DOI: 10.1534/genetics.111.128025

The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution

Duncan C. Ramsey,* Michael P. Scherrer,* Tong Zhou,† and Claus O. Wilke*,[1]

**Center for Computational Biology and Bioinformatics, Institute for Cellular and Molecular Biology and Section of Integrative Biology, University of Texas, Austin, Texas 78712 and* †*Section of Pulmonary, Critical Care, Sleep and Allergy, Department of Medicine and Institute for Personalized Respiratory Medicine, University of Illinois, Chicago, Illinois 60612*

Manuscript received February 21, 2011
Accepted for publication March 16, 2011

ABSTRACT

Recent work with *Saccharomyces cerevisiae* shows a linear relationship between the evolutionary rate of sites and the relative solvent accessibility (RSA) of the corresponding residues in the folded protein. Here, we aim to develop a mathematical model that can reproduce this linear relationship. We first demonstrate that two models that both seem reasonable choices (a simple model in which selection strength correlates with RSA and a more complex model based on RSA-dependent amino acid distributions) fail to reproduce the observed relationship. We then develop a model on the basis of observed site-specific amino acid distributions and show that this model behaves appropriately. We conclude that evolutionary rates are directly linked to the distribution of amino acids at individual sites. Because of this link, any future insight into the biophysical mechanisms that determine amino acid distributions will improve our understanding of evolutionary rates.

THE requirement for successful and efficient protein folding imposes significant biophysical constraints on coding sequences. These constraints shape how sequences evolve. Mutations that interfere with correct folding will generally be removed by purifying selection. Likewise, mutations that do not interfere with folding are often neutral, or nearly so, and accumulate over time. As a consequence of this interaction between protein biophysics and molecular evolution, signatures of protein structure can be found in the divergence patterns of coding sequences (Franzosa and Xia 2008; Lobkovsky *et al.* 2009; Wilke and Drummond 2010).

Mutagenesis experiments have shown that different positions in proteins have widely differing tolerances to amino acid substitutions (Reidhaar-Olson and Sauer 1988; Bowie *et al.* 1990; Lau and Dill 1990; Guo *et al.* 2004; Campbell-Valois *et al.* 2005; Smith and Raines 2006). On average, however, mutations introduced at solvent-exposed sites are less likely to disrupt protein structure and function than mutations introduced at buried sites. The latter tend to destabilize proteins, through steric hindrance and introduction of strained conformations (Chothia and Finkelstein 1990).

The higher tolerance of solvent-exposed sites to amino acid substitutions results in a correlation between the rate at which individual sites in coding sequences accumulate mutations over evolutionary time and the solvent exposure that these sites have in the expressed protein. Studies that have linked evolutionary rate with solvent exposure have consistently found that buried sites are more conserved and evolve slower than exposed sites (Overington *et al.* 1992; Goldman *et al.* 1998; Mirny and Shakhnovich 1999; Bustamante *et al.* 2000; Bloom *et al.* 2006; Conant and Stadler 2009; Franzosa and Xia 2009). At the same time, however, proteins with a larger core (more buried residues) evolve faster than proteins with a smaller core (Bloom *et al.* 2006; Ferrada and Wagner 2008; Zhou *et al.* 2008; Franzosa and Xia 2009). This apparent paradox can be resolved by observing that a larger core allows surface residues to vary more (Shakhnovich *et al.* 2005; Bloom *et al.* 2006; Franzosa and Xia 2009).

Recently, Franzosa and Xia (2009) developed a novel approach to analyze the relationship between evolutionary rate and solvent accessibility. They first mapped a large fraction of the genome of the yeast *Saccharomyces cerevisiae* onto homologous crystal structures from the Protein Data Bank (PDB). On the basis of this mapping, Franzosa and Xia (2009) determined relative solvent accessibility (RSA) for ~300,000 sites in the yeast genome. They then grouped these sites into bins of similar RSA value and calculated for each bin the average evolutionary rate d_N/d_S in a phylogeny of four yeast species. They found a strikingly linear relationship between evolutionary rate and RSA. Every 1% increase in RSA was associated with an increase in d_N/d_S of 0.001 (Franzosa and Xia 2009). Why the

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.111.128025/DC1.

[1]*Corresponding author:* Integrative Biology, 1 University Station–C0930, University of Texas, Austin, TX 78712. E-mail: cwilke@mail.utexas.edu

relationship between evolutionary rate and RSA is linear remains unknown.

Here, we employ mathematical modeling and bioinformatics analysis to explore what mechanism could be responsible for the linear relationship. We first show that a two-allele model in which selection strength correlates with RSA fails to reproduce this relationship. We then develop a more sophisticated model on the basis of amino acid frequencies and show that this model fails as well. The second model fails because amino acid frequencies averaged over many sites with comparable RSA differ dramatically from the distributions of allowed amino acids at individual sites. By building a model on the basis of the latter distributions, we can reproduce an approximately linear relationship between evolutionary rate and RSA.

METHODS

Evolutionary rate as a function of RSA: To verify the linear relationship between evolutionary rate and RSA at the amino acid level, we reproduced Franzosa and Xia's (2009) results using amino acid distance instead of d_N/d_S. First, we obtained orthologs between *S. bayanus* and *S. cerevisiae* from the Saccharomyces Genome Database as in Zhou *et al.* (2008) and aligned sequences with MUSCLE (Edgar 2004). We mapped the *S. cerevisiae* sequences to structures using three iterations of PSI-BLAST against the PDB, requiring a minimum of 80% sequence identity for a match. We ended up with 525 matching structures. For these matched structures we used the program DSSP (Kabsch and Sander 1983) to calculate solvent accessibility at each site. To obtain RSA, we normalized the solvent accessibilities calculated by DSSP with respect to an extended Gly-X-Gly peptide (Creighton 1992). We binned sites by RSA and then calculated evolutionary rate K with the PAML package codeml (Yang 2007), using the Whelan and Goldman (WAG) model for amino acid distance (Whelan and Goldman 2001).

Amino acid distribution over many yeast proteins: To calculate amino acid distributions, we used the same set of *S. cerevisiae* ORFs mapped to protein structures. We binned all sites by RSA as above. (A few residues had RSA > 1 and we treated them as if they had RSA $= 1$.) We then calculated the relative frequency of each amino acid in each RSA bin. For visualization, we ordered amino acids by hydrophobicity using the Fauchere–Pliska octanol scale (Fauchere and Pliska 1983).

Coordination number and RSA correlation: We computed the correlation between normalized coordination number and RSA using the same set of *S. cerevisiae* proteins as above. The coordination number of a site is the number of sites it is in contact with, and we considered two sites to be in contact if any two heavy atoms are within 4.5 Å of each other (excluding sequence neighbors). We used the BioPython module Bio.PDB (Hamelryck and Manderick 2003) to compute coordination numbers, which for each site we normalized by the average over the entire protein.

Variation at individual sites over structural homologs: To compute distributions at individual sites across structurally similar proteins, we employed a PSI-BLAST search of the NCBI nonredundant database (NR) to construct alignments from various seed proteins. As seed proteins, we used the proteins obtained by mapping the yeast genome to the PDB (as described above). We then filtered the alignment given by PSI-BLAST such that the remaining sequences all had between 40% and 80% pairwise sequence similarity with all other sequences in the alignment. This filtering procedure excluded redundant sequences while still ensuring structural similarity (Chothia and Lesk 1986; Holm *et al.* 1992). We retained only filtered alignments that contained at least 50 sequences. Our final data set consisted of 162 distinct alignments. In the filtered alignments, we classified each site by RSA of the seed protein at this site and placed sites into bins of similar RSA. We then calculated the alignment-wide amino acid distribution for every site. At each site, we ranked residues by declining frequency at that site. We then averaged the frequency-sorted amino acid distributions over all sites within each bin.

