

# Hydrophobic forces and the length limit of foldable protein domains

Milo M. Lin and Ahmed H. Zewail<sup>1</sup>

Physical Biology Center for Ultrafast Science and Technology, Arthur Amos Noyes Laboratory of Chemical Physics, California Institute of Technology, Pasadena, CA 91125

Contributed by Ahmed H. Zewail, May 2, 2012 (sent for review March 28, 2012)

**To find the native conformation (fold), proteins sample a subspace that is typically hundreds of orders of magnitude smaller than their full conformational space. Whether such fast folding is intrinsic or the result of natural selection, and what is the longest foldable protein, are open questions. Here, we derive the average conformational degeneracy of a lattice polypeptide chain in water and quantitatively show that the constraints associated with hydrophobic forces are themselves sufficient to reduce the effective conformational space to a size compatible with the folding of proteins up to approximately 200 amino acids long within a biologically reasonable amount of time. This size range is in general agreement with the experimental protein domain length distribution obtained from approximately 1,200 proteins. Molecular dynamics simulations of the Trp-cage protein confirm this picture on the free energy landscape. Our analytical and computational results are consistent with a model in which the length and time scales of protein folding, as well as the modular nature of large proteins, are dictated primarily by inherent physical forces, whereas natural selection determines the native state.**

levinthal paradox | lattice model | kinetics | folding funnel

Ever since the discovery that proteins can spontaneously self-assemble into unique three-dimensional shapes (folds) (1), the mechanism of this folding process has been a focus of biology. It has been shown that random sequences can fold into a unique ground state which is separated from other folds by an energy gap (2). However, assuming the existence of a unique native fold, there is no assurance that the protein can efficiently parse the fold space to find it. In particular, how nature is able to search the exponentially increasing number of folds accessible to proteins of nontrivial length has not been explicitly elucidated; Levinthal famously estimated that for a protein consisting of 150 amino acids, in which each degree of freedom is discretized into only ten possible values, it would take much longer than the age of the universe to sample all the folds even at the limit of molecular motion (3).

The qualitative resolution of the Levinthal paradox has been the concept of the folding funnel, whereby a global bias in the multidimensional energy landscape channels the protein toward the subspace containing the native fold (4). For example, by introducing an artificial search bias in favor of native contacts (5) or designing sequences favoring specific secondary structures (6), fast folding was computationally observed. This suggests that the funnel arises from evolutionary tuning of the intramolecular interactions via sequence mutation and that feasible protein folding times are the result of natural selection.

In contrast to this picture, we show below that a general effect, namely the hydrophobic force, is sufficient to account for the kinetics of fast folding without sequence evolution. This force, which is the global tendency for the chain to collapse to a compact shape and for the residues to segregate in the interior of proteins, has long been recognized as a dominant factor in protein folding (7, 8). Indeed, proteins have been experimentally (9, 10) and computationally (11) shown to undergo hydrophobic collapse in the earliest stage of folding. By considering the degeneracy of all

folds, we demonstrate that hydrophobic collapse together with hydrophobic/hydrophilic residue segregation lead to realistic folding time scales for globular proteins and protein domains, which are independently folding subunits that constitute larger proteins (12), thus quantitatively resolving the paradox. We also find an upper limit, of approximately 200 amino acids, on the length of protein domains for which such hydrophobic packing constraints would allow the native state to be identified within a biologically reasonable timescale through a hypothetical exhaustive search. By comparing to the experimental distribution of protein domain lengths, we find that most protein fall below this “hydrophobic length limit,” although it can be exceeded due to the influence of other processes, besides the hydrophobic force, that affect protein folding.

Many attempts have been made to estimate the reduction of the effective search space due to the hydrophobic force. For a self-avoiding chain (SAC) composed of  $L$  residues on a three-dimensional cubic lattice, the number of unique conformational folds (degeneracy) was found to be  $N_{\text{SAC}} \sim 4.68^L$  (13); if we further restrict the chain to adopt maximally compact folds, as defined in the mean field treatment (14), the degeneracy  $N_{\text{Compact}} = (6/e)^L$ , where  $e$  is the base of the natural logarithm. Although compaction significantly reduces the search space, and is a driving factor for secondary structure formation (15), the degeneracy is still astronomically large even for the smallest proteins.

