Computer Simulation of Protein Refolding Pathways and Intermediates

Prashant Gupta and Carol K. Hall
Dept. of Chemical Engineering, Box 7905, North Carolina State University, Raleigh, NC 27695

Computer simulation studies of refolding pathways and the formation of intermediates for a simple, 2-D lattice protein model are presented. The sequence of the 20-bead model protein chain is chosen so that hydrophobic beads will reside in the protein interior in the native state. Nonbonded hydrophobic beads attract each other with strength ϵ; decreasing the $|\epsilon/kT|$ mimics increasing the concentration of the denaturant. Dynamic Monte Carlo simulations and exhaustive conformational searches have been performed on an isolated model protein sequence at different levels of $|\epsilon|$ (different denaturant concentrations). As the denaturant is withdrawn, the model protein exhibits a transition from a random coil state to a compact native state with a hydrophobic core. The refolding process is observed to be cooperative in that the chain does not start folding until the middle section has folded correctly, and proceeds along preferred pathways that are populated by distinct, partially folded intermediates.

Introduction

Genetically engineered proteins are produced in bacterial cells using recombinant DNA technology. These recombinant proteins aggregate within the cells to form insoluble, biologically inactive particles termed "inclusion bodies." To recover biologically active protein, the inclusion bodies have to be dissolved in harsh denaturant solutions and then the resulting deaggregated protein has to be refolded via removal of denaturant. However, as the denaturant is removed, partially refolded intermediates of protein start aggregating, resulting in low yield of active protein (Gierasch and King, 1990; Mitraki and King, 1989). Understanding the mechanism of aggregation and refolding phenomena is important for improving recombinant protein yield.

The mechanism of aggregate formation is not completely understood; however, it is generally believed that the proteins fold along preferred pathways and that the aggregation occurs when certain partially folded intermediates populated along the refolding pathway participate in an off-pathway step and aggregate (Zettlemeissl et al., 1979; London et al., 1974; Goldberg and Zetina, 1980). Therefore, an understanding of folding intermediates, and folding pathways in general, could shed considerable light on the aggregation phenomena. The work presented here is the first step in a project that is aimed at performing computer simulation studies of the competition between protein refolding and aggregation in a solution consisting of a large number of protein molecules.

In this work, a simple lattice model was used to represent an isolated protein chain in two dimensions. This model captures the essential features of the interactions involved in refolding and aggregation phenomena, namely that in order for the hydrophobic beads to avoid solvent contact they tend to cluster together in a compact hydrophobic core in the native state, yet it is simple enough that the simulation of a multiple chain system should be computationally tractable.

The main purpose of this work is to explore the nature of refolding pathways and folding intermediates. The study focuses on a single sequence chosen for study because of its ability to form multiple, easily recognized, partially folded intermediates. The 20-bead sequence containing 8 hydrophobic beads and 12 polar beads is arranged in an order that ensures that the native state contains a hydrophobic core surrounded by polar beads. The exhaustive conformational search and dynamic Monte Carlo simulations described here show that this hypothetical protein chain exhibits a folding–unfolding transition and that the refolding process proceeds along well-defined folding pathways involving cer-

Correspondence concerning this article should be addressed to C. K. Hall.

tain distinct partially folded intermediates. Some of these intermediates appear capable of associating with one another and may initiate an aggregation reaction in a multiple chain simulation.

Model

The folding process is believed to be largely driven by the hydrophobic effect (Dill, 1990). The tendency of the hydrophobic residues to avoid contact with aqueous media causes them to come together at the core of the molecule. A protein folds largely to hide its hydrophobic residues from solvent contact. The compact hydrophobic core is surrounded by polar residues. In order to carry out computer simulation studies of the refolding and aggregation phenomena, a model representation of a protein molecule has to be chosen that captures this important feature of the folding process.

A two-dimensional square lattice representation of protein molecules proposed by Lau and Dill (1989) is employed. A protein molecule is represented as a continuous chain of n amino acids or groups of amino acids termed "beads" that can be either hydrophobic (H) or polar (P). Each bead occupies one lattice site, and to incorporate the excluded volume interaction, a lattice site can hold no more than one bead at a time. A chain conformation is represented as a walk on the lattice subject to the excluded volume criterion. "Connected neighbors" are defined as a pair of beads that are adjacent to each other along the protein chain, whereas "topological neighbors" are defined as a pair of beads that occupy adjacent lattice sites in a given conformation but are not connected neighbors. When two H beads are topological neighbors, they are assumed to interact with an attractive potential of mean force. The native state is defined as the conformation that has the maximum number of topological HH contacts. A given sequence may have more than one native state.

