Commentary

Gō-ing for the prediction of protein folding mechanisms

Shoji Takada*

Department of Chemistry, Faculty of Science, Kobe University, Kobe, 657-8501, Japan

Protein folding has been a long-lived problem in biophysics. Much important progress has been made in the 90s by focusing on small single-domain proteins (1). In particular, (*i*) site-resolved measurement of the folding transition state ensemble, quantified as ϕ-values (2), made it possible to understand folding mechanisms relatively unambiguously, stimulating interaction between experimentalists and theoreticians; (*ii*) the energy-landscape theory (3) gave us a general framework based on statistical physics; and (*iii*) the finding of a significant correlation between folding rates and native structure topology (4), which suggested that the native topology is a key determinant of folding mechanisms, all lead us to believe that the underlying physics could be relatively simple. These three ingredients are linked together with an almost one-line free energy equation in three papers (5–7), which appeared in a recent issue of PNAS, as well as some previous work (8, 9). The surprise of the three papers is that apparently one can have both simplicity and fair predictability. Papers by Galzitskaya and Finkelstein (5), Alm and Baker (6), and Muñoz and Eaton (7), which are independent but resemble each other greatly, report that even highly simplified theories based on energy-landscape ideas can predict trends in the folding rates for many fast folding proteins.

The calculations of the three papers are easy to summarize. In all three papers, each amino acid (or a few adjacent amino acids grouped) is taken to be either in a native configuration (n) or in a completely random set (r). A reduced protein configuration, or *microstate*, is represented as a sequence of n and r, such as rrrr-nnnnnn-rrrrrr-nnnnnnn. The (globally) native and denatured states correspond to the microstate in which all amino acids are in n and r, respectively. The authors introduce simple free energies for microstates given the native three-dimensional structure. Assuming an elementary step to be a change in one amino acid from r to n, the researchers look for the pathways that connect native and denatured states in the microstate space. On each pathway, there is a maximum along the reaction coordinate number of native-like amino acids. The investigators assign this maximum as a member of transition state ensemble (see Fig. 1). Finally, they compare characteristics of their transition states with those measured experimentally. Galzitskaya and Finkelstein (5) as well as Alm and Baker (6) analyzed the structure of the transition states and compared their estimated ϕ-values with the experimental values. The researchers found significant correlation coefficients ($r > 0.4$) for several proteins but with some exceptions ($r \approx 0$). Muñoz and Eaton (7), in addition to the ϕ-value test, focus more on the free-energy barrier height at the transition state and the corresponding total folding rate measured in kinetic experiments. In their paper, figure 5 shows the excellent agreement of folding rate coefficients with those measured in experiments.

One underlying assumption for these analyses is common to all three papers as well as previous, more elaborate treatments (8, 9): all of these theories take into account only the interactions that are present in the native structure, the so-called native interactions. The theories assume that nonnative interactions do not contribute to the global shape of the folding energy surface. We often call this class of models Gō models, because this type of interaction was first introduced in old lattice simulation work by Gō and coworkers (ref. 10 and references therein). [I also would like to mention that, inspired by Gō's work, Miyazawa and Jernigan (11) developed a theory rather parallel to the three papers (5–7). Without good experiments, however, their work received little attention.] The Gō model is one realization of a central idea of the energy landscape theory that the energy landscape resembles a funnel (3). The landscape theory, however, incorporates both the global funnel shape toward the native basin and the ruggedness of the energy landscape which it views as arising from unavoidable frustration upon refolding. While the Gō model captures the first feature, the latter aspect is entirely ignored. The crudeness of the approximation, which is an extreme limiting case of a folding funnel, would lead one to doubt its applicability for quantitatively calculating properties of the ensemble of folding pathways for specific proteins. But, surprisingly enough, the perfect funnel models give reasonable prediction even up to the site-resolved level.

If the interactions included are all attractive to the native structure, why does a barrier or transition state exist for folding at all? Upon folding of small single-domain proteins, the solvent averaged free energy of a chain configuration E almost always decreases because most interactions are favorable, whereas acquisition of specific order results in the loss of chain entropy S. At temperature T, the free energy $F = E - TS$ has a barrier along the reaction coordinate, because chain entropy is lost at a somewhat earlier stage than energetic stability is gained. In all of the above mentioned theories, folding pathways are determined in such a way that the protein gains a given number of native contacts while keeping entropy loss as small as possible. A local contact is more favorable than a nonlocal one in the early stage of folding, because a smaller number of peptides need to be pinned down by a local contact. For the protein as a whole, the ensemble of pathways is controlled by the distribution of native contacts and the latter is determined by the topology of the native structure based on the perfect funnel approximation.