To characterize these averaged distributions with a single parameter, we fitted the one-parameter exponential function $e^{-\lambda k}$ to the average amino acid frequency as a function of the amino acid rank k.

Analysis scripts and data to reproduce this analysis are provided in supporting information, File S1.

Parameter choices: To study numerically the behavior of our mathematical models of protein evolution, we had to choose suitable values for the parameters N_e (effective population size) and μ (mutation rate). We chose values that are approximately correct for yeast, namely $N_e = 5 \times 10^6$ individuals and $\mu = 3.3 \times 10^{-10}$ mutations per site per generation (Lynch *et al.* 2008; Lancaster *et al.* 2010).

RESULTS

Franzosa and Xia (2009) found a strong linear relationship between d_N/d_S and RSA. While their result was likely driven by selection on the amino acid level, their use of d_N/d_S does not allow us to draw this conclusion *a priori*. Their result could be confounded by varying levels of selection on synonymous sites; synonymous codon usage is not uniform across genes and covaries with protein structure (Akashi 1994; Drummond and Wilke 2008; Zhou *et al.* 2009; Lee *et al.* 2010).

Therefore, to verify that Franzosa and Xia (2009) had indeed identified an effect occurring at the amino acid level, we repeated their analysis with amino acid sequences. We aligned orthologous genes from the two yeasts *S. cerevisiae* and *S. bayanus* and classified sites into

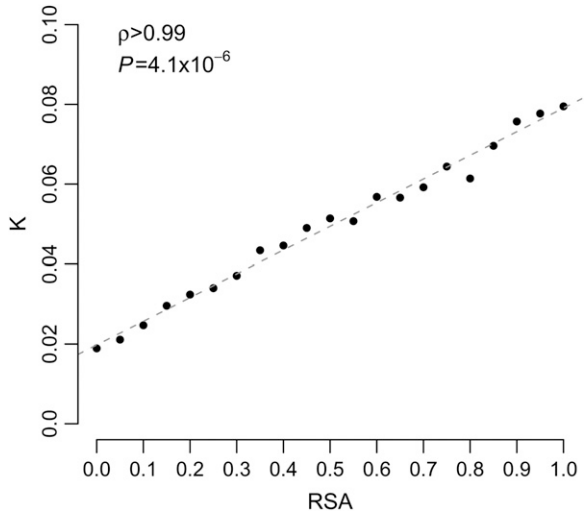

FIGURE 1.—Evolutionary rate K as a function of RSA, for yeast. The dashed line represents the fit of a linear function to the data.

bins of similar RSA values. We then concatenated all sites within each bin and calculated the amino acid distance K between the *S. cerevisiae* and the *S. bayanus* sequence in each bin. Amino acid distance is a measure of evolutionary rate on the amino acid level (WHELAN and GOLDMAN 2001).

We found a near-perfect linear relationship between evolutionary rate K and RSA (Figure 1). We interpret this result as a signal of purifying selection acting on the amino acid sequence. On average, buried sites experience stronger purifying selection than exposed sites and thus evolve slower. The increased selective constraints on buried amino acids presumably reflect the requirement for proteins to fold and function properly.

That buried sites are more constrained than exposed sites is well known. Much existing theory, experiments, and sequence data support the notion that substitutions in the core of a protein are more likely to be disruptive than substitutions in solvent-exposed regions. Yet the perfectly linear relationship between evolutionary rate and RSA is surprising and deserves an explanation. We thus proceeded to explore what kind of evolutionary models could potentially reproduce this observation.

A simple two-allele model: The simplest model we can consider is a multiplicative multisite, two-allele model; in this model, an organism's genome consists of a finite number of sites, each of which can exist in two alleles. All sites contribute multiplicatively to the overall fitness of the organism. At each site i, one of the two alleles is preferred, and we assume it has fitness 1. The second allele is selected against and has fitness $1 - s_i$. We assume that all sites mutate with the same rate μ. In such a model, in equilibrium, sites with larger s_i will evolve slower than sites with smaller s_i. For sufficiently small s_i, sites will evolve neutrally at rate μ.

Here and throughout, we consider haploid, asexual organisms and assume that the product of mutation rate and effective population size N_e is small, $\mu N_e \ll 1$. In this case, and because we consider a multiplicative model, the evolutionary rate of a genome of length L is the average of the evolutionary rates of L single-site models with identical selection coefficients. Therefore, in what follows, we consider only the evolutionary rate at a single site and ask how it changes with selection coefficient s. For simplicity, we drop the site index i.

We refer to the two alleles at a site as A and a. Allele A has fitness 1 and allele a has fitness $1 - s$. The probability that allele a goes to fixation in a background of allele A is given by KIMURA (1962):

$$\pi_{A\to a} = \frac{1 - e^{2s}}{1 - e^{2N_e s}}. \quad (1)$$

Likewise, the probability that allele A goes to fixation in a background of allele a is given by

$$\pi_{a\to A} = \frac{1 - e^{-2s}}{1 - e^{-2N_e s}}. \quad (2)$$

In equilibrium, and averaged over long periods of time, both alleles will be present at the site some fraction of time. We denote these fractions as $F(A)$ and $F(a)$, with $F(A) + F(a) = 1$. We have

$$F(A) = \frac{\pi_{a\to A}}{\pi_{a\to A} + \pi_{A\to a}}, \quad F(a) = \frac{\pi_{A\to a}}{\pi_{a\to A} + \pi_{A\to a}}. \quad (3)$$

Evolutionary rate K is the rate with which mutations originate and go to fixation. Thus, K is given by

$$K = \mu N_e [F(A)\pi_{A\to a} + F(a)\pi_{a\to A}]. \quad (4)$$

The evolutionary rate K is of course a function of s. Thus, we can now ask how K changes as s changes. Assuming $s \ll N_e$ and using standard approximations for the fixation probabilities, we obtain

$$K(s) \approx \frac{4s\mu N_e}{e^{2N_e s} - e^{-2N_e s}}. \quad (5)$$

For $s \leq 1/N_e$, evolution is neutral, and $K(s) \approx \mu$. For larger s, the evolutionary rate declines exponentially in s, $K(s) \approx 4s\mu N_e e^{-2N_e s}$.

