

Thermodynamic prediction of protein neutrality

Jesse D. Bloom^{*†‡}, Jonathan J. Silberg[§], Claus O. Wilke^{¶||}, D. Allan Drummond^{†||}, Christoph Adami^{†||}, and Frances H. Arnold^{*‡}

^{*}Division of Chemistry and Chemical Engineering 210-41, [†]Digital Life Laboratory 136-93, and ^{||}Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125; [§]Department of Biochemistry and Cell Biology, Rice University, Houston, TX 77005; and [¶]Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711

Edited by Alan R. Fersht, University of Cambridge, Cambridge, United Kingdom, and approved December 3, 2004 (received for review September 10, 2004)

We present a simple theory that uses thermodynamic parameters to predict the probability that a protein retains the wild-type structure after one or more random amino acid substitutions. Our theory predicts that for large numbers of substitutions the probability that a protein retains its structure will **decline exponentially with the number of substitutions**, with the severity of this decline determined by properties of the structure. Our theory also predicts that a protein can gain extra robustness to the first few substitutions by increasing its thermodynamic stability. We validate our theory with simulations on lattice protein models and by showing that it quantitatively predicts previously published experimental measurements on subtilisin and our own measurements on variants of TEM1 β -lactamase. Our work unifies observations about the clustering of functional proteins in sequence space, and provides a basis for interpreting the response of proteins to substitutions in protein engineering applications.

mutational robustness | protein evolution | protein stability | directed evolution | β -lactamase

The ability to predict a protein's tolerance to amino acid substitutions is of fundamental importance in understanding natural protein evolution, developing protein engineering strategies, and understanding the basis of genetic diseases. Computational and experimental studies have demonstrated that both protein stability and structure affect a protein's tolerance to substitutions. Simulations have shown that more stable proteins have a higher fraction of folded mutants (1–4) and that some structures are encoded by more sequences than others (5–7). Experiments have demonstrated that proteins can be extremely tolerant to single substitutions; for example, 84% of single-residue mutants of T4 lysozyme (8) and 65% of single-residue mutants of *lac* repressor (9) were scored as functional. For multiple substitutions, the fraction of functional proteins decreases roughly exponentially with the number of substitutions, although the severity of this decline varies among proteins (10–12). Protein mutagenesis experiments have also underscored the contribution of protein stability to mutational tolerance by finding “global suppressor” substitutions that buffer a protein against otherwise deleterious substitutions by increasing its stability (13, 14).

We unify these diverse experimental and computational results into a simple framework for predicting a protein's tolerance to substitutions. A fundamental measure of this tolerance is the fraction of proteins retaining the wild-type structure after a single random substitution, often called the neutrality (15). We extend this concept to multiple substitutions by defining the ***m*-neutrality as the fraction of proteins that fold to the wild-type structure among all sequences that differ from the wild-type sequence at *m* residues**. Because mutants that fail to fold also generally fail to function, the *m*-neutrality provides an upper bound to the fraction of proteins with *m* substitutions that retain biochemical function. We show that a protein's *m*-neutrality can be accurately predicted from measurable thermodynamic parameters, and that these predictions capture the contributions of both stability and structure to determining a protein's tolerance to substitutions.

Methods

Lattice Protein Model. We performed simulations with lattice proteins (16) of length $L = 20$ monomers of 20 types corresponding to the natural amino acids. We folded the proteins on a two-dimensional lattice, allowing them to occupy any of the 41,889,578 possible compact or noncompact conformations. The energy of a conformation C is the sum of the nonbonded nearest-neighbor interactions

$$E(C) = \sum_{i=1}^L \sum_{j=1}^{i-2} C_{ij}(C) \times \varepsilon(\mathcal{A}_i, \mathcal{A}_j),$$

where $C_{ij}(C)$ is one if residues i and j are nearest neighbors in conformation C and zero otherwise, and $\varepsilon(\mathcal{A}_i, \mathcal{A}_j)$ is the interaction energy between residue types \mathcal{A}_i and \mathcal{A}_j , given in table 5 of ref. 17.