The strength of the attractive interaction energy between two H beads, ϵ, is a measure of the amount of denaturant in the solution; ϵ is negative because the HH interaction is attractive in nature. If the concentration of the denaturant in the solution is high, the hydrophobic interactions will be weak and hence the absolute value of ϵ will be low. On the other hand, if the concentration of the denaturant in the solution is low, then the hydrophobic interactions will be strong and the absolute value of ϵ will be high. Therefore, a protein chain can be simulated in solvents containing different amounts of denaturant by changing the $|\epsilon|$ parameter.

Simple lattice models of protein molecules (Šali et al., 1994) can provide us with valuable insights because they capture the important features of folding and aggregation processes, yet are simple enough that the native state can be determined exactly through exhaustive enumeration of conformations. Use of a two-dimensional lattice offers distinct computational advantages over a three-dimensional lattice model (Lau and Dill, 1989):

1. Computer time required to exhaustively enumerate all the possible conformations of a protein chain of length N on a lattice is roughly proportional to N^{z-1}, where z is the coordination number of the lattice ($z = 4$ on a square lattice and $z = 6$ on a cubic lattice). Thus, the computer time required to exhaustively enumerate all of the conformations of a protein chain with length N would be proportional to N^3 on a square lattice and N^5 on a simple cubic lattice.

2. A 20-bead protein chain on a 2-D square lattice can have up to 40% of its beads buried inside the core, away from the solvent contact. On the other hand, a 27-bead protein chain occupying a $3\times3\times3$ cube on a cubic lattice can have only one bead in the core. To bury 40% of the beads in the core on a 3-D cubic lattice would require a chain containing approximately 160 beads.

Clearly, a 2-D representation is far less computationally intensive than its 3-D analog.

Hypothetical sequence studied

Since the long-term objective of this study is to gain insight into aggregation phenomena resulting from association among kinetic intermediates, a model protein chain has been devised that has several distinct partially folded intermediates. Figure 1 shows the sequence of this model protein chain that comprises 8H beads and 12 P beads. Figure 2 shows that the native state of this model protein has eight topological HH

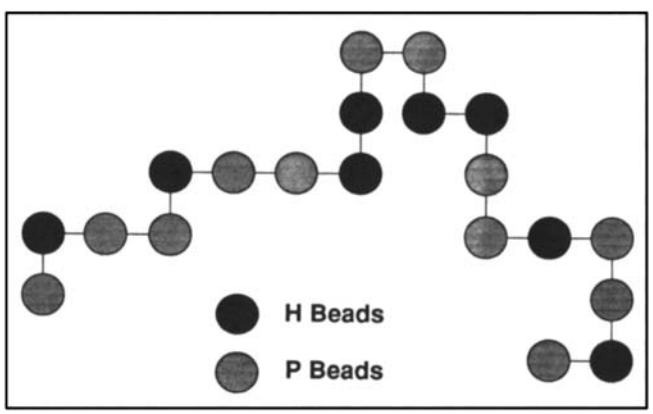

Figure 1. Hypothetical 20-bead protein sequence.
The chain has 12 polar (P) beads and 8 hydrophobic (H) beads.

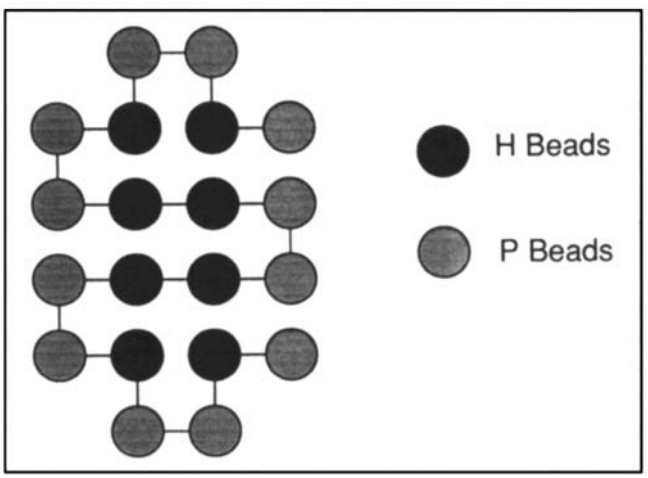

Figure 2. Native state of the 20-bead protein.
The H beads form a compact hydrophobic core that is surrounded by the P beads. There are 8 HH contacts in the native conformation.

