Relationships Between Unfolded Configurations of Proteins and Dynamics of Folding to the Native State

ATTILA GURSOY, OZLEM KESKIN, METIN TURKAY, BURAK ERMAN

College of Engineering and Center for Computational Biology and Bioinformatics, Koç University, İstanbul 34450, Turkey

Received 26 April 2006; revised 18 July 2006; accepted 4 August 2006

DOI: 10.1002/polb.21018

Published online in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: We compare folding trajectories of chymotrypsin inhibitor (CI2) using a dynamic Monte Carlo scheme with Go-type potentials. The model considers the four backbone atoms of each residue and a sphere centered around C^β the diameter of which is chosen according to the type of the side group. Bond lengths and bond angles are kept fixed. Folding trajectories are obtained by giving random increments to the φ and ψ torsion angles with some bias toward the native state. Excluded volume effects are considered. Two sets of 20 trajectories are obtained, with different initial configurations. The first set is generated from random initial configurations. The initial configurations of the second set are generated according to knowledge-based neighbor dependent torsion probabilities derived from triplets in the Protein Data Bank. Compared to chains with randomly generated initial configurations, those generated with neighbor-dependent probabilities (i) fold faster, (ii) have better defined secondary structure elements, and (iii) have less number of non-native contacts during folding. © 2006 Wiley Periodicals, Inc. *J Polym Sci Part B: Polym Phys* 44: 3667–3678, 2006

Keywords: chymotrypsin inhibitor; folding trajectories; go-type model; Monte-Carlo simulation; protein folding dynamics

INTRODUCTION

The unfolded state of proteins has important role on the folding process and on the stability of the folded state.¹ The unfolded state is now being recognized as the ‘other half’ of the protein folding equation.² In this paper, we show, by a dynamic Monte Carlo and statistical mechanical analysis and by using knowledge-based torsion potentials, that the unfolded state is not random. Important correlations exist between the torsion angles in the unfolded state. These correlations are important because they stabilize fragments of secondary structure and reduce the number of initial conformations available to the protein chain. We show

that the conformations that survive in the unfolded state have significantly larger number of native-like contacts compared to those in chains subject to random potentials and this affects the folding rate significantly.

There is growing evidence, both theoretical and experimental, that the unfolded configurations of proteins are not random, and that a significant amount of structure is present that resembles the structure of the folded native state.^{2–4} Recent fluorescence resonance energy transfer experiments on nascent proteins point out to the fact that proteins have native-like conformations even in the ribosome.⁵ The fact that the unfolded state is not as heterogeneous as it was thought to be has been pointed out in several earlier studies.^{6–8} The presence of native-like features in the unfolded protein has important consequences concerning its folding into its native state. As pointed out by several authors,^{2,3,6}

Correspondence to: B. Erman (E-mail: berman@ku.edu.tr)

Journal of Polymer Science: Part B: Polymer Physics, Vol. 44, 3667–3678 (2006)
© 2006 Wiley Periodicals, Inc.

Figure 1. The native conformation of chymotrypsin inhibitor 2 (CI2).

the configuration space of the unfolded state is significantly reduced compared to that of a random chain. In our previous work, we showed that this reduction is mainly determined by the neighbor dependent correlations over the successive torsion angles of the chain and that there is a strong relationship between amino acid sequence and backbone torsion angle preferences.⁹ Recently, Avbelj and Baldwin showed that the backbone angle φ of a residue is more negative when its preceding neighbor is class L (aromatic and β - branched: F, H, W, Y and I, T, V, respectively).¹⁰ They showed that peptide backbone solvation is the probable cause. Avbelj and Baldwin¹⁰ also discuss preferences for helices and β - structures and their dependence on the residue type. These observations are significant indicators of the context dependence and nonrandomness of the configurations of nascent proteins.

Our aim in the present article is to analyze the effect of initial conformations on the folding process. It is shown that initial conformations chosen according to energetically appropriate conditions fold faster to the native state, whereas random initial conditions end up mostly in dead ends along the folding pathway or at best in slow folding pathways. Specifically, we use a dynamic Go model. We show that unfolded proteins whose initial conformations are generated by knowledge-based potentials fold to the native state much faster than those generated randomly do. Recently, Zaman et al.¹¹ showed that the populations and transitions between Ramachandran basins for monomers, dimers, and trimers of amino acids are influenced by the conformation and identity of the nearest

neighbors. The present work is the first treatment, at the full protein scale, of the effect of initial conformation and residue identity on the folding dynamics. We choose a widely studied protein, chymotrypsin inhibitor (CI2), as our example.

In the next section, we describe the protein CI2 in detail and summarize previous experimental and computational work that details the folding dynamics of CI2 and in the section following that we describe our computational model and outline the calculation scheme. In Results of Calculations section, we present the results of the study and show the differences between dynamics of folding from initial configurations from random and knowledge-based potentials.

THE MODEL PROTEIN CI2

Chymotrypsin Inhibitor 2 (CI2) is a 64-residue protein that consists of an α -helix and a four-strand β -sheet as shown in Figure 1. The crystal structure has a resolution of 2.0 Å (PDB code 2CI2). The main hydrophobic core is formed by the packing of the α -helix against the β -sheet. It was the first protein shown to fold by a two-state mechanism.⁷

CI2 contains one α -helix and four β -strands as shown in Table 1 where the secondary structure elements are identified using PDB.¹² The hydrophobic core is formed by packing α -helix to the β -strand 2. In the native state, the residues in contact at the hydrophobic core are 5, 8, 16, 19, 20, 27, 29, 47, 49, 51, 57, and 61. The coordinates for the atoms in the first 18 residues are missing in the PDB file; therefore, we renumbered the residues in our work starting from 1 at the residue number 19 in the original PDB file.

