Andrzej Kolinski[1,3]
Dominik Gront[1]
Piotr Pokarowski[2]
Jeffrey Skolnick[3]
[1] *Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland*

[2] *Institute of Applied Mathematics and Mechanics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland*

[3] *Donald Danforth Plant Science Center, Bioinformatics and Computational Genomics, 893 N. Warson Rd., Saint Louis, MO 63141, USA*

A Simple Lattice Model that Exhibits a Protein-Like Cooperative All-or-None Folding Transition

Received 23 September 2002; accepted 22 January 2003

Abstract: *In a recent paper (D. Gront et al.,* Journal of Chemical Physics, *Vol. 115, pp. 1569, 2001) we applied a simple combination of the Replica Exchange Monte Carlo and the Histogram methods in the computational studies of a simplified protein lattice model containing hydrophobic and polar units and sequence-dependent local stiffness. A well-defined, relatively complex Greek-key topology, ground (native) conformations was found; however, the cooperativity of the folding transition was very low. Here we describe a modified minimal model of the same Greek-key motif for which the folding transition is very cooperative and has all the features of the "all-or-none" transition typical of real globular proteins. It is demonstrated that the all-or-none transition arises from the interplay between local stiffness and properly defined tertiary interactions. The tertiary interactions are directional, mimicking the packing preferences seen in proteins. The model properties are compared with other minimal protein-like models, and we argue that the model presented here captures essential physics of protein folding (structurally well-defined protein-like native conformation and cooperative all-or-none folding transition).* © 2003 Wiley Periodicals, Inc. Biopolymers 69: 399–405, 2003

Keywords: *lattice model; replica Monte Carlo sampling; histogram method*

INTRODUCTION

Simplified protein models are frequently used in studies of protein dynamics and thermodynamics.[1,2] According to the thermodynamic hypothesis of Anfinsen,[3] the native conformation of a globular protein corresponds to the lowest free energy state. It is also known that the protein folding transition is very co-

Correspondence to: Andrzej Kolinski; email: kolinski@chem.uw.edu.pl

Contract grant: Division of General Medical Sciences of the National Institutes of Health (NIH) and Polish Research Council, KBN
Contract grant number: GM-37408 (NIH) and 7-T11F-016-21 (KBN)

Biopolymers, Vol. 69, 399–405 (2003)

operative and has some features of a first-order phase transition.[4,5] When this phenomenon occurs in a finite (or small) system, such a transition is frequently described as an "all-or-none" transition to distinguish it from "true" first-order phase transitions.[5] In the context of simplified protein models, it is possible to study the qualitative factors responsible for this unique behavior of proteins. Reduced models could also be employed in low-resolution structure prediction. Even at low resolution, three-dimensional protein models could be very useful in determining the biological function of newly sequenced proteins.[6] This is a very important goal due to the enormous number of protein sequences of unknown structure and function provided by various genome projects, including the Human Genome Project.[6]

The conformational space of protein-like chain molecules can be explored by a variety of techniques ranging from the enumeration of all compact states in the very simplest of models,[1] through molecular dynamics,[7] genetic algorithms,[8,9] a combination of genetic algorithms with Monte Carlo methods,[10] and numerous Monte Carlo techniques and strategies.[11–23] Here, we employ a combination of the Replica Exchange Monte Carlo (REMC) sampling technique with the Histogram Method (HM) as a means of quickly and accurately estimating the system's thermodynamic properties.[23] For somewhat similar simple protein-like systems, recently we demonstrated[21] that the results are in excellent agreement with Multicanonical Monte Carlo (MMC) sampling[24,25] [or Entropy Sampling Monte Carlo (ESMC)[26]], but the combination of REMC with HM[23] is many times less computationally expensive than MMC. Here, we apply the REMC/HM technique to studies of the thermodynamics of model protein folding. A minimal model that exhibits the important features of an all-or-none collapse transition typical of globular proteins is designed and evaluated. The main purpose of this work is to find a minimal set of interactions that ensure a unique conformation of the collapsed state, fast folding and protein-like folding thermodynamics. In a recent work[23] we achieved the first two goals; however, the cooperativity of the folding transition was too low in comparison with real proteins, as was the free energy barrier between folded and unfolded states. Here, for a modified simple lattice model, we have found highly cooperative, all-or-none folding transition. To the best of our knowledge, this is the first time this has been achieved in a model system having a degenerate native conformation. Thus very important aspects of protein folding were reproduced in a very simple model that could be studied computationally in a great detail.

