Analysis of Multiple Folding Routes of Proteins by a Coarse-Grained Dynamics Model

Burak Erman
Laboratory of Computational Biology, Sabanci University, Faculty of Engineering and Natural Sciences, Tuzla 81474 Istanbul, Turkey

ABSTRACT Langevin dynamics of a protein molecule with Go-type potentials is developed and used to analyze long time-scale events in the folding of cytochrome c. Several trajectories are generated, starting from random coil configurations and going to the native state, that are a few angstroms root mean square deviation (RMSD) from the native structure. The dynamics is controlled, to a large scale, by the two terminal helices that are in contact in the native state. These two helices form very early during folding, and depending on the trajectory, they either stabilize rapidly or break and re-form in going over steric barriers. The extended initial chain exhibits a rapid folding transition into a relatively compact shape, after which the helices are reorganized in a highly correlated manner. The time of formation of residue pair contacts strongly points to the hierarchical nature of folding; i.e., secondary structure forms first, followed by rearrangements of larger length scales at longer times. The kinetics of formation of native contacts is also analyzed, and the onset of a stable globular configuration, referred to as the molten globule in the literature, is identified. Predictions of the model are compared with extensive experimental data on cytochrome c.

INTRODUCTION

Many different pathways are available to a folding protein that starts from a random initial configuration and ends in the native state. Each path is determined by the succession of self-interactions of the elements that make up the protein molecule. In the interest of simplification, pathways may be described in terms of events, introduced explicitly into simulations recently by Hoang and Cieplak (2000a,b). Formation of secondary structure (for example, a helix), its interaction with another helix, the breaking of a helix, entrance of a loop in between two helices, etc., are a few specific examples of such events. Each pathway has its own succession of events. Some pathways consist of events that rapidly move the protein to its native structure. Then the protein is said to go through a funnel-like energy landscape. Much of the recent literature implicitly points to important events in folding (see Wolynes et al., 1995; Chan and Dill, 1998; Laurents and Baldwin, 1998). Some other trajectories consist of a succession of events that are unfavorable for the protein to reach its native state. For example, the entrance of a helix C into the space between two helices A and B is an unfavorable event if A and B are to be in close contact in the native state. Several different paths may evolve from this unfavorable conformation. Helix C may remain between helices A and B. This is a trap on the energy landscape, well surrounded by high hills on all sides, that prevents progress to the native state. On another pathway, either helix A or B or both may break and provide the possibility to helix C to escape from in between. On the energy landscape, the protein goes over a steric barrier. The breaking of helices to provide the flexibility to relieve unfavorable excluded volume interactions slows down the folding process but channels it toward its native state.

On the experimental side, many of the events during folding are already well identified and well documented (Laurents and Baldwin, 1998). Guided by experiments, the specific aim of the present paper is to introduce a coarse-grained model of folding dynamics, to generate several trajectories starting from different initial configurations, and to analyze the essential features of the types of events similar to those briefly mentioned in the preceding paragraph.

The folding dynamics of proteins can be studied at different levels of accuracy and detail. The most accurate and yet the most demanding of these methods is molecular dynamics (MD). Recently, it has been shown that it is possible to fold a 36-residue protein from a random initial configuration to within 1.5-Å root mean square deviation (RMSD) from the known native configuration (Duan and Kollman, 1998). These simulations, and many others along the line, have shown that the basic principles that govern the dynamics of synthetic polymers are essentially the same for proteins. It is only the need to specify the chemical specificity and structural details that makes the latter look more difficult. Currently, the computational speed of present-day computers is not sufficient for generating several trajectories to mimic the ensemble of proteins. Resort to coarse-grained techniques seems obligatory.

A coarse-grained model was introduced recently (Hoang and Cieplak, 2000a,b), where the Langevin equation is solved for a protein chain whose beads are subject to a Go-type potential. In this potential, the interactions between pairs of residues that are in known positions in the native state are assumed known in advance and used in the Lan-

Received for publication 4 June 2001 and in final form 24 August 2001.

Address reprint requests to Dr. Burak Erman, Sabanci University, Laboratory of Computational Biology, Faculty of Engineering and Natural Sciences, Tuzla 81474, Istanbul, Turkey. Tel.: 90-216-483-9505; Fax: 90-216-483-9550; E-mail: erman@sabanciuniv.edu.

0006-3495/01/12/3534/11 $2.00

gevin equation. This approach obviously lacks the generality of MD but yields solutions that are rapid and therefore makes it possible to study folding kinetics at different length and time scales for many trajectories. The use of a Go-type model essentially tells the beads of a protein where to go at the end of the trajectory, but not how to go. The latter is achieved through the solution of the equation of motion. And because to know how the beads move around in a folding protein is essential, the work of Hoang and Cieplak is important. The present work is similar in spirit to that of Hoang and Cieplak. The dynamics of the protein evolves through self-interactions of residues approximated by the α-carbons along the chain. Pairs of beads are attracted toward each other until they reach their separations of the native configuration. If the pair of beads is further brought closer, a strong repulsive force is operated that moves the pair away from each other. Chain connectivity resulting from the covalent bonds along the chain is preserved. During the course of folding, if two beads tend to come closer than the sum of the two bead radii, steric forces resulting from excluded volume interactions are applied to these beads as repulsive noise-like forces.

The Go model, first introduced to investigate on-lattice statistics of proteins (see review in Go, 1983), is now serving as a powerful approach for studying the folding pathways of proteins on energy landscapes. Notably, three recent studies (Galitskaya and Finkelstein, 1999; Munoz and Eaton, 1999; Alm and Baker, 1999) have used different versions of Go-type potentials in studying the unfolding or folding rates of proteins (see the recent commentary by Takada, 1999). All of the three models are able to predict the major trends in the folding rates of fast folding proteins.

In the first section below, the model is described in detail, followed by results of calculations on cytochrome c (Cyt c) and the discussion of these results from the point of view of numerous experiments performed on this system.

THEORY AND CALCULATIONS

The model

We adopt the C^{α} representation of the protein molecule, where the residues are represented by spherical beads along the chain. The position of the ith bead is r_i, expressed with respect to an external Cartesian coordinate frame. The equilibrium length of the covalent bond between a pair of successive beads along the chain is 3.88 Å. The $n - 1$ covalent bonds of the chain are not fixed at 3.88 Å, however, and may exhibit fluctuations about it. In the absence of rigid body translations, a protein of n residues has $3n - 3$ degrees of freedom according to the present representation.

