Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 12972–12975, December 1994
Biophysics

Nonrandomness in protein sequences: Evidence for a physically driven stage of evolution?

(correlations/origin of life)

VIJAY S. PANDE, ALEXANDER Y. GROSBERG*, AND TOYOICHI TANAKA

Department of Physics and Center for Material Sciences and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

Communicated by Herman Feshbach, August 19, 1994

ABSTRACT The sequences, or primary structures, of existing biopolymers—in particular, proteins—are believed to be a product of evolution. Are the sequences random? If not, what is the character of this nonrandomness? To explore the statistics of protein sequences, we use the idea of mapping the sequence onto the trajectory of a random walk, originally proposed by Peng *et al.* [Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. & Stanley, H. E. (1992) *Nature (London)* 356, 168–170] in their analysis of DNA sequences. Using three different mappings, corresponding to three basic physical interactions between amino acids, we found pronounced deviations from pure randomness, and these deviations seem directed toward minimization of the energy of the three-dimensional structure. We consider this result as evidence for a physically driven stage of evolution.

From the molecular point of view, biological evolution implies the change of the set of sequences of existing proteins. In the same spirit, prebiological evolution is also understood as the creation and possibly subsequent change of some primary ensemble of sequences (not necessarily protein sequences). Thus, evolution can be viewed as some walk, search, and optimization in sequence space. This space, however, is astronomically big because the number of possible sequences is exponential in the length of polymer chains involved. For this reason, an exhaustive search in sequence space is well known to be prohibitively time consuming and, therefore, at least some element of randomness seems inevitable for any understandable picture of evolution.

It can be shown mathematically that a random choice of a point in sequence space, with uniform probability distribution over the entire space, is equivalent to a completely random formation of the sequence in a letter-by-letter manner without any correlations. Therefore, delicate deviations of the sequences from pure randomness or correlations between monomers along the sequences might be of great importance, as these changes can yield some fingerprint relating to the process that has created the existing biopolymers.

Similar arguments were used to justify the concept that is imaginatively formulated by the statement "proteins are slightly edited random copolymers" (1). For example, it was shown that the lengths distribution of α-helices in proteins follows accurately what could be expected for just random sequences (1). Some other tests can also be found (ref. 1 and the references therein). We also mention that the small degree of "editing" is closely related to the neutral theory of evolution (2). In the spirit of the concept of "proteins as edited random copolymers," we address here the aspect in which they are "edited."

To look for this nonrandomness, one has to decode the sequence in an appropriate manner. For example, some peculiar correlations between monomers were recently found in purine-pyrimidine representation of DNA sequences (3). As for proteins, we expect that this decoding has to be related to the three-dimensional structure and the folding properties of a protein chain. Indeed, the three-dimensional structure of protein is believed to be completely encoded in the sequence. On the other hand, it is exactly the three-dimensional structure that defines all aspects of a protein's functionality and, therefore, the properties of a protein in competition under evolutionary selective pressure. In other words, the relationship between the sequence and the selective promise of the protein is mediated by the three-dimensional structure. Thus, as the three-dimensional structure can be considered to be "written" in the amino acid sequence in the "language" of the interactions between amino acids, we decode protein sequences according to the role of each particular residue in the determination of the protein's three-dimensional structure. Namely, we consider three ways to decode protein sequence, related to the three most important kinds of volume interactions—Coulomb interaction, hydrophobic/hydrophilic interaction, and hydrogen bonding.

"Brownian Bridge" Representation for Protein Sequences

Technically, we use the idea of Peng *et al.* (3) and map protein sequence onto the trajectory of an artificial one-dimensional random walker. More precisely, we construct for each sequence a one-dimensional walker that makes steps of size σ up and down at discrete time moments i, $0 \leq i \leq L$. The walker is required to return to the origin after the entire trip of L steps, so that the corresponding trajectory is a Brownian bridge. A purely random walker, which corresponds to a random sequence, is expected to travel $\approx\sigma\sqrt{L}$ from the origin on mean-square-average. To reach farther, the walker must go mainly in one direction for the first half-time ($i < L/2$) and mainly back in the second half-time ($i > L/2$), thus approaching the maximal distance of $\sigma L/2$. On the other hand, to keep as close to the origin as possible, the walker must compensate each step to one direction by a subsequent opposite step. Therefore, persistent types of correlations in protein sequences would be manifested in trajectories that go beyond the random one, whereas alternating correlations would lead to the trajectories that do not travel as far.