The final step is the further reduction of the fold space by choosing only those compact folds with hydrophobic residues (H) maximally segregated, in the sense of maximizing the number of H-H contacts, into the interior of the protein, and polar residues (P) on the outside. This minimalist HP representation of proteins has been a mainstay of analytic investigations of protein folding (16). We define  $N_{\text{HP}}(s)$  to be the degeneracy of self-avoiding compact folds with maximum H/P segregation, which is a function of  $s$ , the sequence of H and P residues along the chain. Because hydrophobic residues are empirically randomly distributed along the protein sequence (17), the average conformational degeneracy of a collapsed H/P-segregated protein,  $\langle N_{\text{HP}} \rangle$ , is equal to  $N_{\text{HP}}(s)$  averaged over all possible  $s$  of length  $L$ , where “ $\langle \rangle$ ” denotes averaging over the sequence space. Therefore,  $\langle N_{\text{HP}} \rangle$  represents the size of the effective fold space, on average, when a protein folds in a polar solvent like water.

The chief difficulty in obtaining  $\langle N_{\text{HP}} \rangle$  is the constraint that all monomers must be connected in a chain; this constraint had made  $N_{\text{HP}}(s)$  impossible to analytically compute for any given sequence  $s$ , much less averaged over all  $s$ . Estimates that neglect to enforce the linear sequence of the chain, for example by assigning independent probabilities for each hydrophobic residue to be in the interior of the protein, overlook the crucial role of this constraint in reducing the degeneracy. Consequently,  $\langle N_{\text{HP}} \rangle$  was found to grow almost as quickly as  $N_{\text{Compact}}$  as a function of  $L$

Author contributions: M.M.L. and A.H.Z. performed research and wrote the paper.

The authors declare no conflict of interest.

<sup>1</sup>To whom correspondence should be addressed. E-mail: zewail@caltech.edu.



out” of ways to fold to the optimal map, and maps of increasing energy will become optimal. This turnover of the optimal map causes  $\langle N_{HP} \rangle$ , the optimal map degeneracy, to plateau as a function of  $L$ , in agreement with the explicit 2D calculations of Ref. (18).

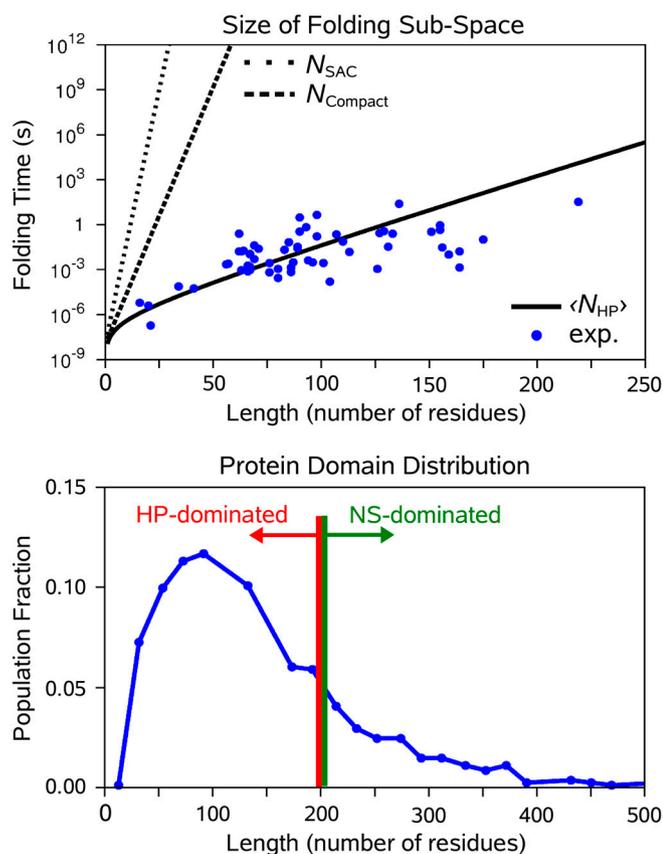
Eq. 3 also correctly predicts lattice simulations in three dimensions. Consistent with the “thousands to millions” of degenerate optimal conformations estimated from explicit enumeration,  $\langle N_{HP} \rangle \approx 10^4$  for  $L = 48$ . At a higher level of chemical accuracy, for a  $L = 80$  lattice chain consisting of the 20 types of amino acids and with a unique native fold, it was found that at most  $10^6$  Monte Carlo steps were required to reach the native fold (22). This is in agreement with Eq. 3:  $\langle N_{HP} \rangle = 5 \times 10^5$  for  $L = 80$ .