It is interesting to compare these papers and related works (5–9) more closely, because there are indeed some differences in formulation that apparently do not perturb agreement with experiments. For interaction energies, all the investigators, except Alm and Baker (6), use pair-contact models as major interactions. Shoemaker *et al*. (8) employ local hydrogen bond interaction, whereas others do not. Alm and Baker (6) employ an accessible surface-area form of interaction without hydrogen bonding. The chain-entropy terms in each case consist of an

See companion articles on pages 11299, 11305, and 11311 in issue 20 of volume 96.

*To whom reprint requests should be addressed. E-mail: stakada@kobe-u.ac.jp.

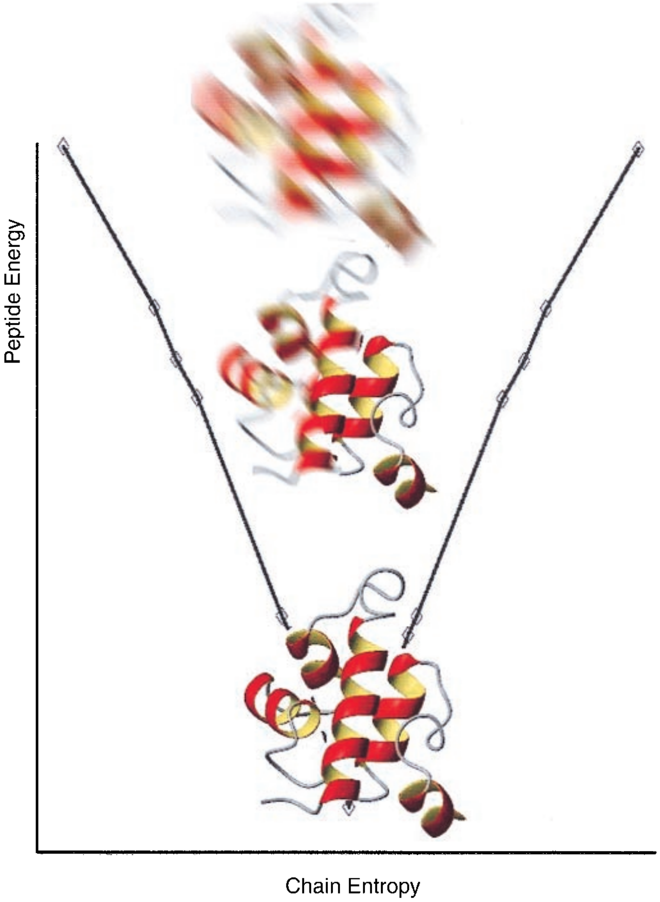

Fig. 1. Schematic pictures of a statistical ensemble of proteins embedded in a folding funnel for a monomeric λ-repressor domain. At the top, the protein is in its denatured state and thus fluctuating wildly, where both energy and entropy are the largest. In the middle, the transition state forms a minicore made of helices 4 and 5 and a central region of helix 1 drawn in the right half, whereas the rest of protein is still denatured. At the bottom is the native structure with the lowest energy and entropy. The figure is based on ref. 9, and the protein structures pictured here were prepared with a graphics package, MOLMOL (22).

entropy loss term for individual amino acids and another term related to the closure of a denatured loop, for which the explicit form differs in each paper. Among the five, the paper by Portman *et al*. (9) is the most microscopic and computationally demanding treatment, in that it starts with a Hamiltonian and computes energy and entropy by the Boltzmann average, whereas others *a priori* assume some explicit free-energy formula. It would seem that the folding barrier height is determined as consequence of subtle cancellation between energy and entropy. One would therefore think the barrier's magnitude to be a very difficult quantity to estimate accurately. Nevertheless, Muñoz and Eaton (7) succeeded in estimating (relative) barrier heights at the transition states for a variety of proteins, whereas the others did not show such a wide range of results. Interestingly, although some other authors are sanguine about the estimation of ϕ-values, showing the structure of the translation state, Muñoz and Eaton (7) are concerned that such a comparison is difficult.

Despite the differences between models and treatments, all of them work to comparable accuracy, implying that the strict form of the interaction may not be crucial and that the ensemble of the folding pathways is quite robust against small perturbations in the interaction. Folding pathways of small single-domain proteins are mostly determined by the native structural information and energy-entropy compensation.

There are still many questions that need to be addressed soon by this class of theories. First, the breadth of the transition states and the movements of the transition state upon change in the solvent condition have been monitored recently in protein engineering experiments (12), and the corresponding protein-specific theory is highly desired. Shoemaker *et al*. (13) have already developed their model in this direction. Second, the compactness of the transition state quantified as Tanford's β (1) may be another challenge for theoreticians to calculate. It is also interesting to investigate the diversity of folding trajectories. Whether there are two or three distinct folding pathways that are well separated or only one broadly distributed pathway would be interesting to investigate (14). How this feature is microscopically determined is also a question.

