Predicting the Tolerance of Proteins to Random Amino Acid Substitution

Claus O. Wilke,*§ Jesse D. Bloom,†§ D. Allan Drummond,‡§ and Alpan Raval*¶
*Keck Graduate Institute of Applied Life Sciences, Claremont, California; †Division of Chemistry and Chemical Engineering, ‡Program in Computation and Neural Systems, §Digital Life Laboratory, California Institute of Technology, Pasadena, California; and ¶School of Mathematical Sciences, Claremont Graduate University, Claremont, California

ABSTRACT We have recently proposed a thermodynamic model that predicts the tolerance of proteins to random amino acid substitutions. Here we test this model against extensive simulations with compact lattice proteins, and find that the overall performance of the model is very good. We also derive an approximate analytic expression for the fraction of mutant proteins that fold stably to the native structure, $P_f(m)$, as a function of the number of amino acid substitutions m, and present several methods to estimate the asymptotic behavior of $P_f(m)$ for large m. We test the accuracy of all approximations against our simulation results, and find good overall agreement between the approximations and the simulation measurements.

INTRODUCTION

A protein's tolerance to random amino acid substitutions is of fundamental importance both in protein engineering and molecular evolution. In molecular evolution, a protein's *neutrality*, that is, the fraction of single amino acid substitutions that do not disrupt the protein's function, has a substantial influence on how this protein evolves and accumulates mutations (1–6). In protein engineering, the knowledge of a protein's tolerance to mutations helps one to optimize the mutagenesis conditions in directed protein evolution (7); several groups have characterized experimentally a protein's loss of function under random mutations (8–11).

Protein mutagenesis studies suggest that a large fraction of deleterious amino acid substitutions disrupt a protein's structure rather than specifically affecting functional residues (12–14). Therefore, the fraction of substitutions that disrupt a protein's structure is a reasonable lower bound to the fraction of substitutions that will disrupt a protein's function.

We (8) have recently proposed a thermodynamic model that allows one to calculate the probability $P_f(m)$ with which a protein retains its structure after m amino acid substitutions. This model uses as input the distribution of free energy changes $\Delta\Delta G$ for individual amino acid substitutions. It is based on the idea that the free energy change caused by one amino acid substitution is independent of the change caused by another such substitution, and that the protein continues to fold correctly as long as its free energy of folding remains below some threshold level. If the protein's free energy of folding is initially a distance C from the threshold, then the fraction of sequences with m substitutions that still fold correctly is given by the fraction of sums $\sum_{i=1}^{m} X_i$ that are less than C, where the X_i are independent, identically distributed random variables taken from the $\Delta\Delta G$ distribution. For a small set of both simulated lattice proteins and real proteins, we (8) have shown that this model has excellent predictive power. Here, we are interested in three questions:

1. How well does this model hold up for a more extensive data set of lattice proteins?
2. Can one make general statements about how $P_f(m)$ behaves for large m, and how is this behavior influenced by the $\Delta\Delta G$ distribution?
3. How can the neutrality be calculated from the distribution of $\Delta\Delta G$ values?

METHODS

Lattice protein simulations

We implemented a maximally compact, 5×5 two-dimensional square lattice model, as previously described (15,5). In short, we folded simulated polypeptide chains of length $L = 25$ residues into a maximally compact structure, representing one of the 1081 possible (16) self-avoiding compact walks of length 25 not related by rotational or reflection symmetry. (We neglected the vanishingly small fraction of palindromic sequences.) We used an alphabet of 20 amino acids, and calculated the contact energies between nonbonded neighboring residues according to Table 3 of Miyazawa and Jernigan (17). We calculated a lattice protein's free energy of folding ΔG_f as described by Taverna and Goldstein (15), and considered the protein to be stably folded if ΔG_f was below a cutoff ΔG_{cut}. We carried out all analyses for three different cutoffs, $\Delta G_{cut} = -4.0$ kcal/mol, -5.0 kcal/mol, and -6.0 kcal/mol.

We first analyzed a dataset of 300 randomly chosen sequences, 100 at each cutoff. We generated these sequences in the following way: First, we generated random sequences and tried to fold them. We kept all those sequences whose free energy of folding was below $\Delta G_{cut} = -4.0$ kcal/mol, and whose native conformation was different from the native conformations of all stably folding sequences we had encountered so far. We repeated this procedure until we had 100 sequences that could stably fold into 100 unique conformations at $\Delta G_{cut} = -4.0$ kcal/mol. For the remaining two cutoffs, we used hill climbing and subsequent neutral evolution to obtain, at each cutoff, 100 additional sequences that could stably fold into the same 100 conformations as the original sequences. Under hill climbing, we repeatedly mutated a sequence, and accepted all mutations that increased the protein's stability without changing the native conformation. Under neutral evolution, we repeatedly mutated a sequence, and accepted all mutations that did not destabilize the protein beyond the chosen cutoff and did not change the

Submitted February 28, 2005, and accepted for publication August 16, 2005.

Address reprint requests to C. O. Wilke at his present address, Section of Integrative Biology, University of Texas at Austin, Texas. E-mail: cwilke@mail.utexas.edu.

0006-3495/05/12/3714/07 $2.00

doi: 10.1529/biophysj.105.062125

native conformation. We always repeated neutral evolution until we had accepted 1000 mutations.