We now assume that the selection coefficient s is a function of RSA. We denote RSA by r in mathematical expressions. An increase in K as r increases corresponds to a greater tolerance to mutation; hence, the selection coefficient $s(r)$ should be a decreasing function of r. We assume that r can take on any value in the interval $[0, 1]$. The function $s(r)$ maps this interval into some

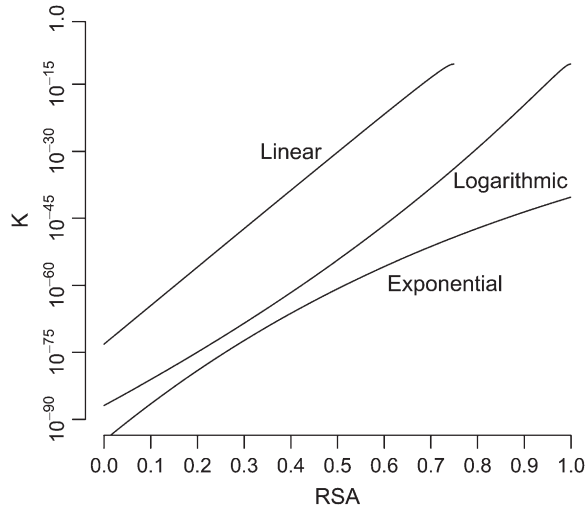

FIGURE 2.—Evolutionary rates K *vs.* RSA in a two-allele model. We mapped RSA r to the selection coefficient s via three functions: a linear one, $s(r) = [-r/5 + 0.15] \times 10^{-4}$; a logarithmic one, $s(r) = \log(2 - r) \times [50{,}000 \times \log(2)]^{-1}$; and an exponential one, $s(r) = \exp[-r + \log(5 \times 10^{-4})]$. We assumed $N_e = 5 \times 10^6$ and $\mu = 3.3 \times 10^{-10}$. Evolutionary rate K is highly nonlinear in all cases. Note that the y-axis uses a logarithmic scale.

interval of s values. Thus, we have to ask: Is there a reasonable mapping from r to $s(r)$ such that $K(s(r))$ is approximately a linear, increasing function in r? We found that generally, such a mapping does not exist. Because K decreases exponentially with s, any function $s(r)$ that might result in approximately linear behavior of K will necessarily have an exponentially small range of possible s values. To illustrate this result, we defined three functions with parameters that give similar ranges in $[0, 1]$: a linearly decaying function whose image is $[0, 1.5 \times 10^{-5}]$, an exponential function with image $[7.3 \times 10^{-6}, 2 \times 10^{-5}]$, and a logarithmic function spanning $\sim[0, 1.8 \times 10^{-5}]$. For all three definitions of $s(r)$, Equation 4 still produces exponentially fast growth of K as a function of r (Figure 2). More generally, we can show that even if the difference in s corresponding to fully buried ($r = 0$) and fully exposed ($r = 1$) sites is only on the order of $1/N_e$, the deviation from linearity is larger than the magnitude of the evolutionary rate K itself (see APPENDIX). We conclude that the two-allele model does not seem to be an appropriate model to describe the effect of relative solvent accessibility on evolutionary rate.

A model based on amino acid frequencies: We believe that the main reason why the two-allele model gives unsatisfactory results is that it replaces 20 different amino acids by only two different states, preferred and unpreferred. In real proteins, it may well be that at one site 3 amino acids are preferred and 17 unpreferred, while at a different site 5 are preferred and 15 unpreferred. All else being equal, the second site will evolve faster than the first. This reasoning suggested to us that we should aim to develop a model on the basis of amino acid frequencies. The sites with the broadest distributions of amino acids should evolve the fastest, and the sites with the narrowest distributions the slowest.

Amino acid distributions in proteins have been studied extensively. The general consensus is that amino acid frequencies follow a Boltzmann distribution. The individual frequencies at sites can be calculated either from stability effects [$\Delta\Delta G$ values (DOKHOLYAN and SHAKHNOVICH 2001; DOKHOLYAN *et al.* 2002; GODOY-RUIZ *et al.* 2004; BLOOM and GLASSMAN 2009; SCHMIDT AM BUSCH *et al.* 2010] or from the protein's connectivity matrix (PORTO *et al.* 2004; BASTOLLA *et al.* 2005; POKAROWSKI *et al.* 2005; WOLFF *et al.* 2008; BASTOLLA *et al.* 2008). In particular, PORTO *et al.* (2004) showed that the frequency of amino acid a is proportional to $e^{-\beta h(a)}$, where β measures properties of the site under consideration and $h(a)$ measures properties of the amino acid. The quantity β can be derived from the protein structure's contact matrix. It varies almost linearly with the site's coordination number normalized by the protein's average. The quantity $h(a)$ is the *interactivity* of amino acid a, a quantity highly correlated with hydrophobicity (BASTOLLA *et al.* 2005).

Because solvent occlusion happens through interresidue contacts, we hypothesized that the normalized coordination number should correlate strongly with RSA and that the theory of PORTO *et al.* (2004) should provide at least a qualitatively correct description of the amino acid distribution in different RSA bins. We found both to be the case in yeast. The normalized coordination number correlated well with RSA (Pearson's $r = 0.66$, $P < 2.2 \times 10^{-16}$). Amino acid distributions were strongly skewed toward hydrophobic residues at low RSA and toward hydrophilic residues at high RSA. For intermediate RSA, corresponding to $\beta = 0$, both hydrophobic and hydrophilic residues had comparable frequencies (Figure S1). Having found this correspondence, we proceeded to obtain the evolutionary rates predicted by the theory of PORTO *et al.* (2004).

The amino acid distribution at a site, combined with effective population size N_e and mutation rate μ, fully specifies the evolutionary rate at the site, under the assumption that sites evolve independently. The link between amino acid distribution and evolutionary rate is established by Sella–Hirsh theory (SELLA and HIRSH 2005). This theory demonstrates that equilibrium frequencies of alleles follow a Boltzmann distribution just like the one found by PORTO *et al.* (2004). Thus, from the equilibrium frequencies of alleles we can infer the relative fitness of alleles and their fixation probabilities in various backgrounds.

According to PORTO *et al.* (2004), the distribution of amino acids is given by

$$F(a) = \frac{\exp[-\beta h(a)]}{\sum_b \exp[-\beta h(b)]}, \quad (6)$$

where a is a specific amino acid as before and the sum in the denominator runs over all 20 amino acids. Fixation probabilities follow as

$$\pi_{a\to b} = \frac{1 - [F(a)/F(b)]^{1/(N_e-1)}}{1 - [F(a)/F(b)]^{N_e/(N_e-1)}} \quad (7)$$

(Sella and Hirsh 2005). (These fixation probabilities are equivalent to the Kimura probabilities used in the previous subsection; see Sella and Hirsh 2005 for details.) We can now express evolutionary rate in terms of amino acid distribution and fixation probabilities as

$$K = \mu N_e \sum_a \left[F(a) \sum_{b\neq a} \pi_{a\to b} \right]. \quad (8)$$

Remember that K is a function of β, and β is an approximate measure of solvent accessibility. Highly buried sites will have a large negative β, highly exposed sites will have a large positive β, and intermediate sites will have a β close to zero. Thus, to be consistent with data (*e.g.*, Figure 1), Equation 8 should be an increasing function of β. Instead, however, we found that Equation 8 predicts K to be maximal at $\beta = 0$ (Figure 3A) and to decline in both directions as the absolute value of β increases. This result makes intuitive sense, as the distribution defined by Equation 6 is the broadest for $\beta = 0$. However, we have to conclude that the theory of Porto *et al.* (2004) cannot be used to explain the linear relationship between evolutionary rate and RSA.

We emphasize that the failure of Equation 8 does not imply that the amino acid distributions calculated by Porto *et al.* (2004) and given by Equation 6 are incorrect. In fact, we used Equations 7 and 8 to predict evolutionary rates from the observed amino acid frequencies in yeast and found similarly that the predicted evolutionary rate peaked at intermediate RSA (Figure 3B).

An alternative model based on amino acid frequencies: The failure of the previous model implies that the model missed some important aspect of protein evolution. We hypothesized that the model failed because Equation 6 was valid only for the entire class of sites with similar β, but not for any individual site in this class. It is entirely possible that the distribution of amino acids at a specific site, when observed over evolutionarily long periods of time, does not agree with Equation 6, even though the average distribution of all sites with similar β or RSA does. Both previously published tests of Equation 6 (Porto *et al.* 2004) and our amino acid distributions as a function of RSA (Figure S1) were obtained by averaging over many sites and thus would not reveal any deviation from Equation 6 at individual sites.

