Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 1229–1233, February 1989
Biophysics

Dynamic Monte Carlo study of the folding of a six-stranded Greek key globular protein

(all-or-none transition/multiple domain protein/plastocyanin model)

JEFFREY SKOLNICK[†‡], ANDRZEJ KOLINSKI[§], AND ROBERT YARIS[†]

[†]Institute of Macromolecular Chemistry, Department of Chemistry, Washington University, Saint Louis, MO 63130; and [§]Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

Communicated by Samuel I. Weissman, November 14, 1988

ABSTRACT **To help elucidate the general rules of equilibrium globular protein folding, dynamic Monte Carlo simulations of a model β-barrel globular protein having the six-stranded Greek key motif characteristic of real globular proteins were undertaken. The model protein possesses a typical β-barrel amino acid sequence; however, all residues of a given type (e.g., hydrophobic residues) are identical. Even in the absence of site-specific interactions, starting from a high-temperature denatured state, these models undergo an all-or-none transition to a structurally unique six-stranded β-barrel. These simulations suggest that the general rules of globular protein folding are rather robust in that the overall tertiary structure is determined by the general pattern of hydrophobic, hydrophilic, and turn-type residues, with site-specific interactions mainly involved in structural fine tuning of a given topology. Finally, these studies suggest that loops may play an important role in producing a unique native state. Depending on the stability of the native conformation of the long loop in the Greek key, the conformational transition can be described by a two-state, three-state, or even larger number of multiple equilibrium states model.**

The ability to predict the three-dimensional (tertiary) structure of a globular protein given the amino acid sequence has been a long desired objective of biochemistry (1–6). Although the equilibrium folding and unfolding of proteins has received substantial attention, a surprising number of questions remain unanswered. Does the folded structure result primarily from local interactions, perhaps involving adjacent residues, or are tertiary interactions involving pieces of the protein that are spatially close but rather far down the chain contour dominant (7–9)? More generally, what is the level of detail required for tertiary structure prediction? Is a very detailed free energy surface required before tertiary structure prediction becomes a reality (10–16)? Or, are the general rules of folding rather robust in that a small number of general criteria are responsible for producing a given topology (17, 18)? If so, then, the myriad of local details would enter in the "fine tuning" of the native conformation. Theoretical models should prove very useful in identifying the essential elements of globular protein folding. This article demonstrates that the conformational transition from the denatured state **1** of Fig. 1*A* to the Greek key motif seen in many β-barrel globular proteins (19) and illustrated in **2** of Fig. 1*A*, in a side view, and Fig. 1*B*, in a top view, can be reproduced from a rather simple protein model, whose details are elaborated on below.

Unlike alternative theoretical approaches that construct a model protein by using the most realistic available potential energy surface (10–16), we have examined a class of models having a minimal set of interactions (17, 18); yet, they reproduce in a qualitative sense a number of the essential features of the glo‿ular protein conformational transition. When using a schematic model, one must be sure that it embodies the essential physics of the real system. For globular proteins, the model must be able to sample all of configuration space; interactions between all of the residues must be allowed and the system must find its way to the native state. For small proteins, the collapse transition must be thermodynamically all-or-none (6, 20–25). That is, the system must spend the majority of its time either in the random coil state or in the completely folded native state, and folding intermediates must be sparsely populated. Furthermore, the collapse must always be to the same native structure. Of course, small local fluctuations in the native state should occur just as in real proteins (26, 27).

In the present paper, we demonstrate that a general pattern of hydrophobic and hydrophilic residues plus the presence of appropriately placed regions that have, based on tertiary interactions, a statistical preference to form loops and turns, are sufficient to fold a six-stranded Greek key β-barrel **2** of Fig. 1, whose topology is very close to that of plastocyanin (19, 28, 29). The requirements for folding to the unique Greek key **2** are similar to those seen in earlier (and simpler) simulations on the folding of a four-stranded β-barrel (18). For a variety of loop and turn stabilities, these models exhibit all the essential features of the equilibrium folding transition found in real globular proteins. The present study once again indicates that the general rules of protein folding are rather robust; a very complicated native state topology can be reproduced without introducing site-specific interactions. Hence, these studies appear to point the way toward a general theory of protein tertiary structure prediction.