Energies

The attractive and repulsive parts of the total energy are considered separately. First, the attractive energy, E_A is calculated as follows:

$$E_A = -\sum_{i=1}^{n-4}\sum_{j=i+4}^{n}\frac{E_{A,ij}}{r_{ij}^{m}} - \sum_{i=1}^{n-1}E_{A_b}r^2_{i,i+1} \qquad (1)$$

Here, the double sum represents the energy between nonbonded pairs. Interactions between pairs separated by four or more beads along the chain are included into the energy. r_{ij} is the scalar distance between the ith and the jth bead, and $E_{A,ij}$ and E_{Ab} are the parameters characterizing attractive inter-residue and bond energies, respectively. The exponent m shows how fast the attractive interaction between a pair of beads decays in space. For the Lennard-Jones (LJ) potential, it is generally taken as 6. However, results of recent calculations (Erman et. al., 1997) show that if two helices in a protein are taken as rigid bodies with residues as centers of interaction, the energy of interaction of the two helices scales with $m = 2$, where r_{ij} in this case is the distance between the centroids of the helices. And, more importantly, the range of interaction extends to over 20 Å in space. In more recent work (Erman and Dill, 2000; Erkip et al., 2001), it was shown that the minimum energy configurations of two-dimensional hydrophobic-polar (HP) models and three-dimensional real proteins may be obtained relatively easily if $m = -2$ for interactions between H pairs. Physically, this corresponds to a linear attractive spring between the H pairs. These observations indicate that protein calculations lead to satisfactory results if the attractive part of the interaction energy is taken as longer ranged than that of the classical LJ potential. Preliminary calculations for the present work in which parameters are optimized by trying different values for the various parameters showed that $m = 3$ is a satisfactory value for the exponent of the attractive energy. The second term in Eq. 1 is the attractive part of the covalent bond between a pair of neighboring beads along the chain.

Second, the repulsive energy, E_R, is calculated as follows:

$$E_R = \sum_{i,j}E_{R,ij}f(d_{ij} - r_{ij}) + \sum_{i}E_{Rb}f(1_b - r_{i,i+1}) \qquad (2)$$

Here, $E_{R,ij}$ and E_{rb} are the coefficients for repulsive inter-residue and bond energies, respectively, l is the bond length, d_{ij} is the diameter of the volume excluded to the pair of beads i and j, and the function $f(x)$ vanishes if x is negative and equates to unity if x is positive. A suitable functional form for f, adopted in the present calculations, is

$$f(x) = \frac{1}{1 + e^{-ax}} \qquad (3)$$

with a large value for the coefficient a.

Equation 2 is essentially an excluded volume condition, where beads i and j are not allowed to be in a volume of diameter d_{ij}. Similarly, two covalently bonded beads cannot be closer than the bond length.

Recently, Clementi et al. (1999) used the LJ potential for attractive and repulsive parts between residues. They used a quartic potential for the bond energy. The attractive and the repulsive parts of the inter-residue potential shown in Eqs. 1 and 3 correspond to the r_{ij}^{-6} and r_{ij}^{-12} terms of the LJ potential, respectively. According to the present choice, the attractive part dies off more slowly and the repulsive part is steeper than that of the LJ potential. The preference for a slowly dying attractive potential follows from our previous calculations on model and real proteins (Erman et al., 1997). The choice of a steeper repulsive potential in the present work is in good agreement with results of MD calculations of Clementi et al. (1999) as may be seen from Fig. 1 of their work. In many cases, it is possible to approximate the repulsive part of the potential by a steep energy well at a given cutoff distance below which the energy is infinitely large. The cutoff distance of 7 Å of the present work is in good agreement with 6.9 Å inferred by Clementi et al. That the steepness of the repulsive part below a minimum distance is significant is also demonstrated in our more detailed inter-residue potential calculations (Erman et al., 1997). The bond potential in the present work is chosen as the superposition of a quadratic attractive and a steep repulsive part as shown in Eqs. 1 and 3. Clementi et al. (1999) chose the superposition of a quadratic and quartic functions for the bond potential. The choice of the quartic part was to remove energy localization in some modes. The use of the strong repulsive bond potential in our work is based on similar arguments.

Random initial configurations

The folding process starts from a random and relatively extended initial configuration. The initial configuration of the chain is generated by expressing the coordinates of the ith bead in the coordinate system of the first bead as the vector $l_i = \mathrm{col}(l\sin\varphi_i \cos\theta_i, l\sin\varphi_i \sin\theta_i, l\cos\varphi_i)$, where col denotes column, and ϕ_i and θ_i are the polar and azimuthal angles of l_i with respect to the laboratory fixed coordinate system to which the first bond is affixed.

The equation of motion

The dynamics of the chain evolves according to

$$\frac{d\mathbf{r}_i}{dt} = -\alpha\nabla_{\mathbf{r}_i}(E_A + E_R), \qquad (4)$$

where $\mathbf{r}_i$ is the position vector of the ith bead and α is the coefficient that modulates the time step. The contribution of forces on the right-hand side of Eq. 4 is separated into two parts, E_A and E_R, coming from the attractive and repulsive parts of the energy, respectively. These two are different in nature. The attractive potential acts on a bead constantly throughout the folding process, whereas the repulsive potential acts only when two beads come closer than the allowed cutoff distance between them. The attractive force between two beads continuously tends to decrease the distance between them to zero. The repulsive force, however, is impulsive. A large value of a in Eq. 3 is needed for an impulsive repulsive force.

Equation 4 is a simplified version of the more general Langevin equation used by Hoang and Cieplak (2000b), which reads as

$$m\ddot{r} = \gamma\dot{r} + F_c + \Gamma, \qquad (4')$$

where m is the mass of the residue, γ is the friction coefficient or the viscosity with dimensions of (force)(time)/(length), F_c is the force derived as the negative gradient of the energy, $-\nabla E$, and Γ is the random force. Equation 4 is recovered after neglecting the inertia term (over-damped approximation) and the random force. Then, α is identified with inverse friction coefficient. Thus, Eq. 4 is a special case of the over-damped Langevin equation where the thermal noise in the latter is replaced by the impulses on a bead coming from other beads that violate the excluded volume condition. In Eq. 4, the coefficient α may be regarded as a scaling factor for time. A small value of α, corresponding to large friction, leads to slower decay of correlations. Calculations of Hoang and Cieplak (2000a) show that folding times of proteins scale linearly with γ.

Parameters used in calculations

The bond length $l = 3.88$ Å. The random initial configurations are generated by taking the polar angle of each bond as 60° and choosing the azimuthal angle randomly in the interval 0° to 360°. The minimum distance of approach between any pair of residues is taken as 7.0 Å. The coefficients in the energy expressions given in Eqs. 1 and 2 are optimized by training the model over several proteins (1d3b, 1c9o, 1dfn, 5pti, 1vii, 1nmf, 1tfg, and 1bla) such that the predicted native conformations have the smallest RMSD from the corresponding known Protein Data Bank structures. The RMSD values for the converged trajectories were less than 3.0 Å for all of these proteins. The values obtained in this manner are $m = 3$, $E_{A,ij} = 0.001$, $E_{Ab} = 0.01$, $E_{R,ij} = 0.001$, $E_{Rb} = 0.01$, and $a = 10$. The parameter α in Eq. 4 is taken as unity in the simulations.