To determine the distribution of amino acids at individual sites, we built large alignments of structurally similar proteins (see methods). We found that the distributions at individual sites were highly variable and looked nothing like Equation 6. In general, at any given site, only a small number of different amino acids were actually present, and there was often no obvious relationship between which amino acids were present and what their hydrophobicity was. However, when averaging over many sites with similar RSA, we could recover distributions comparable to Equation 6.

Even though the specific amino acids preferred at individual sites were highly variable, we found that the frequency distributions at different sites were similar. When we ordered amino acids by their relative frequency at each site, we found that the frequencies were proportional to an exponential, $\exp(-\lambda k)$, where k counts amino acids in descending order of frequency, $k = 0, 1, \ldots, 19$. We averaged the reordered amino acid distributions over all sites within bins of similar RSA (Figure 4A) and fitted $\exp(-\lambda k)$ to these averaged distributions. We thus obtained λ as a function of RSA and found that λ decayed approximately linearly with RSA (Figure 4B and Figure S2).

We carried out this analysis on 162 yeast proteins and found that generally (i) λ was approximately a linear function of RSA and (ii) λ decayed with increasing RSA (Figure 5). For each protein, we fitted a linear function $\lambda(r) = c_1 + c_2 r$ to the data and generally found a negative slope c_2 and a good model fit. The few cases with an apparent positive slope c_2 could be traced back to a single outlying λ-value at the highest RSA bin (see Figure S3 for an example). This bin generally encompassed the fewest number of sites (see also discussion) and thus its λ-value was not always reliable.

On the basis of these findings, we can model the evolutionary process at individual sites such that it produces steady-state amino acid frequencies

$$F(a) = \frac{\exp[-\lambda a]}{\sum_b \exp[-\lambda b]}, \quad (9)$$

where a and b index amino acids, in the appropriate order, and run from 0 to 19. The parameter λ declines with RSA.

As in the previous subsection, we can use the Sella and Hirsh (2005) method to map these steady-state frequencies onto a unique evolutionary process. The fitness values for individual amino acids are given by

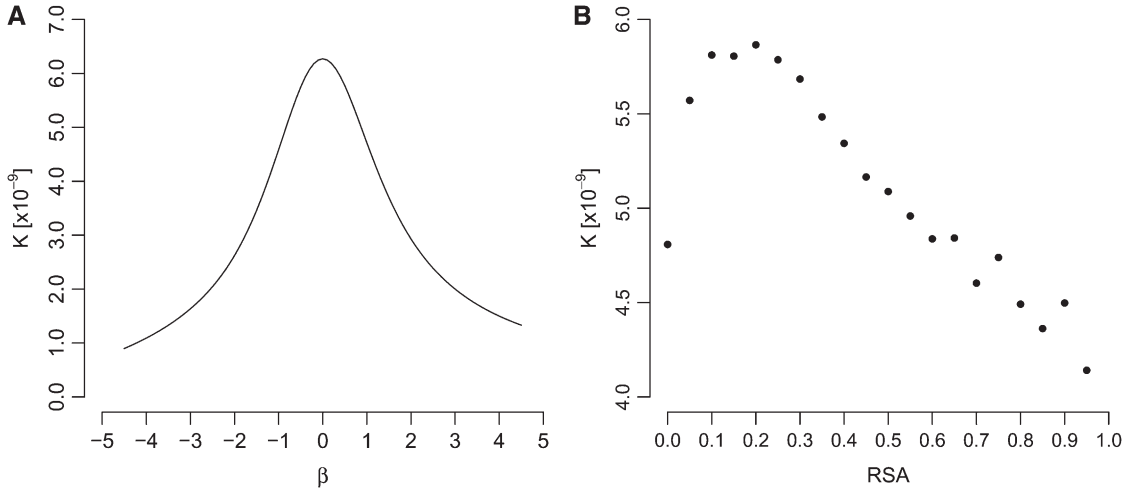

Figure 3.—Evolutionary rates predicted from amino acid distributions. (A) The amino acid distribution used is the one given by Porto *et al.* (2004). The parameter β correlates strongly with RSA. (B) The amino acid distribution used is the observed distribution in yeast; see Figure S1.

$$w(a) = \exp\left[-\lambda \frac{a}{2(N_e - 1)}\right] \approx 1 - \frac{\lambda a}{2N_e}. \tag{10}$$

It might seem disconcerting that we measure fitness here in units of the effective population size N_e. After all, the fitness contribution of a particular amino acid in a particular protein of an organism should not depend on the size of the population of that organism. However, this scaling by population size is merely a mathematical convenience to keep the actually observable quantities (amino acid distributions, evolutionary rates) free of any explicit dependency on N_e. For real organisms, we expect that $w(a)$ is independent of N_e but that λ, $F(a)$, and evolutionary rate K all depend on N_e.

Fixation probabilities follow from Equation 7. Making the approximation $N_e - 1 \approx N_e$, we find

$$\pi_{a\to b} = \frac{1 - \exp[-\lambda(a-b)/N_e]}{1 - \exp[-\lambda(a-b)]}. \tag{11}$$

We obtain the average evolutionary rate for this model by substituting Equations 9 and 11 into Equation 8. We find

$$K = \mu N_e \sum_a \left[\frac{e^{-\lambda a}}{Z} \sum_{b\neq a} \frac{1 - e^{-\lambda(a-b)/N_e}}{1 - e^{-\lambda(a-b)}}\right], \tag{12}$$

where $Z = \sum_b e^{-\lambda b}$ is the partition function.

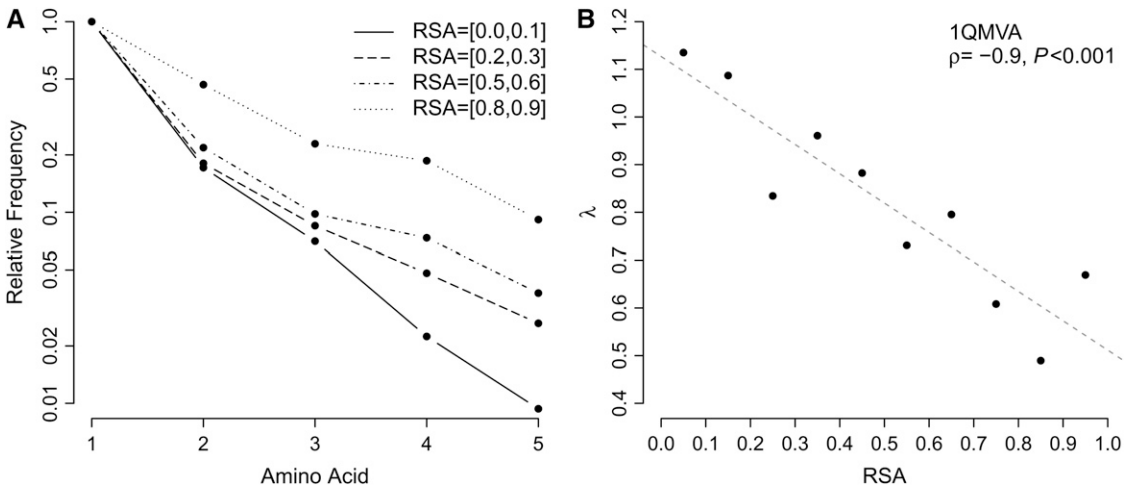

Figure 4.—Variation from primary residue increases with RSA for sequences homologous to thioredoxin peroxidase (PDB identifier 1QMV, chain A). (A) Normalized frequencies of most common residues averaged over all sites in four different RSA bins. (B) The exponential parameter λ approximating these normalized distributions decreases linearly as RSA increases. The dashed line represents the fit of a linear function to the data.