METHODS

Replica Exchange Monte Carlo

In REMC,[13,16,27] a number of copies (Replicas) of the model system placed at different temperatures are simulated by a classical Metropolis algorithm[28–29] (here, an asymmetric sampling scheme is used). The set of temperatures (with a constant temperature increment) covers the range from a random coil (or denatured) state to a compact folded globular state. Occasionally, replicas are exchanged with the probability:

$$P_{m,n} = \min[1, \exp(-\Delta_{m,n})] \quad |m-n| < 5 \tag{1}$$

$$\text{with } \Delta_{m,n} = (1/k_B T_n - 1/k_B T_m)[\mathrm{En}(\mathbf{X}_i) - \mathrm{Em}(\mathbf{X}_j)] \tag{2}$$

where replica $\mathbf{X}_j$ corresponds to temperature T_m. Due to the exponentially decreasing probability of swapping for the remote (in the temperature space) replicas, the attempts were limited to the nearby temperatures $|m - n| < 5$.

As previously demonstrated, this process is very effective in finding the lowest energy conformations.[15,21,30] At low temperatures, the simulations tend to be trapped in local minima on the model energy landscape. The exchange process moves these conformations to a higher temperature where it is much easier to surmount the intervening energy barriers. In addition, another replica has the opportunity to sample a different region of conformational space at lower temperatures.

REMC sampling could be combined with the Histogram Method.[31–35] Histograms of the energy distribution can be accumulated during the simulations. Assuming that at each temperature the distribution of states satisfies the Boltzmann distribution, the density of states could be expressed as follows:

$$P(E) = Z(T)^{-1} \cdot w(E) \cdot \exp(-E/k_B T) \tag{3}$$

where $Z(T)$ is the partition function at temperature T and $w(E)$ is the number of states with conformational energy E.

Consequently, the relative entropy $S(E)$ (with an accuracy to a constant C) could be written as follows:

$$S(E) = k_B \log[w(E)] = k_B \log[P(E)] + C(T) + E/T \tag{4}$$

Overlapping the curves at the various temperatures enables the elimination of $C(T)$.

The idea of combining numerous histograms[29,34,35] is not new and was proposed over ten years ago by Ferrenberg and Swendsen.[31,32] Multiple histograms were also applied in the study of a simple protein model by Ferguson and Garrett.[34,35] They collected histograms at various temperatures in a simulated annealing procedure. What is new in the present approach is the generation of a relatively large number of histograms by the REMC method in a single simulation and a very simple method of merging the multiple histogram data into a single entropy curve. Since the

neighboring temperatures are close to each other, the histograms overlap significantly. The details of this approach have been recently described.[23]

The estimation of the system's entropy enables the calculation of free energy as a function of the conformational energy and entropy for various energetic states, at a temperature T by

$$F(E, T) = E - T \cdot S(E) \quad (5)$$

Consequently, all properties of the model could be calculated as a function of temperature.

Definition of the Model

The test calculations were done on a simple protein lattice model where the polypeptide chain is restricted to a face-centered cubic lattice (fcc). The model studied here has a somewhat different representation from the simple model studied previously.[21,23] In the present model, the extended conformations must be of a zig-zag type, while the previous model also allowed 180° planar angles. This modification allows us to define the direction of the hypothetical side groups that point in the opposite direction of the planar angle bisector. The interaction scheme is also modified to mimic a preference to a "parallel" orientation of the virtual side chains. A lattice vertex corresponds to a single amino acid residue in a real polypeptide. Bonds connecting these united atoms can assume 12 orientations of the fcc lattice vectors. Two types of potentials mimic the main types of molecular interactions in proteins. First, there are short-range interactions that simulate the local conformational stiffness of the chain. We limited ourselves to the case of β-type proteins[36] (in the forthcoming work, we will extend this approach to other classes of globular proteins). Thus, there is an energetic preference towards locally expanded chain conformations. Namely, the model chain energy decreases by an increment ε_β in all cases where, for three subsequent chain bonds (connecting four subsequent lattice vertices), the three following criteria for an expanded state are satisfied:

1. The two corresponding planar angles are equal to 120°.
2. The first and the third consecutive vectors are identical.
3. The third bead of the fragment has assigned an E-state (expanded) preference.