The Go-type potential

We performed two sets of calculations. In the first set, the distances d_{ij} between all pairs i and j making contact in the native structure are assigned. This model is referred to as the

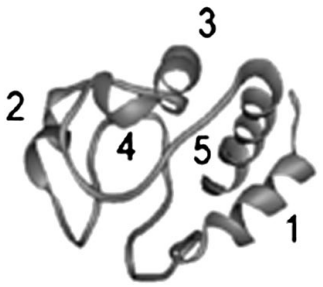

FIGURE 1 Ribbon diagram of the folded state of Cyt c. Numbers from 1 to 5 indicate the five helices.

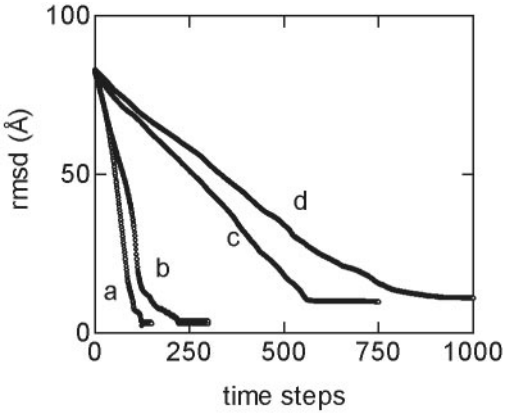

FIGURE 2 RMSD of the protein from its native configuration as a function of time steps. Curves a, b and c, d are obtained with models 2 and 1, respectively.

Go-type model in the literature. In the present calculations, the contact distance is chosen as 7.0 Å. Below, we identify this model as model 1. In the second set of calculations, we assigned the distances d_{ij} between every pair i and j separated by less than 20 Å in the native structure. This also corresponds to a Go-type model with a larger range of potentials prescribed at the outset. We identify this model as model 2.

Exploratory calculations showed that the main differences between the results of the two models are 1) the number of steps required for convergence of model 1 to the native state is three to four times that of model 2 and 2) the majority of the trajectories generated by model 2 converged to the native state, whereas convergence was poorer for model 1. For these two reasons, and for clarity of presentation, we report below mostly the results for model 2.

RESULTS OF CALCULATIONS AND COMPARISON WITH EXPERIMENTS

The ribbon diagram of the folded state of Cyt c (Protein Data Bank code 1CRC) is shown in Fig. 1, on which the helices are numbered from 1 to 5. They are identified in the following discussion as H1 (2–14), H2 (49–55), H3 (60–69), H4 (70–75), and H5 (87–104), the numbers in parentheses denoting the starting and ending residue indices. H1 and H5 are the two long terminal helices, the N- and C-terminal helices, respectively, which are spatially neighboring in the native state. The closest distance between the 6th and 94th C^{α} is 3.69 Å. H2 and H3 are connected by a short loop, H3 and H4 are neighboring along the chain, and H2 and H4 are spatially neighboring, with a distance of 6.1 Å between the 51st and 75th C^{α}.

Approximately one hundred trajectories for Cyt c, starting from random extended initial configurations, are generated for both models 1 and 2. Although each trajectory follows a different route in the phase space of the protein, some general features are readily observed that are present in the majority of the trajectories. Most of the discussion below is based on the trajectories obtained with model 2. Results of only few representative trajectories are presented in the figures.

About 60% of the trajectories generated by model 1, i.e., by the Go potential with 7-Å cutoff, were nonfolding trajectories for which the chain started from a random configuration and got stuck in a configuration from which it cannot escape and reach the folded state. This generally happens when two sequences, say, P and Q, which are in close proximity in the native state, cannot find a route to approach each other during the folding simulation because a third sequence S of beads is between them, and the steric interactions between the beads of P and S and between the beads of Q and S prevent S from being pushed out of the way. The remaining 40% of the trajectories are foldable. All samples in this group fold to within ~6–8 Å of the native structure. Among the trajectories observed in this group, the shortest and the longest folding routes took 600 and 900 calculation steps, respectively.

For model 2, less than 10% of the trajectories were nonfolders. The RMSD values between the predicted and the real native structure varies between 1.0 and 4.0 Å for the folding trajectories. The average number of calculation steps for model 2 is ~150. In Fig. 2, the results for model 1 (curves c and d) and model 2 (curves a and b) are shown. The abscissa represents the number of simulation steps, and the ordinate is the RMSD, in angstroms, from the native structure. Curves a and b, obtained by model 2, are examples of a fast and a slow converging trajectory, respectively. Curves c and d are obtained by model 1 and are examples for a fast and a slow trajectory, respectively. Overall analysis of the trajectories that are folders and nonfolders show a binary behavior: the protein either converges to the native structure smoothly and rapidly, as exemplified by the curves in Fig. 2, or gets trapped in a local minimum and does not converge at all within the time scale of the simulations.

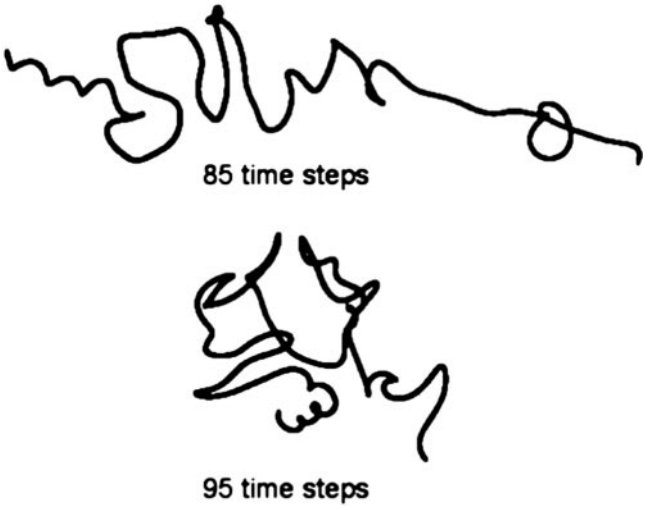

FIGURE 3 Configurations of the chain at the end of 85 (*top*) and 95 (*bottom*) time steps.

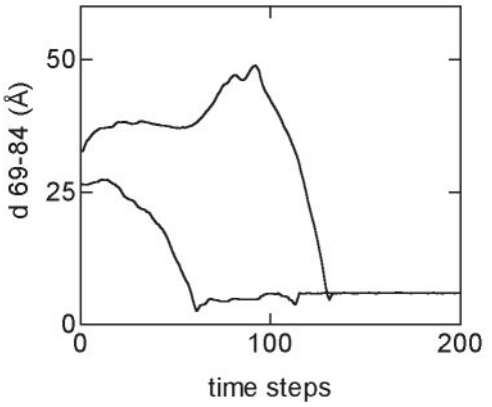

FIGURE 4 Change of the distance between residues GLU69 and GLY84 as a function of time.

In the rest of this section, we report the significant features of the folding of Cyt c as observed in the simulations and compare with experimental data whenever possible.