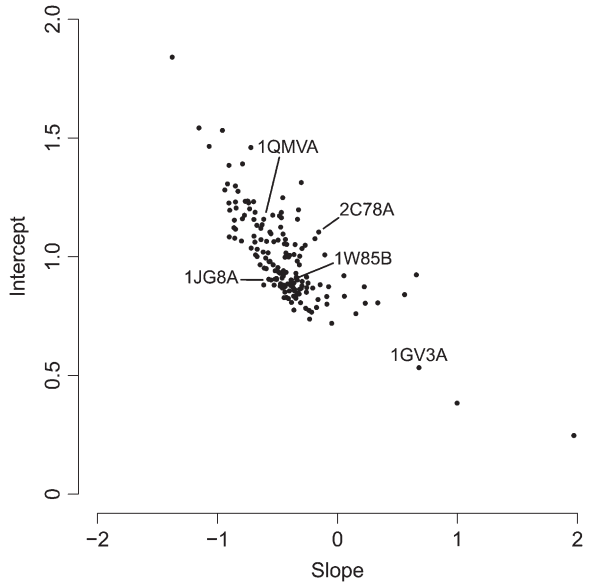

FIGURE 5.—Intercept c_1 and slope c_2 of λ as a function of RSA r, $\lambda(r) = c_1 + c_2 r$, when fitted to 162 yeast proteins. The highlighted proteins are used as examples in Figures 4 and 6 and Figure S2 and Figure S3.

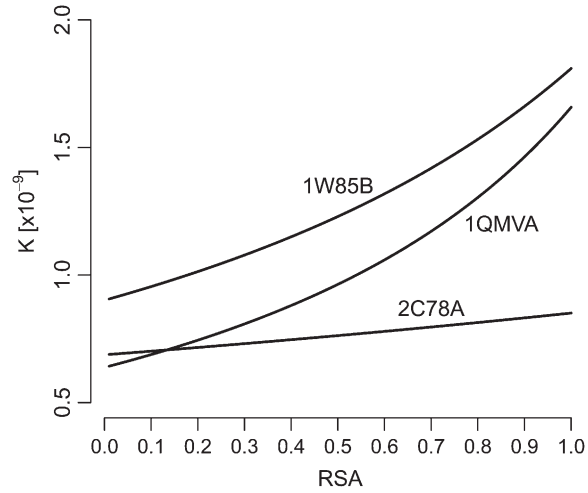

FIGURE 6.—Evolutionary rates predicted from Equation 13 for three different protein structures. Rates were calculated on the basis of fits of $\lambda(r) = c_1 + c_2 r$ to amino acid distributions, as in Figure 4. The fitted constants are given in Table S1.

For large N_e, we can approximate $e^{-\lambda(a-b)/N_e} \approx 1-\lambda(a-b)/N_e$, so that

$$K \approx \mu \sum_a \left[\frac{e^{-\lambda a}}{Z} \sum_{b \neq a} \frac{\lambda(a-b)}{1-e^{-\lambda(a-b)}} \right]. \qquad (13)$$

The absence of N_e from this equation shows that if we scale $w(a)$ with N_e, as in (10), then K is approximately independent of N_e.

To obtain evolutionary rate as a function of RSA, we substitute $\lambda = c_1 + c_2 r$ into Equation 13. Figure 6 shows resulting evolutionary rates for three representative proteins. The curves $K(r)$ are roughly linear and K is approximately of the correct order of magnitude. However, $K(r)$ is not perfectly linear; there is some clear upward curvature. The curvature tends to increase with the absolute magnitude of c_2. We comment on this issue in the DISCUSSION. Also, note that the units for K are not the same in Figure 1 as they are in the other figures. In Figure 1, K is estimated as the number of substitutions per site per unit time. The time unit is the total divergence time between the species that are being compared. By contrast, our mathematical models predict K in units of substitutions per site per generation. We estimate that $\sim 10^{11}$ generations separate *S. cerevisiae* and *S. bayanus*, 40 million yr × 4000 generations/yr.

DISCUSSION

We have shown that the linear relationship between evolutionary rate and RSA reflects a selection pressure on the amino acid level. Further, we have demonstrated that a simple two-allele model and a more elaborate model based on observed mean amino acid frequencies for sites with similar RSA cannot reproduce this linear relationship. The first model fails because it is too simplistic; individual sites in proteins can, at least in principle, assume 1 of 20 different states. The second model fails because amino acid frequencies averaged over many sites are not representative of amino acid frequencies at individual sites. We have found that the latter frequencies follow a Boltzmann distribution that becomes increasingly broad as RSA increases. Finally, we have shown that a mathematical model based on this observation can reproduce the linear relationship between evolutionary rate and relative solvent accessibility.

Our analysis highlights how important it is to distinguish between amino acid frequencies averaged over a large class of sites with similar property (such as RSA) and amino acid frequencies at individual sites. In both cases, frequencies are Boltzmann distributed, and thus it is easy to mistake one for the other. However, the properties of these two distributions are very different. For example, in yeast, at sites with RSA close to 0.2 nearly all amino acids occur at comparable frequencies. Yet at any given site, only a small number of amino acids are actually permissible. Evolutionary rate, which measures the rate at which mutations at individual sites arise and go to fixation, is governed by the amino acid distribution of individual sites, not the average distribution over a broad class of sites.

However, averaging distributions of similarly exposed sites from many proteins seems to agree qualitatively with distributions predicted by PORTO *et al.* (2004). This agreement suggests that any future theory attempting to predict site-specific distributions should also be able to predict average distributions of sites with similar β (or RSA). These average distributions should reduce

to something similar to the theory of Porto *et al.* (2004) and the data shown in Figure S1.

Our model describes the variation in steady-state distribution at sites using the exponential parameter λ, which we defined above as a linear function of RSA: $\lambda(r) = c_1 + c_2 r$. In this way $\lambda(r)$ describes the level of variation in the distribution function (Equation 9) for a given RSA. The intercept of $\lambda(r)$, the largest value it takes, corresponds to the strongest selective pressure and the minimal level of variation for the most buried sites, at $r = 0$. This maximal selective pressure in turn determines the value of the minimal evolutionary rate. Likewise, the slope of $\lambda(r)$ determines the rate of increase of $K(r)$: a steeper slope (more negative c_2) signifies a greater tolerance of alternative residues as r increases compared to a shallower slope (less negative c_2), and greater tolerance of alternative residues implies a greater increase in K as r grows.

We emphasize that different RSA bins contain different numbers of sites (see also Figure 2 from Franzosa and Xia 2009). Bins below RSA values of 0.1 tend to contain more than twice as many sites as bins for RSA values between 0.1 and 0.6. Bins for higher RSA values are even less occupied. In our data set of all yeast genes, we have 69,521 sites in the lowest-RSA bin but only 1452 sites in the highest-RSA bin. Because of the comparative scarcity of high-RSA sites, our estimates for amino acid distributions at these sites are not always reliable, as exemplified in Figure S3. In our experience, the amino acid distributions at high RSA are reliable when a linear model produces a negative slope for $\lambda(r)$ and they are unreliable otherwise.

