Population Size Dependence of Fitness Effect Distribution and Substitution Rate Probed by Biophysical Model of Protein Thermostability

Richard A. Goldstein*

Division of Infection and Immunity, University College London, United Kingdom

***Corresponding author**: E-mail: r.goldstein@ucl.ac.uk.

Accepted: July 19, 2013

Abstract

The predicted effect of effective population size on the distribution of fitness effects and substitution rate is critically dependent on the relationship between sequence and fitness. This highlights the importance of using models that are informed by the molecular biology, biochemistry, and biophysics of the evolving systems. We describe a computational model based on fundamental aspects of biophysics, the requirement for (most) proteins to be thermodynamically stable. Using this model, we find that differences in population size have minimal impact on the distribution of population-scaled fitness effects, as well as on the rate of molecular evolution. This is because larger populations result in selection for more stable proteins that are less affected by mutations. This reduction in the magnitude of the fitness effects almost exactly cancels the greater selective pressure resulting from the larger population size. Conversely, changes in the population size in either direction cause transient increases in the substitution rate. As differences in population size often correspond to changes in population size, this makes comparisons of substitution rates in different lineages difficult to interpret.

Key words: nearly neutral theory, effective population size, substitution rate, protein stability, epistasis, population bottleneck.

Introduction

Novel mutations that appear in a genome can be advantageous, increasing the resulting organism's fitness, deleterious, reducing the fitness, or effectively neutral, having such a small effect on fitness that the fate of the mutation in the population is dominated by random drift. The relative fraction of these three different types of mutations, and the form of the overall distribution of fitness effects ($s = w' - w/w$, where w and w' are the fitness of the wild type and mutant, respectively) caused by such mutations, has been a topic of interest and debate (Bustamante 2005; Eyre-Walker and Keightley 2007). Characterizing this distribution is essential for understanding the nature of genetic variation, including polymorphisms that may cause or influence diseases, as well as characterizing the evolutionary dynamics.

Larger population sizes result in increased magnitude of the selective pressure acting on mutations of a given value of s. The fitness effect and the effective population size N_e generally appear as a product in many evolutionary and genetic calculations, so often equations reference the population-scaled fitness effect $S = 4N_e s$ ($S = 2N_e s$ for haploid organisms). For instance, $P_{\mathrm{Fix}}(s)$, the probability of a new mutation with fitness effect s being fixed in an otherwise homogeneous diploid population, relative to the probability of fixation of a neutral mutation P^0_{Fix}, is given by (Fisher 1930; Kimura 1957, 1962; Crow and Kimura 1970)

$$\frac{P_{\mathrm{Fix}}(s)}{P^0_{\mathrm{Fix}}} = \frac{\left(\frac{1-e^{-2s}}{1-e^{-4N_e s}}\right)}{\left(\frac{1}{2N_e}\right)} \approx \frac{4N_e s}{1-e^{-4N_e s}} = \frac{S}{1-e^{-S}} \qquad (1)$$

where the approximation is valid for small s.

The effective population size affects the substitution rate differently depending up the relative number of advantageous, deleterious, and neutral mutations (Gillespie 1999). If there are a substantial number of adaptive mutations, whose probability of fixation is less dependent on the population size, the substitution rate would be higher in larger populations reflecting the greater number of mutations that arise. Conversely, if the mutations are either neutral or so deleterious as to have negligible fixation probability, as suggested by the neutral theory of molecular evolution (Kimura 1968, 1983), then the substitution rate would be relatively independent of population size, with the greater number of mutations cancelling the decreased probability of fixation. The nearly neutral theory of evolution (Ohta 1973, 1992; Kimura 1983) emphasizes the role of slightly deleterious mutations, whose probability of acceptance is smaller in larger populations. This latter theory predicts that smaller populations should evolve faster.

Observations of the population dependence of the substitution rate are difficult. One approach is to examine the dependence of the substitution rate on the rate of recombination. In regions of the genome with low recombination rates, mutations in linked genes compete for fixation (Hill and Robertson 1966), which has a similar effect as a lower effective population size. High recombination rates reduce this effect, so that regions of the genome that recombine rapidly are characterized by an increased effective population size. It has been observed that regions in *Drosophila* genome with high recombination rates evolve slower than regions of low recombination rates, consistent with the predictions of the nearly neutral theory (Larracuente et al. 2008; Arguello et al. 2010; Campos et al. 2012). Other studies have looked at the difference in the evolution of genes on sex chromosomes, observing higher rates of nonsynonymous substitutions on the nonrecombining chromosome (Wyckoff et al. 2002; Berlin and Ellegren 2006). These types of analysis assume that the mutation process in these two different types of regions, as well as the properties of the encoded proteins (e.g., expression levels, structure, and function), are not systematically dissimilar in ways that affect the substitution rate. There are, for instance, correlations between recombination rate, GC content, mutation rate, and rate of biased gene conversion that remain to be elucidated (Hardison et al. 2003; Duret 2006; Duret and Arndt 2008). Sex chromosomes might also be subject to specific adaptive selection that cannot be easily distinguished from reduced selection.

A more direct approach is to examine how the substitution rate differs in different lineages. For instance, substitution rates have been compared in primates and rodents (Wu and Li 1985; Ohta 1995; Weinreich 2001), although such comparisons are compromised by differences in, for example, generation time, cell division rate, metabolic rate, mating behavior, ecological niche, and DNA repair mechanisms (Bromham et al. 1996; Bromham 2011). Faster evolution has been observed in endosymbiotic bacteria and fungi compared with their free-living relatives (Moran 1996; Woolfit and Bromham 2003). Endosymbiotic bacteria and fungi would have their effective population size reduced by the lower population sizes of their host, and would also be expected to undergo population bottlenecks when relatively few endosymbionts are transmitted to progeny, reducing the intra-host variation; this process is generally modelled as reducing the effective population size (Rispe and Moran 2000). The faster evolution of the endosymbionts is again consistent with the nearly neutral theory, although other biological or ecological characteristics of endosymbionts might complicate the analysis. Comparisons have been made of the rates of evolution of island and mainland populations, with the island populations again having a smaller population size due to the population bottleneck that occurs during colonialization as well as due to habitat restriction; some studies have concluded that the smaller populations evolve faster (Johnson and Seger 2001; Woolfit and Bromham 2005), although other studies have reached different or more nuanced conclusions (Charlesworth and Eyre-Walker 2007; Wright et al. 2009).