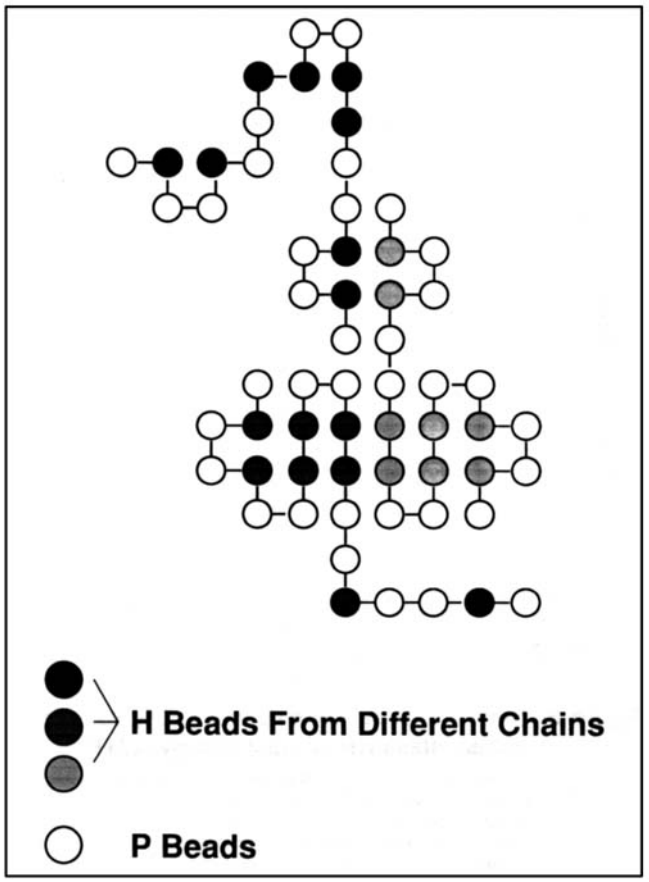

Figure 3. Hypothetical folding intermediates for the 20-bead protein that could cause aggregation in a multiprotein simulation if exposed H beads on one intermediate associate with exposed H beads on another intermediate.

contacts and three topological PP contacts arranged in a conformation with a hydrophobic core surrounded by polar beads. The sequence of this hypothetical protein suggests that it might contain the easily visualized intermediates shown in Figure 3. In a multichain simulation, these intermediates would associate with one another when the exposed hydrophobic beads in one chain encounter the exposed hydrophobic beads on another chain.

Exhaustive Enumerations

Employing a lattice representation of a molecule allows one to generate all possible conformations that the molecule can assume. This exhaustive enumeration information can be used to calculate the ensemble averages of various properties of interest such as the end-to-end distance, radius of gyration, and number of topological contacts. The average of any conformational property, $\langle A \rangle$, is given by

$$\langle A \rangle = \frac{\sum_{i=1}^{N_{\text{conf}}} A_i e^{-n_i(\epsilon/kT)}}{\sum_{i=1}^{N_{\text{conf}}} e^{-n_i(\epsilon/kT)}} = \frac{\sum_{i=1}^{N_{\text{conf}}} A_i e^{-n_i(\epsilon/kT)}}{Z} \tag{1}$$

where A_i is the value of the conformational property in the ith conformation, N_{conf} is the total number of possible conformations, n_i is the number of HH topological contacts in the ith conformation, and Z is the partition function,

$$Z = \sum_{i=1}^{N_{\text{conf}}} e^{-n_i\epsilon/kT}. \tag{2}$$

In the canonical ensemble, the specific heat, C, is given by

$$C = \left(\frac{\langle E^2 \rangle - \langle E \rangle^2}{kT^2} \right) \tag{3}$$

where E is the conformational energy. Since the conformational energy in the ith conformation, E_i, is equal to $n_i\epsilon$, we can rewrite this as

$$C = \frac{\epsilon^2}{kT^2}(\langle n^2 \rangle - \langle n \rangle^2), \tag{4}$$

which leads naturally to the definition of a dimensionless specific heat,

$$C^* = \frac{C}{k}\left(\frac{kT}{\epsilon}\right)^2 = \langle n^2 \rangle - \langle n \rangle^2. \tag{5}$$