All of the trajectories generated show an initial part, within the first 100 time steps, where helices and some non-native tertiary structure form in a relatively open configuration. This early stage is followed by a sharp condensation into a relatively compact configuration. In the condensed state, the tertiary structure reorganizes until the native state is reached. The reorganization stage takes place between 100 and 300 time steps.

The early sequence of events is as follows. At the onset of folding, the helices H1 and H5 are formed independently of each other to within a RMSD of 4–5 Å of their native conformations, and the helix H3 has a very small resemblance to its native state. The independent formation of H1 and H5 into stable helices at the beginning of folding has been observed experimentally (Wu et al., 1993; Kuroda 1993; Sauder and Roder, 1998; Elöve et al., 1994). Two configurations of the chain, at the end of 85 and 95 time steps, respectively, are shown in Fig. 3. Transition from the extended to the globular compact state takes place rather rapidly, within the 10 time steps between 85 and 95. The RMSD values of the two conformations shown in Fig. 3 are 22.5 Å and 11.0 Å, respectively. During condensation, the portion of the chain comprising residues 1–66 is organized into a relatively dense structure. Thus, contrary to the widely adopted view of molten globule formation, only part of the chain condenses first. At the onset of condensation, or the collapse transition, the sequence between residues 69 and 84 is highly extended, and the residues 84–104 form an open, uncondensed conformation at this stage. This feature is observed in the majority of the trajectories generated. The average time of formation of this partial condensation takes place between 80 and 100 time steps. Then, the distance between residues GLU69 and GLY84 starts to decrease sharply, during which the C-terminal approaches the N-terminal. In Fig. 4, the change of the distance between GLU69 and GLY84, denoted by d 69–84 is shown. Two extreme cases are shown, in one of which the native distance of 5 Å is reached in ~60 and the other in 130 time steps. In addition to the decrease in distance, a significant amount of rotation also takes place about the line joining residues GLU69 and GLY84.

The helices of Cyt c have been the subject of several experiments. In the following few figures, we show the RMSD of the helices from their native shapes. In Fig. 5, the RMSD values of H1 are plotted as a function of time steps for several representative trajectories. In Fig. 5 *A*, one sees that within the first 100 time steps, helices H1 fold to within 2.5 Å of their native structure. Between steps 100 and 150, however, their helical structure breaks sharply, and monotonously, and re-forms between 150 and 200 time steps, after which they are stabilized to within a RMSD of 2–2.5 Å, which constitutes the native state. The set of five trajectories shown in Fig. 5 *B* are again for H1, where in this case the helix structure forms within the first 100 time steps and is stabilized from there on. The behavior of H1 during folding, as observed from several independent trajectories, may broadly be partitioned into two classes, as shown in Fig. 5. The early folding of H1, which has also been observed experimentally (Wu et al., 1993; Kuroda, 1993; Sauder and Roder, 1998; Elöve et al., 1994) is possibly due to the fact that being an end-helix, the translational constraints coming from chain connectivity are acting on its one end only. The same behavior is observed for the other end-helix, H5, and not for the internal helices, H2, -3, and –4, where translational constraints are acting on their two ends. The destruction of H1 between time steps 100 and 150, shown in Fig. 5 *A*, results from its steric interaction with helices H5 and the long loop between H4 and H5 during folding. Due to chain connectivity, the loop comes in between H1 and H5.

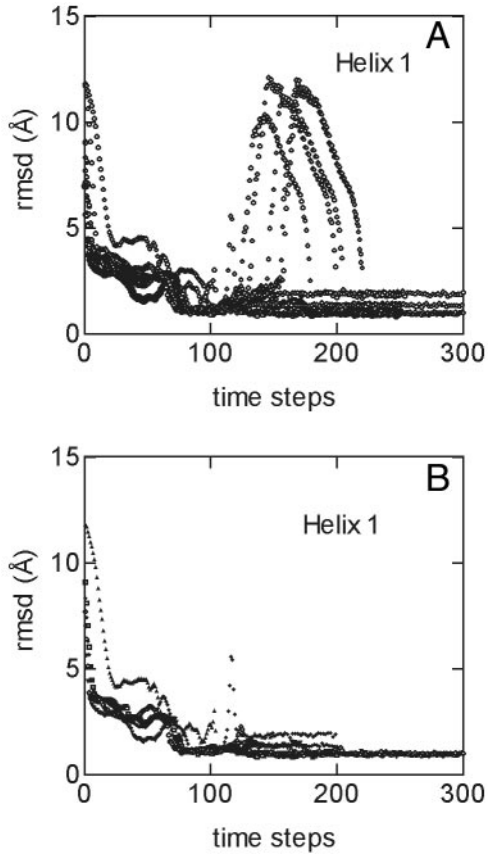

FIGURE 5 The RMSD values for helix H1 from its configuration in the native state as a function of time steps for different representative trajectories. (*A*) Trajectories in which H1 is unfolded in the dense state; (*B*) Stable helices throughout the trajectory.

The latter two should be in contact in the native state. The presence of another part of the chain between H1 and H5 is therefore a steric barrier to the motion of H1 and H5. This barrier is overcome by H1 by breaking its helix structure, which in turn increases its flexibility. Once the barriers are out of the way, H1 takes the helical shape again. The presence of steric barriers of the type presented here depends, however, on the initial configuration of the chain. In the examples shown in Fig. 5 *B*, H1 does not confront such barriers and takes the helical shape within the first 100 time steps, after which it remains helical. In the whole set of simulations, the number of H1 that confronts steric barriers was approximately twice that of those that do not. Thus, two routes may be identified for H1. In the first route, intermediate non-native forms of H1 appear, accumulate, and disappear. In the second route, there are no intermediates. Trajectories with no intermediates, depicted by the type shown in Fig. 5 *B*, for example, are not observed experimentally (see, for example, Mayne and Englander, 2000; Englander, 2000; Baldwin and Rose, 1999a,b)

Among the trajectories generated, more than two-thirds show that the two terminal helices approach each other at the onset of the molten globule stage, and form an H1/H5 intermediate. This observation agrees with earlier kinetic folding experiments of Roder et al. (1988) and more recent measurements of Xu et al. (1998) However, the remaining sequence of events observed from our calculated trajectories do not agree fully with those observed by Xu et al. The sequence of events according to their observations and our simulations are as follows. First, the two terminal helices are stabilized as an accumulating H1/H5 intermediate. Both simulations and experiment agree on this. Xu et al. observed that this is due to trapping of the sequence THR19-HIS33 in between the two terminal helices (Sosnick et al., 1994). Our trajectories mostly indicate that the sequence THR19-HIS33 reaches its native configuration at around the onset of the molten globule state and, in most of the trajectories, is always directed toward the outside the globule. The accumulation of H1/H5 in our case is mostly due to the interaction of the residues of H5 with those in the sequence between the residues THR49 and LYS87. This sequence contains the H3 and the segment PRO71-LYS87. The kinetics of this segment is studied experimentally by Xu et al. In most of the trajectories, H5 is disorganized and undergoes a series of excluded volume interactions with the intervening residues 49–87. In most of the trajectories generated, the sequence between HIS33 and THR49 reached its native configuration rather late during folding. This also agrees with the observations of Xu et al. Close examination of our trajectories shows that this is due to, in most cases, excluded volume interactions of this sequence with the sequence THR19-HIS33 at the early stages and with the sequence PRO71-LYS87 at the later stages of folding.

