These simulations reproduce many recent findings: (*a*) Site-specific rates increase linearly with solvent accessibility (26), (*b*) rates decrease linearly with local packing density (33, 87), (*c*) packing density is a better predictor of rates than solvent accessibility (43, 87), and (*d*) rates increase nonlinearly with flexibility (33). Importantly, the model also predicts a linear increase of site-specific rates with distance to the active site, in agreement with recent findings in a large dataset of over 500 enzymes (34).

5.7. Epistasis and Time Dependence of Site-Specific Patterns

Thus far, we have considered site-specific divergence patterns averaged over time. However, site-specific patterns may change along an evolutionary trajectory because the selection coefficient of a given mutation may depend on the specific sequence background in which it occurs. This phenomenon, called epistasis, has been studied extensively in empirical and experimental research, but only recently has it been investigated using biophysical models of protein evolution (2, 56, 71).

Epistatic effects have been clearly established in an elegant study by Shah et al. (71). They simulated sequence evolution on two fitness landscapes, a (Gaussian) stability optimum and a (semi-Gaussian) soft threshold (see **Figure 1**). For both cases, they found that fixed mutations are nearly neutral and contingent on previous substitutions: The same mutation would be too deleterious to fix if previous permissive substitutions had not occurred. In addition, substitutions become entrenched over time: A mutation that was nearly neutral when fixed becomes increasingly deleterious to revert as subsequent substitutions accumulate.

Entrenchment had previously been demonstrated in a stability-based simulation study by Pollock et al. (56). They found that, as a result of entrenchment, site-specific amino acid distributions are time dependent. However, Bloom and coworkers (2) have called into question the prevalence and magnitude of time-dependent epistatic shifts in site-specific distributions. Using simulations analogous to those of Pollock et al., they found that, although site-specific amino acid propensities do change, they diverge very slowly, so that they are mostly conserved among homologous proteins over realistic timescales.

The degree of variation of site-specific amino acid distributions with time depends on whether deleterious mutations are compensated by reversion or entrenchment. Although entrenchment makes reversion increasingly unlikely (56, 71), reversions may be the most common way of compensating deleterious substitutions (2, 16). A step toward understanding the reversion–entrenchment trade-off is the recent finding that, for a site involved in epistatic interactions, the probability of reverting a deleterious substitution is large just after fixation and then decreases rapidly with time (45).

That epistasis plays a role in protein evolution is widely accepted. However, the prevalence and magnitude of epistatic time-dependent propensity shifts and their implications for phylogenetic modeling of homologs with similar structures and functions are still under debate (16, 29).

6. CONCLUSIONS AND OUTLOOK

Theoretical population genetics studies evolution on fitness landscapes. Protein biophysics aims to understand protein function from the fundamental laws of physics. Insofar as fitness depends on protein function, protein biophysics should be able to derive the fitness landscape from first principles. In recent years, researchers have combined evolutionary models with fitness functions based on biophysical properties to derive biophysical models of protein evolution, giving rise to the emergent field of biophysical protein evolution.

Arguably, the main goal of a theory of protein evolution is to explain and predict patterns of amino acid sequence divergence. Biophysical evolution studies have already provided major advances in our understanding of why evolutionary rates vary among proteins, among sites within proteins, and over time. Regarding rate variation among proteins, the observed rate–expression anticorrelation may be due to several different biophysical mechanisms, including mistranslation-induced misfolding (18), spontaneous misfolding (68, 86), and misinteraction (85). Regarding variation among sites, the increase of rates with solvent accessibility and flexibility and the decrease of rates with local packing density may be explained in terms of folding stability and/or stability of active conformations (21). Finally, variation of rates over time follows from epistatic interactions inherent in selection on stability (2, 56, 71).

Ideally, the fitness landscape should be derived from the biophysical principles that govern protein behavior. However, the fitness functions employed to date remain mostly ad hoc. For example, a fitness function that posits an optimum stability corresponding to maximal fitness represents a biophysical model but does not provide any insight into whether or why fitness may be maximal at that stability value. By contrast, a soft threshold derived from $f \sim P_{\text{folded}}$ can be considered more principled because it represents the assumption that activity is proportional to the abundance of proteins in the native state. Therefore, one broad research direction is the development of more principled biophysical models of fitness.