MODEL

To facilitate the sampling of the important regions of configuration space, an α-carbon representation of a globular protein confined to a diamond lattice is used. The model protein contains $n = 74$ beads, each representing an amino acid residue, which may be hydrophobic, hydrophilic, or inert. Multiple occupancies of all lattice sites are prohibited, thereby implementing excluded volume. Each of the $n - 3$ interior bonds has three allowed rotational states, the planar *trans* (t) and either of the two out of plane *gauche* (g^+ or g^-) states. Since we are interested in forming antiparallel β-proteins, and a sequence of *trans* states produces the β-sheet conformation, ε_g, the intrinsic energy of a *gauche* relative to a *trans* state, is taken to be positive, unless otherwise indicated. The reduced temperature scale is defined as $T^* = k_B T/\varepsilon_g$, where k_B is Boltzmann's constant and T is the absolute temperature. Typical values of $\varepsilon_g/k_B T$ in the transition region are 0.7–0.8.

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.

Abbreviation: MC, Monte Carlo.

[‡]To whom reprint requests should be addressed.

Imagine a nonbonded pair of nearest-neighbor residues. If both are hydrophobic, then ε_h (negative) is the attractive potential of mean force that mimics the hydrophobic interaction. ε_h might also mimic the interaction if the pair forms a salt bridge. If one of the beads is hydrophobic and the other is hydrophilic or if both are hydrophilic, then ε_w (positive) mimics the repulsive potential of mean force. The value of the interaction parameter ε_h or ε_w depends only on the kind of amino acid pair and not their location in the sequence. In addition, a cooperativity parameter, ε_c, which allows for direct conformational coupling between any two nonbonded *trans* states is used. References 17 and 18 present a more detailed discussion of ε_c.

The primary sequence is specified as follows. $B_i(k)$ represents the ith stretch in the primary sequence containing k consecutive residues. A given stretch need not necessarily form a β-pleated sheet. All the ε_g in the $B_i(k)$ are the same. Regions that, based on tertiary interactions, might have a statistical tendency to form tight bends are denoted by b_i and consist of the last two residues of region i and the first bead of region i + 1. For all three residues in b_i, ε_c, ε_h, and ε_w are zero, but ε_g isn't necessarily zero. Finally, the amino acid sequence of the long loop joining β-strands 5 and 6 of 2 in Fig. 1 is specified by $L(k)$. For these putative loop residues ε_c = 0, but ε_g, ε_h, and ε_w need not necessarily be zero.

In all cases, a primary sequence pattern $B_1(11)b_1B_2(11)b_2$-$B_3(11)b_3B_4(12)b_4B_5(11)L(7)B_6(11)$ is used, and the β-strands of the native state contain 11 or 12 residues. For all B_i with i = 1, . . . 5, the odd (even) residues are all hydrophobic (hydrophilic). Because of the lattice, the even hydrophilic residues on strand 2 are nearest neighbors to the odd hydrophobic residues on strand 5. Similarly, the even hydrophilic residues of strands 1 and 3 are nearest neighbor to the even hydrophobic residues of strand 6. However, in the native conformation all the hydrophobic residues of strands 5 and 6 have the apex of their bond pointing in the direction of the protein interior, whereas the hydrophilic-type residues point out into the solvent; thus, we set the interactions between these residues for all (and not necessarily native) geometries to be attractive and equal to ε_h. The loop has a uniform attractive interaction of magnitude $-\varepsilon_g$ with the residues at the start of the turn between strands 3 and 4 (residues 33 and 34) and is repulsive with all the hydrophilic residues in strands 1, 2, 5, and 6. The results reported below are for a few representative choices of parameters; qualitatively identical behavior is observed for a broad range of parameters. More specifically, ε_c was set equal to zero, as well as $-\varepsilon_g/2$; ε_h ranged over values from $-\varepsilon_g/8$ to $-\varepsilon_g/2$.