The primary advantage of using lattice proteins is that we can exactly compute the stability of a conformation C_i as

$$\Delta G_f(C_i) = E(C_i) + T \ln\{Q(T) - \exp[-E(C_i)/T]\},$$

where $Q(T)$ is the partition sum.

$$Q(T) = \sum_{\{C_i\}} \exp[-E(C_i)/T]$$

over all conformations, made tractable by noting that there are only 910,972 unique contact sets. All simulations were performed at a reduced temperature of $T = 1.0$.

TEM1 β -Lactamase Mutant Libraries. To examine the effects of mutations on the retention of protein function, we constructed mutant libraries of wild-type and the thermostable M182T variant of TEM1 β -lactamase. The 861-bp genes (a kind gift from Brian Shoichet, Northwestern University School of Medicine, Chicago; ref. 18) were subcloned into the pMON:1A2 plasmid (19) with *SacI* and *HindIII* by using PCR primers 5'-GCGGCGGAGCTCATGAGTATTCAACATTTCGGT-GTCGC-3' and 5'-GCGGCGAAGCTTTTACCAATGCTTA-ATCAGTGAGGCAC-3' (restriction sites are underlined). We first created a control unmutated library by cutting the gene directly from the plasmid. This unmutated gene was used as the template for a round of error-prone PCR with 100- μ l reactions containing 3 ng of template, 0.5 μ M each of the above primers, 7 mM MgCl₂, 75 μ M MnCl₂, 200 μ M dATP and dGTP, 500 μ M dTTP and dCTP, 1 \times Applied Biosystems PCR buffer without MgCl₂, and 5 units of Applied Biosystems *Taq* DNA polymerase. The PCR conditions were 95°C for 5 min, and then 14 cycles of 30 s each at 95°C, 50°C, and 72°C. The product from this PCR was digested with *SacI/HindIII* and gel purified, and then used as the template for another identical round of error-prone PCR.

This paper was submitted directly (Track II) to the PNAS office.

[†]To whom correspondence may be addressed: E-mail: bloom@caltech.edu or frances@cheme.caltech.edu.

© 2005 by The National Academy of Sciences of the USA

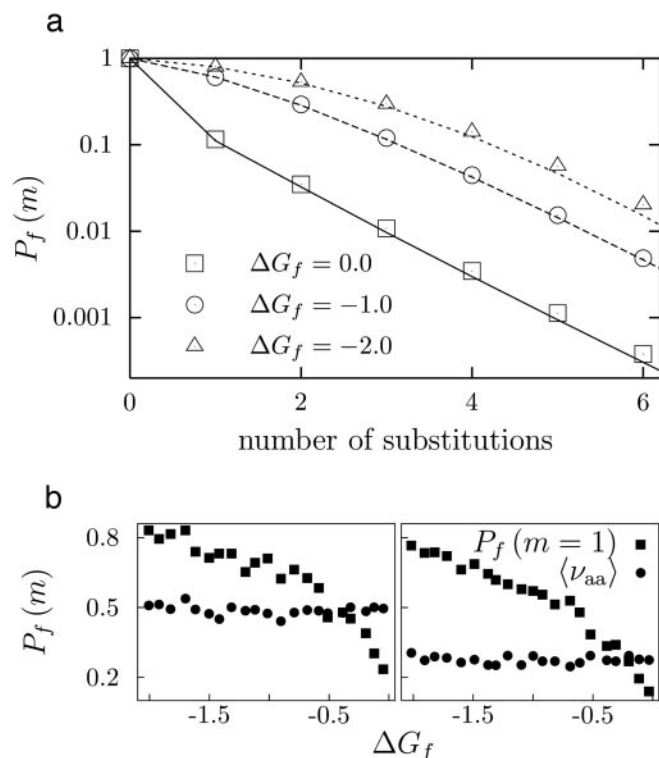


Fig. 2. Lattice proteins with the same structure but different stabilities have different 1-neutralities but have the same average neutrality (ν_{aa}). (a) Predicted (lines) and measured (symbols) m -neutralities for proteins with different stabilities and the same structure (III in Fig. 1). (b) Measured values of the 1-neutralities (squares) and average neutralities (circles) for proteins with different stabilities but the same structures (the plots at left and right are for structures I and IV from Fig. 1, respectively). The sequences were generated by finding a sequence with $\Delta G_f = -2.0$ by using the procedure described in Fig. 1, and then using this sequence as a starting point for neutral evolution selecting for the indicated target stabilities. The proteins with different stabilities are highly diverged, with average pairwise sequence identities of 15% and 41% for the structures at left and right, respectively. The m -neutralities were computed as in Fig. 1, and $\langle \nu_{aa} \rangle$ was computed as the square root of the 6-neutrality divided by the 4-neutrality.