For all 300 sequences, we estimated $P_f(m)$, the fraction of mutant proteins that fold stably to the original native conformation after m amino-acid substitutions, by randomly sampling mutants according to the following procedure: We carried out all single-point mutations, and sampled 10^4, 5×10^4,10^5,...10^7 multiple-point mutations for $m = 2,3,4,\ldots,8$. We then calculated $P_f(m)$ by dividing the number of correctly folded sequences that we found at the given mutational distance m by the total number of mutants we tried at that distance. We defined a protein as correctly folded if its minimum free energy was below the chosen cutoff ΔG_{cut} and if its native conformation was identical to that of the starting sequence. In the vast majority of these 300 replicates, we found between several hundred and several thousand correctly folded proteins at each mutational distance m. Consequently, our estimate for $P_f(m)$ in lattice proteins is highly accurate.

We measured the $\Delta\Delta G$ distribution of each of the 300 sequences by carrying out all possible single-point mutations, and then calculating the differences between the minimum free energy of the original sequence and the mutated sequences.

We calculated the prediction for $P_f(m)$ from the $\Delta\Delta G$ distribution as described (8). In short, we first binned the $\Delta\Delta G$ distribution into bins of width 0.01 kcal/mol, and then calculated the m-fold convolution of this binned distribution using the fast Fourier transform of the software package *R*, version 1.9.1 (18). Finally, we numerically integrated the convolved distribution from $-\infty$ to C to obtain $P_f(m)$.

We carried out a second set of simulations to determine the influence of the starting sequence on the neutrality $\langle\nu\rangle$. We selected the sequences of 10 representative conformations (among the 100 unique conformations of the first data set), and generated, through neutral evolution as before, for each conformation at each cutoff nine additional sequences folding stably into this conformation. We measured then both $P_f(m)$ and the $\Delta\Delta G$ distribution for these additional 270 sequences as described above.

Calculation of $\langle\nu\rangle$

$P_f(m)$ decays approximately as $\langle\nu\rangle^m$ for large m. We estimated $\langle\nu\rangle$ from the measured $P_f(m)$ by carrying out a linear regression of $\ln P_f(m)$ versus m, where we restricted the range of m from 4 to 8 to capture the asymptotic behavior of $P_f(m)$. The neutrality $\langle\nu\rangle$ followed then as $\langle\nu\rangle = e^a$, where a is the slope of the regression line.

We also calculated $\langle\nu\rangle$ in the context of a number of approximation schemes, described in Appendices A–D, and summarized in Results, below. For the Cramér approximation (Appendix B), we numerically minimized the moment-generating function $\phi(t)$ of the $\Delta\Delta G$ distribution. Let $\{\Delta\Delta G_i\}$ be the set of free energy changes caused by all single point mutations. Then, $\phi(t) = \sum_i e^{\Delta\Delta G_i t}$, and its derivative $\phi'(t) = \sum_i \Delta\Delta G_i e^{\Delta\Delta G_i t}$. We numerically found the value t^* at which $\phi'(t^*) = 0$, and then set $\langle\nu\rangle = \phi(t^*)$.

For the Markov chain approximation (Appendix C), we constructed the matrix W_{ij} using bins of width 0.015 kcal/mol, and spanning a range of 25.0 kcal/mol, from ΔG_{cut} to ΔG_{cut} −25.0 kcal/mol. We calculated the largest eigenvalue of this matrix by repeatedly multiplying W_{ij} to a vector (with all components initially set to one), and then renormalizing the vector to unit length, until the vector had converged to the dominant eigenvector of W_{ij}. We then obtained the quantity $\langle\nu\rangle$ from the change in length in the dominant eigenvector of W_{ij} after a single multiplication with W_{ij}.

RESULTS

First, we assess how well our method to predict $P_f(m)$ works in a large data set. We (8) have previously studied only a handful of noncompact lattice proteins and three real proteins. Overall, we find that the method works very well for the compact lattice proteins we study here. Fig. 1 shows several typical examples. In many cases, we find that the prediction of $P_f(m)$ is highly accurate up to $m = 8$, which is the largest number of mutations we consider (Fig. 1, *A–D*). In those cases that show some discrepancy between the predicted and the measured $P_f(m)$, we typically find that the prediction works well up to $m = 3$ or 4, but starts to deviate from the measured results for larger m. There is no clear tendency toward either over- or underestimation of the measured results by the prediction (Fig. 1, *E–H*). Note that our data set covers a wide range of different conformations, as all 100 sequences at a given cutoff fold into a unique conformation.

We can quantify the performance of our prediction using the root-mean-squared (RMS) deviation of the log-transformed $P_f(m)$. Let $P_f^{pred}(m)$ be the predicted fraction of mutants that fold correctly, and $P_f(m)$ the corresponding measured value. Then, we define the logarithmic RMS deviation ρ as

$$\rho = \left(\sum_{m=1}^{8} \left[\ln P_f(m) - \ln P_f^{pred}(m) \right]^2 \right)^{1/2}. \quad (1)$$

For our data, cases in which the prediction agrees well with the measured $P_f(m)$ have RMS values well below 1.0,