We see several other major avenues for future work. First, the existing models must be studied more exhaustively across a larger range of scenarios. For example, most of the models that are critical to understanding variation of rates among proteins, such as models of mistranslation, misfolding, and misinteraction, have not been used to study site-specific patterns, although these processes will likely affect site-specific patterns as well. Second, we must develop better models of functional constraints. For instance, the reasons why enzyme active sites impose long-range evolutionary constraints on most protein sites are not yet fully understood. Third, the ways in which epistasis among protein sites affects the substitution process at individual sites remain poorly understood and deserve further study. Finally, biophysical models of protein evolution can be integrated directly into phylogenetic methods of sequence analysis, alleviating the need to first calculate summary statistics of the evolutionary process (such as rates or entropies) and then correlate these statistics with model predictions. Although some progress has been made in this direction (1, 12, 38, 47, 60, 61, 64), most of the models proposed to date have been computationally cumbersome and ultimately not yet useful.

SUMMARY POINTS

1. Biophysical models of evolution combine population genetics models of evolution on fitness landscapes with biophysical models of the fitness landscape.
2. Most current biophysical models posit ad hoc fitness functions that depend on biophysical traits that are assumed to affect protein function or toxicity.
3. In spite of their mostly ad hoc nature, biophysical modeling studies have significantly advanced our understanding of why the rate of evolution varies among proteins, among sites within proteins, and over time.

FUTURE ISSUES

1. Less ad hoc fitness functions need to be derived from first principles of molecular biophysics.
2. Fitness models that consider the joint effects of distinct biophysical mechanisms need to be developed.
3. Better models of functional constraints, especially for enzymes, need to be developed.
4. The ways in which epistasis shapes evolutionary divergence at individual protein sites need to be clarified.

5. Biophysical protein-evolution models need to be directly integrated into phylogenetic methods of sequence analysis.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported in part by National Institutes of Health grants R01 GM088344 and R01 AI120560, National Science Foundation cooperative agreement DBI-0939454 (BEACON Center), and Army Research Office grant W911NF-12-1-0390.

LITERATURE CITED

1. Arenas M, Sanchez-Cobos A, Bastolla U. 2015. Maximum-likelihood phylogenetic inference with selection on protein folding stability. *Mol. Biol. Evol.* 32:2195–207
2. Ashenberg O, Gong LI, Bloom JD. 2013. Mutational effects on stability are largely conserved during protein evolution. *PNAS* 110:21071–76
3. Bastolla U, Porto M, Roman HE, Vendruscolo M. 2004. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins* 58:22–30
4. Bastolla U, Porto M, Roman HE, Vendruscolo M. 2006. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. *BMC Evol. Biol.* 6:43
5. Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* 23:1751–61
6. Bloom JD, Glassman MJ. 2009. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLOS Comput. Biol.* 5:e1000349
7. Bloom JD, Raval A, Wilke CO. 2007. Thermodynamics of neutral protein evolution. *Genetics* 175:255–66
8. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. *PNAS* 102:606–11
9. Bloom JD, Wilke CO, Arnold FH, Adami C. 2004. Stability and the evolvability of function in a model protein. *Biophys. J.* 86:2758–64
10. Bornberg-Bauer E, Chan HS. 1999. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *PNAS* 96:10689–94
11. Cherry JL. 2010. Expression level, evolutionary rate, and the cost of expression. *Genome Biol. Evol.* 2:757–69
12. Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. *Mol. Biol. Evol.* 24:1769–82
13. Deeds EJ, Ashenberg O, Shakhnovich EI. 2006. A simple physical model for scaling in protein-protein interaction networks. *PNAS* 103:311–16
14. DePristo MA, Weinreich DM, Hartl DL. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* 6:678–87
15. Ding F, Dokholyan NV. 2006. Emergence of protein fold families through rational design. *PLOS Comput. Biol.* 2:725–33
16. Doud MB, Ashenberg O, Bloom JD. 2015. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.* 32:2944–60
17. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *PNAS* 102:14338–43

18. Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–52

19. Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* 10:715–24

20. Echave J, Jackson EL, Wilke CO. 2015. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Phys. Biol.* 12:025002

21. Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* 17:109–21

22. England JL, Shakhnovich EI. 2003. Structural determinant of protein designability. *Phys. Rev. Lett.* 90:218101

23. Faure G, Koonin EV. 2015. Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Phys. Biol.* 12:035001

24. Fornasari MS, Parisi G, Echave J. 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol. Biol. Evol.* 19:352–56

25. Fornasari MS, Parisi G, Echave J. 2007. Quaternary structure constraints on evolutionary sequence divergence. *Mol. Biol. Evol.* 24:349–51

26. Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* 26:2387–95

27. Franzosa EA, Xia Y. 2012. Independent effects of protein core size and expression on residue-level structure-evolution relationships. *PLOS ONE* 7:e46602

28. Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, Drummond DA. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *PNAS* 108:680–85

29. Goldstein RA, Pollock DD. 2016. The tangled bank of amino acids. *Protein Sci.* 25:1354–62

30. Gout JF, Kahn D, Duret L. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLOS Genet.* 6:e1000944

31. Grahnen JA, Nandakumar P, Kubelka J, Liberles DA. 2011. Biophysical and structural considerations for protein sequence evolution. *BMC Evol. Biol.* 11:361

32. Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320:369–87

33. Huang TT, Marcos ML, Hwang JK, Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evol. Biol.* 14:78

34. Jack BR, Meyer AG, Echave J, Wilke CO. 2016. Functional sites induce long-range evolutionary constraints in enzymes. *PLOS Biol.* 14:e1002452

35. Jackson EL, Ollikainen N, Covert AW, Kortemme T, Wilke CO. 2013. Amino-acid site variability among natural and designed proteins. *PeerJ* 1:e211

36. Juritz EI, Palopoli N, Fornasari MS, Fernandez-Alberti S, Parisi GD. 2012. Protein conformational diversity modulates protein divergence. *Mol. Biol. Evol.* 30:79–87

37. Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. 2015. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* 348:921–25

38. Kleinman CL, Rodrigue N, Lartillot N, Philippe H. 2010. Statistical potentials for improved structurally constrained evolutionary models. *Mol. Biol. Evol.* 27:1546–60

39. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, et al. 2006. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* 34:D204–6

40. Larson SM, Ruczinski I, Davidson AR, Baker D, Plaxco KW. 2002. Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J. Mol. Biol.* 316:225–33

41. Liberles DA, Tisdell MD, Grahnen JA. 2011. Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. *Proc. Biol. Sci.* 278:1930–35

42. Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. *PNAS* 107:2983–88

43. Marcos ML, Echave J. 2015. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ* 3:e911

44. McCandlish DM, Epstein CL, Plotkin JB. 2015. Formal properties of the probability of fixation: identities, inequalities and approximations. *Theor. Popul. Biol.* 99:98–113

45. McCandlish DM, Shah P, Plotkin JB. 2016. Epistasis and the dynamics of reversion in molecular evolution. *Genetics* 203:1335–51

46. McCandlish DM, Stoltzfus A. 2014. Modeling evolution using the probability of fixation: history and implications. *Q. Rev. Biol.* 89:225–52

47. Meyer AG, Wilke CO. 2013. Integrating sequence variation and protein structure to identify sites under selection. *Mol. Biol. Evol.* 30:36–44

48. Michnick SW, Shakhnovich E. 1998. A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des.* 3:239–51

49. Miller DW, Dill KA. 1997. Ligand binding to proteins: the binding landscape model. *Protein Sci.* 6:2166–79

50. Naganathan AN, Muñoz V. 2010. Insights into protein folding mechanisms from large scale analysis of mutational effects. *PNAS* 107:8611–16

51. Nelson ED, Grishin NV. 2016. Evolution of off-lattice model proteins under ligand binding constraints. *Phys. Rev. E* 94:022410

52. Ollikainen N, Kortemme T. 2013. Computational protein design quantifies structural constraints on amino acid covariation. *PLOS Comput. Biol.* 9:e1003313

53. Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.* 18:750–56

54. Parisi G, Echave J. 2004. The structurally constrained protein evolution model accounts for sequence patterns of the LβH superfamily. *BMC Evol. Biol.* 4:41

55. Parisi G, Echave J. 2005. Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. *Gene* 345:45–53

56. Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary Stokes shift. *PNAS* 109:E1352–59

57. Porto M, Roman HE, Vendruscolo M, Bastolla U. 2005. Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. *Mol. Biol. Evol.* 22:630–38

58. Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* 188:479–88

59. Rastogi S, Reuter N, Liberles DA. 2006. Evaluation of models for the evolution of protein sequences and functions under structural constraint. *Biophys. Chem.* 124:134–44

60. Robinson D, Jones D, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.* 20:1692–704

61. Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347:207–17

62. Saito S, Sasai M, Yomo T. 1997. Evolution of the folding ability of proteins through functional selection. *PNAS* 94:11324–28

63. Sasaki TN, Sasai M. 2002. Correlation between the conformation space and the sequence space of peptide chain. *J. Biol. Phys.* 28:483–92

64. Scherrer MP, Meyer AG, Wilke CO. 2012. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol. Biol.* 12:179

65. Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. 2005. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *PNAS* 102:10147–52

66. Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *PNAS* 102:9541–46

67. Serohijos AWR, Lee SYR, Shakhnovich EI. 2013. Highly abundant proteins favor more stable 3D structures in yeast. *Biophys. J.* 104:L1–3

68. Serohijos AWR, Rimas Z, Shakhnovich EI. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* 2:249–56

69. Serohijos AWR, Shakhnovich EI. 2014. Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. *Mol. Biol. Evol.* 31:165–76

70. Serohijos AWR, Shakhnovich EI. 2014. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. *Curr. Opin. Struct. Biol.* 26:84–91

71. Shah P, McCandlish DM, Plotkin JB. 2015. Contingency and entrenchment in protein evolution under purifying selection. *PNAS* 112:E3226–35

72. Shakhnovich EI, Abkevich V, Ptitsyn O. 1996. Conserved residues and the mechanism of protein folding. *Nature* 379:96–98

73. Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. *PNAS* 90:7195–99

74. Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface* 11:20140419

75. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites of residues in proteins. *PLOS ONE* 8:e80635

76. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. 2007. The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* 369:1318–32

77. Tseng YY, Liang J. 2004. Are residues in a protein folding nucleus evolutionarily conserved? *J. Mol. Biol.* 335:869–80

78. van Nimwegen E, Crutchfield JP, Huynen M. 1999. Neutral evolution of mutational robustness. *PNAS* 96:9716–20

79. Wilke CO. 2004. Molecular clock in neutral protein evolution. *BMC Genet.* 5:25

80. Wilke CO, Bloom JD, Drummond DA, Raval A. 2005. Predicting the tolerance of proteins to random amino acid substitution. *Biophys. J.* 89:3714–20

81. Wilke CO, Drummond DA. 2006. Population genetics of translational robustness. *Genetics* 173:473–81

82. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *PLOS Comput. Biol.* 2:e69

83. Williams PD, Pollock DD, Goldstein RA. 2001. Evolution of functionality in lattice proteins. *J. Mol. Graph. Model.* 19:150–56

84. Wright S. 1988. Surfaces of selective value revisited. *Am. Nat.* 131:115–23

85. Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *PNAS* 109:E831–40

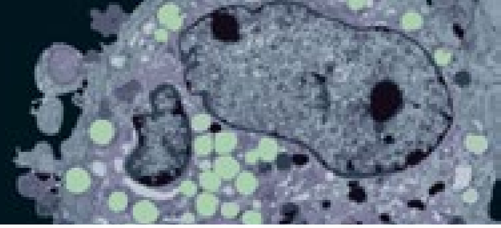
86. Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol. Syst. Biol.* 6:421

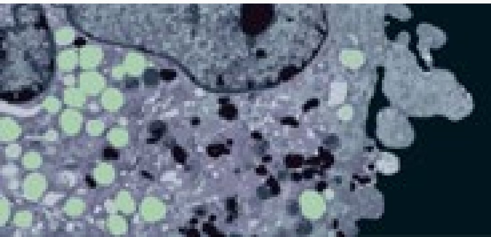
87. Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. 2014. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol. Biol. Evol.* 31:135–39

88. Yomo T, Saito S, Sasai M. 1999. Gradual development of protein-like global structures through functional selection. *Nat. Struct. Biol.* 6:743–46

89. Zhang J, Maslov S, Shakhnovich EI. 2008. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Mol. Syst. Biol.* 4:210

90. Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* 16:409–20





Annual Review of Biophysics
Volume 46, 2017

Contents