$$f(m_{\text{nt}}) = (1 + \lambda)^{-N} \sum_{k=0}^N \binom{N}{k} \lambda^k \frac{(kx)^{m_{\text{nt}}} e^{-kx}}{m_{\text{nt}}!},$$

where $x = \langle m_{\text{nt}} \rangle (1 + \lambda)/(N\lambda)$ (36, 37). Subtilisin was mutagenized by using 13 PCR cycles with 10 effective doublings (10), so N is 13 times the number of rounds of error-prone PCR and $\lambda = 0.77$. TEM1 was mutagenized by using 14 PCR cycles with 10 effective doublings, so N is 14 times the number of rounds and $\lambda = 0.71$. We confirmed that $f(m_{\text{nt}})$ accurately describes the distribution of mutations in our libraries (Fig. 5, which is published as supporting information on the PNAS web site).

The expected fraction of folded sequences in a mutant library is easily calculated from $f(m_{\text{nt}})$ and the probability $P_f(m_{\text{nt}})$ that a sequence is still functional after m_{nt} nucleotide mutations as

$$\mathcal{F} = \sum_{m_{\text{nt}}=0}^{\infty} f(m_{\text{nt}}) \times P_{\text{f}}(m_{\text{nt}}).$$

We calculated the probability $P_f(m_{nt})$ that a sequence was still folded after m_{nt} nucleotide mutations by using two existing computer programs for estimating the $\Delta\Delta G$ values for single substitutions to proteins with known structures (Protein Data

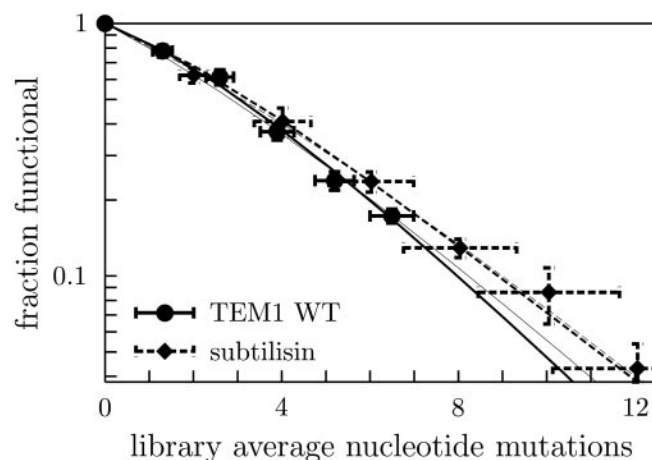


Fig. 3. Theoretical predictions and fractions of functional proteins in mutant libraries of subtilisin (dashed lines) and TEM1 β -lactamase (solid lines) genes. Thick lines show predictions made by using POPMUSIC (38), and thin lines show predictions made by using FOLDEF (39). The TEM1 measurements are from Table 1, normalized by the values from the control unmutated library.