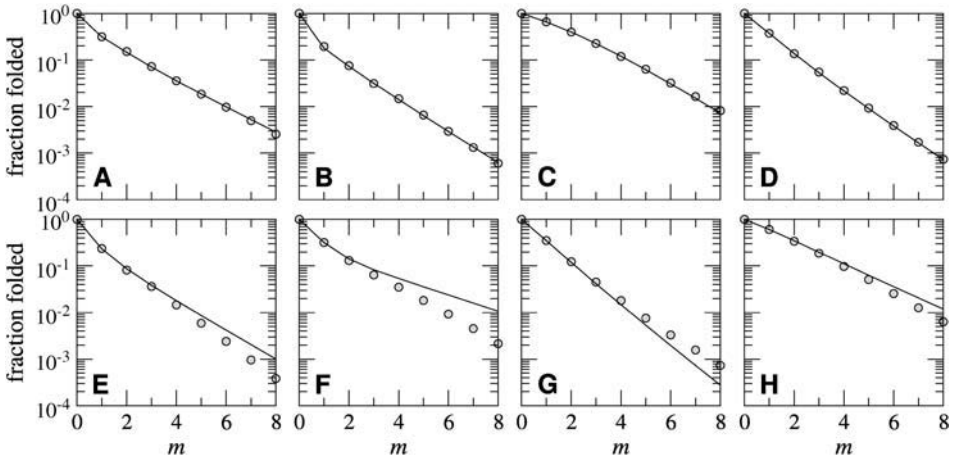

FIGURE 1 Fraction of correctly folded mutants P_f as a function of the number of mutations m, for eight lattice proteins that stably fold at $\Delta G_{cut} = -5.0$ kcal/mol. Points indicate measurement results, and solid lines indicate the prediction derived from the m-fold convolution of the $\Delta\Delta G$ distribution. *A–D* show cases for which the prediction works excellently, and *E–H* show cases for which there is some disagreement between the prediction and the measured values. The RMS values for these eight cases are (*left* to *right* and *top* to *bottom*): 0.127, 0.112, 0.157, 0.065, 1.423, 2.407, 1.406, and 0.902.

whereas cases in which the prediction shows some clear deviation have RMS values of ~1.0 or higher (Fig. 1). When we consider all 300 replicates, we find that the majority of the cases have an RMS value below 1.0, and only rarely does the RMS value exceed 2.0 (Fig. 2). There seems to be a slight tendency for the RMS value to increase as the cutoff value becomes more stringent (i.e., from $\Delta G_{cut} = -4$ kcal/mol to – 6 kcal/mol).

Next, we are interested in asymptotic expressions of $P_f(m)$ for small and large m. For small m, we can approximate $P_f(m)$ using the Edgeworth expansion (Appendix A). The Edgeworth expansion provides correction terms to the central limit theorem for finite sums of random variables. These correction terms take into account successively higher moments of the $\Delta\Delta G$ distribution. Fig. 3 shows how the Edgeworth expansion provides an increasingly more accurate approximation of $P_f(m)$ as higher-order terms are included. However, whereas in some cases the Edgeworth expansion works very well with only three additional moments beyond mean and variance (Fig. 3 *A*), in other cases the Edgeworth expansion deviates significantly from $P_f(m)$ in all orders we have considered (Fig. 3 *B*). Furthermore, because the Edgeworth expansion leads to a normal distribution function multiplied by a polynomial (Eq. 3), it must inevitably break down as m becomes large.

For large m, empirical observations show that $P_f(m)$ decays approximately as $\langle\nu\rangle^m$ ((8–10) and Fig. 1). The value $\langle\nu\rangle$ can vary substantially among sequences, but generally tends to increase with the cutoff (Fig. 4). We can interpret $\langle\nu\rangle$ intuitively as the average neutrality of all sequences that stably fold into the given structure. We give a formal argument for this interpretation in Appendix C. An exponential decay of the form $P_f(m) \approx \langle\nu\rangle^m$ follows from the Gaussian term in the Edgeworth expansion (Appendix A). However, the value of $\langle\nu\rangle$ predicted by this term is not very accurate (data not shown). The Gaussian approximation fails because, for large m, $P_f(m)$ is extremely sensitive to small deviations from normality in the tail of the m-fold convolved $\Delta\Delta G$ distribution.

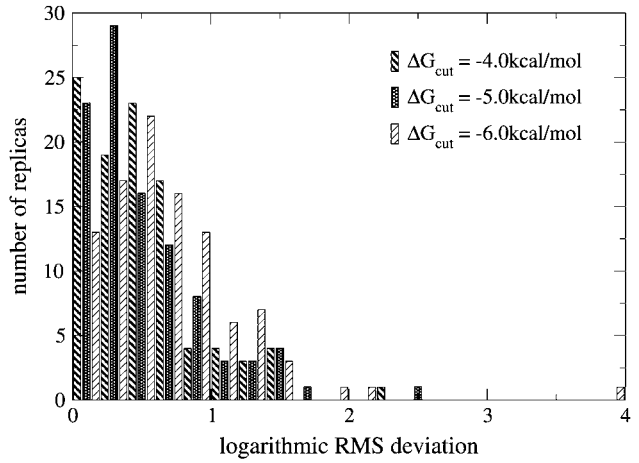

FIGURE 2 Histogram of RMS values for 100 randomly chosen sequences each, at three cutoff levels.

Numerically, we can estimate $\langle\nu\rangle$ by first calculating the prediction for $P_f(m)$ using the m-fold convolution of the $\Delta\Delta G$ distribution, and then obtaining $\langle\nu\rangle$ from a log-linear regression in the same way in which we estimate it from the measured $P_f(m)$ (see Calculation of $\langle\nu\rangle$, above). In the following, we refer to this method as the *convolution method*. The convolution method does not generate any new insight into what determines the value of $\langle\nu\rangle$, but it serves as a useful test case. First, by comparing for a large set of proteins the measured $\langle\nu\rangle$ to the $\langle\nu\rangle$ predicted by the convolution method, we obtain an overall estimate of how well our model performs. Second, the convolution method is the correct benchmark for all other methods of estimating $\langle\nu\rangle$: Because any deviation between the prediction from the convolution method and the measured $\langle\nu\rangle$ is an inherent shortcoming of our model, we can only expect that any approximate method to estimate $\langle\nu\rangle$ will work at most as well as the convolution method, and will generally perform worse. Fig. 5 *A* shows that the $\langle\nu\rangle$ predicted by the convolution method correlates strongly with the measured (overall R^2 for all 300 data points $R^2 = 0.789$, $p < 10^{-15}$), in agreement with our earlier observation that, overall, our model works very well.

