Why Are Proteins Marginally Stable?

Darin M. Taverna[1] **and Richard A. Goldstein**[1,2]*
[1]*Biophysics Research Division, University of Michigan, Ann Arbor, Michigan*
[2]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **Most globular proteins are marginally stable regardless of size or activity. The most common interpretation is that proteins must be marginally stable in order to function, and so marginal stability represents the results of positive selection. We consider the issue of marginal stability directly using model proteins and the dynamical aspects of protein evolution in populations. We find that the marginal stability of proteins is an inherent property of proteins due to the high dimensionality of the sequence space, without regard to protein function. In this way, marginal stability can result from neutral, non-adaptive evolution. By allowing evolving protein sub-populations with different stability requirements for functionality to complete, we find that marginally stable populations of proteins tend to dominate. Our results show that functionalities consistent with marginal stability have a strong evolutionary advantage, and might arise because of the natural tendency of proteins towards marginal stability. Proteins 2002;46:105–109.**

INTRODUCTION

Proteins have three major evolutionary constraints: they must fold to a structure in a reasonable time, the structure they fold to must perform a function, and the folded structure must be stable enough to perform that function reliably while resisting side-reactions such as aggregation and proteolysis. It has been noticed that most globular proteins are marginally stable, with a $\Delta G_{\text{folding}}$ of about -10 kcal/mol.[1–5] It has been repeatedly suggested that this observed marginal stability represents an adaptation for increased functionality, as marginal stability would be correlated with increased protein flexibility.[6–9] The basic perspective behind this interpretation is that, as proteins are adapted for specific traits including functionality, we can understand the observed properties of proteins by asking how these properties contribute to the required traits.

Starting with the neutral theory of evolution proposed by Kimura[10] and King and Jukes,[11] there has been increased interest in how evolutionary dynamics are affected by random processes and how much evolutionary change is non-adaptive. This perspective has brought about a challenge to the paradigm that properties of biological systems are explainable in terms of positive selection. According to this new viewpoint, identifying biological characteristics as adaptive is only appropriate if other evolutionary mechanisms can be rejected. Gould has also described how evolution can take advantage of characteristics that arise for non-adaptive reasons and use them for adaptive purposes; he calls these particular characteristics "spandrels."[12] In this way, we cannot even conclude that characteristics that fulfill an obviously adaptive purpose arose through positive selection. The rationalization of observed properties based on evolutionary adaptation may represent a "Panglossian paradigm," based on the character of Voltaire's *Candide* who characterizes our nose as an obvious adaptation to our need for spectacles and our legs as an adaptation for wearing trousers.[12] There has been increased interest by a number of different researchers in the role of neutral evolution in understanding the evolutionary process in biological macromolecules such as proteins and RNA.[13–22] Given these more recent theories, it is important to investigate whether common properties of proteins (such as marginal stability) can be rationalized as other than the result of positive selection. If so, the role of marginal stability in protein function and its importance in protein engineering must be re-examined.

In the past few years, we (and others) have concentrated on how the observed properties of proteins can often be explained by means of such concepts as "sequence entropy."[23–30] In this viewpoint, the total number of sequences (genotypes) consistent with a given property (phenotype) can have a strong effect on the probability of that property arising. In our work, we considered the constraints on a protein that it had to be able to fold on a reasonable timescale. We concentrated on a given thermodynamic characteristic ("foldability" $\mathcal{F}$) as a useful measure of this folding ability, and considered the mapping of sequences to foldability.[23] According to our models, we considered that there was a minimal "critical foldability" $\mathcal{F}_{\text{crit}}$ required for adequate folding, so that all sequences with $\mathcal{F} > \mathcal{F}_{\text{crit}}$ were considered viable (fitness equal to one) and all sequences with $\mathcal{F} < \mathcal{F}_{\text{crit}}$ were considered unviable (fitness equal to zero). Using computational and analytical models, we described how many more sequences would fold into certain structures compared with others, explaining why some structures were so common,[24] why proteins

Grant sponsor: NIH; Grant numbers: LM05770, GM08270; Grant sponsor: NSF; Grant number: BIR9512955.

Darin M. Taverna's present address is Protein Pathways, Inc., 1145 Gayley Ave., Suite 304, Los Angeles, CA 90024.