Exhaustive enumerations of all possible chain conformations on a two-dimensional lattice have been carried out for the 20-bead model protein. At low values of the HH attraction energy, which corresponds to high denaturant concentration, the chain exists in a random coil state. As the HH attraction energy is increased, which corresponds to removing the denaturant, the chain undergoes a folding transition from an unfolded to a compact folded state. This process is illustrated in Figure 4, which shows the average number of topological HH contacts vs. the absolute value of the HH contact energy, $|\epsilon/kT|$. Figure 5 shows the dimensionless specific heat, C^*, vs. the $|\epsilon/kT|$ curve; the peak marks the transition point for the chain. The average radius of gyration, $\langle R_g^2 \rangle^{1/2}$, and the average end-to-end distance, $\langle R^2 \rangle^{1/2}$, are also plotted against $|\epsilon/kT|$ in Figure 6. At low values of the HH attraction energy, the chain exists in an unfolded state characterized by high values of $\langle R_g^2 \rangle^{1/2}$ and $\langle R^2 \rangle^{1/2}$. As the HH attraction is made stronger, the chain collapses into the native state.

Description of the Simulation Algorithm

Dynamic Monte Carlo simulations of the two-dimensional model protein chain can give us some insight into the folding pathway. The Monte Carlo algorithm proceeds by constructing a Markov chain of conformational states of the molecule so that in the limit of an infinitely long run, all the conformational states occur with probabilities determined by the Boltzmann factor (Metropolis et al., 1953). The Monte Carlo method of Metropolis et al. is designed to calculate the ensemble average for equilibrium systems. It does not simulate the time evolution of a system. However, if the *a priori* transition probabilities are chosen properly, the method can be used to approximately simulate the time behavior of a system (Taketomi et al., 1975).

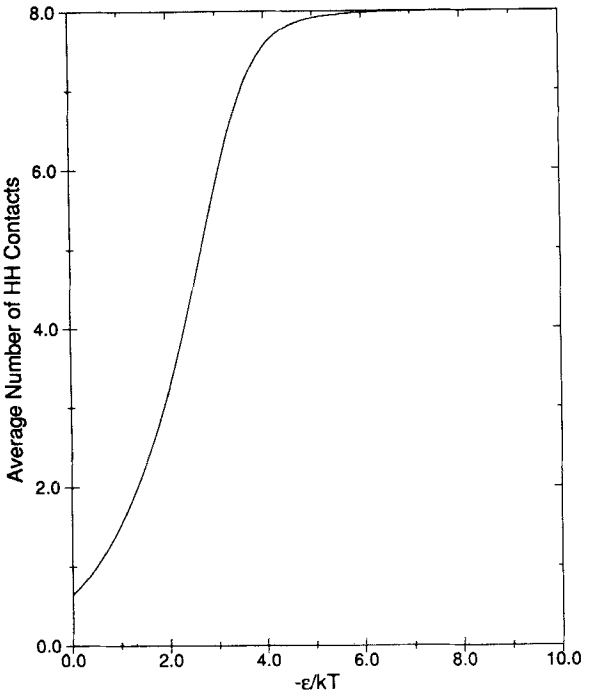

Figure 4. Average number of HH contacts vs. the dimensionless HH contact energy, ϵ/kT.

At low values of $|\epsilon/kT|$ (high denaturant concentration) the chain is in an unfolded state. As $|\epsilon/kT|$ is increased (denaturant withdrawn), the chain undergoes a transition to a native state that has 8 HH contacts.