Theoretical models of these effects are often based on simple models of the fitness landscape, such that the distribution of fitness effects is constant (Ohta 1977) or that the fitness of the mutant alleles has a fixed distribution (Kingman 1978). These models can break down if the population is far from a fitness optimum due to mutation–selection balance, where the preponderance of deleterious mutations is balanced by the greater fixation probabilities of advantageous mutations (Hartl et al. 1985; Cherry 1998; Wylie and Shakhnovich 2011; Charlesworth 2013). Mutation–selection balance cannot be achieved if the distribution of selective coefficients is independent of fitness; a stable equilibrium requires that there be an increased tendency toward acceptance of deleterious mutations as the fitness increases (Cherry 1998). In the case of a fitness function that plateaus as the fitness increases, increasing the population size would result in a higher equilibrium fitness, which can reduce the fitness impact of mutations, resulting in a narrower distribution of fitness effects. Under some conditions, this contraction of the distribution in s can exactly cancel the explicit population size dependence of S, so that the distribution of population-scaled fitness effects [$\rho_S(S)$], and therefore the substitution rate, is independent of population size, even in the nearly neutral model (Cherry 1998; Charlesworth 2013).

The evolutionary process involves modifications of interacting biological macromolecules. By creating evolutionary models that explicitly include the properties of these evolving biomolecules, we can develop more realistic models of the evolutionary process, better understand how the evolutionary dynamics depends on biological context, and improve our understandings of how the properties of these biological molecules arose. To fulfill these ambitions, we need to create computational models that capture the salient aspects of the biology while still being computationally tractable.

It has been noted that much of the selection pressure on coding regions involves maintaining an adequate degree of thermodynamic stability for the resulting expressed proteins (Wang and Moult 2001; Zeldovich et al. 2007; Drummond and Wilke 2008; Serohijos et al. 2012). This has led to studies investigating how these genetic regions would evolve where the fitness corresponds to a simple function of stability, such as the fraction of proteins that would be folded at equilibrium (Williams et al. 2006; Chen and Shakhnovich 2009; Goldstein 2011; Wylie and Shakhnovich 2011; Pollock et al. 2012). We investigate the distribution of selective effects generated by such a simple model, including its dependence on the effective population size. Our results correspond roughly to the nearly neutral theory. We find that the distribution of population-scaled fitness effect [$\rho_S(S)$] is essentially independent of the effective population size, suggesting that evolutionary dynamics, at least of regions of protein-coding genes where selection for stability dominates, should be similarly independent of population size. We observe, however, a strong dependence of the distribution of fitness effects, as well as overall substitution rate, on fluctuations in population size. This result can explain why differences in substitution rates have been observed in different lineages, and indicate that the effect of population bottlenecks on substitution rates cannot be modeled through an appropriate modification of the effective population size.

Materials and Methods

Protein Model

We consider a genome containing a 900-nucleotide gene, coding for a 300-residue protein, as described in earlier publications (Williams et al. 2006; Goldstein 2011; Pollock et al. 2012). We use a simple fitness function based on protein stability, where the fitness w of a protein is equal to the probability that the protein is folded at equilibrium P_{Folded}, given by

$$w \equiv P_{\mathrm{Folded}} = \frac{1}{1+e^{\Delta G/kT}} \qquad (2)$$

where ΔG is the difference in free energy between the folded and unfolded states, T is the temperature, and k is Boltzmann's constant. (Note that more negative values of ΔG correspond to higher stability.)

The free energy $G(S, C_k)$ of a protein with sequence $S = \{a_1, a_2, a_3 \ldots a_{300}\}$ in any given conformation C_k is computed by summing the contact energies of all of the pairs of residues which are in contact in that conformation, where we use the contact energies determined by Miyazawa and Jernigan (1985) based on frequencies of contacts in known protein structures; residues are in contact if their C_β atoms (C_α for glycine) are closer than 7 Å to each other. We consider the native state of the protein to be the conformation of the purple acid phosphatase (PDB 1QHW; Lindqvist et al. 1999); the free energy for a given sequence in this particular native state is designated $G_{\mathrm{NS}}(S)$. We assume that the distribution of free energies for the large ensemble of N_{U} unfolded states can be represented as a Gaussian distribution with mean $\bar{G}(S)$ and variance $\sigma(S)^2$. We estimate $\bar{G}(S)$ and $\sigma(S)^2$ by calculating the free energy of the sequence in a set of 55 alternative structures. The free energy difference between the folded and unfolded states is then given by

$$\Delta G(S) = G_{\mathrm{NS}}(S) + \frac{\sigma(S)^2 - 2kT\overline{G}(S)}{2kT} + kT \ln N_{\mathrm{U}} \qquad (3)$$

N_{U} is set to 10^{160}. T is set to 20 °C. The probability of folding, and thus the fitness, is then calculated using equation (2).

Evolutionary Model

We initialize a nucleic acid sequence to a set of 300 random codons (excluding stop codons). The codons are translated into a protein sequence using the standard genetic code, and the free energy of folding (and organismal fitness) calculated as described earlier. We simulate evolutionary dynamics where we assume that the mutation rate is slow relative to the fixation time, so that population variation can be ignored. We calculate the rate of all 3 × 900 possible single nucleotide substitutions, equal to the rate of mutation (using a K80 nucleotide substitution model [Kimura 1980] with a transition-transversion ratio of 2.0) times the probability of fixation of the mutation, calculated by computing the free energy of folding of the mutant and using equation (1). (Mutations resulting in stop codons are considered lethal.) We estimate the time to the next substitution by drawing from an exponential distribution with decay rate equal to the sum of all of the individual substitution rates, and choose a mutation to accept with probability proportional to its substitution rate. The protein sequence evolves with increasing stability (decreasing ΔG) until the point of mutation–selection balance, where there is no further long-term change of stability. The simulation is then extended to an evolutionary interval of 10 nucleotide substitutions expected per nucleotide location for neutral substitutions. Only the data subsequent to the establishment of mutation–selection balance are used in the subsequent analysis. These simulations are repeated 100 times with $N_{\mathrm{e}} = 10^4$, $N_{\mathrm{e}} = 10^6$, and $N_{\mathrm{e}} = 10^8$.