To surmount the multiple minima problem inherent in protein folding, a highly efficient dynamic Monte Carlo (MC) technique with an asymmetric Metropolis scheme is used (30, 31). For details, we refer to previous work (18). For the parameters described below, at least three independent cooling and heating sequences composed of a minimum of 2 × 10^6 MC cycles per temperature were run. In the transition region, the temperature gradient between runs is less and the number of MC cycles is increased up to 3 × 10^7 cycles. Great care has been taken to generate and sample highly equilibrated systems. This allows us to accurately describe the equilibrium properties of the native and denatured states. However, at a given temperature in the transition region, due to limitations of computer time, we have only been able to observe a small number (on the order of 10) of jumps between the native and denatured state. Thus, the equilibrium fraction of native and denatured states in the transition region is not very well characterized. Better computation of the equilibrium constant requires alternative sampling techniques whose development is under way.

RESULTS

The simplest amino acid sequence pattern that gives the structurally unique Greek key 2 of Fig. 1 and whose transition is thermodynamically all-or-none, is model A: $B_1(11)b_1^0$-$B_2(11)b_2^0B_3(11)b_3^0B_4(12)b_4^0B_5(11)L(7)B_6(11)$. For all hydrophilic residues in B_i, $\varepsilon_w = \varepsilon_g$; for all hydrophobic residues, ε_h $= -\varepsilon_g/4$ and for all residues in B_i, $\varepsilon_c = -\varepsilon_g/2$. The superscript zero on b indicates that $\varepsilon_g = 0$ for residues 10–12, 21–23, 32–34, and 44–46. Hence, the native turn configuration $g^+g^-g^+$ associated with b_1^0 to b_4^0 based on short range interactions is but one of 27 possible weighted configurations. For the putative loop residues numbered 57–63, the local energetic preferences are $-2\varepsilon_g$, $2\varepsilon_g$, $-2\varepsilon_g$, $-2\varepsilon_g$, $2\varepsilon_g$, $2\varepsilon_g$, and $-2\varepsilon_g$, respectively. The $-2\varepsilon_g$ ($+2\varepsilon_g$) indicates that the g^+ or $g^-(t)$ state is favored. Without long-range interactions, the native loop conformation $g^-tg^-g^-ttg^+$ is one of 16 equally weighted conformations of the *gauche* states. Thus, based on intrinsic stability, a native-like conformation of the loop is not enforced. Finally, we point out that because this model has a spherical potential, both the native conformation and its mirror image are isoenergetic and both have been obtained.

The curve denoted by the solid diamonds in Fig. 2 presents the mean square radius of gyration, $\langle S^2 \rangle$, vs. T^* obtained as an average over four cooling runs. The fraction of *trans* states, f_t, increases from ≈0.55 in the denatured state to 0.77 in the native conformation. The f_t of the pure native Greek key 2 of Fig. 1 equals 0.78. Tertiary interactions induce both additional secondary structure and the location of the tight bends and the long loop. The desired native conformation is the only collapsed state obtained on cooling and has been obtained from the denatured state over 30 times.

In Fig. 3, the average number of native contacts, N_c, as a function of time is plotted at T^* = 1.333, 1.220, and 1.111, respectively. (The pure native state has 74 contacts.) Each