Bank ID 1IAV for subtilisin and 1BTL for TEM1): Gilis and Rooman’s POPMUSIC potential (38) and Serrano and coworkers’ FOLDEF potential (39) with van der Waals clash energies. Because the genetic code makes nucleotide mutations more likely to induce some amino acid substitutions than others, and because error-prone PCR introduces a nonrandom distribution of nucleotide mutations, we weighted each $\Delta\Delta G$ value by the probability that it would be induced by a single nucleotide mutation made according to the observed error-prone PCR nucleotide mutation frequencies (given in table 1 of ref. 10 for subtilisin and Table 2 for TEM1). We assigned a $\Delta\Delta G$ of zero to synonymous nucleotide mutations because they do not cause an amino acid substitution, and we assigned a $\Delta\Delta G$ of 25 kcal/mol to frameshift and nonsense mutations because premature truncation is expected to inactivate the protein. We ignored the small number of substitutions for which POPMUSIC failed to calculate a $\Delta\Delta G$. With this weighted $\Delta\Delta G$ distribution for nucleotide mutations, all we needed to construct $P_f(m_{nt})$ according to Eq. 1 was the value of $\Delta G_f^{\text{extra}}$. This value cannot be measured directly because we do not know the minimal stability threshold. However, because $\Delta G_f^{\text{extra}}$ only influences the initial behavior of the m -neutrality and does not affect the limiting decline (Fig. 2), and because we have six data points for each protein, we could do a least-squares fit of $\Delta G_f^{\text{extra}}$ to the data and still test the ability of the theory to predict the decline in the fraction of functional proteins.

Fig. 3 shows the measured fractions of functional proteins for subtilisin and wild-type TEM1 versus the theoretical predictions made with POPMUSIC and FOLDEF. The theoretical predictions closely match the measured fractions of functional proteins in all cases, with subtilisin exhibiting slightly higher m -neutralities than TEM1.

The second major prediction of our theory is that, among proteins with the same structure, more stable variants will exhibit higher initial m -neutralities, but converge to same average neutrality. To test this prediction, we compared the fractions of functional proteins in mutant libraries of wild-type and the M182T variant of TEM1. The M182T variant differs from wild-type by only a single substitution, yet is 2.7 kcal/mol more stable (18), so we predict that it should exhibit a higher fraction of functional proteins at the same level of mutation. Fig. 4 shows the measured fractions functional for wild-type and the M182T variant, as well as the theoretical predictions made with both

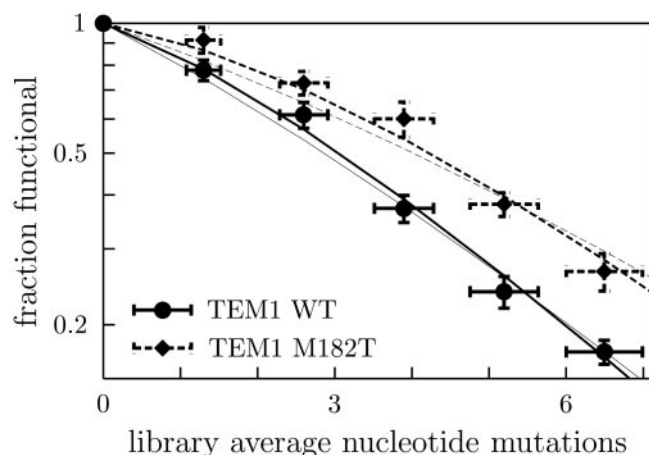


Fig. 4. The more stable M182T variant of TEM1 β -lactamase (dashed lines) exhibits a higher fraction of functional mutants relative to wild type (solid lines), as predicted. Thick lines show predictions made by using POPMUSIC (38), and thin lines show predictions made by using FOLDEF (39). The measurements are from Table 1, normalized by the values from the control unmutated library.

POPMUSIC and FOLDEF. As predicted, the M182T variant exhibits a higher fraction of functional proteins, and once again the predictions made with both potentials are in good agreement with the experimental measurements.

To further explore the range of possible neutralities for different proteins, we used $\Delta\Delta G$ values from POPMUSIC to predict the expected average neutralities to both amino acid substitutions ($\langle\nu_{aa}\rangle$) and nucleotide mutations ($\langle\nu_{nt}\rangle$) for proteins chosen from several different CATH (40) protein structure classifications. Because we do not know $\Delta G_f^{\text{extra}}$ for these proteins, we computed the fraction of proteins expected to be inactivated by the 10th mutation because, after this many mutations, effects of the initial protein stability should be small. Table 3 shows the predicted average neutralities to both random amino acid substitutions and nucleotide mutations made according to the mutation probabilities of our TEM1 mutagenesis. The predicted average neutralities differ considerably, showing that our theory predicts that different proteins can have substantially different neutralities.