At each time point of the simulation, we calculate the effect of every possible single nucleotide mutation, and use all of these mutations to calculate the distribution of population-scaled fitness effects [$\rho_S(S)$] as well as the instantaneous substitution rate, represented by the ratio of nonsynonymous to synonymous substitution rates (dN/dS).

Explorations of Alternative Models

How sensitive are the results to a particular model? Two aspects of this model might be particularly relevant: 1) the specific relationship between fitness and stability; 2) the epistasis between various locations in the protein in calculating the fitness. We examine these aspects sequentially.

The relationship between protein stability and organismal fitness is still unclear and is possibly complicated (Bershtein et al. 2012). In particular, there are indications that avoiding aggregation may be more important than the concentration of the folded state (Chen and Dokholyan 2008; Zhang et al. 2008; Johnson and Hummer 2011; Levy et al. 2012; Yang et al. 2012). A linear relationship has been observed between fitness cost and fraction of aggregated proteins (Geiler-Samerotte et al. 2011); such an effect would not greatly change the model discussed earlier, as it still results in a linear relationship between fraction of folded protein and fitness, albeit with a different slope. To see how sensitive the results are to variations in the fitness function, we consider a different fitness function representing a fitness penalty for aggregation, which we model using a quadratic function of the amount of unfolded protein

$$w \equiv 1 - \xi(1 - P_{\mathrm{Folded}})^2 \tag{4}$$

where $\xi = 3 \times 10^3$ is chosen so that $dw/d\Delta G$ at $\Delta G = 9\,\mathrm{kcal/mol}$ is roughly similar to that in the original model represented by equation (2). Thirty simulations are made for each of the three values of N_e using this fitness function.

Removing Epistasis

The contribution of every residue to the fitness depends on the amino acids at every other location in the protein. This is due to two different aspects of the model. First, the energetics are based on contact potentials, which are a function of pairs of amino acids which are in contact in the native or alternative structures. Second, the fitness is a nonlinear function of the free energy of folding, as indicated by equation (2). In particular, an amino acid substitution at one location in the protein will cause a change in the protein's stability, but the effect of this change on the protein fitness will depend on the prior stability, which depends on the amino acids found in all other locations. This can be seen if we represent s as a function of the initial stability and change in stability (Wylie and Shakhnovich 2011)

$$s = \frac{\frac{1}{1+e^{(\Delta G+\Delta\Delta G)/kT}} - \frac{1}{1+e^{\Delta G/kT}}}{\frac{1}{1+e^{\Delta G/kT}}} \approx e^{\Delta G/kT}\left(1 - e^{\Delta\Delta G/kT}\right). \tag{5}$$

To remove these sources of epistasis, we construct a model where the fitness contribution of every amino acid at each location is fixed. We first compute the fitness w_{Ref} of the reference sequence. We then calculate the change in fitness for every single amino acid change, to generate $\Delta w_l(a') = w(\{a_1, a_2, \ldots a'_l, \ldots a_{300}\}) - w_{\mathrm{Ref}}$, where $w(\{a_1, a_2, \ldots a'_l, \ldots a_{300}\})$ is the fitness of an amino acid sequence differing from the reference sequence by the single replacement of a'_l for a_l. The fitness of any arbitrary sequence is represented as

$$w(\{a'_1, a'_2, \ldots a'_l, \ldots a'_{300}\}) = \sum_l \Delta w_l(a'_l) + \Delta w_{\mathrm{Ref}}. \tag{6}$$

The fitness effect $s_l(a' \rightarrow a'')$ corresponding to a mutation from amino acid a' to a'' at position l is given by

$$s_l(a' \rightarrow a'') = \frac{\Delta w_l(a'') - \Delta w_l(a')}{w_l(a')} \approx \Delta w_l(a'') - \Delta w_l(a'), \tag{7}$$

where we have taken advantage of the fact that the fitnesses during the simulation are all close to unity.

For calculating the resulting distribution of population-scaled fitness effects, we take advantage of the simplicity of the model to calculate the distributions by summing over all possible mutations from all possible codons at each location, weighted by the equilibrium probability of the original codon and the mutation, using the approaches described in Tamuri et al. (2012). We averaged over the results obtained with 10 different reference sequences.

Variations in Effective Population Size: Bottlenecks

We also perform simulations with the original model (eq. 2) where the effective population size fluctuates between 10^6 and 10^4, representing periodic population bottlenecks, with equal amounts of evolutionary time spent at each population level. The period of the oscillations in units of evolutionary time vary between 0.001 and 1.0 expected neutral substitutions per location, with 10 simulations performed for each period.

Results

After an initial period, the free energy of folding of the proteins reached values of approximately $\Delta G \sim -7$ for $N_e = 10^4$ to $-12\ \mathrm{kcal\,mol^{-1}}$ for $N_e = 10^8$. This degree of stability is roughly similar to that observed in real proteins. It is important to note that this stability is far from optimum; we can use hill-climbing algorithms to find sequences with stabilities in the order of -118 kcal/mol (Goldstein 2011). This marginal stability is also observed with natural proteins, which can be modified to have higher stabilities while retaining native-like activities (Serrano et al. 1993; Giver et al. 1998; Van den Burg et al. 1998; Zhao and Arnold 1999; Korkegian et al. 2005). The stability of these modelled proteins represents mutation–selection balance, where the greater number of destabilizing mutations is compensated by the higher acceptance rate for stabilizing mutations (Goldstein 2011).