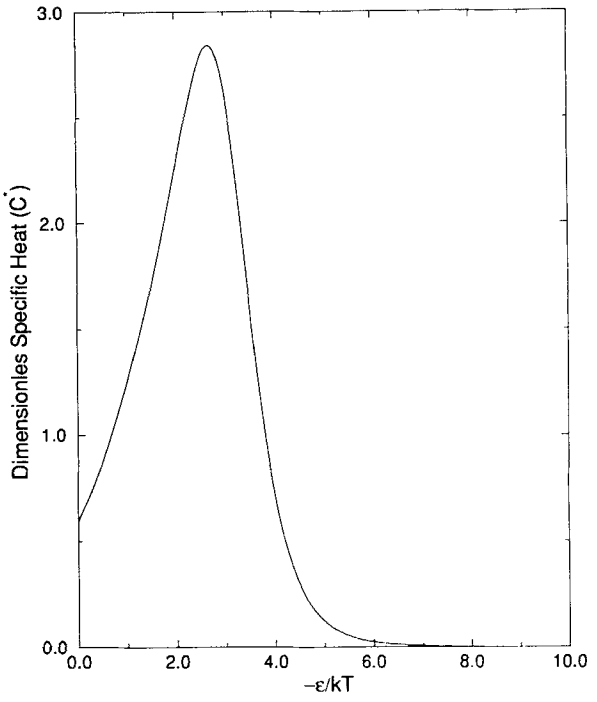

Figure 5. Dimensionless specific heat, C^*, vs. the dimensionless HH contact energy, ϵ/kT.

At low values of $|\epsilon/kT|$ (high denaturant concentration) the chain is in an unfolded state. As $|\epsilon/kT|$ is increased (denaturant withdrawn), the chain undergoes a transition to a native state. The maximum in the curve represents the transition point for the folding-unfolding transition.

The set of moves employed in the dynamic Monte Carlo simulation of protein folding reported here has been taken from Kolinski et al. (1986) and consists of chain-end rearrangements, three-bond moves, and four-bond moves. A bond, $\mathbf{b}_i$, is defined as a vector drawn from the center of bead i to the center of bead $i+1$. In a three-bond move, randomly selected bonds $\mathbf{b}_i$ and $\mathbf{b}_{i+2}$ are interchanged, provided $\mathbf{b}_i \neq \mathbf{b}_{i+2}$. In a four-bond move, two randomly selected bonds, $\mathbf{b}_i$ and $\mathbf{b}_{i+3}$ are reoriented randomly, provided $\mathbf{b}_i = -\mathbf{b}_{i+3}$. The four-bond move helps create a new local conformation in the middle of the chain by moving three beads at a time. The *a priori* transition probabilities used by Kolinski et al. (1986) have been employed. A few examples of the three- and four-bond moves are illustrated in Figure 7.

Determination of the Folding Pathway

The folding pathway for the 20-bead chain has been explored using dynamic Monte Carlo simulation in the canonical ensemble. A folding pathway is simply an ordered sequence of conformations that the chain assumes as it proceeds from a random coil conformation to the native conformation. To describe the folding pathways, we define a "conformation" to be completely specified by all of its topological HH contacts. This means that two conformations are indistinguishable if they contain identical topological HH contacts but different HP or PP contacts.

Simulations are carried out in the canonical ensemble at a specific value of ϵ/kT. A random initial conformation is generated at the start of the simulation. The chain is then subjected to the set of moves described earlier until the native state is reached. The sequence of conformations traversed from the initial state to the final state is recorded. In order to determine the pathway, the sequence of conformations is examined for the presence of "loops." If the chain exists in a given conformation at time step t, and the chain exists in the same conformation at a later time step $t+l$, then the sequence of conformations is said to contain a "loop" of length l. All of the l conformations making the loop are deleted from the conformation sequence. When all the loops are removed from the conformation sequence, what remains is the path taken by the protein from a random state to the native state.

One hundred simulation runs were carried out, starting from different random initial conformations, at an ϵ/kT value of -4.0. The pathways obtained exhibit certain common features. The ends of the chain are seen to rapidly fold and unfold throughout the folding process. However, the folded ends are not stable by themselves until the middle part of the chain folds, that is, when the beads labeled 9 and 12 in Figure 8 form an HH contact. Thus, the folding of the chain starts when the middle part of the chain folds correctly. After the middle part of the chain is folded, one of the ends folds in while the other one may keep fluctuating for some time. Eventually both the ends fold giving rise to the native state.