The folding of CI2 is well studied both experimentally^{13–18} and computationally.^{17,18} The idea of studying this protein was a turning point, which gave insight into protein folding and new experi-

Table 1. Secondary Structure Elements and Corresponding Residue Numbers for CI2

Residue Numbers	Type of Secondary Structure
3–5	β -Strand 1
12–24	α -Helix
28–34	β -Strand 2
35–44	Recognition loop
45–52	β -Strand 3
60–64	β -Strand 4

Figure 2. Five-atom representation of the $(i - 1)$ th, i th, and $(i + 1)$ th residues of the protein chain.

mental techniques. CI2 was found to fold and unfold as a simple two-state system with no kinetic intermediates.¹⁹ Mutational analysis, Φ -value^{19,20} allows the investigation of the role of individual residues in the TS (transition state) for folding. This method has also been used by many theoreticians to understand the TS structure and the folding mechanism of simple protein models.^{21–27} Thus, the method has stimulated interactions between experimentalists and theoreticians as it allows the understanding of folding mechanisms.²⁷ The all-atom MD simulations of unfolding trajectories of CI2 at high temperatures gave insights into the nature of the TS and denatured state ensemble.^{17,28} The folding dynamics and kinetics of CI2 by using an all-atom MD with a Go-potential was studied.²⁸ They separated the TS ensembles into two sets, (1) with α -helices disrupted and β -strands intact, (2) with α -helices intact and β -strands disrupted. They found that the first set has a much higher success rate to fold. The unfolding MD trajectories of CI2 were also analyzed to clarify the TS.²⁴ It was observed that the TS region for folding and unfolding occurs early with only 25% of the native contacts. Another interesting result from the latter simulations was the emergence of the statistically preferred unfolding pathway. A C^α model was used to investigate the CI2 model.²¹ The simulations revealed that a few well-defined contacts were formed with high probability in the TS that drives the protein into its folded conformation.

Experimental and theoretical studies suggest that CI2 folds by a nucleation-condensation/collapse mechanism.⁷ The earliest folding events were suggested as the clustering of hydrophobic residues (19–21, 29–35) using NMR experiments.^{17,28} Molecular dynamics simulations also suggested that residues 49–51 may participate in

the clustering of hydrophobic residues.²⁹ The main event during folding was suggested as the closing of the α -helix to β -strand 3.³⁰ A combination of NMR experiments and molecular dynamics simulations also suggested nucleation around residues 16, 49, and 57.¹⁷ In addition, the root-mean squared deviation (rmsd) from the native structure during folding suggests sudden collapse after contacts between the α -helix and β -strand 1 are formed.³¹

Monitoring the native contacts may help understanding of the folding mechanism using targeted molecular dynamics methods. Ferrara et al. performed 128 folding and 45 unfolding molecular dynamic runs on CI2 and reported that majority of the native contacts are formed at the end of folding.³² Similar results are confirmed for wild-type and circular-permuted CI2.³³

MODEL AND CALCULATIONS

Description of Chain Kinematics

We adopt the five-atom representation of the protein chain, i.e., the four atoms N, C^α , C, and carbonyl O in the backbone, and a C^β atom representing the side group. The total number of atoms of the model, N , is equal to $5n$ where n is the number of residues. The side groups are considered as hard spheres with different van der Waals radii centered on the C^β atoms. The atomic radii given in Rowland and Taylor for each amino acid are used in the model.³⁴ The bond lengths of the atoms in the protein backbone are assumed to be constant during folding and the default values for these bond lengths are used.³⁵ In addition, the bond angles of the amino acids are also assumed to be fixed. The $i - 1$ st, i th and $(i + 1)$ th residues are shown in Figure 2.

The torsion angles around the bond between atoms $C_i - N_{i+1}$, that is represented as ω_i is fixed at π radians due to the partial double bond nature of the corresponding bonds. Only the two torsion angles, φ_i and ψ_i change.

The position of atom j is given by $r_j = [x_j \ y_j \ z_j]^T$ in the Cartesian coordinate system. Configurations of the protein chains are generated according to the rotational isomeric state formalism and successive matrix multiplication scheme.³⁶ The position vector r_j of the j th atom can be expressed as,

$$r_j = A_{[1]} \left[\prod_{k=2}^{j-1} A_k \right] A_{[j]} \quad (1)$$

where,

$$A_{[1]} = [T_1 \ I_1], \quad A_k = \begin{bmatrix} T_k & I_k \\ 0 & 1 \end{bmatrix}, \quad A_{[j]} = \begin{bmatrix} I_j \\ 1 \end{bmatrix} \quad (2)$$

In eq 2, T_k is the transformation matrix for the k th bond, I_k is the k th bond vector expressed in the coordinate frame of the bond, and 0 is the row vector $[0 \ 0 \ 0]$. The position vector is expressed in the internal coordinate system of the protein which is established by setting the x -axis along the direction of the first bond between N_1 and C_1^α atoms. The torsion angle of the first bond is fixed at 0° and the y -axis is chosen in the plane defined by the bonds 1 and 2. The z -axis of the internal coordinate system is chosen to complete the right-handed system.