Figure 1 shows the distribution of population-scaled fitness effects of nonsynonymous mutations for three different population sizes varying over four orders of magnitude. As can be

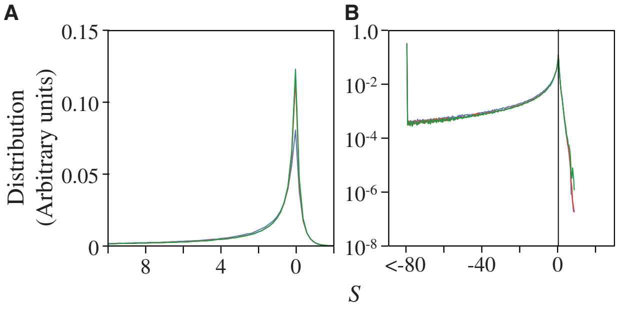

Fig. 1.—Distribution of population-scaled fitness effects for nonsynonymous mutations when fitness is proportional to the fraction of proteins folded at equilibrium, calculated using equation (2), for $N_e = 10^4$ (green), $N_e = 10^6$ (blue), and $N_e = 10^8$ (red), on linear (*A*) and log (*B*) scales.

seen, the distributions are extremely similar. All of the distributions of deleterious mutations are strongly leptokurtic, fitting an inverted Gamma distribution with shape parameter $\alpha = 0.08$. Approximately 25% of mutations are effectively neutral ($-1 < S < 1$), approximately 25% are mildly deleterious ($-10 < S < -1$), and 50% are strongly deleterious ($S < -10$). Although the distribution corresponds to the near-neutral theory, the substitution rate is nearly independent of population size, in agreement with the predictions of Cherry (1998): the average ratio of nonsynonymous to synonymous substitution rates ($\langle dN/dS \rangle$) only changes minimally, from 0.350 for $N_e = 10^4$ to 0.338 for $N_e = 10^8$.

Explorations of Alternative Models

To see how the results vary with the fitness function, we perform simulations using an alternative model based on aggregation, as represented by equation (4). The resulting distribution of population-scaled fitness effects (fig. 2*A*) is similar to the earlier model, with a rate of evolutionary change essentially independent of effective population size, with $\langle dN/dS \rangle$ changing from 0.327 for $N_e = 10^4$ to 0.323 for $N_e = 10^8$.

Removing epistasis through the use of the model represented by equation (7) results in a strong dependence of the distribution of *S* on the population size, as shown in figure 2*B*. There is also an extremely strong dependence of the substitution rate with effective population size, with $\langle dN/dS \rangle$ changing from 0.725 for $N_e = 10^4$ to 0.357 for $N_e = 10^6$ to 0.027 for $N_e = 10^8$.

Variations in Effective Population Size: Bottlenecks

In contrast to differences in effective population size, there can also be fluctuations in effective population size. We perform simulations where the effective population size alternates between $N_e = 10^6$ and $N_e = 10^4$, as illustrated in the bottom of figure 3*A*, representing repetitive population bottlenecks. The resulting distribution of population-scaled fitness effects, for various timescales of population changes, is shown in figure 2*C*. With faster changes in population, there is a substantial increase in the number of advantageous mutations. The dependence of the rate of evolution on the period of the population changes is shown in figure 3*B*. With faster and faster changes, there is approximately a doubling of the average value of d*N*/d*S*. The instantaneous value of d*N*/d*S* for a period of fluctuation of 0.1, averaged over 1,000 cycles, is shown in figure 3*A*. Both increases and decreases in effective population sizes cause increases in the instantaneous value of d*N*/d*S*; when the population size increases, there is increased

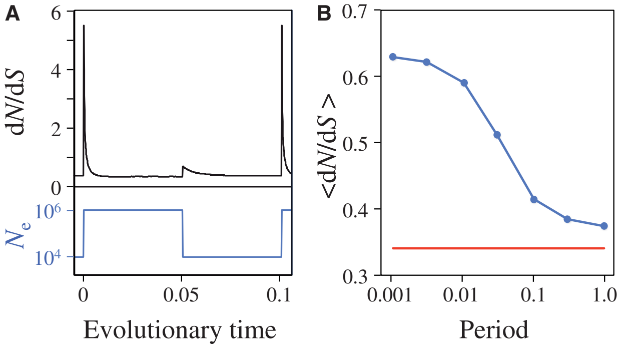

Fig. 3.—(*A*) Instantaneous values of d*N*/d*S* during fluctuations in effective population size, when the fluctuations have a period of 0.1. Both increases and decreases in population size cause transient increases in the rate of evolution. Periods represent durations in evolutionary time corresponding to expected number of base substitutions per nucleotide position under conditions of neutral evolution. (*B*) Averaged values of $\langle dN/dS \rangle$ for fluctuating population sizes, as a function of the period of the fluctuations (blue). The value of $\langle dN/dS \rangle$ when the population is fixed at $N_e = 10^6$ is shown for comparison (red).

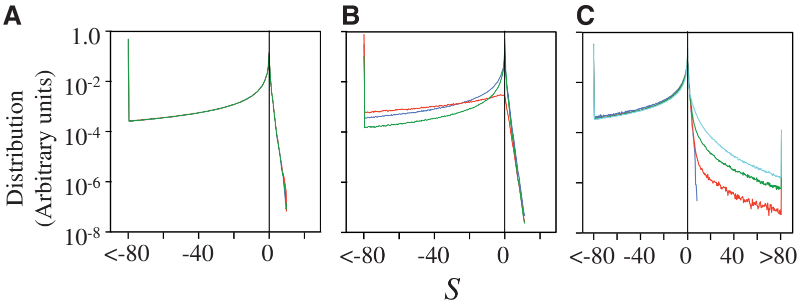

Fig. 2.—(*A*) Distribution of population-scaled fitness effects for nonsynonymous mutations using a fitness model penalizing unfolded protein, based on equation (4), for $N_e = 10^4$ (green), $N_e = 10^6$ (blue), and $N_e = 10^8$ (red). (*B*) Distribution of population-scaled fitness effects for a model where epistasis has been removed, based on equation (7), with the same color scheme as (*A*). (*C*) Distribution of population-scaled fitness effects for effective population size oscillating between 10^4 and 10^6, with the fitness calculated using equation (2), for various periods of the fluctuation: fixed at $N_e = 10^6$ (blue), period = 0.1 (red), 0.01 (green), and 0.001 (cyan). Periods represent durations in evolutionary time corresponding to expected number of base substitutions per nucleotide position under conditions of neutral evolution.

selection for greater stability, resulting in an increase in the number of stabilizing (adaptive) substitutions, while decreases in the population size results in a decrease in selective constraints, resulting in increased acceptance of slightly destabilizing substitutions.