*Correspondence to: Richard A. Goldstein, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055. E-mail: richardg@umich.edu

Received 23 March 2001; Accepted 10 August 2001

DOI 10.1002/prot.10016

that fold under kinetic control would most likely evolve so that the native state was the state of lowest free energy fulfilling the so-called "thermodynamic hypothesis,"[31] and why structures are so robust to changes in sequence.[32,33] We also demonstrated that the observed properties are highly influenced by the fact that evolutionary change involves population effects.[34] One specific observation of our model was the observation that most viable sequences are minimally-viable, that the vast majority of sequences with $\mathcal{F} > \mathcal{F}_{\text{crit}}$ have $\mathcal{F} \approx \mathcal{F}_{\text{crit}}$.[32] This resulted in the prediction that most proteins would be marginally foldable.

In this paper, we turn our attention to protein stability, a quantity related to foldability. In particular, we are interested in using our computational models to consider whether marginal stability can arise for non-adaptive reasons in the same way that marginal foldability developed in our previous models. In addition, we examine the role that population effects play in the resulting distribution of stabilities. Finally, we consider how functionality might arise in the context of marginally-stable proteins, and how marginal stability might represent a spandrel, a non-adaptive property later utilized for an adaptive purpose. In this manner, protein functionality that requires conditions of marginal stability may be the result of proteins being naturally marginally stable, rather than functionality imposing a selective pressure towards marginal stability.

MODEL

Our protein model consists of a chain of 25 monomers confined to a 5×5 two-dimensional maximally compact square lattice, with each monomer located at one lattice point. This provides us with 1,081 possible conformations represented by the 1,081 self-avoiding walks on this lattice not including structures related by rotation, reflection, or inversion. We assume that the energy of any sequence in conformation k is given by a simple contact energy of the form $E = \Sigma_{i<j} \gamma(\mathcal{A}_i\mathcal{A}_j) \Delta_{ij}^k$ where Δ_{ij}^k is equal to 1 if residues i and j are not covalently connected but are on adjacent lattice sites in conformation k and 0 otherwise, and $\gamma(\mathcal{A}_i\mathcal{A}_j)$ is the contact energy between amino acids $\mathcal{A}_i$ at location i and $\mathcal{A}_j$ at location j in the sequence. These contact energies represent potentials of mean force derived by Miyazawa and Jernigan based on a statistical analysis of the database of known proteins, and implicitly includes the enthalpic and entropic contributions due to interactions of the protein with the solvent.[35] There are 132 pairs of residues that can possibly come into contact with 16 of these contacts present in any given compact structure.

A characteristic universal to most all proteins is the ability to be stable in their folded state so as to function while avoiding proteolysis, aggregation, or initiation of an immune response. We characterize a protein structure's stability with its $\Delta G_{\text{folding}}$, defined as $\Delta G_{\text{folding}} = -kT \ln(P_f/P_u)$, where P_f and P_u refer to the probability of finding a given sequence folded in its native state or in the ensemble of unfolded states, respectively. We make the assumption that the thermodynamic hypothesis is obeyed and that the lowest energy structure is the native state.[31] The other 1,080 possible structures represent the ensemble of unfolded states. (The non-compact states were neglected in order to allow for a reasonable number of stable sequences. Alternatively, we would expect the non-compact states to be neglectible as long as the contact energies were sufficiently attractive. The fact that most protein structures are reasonably compact makes this assumption not too unreasonable). Assuming a Boltzmann distribution, we can express the free-energy of folding as equal to $\Delta G_{\text{folding}} = E_f + kT \ln(Z - \exp(-E_f/kT))$ where E_f is the energy of the folded state and Z is the partition function.

As mentioned above, proteins have to be sufficiently stable. We model this by considering that a viable protein requires $\Delta G_{\text{folding}}$ less than some specified ΔG_{crit}. The use of lattice models allows us to analyze evolutionary processes ignoring functional constraints. In order to explore the relationship between sequence space and corresponding stability, we followed the evolution of a single protein diffusing about the sequence landscape for 10 million generations under the constraint that the native state remain fixed at a predetermined structure. Starting with an initial sequence chosen at random from all viable sequences, amino acids were randomly mutated with the number of mutations chosen from a Poisson distribution with an average of one mutation per generation. The stability of the new sequence was calculated; if $\Delta G_{\text{folding}}$ was larger than $\Delta G_{\text{crit}} = 0$ or the structure had changed, the mutation was rejected and the original sequence retained. Generations where no mutations occurred were not counted. This is analogous to random-walk models in which the particle has average zero velocity when a boundary is encountered. This was done 5 times, each with a different seed sequence, but using the same structure.