The rotational isomeric state representation of the protein chain and the internal coordinate system adopted in the model is computationally efficient, the precise location of any atom in space can be calculated for any given set of torsion angles, φ_i , $i = 2, \dots, n$, and ψ_i , $i = 1, \dots, n - 1$ leading to $2n - 2$ for the degrees of freedom for a protein chain.

The Go-Model and Simulation of Folding Trajectories

We use a Monte Carlo-Go model for generating the folding trajectories.^{37–39} The model assumes the native state of the protein as the most stable configuration and drives the chain from a given initial state to the native. Since the native conformations of the protein are resolved in water, the effect of water is implicit in the Go-potentials used. The trajectories are generated by applying small perturbations to randomly chosen φ and ψ angles. At any step, a small number of angles (10% of the total number of angles) are subjected to perturbations with a maximum magnitude of 0.05 radians.

This small magnitude of perturbations is chosen to reduce the number of atom pairs that violate excluded volume condition at each move. Each perturbation of the torsion angles contains a random component and a bias so that the dihedral angles are driven to their native values. At the end of each step, the atom pairs are checked for excluded volume given by the van der Waals radii. The calculation of the distances between all atom pairs, of the order n^2 in number, is computationally expensive. We employed a grid-based range search algorithm to speed up the collision detection. The algorithmic complexity of this method is linear in the number of neighboring atoms on the average.⁴⁰ If the distance between two atoms, i and j , is less than the sum of their van der Waals radii, then three torsion angles along the backbone between atoms i and j are modified until all of atom pairs satisfy the excluded volume constraints given by the van der Waals radii. When all collisions among atom pairs are eliminated, the simulation of the folding trajectory continues with the next timestep in the Go model. Simulations were terminated when the rmsd between the native and computed structure was less than 0.3 Å.

Determination of the Initial Conformations

Initial configurations were generated in two ways: (i) in the first method, each initial configuration was generated by choosing each φ and ψ angle randomly. All initial configurations were relaxed to eliminate steric clashes. (ii) The second method was a knowledge-based approach in which triplets from Protein Data Bank structures were used to generate energy maps for $\psi_{i-1} - \varphi_i$, $\varphi_i - \psi_i$, and $\psi_i - \varphi_{i+1}$ pairs of dihedral angles. Keskin et al. gives the details of the method for extracting pair wise dependent probabilities.⁹

Residue-specific conformational potentials are developed on the basis of the probabilities of observations, $P_{XYZ}(\varphi_i, \psi_i)$ and $P_{XYZ}(\psi_i, \varphi_{i+1})$, of the configurations of each triplet XYZ . Here, $P_{XYZ}(\varphi_i, \psi_i)$ is the probability of observing the middle residue (type Y) to be in state (φ_i, ψ_i) and $P_{XYZ}(\psi_i, \varphi_{i+1})$ is the probability of observing residue type Y to be in state (ψ_i) and Z to be in state (φ_{i+1}) . $P_{XYZ}(\varphi_i, \psi_i)$ is a measure of the intraresidue correlation of torsion angles and $P_{XYZ}(\psi_i, \varphi_{i+1})$ is a measure of the inter-residue correlation of two successive torsion angles. We use the discrete state formalism for the torsion angles. Each φ and ψ angle is divided into intervals of 30° ; thus, 12 representative torsion states are assigned to each bond.^{41–43} The statis-

tics for the $\varphi_i\psi_i$ and $\psi_i\varphi_{i+1}$ pairs could as well be obtained from pairs of residues, YZ without incorporating the $i - 1$ th residue, X, along the sequence. This would improve the statistics in the nonredundant database significantly. However, energy maps obtained for doublets and triplets exhibit significant differences. Most importantly, energy maps based on doublets contain different numbers of minima than those based on triplets.

The pairwise dependent probabilities $P_{XYZ}(\varphi_i, \psi_i)$ and $P_{XYZ}(\psi_i, \varphi_{i+1})$ are calculated according to,

$$\begin{aligned} P_{XYZ}(\varphi_i, \psi_i) &= N_{XYZ}(\varphi_i, \psi_i) / \sum N_{XYZ} \\ P_{XYZ}(\psi_i, \varphi_{i+1}) &= N_{XYZ}(\psi_i, \varphi_{i+1}) / \sum N_{XYZ} \end{aligned} \quad (3)$$

where $N_{XYZ}(\varphi_i, \psi_i)$ indicates the number of residue triplets observed having the indicated values of the argument. The definition for $N_{XYZ}(\psi_i, \varphi_{i+1})$ is similar to the definition of $P_{XYZ}(\psi_i, \varphi_{i+1})$. The term $\sum N_{XYZ}$ in the denominator is the total number of observed triplets of XYZ in all possible states.