In our analysis of amino acid distributions at individual sites in individual proteins, we generally observed only a few (on the order of 5) different amino acids at each site. This outcome was expected for Boltzmann-distributed amino acid frequencies. For the proteins we investigated, we found that λ typically fell somewhere between 0.3 and 1.2. Even for the smallest λ in this range, $\lambda = 0.3$, the expected frequency of the 10th most abundant amino acid under a Boltzmann distribution is only ~0.02, and the expected frequency of the 20th most abundant amino acid is ~0.001. For larger λ, the expected frequencies are much smaller. In our alignments, which mostly ranged from 50 sequences to ~200 sequences, with a few cases going up to 360 sequences, we could not properly sample amino acids that have such low expected abundances. In fact, in our distributions, the least abundant amino acid generally has absolute frequencies in low single digits, and thus we cannot expect to see any other amino acids that should, according to theory and the overall pattern we see, arise at less than single-digit frequencies.

By measuring fitness in units of N_e for ease of analysis, we have implicitly made $\lambda(r) = N_e\hat{\lambda}(r)$, where $\hat{\lambda}(r) = \hat{c}_1 + \hat{c}_2 r$. What we have then is a relation linking the original $\lambda(r)$ to N_e and the parameters $\hat{c}_1$ and $\hat{c}_2$. Note that the original $\lambda(r)$ is a statistically measurable function describing variation at sites by RSA. If we could obtain estimates of $\hat{c}_1$ and $\hat{c}_2$ independently of K, say from an *ab initio* model of protein folding, and then the relationships were formally attached to biophysical quantities that proved reliably measurable, this relationship $\lambda(r) = N_e\hat{\lambda}(r)$ could provide a novel method by which to estimate effective population size.

While our final model produces an approximately linear relationship between evolutionary rate and RSA, the model predictions are not perfectly linear. In particular for proteins with larger absolute c_2 values, we see a clear upward curvature in evolutionary rate as a function of RSA (Figure 6). In our modeling approach, we made several approximating assumptions, and each of them could potentially be the source of the curvature. First, we assumed that amino acid distributions are Boltzmann distributed. This assumption may not be entirely correct. In fact, if amino acid distributions were perfectly Boltzmann distributed, then the data in Figure 4A should be perfectly linear. Instead, they seem to display a moderate amount of curvature. Second, λ may not be a linear function of RSA. We did see a fair amount of noise in λ for some proteins (*e.g.*, Figure S2D), but we did not see any systematic deviation from the linear trend. Third, when modeling how amino acid distributions relate to evolutionary rate, we completely neglected any interactions among sites. While models without interactions have been successful in related studies, *e.g.*, in predicting the effect of multiple mutations on protein stability (Bloom *et al.* 2005) and in linking mutation frequencies to stability effects ($\Delta\Delta G$ values) (Godoy-Ruiz *et al.* 2004; Zeldovich *et al.* 2007; Bloom and Glassman 2009), epistatic interactions among sites in proteins are well documented and may be important for precise prediction of evolutionary rates.

We thank Markus Porto and Eugene Shakhnovich for helpful comments on this work. This work was supported by National Institutes of Health grant R01 GM088344 and by the National Science Foundation under Cooperative Agreement DBI-0939454.

LITERATURE CITED

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927–935.

Bastolla, U., M. Porto, H. E. Roman and M. Vendruscolo, 2005 Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins **58:** 22–30.

Bastolla, U., A. R. Ortíz, M. Porto and F. Teichert, 2008 Effective connectivity profile: a structural representation that evidences the relationship between protein structures and sequences. Proteins **73:** 872–888.

Bloom, J., D. A. Drummond, F. H. Arnold and C. O. Wilke, 2006 Structural determinants of the rate of protein evolution in yeast. Mol. Biol. Evol. **23:** 1751–1761.

Bloom, J. D., and M. J. Glassman, 2009 Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. PLoS Comp. Biol. **5:** e1000349.

Bloom, J. D., J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami *et al.*, 2005 Thermodynamic prediction of protein neutrality. Proc. Natl. Acad. Sci. USA **102:** 606–611.

Bowie, J. U., J. F. Reidhaar-Olson, W. A. Lim and R. T. Sauer, 1990 Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science **247:** 1306–1310.

Bustamante, C. D., J. P. Townsend and D. L. Hartl, 2000 Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. Mol. Biol. Evol. **17:** 301–308.

Campbell-Valois, F. X., K. Tarassov and S. W. Michnick, 2005 Massive sequence perturbation of a small protein. Proc. Natl. Acad. Sci. USA **102:** 14988–14993.

Chothia, C., and A. V. Finkelstein, 1990 The classification and origins of protein folding patterns. Annu. Rev. Biochem. **59:** 1007–1039.

Chothia, C., and A. M. Lesk, 1986 The relation between the divergence of sequence and structure in proteins. EMBO J. **5:** 823–826.

Conant, G. C., and P. F. Stadler, 2009 Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol. Biol. Evol. **26:** 1155–1161.

Creighton, T. E., 1992 *Proteins: Structures and Molecular Properties*. W. H. Freeman, New York

Dokholyan, N. V., and E. Shakhnovich, 2001 Understanding hierarchical protein evolution from first principles. J. Mol. Biol. **312:** 289–307.

Dokholyan, N. V., L. A. Mirny and E. Shakhnovich, 2002 Understanding conserved amino acids in proteins. Physica A **314:** 600–606.

Drummond, D. A., and C. O. Wilke, 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell **134:** 341–352.

Edgar, R., 2004 Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:** 1792–1797.

Fauchere, J. L., and V. Pliska, 1983 Hydrophobic parameters pi of amino acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem. **18:** 369–375.

Ferrada, E., and A. Wagner, 2008 Protein robustness promotes evolutionary innovations on large evolutionary time-scales. Proc. R. Soc. B **275:** 1595–1602.

Franzosa, E., and Y. Xia, 2008 Structural perspectives on protein evolution. Ann. Rep. Comp. Chem. **4:** 3–21.

Franzosa, E. A., and Y. Xia, 2009 Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. **26:** 2387–2395.

Godoy-Ruiz, R., R. Perez-Jimenez, B. Ibarra-Molero and J. M. Sanchez-Ruiz, 2004 Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations. J. Mol. Biol. **336:** 313–318.

Goldman, N., J. L. Thorne and D. T. Jones, 1998 Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149:** 445–458.

Guo, H., J. Choe and L. Loeb, 2004 Protein tolerance to random amino acid change. Proc. Natl. Acad. Sci. USA **101:** 9205–9210.

Hamelryck, T., and B. Manderick, 2003 PDB parser and structure class implemented in python. Bioinformatics **19:** 2308–2310.

Holm, L., C. Ouzounis, C. Sander, G. Tuparev and G. Vriend, 1992 A database of protein structure families with common folding motifs. Protein Sci. **1:** 1691–1698.

Kabsch, W., and C. Sander, 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22:** 2577–2637.

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

Lancaster, A. K., J. P. Bardill, H. L. True and J. Masel, 2010 The spontaneous appearance rate of the yeast prion [psi+] and its implications for the evolution of the evolvability properties of the [psi+] system. Genetics **184:** 393–400.

Lau, K., and K. Dill, 1990 Theory for protein mutability and biogenesis. Proc. Nati. Acad. Sci. USA **87:** 638–642.

Lee, Y., T. Zhou, G. G. Tartaglia, M. Vendruscolo and C. O. Wilke, 2010 Translationally optimal codons associate with aggregation-prone sites in proteins. Proteomics **10:** 4163–4171.