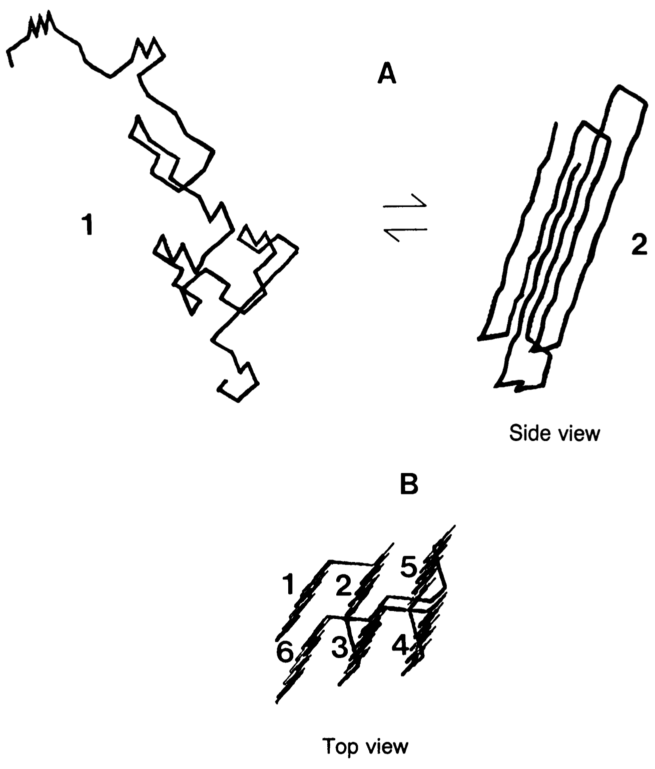

FIG. 1. (*A*) Representative configuration of a random coil state **1**, in equilibrium with the six-stranded β-barrel Greek key **2**, for which a side view is shown. (*B*) Top view of the native Greek key **2** that shows the numbering convention for the β-strands used in the text.

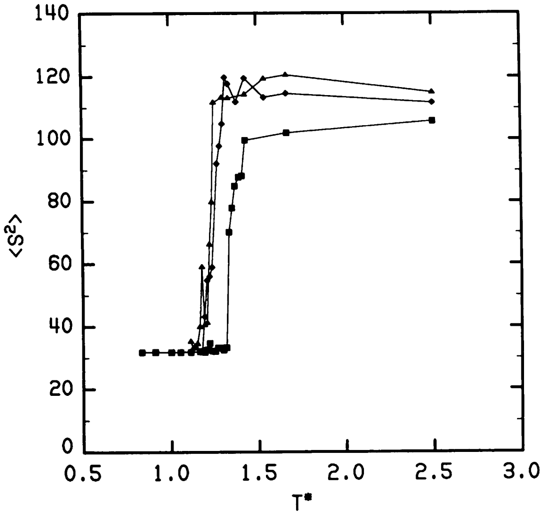

FIG. 2. Plot of the mean square radius of gyration, $\langle S^2\rangle$, vs. T^* for models A, B, and C in the curves denoted by ◆, ■, and ▲, respectively.

time unit represents 20,000 MC cycles. At $T^* = 1.333$, the system is under strongly denaturing conditions. As is apparent from Fig. 3*B*, at $T^* = 1.220$, where the system is in the transition region, the conformational transition is all-or-none, with the system spending <2.8% of its time in intermediate conformations. The oscillations in N_c between 55 and 74 involve fluctuations of the ends of strands 1 and 6 in the native state, an entirely expected result (26, 27). On further cooling, as in Fig. 3*C* to $T^* = 1.111$, the system is under strongly renaturing conditions. Thus, it appears that model A is a rather good schematic model of a globular protein, as a major feature of the thermodynamics of the folding transition has been reproduced. Such folding intermediates, which do occur, involve a four-member β-barrel involving strands 2–5, perhaps with β-strand 1 or β-strand 6. These intermediates can be made more populous by decreasing the strength of the attraction of the loop for the bottom of the native structure. This transforms the two-state model into a three-state model, where strands 1–5 form an equilibrium folding intermediate. Here too, the low-temperature conformation is the unique Greek key 2 of Fig. 1.

We next investigate some features that might enhance the range of thermal stability. One simple way is to augment the statistical preference for the native bend formation. The plausibility of this is supported by the elegant NMR studies of Wright *et al.* (32, 33), who find substantial populations of native-like β-turns in peptide fragments in the denatured state. We emphasize that such a local preference for turn formation is not required by the model.

Model B has the primary sequence $B_1(11)b_1B_2(11)b_2B_3(11)b_3B_4(12)b_4B_5(11)L(7)B_6(11)$; the turn neutral regions of model A are replaced by residues having an energetic preference for any *gauche* state equal to $-2\varepsilon_g$. For the bends b_i, $i = 1, \ldots 4$ based on their local intrinsic stability, any one of the 8 triplets of *gauche* states is equally likely. The native $g^+g^-g^+$ state should be stabilized from the tertiary interactions. In Fig. 2, for model B, $\langle S^2\rangle$ vs. T^* obtained as an average over five cooling sequences is given in the curve denoted by the solid squares. As expected, because of the lower free energy of the native Greek key as compared to model A, the transition has shifted to substantially higher temperature. f_t increases from ≈0.47 in the denatured state to 0.76 in the native conformation. Because of the enhanced preference for the *gauche* state, conformations in the denatured state, $\langle S^2\rangle$ in model B is less than in model A.