Discussion

Using a simple but reasonable model of protein thermodynamics to provide a fitness function, we find that the distribution of the population-scaled fitness effects and the substitution rate are remarkably unaffected by the effective population size. In contrast to the small dependence of these evolutionary parameters on N_e, we find a strong effect from time varying effective population sizes. There is a large transient increase in the number of adaptive substitutions when the population size increases, as the protein adapts to the greater degree of selective pressure; there is also a transient increase in the number of slightly deleterious substitutions when the population size decreases, as the selective pressure relaxes and the protein evolves to lower stabilities. This effect depends on the timescale of the population fluctuations, but is significant over a wide range.

Why the Lack of Dependence on Effective Population Size?

When a mutation occurs, the values of s and S corresponding to a given value of $\Delta\Delta G$ is approximately given by

$$\begin{aligned} s &\approx \frac{dw}{d(\Delta G)}\Delta\Delta G \\ S &\approx \left(4N_e\frac{dw}{d(\Delta G)}\right)\Delta\Delta G \end{aligned} \tag{8}$$

where we have assumed that the fitness of the wild type is close to unity (true of these simulations) and that the magnitude of ΔG is sufficiently small that a Taylor expansion is justified. (This is also not a bad assumption, as mutations with very large destabilizing effects will be evolutionarily unimportant, and mutations with very large stabilizing effects are extremely rare.) (A more exact but less general calculation can be performed starting with eq. 5.) For any given values of N_e and $dw/d(\Delta G)$, S is proportional to $\Delta\Delta G$, so that distribution of population-scaled selective effects, $\rho_S(S)$, is then a stretched version of $\rho_{\Delta\Delta G}(\Delta\Delta G)$ given by

$$\rho_S(S) \approx \frac{1}{|\lambda|}\,\rho_{\Delta\Delta G}\left(\frac{S}{\lambda}\right) \tag{9}$$

where λ is given by

$$\lambda = 4N_e\frac{dw}{d(\Delta G)}. \tag{10}$$

The stability is based on a large number of stabilizing and destabilizing interactions, involving residues throughout the protein. The magnitude of these interactions is on the same scale as the total free energy of folding, so that significant changes in ΔG can be caused by modifying only a few of these interactions. As a result the distribution of changes of free energy of folding, $\rho_{\Delta\Delta G}(\Delta\Delta G)$, is relatively unaffected by the stability ΔG, as long as the protein is not excessively (i.e., unrealistically) stable (Goldstein 2011), a conclusion that has been verified both by other simulations and experimental measurements (Bloom et al. 2005, 2006, 2007; Tokuriki et al. 2007).

Although $\rho_{\Delta\Delta G}(\Delta\Delta G)$ is independent ΔG, the slope of the fitness function of equation (2)

$$\frac{dw}{d(\Delta G)} = \frac{-e^{\Delta G/kT}}{kT\left(1+e^{\Delta G/kT}\right)^2} \tag{11}$$

will be dependent on the protein stability, becoming closer to zero as the protein stability increases, so that λ is dependent on ΔG as well as N_e.

As a protein evolves toward higher stability, the distribution of $\Delta\Delta G$ is constant but the selective pressure relaxes until the expected change in fitness, or alternatively the equilibrium average value of S for accepted mutations, is approximately zero. We can describe this equilibrium condition as

$$\begin{aligned} \langle S\rangle_{\text{Fixed}} &= \int S\,P_{\text{Fix}}(S)\,\rho_S(S)\,dS \approx 0 \\ &= \frac{1}{|\lambda|}\int S\,P_{\text{Fix}}(S)\,\rho_{\Delta\Delta G}\left(\frac{S}{\lambda}\right)dS \end{aligned} \tag{12}$$

where $\langle S\rangle_{\text{Fixed}}$ represents the value of S averaged over fixed substitutions, and $P_{\text{Fix}}(S)$ is the fixation probability, which we are assuming, as in equation (1), is only a function of S. Note that, as long as $\rho_{\Delta\Delta G}(\Delta\Delta G)$ is fixed, the only adjustable parameter in equation (12) is λ. There will be a certain value of $\lambda = \lambda_{\text{Eq}}$ where equation (12) is satisfied. (For the current model, this value is approximately $\lambda_{\text{Eq}} = -0.766$.) More negative values result in a positive $\langle S\rangle_{\text{Fixed}}$, moving the system to a flatter region of the fitness curve, making λ less negative, while less negative values result in a positive $\langle S\rangle_{\text{Fixed}}$. The result is that the free energy of folding will change, modifying $dw/d(\Delta G)$ until $\lambda = \lambda_{\text{Eq}}$, so that the change in the slope of the fitness landscape cancels the effect of the changing population size. At this equilibrium, the distribution of population-scaled fitness effects will be given by equation (9) with $\lambda = \lambda_{\text{Eq}}$. The resulting distribution of S will be dependent on the value of λ_{Eq}, which will depend on the forms of $\rho_{\Delta\Delta G}(\Delta\Delta G)$ and $P_{\text{Fix}}(S)$, but not on the value of N_e.

The generality of this argument indicates that this observation should not be dependent on a specific fitness function. What is required is for the fitness to be a concave function of some parameter and that this parameter is what Cherry terms an equimutable parameter, where the distribution of changes of this parameter with mutations is independent of its current value (Cherry 1998); for the current model, this is satisfied by the observed independence of $\rho_{\Delta\Delta G}(\Delta\Delta G)$

on ΔG. It is likely that the characteristics of protein stability that provide for equimutability—that stability is a composite function that depends on contributions from many appropriately sized terms, where the rapidly declining number of increasingly stable sequences means the stability is far from optimal—is common in biology. We use fraction of proteins folded (eq. 2) as our fitness function, but alternative formulations (avoiding of self aggregation, eq. 4) give similar results. The calculation of the fitness is highly epistatic, where the contribution of each amino acid to the fitness depends on the rest of the protein sequence. This epistasis is required for this population size independence. When the epistasis is removed and the fitness becomes the sum of a large number of contributions from simple states, the fitness function ceases to be a concave function of a composite property, and a strong population size dependence results, as shown in figure 2*B*.