A closer analysis and comparison of folding and nonfolding trajectories shows that the kinetic barriers to folding in both are of similar origin. A frequently observed pattern of events is as follows. Two helices P and Q (mostly H1 and H2 in the case of Cyt c) that have formed in earlier stages get entangled with another segment of the chain that constitutes a steric barrier to the final approach of P and Q to form their native contacts. P and Q attempt to break their helical structures. This is essentially a move for a higher degree of flexibility. During this stage, several non-native contacts are formed. This is an intermediate state. In the event that P and Q disentangle themselves from the intervening segment and proceed toward the native state, this intermediate is called on-pathway. In the event that P and Q cannot escape from the sequence located between them, the trajectory cannot progress toward the native state. This is referred to as an off-pathway intermediate. We have not waited long enough in the present simulations to see whether an off-pathway intermediate can result in the native configuration. More detailed analysis of some of the nonfolding trajectories showed, however, that as time passes while H1 and H5 unsuccessfully attempt to relieve themselves from the intervening segment, the other parts of the

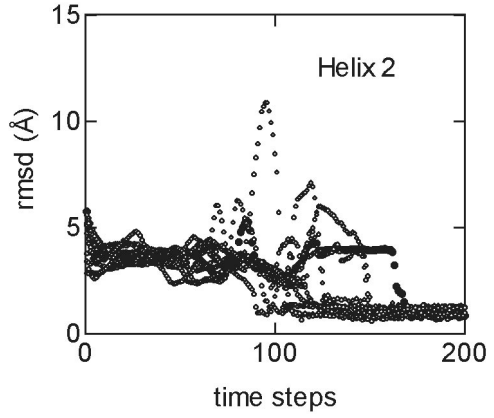

FIGURE 6 The RMSD values for the trajectories of H2 presented as a function of time steps. Trajectories are represented by the sequences of points.

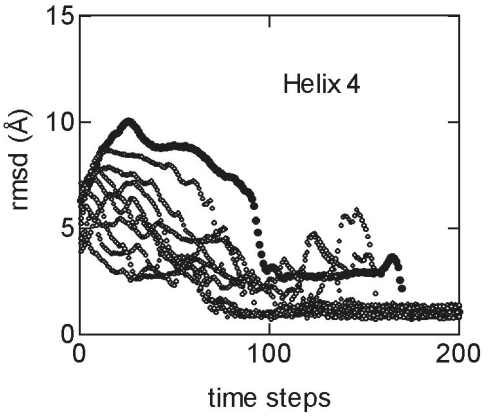

FIGURE 8 The RMSD values for the trajectories of H4 presented as a function of time steps. Trajectories are represented by the sequences of points.

molecule start drifting away into unfavorable configurations that are far from the native. Chan and Dill (1998) pointed out earlier that kinetic traps can account for both on- and off-pathways, similar to our observation of the events confronted by the two entangled helices P and Q.

In Fig. 6 representative trajectories for H2 are shown. Contrary to H1, H2 does not reach its native state until much later during folding. This is apparently due to the translational constraints acting at the two ends of H2. In some of the trajectories, helix breaking is observed, similar to that of H1, also resulting from the presence of steric barriers. Figs. 7 and 8 are for H3 and H4, respectively. The multi-peaked trajectories in Fig. 7 show that H3 performs several attempts for obtaining its helical structure. Most of these attempts are during the interval of time steps 0–100, where the chain has a relatively open configuration. Similar observations hold for H4 of Fig. 8. It is interesting to note from Fig. 7 that different trajectories result in different equilibrium configurations for H3, one of which has a RMSD of 1.2 Å and the other has 2.4 Å. In Fig. 8, the trajectory denoted by the darker circles shows that H4 is trapped in a steric barrier plateau between time steps 100 and 175. During this time, in which the protein is in a compact globular state, H4 is frozen in its non-native state, whereas the other helices rearrange in the attempt to reach their final conformations. This waiting in line of H4 while the others seek their native conformations is not uncommon and is seen in the other trajectories of the internal helices. In all of the trajectories, the helices H2, H3, and H4 exist in a partly formed state and exhibit large-scale fluctuations during the molten globule stage. These fluctuations are between 5 and 10 Å RMSD of the helices from their corresponding native conformations. The fact that H2, H3, and H4 are not formed completely in the molten globule state has recently been reported by Hostetter et al. (1999) based on the saturation mutagenesis at the evolutionarily invariant residue LEU68.

In Fig. 9, the RMSD values of H5 obtained from different

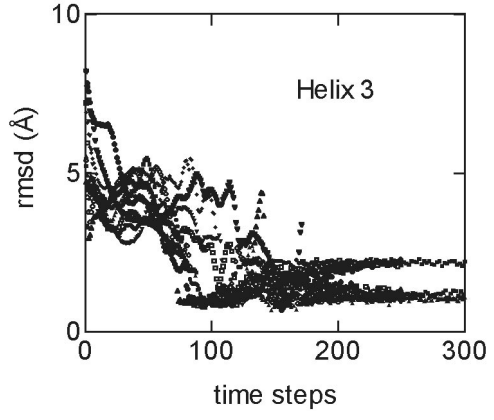

FIGURE 7 The RMSD values for the trajectories of H3 presented as a function of time steps. Trajectories are represented by the sequences of points.

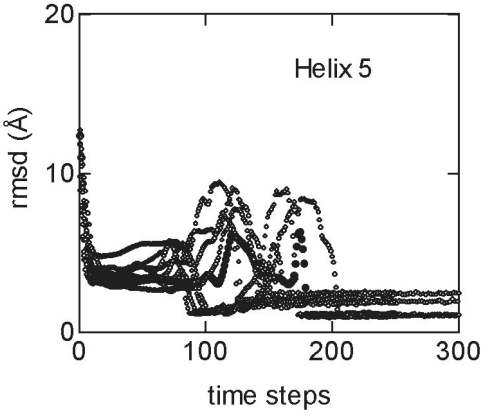

FIGURE 9 The RMSD values for the trajectories of H5 presented as a function of time steps. Trajectories are represented by the sequences of points.