A straightforward method to predict $\langle\nu\rangle$ from the $\Delta\Delta G$ distribution follows from large-deviation probability theory. Cramér's theorem implies that $P_f(m)$ must decay exponentially, and implies that $\langle\nu\rangle$ is approximately given by the unique minimum of the moment-generating function of the $\Delta\Delta G$ distribution (Appendix B). In Fig. 5 *B*, we compare the $\langle\nu\rangle$ predicted by the Cramér approximation to the measured $\langle\nu\rangle$. We see that the Cramér approximation performs almost as well as the convolution method. The correlation between the $\langle\nu\rangle$ values predicted according to the convolution method and the Cramér approximation is very strong (overall R^2 for all 300 data points $R^2 = 0.971$, $p < 10^{-15}$).

The intuitive explanation for why $P_f(m)$ decays approximately as $\langle\nu\rangle^m$ is that each correctly folded sequence has, on average, a fraction $\langle\nu\rangle$ of correctly folded single-point neighbors, so that with each mutational step the total $P_f(m)$ is reduced by a factor of $\langle\nu\rangle$. We can make this reasoning more precise with the Markov chain approximation. The Markov chain approximation is based on the assumption that single-point mutants to sequences at distance m that do not fold correctly do not contribute to $P_f(m + 1)$. With this assumption, $\langle\nu\rangle$ turns out to be the largest eigenvalue of a matrix W_{ij} that contains the transition probabilities from any stable protein to any other stable protein under single-point mutations (Appendix C). We do not present results from the Markov chain approximation in Fig. 5, because they are very similar to those found with the Cramér approximation (overall R^2 for all 300 data points $R^2 = 0.9992$, $p < 10^{-15}$). However, the $\langle\nu\rangle$ values predicted by the Markov chain approximation

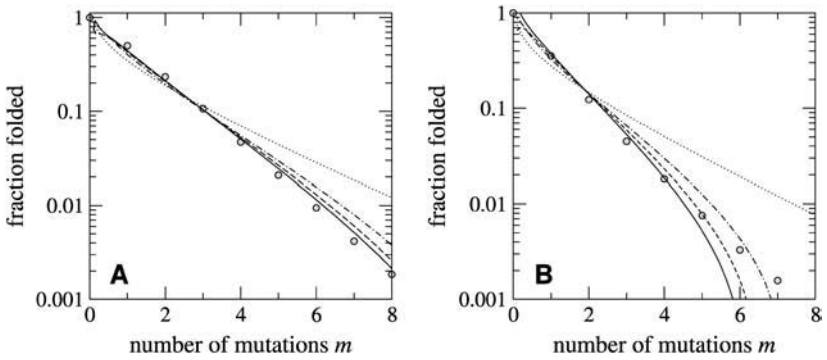

FIGURE 3 Prediction of $P_f(m)$ according to the Edgeworth expansion. Points indicate the measured $P_f(m)$, and lines indicate the Edgeworth expansion to various orders (*dotted lines*, normal term only; *dot-dashed lines*, normal term plus first-order corrections; *dashed lines*, normal term plus first- and second-order corrections; and *solid lines*, normal term plus first-, second-, and third-order corrections). (*A*) Example of a case where the expansion works well up to $m = 8$. (*B*) Example of a case where the expansion works poorly.

tend to be slightly smaller than those predicted by the Cramér approximation, the reason being that the Markov chain approximation neglects mutations that stabilize previously unstable sequences (Appendix C).

The last method we consider is the mean-field approximation. The mean-field approximation is based on the idea that we can replace the distribution of proteins with different neutralities by a single protein with an effective neutrality that equals $\langle\nu\rangle$, and is extremely simple to calculate (Appendix D). Fig. 5 *C* shows that the mean-field approximation performs only slightly worse than the Cramér approximation. The correlation between the $\langle\nu\rangle$ values predicted from the convolution method and the mean-field approximation is also strong (overall R^2 for all 300 data points $R^2 = 0.939$, $p < 10^{-15}$).

Finally, we have generated an additional data set of 10×10 sequences that fold into the same structure, to assess to what extent $\langle\nu\rangle$ depends on the initial sequence or the structure. We find that although there is some spread in the estimated $\langle\nu\rangle$ for different sequences folded into the same structure, the $\langle\nu\rangle$ values for the different starting sequences clearly cluster around a mean value $\overline{\langle\nu\rangle}$ that is determined by the structure. Fig. 6 shows data for a representative five of the 10 structures we considered for this additional data set. We carried out a pairwise *t*-test for all 45 possible pairings of the 10 structures, at each cutoff, and found that (after applying the false-discovery-rate correction for multiple testing (19)) only 12, 9, and 5 of the 45 pairs at cutoffs $\Delta G_{cut} = -4.0$ kcal/mol, -5.0 kcal/mol, and -6.0 kcal/mol do not have a statistically significant (at the 5% level) difference in $\overline{\langle\nu\rangle}$.

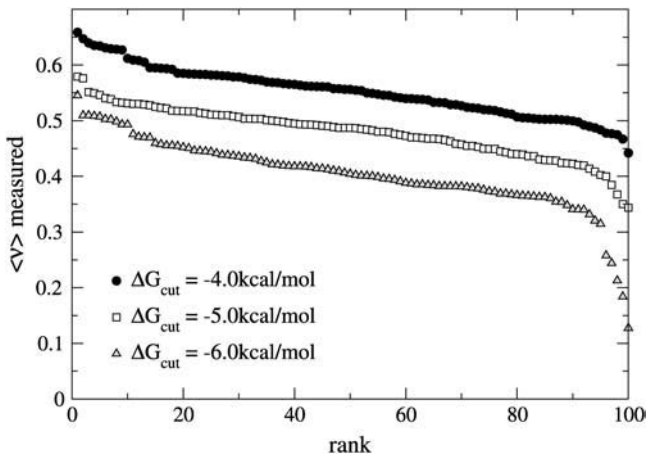

FIGURE 4 Asymptotic neutralities $\langle\nu\rangle$ (measured) at the three different cutoffs, sorted by magnitude and plotted against their rank.