The results presented here indicate that changes in population size affect evolutionary dynamics quite differently from differences in population size, as has been suggested by Charlesworth and Eyre-Walker (2007) and Cherry (1998). In particular, they noted that large increases in population size can cause adaptive bursts that overcome the decrease in substitution rate due to the stronger selective constraints. The work presented here also emphasizes the role of population changes, except in our model it is only the changes in population size that cause significant changes in the substitution rate. As a result, population increases of arbitrary size will cause increases in the substitution rate, as there is no decrease in the substitution rate to be overcome. Similarly, decreases in the population size will cause increases in the substitution rate, but this will only be a transient effect resulting from the change in population size, rather than the difference in population size.

How Does the Substitution Rate Depend on the Distribution of Mutational Effects?

In addition to being largely independence of population size, the substitution rate will also be independent of the magnitude of the effect of mutations on the protein stability; that is, scaling all of the values of $\Delta\Delta G$ by a constant factor α will result in a change the stability of the protein so as to scale $\mathrm{d}w/\mathrm{d}(\Delta G)$ by $1/\alpha$, resulting in the same distribution of $\rho_S(S)$ and thus the same substitution rate. The substitution rate, however, is dependent on the shape of $\rho_{\Delta\Delta G}(\Delta\Delta G)$ To explore this dependence, we constructed a simpler model where a fraction p_- of all mutations is destabilizing (with change in free energy of folding $\Delta\Delta G_-$), a fraction p_0 is neutral ($\Delta\Delta G_0 = 0$), and a fraction p_+ is stabilizing (with change in free energy of folding $\Delta\Delta G_+$). The protein stability ΔG is adjusted until equation (12) is satisfied, and the relative fitness of the three different types of mutants and the corresponding acceptance rates calculated using equation (1). Figure 4*A* shows the dependence of the substitution rate $\langle \mathrm{d}N/\mathrm{d}S \rangle$ on the fraction p_- and effect $\Delta\Delta G_-$ of the destabilizing mutations ($p_+ = 0.05$, $\Delta\Delta G_+ = -0.5\ \mathrm{kcal\,mol}^{-1}$, $p_0 = 1 - p_+ - p_-$). As shown, the rate is relatively insensitive to the magnitude of the destabilization, but extremely sensitive to the relative fraction. Increasing $\Delta\Delta G_-$ results in a corresponding stabilization of the protein, resulting in a movement to the flatter part of the fitness curve, reducing the impact of these deleterious mutations on the fitness. Compensation resulting from changes in p_- are much weaker. Modifying the fraction and effect of stabilizing mutations shows a smaller effect, with the substitution rate increasing both with the fraction and magnitude of the stabilization, as shown in figure 4*B* ($p_- = 0.3$, $\Delta\Delta G_- = 2.0\ \mathrm{kcal\,mol}^{-1}$, $p_0 = 1 - p_+ - p_-$).

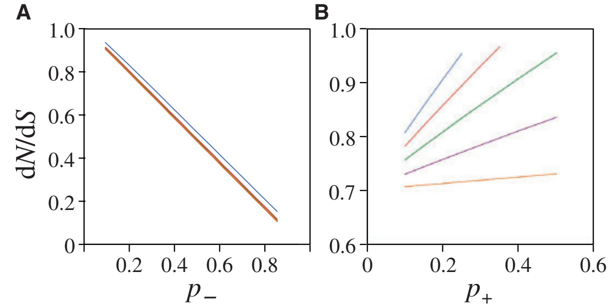

Fig. 4.—(*A*) Effect on d*N*/d*S* of changing the fraction of mutations that are deleterious (p_-), computed using a simple model, where the effect of the mutation on the free energy of folding is equal to $\Delta\Delta G_- = 1\ \mathrm{kcal\,mol}^{-1}$ (blue), $\Delta\Delta G_- = 2\ \mathrm{kcal\,mol}^{-1}$ (red), $\Delta\Delta G_- = 3\ \mathrm{kcal\,mol}^{-1}$ (green), $\Delta\Delta G_- = 4\ \mathrm{kcal\,mol}^{-1}$ (purple), and $\Delta\Delta G_- = 5\ \mathrm{kcal\,mol}^{-1}$ (orange). Other parameters as defined in the text. (*B*) Effect on d*N*/d*S* of changing the fraction of mutations that are advantageous (p_+), computed using a simple model, where the effect of the mutation on the free energy of folding is equal to $\Delta\Delta G_+ = -2\ \mathrm{kcal\,mol}^{-1}$ (blue), $\Delta\Delta G_+ = -1.5\ \mathrm{kcal\,mol}^{-1}$ (red), $\Delta\Delta G_+ = -1\ \mathrm{kcal\,mol}^{-1}$ (green), $\Delta\Delta G_+ = -0.5\ \mathrm{kcal\,mol}^{-1}$ (purple), and $\Delta\Delta G_+ = -0.1\ \mathrm{kcalmol}^{-1}$ (orange). Other parameters as defined in the text.

Comparison with Experimental Observations

The results presented here seem in contradiction with the observations described in the Introduction, where faster substitution rates are observed in 1) regions of the genome undergoing slow recombination compared with regions undergoing fast recombination, 2) endosymbionts compared with their free living relatives, and 3) island populations compared with mainland populations. In each of these three cases, there seems to be faster evolution in smaller populations, as would be predicted based on the nearly neutral model.

As remarked earlier, there are many complicated issues in these comparisons, so that it is difficult to conclude that the only relevant differences between these two sets are differences in effective population size. We also note that the differences in substitution rates observed in these comparisons are generally quite modest. For instance, Woolfit and Bromham (2005) observed a median increase in d*N*/d*S* of only 20% in island populations compared with mainland populations, with no significant increase in overall substitution rate. Campos et al. (2012) observed that autosomal genes in nonrecombining regions in *Drosophila* had a d*N*/d*S* ratio of approximately 45% higher than similar genes in recombining regions of the genome. Although it is difficult to make quantitative comparisons with the simple models presented here, and it is difficult to estimate differences in effective population sizes (Gossmann et al. 2011, 2012), position-specific measures of fitness, as represented by equation (7), result in d*N*/d*S* increasing by a over factor of over 26 as the population size is reduced from 10^8 to 10^4, which suggests that it may be the weakness of effect of population size on substitution rate that requires an explanation.