Second, the long-range (between the nonbonded chain units) interactions between model residues have the form of a contact potential. There are two types of residues: nonpolar hydrophobic (H) and polar (P). The assumed contact potentials are as follows: ε_{HH}, ε_{PP}, and ε_{HP}. PP interactions were counted only for the residues contacting in a parallel fashion (the bisectors of the chain valence angles for the interacting residues must point in the exactly the same direction), reflecting the tendency of the parallel packing of polar side chains on the surface of a protein. Such a pattern is typical for all water-soluble single-domain globular proteins. Later we will show that this tendency for parallel packing of the side chains is essential for uniqueness of the folded structure of the model and that the cooperativity of the folding transitions is strictly related to this feature of the model. The numerical values of the potential are (in dimensionless k_BT units):

$$\varepsilon_{HH} = -4, \quad \varepsilon_{PP} = -4, \quad \varepsilon_{HP} = 4, \quad \varepsilon\beta = -7 \quad (6)$$

The "sequence" of the model protein is $[(PH)_4P_3(PH)_4P]_3$. On the level of secondary structure (chain stiffness), it reads as $(C_2E_7C)_6$, where E stands for the extended, β-type short-range preferences (local stiffness), and C denotes the flexible coil, or loop, regions. Loops contain polar (P) residues, while the strands have an alternating regular HP pattern of polar and hydrophobic residues. Such design of sequence is consistent with the assumed target structure of the folded conformation of the model chain. In the putative β-strands the hydrophobic residues points toward the interior of the globule an alternating polar residues points outside, into (implicit) surrounding solvent. Flexible loops have different length. The shorter ones are within the sheets and the longer ones cross over from one sheet to another. This is again consistent with the assumed structure of the two-sheet, Greek-key structure of the model β-barrel. Such a structure is typical for a large number of β-type globular proteins.[36]

Conformational updating was performed by a random sequence of local micromodifications (two-bond moves) of the chain geometry.

RESULTS AND DISCUSSION

Due to the large number of significantly overlapping histograms (we used 15 replicas) from REMC, and to obtain as accurate an entropy estimation as possible, we ignored the wings of distributions for which the observed density of states was below 1/80 of the maximum density. We gave a detailed description of the entropy calculations in our previous work.[23] Here, for the reader's convenience, we provide just a brief summary. We employed 15 replicas with the following set of temperatures: 10^7, 100, 10, 4.0, 3.5, 3.4, 3.3, 3.2, 3.1, 3.05, 3.0, 2.95, 2.9, 2.8, and 2.7. This set covers the high temperature range (where the system very effectively samples conformational space) and a couple of temperatures below the transition point, estimated here as $T_f = 2.93$. The program performed about 0.13 billion iterations (near one million per single replica). A single iteration consisted of N (where N is the chain length), randomly distributed along the chain, attempts to change the local conformation. Single bead changes were employed. Swaps of the replicas were attempts every 100 iterations of the algorithm. As demonstrated in our previous studies[23] of a very similar model, these parameters en-

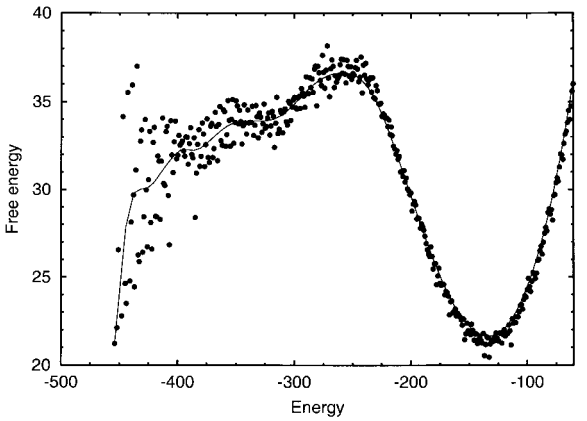

FIGURE 1 Free energy as a function of energy at a temperature $T = 2.93$. The high free energy barrier (about 17 k_BT) between the random coil region of the energy and the energy of the collapsed states indicates a very cooperative transition.

sured full convergence. All the degenerate lowest energy states were visited many times during the simulations. This provides an additional test for very good convergence.