We define a conformational energy for a given residue Y (in the triplet XYZ) along the primary sequence of the protein as:

$$\begin{aligned} E_{XYZ}(\varphi_i, \psi_i) &= -RT \ln [P_{XYZ}(\varphi_i, \psi_i) / P_{XYZ}^0(\varphi_i)P_{XYZ}^0(\psi_i)] \\ E_{XYZ}(\psi_i, \varphi_{i+1}) &= -RT \ln [P_{XYZ}(\psi_i, \varphi_{i+1}) / P_{XYZ}^0(\psi_i)P_{XYZ}^0(\varphi_{i+1})] \end{aligned} \quad (4)$$

where, $P_{XYZ}^0(\varphi_i)$, $P_{XYZ}^0(\psi_i)$, and $P_{XYZ}^0(\varphi_{i+1})$ are the uniform distribution probabilities, that is, those valid when all angles are equally probable. In continuous space, they are equal to $1/2\pi$, and; in the discrete state formalism, they are directly proportional to the size of the angular intervals of the states ($1/12$).⁴¹

The statistical weight matrix $u_{\eta\zeta,i}$ for a given bond pair $i - 1$ and i is given as:⁴⁴⁻⁴⁶

$$U_i = u_{\eta\zeta,i} = \exp\left(-\frac{E_{\eta\zeta,i-1,i}}{RT}\right) \quad (5)$$

Here the $\eta\zeta$ th element of U_i indicates the statistical weight for bond i when it is in state ζ while the bond $i - 1$ is in state η .

The partition function, Z , of a chain that is n residues long is obtained according to,

$$Z = J^* \left[\prod_{i=2}^{2n} U_i \right] J \quad (6)$$

where $J^* = [1 \ 0 \ \dots \ 0]$; $J = \text{column}[1 \ 1 \ \dots \ 1]$

Journal of Polymer Science: Part B: Polymer Physics
DOI 10.1002/polb

The probability $P_{\zeta,i}$ That bond i will be in state ζ is given by

$$p_{\zeta,i} = Z^{-1} J^* \left[\prod_{m=2}^{i-1} U_m \right] U'_{\zeta,i} \left[\prod_{m=i+1}^{2n} U_m \right] J \quad (7)$$

Here, $U'_{\zeta,i}$ is the matrix obtained by equating the entries of all of its columns to zero except those of column ζ . The joint probability $P_{\eta\zeta,i-1,i}$ that bond $i - 1$ is in state η and bond i is in state ζ is given by,

$$p_{\eta\zeta,i-1,i} = Z^{-1} J^* \left[\prod_{m=2}^{i-1} U_m \right] U'_{\eta\zeta,i} \left[\prod_{m=i+1}^{2n} U_m \right] J \quad (8)$$

The conditional probability $q_{\eta\zeta,i-1,i}$ that bond i will be in state ζ , given that bond $i - 1$ is in state η is obtained from,

$$q_{\eta\zeta,i-1,i} = \frac{p_{\eta\zeta,i-1,i}}{p_{\eta,i-1}} \quad (9)$$

An initial conformation for the CI2 molecule with knowledge-based potentials is generated by adopting the conditional probabilities given in eq 9. Random initial configurations, on the other hand, are generated by randomly assigning each torsion angle into one of the 12 intervals between -180° and 180° .

RESULTS OF CALCULATIONS

The method of initial chain conformation generation described in the preceding section is used to obtain 10^4 initial conformations for the CI2 molecule. The statistical analysis of the initial conformations is made on this set of 10^4 conformations. This set is referred to as the conditionally generated (CG) set. Out of these 10^4 initial conformations, 20 are chosen randomly and the folding trajectory is generated for each, using the Go-model described above. The same analysis is repeated for 10^4 initial conformations generated randomly, which we refer as the randomly generated (RG) set.

Initial Conformations

The rmsd of the initial structures from the native structure range between 8 and 36 Å for both sets of initial conformations, RG and CG. Both sets show similar distributions. The CG set has a mean of 16.2 ± 3.2 Å and the RG set has a mean of 17.0 ± 3.4 Å. CG's and RG's have a mean squared radius of gyration (R_g^2) of 356 ± 150 Å² and 323 ± 144 Å²,

Figure 3. Components of the persistence vector for RG (○) and CG (●).

respectively. For both sets, the squared radii of gyration range between 75 and 1225 Å². Although the rmsd and R_g^2 distributions of the initial structures are comparable for the RG and CG populations, the persistence lengths for the two sets differ significantly. The persistence length is the magnitude of the persistence vector, $\mathbf{a}_n$, defined as

$$\mathbf{a}_n = \langle x_{0n} \rangle \mathbf{i} + \langle y_{0n} \rangle \mathbf{j} + \langle z_{0n} \rangle \mathbf{k} \quad (10)$$

where, the terms in brackets represent the x , y , and z components of the averaged end-to-end vector, and $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are the unit vectors along the x , y , and z axes, respectively. The persistence vector reflects the anisotropy of chain dimensions when the first two bonds are held fixed in the x - y plane.^{36,45–47} In Figure 3, we compare the different components of the persistence vector for RG and CG. Calculations are performed as follows: the first bond of each chain (the bond between the first and second atoms, N and C $^\alpha$ atoms) is fixed along the x -axis of a Cartesian coordinate system O_{xyz} . The second bond (the bond between the C $^\alpha$ and C atoms along the backbone) is fixed in the xy plane. The persistence vector, $\langle \mathbf{r}_{0i} \rangle$, between the zeroth atom fixed at the origin and the i th atom is obtained by averaging over the whole ensemble of chains generated. In Figure 3, the ordinate values are the components of the persistence vector between the zeroth and the i th atom of the chain, in angstroms, plotted as a function of atom indices. Filled circles represent results for the CG chains, the three components of the vector being shown on each curve. Similarly, the empty circles denote results for the RG chains. The x , y , and z components of the RG chains are 3.59, 4.98, and 0.09 Å, respectively. For the CG chains these values