trajectories are presented as a function of time steps. Similar to the behavior of H1, H5 exhibits an initial helix formation within a few time steps to a RMSD of 3–5 Å. These RMSD values are approximately equal to those of H1. However, the initial formation of H5 is faster than that of H1. In most of the trajectories of Fig. 9, a disruption of the helix structure is noticeable between 90 and 200 time steps, a behavior very similar to that of H1. The equilibrium RMSDs cluster into three different values for H5 at 1.0, 2.0, and 2.5 Å. The breaking of the helix structure observed for H1 and H5 results from the cooperative nature of their motions during the time interval that follows the rapid folding transition of the protein into a compact globular shape. The time up to ~100 time steps may be viewed as the burst phase where folding occurs via the rapid emergence of secondary and tertiary structure, preceding the relatively slow stretch between 100–200 time steps that results in the native state (Parker and Marqusee, 2000). The folding of Cyt c may be divided into two distinct stages. In the first stage, H1 and H5 approach each other without interaction. In the second stage they interact strongly. It is this highly correlated stage where the important contacts between residues of H1 and H5 are established. Previous experimental and theoretical work (Elöve et al., 1992; Colon et al., 1996; Colon and Roder, 1996; Roder et al., 1988) have shown that H1 and H5 play a critical role in the folding of Cyt c. The motions of H2, H3, and H4 may be regarded as dynamic reactions to the motions of H1 and H5, the major determinants of the dynamics.

Our simulations show that the chain folds in two discrete stages. Both of the stages consist of highly cooperative motions. However, the nature of cooperativity is different in each. In the first stage, including the beginning of collapse, where the chain is highly extended, part of the chain arranges into a relatively compact form and the remaining part rapidly condenses on the first part. This is possible only by the dominance of 1) long-range attractive forces, 2) nonrandom hydrophobic interactions, and 3) local concerted motions resulting from specific inter-residue interactions. These motions may be identified with a kinetically controlled search for a minimum free energy structure. Recently, Hagen and Eaton (2000) analyzed the rapid folding of Cyt c at nanosecond time resolution and observed the effects of cooperativity during the collapse. In the second stage, the molten globule state, portions of the chain reorganize in a highly cooperative way, in which the cooperativity is dominated by excluded volume interactions. The Go potentials chosen in the present simulations try to drive the molecule to its native state. However, due to the denseness of the configurations in the molten state, any move toward the native state is counteracted by an excluded volume reaction. Therefore the movement of the topology toward that of the native one is slow in the second stage. It is this second stage where a major misligation with the Haem group may occur, as discussed below, that slows down the folding process.

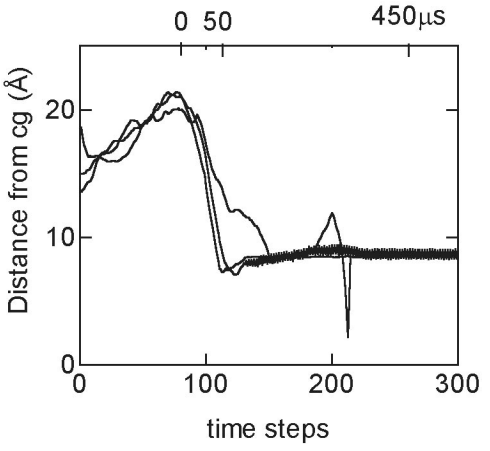

FIGURE 10 Time dependence of the distance of TRP59 from the instantaneous centroid of the chain.

The fact that the chain folds in two distinct stages has been shown by experiments from Roder group. The folding kinetics of proteins are conveniently studied by capillary mixing experiments. The earlier instruments used were of the stop-flow mixing type. However, many of the fast events in protein folding fall within the dead time of these instruments. Recently, Roder introduced a continuous flow rapid mixing device (Shastry et. al., 1998) which allows the detection of submillisecond events. Observations on acid-unfolded Cyt c with folding initiated by a rapid pH jump to native conditions were made in the time range of 45–900 μs.(Shastry and Roder, 1998) Based on the fluorescence decay of TRP59, Shastry and Roder observed a rapid (~50 μs) formation of a compact state, followed by a longer period (~450 μs) of reorganization. The presence of an extended stage where the tertiary structure exhibits extensive reorganization, observed by the continuous flow experiments and the present calculations, is clearly inconsistent with a two-state mechanism proposed earlier. In Fig. 10, some representative trajectories of the distance of TRP59 from the instantaneous centroid of the chain are shown. In a significant number of the trajectories generated, the behavior before 100 time steps was similar, first exhibiting a small increase of the distance of TRP59 from the centroid of the chain and then exhibiting a sharp approach to the centroid around 100 time steps. A closer inspection of the trajectories showed that the initial increase of the distance from the centroid results from the condensation of the portion of the chain between residues 1 and 59, as a result of which the centroid of the whole chain shifts toward the region occupied by the open, coiled section of residues 60–104. An absolute time scale cannot be ascribed to the folding of Cyt c in the laboratory due to 1) the uncertainty of defining an initial configuration at time $t = 0$ and 2) the

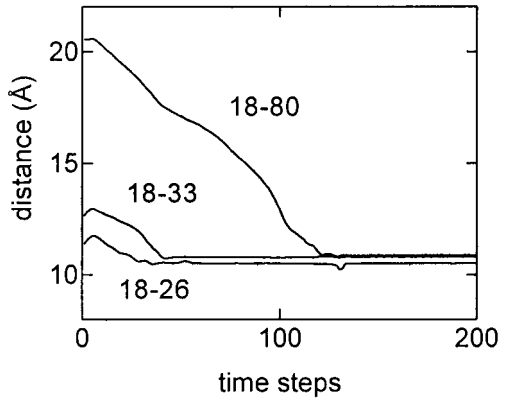

FIGURE 11 Time dependence of the distance between HIS18 and the three residues HIS26, HIS33, and MET80.

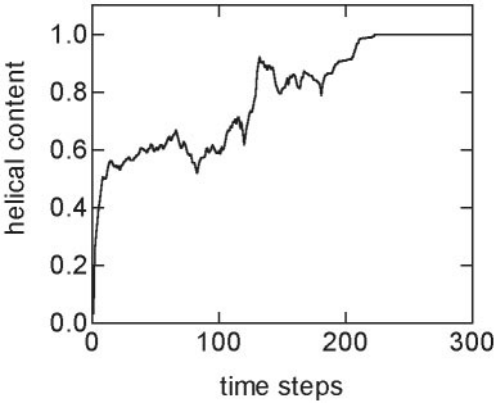

FIGURE 12 Evolution of the helix content of Cyt c with time.

fact that folding rates depend strongly on the pH (Sosnick et al., 1994). It is, however, possible to make an order of magnitude estimate of the time scale of the simulations by comparing Fig. 10 with results of the Roder experiments. Accordingly, the estimated absolute time scale is shown on the upper horizontal axis of Fig. 10. The onset of folding is chosen at the peak of the curves in Fig. 10, where the chain is still in a highly extended conformation. Between time 0 and 50 μs, TRP59 moves abruptly to the centroid of the chain, as observed by the Roder group. From then on until 400 μs, the condensed state is observed during which the portions of the chain reorganizes. Comparison of the lower and upper abscissas shows that roughly one computational time step corresponds to $\sim$2–3 μs. This comparison should be taken as a very crude order of magnitude estimate, however.