We have shown previously how population dynamics can effect evolutionary processes.[34] To demonstrate this effect, we modeled population dynamics by constructing an initial population of $N = 3{,}000$ identical sequences. The population was allowed to equilibrate for 30,000 generations before data was recorded; data was then accumulated for 30,000 additional generations. For all post-equilibrium generations, each residue in every protein was chosen with probability 0.2% to be mutated to another random residue; both the population size and mutation rate were chosen to be comparable to previous analytical models of evolution processes.[36–38] The stability of each protein in the population was then calculated. The N' sequences with $\Delta G_{\text{folding}} < \Delta G_{\text{crit}} = 0$ and a conserved native state structure were considered viable and capable of reproducing. The next generation of N sequences was chosen from the N' surviving sequences randomly with replacement, representing the stochastic process of reproduction.

The results of these studies, described more fully below, indicate that the marginal stability of observed proteins can be adequately explained by considerations independent of protein functionality. We are then interested in understanding how functionality would evolve in the

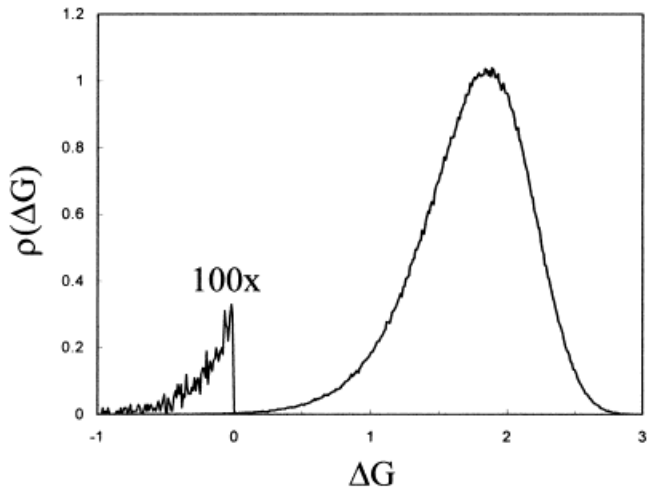

Fig. 1. Distribution of stabilities of randomly-chosen sequences. The distribution of stabilities in the region $\Delta G_{\text{folding}} < 0$ are shown multiplied by a factor of 100.

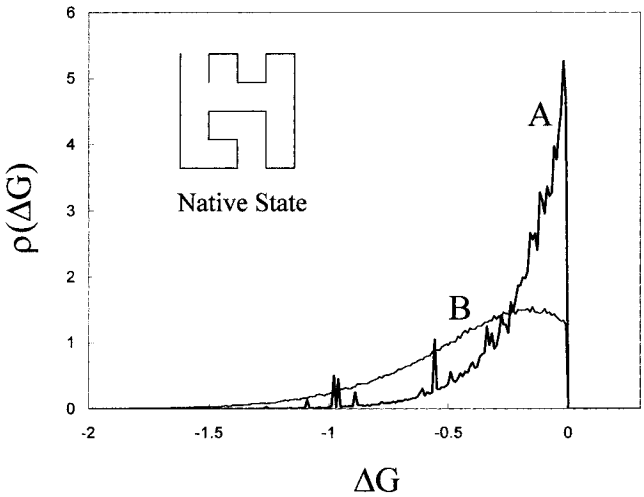

Fig. 2. **A:** Distribution of stabilities produced by single-sequence walks with structure conserved. **B:** Distribution of stabilities produced by population evolution with structure conserved. The structure used is shown in the upper left corner.

context of marginally stable proteins. If functionality required marginal stability, then the natural propensity towards such marginal stability would assist in biological evolution. Conversely, we can imagine that the natural proclivity towards marginal stability would give a great evolutionary advantage to mechanisms consistent with this property. We can demonstrate the latter effect by considering the evolutionary competition between alternative functionalities with differing stability requirements, keeping other requirements (such as native-state structure) constant. Specifically, we considered three sub-populations (Φ_1, Φ_2, Φ_3) with three different functional mechanisms with stability requirements $0 \geq \Delta G_{\text{folding}}(\Phi_1) > -1$, $-1 \geq \Delta G_{\text{folding}}(\Phi_2) > -2$, $-2 \geq \Delta G_{\text{folding}}(\Phi_3) > -3$, respectively. Each sub-population was created with $N = 3{,}000$ identical sequences, the population dynamics described above were implemented, and each sub-population was independently equilibrated for 30,000 generations. The subpopulations were then combined into one population of size $3N$. Population dynamics were resumed with this larger set of proteins, with the sequences retaining a memory of their original sub-population so that only sequences with stabilities compatible with the requirements for the appropriate sub-population were considered viable and capable of reproducing. The next generation of $3N$ sequences was chosen randomly from all of the various viable members of the three sub-populations. The number of sequences in each sub-population was measured until two sub-populations became extinct as one sub-population overcame the entire population. This was done for 5 different initial populations; 5 independent trials were run for each population.