To use this test of nonrandomness, we have calculated for each of the amino acid sequences obtained from a data bank (8) the trajectories of three different artificial walkers, each related to a kind of physical interactions between residues—hydrophobic (A), hydrogen bonds (B), and Coulomb (C). The subsequent steps of each walker are given by the numbers $\{\xi_i\}$ defined as follows: for A, $\xi_i = +1$ if monomer number i in the given sequence is highly hydrophilic (lysine, arginine, histidine, aspartate, and glutamate) or $\xi_i = -1$ in any other case; for B, ξ_i may be $+1$ or -1 for monomers capable (asparagine,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*On leave from: Institute of Chemical Physics, Russian Academy of Sciences, Moscow 117977, Russia.

glutamine, serine, threonine, tryptophan, and tyrosine) or incapable (all others) of hydrogen bonding (4); for C, ξ_i may be $+1$, -1, or 0 for positively (lysine, arginine, and histidine) or negatively charged (asparagine and glutamate) and neutral (all others) monomer i, respectively (4).

To look for correlations by comparing the trajectories, we have to exclude the dependencies on protein length, overall composition, and the step size of the walker. This is done by the following definition of trajectories:

$$r(\lambda) \equiv \left\langle \left[\sum_{i=0}^{\lceil \lambda L_p \rceil} \frac{\Delta\xi_i^{(p)}}{\sigma^{(p)}} \right]^2 \right\rangle_p , \qquad [1]$$

where p denotes a given protein, $\langle \ldots \rangle_p$ means average over the set of proteins, $\lceil \ldots \rceil$ means take the next highest integer, and L_p is the total number of amino acids in p. (*i*) To exclude L_p-dependence, we rescale the number of steps taken (l) as $\lambda = l/L_p$, $0 \leq \lambda \leq 1$; (*ii*) to exclude the walker's drift due to the protein overall composition, we subtract the term linear in λ for each protein by $\Delta\xi_i^{(p)} = \xi_i^{(p)} - \overline{\xi^{(p)}}$, $\overline{\xi^{(p)}} = (1/L_p)\sum_{i=0}^{L_p} \xi_i^{(p)}$ (in this way the trajectory is brought to the bridge shape); (*iii*) to exclude the step-size dependence, we divide by $\sigma^{(p)} = \sqrt{\sum_{i=0}^{L_p} [\xi_i^{(p)} - \overline{\xi_i^{(p)}}]^2}$. In other words, $r(\lambda)$ is the distance traveled by the effective walker (i.e., with the mean drift removed) after taking $\lceil \lambda L_p \rceil$ steps of size σ.

Our procedure to construct the walkers is thus a modification of the original Peng *et al.* (3) procedure, in such a way, that (*i*) we average over an ensemble of different proteins rather than along the chain and (*ii*) all the trajectories are bridges.

The trajectories $r_A(\lambda)$, $r_B(\lambda)$, and $r_C(\lambda)$, along with the theoretically found trajectory

$$r_{\text{rand}}(\lambda) = \frac{1}{\lambda^{-1} + (1-\lambda)^{-1}} , \qquad [2]$$

for purely random case, are shown in Fig. 1 for a set of globular proteins [those coded as catalysts in the Data Bank (8)]. The $r_A(\lambda)$ and $r_B(\lambda)$ bridges are clearly over $r_{\text{rand}}(\lambda)$, manifesting pronounced persistent correlations in the distribution of hydrophobicity. Alternating correlations are found between electrical charges on protein chains because $r_C(\lambda)$ is definitely under $r_{\text{rand}}(\lambda)$. This is the main finding of the work.

Brownian Bridges for Some Particular Sets of Proteins

Some developments of this main result are as follows. When we look at early forms of life, such as prokaryotes, we find that the corresponding Brownian bridges shown in Fig. 2 fit quite well to a phenomenological scaling generalization of Eq. 2 of the form

$$r(\lambda) = \frac{L_0^{2\alpha-1}}{\lambda^{-2\alpha} + (1-\lambda)^{-2\alpha}} , \qquad [3]$$

yielding quantitative results of $\alpha_A = 0.520 \pm 0.005$, $\alpha_B = 0.520 \pm 0.005$, and $\alpha_C = 0.470 \pm 0.005$ for prokaryotes. Clearly, $\alpha > 1/2$ and $\alpha < 1/2$ means persistent and alternating type of correlations, respectively. To exclude small polypeptides as well as multiglobular proteins, we have examined only proteins with lengths between 110 and 750 amino acids. For simplicity, we take $L_0 = 110$—i.e., the shortest chain in the ensemble, but we have found no special qualitative dependance on L_0.