are 8.66, 11.35, and -4.98 Å. One sees that for the CG chains, the components of the persistence vector exhibit significant differences, indicating different preferences along different directions relative to the three Cartesian coordinates. Differences in preference are much less pronounced for the RG chains. The effects of holding the first and second bonds in the xy -plane are exhibited in the x and y components. This effect dies off approximately when the 20th atom is reached for the RG chains, corresponding to four residues. However, the effect of fixing the first two bonds in the xy -plane reaches up to 25–75 bonds, or 8–25 residues, for the CG chains. Thus, correlations between residues along the CG generated protein chain reach over very large distances. The z component of the persistence vector dies off readily for the RG chains, indicating that correlations are rapidly randomized in this direction. However, the z -component strongly persists, in the negative direction, for the CG chains. The magnitude of the persistence vector for the full CG chains is 15.1 Å while it is 6.1 Å for the RG chains.

In Figure 4, the mean-squared end-to-end distance between the zeroth atom and the i th atom of the RG (empty circles) and CG (filled circles) chains are shown. Characteristic ratios, C_i , given in the ordinate are obtained by dividing the mean-squared end-to-end distances by il^2 , where i is the atom index and l is the bond length taken as 1.47 Å. The characteristic ratio for the RG set rapidly smooths out, indicating the absence of local correlations. The characteristic ratio for the CG set exhibits fluctuations. The characteristic ratio C_∞ for the whole RG and CG chains are 4.79 and 5.38,

Figure 4. Characteristic ratio for the i th atom of the chains. See legend for Figure 3.

respectively. These two values are strikingly close to each other. The apparently unexpected result showing that the persistence vectors are significantly different from each other for RG and CG chains while the characteristic ratios are similar is due to presence of strong torsion angle correlations in the CG chains as a result of which they are not yet in the scaling regime. One may suitably analyze the chain in terms of the worm-like or the Porod-Kratky chain for which the mean-squared end-to-end distance $\langle r^2 \rangle$ and the projection a of the persistence vector along the first bond are related according to the expression⁴⁵

$$\frac{\langle r^2 \rangle_0}{L} = 2a \left[1 - \left(\frac{a}{L} \right) \left(1 - e^{-L/a} \right) \right] \quad (11)$$

where L is the contour length and the subscript zero indicates the absence of long range correlations. The contour length L for CI2 is 276 Å. For the RG chains, $\langle r^2 \rangle_0 = 4.79$. Substitution of these values into eq 11 and solving for a yields $a_{\text{RG}} = 3.64$ Å, which agrees well with the value of 3.59 Å calculated from the simulations. For the CG chains, $\langle r^2 \rangle_0 = 5.38$ Å. Substitution into eq 11 and solving for the persistence length yields $a_{\text{CG}} = 4.10$ Å. Calculations from the simulations give, however, $a_{\text{CG}} = 8.66$ Å, which is more than twice of the Porod-Kratky chain. This indicates that the CI2 chain is not yet in the scaling regime and strong conformational correlations inside the chain dominate the statistics of the overall chain.

Bond Correlations

In this section, we investigate the statistical correlations between two bonds of the chain. The Flory Isolated Residue Pair hypothesis^{45,48,49} states that the conformations of a given residue are statistically independent of those of the neighboring residues in the random configurations of the protein. Subsequent work, however, showed that certain conformations of neighboring bonds of proteins are correlated.^{9,11,50} The presence of such correlations (or anticorrelations) is important since they may possibly affect the folding of the peptide chain and the stabilization of secondary structure.

The state η of bond i and the state ζ of bond j are uncorrelated if the joint probability $P_{\eta\zeta;ij}$ is equal to the product of the corresponding singlet probabilities, $P_{\eta;i}$ and $P_{\zeta;j}$. The ratio

$$C_{\eta\zeta;ij} = \frac{P_{\eta\zeta;ij}}{P_{\eta;i}P_{\zeta;j}} \quad (12)$$

is a very sensitive measure of correlations in this respect. If the occurrence of state η in bond i

enhances (decreases) the occurrence of state ζ in bond j , then the ratio in eq 12 is larger (smaller) than unity and the two states of the two bonds are correlated (anticorrelated). In a random chain, the ratio is unity for all pairs of states.

The neighbor dependent potentials adopted in this study lead to strong dependences between the specific states of neighboring bonds. To answer the more meaningful and more general question of whether two bonds of the chain are correlated or not, one needs to sum eq 12 over all states of the two bonds. For this purpose, we adopted the χ^2 test for estimating the degree of correlation between the i th and j th bonds. To emphasize the effect of secondary structure on correlations, we grouped each ψ angle into three states: State 1 is the helix state with $-150^\circ \leq \psi \leq -70^\circ$. State 2 is the β state with $90^\circ \leq \psi \leq 180^\circ$. State 3 comprises all the remaining values of the ψ angle. All the states have $-180^\circ \leq \phi \leq -70^\circ$. The χ^2 -test is then performed according to the expression⁵⁰

$$\chi_{ij}^2 = \sum_{\zeta=1}^3 \sum_{\eta=1}^3 \frac{(\eta_{\zeta\eta;ij} - \varepsilon_{\zeta\eta;ij})^2}{\varepsilon_{\zeta\eta;ij}} \quad (13)$$