Multiplying by the experimentally determined 10-ns conformational rearrangement time ( $\tau_{\text{sampling}}$ ) (23), Eq. 4 converts the degeneracy into the folding time  $\tau_{\text{folding}}$ . Fig. 2 (Top) plots the folding (or sampling) time of the self-avoiding chain  $N_{\text{SAC}}$ , compact self-avoiding chain  $N_{\text{Compact}}$ , and the average collapsed H/P-segregated chain  $\langle N_{HP} \rangle \approx \langle N_{HP}^* \rangle$  as a function of  $L$  in three dimensions. In accordance with Levinthal, the number of conformations, even if restricted to the compact subset, becomes astro-

nomically large for very short chains ( $N_{\text{SAC}}$  and  $N_{\text{Compact}}$ ). Nevertheless, if confined to  $\langle N_{HP} \rangle$  by hydrophobic segregation, exhaustive search of this subspace can be accomplished in biological time (nanoseconds to minutes) for  $L < 200$ . Because  $\langle N_{HP} \rangle$  grows exponentially with  $L$ , beyond this length proteins cannot complete an exhaustive search of the hydrophobic subspace; this is the exhaustive hydrophobic search length limit for protein domains.

Also shown in Fig. 2 (Top) are the experimentally measured folding times for 65 single domain proteins (24, 25). For  $L < 100$ , the folding time agrees with the exhaustive sequence-averaged folding time  $\tau_{\text{folding}}$ , with the variance arising from the particular protein sequence. For  $L > 100$ , the average folding time falls below  $\tau_{\text{folding}}$ , indicating the onset of other factors such as sequence selection in order to evolve faster kinetics, despite the overall folding timescale for  $L < 200$  being dominated by the H/P collapse.

Although proteins often consist of more than 1,000 amino acids, protein domains are on average 100 amino acids long, typically ranging from 50 to 200 (26, 27), with 90% being less than 200 (28), which we have established as the length regime for which hydrophobic-polar interactions are sufficient for fast folding (HP-dominated). Besides being composed of sequences with smaller degeneracy than that of the average sequence, or whose energy landscapes allow fast folding due to other forces besides hydrophobic force, longer domains often consist of periodic local structures as a result of repeat insertion mutations (29), and molecular chaperones can also assist in folding (30). These types of evolutionary selection allow for the existence of domains that exceed the hydrophobic length limit; the fast folding of proteins in this second regime is therefore consistent with the effect of natural selection (NS-dominated). Fig. 2 (Bottom) shows the domain length distribution of a representative sample of 1236 proteins (21) and its partitioning into the two regimes. Consistent with the folding data of Fig. 2 (Top), the population fraction of proteins with length  $L$  begins to decay near  $L = 100$ , consistent with the onset of evolutionary pressure at this length scale to select for sequences that fold faster than exhaustive search. However, since the folding degeneracy increases exponentially with  $L$ , the fraction of protein domains exceeding the  $L = 200$  hydrophobic length limit is small. This may have forced most proteins with  $L > 200$  to evolve as modular combinations of smaller domains.



**Fig. 2.** Hydrophobic length and folding time limits of folding. Degeneracy of conformations on a cubic lattice is plotted as a function of chain length (Top). Conformational degeneracies of self-avoiding chain (SAC), self-avoiding compact chain, and the sequence-averaged lowest energy HP chain are shown with dotted, dashed, and solid lines, respectively. The degeneracies are multiplied by the 10-ns residue reorganization time to obtain the exhaustive folding times. The predicted limit (above which folding cannot occur at biologically relevant timescales) is  $L \sim 200$  amino acids. Experimentally measured folding times taken from refs. 24 and 25 are also shown, indicating that faster-than-exhaustive-search folding occurs for  $L > 100$ . The experimental domain length distribution of a representative set of 1236 proteins (data from ref. (26)) shows that for  $L > 100$ , the population fraction begins to decay (Bottom). The protein populations are divided into the HP-dominated regime for  $L$  below the exhaustive search length limit (red arrow), and the natural selection (NS)-dominated regime for  $L$  above the length limit (green arrow). See text for further description of the regimes.