Additionally, these comparisons often interpret changes in effective population size as equivalent to differences in effective population size. Island populations undergo severe population bottlenecks, and bottlenecks are generally considered to reduce the effective population size. The analysis presented here suggests that population bottlenecks affect evolutionary dynamics quite differently from constant differences in population size, and it might be the population bottlenecks, with the resulting decrease and increase in selective constraints, that are affecting the substitution rate, while a static lower effective population size would have no such effect. Charlesworth and Eyre-Walker (2007), for instance, observed that differences in substitution rate between island and mainland populations depend upon whether a mainland population colonized an island (population size decrease in the island population) or an island population colonized a mainland (population size increase in the mainland population). Significantly, in the latter case, the mainland population generally had a higher rate of evolution than the island population, as would be predicted by the model presented here. This indicates that comparisons between the evolution of different lineages should be interpreted with care, as it would be difficult to disentangle the very different ways that static population size differences and population size fluctuations contribute to substitution rates. This is an inherent problem with this type of comparisons, as related lineages with different effective population sizes must have experienced the changes in population size that caused these differences.

A similar argument can be made comparing the effect of recombination rates on effective population size. Competition between mutations occurring at different points on a genetically linked region of the genome may correspond to reduced effective population sizes, but these mutations would occur sporadically. In this case, there would be temporal fluctuations in this effective population size, as other mutations with different fitness effects occur in nearby genes. In this way, lack of recombination would result in variations in effective population size, increasing the rate of evolutionary change, as has been observed experimentally (Larracuente et al. 2008; Arguello et al. 2010; Campos et al. 2012). Regions of low or no recombination would also be more subject to selective sweeps, providing a further mechanism for rapid changes in effective population size. Again, as with lineage-specific substitution rates, it is difficult to disentangle differences from fluctuations in effective population sizes.

Why Do Some Proteins Evolve Faster than Others?

As pointed out by Cherry (1998), with the exception of conspicuous outliers, differences in the substitution rates in different proteins is surprisingly modest, varying by approximately an order of magnitude (Grishin et al. 2000). Differences in the mutation rate in different parts of the genome would contribute to this rate variation, as would differences in the number of sites under nonthermodynamic constraints such as the requirements of functionality (Zuckerkandl 1976). In addition, as described earlier, the substitution rate is strongly dependent on $\rho_{\Delta\Delta G}(\Delta\Delta G)$, in particular on the fraction of destabilizing mutations. It is likely that this distribution is dependent on the size, structure, composition, and environment of the protein, leading to variation in the substitution rates. Finally, there has been significant interest in proteins that are unfolded under physiological conditions, or contain significant unstructured regions (Wright and Dyson 1999; Dunker et al. 2008). The selective constraints on these proteins and regions are still poorly understood.

Limitations of the Model

Evolutionary dynamics are dependent on the mapping between genotype, phenotype, and fitness. In this article, we analyze a model of protein thermodynamics that provide a reasonable mapping between these quantities. It is known that achieving sufficient thermodynamic stability is an important selective constraint for many proteins (Wang and Moult 2001; Zeldovich et al. 2007; Drummond and Wilke 2008; Serohijos et al. 2012). Although our model of thermodynamics is, by necessity, simplistic, it does include many realistic aspects, such as the need for considering differences between the free energy of the native state and a large ensemble of alternative states and the stability being a holistic function of many epistatic energetic interactions. Importantly, it reproduces many known properties of proteins and their evolution, such as the observed marginal stability, the higher rate of evolution of exposed locations compared with buried locations, the tendency for hydrophobic residues to cluster in the interior, the dependence of protein stability on population size, and over-dispersion of the molecular clock (Goldstein 2011).

Our measure of organismal fitness, the fraction of proteins folded at equilibrium, is certainly overly simple. There are likely to be specific requirements at particular locations in the protein, necessary for achieving functionality. These requirements on the protein sequence will, however, likely be sufficiently rigid so that changes in these locations would contribute minimally to the evolutionary dynamics. Other properties, such as resistance to aggregation, are also likely significant (Chen and Dokholyan 2008; Zhang et al. 2008; Johnson and Hummer 2011; Levy et al. 2012; Yang et al. 2012). Using a different fitness function based on avoiding aggregation, as represented by equation (4), did not significantly change the results. As long as the fitness function is a concave function of free energy of folding, increasing the population size will move the protein to a higher, and correspondingly flatter, region of the fitness landscape, and will result in the population-independent substitution rates described here. This is likely true even if the fitness is a concave function of some other quantity besides protein stability (such as saturation kinetics in biochemical reactions; Hartl et al. 1985), as long as this quantity is an aggregate quantity dependent on overall properties of the protein sequence that fulfils the equimutability criterion.

An important caveat of this analysis is the assumption of a slow mutation rate, so that the time for fixation (or elimination) is short relative to the length of evolutionary time between mutations. Genetic variation in the population would affect the substitution rate and would also be dependent on the population size. Wylie and Shakhnovich (2011), for instance, have observed in a simple model that the distribution of fitness effects depends on the mutation rate, indicating that the presence of multiple mutations in the population has an effect. Similar complications can emerge if the timescale for fluctuations in population size become comparable with or shorter than the fixation time (Otto and Whitlock 1997). Neglecting this effect is a limitation of this work.

Acknowledgments

The author thanks Brian Charlesworth and Adam Eyre-Walker for helpful discussions. This work was supported by the NIMR (MRC).

Literature Cited

Arguello JR, et al. 2010. Recombination yet inefficient selection along the *Drosophila melanogaster* subgroup's fourth chromosome. Mol Biol Evol. 27:848–861.

Berlin S, Ellegren H. 2006. Fast accumulation of nonsynonymous mutations on the female-specific W chromosome in birds. J Mol Evol. 62: 66–72.

Bershtein S, Mu W, Shakhnovich EI. 2012. Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. Proc Natl Acad Sci U S A. 109:4857–4862.

Bloom JD, Labthavikul ST, Otey CR, Arnold FH. 2006. Protein stability promotes evolvability. Proc Natl Acad Sci U S A. 103:5869–5874.

Bloom JD, Raval A, Wilke CO. 2007. Thermodynamics of neutral protein evolution. Genetics 175:255–266.

Bloom JD, et al. 2005. Thermodynamic prediction of protein neutrality. Proc. Natl Acad Sci U S A. 102:606–611.