In the simulations, the initial extended configurations of the chains were generated randomly. It is interesting to note that the distance of TRP59 from the centroid of the chain behaves similarly in the different simulations. This is in agreement with experimental observations of the Roder group. Based on the fluorescence decay of TRP59, Shastry and Roder (1998) showed that the rate of collapse is independent of initial conditions.

In the native state of Cyt c, HIS18 and MET80 are ligated to a heme group. The distance between the α carbons of the two ligands is 10.9 Å. During the molten state, HIS18 remains intact with the heme group, but mis-ligation occurs by the ligation of HIS26 or HIS33 in place of MET80. The trajectories here show that the distance between HIS18 and HIS33 reaches the value of 10.9 Å even before the collapsed state and remains at this value, as shown for a representative trajectory in Fig. 11. The distance between HIS18 and HIS26 is less than 10.9 Å at all times. The distance between HIS18 and MET80 reaches 10.9 Å only after the molten globule state is reached. We conclude from these observations that a HIS18/HIS33 mis-ligation is possible even before the chain collapses to its ligand exchange phase. Mis-ligation by HIS33 substitution has been observed in submillisecond folding experiments of Yeh et al. (1997) at suitable pH values. Mis-ligation by HIS33 is thought to slow down the folding reactions in the molten globule stage. Without an explicit specific potential that favors the stated ligation in Cyt c, whether such a ligation will form or not cannot be determined with the present model. The model in its present state tells only that if mis-ligation can occur, it will occur. It is to be noted however that Akiyama et al. (2000) have reached the conclusion that mis-ligation is not the determining factor that causes the slowness of the second stage at the native conditions of pH 4.5.

Results of the present simulations shed some light on the controversy existing on the question of whether folding of Cyt c fits the framework model or the hydrophobic collapse model. According to the framework model, collapse results in a highly structured ensemble of states that are on the folding pathway. According to the hydrophobic collapse model, the collapsed state lacks any secondary or tertiary structure and is simply a response to the change in solvent conditions (Yeh and Rousseau, 2000). The present calculations show that the helical content (for H1 and H5) at the initial stages, which usually falls within the dead time of many stop-flow experiments, is significant. In Fig. 12, the helical content is plotted as a function of time steps. Helical content is obtained as follows: if the distances between residues i, $i + 3$, and i, $i + 4$ are both within 10% of their native values in the helices H1–H5, then the segment i, $i + 4$ is accepted to be in the helical state. The curve is an average over all trajectories of model 2. The helical content increases sharply to $\sim$55% within the first 15 time steps. This value is more than the 20% that is observed in the recent submillisecond measurements by Akiyama et al. (2000). Between time steps 70 and 100, during which the chain exhibits a sharp collapse transition, helical content decreases. Between time steps 100 and 200, helical content exhibits large fluctuations. In the majority of the trajectories, the helices have to go through a significant amount of

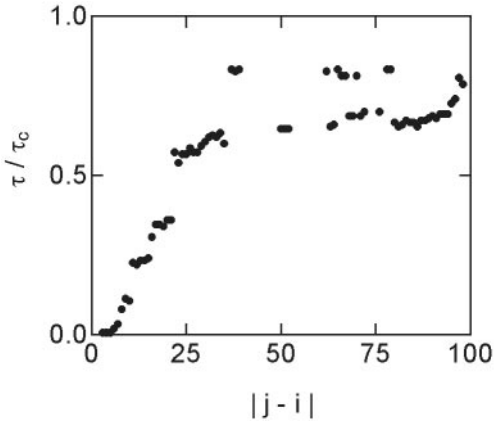

FIGURE 13 Time of formation, τ, of a native contact as a function of the distance $|i - j|$ between residues i and j. τ_c normalizes the ordinate values.

distortion during this time span to relieve the steric, excluded volume, barriers on their way to the final configuration. The distortion of the helices proceeds, in general, by sharp bending (similar to jack-knifing), and a loss of helicity of one part of the bent shape. These observations are in qualitative but not quantitative agreement with the recent far-UV time-resolved circular dichroism measurements of Chen et al. (1999), where the helix content formed in the first 5 μs of folding is destroyed almost totally between 16 μs and 1 ms and re-formed around 320 ms. In the present simulations, the destruction of helices, although is present due to the mechanisms stated above, is not as extensive as observed by Chen et al. In conclusion, the collapsed state contains a significant amount of secondary structure, which reorganizes until the native structure is obtained.

The formation of native contacts in a folding protein may be characterized by introducing a contact correlation time for each pair that are in contact in the native state. We define the contact correlation time, τ, as the first time when a pair of C^α come to within 10% of the contact distance of 7 Å. The time τ_c of folding is the time when the protein reaches its native state during the simulation. In Fig. 13, τ/τ_c is plotted as a function of separation between two C^α along the backbone. The points are averages over all trajectories of model 2. The contact correlation time increases approximately linearly with residue separation $|i - j|$, up to a value of $|i - j| = 30$, and then levels off and becomes independent of residue separation. This is in agreement with the findings of Baker and colleagues (Plaxco et al., 1998). The linear dependence agrees with recent findings of Hoang and Cieplak (2000b). It may be concluded from this that first the helices form, followed by their rearrangements in space. The simulations above showed that some helices form and break and re-form during folding. Our definition of τ here ignores the breaking and re-forming and adopts the time of first formation.

CONCLUSION AND SUMMARY

We presented a coarse-grained off-lattice folding model for proteins in the C^α representation, subject to Go-type potentials. Several trajectories are generated and some common features of folding events are identified. A major advantage of coarse-grained simulations is their fast convergence, as a result of which several trajectories may be generated and a statistical analysis can be made.

The results obtained from the simulations can be summarized as follows:

1. The folding times of the trajectories are narrowly distributed. The maximum deviations from average folding times are $\sim\pm30\%$.
2. The protein starts folding by partial condensation of its residues between 1 and 66. The sequence between residues 66 and 84 is highly extended, and the residues 84–104 form an open, coil-like conformation. Helices H1 and H5 and some tertiary non-native structure between residues 1 and 66 form during this initial stage. The helical content is in the order of 50% at this stage.
3. This early stage is followed by a sharp condensation into a relatively compact configuration. Helix content decreases significantly during the collapse transition. The rate of collapse is independent of initial conditions.
4. Collapse transition takes place by the decrease of the distance between residues GLU69 and GLY84 and the approach of the C-terminal to the N-terminal.
5. In the condensed state, the tertiary structure exhibits extensive reorganization. This takes place through the following mechanisms: 1) destruction and reformation of helices (mostly H1 and H5) and 2) translation and rotation of helices, to relieve steric hindrances imposed by non-native contacts.
6. In the condensed state, the distance between THR19 and HIS33 remains around 10.9 Å and is suitable for forming a mis-ligated complex with the heme group. The distance between THR19 and MET80 reaches 10.9 Å much later in the condensed state.
7. Native contacts form sequentially. First, residues closer along the chain form contact. Residues further away along the chain form native contacts later. Accordingly, the helices form first. Then the helices rearrange in space to form the final native structure.