Free energy curves can be used to detect phase transitions. For example, at $T_f = 2.93$, the free energy plot (see Figure 1) exhibits two minima corresponding to two states at equilibrium. It should be pointed out that the scatter of the data in the plot is the result of the lattice approximation, *not* the effect of statistical error. The dots correspond to all integer values of the system's energy. Our model has larger values of the interaction parameters. When the energy is averaged over a window of 4, the fluctuations decrease to the range of the symbol size. The solid line corresponds to such averaged data. The curves taken from shorter subruns are virtually indistinguishable. Another test of consistency of our calculations comes from comparison of the free energy plots generated from single histograms (instead combining all histograms, as it was done for the data presented in Figure 1). For instance, pairwise correlation coefficients between the free energy curves calculated from histograms at $T = 3.0$, $T = 2.95$, and $T = 2.9$, are equal to 0.9943 (the 3.0, 2.95 pair), 0.9850 (2.95, 2.9 pair), and 0.9852 (3.0, 2.9 pair), respectively.

In Figure 2, examples of conformations of the model chain are shown at various characteristic temperatures. Conformations corresponding to the right minimum in a free energy curve are placed in an upper row. In the middle row, typical conformations from the center of the free energy barrier are shown. The lowest row presents "native-like" conformations of lowest energy. These conformations correspond to the left-hand minimum of the free-energy curve.

The completely folded state has 20 conformers. All have the same energy (−454), the same topology of the Greek-key fold, and exactly the same pattern of long-range and short-range interactions. Particular conformations differ only by small geometrical details (due to the various orientations of particular strands on the lattice) or by a mirror image transformation. The set of native-like topological diagrams are given in Figure 3 and show five different structures when viewed from the top of the barrel. Each structure has an isomer in which one of the sheets is located lower (along the barrel axis) than the opposite sheet. All ten structures have mirror image conformations due to the lack of any chiral component in the present model. This degeneracy of the ground state of the model is probably quite physical. Indeed, real proteins have quite substantial conformational entropy in the native state. The degeneracy of the folded state may contribute to the cooperativity of the folding process, thereby increasing the free energy barrier between the folded state(s) and the manifold of unfolded conformations. A high free-energy barrier is located between the random coil conformations and fully folded structures (as shown in the third row of snapshots in Figure 2), taken for structures with E = −454 The free-energy barrier is large, about 17 in k_BT units, demonstrating a pseudo first-order transition with negligible equilibrium population of folding intermediates. Analysis of the heat capacity also reveals just one very sharp peak at $T = 2.93$. Intermediates (second row of snapshots in Figure 2) have some features of the molten globule state. They are quite compact and exhibit high secondary structure content, mostly consistent with the secondary structure of the folded state; however, it is rapidly fluctuating and incomplete. Comparison with our previous work shows that while a weakly cooperative model results from a competition between short-range conformational stiffness and long-range interactions, for well-defined cooperativity and a two-state transition, orientation-dependent interactions may be necessary. In real proteins, main chain hydrogen bonds play such a role. In the present model, the orientation-dependent interactions of polar groups account, to some extent, for such structural ordering. Additionally, this orientation dependence provides a means for the separation of the hydrophobic core from the polar surface in the model globule. Similarly, in real proteins, polar side groups are composed of hydrophobic and polar fragments. This leads to a parallel packing of the side chains on the surface of water-soluble proteins.

SUMMARY

The proposed minimal protein model exhibits a number of essential features of real proteins:

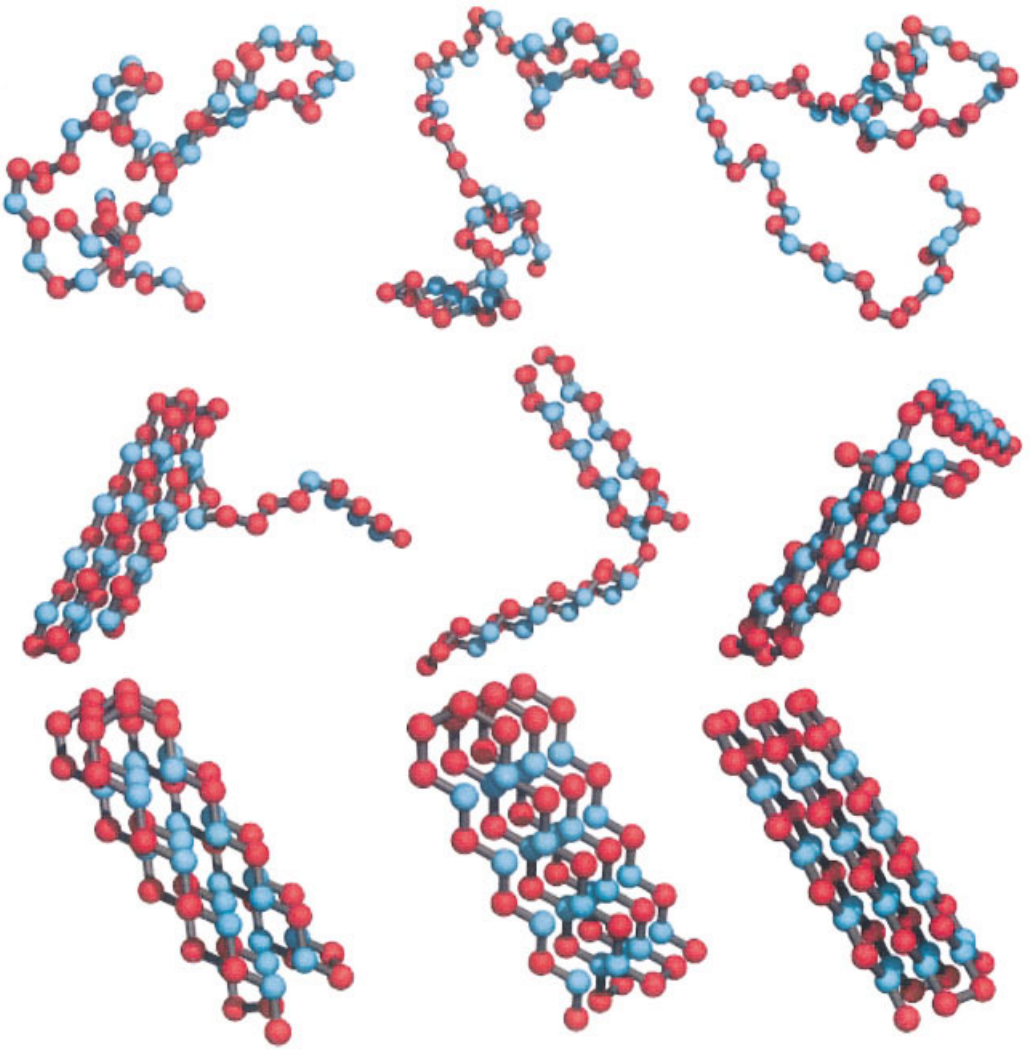

FIGURE 2 Representative chain conformation snapshots. The upper row shows examples of the random coil state (all have the same conformational energy of $-140\ k_BT$) The second row shows partly folded, collapsed structures with a conformational energy of $-365\ k_BT$. The bottom row shows three selected completely folded structures for which $E = -454\ k_BT$.

1. There is a characteristic interplay between short- and long-range interactions; protein chains have limited local flexibility. A position-dependent flexibility and three types of long-range interactions mimic (in a minimalist approach) this effect.
2. Very important aspects of protein physics may be encoded in orientation-dependent interactions of the polar groups. Such a potential mimics the combined effect of the tendency of the polar side chains to pack in a parallel fashion on the protein's surface, as well as the orientation effect of the main chain hydrogen bonds.
3. The model protein has a well-defined hydrophobic core with a well-defined ground state (ten very similar conformations with the same energy, when the mirror image folds are neglected). All have the relatively complex Greek-key topology, typical of a large fraction of β-type globular proteins.
4. Interestingly, a slightly more flexible model ($\varepsilon_\beta = -6$) folds into the same topology, however, with numerous secondary structure "defects."
5. The folding transition is very cooperative and exhibits features of first-order ("all-or-none") phase transitions with a relatively high free-energy barrier (see Figure 1) between the folded state and the multiplicity of random-coil conformations.