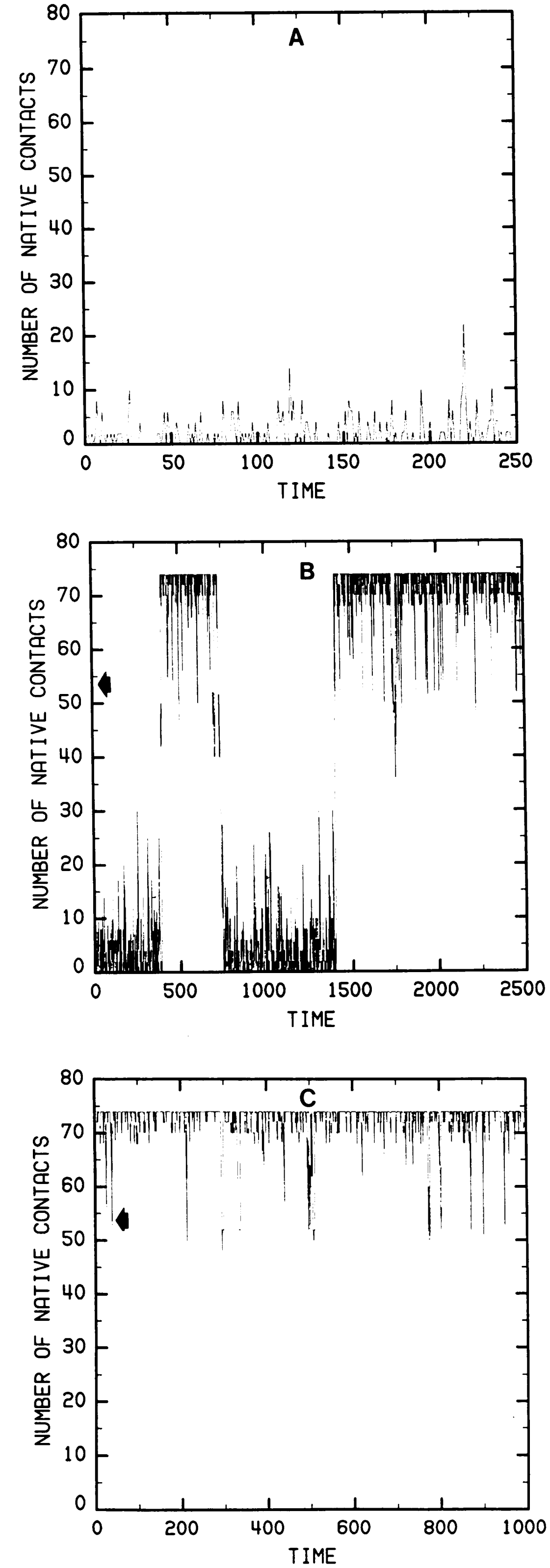

FIG. 3. (*A*–*C*) Plot of the number of native state contacts, N_c, vs. time for model A at $T^* = 1.333$, 1.220, and 1.111, respectively. Each time unit corresponds to 20,000 MC cycles. All values of N_c above the arrow reflect minor oscillations of β-strands 1 and/or 6 about the native conformation.

A similar analysis as presented in Fig. 3 reveals that this model too has an all-or-none transition. Basically, in models A and B the only intermediate states of any significance involve the four-member β-barrel of strands 2–5, sometimes with strand 1 or 6. By increasing the stability of the native conformation (now the free energy of the native turn is lower), this further diminishes the importance of folding intermediates in comparison to model A.

An interesting but probably somewhat artificial model is provided by model C, having the primary sequence pattern: $B_1(11)b_1^0B_2(11)b_2^0B_3(11)b_3^0B_4(12)b_4^0B_5(11)L^0(7)B_6(11)$. Model C is identical to model A in that all the b_i are turn neutral, but the loop region of model A has been replaced by a region where $\varepsilon_g = 0$ for residues 57–63. Hence, the native loop configuration $g^-tg^-g^-ttg^+$ is one of 2187 equally weighted configurations in the absence of long-range interactions. Once again, the low-temperature state is the unique Greek key **2** of Fig. 1. However, in the transition region, the four-member β-barrel intermediate is substantially populated, as are out of register conformations of strand 6 in the native-like states. This transition is not all-or-none. Model C points out the crucial role of the loop in determining the uniqueness and stability of the native state. In Fig. 2, for model C, the curve containing the solid triangles presents a plot of the $\langle S^2\rangle$ vs. T^*. The fraction of *trans* states, f_t, increases from ≈0.58 prior to collapse to 0.73 for the native-like states.