We stress here that $\alpha \neq 1/2$ does not imply any fractal interpretation, contrary to the DNA case, because we average over the ensemble of different sequences rather than over the sliding window in one sequence.

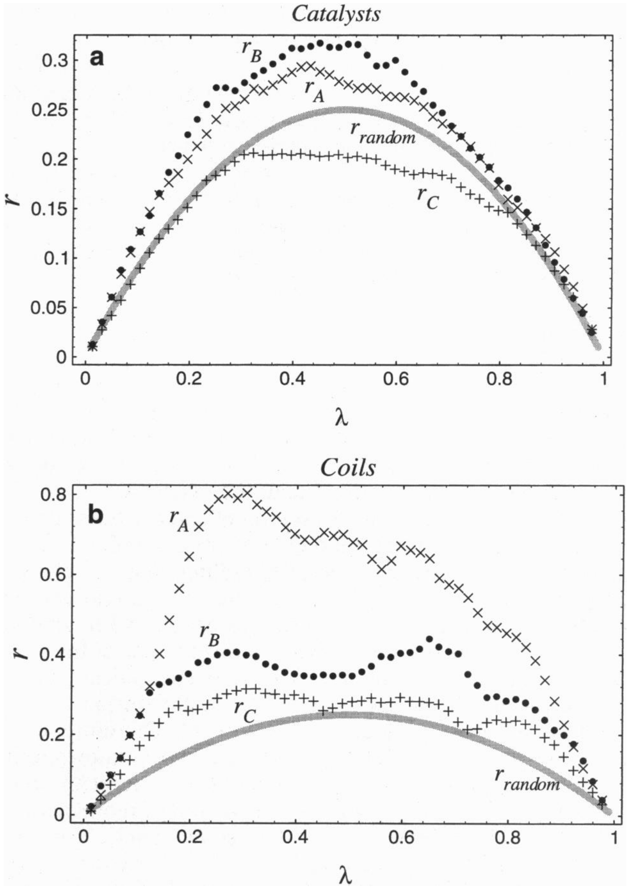

FIG. 1. Brownian bridges for hydrophilic (×), hydrogen bonding (●), and Coulomb (+) mappings of sequences of proteins with catalytic activity and, therefore, globular structure (*a*) and coiled structure (*b*). (*a*) The general qualitative behavior for catalysts ($\alpha_A > 1/2$, $\alpha_B > 1/2$, and $\alpha_C < 1/2$) is seen, when compared with the bridge corresponding to an ensemble of random sequences r_{rand} (thick gray curve)—i.e., $\alpha = 1/2$. (*b*) Persistent correlations are found in all mappings for coils.

Of course, the statistical errors are greater for smaller subsets of sequences. Nevertheless, the main qualitative finding (α_A, $\alpha_B > 1/2$, $\alpha_C < 1/2$) remains valid for all of the considered groups of globular proteins. At the same time, we have to mention, that

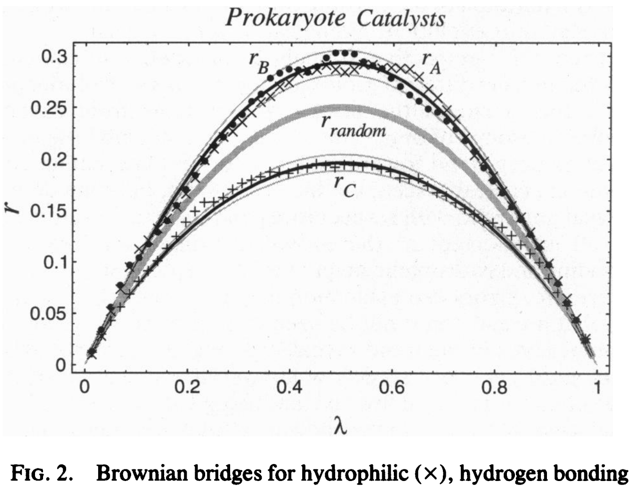

FIG. 2. Brownian bridges for hydrophilic (×), hydrogen bonding (●), and Coulomb (+) mappings of sequences of prokaryote proteins sequences. We find that these bridges fit well to Eq. 3 with $\alpha_A = 0.520 \pm 0.005$, $\alpha_B = 0.520 \pm 0.005$, and $\alpha_C = 0.470 \pm 0.005$ ($L_0 = 110$). The thin gray lines bounding a given bridge give the error spread specified above.