The features (a well-defined secondary structure, a well-defined complex topology of the folded state,

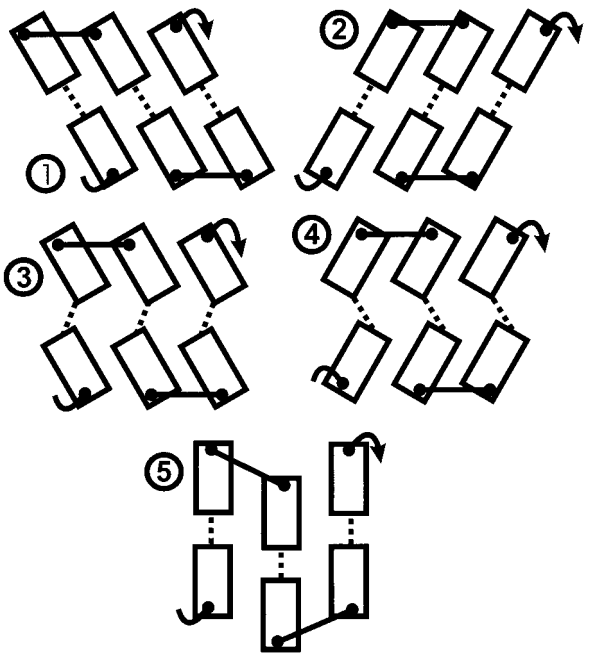

FIGURE 3 A schematic illustration of the five distinct geometrical realizations of the six-stranded Greek-key fold for the fcc-lattice model protein. All structures have the same pattern of all types of interactions. The orientation of the rectangular boxes reflects the orientation of the β-strands. Solid lines indicate loops above the barrel while dashed ones correspond to loops below the barrel.

and protein-like folding thermodynamics) of the model studied here are not present in other minimal protein models (see an excellent review on simple-exact models of proteins by Dill et al.[1]). For instance, in the HP-type, simple cubic lattice chains studied in many works,[1,2] secondary structure preferences are ignored and protein-like thermodynamics is very difficult to achieve for simple sequences. Some 15 years ago, we demonstrated via Monte Carlo simulations that a finite length semiflexible homopolymer model undergoes a pseudo first-order collapse transition.[2,37,38] Thus the competition between short-range and long-range interactions seems be an important factor of cooperatively folding polymers. Indeed, formation of relatively rigid elements of secondary structure partitions the conformational space and is very likely the main means of beating the Levintal's paradox.[39] Yet, in the design of the majority of the simple HP-type models, this is entirely ignored. The HP sequence of the present model is very simple, yet thanks to the orientation-dependent interaction of the polar groups and due to sequence-dependent, local chain stiffness the model adopts a quite complex topology. The folding transition (in contrast to most of the other minimal models) is highly cooperative, with a large free-energy barrier between the folded and unfolded states. At equilibrium, the population of folding intermediates is extremely low, as it is in the real systems. Some contribution to the high free-energy barrier comes probably from entropic stabilization of the folded state. Indeed, the Greek-key fold of the present model has 20 variants that have exactly the same interaction pattern and consequently the same conformational energy. This nicely simulates the small conformational fluctuations present in real native structures. Thus, it appears that this very simple model captures many basic properties of globular proteins.[40,41] In forthcoming work, we will show that the design of the present model is indeed a minimal one, i.e., that all its features are necessary for protein-like behavior (unpublished).

The Replica Exchange Monte Carlo sampling technique, when combined with the Histogram method, provides a very convenient and computationally effective means of estimating the entropy of this protein model system, thereby providing all interesting thermodynamic quantities as a function of temperature. Alternatively, in very conservative computations, the proposed method can be used to build an initial entropy histogram as a function of energy for Multicanonical Ensemble, MC, or Entropy Sampling Monte Carlo, thereby significantly reducing the computational cost of these methods. Recently, we demonstrated that the REMC method could be efficiently employed in the search for the lowest energy conformations of much more complex models of proteins.[42] This has very important implications for protein structure and function prediction.[43]

This research was supported in part by grant no. GM-37408 of the Division of General Medical Sciences of the National Institutes of Health. Piotr Pokarowski was partially supported by Polish Research Council, KBN, grant no. 7-T11F-016-21.

REFERENCES

1. Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; and Chan, H. S. Protein Sci 1995, 4, 561–602.
2. Kolinski, A.; Skolnick, J. Lattice Models of Protein Folding, Dynamics and Thermodynamics; R. G. Landes: Austin, TX, 1996.
3. Anfinsen, C. B. Science 1973, 181, 223.
4. Creighton, T. E. Proteins: Structures and Molecular Properties; W. H. Freeman and Company: New York, 1993
5. Scheraga, H. A.; Hao, M.-H.; Kostrowicki, J. In Atassi, M. Z. Appela, E., eds.; Methods in Protein Structure Analysis; Plenum Press: New York, 1995.