DISCUSSION

For the first time in a computer simulation, the folding to the complicated structural motif of a β-barrel Greek key has been obtained. These systems start out in the completely unfolded state and must hunt through all of configuration space to find the native structure that they prefer. An odd/even pattern of hydrophobic/hydrophilic residues, a loop that has a local energetic bias for a small subset of states, which includes the native conformation and the presence of residues that are intrinsically indifferent to tight turn formation, is sufficient to fold to a unique six-stranded Greek key **2** of Fig. 1. The origin of the uniqueness is as follows: Unlike turn neutral residues, if the hydrophobic residues occur in a bend this costs free energy—i.e., tertiary interactions induce bend and loop formation. Juxtaposition of this effect with the free energy cost of having nonbonded hydrophobic and hydrophilic residues as nearest neighbors produces only in register unique native conformations. Furthermore, the conformational transition is well approximated by a two-state model. If the long loop weakly interacts with the remainder of the assembled protein, folding intermediates comprised of the four-member β-barrel involving strands 2–5, perhaps with strand 1 or 6 also in the native conformation are observed and a two-domain model protein results. The transition from the random coil to the four-stranded β-barrel is all-or-none, as is the transition from the β-barrel intermediate to the Greek key. We therefore conclude that loops can play a very important role in determining the thermodynamics of the conformational transition, and their native state secondary structure, while irregular, is nevertheless very important (34–36). If the loop is made intrinsically indifferent to the native state conformation, while the Greek key structure **2** of Fig. 1 is reproduced at low temperatures, a whole spectrum of equilibrium folding intermediates emerge in the transition region. If there are regions having substantial variations in tertiary interactions (for example, loops that have to overcome a large entropic barrier for native state formation and that energetically marginally favor the native state), then multiple-domain proteins can emerge. Thus, the unique native conformation results from the interplay of short-range interactions, which impart a marginal stability to a given element of secondary structure, and tertiary interactions, which stabilize the secondary structure found in the native state.

The complicated six-stranded Greek key native state is obtained without invoking site-specific interactions—i.e., a very detailed model is not required. An identical conclusion was arrived at in previous folding simulations of four-stranded β-barrels (18) and four helix bundles having tight bends (ref. 37; unpublished data). However, unlike these cases, the Greek key topology reproduced here has a reversal of strand direction (19). Thus, the conjecture that the folding rules are extremely robust and, more specifically, that the gross native state topology depends on the hydrophobic/hydrophilic amino acid sequence pattern plus the presence of regions that, based on tertiary interactions, have a statistical preference to form loops or bends is again borne out, this time in the β-barrel Greek key topology. Hence, these models of protein folding, while schematic, seem to embody the necessary physics of globular protein folding. They reproduce both the thermodynamics and topology of real globular proteins. Of course, refinements in these models allowing the treatment of local details are required. We believe that these details of the free energy surface will produce small scale modifications of the tertiary structure—e.g., they might modify side-chain packing. Perhaps the most restrictive feature of the diamond lattice models is that α-helices and β-sheets cannot be parallel to each other. To accommodate the motif of α/β proteins, a more flexible lattice is required. Thus, the development of a more general lattice model is now under way.

This research was supported in part by National Institutes of Health Grant GM37408 from the Division of General Medical Sciences, U.S. Public Health Service.