Discussion

We have presented a theory for calculating the probability that a protein will retain its structure after random amino acid substitutions, and have confirmed the main theoretical predic-

tions with simulations and experiments. Our theory naturally separates a protein's m -neutrality into components caused by structure and stability. The eventual severity of the exponential decline in m -neutrality with the number of substitutions is a property of a protein's structure. On the other hand, increased stability confers greater tolerance to the first few substitutions, in effect allowing a protein to "take a few hits" before it is pushed into the inevitable structurally determined exponential decline in m -neutrality. This increased tolerance to mutations due to extra stability is probably also the underlying reason for the existence of global suppressor mutations (13, 14) that buffer proteins against otherwise deleterious mutations.

The major assumption underlying our theory is that the thermodynamic effects of substitutions are additive. This assumption is clearly not strictly true because protein residues do interact. Substitutions are most likely to be nonadditive if the mutated residues are in close contact in a protein's structure (24, 25). Because proteins are large, two randomly chosen residues will rarely contact each other, and so although the additivity assumption is certainly violated for some specific combinations of substitutions, it is accurate when averaged over all possible substitutions. When we apply our theory to measurements of the fraction of mutant proteins that retain function, we are making a second assumption by ignoring the possibility that some substitutions may disrupt a protein's function in ways other than affecting its stability. Therefore, for proteins with a high fraction of functional residues, our theory provides only an upper bound on the fraction of functional proteins. However, our theory's remarkable success for both the subtilisin and the TEM1 mutant libraries suggests that this assumption is also valid.

Our theory provides a quantitative rationale for earlier work with lattice proteins on the organization of functional proteins in sequence space. Bornberg-Bauer and Chan (2) proposed that proteins are located in superfundels in sequence space, with the most stable sequence having the most neutral neighbors; others have reported that folded proteins surround highly stable prototype sequences in sequence space (3, 4, 41), and Shakhnovich and coworkers (1) showed that proteins with a large energy gap between the lowest and second lowest energy conformations are stabilized against mutations. We provide a clear explanation: more stable proteins are able to tolerate more of the possible mutations before unfolding, and so a higher fraction of their neighboring sequences fold.

In addition to these stability-based effects, different protein structures have different inherent designabilities, with more sequences folding into some structures than others (5, 42, 43). Proteins with more designable structures might be expected to show a higher average neutrality because their structures occupy a larger fraction of sequence space. Therefore, the average

Table 3. Predicted average neutralities

PDB	Protein	CATH architecture	Length, bp	$\langle\nu_{nt}\rangle$	$\langle\nu_{aa}\rangle$
1IAV	Subtilisin	$\alpha\beta$ 3-layer sandwich	269	0.65	0.55
1B9C	GFP	β -barrel	236	0.62	0.56
1BTL	TEM1 β -lactamase	$\alpha\beta$ 3-layer sandwich	263	0.58	0.46
1RLV	tRNA endonuclease	Not classified	305	0.55	0.44
1HZW	Thymidylate synthase	$\alpha\beta$ 2-layer sandwich	290	0.50	0.41
2BNH	Ribonuclease inhibitor	$\alpha\beta$ horseshoe	457	0.45	0.35
1HEL	Hen lysozyme	α orthogonal bundle	129	0.43	0.38

Predictions of the average neutralities of various proteins to both nucleotide mutations ($\langle\nu_{nt}\rangle$) and amino acid substitutions ($\langle\nu_{aa}\rangle$). The Protein Data Bank ID codes (PDB) and the CATH (40) architectures are shown along with the lengths of the protein chains in the PDB structures (in all cases we consider chain A). The average neutralities are computed by calculating the fraction of sequences predicted to be inactivated by the 10th mutation or substitution, by using $\Delta\Delta G$ values from POPMUSIC (38) and assuming that the proteins all have the same value of $\Delta G_f^{\text{extra}}$ as wild-type TEM1 β -lactamase. The values of $\langle\nu_{nt}\rangle$ are computed assuming that nucleotide mutations are made according to the error-prone PCR mutation frequencies of Table 2.