**Computational.** To complement the general, yet coarse-grained, results above, we also performed ensemble-convergent molecular dynamics (MD) simulations using the CHARMM suite of programs and force field (31) on a single polypeptide to gain insight at the atomistic level. We chose the 20-residue Trp-cage (32), which despite its small size contains both secondary and tertiary structure, in particular the burial of the large hydrophobic tryptophan side-chain in the interior. Vacuum phase simulations were performed for 76 independent trajectories, each lasting 2  $\mu\text{s}$ . Solution phase simulations were performed explicitly represented solvent for 38 independent trajectories, each lasting 60 ns. All simulations were coupled to a Nose thermostat set at 28  $^{\circ}\text{C}$  to ensure a canonical room temperature ensemble, and the simulation time was long enough such that doubling of the simulation time did not significantly affect the results.

Fig. 3 shows the free energy landscape in both environments as a function of rmsd, the root-mean-squared deviation from the original experimentally determined structure and the solvent accessible surface area of the hydrophobic residues (exposed H area). In water, the peptide is confined to a free energy basin with burial of hydrophobic residues, including tryptophan (low exposed H area) and structure similar to that found experimentally (low rmsd). In the absence of water, the hydrophobic residues are more solvent-exposed; there are multiple conformational basins distributed throughout the free energy landscape, with



quantifies the effect of protein sequence on folding speed is the ratio of a protein's folding time to  $\tau_{\text{folding}}$ , the exhaustive search time of an average HP chain. In the case of protein design, we predict that for  $L < 200$  it is not necessary to engineer a kinetic pathway which leads to the desired native state; as long as the native state is thermodynamically stable and the roughness of the folding energy landscape is sufficiently low, the protein will fold in a reasonable time.

## Methods

**Lattice Model.** We denote a sequence "optimal" if it can achieve the global optimal map, and "suboptimal" otherwise. We define  $p$  to be the fraction of sequences that can only fold into suboptimal maps,  $\langle N_s \rangle$  to be the average ground state degeneracy over all such suboptimal sequences, and  $\langle N_o \rangle$  the average ground state degeneracy over all optimal sequences. To show that  $\langle N_{\text{HP}} \rangle \approx \langle N_{\text{HP}}^* \rangle$ , we note that  $\langle N_{\text{HP}} \rangle = p \langle N_s \rangle + (1-p) \langle N_o \rangle = p \langle N_s \rangle + \langle N_{\text{HP}}^* \rangle$ ; it is therefore sufficient to show that  $p \langle N_s \rangle < \langle N_{\text{HP}}^* \rangle$ . To this end, note that if a sequence is suboptimal, there exist lower energy states that it cannot achieve by locally perturbing any of its ground state folds. Being a globally suboptimal sequence, each ground state fold  $f$  of the sequence  $s$  contains a minimal-sized subvolume  $n(f, s)$  of the lattice in which the number of H–H contacts can be increased (and thus the energy decreased) by changing the positions of H and P residues to form a new map with lower energy in the subvolume. For each ground state fold  $f$ , assuming that there exists at least one other fold besides the starting fold which preserves the chain connectivity to the outside of the subvolume, define  $P_f$  to be the probability that at least one such fold can achieve a lower energy. Then, the probability that each residue of  $n(f, s)$  matches that of the lower-energy map is  $(1/2)^{n(f, s)}$ . Therefore,  $P_f > (1/2)^{n(f, s)}$ , where the greater-than sign is due to the possibility of multiple conformations, multiple lower energy maps, and unequal numbers of H and P residues within the subvolume which can all increase this probability. Then, the probability that none of the ground state folds can be locally perturbed to achieve a lower energy is:

$$\prod_{i=1}^{N_s} (1 - P_f) \approx 1 - \sum_{f=1}^{N_s} P_i < 1 - \sum_{f=1}^{N_s} (1/2)^{n(f, s)} < 1 - \langle N_s \rangle (1/2)^{\langle n \rangle}, \quad [5]$$

where  $\langle n \rangle$  is the average size of the minimal volume over all sequences and over all ground state folds of each sequence. The approximation in Eq. 5 is justified because  $P_f N_s < 1/2$  at the locally optimal fold. Since probabilities must be non-negative, we obtain from Eq. 5 that  $\langle N_s \rangle < 2^{\langle n \rangle}$ .