some of the bridges—for example, $r_A(\lambda)$ for enzymes from plants—exhibit clear irregularities and asymmetries, which remain unexplained. For the subset of coil-like proteins (i.e., denoted to be coiled in a comment or keyword of the data base), we found α_A, α_B, and $\alpha_C > 1/2$; this is easily related to the known periodicity of fibrillar protein sequences.

To insure that these results are not artifacts of the procedure used, we performed several control tests. In particular, artificial shuffling of the units along the chain as well as randomly shuffled versions of the maps *A*, *B*, and *C* all lead to random sequences ($\alpha = 0.5 \pm 0.0025$).

Discussion

To conclude, we speculate on the possible explanations for the nonrandomness of protein sequences. As mentioned in the Introduction, we believe that the deviations from randomness seen are the fingerprints of an evolutionary process, biological or prebiological. On the other hand, the results α_A, $\alpha_B > 1/2$, $\alpha_C < 1/2$ appear to be a manifestation of some process driven by physical interactions among monomers. Indeed, a sequence with a tendency toward alterating signs of charges along the chain ($\alpha_C < 1/2$) has, at the same conformation, obviously lower Coulomb energy compared with another hypothetical sequence with blocks of the charges of the same sign. Analogously, hydrophilic monomers energetically prefer to concentrate at the loops that are on the surface of the globule and thus in contact with the solvent. Therefore, there is the coincidence: the set of protein sequences, known to be a product of evolution, looks similar to the result of some physical game with repulsion and attraction of monomers.

What could be the reason for this coincidence? Consider recent works (5, 6), where two different procedures were suggested to prepare or, at least, to imitate the preparation of heteropolymers with sequences capable of renaturation into a given molecular fold. One of them (5) is based on annealing of the sequence of the polymer with a chosen target conformation. Another procedure (6) implies, before polymerization, prearrangement of monomers in space due to the interplay of repulsive and attractive interactions. These processes are both driven physically and lead, therefore, to α_A, $\alpha_B > 1/2$, and $\alpha_C < 1/2$. We have analyzed correlations along the artificial sequences produced by our model of polymerization (6) and found very reasonable agreement with the data for real proteins (e.g., prokaryotes). We conclude from this consideration that some physically driven process, where the same set of monomer-to-monomer interactions is used as in the renaturation of the existing proteins, is likely to be one of the stages of evolution, biological or prebiological.

From this perspective, it might be instructive to compare correlations in different groups of organisms vs. evolutionary age. Fig. 3 shows the bridges for proteins from several different groups of organisms. As to the Coulomb bridge, an evolutionary trend toward larger α_C or less alternating correlations is clearly seen. On the other hand, our data do not reveal any trend with respect to α_A and α_B. This result is not at all unexpected, as the Brownian bridges for hydrogen bonding and hydrophilic mappings had greater variation and, therefore, errors in α estimation than the Coulomb mapping, so that a trend might not be seen even if there were one. If one believes in the trend revealed by Fig. 3*a*, it implies that biological evolution somehow allows the elimination of the correlations imposed by the prebiological creation of sequences. We must stress, however, that this question remains of much more speculative character than our main finding, shown in Fig. 1.

One might consider our main results as only the reflection of physical constraints involved with the formation of heteropolymers with a unique structure (similar to, for example, obvious constraint that the total charge of the chain cannot be too large)—i.e., the correlations obtained represent the fact that certain sequences are more favorable due to physical criteria. However, the sheer fact that correlations are seen in the ensemble of proteins, which are assumed to be a product of evolution, is exactly how we understand our statement that at least some stage of biological or prebiological evolution has selected protein sequences based upon physical criteria.