where, χ_{ij}^2 denotes the value for the two bonds i and j , $n_{\zeta\eta;ij}$ is the number of configurations for which the i 'th bond is in state ζ and the j 'th bond is in state η , and $\varepsilon_{\zeta\eta;ij} = n p_{\zeta;i} p_{\eta;j}$. In the last expression, n is the total number of configurations. The total number, 10^4 , of configurations generated was sufficient to ensure⁵¹ that $\varepsilon_{\zeta\eta;ij} \geq 5$ for all $\zeta\eta$, ij pairs of the chain, both for the RG and the CG sets. The critical value of χ_{ij}^2 is 18.7 for a confidence level of 99.9% for the system of three states: α -, β - and all others.⁵¹ The χ_{ij}^2 values ranged between 0.022 and 18.27 for all bond pairs of the RG set. Therefore, within this confidence level, any two bonds of the random set are independent. For the CG set, the χ_{ij}^2 values ranged between 0.016 and 4654. The values of χ_{ij}^2 between 0.016 and 18.7 indicate that the corresponding bonds i and j are uncorrelated for the confidence level of 99.9%. The spectrum of χ_{ij}^2 values larger than 18.7 was approximately continuous up to 600 and showed a large gap between 600 and 1000. The strongest correlations are indicated by χ_{ij}^2 values larger than 1000. In Figure 5, the regions with χ_{ij}^2 values larger than 1000 are shown as a function of residue index. The upper band shows the six secondary structures of the native protein, (β , α , $\beta 2$, recognition loop, $\beta 3$ and $\beta 4$). The lower four black strips indicate the groups of residues having significant correlations among the ψ -bonds. The first group of correlations

Figure 5. Correlations among ψ -bonds (shown by the lower black strips), and secondary structure regions of the native protein (shown by the upper black strips). Rec. loop denotes the recognition loop.

is among the ψ -bonds between residues 18 and 24. This region corresponds to the α -helix. The second group of correlations is among the ψ -bonds between residues 29 and 33 in the β_2 structure. The third group of correlations is among the ψ -bonds between residues 37 and 41 in the recognition loop. The fourth group of correlations is among the ψ -bonds between residues 46 and 53 comprising the β_3 structure of CI2. Thus, correlations among torsion angles in the unfolded state are harbingers of secondary structure regions of the native protein. Furthermore, the most probable values of more than 90% of the correlated ψ -bonds, obtained using eq 7, coincide with the native value of the corresponding angles.⁹ We therefore conclude that correlations among bonds in the unfolded state have the effect of stabilizing the secondary structures.

Simulation Times

We performed 20 simulations starting with CG chains and 20 simulations starting with RG chains. Figure 6 displays the simulation times in increasing order, for each set. Solid bars represent

the simulation time for CGs and empty bars represent those for RGs.

Three of the 20 simulations were unsuccessful (they were nonfolders) in the CG set. On the other hand, 8 of the 20 simulations of the RG set did not reach to the folded structure. Success rates are 0.85 and 0.60 for the two cases. The average folding time for the first set is 11×10^6 time steps. This number increases to 20×10^6 time steps for the second set of simulations. These indicate that the structures starting with conditional probabilities are favored in the folding process compared to the ones starting with random structures. These may suggest that the conditional probability matrices are important to guide the initial coordinates.

Trajectories

Twenty simulations performed starting with CG chains display distinct behaviors along the folding trajectory. Two of the examples are shown in Figure 7 where the rmsd from the native structure is presented as a function of simulation time. Part (A) is a typical example where the simulation starts and a sudden decrease in the rmsd in the early folding times is observed. Eleven of the simulations show such behavior. There is a pseudoequilibrium region where the residues fluctuate and all the structures have rmsd around 6 Å. Then, there is another sudden decrease in the rmsd toward the end of simulation, where the native state is reached. The final folding is always observed to be very fast in a couple of simulation times. Usually the rmsd stays around 6 Å, and suddenly drops to an rmsd (0.3 Å) at which the simulations are terminated. Final adjustments are completed just after the folding nucleus was formed during the pseudoequilibrium region. Part (B) is another example where the structure exhibits several transitions along its folding pathway. Large deviations from the native structure result from attempts for

Figure 6. Folding times for 20 different conditional and random simulations.

Figure 7. Rmsd from the native structure during a simulation process starting with conditionally generated set.

relieving local contacts of residue pairs as discussed in more detail latter.

Residue Contacts

Figure 8 displays the behavior of contacts during two of the simulations. The x -axis is the simulation time during a trajectory and the y -axis represents the residue numbers in CI2. A dot is placed in Figure 8 for residue i if it is in contact with another residue at a given simulation time. Results from the two trajectories are presented in Figure 8(A,B). In Figure 8(A) the four bands around residues 13–17, 28–32, 47–51, and 56–59 are the most discernible ones. In Figure 8(B), three bands around residues 28–32, 47–51, and 56–59 are the most persistent ones. All other trajectories conform to either Figure 8(A or B). Fersht and Daggett reported that there is no compact structure until residues 49–57 pack against each other and stabilize the native helix.⁵ Then these residues contact with residue 16 to form the folding core. Thus, residues 16, 49, and 57 form the folding nucleus.⁴⁵ It is interesting to see that two residues 49 and 57 are always in contact during the simulation. This is just one of the examples of the simulations, indeed in most cases (13 out of 17 folders) these residues form their contacts in early stages of folding and stay in contact during the simulations.

Cyclization Probabilities

Another observation from our simulations is that the RG chains exhibit higher number of non-

native contacts throughout their folding trajectories. On the other hand, the CG chains have a higher number of native contacts compared to the

Figure 8. Residues that are in contact during the simulation.