DISCUSSION

We have extensively tested a model introduced earlier to describe and explain the tolerance of proteins to amino-acid substitutions (8). These tests were performed on an array of 100 structures and three cutoff levels. The model performs well across this data set, which gives strong support for the model's central claims, its generality, and its theoretical underpinnings. The predicted emergence of an exponential decline in the $P_f(m)$ that is parameterized by the mean neutrality $\langle\nu\rangle$ is both observed and estimated by several independent methods, and the preliminary finding that $\langle\nu\rangle$ is principally a structural property receives computational support through tests across 10 structures. Using a Markov chain method, we also explain why the rate of the asymptotic decay of $P_f(m)$, as measured by $\langle\nu\rangle$, is in fact related to the average neutrality of all sequences that can stably fold into the native conformation.

For computational efficiency, we have used maximally compact two-dimensional lattice proteins (with the full amino-acid alphabet). Compact lattice proteins have the drawback that the additional constraint of maximal compactness allows many more sequences to stably fold than otherwise would; also, noncompact lattice proteins rarely fold into maximally compact formations (20, 21). However, in previous work (8), we had tested the model against a small set of two-dimensional noncompact lattice proteins, as well as two real proteins, and found the model to perform well in these cases. It therefore seems unlikely that the results that we report here are artifacts of the additional constraint of maximal compactness. Likewise, three-dimensional lattice proteins have substantially more conformations at the same sequence length than two-dimensional lattice proteins, and

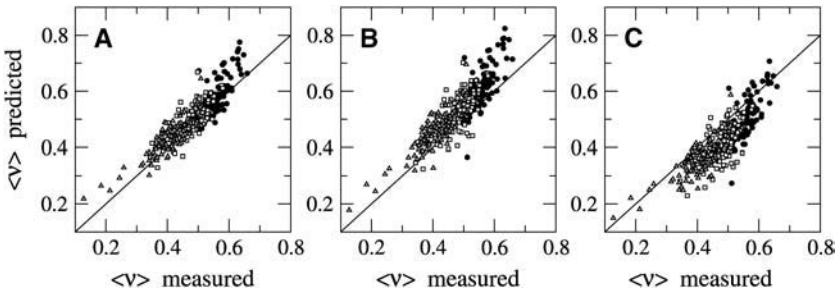

FIGURE 5 Prediction for $\langle\nu\rangle$ according to three different methods. The solid lines indicate perfect agreement between predicted and measured $\langle\nu\rangle$. Symbols indicate $\Delta G_{cut} = -4.0$ kcal/mol (*solid circles*), $\Delta G_{cut} = -5.0$ kcal/mol (*open squares*), and $\Delta G_{cut} = -6.0$ kcal/mol (*shaded triangles*). (*A*) Convolution method. Correlations with measured $\langle\nu\rangle$, in order of increasing absolute cutoff value: $R^2 = 0.642$, $R^2 = 0.540$, and $R^2 = 0.757$. Overall correlation, $R^2 = 0.789$ (all $p < 10^{-15}$). (*B*) Cramér approximation. Correlations with measured $\langle\nu\rangle$, in order of increasing absolute cutoff value: $R^2 = 0.584$, $R^2 = 0.467$, and $R^2 = 0.747$. Overall correlation, $R^2 = 0.751$ (all $p < 10^{-14}$). (*C*) Mean-field approximation. Correlations with measured $\langle\nu\rangle$, in order of increasing absolute cutoff value: $R^2 = 0.590$, $R^2 = 0.410$, and $R^2 = 0.650$. Overall correlation, $R^2 = 0.716$ (all $p < 10^{-12}$).

our model could, in principle, break down in three dimensions. We have no specific reason to believe that our model would perform substantially worse for three-dimensional lattice proteins than for two-dimensional lattice proteins, but this hypothesis remains to be tested.

A key advantage of our model is its extreme simplicity. Our finding that $\langle\nu\rangle$ can be trivially computed with reasonable accuracy using either a mean-field approximation or a generating function approach that extends the model's utility. Our finding that the Gaussian term in the Edgeworth expansion cannot accurately describe the data suggests that a Gaussian approximation for the initial $\Delta\Delta G$ distribution is simply not adequate for the estimation of $\langle\nu\rangle$. Thus our model, although simple, is sensitive to the detailed form of the $\Delta\Delta G$ distribution, rather than just its mean and variance.

Whether these results extend to an equally broad class of naturally occurring proteins remains an open question. A useful feature of our model is that it depends, in a direct and relatively simple manner, on the distribution of the $\Delta\Delta G$ values, which are routinely measured in natural proteins and can be computationally estimated from crystal structures. In general, we do not know the difference C between the native stability of proteins and their minimum free energy cutoff. However, the existence of a cutoff is indicated by diverse observations such as the abundance of temperature-sensitive mutations and the steep (exponential) dependence on stability of the folded and unfolded protein concentrations at equilibrium. We do not know whether the cutoff is consistent across proteins or varies, like $\langle\nu\rangle$, from structure to structure.

An important practical implication of our model is that the fraction of mutant proteins retaining fold can be increased in a predictable fashion by modest increases in wild-type protein stability. Mutagenesis experiments aimed at discovering functionally improved proteins may thus have stability-dependent optimal mutation rates (7) which, at least in principle, may be estimated using our model. Our results here offer strong support to the suggestion (8) that stability is a critical, but generally overlooked, parameter in directed evolution.