Lobkovsky, A., Y. Wolf and E. Koonin, 2009 Universal distribution of protein evolution rates as a consequence of protein folding physics. Proc. Natl. Acad. Sci. USA **107:** 2983–2988.

Lynch, M., W. Sung, K. Morris, N. Coffey and C. R. Landry, 2008 A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. USA **105:** 9272–9277.

Mirny, L. A., and E. I. Shakhnovich, 1999 Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J. Mol. Biol. **291:** 177–196.

Overington, J., D. Donnelly, M. S. Johnson, A. Sali and T. L. Blundell, 1992 Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. **1:** 216–226.

Pokarowski, P., A. Kloczkowski, R. Jernigan, N. Kothari, M. Podarowski *et al.*, 2005 Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins **59:** 49–57.

Porto, M., H. E. Roman, M. Vendruscolo and U. Bastolla, 2004 Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. Mol. Biol. Evol. **22:** 630–638.

Reidhaar-Olson, J. F., and R. T. Sauer, 1988 Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. Science **241:** 53–57.

Schmidt am Busch, M., S. Sedano and T. Simonson, 2010 Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition. PLoS One **5:** e10410.

Sella, G., and A. Hirsh, 2005 The application of statistical physics to evolutionary biology. Proc. Natl. Acad. Sci. USA **102:** 9541–9546.

Shakhnovich, B. E., E. Deeds, C. Delisi and E. Shakhnovich, 2005 Protein structure and evolutionary history determine sequence space topology. Genome Res. **15:** 385–392.

Smith, B., and R. Raines, 2006 Genetic selection for critical residues in ribonucleases. J. Mol. Biol. **362:** 459–478.

Whelan, S., and N. Goldman, 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18:** 691–699.

Wilke, C. O., and D. A. Drummond, 2010 Signatures of protein biophysics in coding sequence evolution. Cur. Opin. Struct. Biol. **20:** 385–389.

Wolff, K., M. Vendruscolo and M. Porto, 2008 Stochastic reconstruction of protein structures from effective connectivity profiles. PMC Biophys. **1:** 5.

Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:** 1586–1591.

Zeldovich, K. B., P. Chen and E. I. Shakhnovich, 2007 Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc. Natl. Acad. Sci. USA **104:** 16152–16157.

Zhou, T., D. A. Drummond and C. O. Wilke, 2008 Contact density affects protein evolutionary rate from bacteria to animals. J. Mol. Evol. **66:** 395–404.

Zhou, T., M. Weems and C. O. Wilke, 2009 Translationally optimal codons associate with structurally sensitive sites in proteins. Mol. Biol. Evol. **26:** 1571–1580.

Communicating editor: L. M. Wahl

APPENDIX

In the main body of this article, we have shown that in the two-allele model the evolutionary rate declines exponentially in s for $s > 1/N_e$. We may ask whether it is possible for $s(r)$ to map r to a sufficiently small range $[s_1, s_2] \subset [0, 1]$ so that $K(r)$ is approximately linear over that range. To this end, take s_1, s_2 with $1/N_e < s_1 < s_2$.

We judge linearity in the range $[s_1, x]$ by the magnitude of the function

$$D(x) = L(x) - K(x), \tag{A1}$$

where $L(x) = K'(s_1)(x - s_1) + K(s_1)$ is the line tangent to K at s_1. We examine the behavior of $D(s_2)$ for a fixed distance $\varepsilon = s_2 - s_1$. Substituting $K(s) \approx 4s\mu N_e e^{-2N_e s}$, $K'(s) \approx 4\mu N_e(1-2N_e s)e^{-2N_e s}$, and $s_2 = s_1 + \varepsilon$, we find for Equation A1,

$$D(s_2) = K'(s_1)\varepsilon + K(s_1) - K(s_2) \tag{A2}$$

$$\approx 4\mu N_e e^{-2N_e s_1}\left[s_1\left(1 + 2N_e\varepsilon - e^{-2N_e\varepsilon}\right) - \varepsilon\left(1 + e^{-2N_e\varepsilon}\right)\right]. \tag{A3}$$

This function decreases with both ε and s_1. Setting $s_2 = s_1 + 1/N_e$ gives us

$$D(s_2) = 4\mu N_e e^{-2N_e s_1}\left[s_1\left(3 - e^{-2}\right) + \frac{1}{N_e(1 + e^{-2})}\right] \tag{A4}$$

$$> 4\mu N_e s_1 e^{-2N_e s_1} = K(s_1). \tag{A5}$$

Wherever the approximations based on $s_1 > 1/N_e$ are valid, a value of ε on the order of $1/N_e$ gives a difference between $K(s_2)$ and $L(s_2)$ larger than the magnitude of $K(s_1)$ itself. Thus, in the two-allele model, the relationship between K and RSA remains highly nonlinear even if the difference in selection pressure at fully exposed and fully buried sites becomes minute.

GENETICS

Supporting Information

http://www.genetics.org/cgi/content/full/genetics.111.128025/DC1

The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution

Duncan C. Ramsey, Michael P. Scherrer, Tong Zhou and Claus O. Wilke

DOI: 10.1534/genetics.111.128025

File S1

Supporting Data

File S1 is available for download as a compressed folder at http://www.genetics.org/cgi/content/full/genetics.111.128025/DC1.

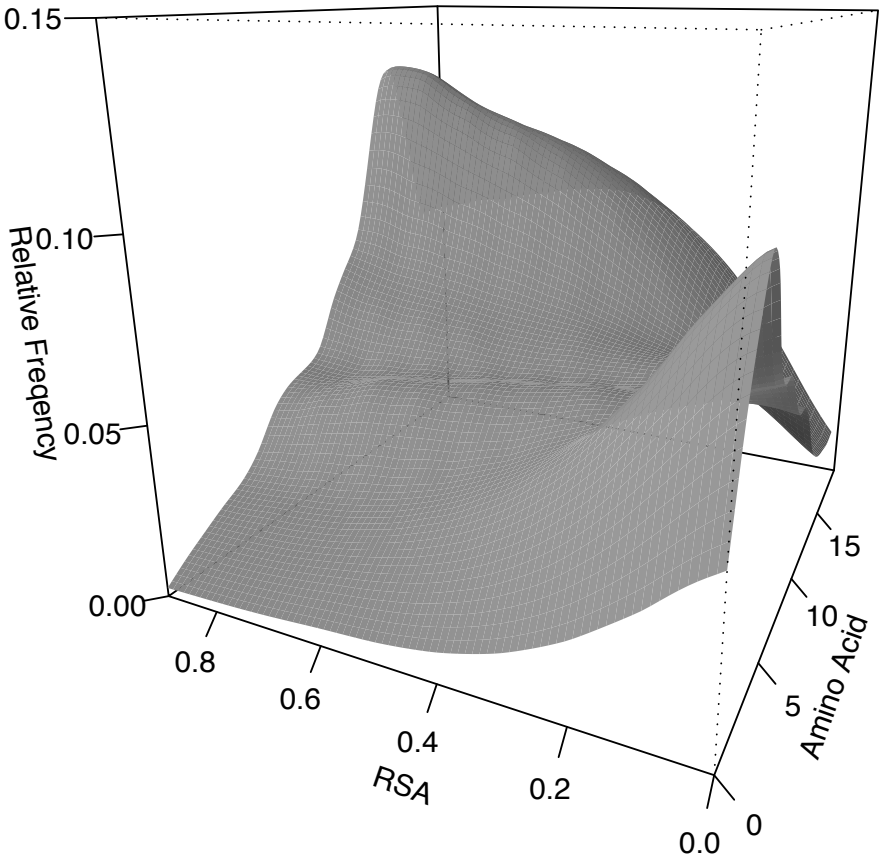

FIGURE S1.—Relative frequencies of amino acids across 525 yeast proteins, binned by RSA and ordered by decreasing hydrophobicity. Cysteine and methionine are omitted. Due to their specialized function, their average frequencies across all bins were 1.5% and 0.6%, respectively.