RESULTS

Figure 1, which shows the distribution of values of $\Delta G_{\text{folding}}$ of 1 million randomly generated protein sequences, demonstrates that most proteins are naturally unstable in this model. Only 0.15% of the sequences have $\Delta G_{\text{folding}} \leq 0$. The distribution of values of $\Delta G_{\text{folding}}$ for a single sequence diffusing with a fixed structure is shown in Figure 2 (curve A). This measures the distribution of stabilities of sequences corresponding to this structure. The distributions in Figure 1 and 2 (curve A) show how most stable random sequences lie close to the ΔG_{crit} barrier, independent of structural considerations. Figure 2 (curve B) shows the corresponding distribution of stabilities during the population evolution. Negligible differences in the shape of this distribution were found when the population size N was varied between 1,000 and 10,000 (data not shown). The random distribution of Figure 2 (curve A) is not reproduced exactly when population effects are included, as shown in Figure 2 (curve B). This is due to the greater probability of a sequence with $\Delta G_{\text{folding}} \sim \Delta G_{\text{crit}}$ to mutate to a non-viable sequence with $\Delta G_{\text{folding}} > \Delta G_{\text{crit}}$ and, therefore, being removed from the population; there is an effective population "sink" at this critical barrier. This effect notwithstanding, the result of these dynamics is a flux of the population to $\Delta G_{\text{folding}}$ slightly below ΔG_{crit}.

When multiple sub-populations competed with different stability requirements, subpopulation Φ_1 became the sole remaining sub-population in 24 of the 25 runs. Figure 3(a) shows the results of a typical run in which only subpopulation Φ_1 survived. Figure 3(b) shows a run with the same end-result, but where the randomness of the population evolution resulted in a competition lasting six times longer than normal. Figure 3(c) shows the one atypical run in which Φ_2 became the sole surviving sub-population. Figure 3(d) shows the average of all 25 runs.

DISCUSSION

We can understand the results of these simulations by considering the high-dimensional space of all possible sequences. In this space, each dimension represents one