Figure 9. The ratio of native contacts for a random (RG) and conditional (CG) simulation along the time steps.

non-native contacts. The fraction of native contacts, Q , can be used to describe the folding process.⁵² The Q variable has a value very close to zero for highly denatured protein chains and reaches unity in the native state. Contact of two residues separated by at least two other residues are used in defining cyclization probabilities. Cyclization probabilities may be divided into two, cyclization relating to native and the other relating to non-native contacts.

The total number of contacts (both native and non-native) in a chain at any time is $n_c = n_n + n_{nn}$. Here, n_c , n_n , n_{nn} are the numbers of total, native, and non-native contacts in a chain, respectively. The value of Q equates to $Q = \frac{n_n}{n_n + n_{nn}}$. In Figure 9, we compare the evolution of Q along the folding trajectory for an RG (empty circles) and a CG chain (filled circles) whose folding times are close to each other. In the initial state, the ratio of the native contacts, Q , for the two chains are equal and around 0.57. However, the Q value of the CG chain becomes larger than 0.9 at the beginning of the folding simulation. The points for the RG chain remain below that of the CG chain until the late stages of the trajectory.

In Figure 10, the average of Q is presented for each trajectory. The abscissa represents the twenty trajectories. Each point is obtained by averaging the Q values over a full trajectory. The filled and empty circles correspond to results of CG and RG simulations, respectively. Both CG and RG Q values are rearranged in increasing order to illustrate the difference between the two. The figure

shows that CG Q values are always above that of the RG chains. The means of the CG and RG Q values are $0.88 (\pm 0.06)$ and $0.81 (\pm 0.10)$, respectively. We also performed the Wilcoxon nonparametric rank test to see if the difference between medians of these two datasets is statistically significant. We obtained a P -value of 0.02 (at $\alpha = 0.05$). Therefore, we concluded that the CG chains behave statistically different during the simulations compared to the RG chains.

The differences between the average native contact probabilities presented in Figure 10 result from two important factors. First, the initial configurations of neither the RG nor the CG chains have knots. The number of repeat units of the CI2 chain is too small for the chain to exhibit knots while being produced from the amino to carboxyl direction. The persistence length of the RG chains, being 6.1 Å, is approximately equal to that of polyethylene chains.⁴⁷ The persistence length of the CG chains, on the other hand, is 15.1 Å, which is significantly higher than that of the random chains. The value obtained for the CG chains is somewhat smaller than the persistence length for polyaniline calculated as 22.5 Å by Conrad and Flory.⁵³ This results from the specific $\phi - \psi$ propensities inherent in the non-native state of the CG chains. As a result, dense nonlocal non-native contacts form in the initial stages of the RG chains and persist throughout the complete folding trajectory. Secondly, our simulations indicate that native-like blocks of several consecutive amino acids form at the early stages of the trajectories

Figure 10. The average ratio of native contacts for the 20 random (RG) and 20 conditional (CG) simulations. The simulations are sorted with increasing Q values.

resulting from local rearrangements for CG chains. During the remaining stages of the trajectories, these blocks move toward their relative positions that obtain in the native state. These second types of motions dominate most of the trajectories since they are large-scale motions and require more time. The time of native contact formation of two residues i and j increases linearly with the number of residues $|j - i + 1|$ between them.⁵⁴ Regions connecting such blocks play crucial roles in folding. In RG chains, the connecting regions form tight turns that, once formed, require a long time to break and transform into the native configuration. The most time consuming motions in the Go-model simulations are those that transform one tightly formed region into another. Here, we use the term 'tightly formed' to indicate a configuration that contains local contacts. In CG chains, on the other hand, the connecting regions in general do not contain local contacts, in the earlier stages of their trajectories, and therefore require less time to change their shapes from one configuration to another.

CONCLUSION AND DISCUSSION

The following are the major conclusions of the present study of CI2:

1. Folding times of CG chains are at least two times faster than RG chains. There are eight nonfolders in the RG set compared to the three in the CG set. These numbers indicate the importance of initial structures during folding simulations.
2. Although the characteristic ratios of the RG and CG chains are close to each other, their persistence lengths differ significantly, resulting from the presence of neighbor dependent correlations between torsion potentials of the CG chains which are absent in the RG chains.
3. The number of native and non-native contacts during folding is controlled by the initial conformations. The number of native contacts of the chains generated by the knowledge-based potentials is significantly larger than that of the random chains at the initial configurations and throughout the folding process.
4. There are specific contacts (e.g., residues 49 and 57) that appear throughout most of the trajectories.

5. Dominant bottlenecks during folding are due to local non-native contacts in the regions of the chain that join two native-like blocks that form but are not in the correct position relative to each other.
6. The neighbor dependent torsion potentials built into the CG chains result in strong correlations among bonds that stabilize fragments of secondary structure at the onset of folding.