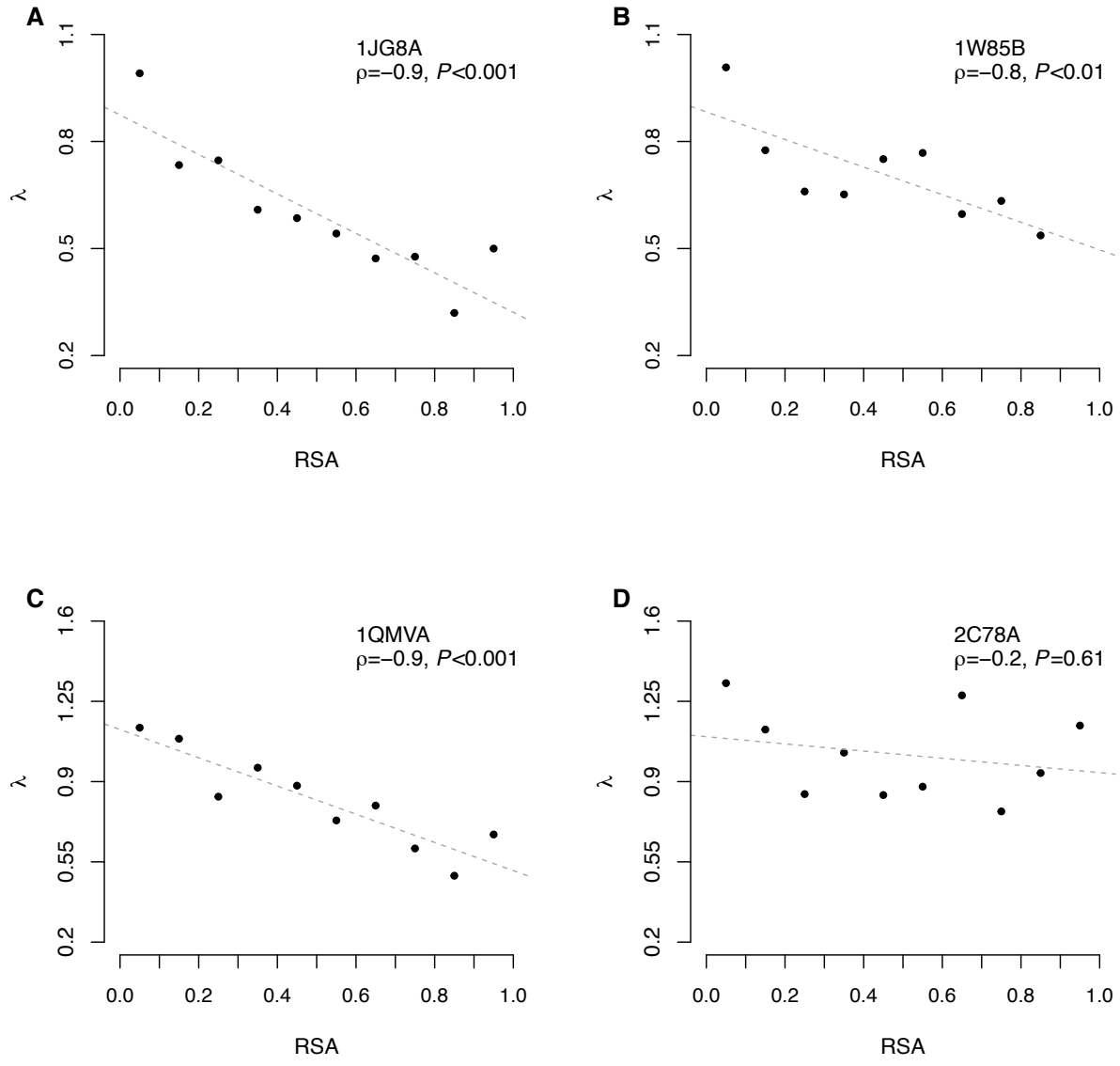

FIGURE S2.—The exponential parameter λ as a function of RSA, for four different protein structures.

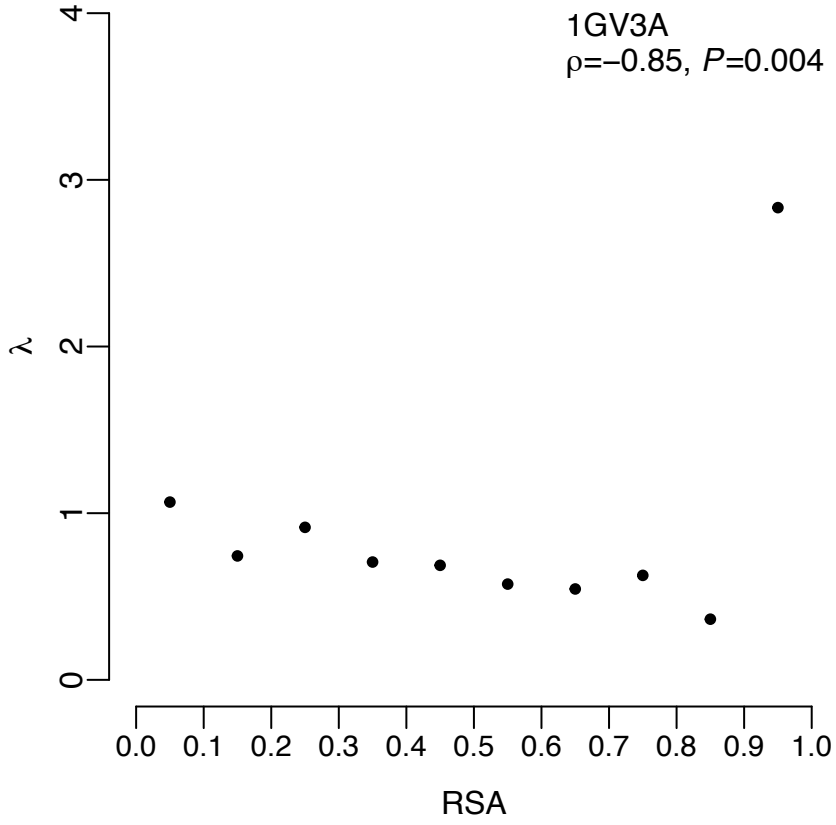

FIGURE S3.—The protein manganese superoxide dismutase (PDF identifier 1GV3, chain A) shows a clear linear decrease of λ with RSA, except for the highest RSA bin. Even though a non-parametric correlation analysis shows a strong negative correlation, a linear model (dashed line) infers a positive slope because of the one outlying data point.

TABLE S1

Fitted constants c_1 and c_2 for protein structures highlighted in Figure 5.

PDB id	c_1	c_2
1GV3A	0.53	0.68
1JG8A	0.90	-0.55
1QMVA	1.16	-0.62
1S4OA	0.90	-0.57
1W85B	0.90	-0.39
2C78A	1.10	-0.16
2GLFA	0.87	-0.37