How far can we go with Gō models? There is much evidence that indicates there are limits to the perfect funnel model description. First, there is a small but nonnegligible number of residues that have ϕ-values larger than unity or smaller than zero. These have been interpreted as the effects of interactions that are not, or are weaker, at the native structure but are important in the transition ensemble. This is a sign of the existence of frustration. It should be noted that these are often found when mutations change the stability only slightly. Perhaps, in estimating the experimental ϕ-value, we suffer from small denominators that can cause significant errors in ϕ-values. Therefore, how much nonnative pairs affect measured ϕ-values is not experimentally clear. In these cases, we can incorporate some specific nonnative interactions in the Gō model for clarifying their roles. Another class of evidence that indicates limits of perfect funnel models is the existence of clearly nonnative intermediates. Some mainly β-proteins, such as β-lactoglobulin (15), are known to have transiently more α-helical contents. In this case, nonnative structures appear globally in the protein, and simple Gō models used so far cannot describe this.

Finally, it is worth contemplating the questions of nonnative interactions in general, that is, how does specificity arise? As mentioned above, the energy-landscape theory includes two aspects; the global funnel shape toward the native structure, caused by native interactions, and the ruggedness of the energy surface, caused by nonnative interactions. Clearly, the success of the prediction of folding pathways with Gō models demonstrates the primacy of the former aspect. The agreement with experiment suggests that for fast folding *natural* proteins under physiological conditions, the funnel aspect seems to be primarily important. Interestingly, the other aspect, the ruggedness of the landscape, seems to be more important for *human scientists* trying to imitate real proteins both in the laboratory and on computers (16, 17). In the *de novo* design of protein-like peptides, the so-called *negative* design is crucial for success in finding a peptide sequence which can have unique structure. This negative design is closely related to minimizing the ruggedness of the energy surface in the language of the energy-landscape theory. "Realistic" simulations, i.e., currently available protein models without Gō bias, from full-atom models (18, 19) to minimal models, seem to have an energy landscape that is more rugged than experiment would suggest. Apparently, real proteins have almost perfectly favorable interactions of many different kinds near the native structure, and errors in modeling any interaction almost always destabilizes the native structure relative to the true interaction. For better modeling, the potential functions need to be optimized such that the ruggedness is minimized while the depth of the native basin is kept sufficiently large (20, 21). Although making models that can recognize the native structure is an ultimate goal for practical people interested in structure prediction, as of this writing, Gō models may still be closer to reality than our current *realistic* models, at least for describing kinetics of fast folding proteins.

1. Fersht, A. R. (1999) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York).
2. Itzhaki, L. A., Otzen, D. E. & Fersht, A. R. (1995) *J. Mol. Biol.* **254**, 260–288.
3. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995) *Proteins Struct. Funct. Genet.* **21**, 167–195.
4. Plaxco, K. W., Simons, K. T. & Baker, D. (1998) *J. Mol. Biol.* **277**, 985–994.
5. Galzitskaya, O. V. & Finkelstein, A. V. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 11299–11304.
6. Alm, E. & Baker, D. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 11305–11310.
7. Muñoz, V. & Eaton, W. A. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 11311–11316.
8. Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 777–782.
9. Portman, J. J., Takada, S. & Wolynes, P. G. (1998) *Phys. Rev. Lett.* **81**, 5237–5240.
10. Gō, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.
11. Miyazawa, S. & Jernigan, R. L. (1982) *Biochemistry* **21**, 5203–5213.
12. Oliveberg, M., Tan, Y.-J., Silow, M. & Fersht, A. R. (1998) *J. Mol. Biol.* **277**, 933–943.
13. Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1999) *J. Mol. Biol.* **287**, 675–694.
14. Pande, V. S., Grosberg, A. Y., Tanaka, T. & Rokhsar, D. S. (1998) *Curr. Opin. Struct. Biol.* **8,** 68–79.
15. Hamada, D., Segewa, S. & Goto, Y. (1996) *Nat. Struct. Biol.* **3**, 868–873.
16. Street, A. G. & Mayo, S. L. (1999) *Structure* **7**, R105–R109.
17. Shakhnovich, E. I. (1998) *Folding Des.* **3**, R45–R58.
18. Lazardis, T. & Karplus, M. (1997) *Science* **278,** 1928–1931.
19. Duan, Y. & Kollman, P. A. (1998) *Science* **282,** 740–744.
20. Koretke, K. K., Luthey-Schulten, Z. & Wolynes, P. G. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 2932–2937.
21. Hao, M.-H. & Scheraga, H. A. (1999) *Curr. Opin. Struct. Biol.* **9**, 184–188.
22. Koradi, R., Billeter, M. & Wüthrich, K. (1996) *J. Mol. Graphics* **14,** 51–55.