1. Anfinsen, C. B. (1973) *Science* **181**, 223–230.
2. Jaenicke, R., ed. (1980) *Protein Folding*, Proceedings of the 28th Conference of the German Society (Elsevier/North Holland, Amsterdam).
3. Ptitsyn, O. B. & Finkelstein, V. A. (1980) *Q. Rev. Biophys.* **13**, 339–386.
4. Go, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.
5. Wetlaufer, D., ed. (1984) *The Protein Folding Problem* (Westview, Boulder, CO).
6. Creighton, T. E. (1985) *J. Phys. Chem.* **89**, 2452–2459.
7. Lewis, P. N., Go, N., Go, M., Kotelchuck, D. & Scheraga, H. A. (1970) *Proc. Natl. Acad. Sci. USA* **67**, 810–815.
8. Fasman, G. (1987) *Biopolymers* **26**, S59–S79.
9. Rose, G. D., Winters, R. H. & Wetlaufer, D. B. (1976) *FEBS Lett.* **63**, 10–16.
10. McCammon, J. A. (1984) *Rep. Prog. Phys.* **47**, 1–46.
11. Kollman, P. & van Gunsteren, W. F. (1987) *Methods Enzymol.* **154**, 430–449.
12. Karplus, M. & McCammon, J. A. (1981) *CRC Crit. Rev. Biochem.* **9**, 293–315.
13. McCammon, J. A. & Harvey, S. C. (1987) *Dynamics of Proteins and Nucleic Acids* (Cambridge Univ. Press, Cambridge, U.K.).
14. McCammon, J. A. (1987) *Science* **238**, 486–491.
15. Vásquez, M. & Scheraga, H. A. (1985) *Biopolymers* **24**, 1437–1447.
16. Li, Z. & Scheraga, H. A. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 6611–6615.
17. Kolinski, A., Skolnick, J. & Yaris, R. (1987) *Biopolymers* **26**, 937–962.
18. Skolnick, J., Kolinski, A. & Yaris, R. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 5057–5061.
19. Richardson, J. S. (1981) *Adv. Protein Chem.* **34**, 167–339.
20. Schellman, J. A. (1952) *CR Trav. Lab. Carlsberg Ser. Chem.* **29**, 230–259.
21. Tanford, C. (1962) *J. Am. Chem. Soc.* **84**, 4240–4247.
22. Brandts, J. & Lumry, R. (1963) *J. Phys. Chem.* **67**, 1484–1494.
23. Wetlaufer, D. B., Malik, S. K., Stoller, L. & Coffin, R. L. (1964) *J. Am. Chem. Soc.* **86**, 508–514.
24. Tanford, C. (1968) *Adv. Protein Chem.* **23**, 121–282.
25. Privalov, P. L. (1979) *Adv. Protein. Chem.* **33**, 167–241.
26. Ansari, A., Berendzen, J., Bowne, S. F., Frauenfelder, H., Iben, I. E. T., Sauke, T. B., Shyamsunder, E. & Young, R. D. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 5000–5004.
27. Elber, R. & Karplus, M. (1987) *Science* **235**, 318–321.

28. Guss, J. M. & Freeman, H. C. (1983) *J. Mol. Biol.* **169,** 521–563.
29. Moore, J. M., Case, D. A., Chazin, W.-J., Gippert, G. P., Havel, T., Powls, R. & Wright, P. E. (1988) *Science* **240,** 314–317.
30. Kremer, K., Baumgartner, A. & Binder, K. (1981) *J. Phys. A* **15,** 2879–2883.
31. Binder, K., ed. (1986) *Monte Carlo Methods in Statistical Physics* (Springer, Berlin), pp. 1–45.
32. Wright, P. E., Dyson, H. J., Rance, M., Houghten, R. A. & Lerner, R. A. (1987) *Protides Biol. Fluids* **35,** 477–480.
33. Dyson, H. J., Rance, M., Houghten, R. A., Lerner, R. A. & Wright, P. E. (1988) *J. Mol. Biol.* **201,** 161–200.
34. Venkatachalam, C. H. (1968) *Biopolymers* **6,** 1425–1436.
35. Chou, P. Y. & Fasman, G. D. (1977) *J. Mol. Biol.* **115,** 135–175.
36. Rose, G. D., Gierasch, L. M. & Smith, J. A. (1985) *Adv. Protein Chem.* **37,** 1–109.
37. Abdel-Meguid, S. S., Shieh, H. S., Smith, W. W., Dayringer, H. E., Violand, B. N. & Bentle, L. A. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6434–6437.