We can estimate  $\langle n \rangle$  by noting that there are three generic types of H–P swapping to achieve lower energy. First, there is the most likely case in which at least one extra H–H contact can be made by rearranging one surface of an H-region. The volume  $n(f, s)$  is therefore a path on an H-surface such that the H residue(s) at one end of the path swaps with the P residue(s) at the other end; the volume is thus at most twice the length of the cubic lattice:  $n(f, s) < 2L^{1/3}$  (see Fig. 4, *Top*) for any fold  $f$  and sequence  $s$ .

In the second case, rearrangement may require multiple H residues on one face of an H-region to swap with P residues on a different face. In this case (see Fig. 4, *Center*) the maximum  $n$  is limited to the area of a face:  $n(f, s) < L^{2/3}$  for any fold  $f$  and sequence  $s$ .

Finally, there is the case which requires two separate H-regions to connect. In this case, all H-regions must be cubes, otherwise case 1 or case 2 would apply. The maximum  $n(f, s)$  corresponds to the case in which the lattice is divided into a three-dimensional checkerboard, with each H-region being a cube of sides at most  $L^{1/3}/2$ . Thus, the maximum  $n(f, s)$  is equal to this cube plus one face of the adjacent cube, so that the cube may be shifted by one and thereby make contact with another cube:  $n(f, s) < (L^{1/3}/2)^2 * (L^{1/3}/2 + 1) = L/8 + L^{2/3}/4$  (see Fig. 4, *Bottom*) for any fold  $f$  and sequence  $s$ .

In all three cases,  $2^{n(f, s)} < \langle N_{\text{HP}}^* \rangle$  for any fold  $f$  and sequence  $s$ . Since this is true for any  $f$  and  $s$ , it is true when  $n(f, s)$  is averaged over all  $f$  and  $s$ :  $2^{\langle n \rangle} < \langle N_{\text{HP}}^* \rangle$ . For example, for  $L = 200$ ,  $\langle N_{\text{HP}}^* \rangle \sim 10^{12}$ , whereas  $2^{\langle n \rangle} < 10^5$ ,  $10^{10}$ , and  $10^{10}$  for the worst-case scenarios of the three cases, respectively. Therefore,  $p \langle N_s \rangle < 2^{\langle n \rangle} < \langle N_{\text{HP}}^* \rangle$ , and thus  $\langle N_{\text{HP}} \rangle \approx \langle N_{\text{HP}}^* \rangle$ .

**MD Simulations.** The solution phase simulations were performed on the peptide, 3914 TIP3P (37) water molecules, and one chlorine atom for neutrality. The system was restricted to a cubic box with initial sides of 50 Å and equilibrated at constant temperature (298 K) and pressure (1 atm) with periodic boundary conditions. The trajectories were seeded from the 38 NMR structural variants. In the case of vacuum simulations, 1-ns solution phase simulations were performed on each of the 38 initial structures, and the final conformations from these simulations were used to double the number of trajectories. We also performed the vacuum-phase simulations without coupling to a thermal bath (microcanonical ensemble) and confirmed that the free energy landscape is not significantly affected.

**ACKNOWLEDGMENTS.** We thank Dr. David Shaw for thoroughly going over the entire manuscript; we value his comments, which added to the clarity of the work presented. We also appreciate the helpful feedback on the preliminary manuscript from Prof. Thomas Miller III, Prof. Rob Phillips, Prof. Eugene Shakhnovich, Prof. Kenneth Dill, Prof. Peter Wolynes, and Prof. Vijay Pande. We are grateful to the National Science Foundation for funding of this research at Caltech. M.M.L. acknowledges financial support from the Krell Institute and the US Department of Energy for a DoE CSGF graduate fellowship (Grant DE-FG02-97ER25308) at Caltech.

- Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181:223–230.
- Shakhnovich EI, Gutin AM (1990) Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* 346:773–775.
- Levinthal C (1969) *Mossbauer Spectroscopy in Biological Systems*, eds P Debrunner, JCM Tsibris, and E Munck (University of Illinois Press, Urbana), pp 22–24.
- Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254:1598–1603.
- Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. *Proc Natl Acad Sci USA* 89:20–22.
- Dinner AR, Sali A, Karplus M (1996) The folding mechanism of larger model proteins: Role of native structure. *Proc Natl Acad Sci USA* 93:8356–8361.
- Kauzmann W (1959) Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 14:1–63.
- Tanford C (1968) Protein denaturation. *Adv Protein Chem* 23:121–282.
- Agashe VR, Shastry MC, Udgaonkar JB (1995) Initial hydrophobic collapse in the folding of barstar. *Nature* 377:754–757.
- Dasgupta A, Udgaonkar JB (2010) Evidence for initial non-specific polypeptide chain collapse during the refolding of the SH3 domain of PI3 kinase. *J Mol Biol* 403:430–445.
- Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282:740–744.
- Richardson JS (1981) The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34:167–339.
- Fisher ME, Hiley BJ (1961) Configuration and free energy of a polymer molecule with solvent interaction. *J Chem Phys* 34:1253–1267.
- Orland HI, de Dominicis C (1985) An evaluation of the number of hamiltonian paths. *J Physique Lett* 46:353–357.
- Chan HS, Dill KA (1989) Compact polymers. *Macromolecules* 22:4559–4573.
- Dill KA (1985) Theory for the folding and stability of globular proteins. *Biochemistry* 24:1501–1509.
- White SH, Jacobs RE (1990) Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution. *Biophys J* 57:911–921.
- Camacho CJ, Thirumalai D (1993) Minimum energy compact structures of random sequences of heteropolymers. *Phys Rev Lett* 71:2505–2508.
- Thirumalai D, O'Brien EP, Morrison G, Hyeon C (2010) Theoretical perspectives on protein folding. *Annu Rev Biophys* 39:159–183.
- Yue K, Dill KA (1995) Forces of tertiary structural organization in globular proteins. *Proc Natl Acad Sci USA* 92:146–150.
- White SH, Jacobs RE (1993) The evolution of proteins from random amino acid sequences 1. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J Mol Evol* 36:79–95.
- Shakhnovich EI (1994) Proteins with selected sequences fold into unique native conformation. *Phys Rev Lett* 72:3907–3910.
- Zana R (1975) on the rate-determining step for helix propagation in the helix-coil transition of polypeptides in solution. *Biopolymers* 14:2425–2428.
- Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277:985–994.
- Ivankov DN, et al. (2003) Contact order revisited: Influence of protein size on the folding rate. *Protein Sci* 12:2057–2062.
- Wheeler SJ, Marchler-Bauer A, Bryant SH (2000) Domain size distributions can predict domain boundaries. *Bioinformatics* 16:613–618.
- Doolittle RF (1995) The multiplicity of domains in proteins. *Annu Rev Biochem* 64:287–314.
- Islam SA, Luo J, Sternberg MJ (1995) Identification and analysis of domains in proteins. *Protein Eng* 8:513–525.
- Sandhya S, et al. (2009) Length variations amongst protein domain superfamilies and consequences on structure and function. *PLoS One* 4(3):e4981.
- Ellis RJ, van der Vies SM (1991) Molecular chaperones. *Annu Rev Biochem* 60:321–347.
- Brooks BR, et al. (1983) Charmm—A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4:187–217.

32. Qiu LL, Pabit SA, Roitberg AE, Hagen SJ (2002) Smaller and faster: The 20-residue Trp-cage protein folds in 4  $\mu$ s. *J Am Chem Soc* 124:12952–12953.
33. Wolynes PG (1995) Biomolecular folding in vacuo!!!(?). *Proc Natl Acad Sci USA* 92:2426–2427.
34. Salisbury FB (1969) Natural selection and the complexity of the gene. *Nature* 224:342–343.
35. Smith JM (1970) Natural selection and the concept of a protein space. *Nature* 225:563–564.
36. Heringa J, Taylor WR (1997) Three-dimensional domain duplication, swapping and stealing. *Curr Opin Struct Biol* 7:416–421.
37. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926–935.