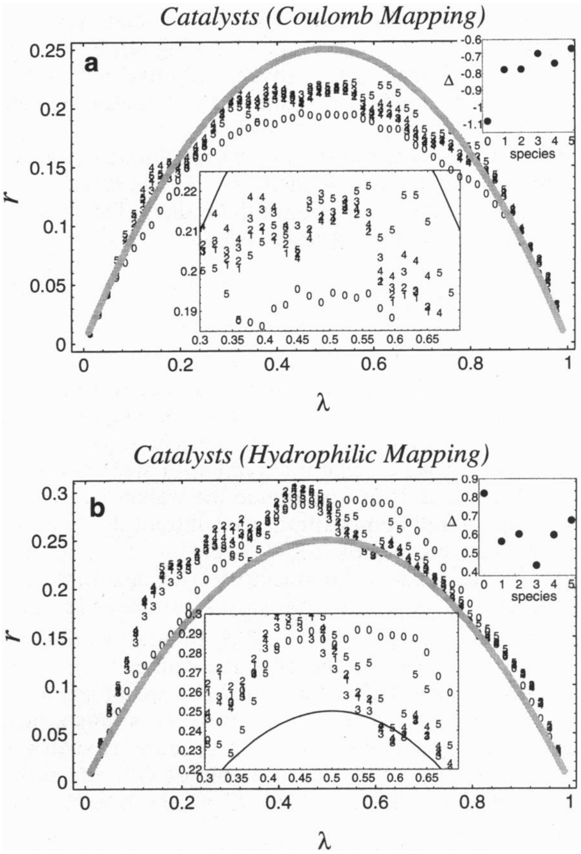

FIG. 3. Brownian bridges for a series of evolutionary groups: 0, Prokaryota; 1, Chordata; 2, Tetrapoda; 3, Metazoa; 4, Mammalia; and 5, Rodentia. (*a*) Coulomb mapping, with a magnified region 0.3 ≤ λ ≤ 0.7 in the lower center. There is a clearly seen trend, such that the younger (larger label numbers) evolutionary groups have bridges closer to r_{rand} (thick gray curve). This trend can be characterized by computing the difference (Δ) between the area under the Brownian bridge for a given species and the area under the bridge for random sequences. We have chosen the domain (0.3,0.7) for integration (*Inset*) as the error becomes great outside of this range. The result is seen in the upper right-hand corner. Another quantitative measure of the evolutionary trend would be to fit each bridge with Eq. 3 and plot α_i vs. i; qualitatively, this approach leads to the same conclusion, but because individual bridges do not necesarily fit very well to Eq. 3, except for prokaryotes, this fit introduces some artificial errors. (*b*) Using the hydrophilic mapping, again the prokaryote bridge fits well to Eq. 3 with $\alpha > 1/2$. As in the Coulomb case, the bridges for the other evolutionary groups deviate more from Eq. 3 than the prokaryote bridge; however, the evolutionary trend found with the hydrophilic mapping is not seen as clearly, as shown in the plot of Δ_i vs. i in the upper right-hand corner.

Appendix A

Derivation of Eq. 1. We start with a given ensemble of protein sequences. With the decoded sequence $\{\xi_1, \xi_2, \ldots, \xi_L\}$, we map it onto the trajectory as

$$x(l) = \sum_{t=1}^{l} \xi_t. \qquad \textbf{[A1]}$$

The walker defined by Eq. **A1** may have a strong drift, so that the leading term in $x(l)$ might be linear in l; this is related simply to the mean composition of the chain considered. Because overall composition is beyond our interest here, we define the reduced trajectory:

$$y(l) = x(l) - (l/L)x(L), \qquad \textbf{[A2]}$$

L being the total number of links in the entire polymer chain. Obviously, the y-walker returns back to the origin after the entire "trip." The corresponding trajectory $y(l)$ is called a "Brownian bridge."

In principle, y is expected to scale as L^{α} with chain length. For example, we have considered $y^2(L/2)$ for each protein and made the logarithm–logarithm plot, where each point corresponds to one particular protein and has coordinates L, $y^2(L/2)$. These plots indicate clearly the tendency toward power-law dependence of the type $y^2(L/2) \sim L^{2\alpha}$. However, because of restricted statistics available and great fluctuations, it is hard to come to convincing conclusions with this approach.

To collect all data in a comparable form, we have rescaled all the Brownian bridges compensating for different proteins with different lengths and variances of ξ distribution by

$$z^2(\lambda) = \frac{y^2}{L\overline{(\xi - \bar{\xi})^2}}, \qquad \textbf{[A3]}$$

where $\overline{(. . .)}$ = averaging over a given protein sequence (e.g., $\bar{\xi} = (1/L)\Sigma_{i=1}^{L}\xi_i$), and to exclude L-dependence, we rescale the number of steps taken (l) as $\lambda = l/L$, where $0 \le \lambda \le 1$.