The resulting pathmap, which displays the important parts of the pathways, is shown in Figure 9. The middle contact is already formed in the three conformations, (a), (b), and (c)

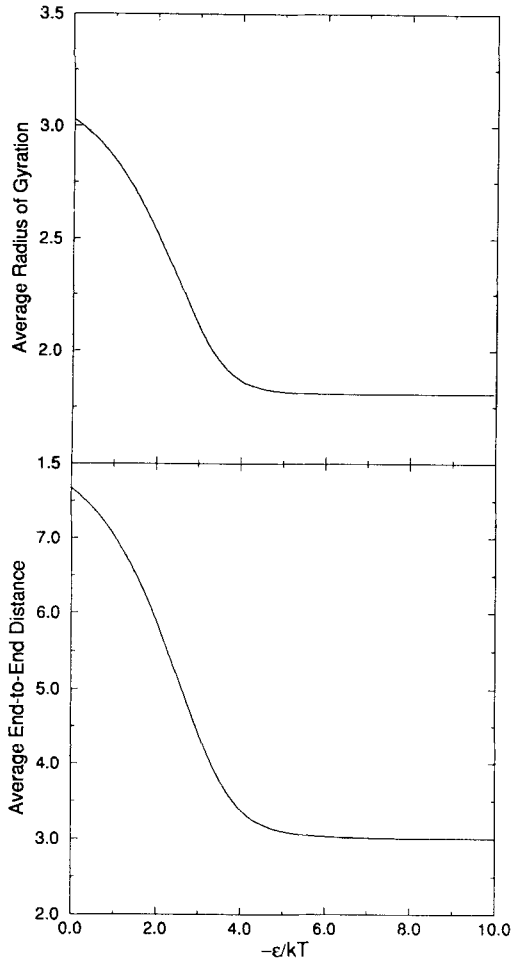

Figure 6. Average value of the radius of gyration ($\langle R_g \rangle^{1/2}$) and average value of the end-to-end distance ($\langle R^2 \rangle^{1/2}$) vs. the dimensionless HH contact energy, ϵ/kT.

At low values of $|\epsilon/kT|$, the chain exhibits its random coil $\langle R_g^2 \rangle^{1/2}$ and $\langle R^2 \rangle^{1/2}$ values. As the HH attractions are made stronger, the chain folds into a compact state, characterized by comparatively small values of $\langle R_g^2 \rangle^{1/2}$ and $\langle R^2 \rangle^{1/2}$.

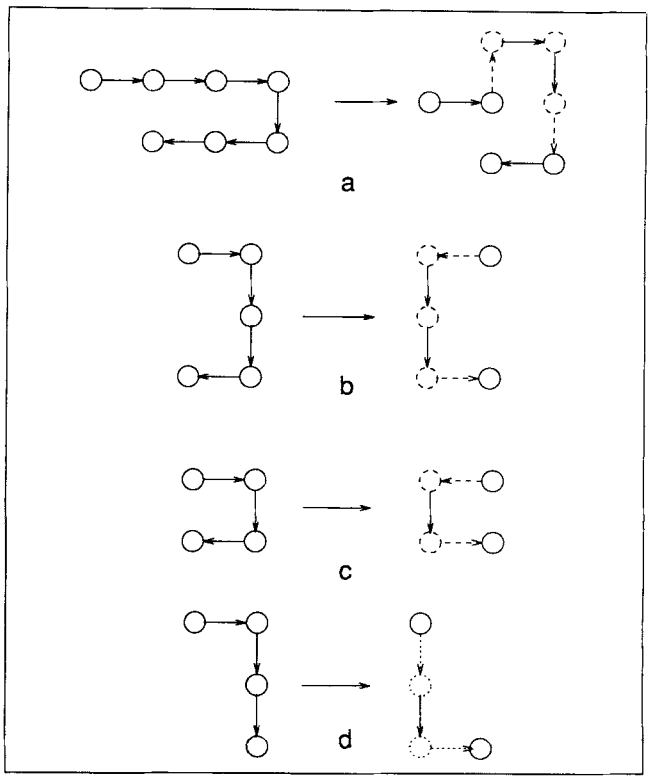

Figure 7. Examples of the three- and four-bond moves employed in the dynamic Monte Carlo simulations.