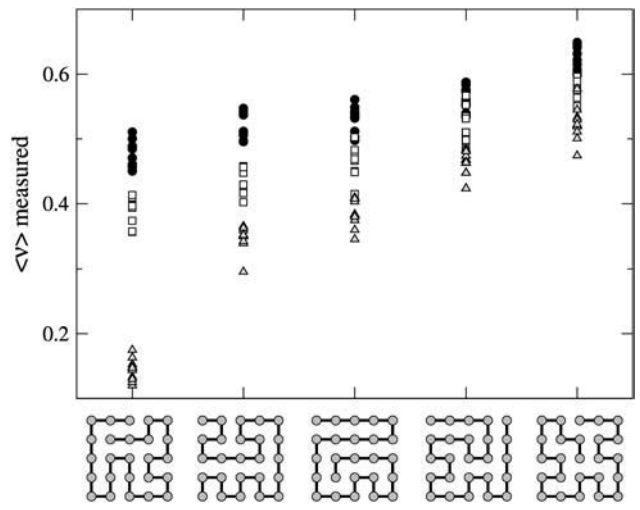

FIGURE 6 Asymptotic neutralities $\langle\nu\rangle$ (measured) for five different structures. Symbols indicate $\Delta G_{cut} = -4.0$ kcal/mol (*solid circles*), $\Delta G_{cut} = -5.0$ kcal/mol (*open squares*), and $\Delta G_{cut} = -6.0$ kcal/mol (*shaded triangles*).

APPENDIX A: EDGEWORTH EXPANSION

We wish to estimate the probability $P_f(m) = \text{Prob}\left(\sum_{i=1}^{m} X_i < C\right)$, where X_i are independent, identically distributed random variables distributed according to the $\Delta\Delta G$ distribution, and C is the distance to the free-energy cutoff beyond which the protein does not stably fold. It is convenient to introduce the standardized random variable $Z = (S_m - m\mu)/(\sqrt{m}\sigma)$, where $S_m = \sum_{i=1}^{m} X_i$, and μ and σ are the mean and standard deviation of the $\Delta\Delta G$ distribution, respectively. Let κ_n be the n^{th} cumulant (see Appendix E) of the $\Delta\Delta G$ distribution. We define $\lambda_n = \kappa_n/\sigma^n$, and write the standard normal distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{2}$$

Then, the Edgeworth expansion of the distribution function $F(z)$ of the random variable Z is given by (22)

$$F(z) = \Phi(z) - \frac{1}{3!}\frac{\lambda_3}{m^{1/2}}\Phi^{(3)}(z) + \frac{1}{4!}\frac{\lambda_4}{m}\Phi^{(4)}(z) + \frac{10}{6!}\frac{\lambda_3^2}{m}\Phi^{(6)}(z) - \frac{1}{5!}\frac{\lambda_5}{m^{3/2}}\Phi^{(5)}(z) - \frac{35}{7!}\frac{\lambda_3\lambda_4}{m^{3/2}}\Phi^{(7)}(z) - \frac{280}{9!}\frac{\lambda_3^3}{m^{3/2}}\Phi^{(9)}(z) + \cdots. \tag{3}$$

An efficient algorithm to generate higher-order terms of the expansion has been presented by Blinnikov and Moessner (23). From $F(z)$, we obtain $P_f(m)$ via

$$P_f(m) = F\left(\frac{C - m\mu}{\sqrt{m\sigma^2}}\right). \quad (4)$$

The zero[th]-order term of the Edgeworth expansion (the Gaussian term, which takes into account only the mean and variance of the $\Delta\Delta G$ distribution) is

$$P_f(m) = \frac{1}{2}\mathrm{erfc}\left(-\frac{C - m\mu}{\sqrt{2m\sigma^2}}\right), \quad (5)$$

where erfc(t) is the complementary error function. This expression predicts for large m that $P_f(m) \approx \langle\nu\rangle^m$ with $\langle\nu\rangle = \exp[-\mu^2/(2\sigma^2)]$.

APPENDIX B: CRAMÉR APPROXIMATION

We can calculate the asymptotic behavior of $P_f(m)$ for large m from large-deviation theory. According to the central limit theorem, for large m the sum $S_m = \sum_{i=1}^{m} X_i$ (as introduced in Appendix A) is approximately normally distributed with mean $m\mu$ and variance $m\sigma^2$, where μ and σ^2 are the mean and variance of the $\Delta\Delta G$ distribution. The probability $P_f(m) = \mathrm{Prob}\left(\sum_{i=1}^{m} X_i < C\right)$ is therefore a tail probability that becomes vanishingly small as m approaches infinity. Cramér's theorem (24) for large deviation probability states that, for $a < \mu$,

$$\lim_{m\to\infty} \frac{1}{m} \ln \mathrm{Prob}(S_m/m \le a) = \ln\phi(t^*) - at^*, \quad (6)$$

where $\phi(t)$ is the moment-generating function of the distribution of X_i, and t^* is the value of t at which $\phi(t) - at$ attains its minimum.

Cramér's theorem can be used as a basis for approximating the asymptotic behavior of Prob($S_m/m \le a$), namely, for large m,

$$\mathrm{Prob}(S_m/m \le a) \approx \left(e^{-at^*}\phi(t^*)\right)^m. \quad (7)$$

The theorem therefore gives a theoretical justification for the exponential decay of tail probabilities. For the case of interest, we have $a = C/m$, which is small when m is large. For large m, we may therefore consider an approximation to $P_f(m)$ of the form

$$P_f(m) = \mathrm{Prob}(S_m/m \le C/m) \approx \mathrm{Prob}(S_m/m \le 0) \approx \phi(t^*)^m, \quad (8)$$

and therefore estimate the average neutrality as $\langle\nu\rangle \approx \phi(t^*)$.