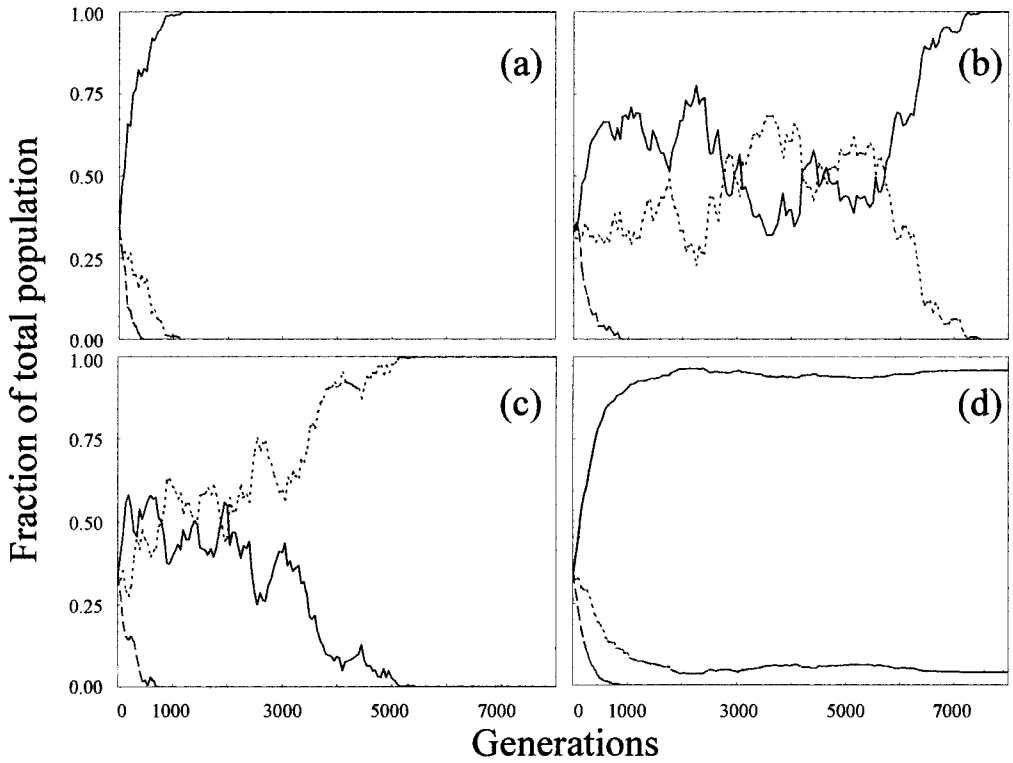

Fig. 3. **a:** Typical sub-population competition trial in which Φ_1 survived. **b:** Sub-population trial in which Φ_1 survived, yet competed with Φ_2 six times longer than normal. **c:** Single instance where Φ_2 survived. **d:** Average of all 25 sub-population competition trials. In each sub-figure, the data shown is represented by Φ_1 (——), Φ_2 (- - - -), and Φ_3 (– – – –).

location in the sequence, so that there are as many dimensions as residues in the protein. There exist clusters in this space corresponding to stable sequences, with $\Delta G_{\text{folding}} \leq \Delta G_{\text{crit}}$. We can consider the fringes of the cluster, where the sequences have decreased similarity with the average, prototypical, or maximally stable sequences; assuming that the stability of this space is somewhat continuous, we would expect that these sequences would have values of $\Delta G_{\text{folding}}$ close to the value of ΔG_{crit}, that is, to have marginal stability. Due to the high-dimensionality of the protein space, the vast majority of the viable sequences are clustered in this marginally stable fringe region.

As can be seen in Figures 1 and 2, the number of available sequences decreases quickly as the stability is increased, and increases sharply with decreasing stability. This results in a high probability that any attempted mutation will be destabilizing compared to a rarity of stabilizing mutations, as is experimentally observed. We would expect this bias to be smaller for mutations that cause small changes in stability, and larger for more significant mutations. The rationalization for this observation is highly related to the fact that the probability of a mutation resulting in improved fitness is a strongly-decreasing function of the size of the fitness change, an argument first made by Fisher in 1930.[39]

This tendency is independent of any need for functionality. In fact, as shown in Figure 3, the intrinsic tendency of proteins to evolve towards marginal stability results in a selective pressure that favors functionality consistent with marginal stability. Because of this effect, even if there were a variety of mechanisms possible for protein functionality, the observed functionalities would likely be consistent with (and possibly require) marginal stability. In fact, the entire system of biochemistry with its emphasis on weak and subtle non-covalent molecular interactions, might represent nature's method to take advantage of this natural tendency. The larger number of sequences available to marginally stable proteins would increase sequence plasticity, and due to the greater variations in residue composition, the ability of proteins to acquire new functions would also be favored. Marginal stability may represent a "spandrel," a naturally occurring tendency that biology can use for its own advantage. If so, this tendency has provided biology with a robust system characterized by easy adaptation of new functionalities.

ACKNOWLEDGMENTS

We thank Ting-Lan Chiu and John Gland for helpful comments and Todd Raeker for computational assistance. Financial backing was provided in part by NSF shared-equipment grant BIR9512955.

REFERENCES

1. Savage HJ, Elliot CJ, Freeman CM, Finney JL. Lost hydrogen-bonds and buried surface-area: Rationalizing stability in globular-