The dotted lines represent the new configuration after the move is completed. (a), (b) four-bond move; (c), (d) three-bond move. The four-bond move generates a new local conformation in the middle of the chain.

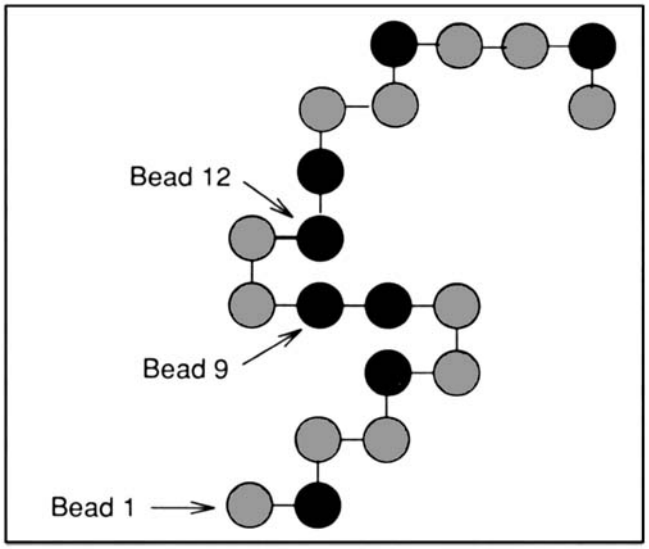

Figure 8. Protein chain with the middle section folded correctly.

The middle of the chain is said to be correctly folded when beads 9 and 12 form an HH contact. After the middle part of the chain folds, the ends come together to give rise to the native conformation.

shown at the bottom of the Figure 9. After the middle contact is formed, the chain proceeds along the pathways indicated toward the native state as shown. A few of the partially folded intermediates along the way (such as conformations (d) and (e)) have exposed hydrophobic sites. In a multichain environment, these exposed hydrophobic sites could bind to exposed hydrophobic sites on other partially folded chains. Therefore, "aggregation prone" intermediates exist for the hypothetical 20-bead protein and could cause aggregate formation.

Conclusion

The simple 2-D protein model seems to mimic real proteins in several respects. The protein chain exhibits a relatively sharp folding–unfolding transition and folds along preferred pathways via intermediates. The intermediates are

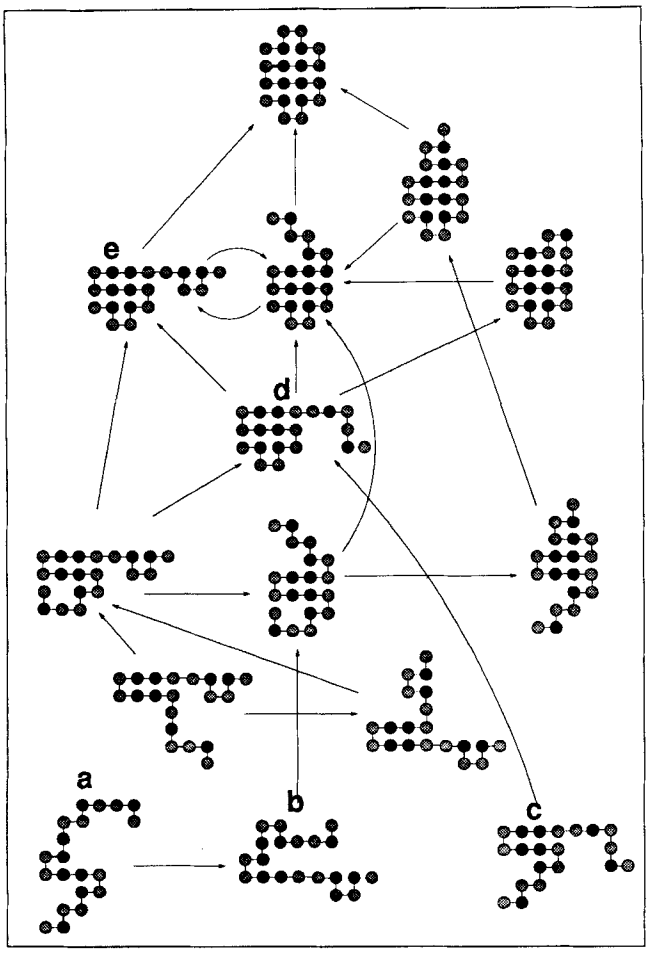

Figure 9. Partial pathmap for the 20-bead chain.