Further refinements to Cramér's theorem, especially in the context of placing bounds on tail probabilities for finite m, have been the subject of recent advances in large deviation probability theory (see, for example, Hahn and Klass (25) and references therein) and may be used to obtain more accurate estimates. For our purposes, Cramér's theorem gives a simple and reasonably accurate estimate of $P_f(m)$.

APPENDIX C: MARKOV CHAIN APPROXIMATION

An alternative method to estimate the asymptotic slope $\langle\nu\rangle$ of $P_f(m)$ is based on calculating the steady-state solution of a suitable Markov process. First, we subdivide the range of free energies of folding into discrete bins of width b. We number the bins consecutively and in such a way that all bins with index $i \ge 0$ represent stable proteins, and all other bins represent unstable proteins. Now, let $p_i(m)$ be the fraction of proteins at mutation distance m in bin i. Clearly, we have $P_f(m) = \sum_{i=0}^{\infty} p_i(m)$. Next, we introduce the matrix M_{ij}, which gives the probability that a single mutation to a protein in bin j moves that protein into bin i. (Note that under the assumptions of our theory, M_{ij} does not depend on m, and furthermore depends only on the difference $i - j$, but not on the specific values of i or j. The first assumption is necessary for the development of the Markov approximation; the second assumption of stationarity of the transition matrix could be, in principle, relaxed.) Then, we can write $P_f(m + 1)$ as

$$P_f(m+1) = \sum_{i=0}^{\infty}\sum_{j=-\infty}^{\infty} M_{ij}p_j(m). \quad (9)$$

Our goal is to express the right-hand side in terms of $P_f(m)$, so that we obtain a recursion relation for $P_f(m)$. Unfortunately, the second sum on the right-hand side spans all values of j, positive as well as negative, whereas $P_f(m)$ contains only information about $p_j(m)$ with positive index j. Therefore, we now make the approximation that

$$\sum_{j=-\infty}^{\infty} M_{ij}p_j(m) \approx \sum_{j=0}^{\infty} M_{ij}p_j(m) \quad \text{for} \quad i \ge 0. \quad (10)$$

This approximation is based on the assumption that mutations that stabilize an unstable protein are rare in comparison to mutations that do not destabilize a stable protein, and is the main difference between the Markov chain approximation and the m-fold convolution of the $\Delta\Delta G$ distribution. From here on, we will refer to the submatrix of M_{ij} for which the indices i and j are non-negative as W_{ij}. The distinction between M_{ij} and W_{ij} will become important when we discuss eigenvectors and eigenvalues of W_{ij}.

We can interpret $\sum_{i=0}^{\infty} W_{ij}$ as the neutrality of a protein in bin j (for $j \ge 0$), and the average neutrality of all proteins at distance m is given by

$$\langle\nu(m)\rangle = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} W_{ij}p_j(m) \Big/ \sum_{j=0}^{\infty} p_j(m). \quad (11)$$

Since we can replace $\sum_{j=0}^{\infty} p_j(m)$ with $P_f(m)$, we find that $P_f(m + 1)$ and $P_f(m)$ are related to each other via

$$P_f(m+1) = \langle\nu(m)\rangle P_f(m), \quad (12)$$

and, assuming that $\langle\nu(m)\rangle$ approaches a limiting value $\langle\nu\rangle$ for large m, we have

$$P_f(m+1) = \langle\nu\rangle P_f(m) \quad \text{for large } m. \quad (13)$$

This equation implies that for large m, $P_f(m)$ is proportional to $\langle\nu\rangle^m$.

From $p_i(m+1) = \sum_{j=0}^{\infty} W_{ij}p_j(m)$, we see that for large m, the p_i are proportional to the dominant eigenvector of W_{ij}, by virtue of the Frobenius-Perron theorem (26). (The Frobenius-Perron theorem holds if W_{ij} is primitive—the case whenever there is a path of mutations that leads from any bin i to any other bin j, and $W_{ii} > 0$ for at least one i.) Furthermore, Eq. 11 implies

$$\langle\nu\rangle p_i^* = \sum_{j=0}^{\infty} W_{ij}\, p_j^*, \quad (14)$$

where p_i^* is the dominant eigenvector of W_{ij}. Consequently, $\langle\nu\rangle$ corresponds to the dominant eigenvalue of W_{ij}.

APPENDIX D: MEAN-FIELD APPROXIMATION

A third method to calculate $\langle\nu\rangle$ is the mean-field approximation. The idea of this approximation is that we can replace the distribution of proteins of different stabilities with a single protein of typical stability. The neutrality of this protein should correspond to the average neutrality of all stable proteins. We choose the stability of this protein such that its free energy of folding is identical to the average free energy of folding of all possible single-point mutants that fold correctly. In other words, the average change in free energy of a single mutation that does not destroy the protein's ability to fold is zero.