With the rescaled trajectories $z^2(\lambda)$, we perform averaging over the ensemble of proteins:

$$r(\lambda) = \langle z^2(\lambda)\rangle_{\text{ensemble}}, \qquad \textbf{[A4]}$$

which, when combined with Eqs. **A1–A3**, yields Eq. **1**.

Appendix B

Derivation of Eq. 3. A Brownian bridge is generally the trajectory of a random walk that starts and terminates at the same point in space—say, in the origin. Let us consider first the simplest case of a random walk without correlations, and let us evaluate the probability distribution for the walker displacement z as a function of "time" l, $\mathcal{P}_l(z)$. This can be considered as the probability for two walkers to meet each other at the point z at the "moment" l: both of them start from the origin, but the first begins at zero time and walks for the time l, whereas the second begins at the time L and walks back in time for the period $L - l$. For the uncorrelated process, we have thus

$$\mathcal{P}_l(z) = p_l(z)\cdot p_{L-l}(z). \qquad \textbf{[B1]}$$

Because there are no correlations, $p_l(z)$ is simply the standard Gaussian distribution

$$p_l(z) = (la\pi)^{-1/2}\exp\left[-\frac{z^2}{la}\right], \qquad \textbf{[B2]}$$

where a is a parameter. We see, therefore, that in this case

$$\mathcal{P}_l(z) = \text{const}\cdot\exp\left[-\frac{z^2}{a}\left(\frac{1}{l} + \frac{1}{L-l}\right)\right], \qquad \textbf{[B3]}$$

where const is normalization factor, and $r(l) = \langle z^2(l)\rangle = \int z^2\mathcal{P}_l(z)dz$ thus obeys Eq. **2**.

We now return to a more general case. Scaling arguments imply that the distribution $p_l(z)$ is of the form

$$p_l(z) = \text{const}\cdot\exp\left[-\left(\frac{z}{al^{\alpha}}\right)^{\beta}\right], \qquad \textbf{[B4]}$$

where α and β are critical exponents. Supposing Eq. **B1** is valid (which generally may not be true), one easily gets the expression for $\mathcal{P}_l(z)$ and then for $r(l) = \langle z^2(l)\rangle$. At $\beta = 2$ we recover exactly Eq. **3**. It is clear from the derivation that applicability of Eq. **3** is restricted from two sides—namely, the validity of Eq. **B1** and the supposition $\beta = 2$. Our statistical analysis shows no need in trying other values of β, as well as in consideration of any generalization of Eq. **B1**. The simple variant of Eq. **3**, considered as purely phenomenological, works reasonably well.

To understand the physical meaning of critical exponent α, one has to look at Eq. **B4**. In terms of random-walk representation, Eq. **B4** implies that rms displacement of the walker scales as l^{α} with "time" l. Certainly, it is analogous to the excluded volume problem in polymer physics, where the size of polymer chain is known to scale as l^{ν} with chain length l, where $\nu > 1/2$ [3/5 in classical Flory theory (7)] or $\nu < 1/2$ (1/3 for dense globule), depending on the prevailing of repulsive or attractive monomer-to-monomer interactions, respectively. Therefore, α is analogous to the critical exponent of correlation radius. It is worthwhile to mention here that $\alpha > 1/2$ was found for DNA sequences (3).

The work was supported by National Science Foundation Grant DMR 90-22933. V.S.P. acknowledges the support of a National Science Foundation Fellowship. A.Y.G. acknowledges the support of a Kao Fellowship.

1. Ptitsyn, O. B. & Volkenstein, M. V. (1986) *J. Biomol. Struct. Dynamics* **4,** 137–156.
2. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, New York).
3. Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. & Stanley, H. E. (1992) *Nature (London)* **356,** 168–170.
4. Dressler, D. & Potter, H. (1990) *Discovering Enzymes* (Sci. Am. Library, New York).
5. Shakhnovich, E. I. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7195–7199.
6. Pande, V., Grosberg, A. Y. & Tanaka, T. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12976–12979.
7. Flory, P. J. (1953) *Principles of Polymer Chemistry* (Cornell Univ. Press, Ithaca, NY).
8. Bairoch, A. & Boeckmann, B. (1992) *Nucleic Acids Res.* **20,** 2019–2022.