The chain forms distinct folding intermediates that appear to be capable of aggregating among themselves. Conformations (a), (b), and (c) have correctly folded middle sections. Conformations (d) and (e) have exposed hydrophobic sites that can cause aggregation.

partially folded compact states and appear capable of associating among themselves to cause aggregation. The folding process seems to be cooperative in that the chain does not start folding until the middle section has folded correctly even though the ends continuously fold and unfold. When the middle fold is finally formed, the rest of the chain "zips up" into the native conformation in a relatively short time.

The folding pathways and the observed intermediates are naturally dependent on the choice of moves employed in the simulation. However, the very existence of favored folding pathways and intermediates is probably an attribute of the folding process itself and has been noted earlier by Miller et al. (1992).

The model that we have investigated is limited in that the solvent molecules are not explicitly represented. Also, unfolding at low temperatures (Dill, 1990) is not predicted and site-specific interactions such as disulfide bonds are not included. Despite these limitations, the model is adequate for our purposes because it provides insights on the nature of protein refolding pathways and intermediates.

Acknowledgments

This work was supported by the National Science Foundation (grant CTS 9208590) and by the North Carolina Biotechnology Center (grant 9113-ARG-0205). The authors are grateful to Professor Ken Dill and Dr. Jeffrey Cleland for the guidance that they have provided.

Notation

k = Boltzmann constant
T = absolute temperature
ϵ = hydrophobic contact energy

Literature Cited

Dill, K. A., "Dominant Forces in Protein Folding," *Biochemistry*, **29**, 7133 (1990).

Gierasch, L. M., and J. King, eds., *Protein Folding: Deciphering the Second Half of the Genetic Code*, Amer. Assoc. for the Adv. of Sci. (1990).

Goldberg, M. E., and C. R. Zetina, "Importance of Interdomain Interactions in the Structure, Function and Stability of the F_1 and F_2 Domains Isolated from the β_2 Subunit of *E. coli* Tryptophan Synthase," *Protein Folding*, R. Jaenicke, ed., Elsevier North-Holland, Amsterdam, p. 469 (1980).

Kolinski, A., J. Skolnick, and R. Yaris, "Can Reptation Describe the Dynamics of Entangled, Finite Length Polymer Systems? A Model Simulation," *J. Chem. Phys.*, **86**, 1567 (1986).

Lau, K. F., and K. A. Dill, "A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins," *Macromol.*, **22**, 3986 (1989).

London, J., C. Skrzynia, and M. Goldberg, "Renaturation of *E. coli* Tryptophanase after Exposure to 8M Urea. Evidence for the Existence of the Nucleation Centers," *Eur. J. Biochem.*, **47**, 409 (1974).

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller, "Equation of State Calculations by Fast Computing Machines," *J. Chem. Phys.*, **21**, 1087 (1953).

Miller, R., C. A. Danko, M. J. Fasolka, A. C. Balazs, H. S. Chan, and K. A. Dill, "Folding Kinetics of Proteins and Copolymers," *J. Chem. Phys.*, **96**, 768 (1992).

Mitraki, A., and J. King, "Protein Folding Intermediates and Inclusion Body Formation," *Bio/Technol.*, **7**, 690 (1989).

Šali, A., E. Shaknovich, and M. Karplus, "How does a Protein Fold?" *Nature*, **369**, 258 (1994).

Taketomi, H., Y. Ueda, and N. Gō, "Studies of Protein Folding, Unfolding and Fluctuations by Computer Simulation," *Int. J. Pept. Prot. Res.*, **7**, 445 (1975).

Zettlemeissl, G., R. Rudolph, and R. Jaenicke, "Reconstitution of Lactic Dehydrogenase—Noncovalent Aggregation vs. Reactivation: Physical Properties and Kinetics of Aggregation," *Biochemistry*, **18**, 5567 (1979).

Manuscript received May 12, 1994, and revision received Sep. 23, 1994.