The neutrality of this protein is then the fraction of mutations that cause a change in free energy below a certain cutoff, where the cutoff is chosen such that the average change in free energy for all mutations below the cutoff is as close as possible to zero. We can formalize this condition as follows. Assume that the set $\{\Delta\Delta G_i\}$ contains the free-energy changes caused by all possible single-point mutations (of which there are n), and that the set is ordered such that $\Delta\Delta G_i < \Delta\Delta G_{i+1}$ for all i. Then, we have

$$\langle \nu \rangle \approx k/n, \quad \text{where} \quad k = \min_{j \in \{1,\ldots,n\}} \left\{ j \left| \sum_{i=1}^{j} \Delta\Delta G_i \geq 0 \right. \right\}. \quad (15)$$

APPENDIX E: UNBIASED ESTIMATORS OF CUMULANTS

Let $\{X_1, \ldots, X_n\}$ be a set of n measurements, and define

$$S_k = \sum_{i=1}^{n} X_i^k. \quad (16)$$

According to Dressel (27), the following are unbiased estimators for the first five cumulants κ_1–κ_5 (note that κ_1 is the sample average, and κ_2 is the sample variance):

$$\kappa_1 = S_1/n, \quad (17)$$

$$\kappa_2 = (nS_2 - S_1^2)/[n(n-1)], \quad (18)$$

$$\kappa_3 = (2S_1^3 - 3nS_1S_2 + n^2S_3)/[n(n-1)(n-2)], \quad (19)$$

$$\kappa_4 = [-6S_1^4 + 12nS_1^2S_2 - 3n(n-1)S_2^2 - 4n(n+1)S_1S_3 + n^2(n+1)S_4]/[n(n-1)(n-2)(n-3)], \quad (20)$$

$$\kappa_5 = [24S_1^5 - 60nS_1^3S_2 + 30n(n-1)S_1S_2^2 + 20n(n+2)S_1^2S_3 - 10n^2(n-1)S_2S_3 - 5n^2(n+5)S_1S_4 + n^3(n+5)S_5]/[n(n-1)(n-2)(n-3)(n-4)]. \quad (21)$$

This work was supported by National Institutes of Health NRSA No. 5 T32 MH19138 to D.A.D., and by a Howard Hughes Medical Institute predoctoral fellowship to J.D.B. C.O.W. was supported in part by National Institutes of Health grant AI 065960.

REFERENCES

1. Bastolla, U., M. Porto, H. E. Roman, and M. Vendruscolo. 2002. Lack of self-averaging in neutral evolution of proteins. *Phys. Rev. Lett.* 89: 208101.
2. Bornberg-Bauer, E., and H. S. Chan. 1999. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. USA.* 96:10689–10694.
3. Broglia, R. A., G. Tiana, H. E. Roman, E. Vigezzi, and E. Shakhnovich. 1999. Stability of designed proteins against mutations. *Phys. Rev. Lett.* 82:4727–4730.
4. Chan, H. S., and E. Bornberg-Bauer. 2002. Perspectives on protein evolution from simple exact models. *Appl. Bioinformat.* 1:121–144.
5. Wilke, C. O. 2004. Molecular clock in neutral protein evolution. *BMC Genet.* 5:25.
6. Xia, Y., and M. Levitt. 2004. Simulating protein evolution in sequence and structure space. *Curr. Opin. Struct. Biol.* 14:202–207.
7. Drummond, D. A., B. L. Iverson, G. Georgiou, and F. H. Arnold. 2005. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *J. Mol. Biol.* 350:806–816.
8. Bloom, J. D., J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold. 2005. Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. USA.* 102:606–611.
9. Daugherty, P. S., G. Chen, B. L. Iverson, and G. Georgiou. 1999. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc. Natl. Acad. Sci. USA.* 97:2029–2034.
10. Guo, H. H., J. Choe, and L. A. Loeb. 2004. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. USA.* 101:9205–9210.
11. Shafikhani, S., R. A. Siegel, E. Ferrari, and V. Schnellenberger. 1997. Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. *Biotechniques.* 23:304–310.
12. Loeb, D. D., R. Swanstrom, L. Everitt, M. Manchester, S. E. Stamper, and C. A. Hutchison. 1989. Complete mutagenesis of the HIV-1 protease. *Nature.* 340:397–400.
13. Pakula, A. A., V. B. Young, and R. T. Sauer. 1986. Bacteriophage λ *cro* mutations: effects on activity and intracellular degradation. *Proc. Natl. Acad. Sci. USA.* 83:8829–8833.
14. Shortle, D., and B. Lin. 1985. Genetic analysis of staphylococcal nuclease: identification of three intragenic ''global'' suppressors of nuclease-minus mutations. *Genetics.* 110:539–555.
15. Taverna, D. M., and R. A. Goldstein. 2002. Why are proteins so robust to site mutations? *J. Mol. Biol.* 315:479–484.
16. Kloczkowski, A., and R. L. Jernigan. 1997. Computer generation and enumeration of compact self-avoiding walks within simple geometries on lattices. *Comput. Theor. Polym. Sci.* 7:163–173.
17. Miyazawa, S., and R. L. Jernigan. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256:623–644.
18. Venables, W. N., and D. M. Smith. 2002. The *R* Development Core Team. An Introduction to R. Network Theory Ltd., Bristol, UK.
19. Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B.* 57:289–300.
20. Chan, H. S., and K. A. Dill. 1996. Comparing folding codes for proteins and polymers. *Proteins Struct. Funct. Genet.* 24:335–344.
21. Irbäck, A., and C. Troein. 2002. Enumerating designing sequences in the HP model. *J. Biol. Phys.* 28:1–15.
22. Cramér, H. 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ.
23. Blinnikov, S., and R. Moessner. 1998. Expansions for nearly Gaussian distributions. *Astron. Astrophys. Suppl. Ser.* 130:193–205.
24. Cramér, H. 1938. On a new limit theorem in the theory of probability. *In* Colloquium on the Theory of Probability. Hermann, Paris, France.
25. Hahn, M. G., and M. J. Klass. 1997. Approximation of partial sums of arbitrary i.i.d. random variables and the precision of the usual exponential upper bound. *Annals Prob.* 25:1451–1470.
26. Varga, R. S. 2000. Matrix Iterative Analysis, 2nd Ed. Springer-Verlag, New York.
27. Dressel, P. L. 1940. Statistical semivariants and their estimates with particular emphasis on their relation to algebraic invariants. *Annals Math. Stat.* 11:33–57.