proteins. J Chem Soc Faraday Trans 1993;89:2609–2617.
2. Vogl T, Jatzke C, Hinz HJ, Benz J, Huber R. Thermodynamic stability of annexin V E17G: Equilibrium parameters from an irreversible unfolding raction. Biochemistry 1997;36:1657–1668.
3. Ruvinov S, Wang L, Ruan B, Almog O, Gilliland GL, Eisenstein E, Bryan PN. Engineering the independent folding of the subtilisin bpn′ prodomain: Analysis of two-state folding verses protein stability. Biochemistry 1997;36:10414–10421.
4. Giver L, Gershenson A, Freskgard PO, Arnold FH. Directed evolution of a thermostable esterase. Proc Nat Acad Sci USA 1998;95:12809–12813.
5. Privalov PL, Khechinashvili NN. A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. J Mol Biol 1974;86:665–684.
6. Rasmussen BF, Stock A, Ringe D, Petsko GA. Crystalline ribonuclease A loses function below the dynamical transition at 220k. Nature 1992;357:423–424.
7. Zavodszky P, Jozsef K, Svingor A, Petsko GA. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. Proc Nat Acad Sci USA 1998;98:7406–7411.
8. Tsou CL. Active site flexibility in enzyme catalysis. Enzyme Eng XIV 1998;864:1–8.
9. Li H, Tang C, Wingreen N. Are protein folds atypical? Proc Nat Acad Sci USA 1998;95:4987.
10. Kimura M. Evolutionary rate at the molecular level. Nature (Lond) 1968;217:624–626.
11. King JL, Jukes TH. Non-Darwinian evolution. Science 1969;164: 788–798.
12. Gould SJ, Lewontin RC. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. Proc R Soc London B 1979;205:581–598.
13. Lipman DJ, Wilbur WJ. Modelling neutral and selective evolution of protein folding. Proc R Soc London B 1991;245:7–11.
14. Schuster P, Fontana W, Stadler PF, Hofacker IL. From sequences to shapes and back: A case study in RNA secondary structures. Proc R Soc London B 1994;255:279–284.
15. Bornberg-Bauer E. How are model protein structures distributed in sequence space? Biophys J 1997;73:2393–2403.
16. Babajide A, Hofacker IL, Sippl MJ, Stadler PF. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. Fold Design 1997;2:261–269.
17. Fontana W, Schuster P. Continuity in evolution: On the nature of transitions. Science 1998;280:1451–1455.
18. Bastolla U, Roman HE, Vendruscolo M. Neutral evolution of model proteins: Diffusion in sequence space and overdispersion. J Theor Biol 1999;200:49–64.
19. Bourdeau V, Ferbeyre G, Pageau M, Paquin B, Cedergren R. The distribution of RNA motifs in natural sequences. Nucl Acids Res 1999;27:4457–67.
20. Ancel LW, Fontana W. Plasticity, evolvability, and modularity in RNA. J Exp Zool 2000;288:242–283.
21. Forst CV. Molecular evolution of catalysis. J Theor Biol 2000;205: 409–231.
22. Reidys C, Forst CV, Schuster P. Replication and mutation on neutral networks. Bull Math Biol 2001;63:57–94.
23. Govindarajan S, Goldstein RA. Searching for foldable protein structures using optimized energy functions. Biopolymers 1995;36: 43–51.
24. Govindarajan S, Goldstein RA. Why are some protein structures so common? Proc Nat Acad Sci USA 1996;93:3341–3345.
25. Finkelstein AV, Ptitsyn OB. Why do globular proteins fit the limited set of folding patterns. Prog Biophys Mol Biol 1987;50:171–190.
26. Finkelstein AV, Reva B. A search for the most stable folds of protein chains. Nature (Lond) 1991;351:497–499.
27. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science 1996;273: 666–669.
28. Shakhnovich EI. Protein design: A perspective from simple tractable models. Fold Design 1998;3:R45–R58.
29. Kussell EL, Shakhnovich EI. Analytical approach to the protein design problem. Phys Rev Lett 1999;83:4437–4439.
30. Pande VS, Grosberg AY, Tanaka T. Heteropolymer freezing and design: Towards physical models of protein folding. Rev Mod Phys 2000;72:259–314.
31. Govindarajan S, Goldstein RA. On the thermodynamic hypothesis of protein folding. Proc Nat Acad Sci USA 1998;95:5545–5549.
32. Govindarajan S, Goldstein RA. The foldability landscape of model proteins. Biopolymers 1997;42:427–438.
33. Govindarajan S, Goldstein RA. Evolution of model proteins on a foldability landscape. Proteins 1997;29:461–466.
34. Taverna D, Goldstein RA. The distribution of structures in evolving protein populations. Biopolymers 2000;53:1–8.
35. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 1985;18:534–552.
36. Ohta T. Multigene and supergene families. Oxford Surv Evol Biol 1988;5:41–65.
37. Kimura M. The neutral theory of molecular evolution. Sci Am 1979;241:98–126.
38. Ohta T. Simulating evolution by gene duplication. Genetics 1987; 115:207–213.
39. Fisher RA. The genetical theory of natural selection. Oxford: Oxford University Press; 1930.