REFERENCES

Akiyama, S., S. Takahashi, K. Ishimori, and I. Morishima. 2000. Stepwise formation of alpha helices during cyt c folding. *Nat. Struct. Biol.* 7:514–520.

Alm, E., and D. Baker. 1999. Matching theory and experiment. *Curr. Opin. Struct. Biol.* 9:189–196.

Baldwin, R. L., and G. D. Rose. 1999a. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biol. Sci.* 24:26–33.

Baldwin, R. L., and G. D. Rose. 1999b. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biol. Sci.* 24:77–83.

Chan, H. S., and K. A. Dill. 1998. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins Struct. Funct. Genet.* 30:1–33.

Chen, E., P. Wittung-Stafshede, and D. S. Kliger. 1999. Far-UV time-resolved circular dichroism detection of electron transfer triggered cytochrome c folding. *J. Am. Chem. Soc.* 121:3811–3817.

Clementi, C., M. Vendruscolo, A. Maritan, and E. Domany. 1999. Proteins: structure, function and genetics. Folding Lennard-Jones proteins by a contact potential. 37:544–553.

Colon, W., G. A. Elöve, L. P. Wakem, F. Sherman, and H. Roder. 1996. Side chain packing of the N- and C-terminal helices plays a critical role in the kinetics of cytochrome c folding. *Biochemistry.* 35:5538–5549.

Colon, W., and H. Roder. 1996. Kinetic intermediates in the formation of cytochrome c molten globule. *Nat. Struct. Biol.* 3:1019–1025.

Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a1-microsecond simulation in aqueous solution. *Science.* 282:740–744.

Elöve, G. A., A. K. Bhuyan, and H. Roder. 1994. Kinetic mechanism of cytochrome c folding: involvement of the heme and its ligands. *Biochemistry.* 33:6925–6935.

Elöve, G. A., A. F. Chafotte, H. Roder, and M. E. Goldberg. 1992. Early steps in cytochrome c folding probed by time-resolved circular dichroism and fluorescence spectroscopy. *Biochemistry.* 31:6876–6883.

Englander, S. W. 2000. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomol. Struct.* 29:213–238.

Erkip, A., B. Erman, C. Seok, and K. A. Dill. 2001. Parameter optimization for the Gaussian model of protein folding. *Polymer.* In press.

Erman, B., I. Bahar, and R. L. Jernigan. 1997. Equilibrium states of rigid bodies with multiple interaction sites: application to protein helices. *J. Chem. Phys.* 107:2046–2059.

Erman, B., and K. A. Dill. 2000. Gaussian theory of protein folding. *J. Chem. Phys.* 112:1050–1056.

Galitskaya, O. V., and A. V. Finkelstein. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. U.S.A.* 96:11299–11304.

Go, N. 1983. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12:183–210.

Hagen, S. J., and W. A. Eaton. 2000. Two-state expansion and collapse of a polypeptide. *J. Mol. Biol.* 301:1019–1027.

Hoang, T. X., and M. Cieplak. 2000a. Molecular dynamics of folding of secondary structures in Go-type models of proteins. *J. Chem. Phys.* 112:6851–6862.

Hoang, T. X., and M. Cieplak. 2000b. Sequencing of folding events in Go-type proteins. *J. Chem. Phys.* 113:8319–8328.

Hostetter, D. R., G. T. Weatherly, J. R. Beasley, K. Bortone, D. S. Cohen, S. A. Finger, P. Hardwidge, D. S. Kakouras, A. J. Saunders, S. K. Trojak, J. C. Waldner, and G. J. Pielak. 1999. Partially formed native tertiary interactions in the A-state of cyt c. *J. Mol. Biol.* 289:639–644.

Kuroda, Y. 1993. Residual helical structure in the C-terminal fragment of cytochrome c. *Biochemistry.* 32:1219–1224.

Laurents, D. V., and R. L. Baldwin. 1998. Protein folding: matching theory and experiment. *Biophys. J.* 75:428–434.

Mayne, L., and W. Englander. 2000. Two-state vs multistate protein unfolding studies by optical melting and hydrogen exchange. *Protein Sci.* 9:1873–1877.

Munoz, V., and W. E. Eaton. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A.* 96:11311–11316.

Parker, M. J., and S. Marqusee. 2000. A statistical appraisal of native state hydrogen exchange data: evidence for a burst phase continuum? *J. Mol. Biol.* 300:1361–1375.

Plaxco, K. W., K. T. Simons, and D. Baker. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–994.

Portman, J. J., S. Takada, and P. G. Wolynes. 2001. Microscopic theory of protein folding rates. I. Fine structure of the free energy profile and folding routes from a variational approach. *J. Chem. Phys.* 114: 5069–5081.

Roder, H., G. Elöve, and S. W. Englander. 1988. Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. *Nature.* 335:700–704.

Sauder, J. M., and H. Roder. 1998. Amide protection in an early folding intermediate of cytochrome c. *Folding Design.* 3:293–301.

Shastry, M. C. R., S. D. Luck, and H. Roder. 1998. A continuous flow capillary mixing method to monitor reactions on the microsecond time scale. *Biophys. J.* 74:2714–2721.

Shastry, M. C. R., and H. Roder. 1998. Evidence for barrier-limited protein folding kinetics on the microsecond time scale. *Nat. Struct. Biol.* 5:385–392.

Sosnick, T. R., L. Mayne, R. Hiller, and S. W. Englander. 1994. The barriers in protein folding. *Nat. Struct. Biol.* 1:149–156.

Takada, S. 1999. Go-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* 96:11698–11700.

Wolynes, P. G., J. N. Onuchic, and D. Thirumalai. 1995. Navigating the folding routes. *Science.* 267:1619–1620.

Wu, L. C., P. B. Laub, G. A. Elöve, J. Carey, and H. Roder. 1993. A noncovalent peptide complex as a model for an early folding intermediate of cytochrome c. *Biochemistry.* 32:10271–10276.

Xu, Y., L. Mayne, and S. W. Englander. 1998. Evidence for an unfolding and refolding pathway in cytochrome c. *Nat. Struct. Biol.* 5:774–778.

Yeh, S. R., and D. L. Rousseau. 2000. Hierarchical folding of cyt c. *Nat. Struct. Biol.* 7:443–445.

Yeh, S. Y., S. Takahashi, B. Fan, and D. L. Rousseau. 1997. Ligand exchange during cytochrome c folding. *Nat. Struct. Biol.* 4:51–56.