6. Montelione, G. T.; Anderson, S. Nature Struct Biol 1999, 6, 11–12.
7. Brooks, C. L. I., Karplus, M.; Pettitt, B. M. Proteins: A Theoretical Perspective of Dynamics Structure and Thermodynamics; Wiley: New York, 1988.
8. Dandekar, T.; Argos, P. Protein Eng 1992, 5, 637–645.
9. Unger, R. and Moult, J. J Mol Biol 1993, 231, 75–81.
10. Sun, S. Protein Sci 1993, 2, 762–785.
11. Hao, M.-H.; Scheraga, H. A. J Phys Chem 1994, 98, 9882–9893.
12. Hao, M.-H.; Scheraga, H. A. J Chem Phys 1995, 102, 1334–1348.
13. Hukushima, K.; Nemoto, K. J Phys Soc Jpn 1996, 65, 1604–1608.
14. Mohanty, D.; Kolinski, A.; Skolnick, J. Biophys J 1999, 77, 54–69.
15. Hansmann, U. H. E.; Okamoto, Y. Curr Opin Struct Biol 1999, 9, 177–181.
16. Sugita, Y.; Okamoto, Y. Chem Phys Lett 1999, 314, 141–151.
17. Hansmann, U. H. E.; Okamoto, Y. J Comput Chem 1993, 14, 1333–1338.
18. Hansmann, U. H. E. Chem Phys Lett 1997, 281, 140–150.
19. Kolinski, A.; Galazka, W.; Skolnick, J. Proteins 1996, 26, 271–287.
20. Kolinski, A.; Galazka, W.; Skolnick, J. J Chem Phys 1998, 108, 2608–2617.
21. Gront, D.; Kolinski, A.; Skolnick, J. J Chem Phys 2000, 113, 5065–5071.
22. Scheraga, H. A.; Hao, M.-H. Adv Chem Phys 1999, 105, 243–272.
23. Gront, D.; Kolinski, A.; Skolnick, J. J Chem Phys 2001, 115, 1569–1574.
24. Berg, B. A.; Neuhaus, T. Phys Rev Lett 1991, 68, 9–12.
25. Lee, J. Phys Rev Lett 1993, 71, 211–214.
26. Hao, M.-H.; Scheraga, H. A. J Phys Chem 1994, 98, 4940–4948.
27. Swendsen, R. H.; Wang, J. S. Phys Rev Lett 1986, 57, 2607–2609.
28. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. J Chem Phys 1953, 51, 1087–1092.
29. Newman, M. E. J.; Barkema, G. T. Monte Carlo Methods in Statistical Physics; Clarendon Press: Oxford, 1999.
30. Hansmann, U. H. E.; Okamoto, Y. J Comput Chem 1997, 18, 920–933.
31. Ferrenberg, A. M.; Swendsen, R. H. Phys Rev Lett 1988, 61, 2635–2637.
32. Ferrenberg, A. M.; Swendsen, R. H. Phys Rev Lett 1989, 63, 1195–1198.
33. Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. J Comput Chem 1992, 13, 1011–1021.
34. Ferguson, D. M.; Garrett, D. G. Adv Chem Phys 1999, 105, 311–336.
35. Ferguson, D. M.; Marsh, A.; Metzger, T.; Garrett, D. G.; Kastella, K. J Global Opt 1993, 4, 209–227.
36. Branden, C.; Tooze, J. Introduction to Protein Structure; Garland Publishing: New York and London, 1991.
37. Kolinski, A.; Skolnick, J.; Yaris, R. J Chem Phys 1986, 85, 3585–3597.
38. Kolinski, A.; Skolnick, J.; Yaris, R. Biopolymers 1987, 26, 937.
39. Zwanzig, R.; Szabo, A.; Bagchi, B. Proc Natl Acad Sci 1982, 89, 20.
40. Thirumalai, D. J Phys I France 1995, 5, 1457–1467.
41. Thirumalai, D.; Woodson, S. A. Accts Chem Res 1996, 29, 433–439.
42. Kihara, D.; Zhang, Y.; Kolinski, A.; Skolnick, J. Proc Natl Acad Sci USA 2002, 99, 5993–5998.
43. Skolnick, J.; Kolinski, A. Adv Chem Phys 2002, 120, 131–192.