REFERENCES AND NOTES

1. Dill, K. A.; Shortle, D. *Annu Rev Biochem* 1991, 60, 795.
2. Zagrovic, B.; Pande, V. S. *Nat Struct Biol* 2003, 10, 955.
3. Plaxco, K. W.; Gross, M. *Nat Struct Biol* 2001, 8, 659.
4. Shortle, D.; Ackerman, M. S. *Science* 2001, 293, 487.
5. Etchelles, S. A.; Hartl, F. U. *Nat Struct Biol* 2004, 11, 391.
6. Dill, K. A.; Chan, H. S. *Nat Struct Biol* 1997, 4, 10.
7. Fersht, A. R.; Daggett, V. *Cell* 2002, 108, 573.
8. van Gunsteren, W. F.; Burgi, R.; Peter, C.; Daura, X. *Angew Chem Int Ed Engl* 2001, 40, 351.
9. Keskin, O.; Yuret, D.; Gursoy, A.; Turkay, M.; Erman, B. *Proteins* 2004, 55, 992.
10. Avbelj, F.; Baldwin, R. L. *Proc Natl Acad Sci USA* 2004, 101, 10967.
11. Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F.; Sosnick, T. R. *J Mol Biol* 2003, 331, 693.
12. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res* 2000, 28, 235.
13. Jackson, S. E.; elMasry, N.; Fersht, A. R. *Biochemistry* 1993, 32, 11270.
14. Jackson, S. E.; Fersht, A. R. *Biochemistry* 1991, 30, 10428.
15. Jackson, S. E.; Fersht, A. R. *Biochemistry* 1991, 30, 10436.
16. Jackson, S. E.; Moracci, M.; elMasry, N.; Johnson, C.; Fersht, A. R. *Biochemistry* 1993, 32, 11259.
17. Kazmirski, S. L.; Wong, K. B.; Freund, S. M. V.; Tan, Y. J.; Fersht, A. R.; Daggett, V. *Proc Natl Acad Sci USA* 2001, 98, 4349.
18. Pan, Y. P.; Daggett, V. *Biochemistry* 2001, 40, 2723.
19. Otzen, D. E.; Fersht, A. R. *Biochemistry* 1998, 37, 8139.
20. Matouscheck, A.; Kellis, J. T.; Serrano, L.; Fersht, A. R. *Nature* 1989, 340, 122.
21. Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J Mol Biol* 2000, 298, 937.
22. Daggett, V.; Li, A.; Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R. *J Mol Biol* 1996, 257, 430.
23. Klimov, D. K.; Thirumalai, D. *J Mol Biol* 1998, 282, 471.

24. Lazaridis, T.; Karplus, M. *Science* 1997, 278, 1928.
25. Li, L.; Mirny, L. A.; Shakhnovich, E. I. *Nat Struct Biol* 2000, 7, 336.
26. Nymeyer, N.; Socci, N. D.; Onuchic, J. N. *Proc Natl Acad Sci USA* 2000, 97, 634.
27. Takada, S. *Proc Natl Acad Sci USA* 1999, 96, 11698.
28. Li, L.; Shakhnovich, E. I. *Proc Natl Acad Sci USA* 2001, 98, 13014.
29. Torshin, I. Y.; Harrison, R. W. *Proteins* 2001, 43, 353.
30. Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R. *J Mol Biol* 1995, 254, 260.
31. Li, A.; Daggett, V. *J Mol Biol* 1996, 257, 412.
32. Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* 2000, 39, 252.
33. Clementi, C.; Jennings, P. A.; Onuchic, J. N. *J Mol Biol* 2001, 311, 879.
34. Rowland, R. S.; Taylor, R. *J Phys Chem* 1996, 100, 7384.
35. Gerstein, M.; Richards, F. M. In *International Tables for Crystallography*; Rossmann, M.; Arnold, E., Eds.; Kluwer: Dordrecht, 2001; pp 531–539.
36. Flory, P. J. *Proc Natl Acad Sci USA* 1973, 70, 1819.
37. Go, N. *Macromolecules* 1976, 9, 535.
38. Hoang, T. X.; Cieplak, M. *J Chem Phys* 2000, 112, 6851.
39. Hoang, T. X.; Cieplak, M. *J Chem Phys* 2000, 113, 8319.
40. Sedgewick, R. *Algorithms*; Addison-Wesley: New York, 1988.
41. Bahar, I.; Kaplan, M.; Jernigan, R. L. *Proteins* 1997, 29, 292.
42. Keskin, O.; Bahar, I. *Fold Des* 1998, 3, 469.
43. Keskin, O.; Bahar, I.; Badretdinov, A. Y.; Ptitsyn, O. B.; Jernigan, R. L. *Protein Sci* 1998, 7, 2578.
44. Baldwin, R. L.; Zimm, B. H. *Proc Natl Acad Sci USA* 2000, 97, 12391.
45. Flory, P. J. *Statistical Mechanics of Chain Molecules*; Wiley: New York, 1969.
46. Mattice, W. L.; Suter, U. W. *Conformational Theory of Large Molecules*; Wiley Interscience: New York, 1994.
47. Yoon, D. Y.; Flory, P. J. *J Chem Phys* 1974, 61, 5366.
48. Brant, D. A.; Flory, P. J. *J Am Chem Soc* 1965, 87, 663.
49. Brant, D. A.; Flory, P. J. *J Am Chem Soc* 1965, 87, 1175.
50. Fitzkee, N. C.; Rose, G. D. *J Mol Biol* 2005, 353, 873.
51. Devore, J. L. *Probability and Statistics for Engineering and Sciences*; 6th ed., Duxbury Press: Pacific Grove, CA, 2004.
52. Dobson, C. M.; Sali, A.; Karplus, M. *Angew Chem Int Ed Engl* 1998, 37, 868.
53. Conrad, J. C.; Flory, P. J. *Macromolecules* 1976, 9, 41.
54. Erman, B. *Biophys J* 2001, 81, 3534.