Bromham L. 2011. The genome as a life-history character: why rate of molecular evolution varies between mammal species. Philos Trans R Soc Lond B Biol Sci. 366:2503–2513.

Bromham L, Rambaut A, Harvey PH. 1996. Determinants of rate variation in mammalian DNA sequence evolution. J Mol Evol. 43: 610–621.

Bustamante CD. 2005. Population genetics of molecular evolution. In: Nielsen R, editor. Statistical methods in molecular evolution. New York: Springer-Verlag. p. 63–99.

Campos JL, Charlesworth B, Haddrill PR. 2012. Molecular evolution in nonrecombining regions of the *Drosophila melanogaster* genome. Genome Biol Evol. 4:278–288.

Charlesworth B. 2013. Stablizing selection, purifying selection, and mutational bias in finite populations. Genetics 194:955–971.

Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. Proc Natl Acad Sci U S A. 104:16992–16997.

Chen P, Shakhnovich EI. 2009. Lethal mutagenesis in viruses and bacteria. Genetics 183:639–650.

Chen Y, Dokholyan NV. 2008. Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. Mol Biol Evol. 25:1530–1533.

Cherry JL. 1998. Should we expect substitution rate to depend on population size? Genetics 150:911–919.

Crow JF, Kimura M. 1970. An introduction to population genetics theory. New York: Harper & Row.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.

Dunker AK, Silman I, Uversky VN, Sussman JL. 2008. Function and structure of inherently disordered proteins. Curr Opin Struct Biol. 18: 756–764.

Duret L. 2006. The GC content of primates and rodents genomes is not at equilibrium: a reply to Antezana. J Mol Evol. 62:803–806.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8:610–618.

Fisher RA. 1930. The distribution of gene ratios for rare mutations. Proc R Soc Edinburgh. 50:205–220.

Geiler-Samerotte KA, et al. 2011. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci U S A. 108:680–685.

Gillespie JH. 1999. The role of population size in molecular evolution. Theor Popul Biol. 55:145–156.

Giver L, Gershenson A, Freskgard PO, Arnold FH. 1998. Directed evolution of a thermostable esterase. Proc Natl Acad Sci U S A. 95: 12809–12813.

Goldstein RA. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins 79:1396–1407.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol. 4:658–667.

Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective population size within a genome. Genetics 189: 1389–1402.

Grishin NV, Wolf YI, Koonin EV. 2000. From complete genomes to measures of substitution rate variability within and between proteins. Genome Res. 10:991–1000.

Hardison RC, et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13:13–26.

Hartl DL, Dykhuizen DE, Dean AM. 1985. Limits of adaptation: the evolution of selective neutrality. Genetics 111:655–674.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet Res. 89:311–336.

Johnson KP, Seger J. 2001. Elevated rates of nonsynonymous substitution in island birds. Mol Biol Evol. 18:874–881.

Johnson ME, Hummer G. 2011. Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. Proc Natl Acad Sci U S A. 108:603–608.

Kimura M. 1957. Some problems of stochastic processes in genetics. Ann Math Stat. 28:882–901.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47:713–719.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217: 624–626.

Kimura M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kingman JFC. 1978. A simple model for the balance between selection and mutation. J Appl Probab. 15:1–12.

Korkegian A, Black ME, Baker D, Stoddard BL. 2005. Computational thermostabilization of an enzyme. Science 308:857–860.

Larracuente AM, et al. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 24:114–123.

Levy ED, De S, Teichmann SA. 2012. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci U S A. 109:20461–20466.

Lindqvist Y, Johansson E, Kaija H, Vihko P, Schneider G. 1999. Three-dimensional structure of a mammalian purple acid phosphatase at 2.2 A resolution with a mu-(hydr)oxo bridged di-iron center. J Mol Biol. 291:135–147.

Miyazawa S, Jernigan R. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A. 93:2873–2878.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96–98.

Ohta T. 1977. Extension to the neutral mutation random drift hypothesis. In: Kimura M, editor. Molecular evolution and polymorphism. Mishima (Japan): National Institute of Genetics. p. 148–167.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst. 23:263–286.

Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol. 40:56–63.

Otto SP, Whitlock MC. 1997. The probability of fixation in populations of changing size. Genetics 146:723–733.

Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci U S A. 109: E1352–E1359.

Rispe C, Moran NA. 2000. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. Am Nat. 156:425–441.

Serohijos AW, Rimas Z, Shakhnovich EI. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep. 2:249–256.

Serrano L, Day AG, Fersht AR. 1993. Step-wise mutation of barnase to binase—a procedure for engineering increased stability of proteins and an experimental-analysis of the evolution of protein stability. J Mol Biol. 233:305–312.

Tamuri AU, dos Reis M, Goldstein RA. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. Genetics 190:1101–1115.

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. 369:1318–1332.

Van den Burg B, Vriend G, Veltman OR, Eijsink VGH. 1998. Engineering an enzyme to resist boiling. Proc Natl Acad Sci U S A. 95: 2056–2060.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. Hum Mutat. 17:263–270.

Weinreich DM. 2001. The rates of molecular evolution in rodent and primate mitochondrial DNA. J Mol Evol. 52:40–50.

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

Woolfit M, Bromham L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. Mol Biol Evol. 20:1545–1555.

Woolfit M, Bromham L. 2005. Population size and molecular evolution on islands. Proc Biol Sci. 272:2277–2282.

Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 293:321–331.

Wright SD, Gillman LN, Ross HA, Keeling DJ. 2009. Slower tempo of microevolution in island birds: implications for conservation biology. Evolution 63:2275–2287.

Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci U S A. 82:1741–1745.

Wyckoff GJ, Li J, Wu CI. 2002. Molecular evolution of functional genes on the mammalian Y chromosome. Mol Biol Evol. 19:1633–1636.

Wylie CS, Shakhnovich EI. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci U S A. 108:9916–9921.

Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A. 109:E831–E840.

Zeldovich KB, Chen P, Shakhnovich EI. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci U S A. 104:16152–16157.

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.

Zhao HM, Arnold FH. 1999. Directed evolution converts subtilisin E into a functional equivalent of thermitase. Protein Eng. 12:47–53.

Zuckerkandl E. 1976. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. J Mol Evol. 7: 167–183.

Associate editor: Bill Martin