JCTC Journal of Chemical Theory and Computation

Temperature-Dependent Probabilistic Roadmap Algorithm for Calculating Variationally Optimized Conformational Transition Pathways

Haijun Yang,[†] Hao Wu,[†] Dawei Li,[†] Li Han,*[,‡] and Shuanghong Huo*[,†]

Gustaf H. Carlson School of Chemistry and Biochemistry and Department of Mathematics and Computer Science, Clark University, 950 Main Street, Worcester, Massachusetts 01610

Received August 18, 2005

Abstract: In this paper we present a method to calculate a temperature-dependent optimized conformational transition pathways. This method is based on the maximization of the flux derived from the Smoluchowski equation and is implemented with a probabilistic roadmap algorithm. We have tested the algorithm on four systems—the Müller potential, the three-hole potential, alanine dipeptide, and the folding of β-hairpin. Comparison is made with existing algorithms designed for the calculation of protein conformational transition and folding pathways. The applications demonstrate the ability of the algorithm to isolate a temperature-dependent optimal reaction path with improved sampling and efficiency.

I. Introduction

Protein/peptide conformational transitions are closely related to their biological function. Here, the conformational transition is broadly defined, including the transition from an unfolded state to the folded state, from a partially unfolded intermediate state to the native state, or the disorder-to-order transition for intrinsically disordered proteins. It is of great challenge and significance to describe this long-time process accurately and effectively using computational approaches.[1] This process can be investigated using the "chain of states" methods to identify all intermediates and transition states between the two-end conformations concomitantly.[2−22] A number of these methods search for the steepest descent path or minimum-energy path. While the steepest descent path may make significant contributions to the reaction rate, it has been noticed that, in general, the observed average dynamics is not determined solely by the steepest descent path.[13,23] Therefore, including temperature effect in the calculation of transition pathway is desired. Methods have been developed to generate an ensemble of transition pathways.[24−27] In principle, the transition-path sampling (TPS) method can be applied to any system to provide detailed information about the transition at a given temperature. However, in practice, they are too computationally demanding for larger systems, e.g. a collection of a few hundred unfolding pathways of a 16-residue peptide took ∼3 months on an 8-node, 1-GHz Athlon PC cluster.[28] Furthermore, the quality of the path sampling relies on a proper definition of the stable states by order parameters. It may need an even longer time to fulfill the requirements of order parameters by trial and error.

Elber and Shalloway have developed a cost-effective method to include temperature effect based on a classical expansion of path integral for a stochastic process. However, in their method, the time for transition is a variable.[29] Huo and Straub have developed a time-independent algorithm to search for protein/peptide conformational transition pathways at a well-defined temperature.[30] This algorithm is based on the work of Berkowitz and co-workers who derived variational formulas for the optimal transition pathway connecting the reactant and the product.[3] Suppose that the conformational transition is a stochastic process, the probability distribution of the system is well described by the Smoluchowski equation.[31] Assuming that the friction (γ) of the system is isotropic and spatially independent, the optimal

* Corresponding author fax: (508) 793-8861; e-mail: shuo@clarku.edu (S.H.) and fax: (508) 421-3715; e-mail: lhan@clarku.edu (L.H.).

[†] Gustaf H. Carlson School of Chemistry and Biochemistry.

[‡] Department of Mathematics and Computer Science.

10.1021/ct0502054 CCC: $37.00
Published on Web 11/15/2006

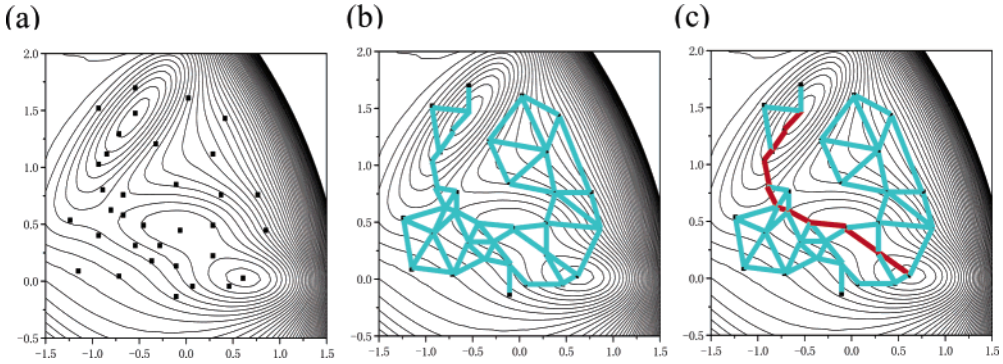

Figure 1. Schematic representation of the PRM approach applied to the Müller potential.[40,41] (a) Node generation. Each black dot represents a node. (b) Roadmap connection. All of the blue lines are edges connecting the nodes. (c) Roadmap query. The shortest path is denoted by the red line.

reaction pathway is defined as the one that maximizes the reactive flux (j) between the reactant ($\{\mathbf{r}_R\}$) and the product ($\{\mathbf{r}_P\}$), $Max\{j \propto [1/\gamma \int e^{\beta U(\mathbf{r})} dl(\mathbf{r})]\}$, or minimizes the mean first passage time (τ),[3,30] $Min\{\tau \propto \gamma \int e^{\beta U(\mathbf{r})} dl(\mathbf{r})\}$. The objective is to minimize the line integral

$$P = \int_{\Gamma_R}^{\mathbf{r}_P} e^{\beta U(\mathbf{r})} dl(\mathbf{r}) \tag{1}$$

In eq 1, $\beta = 1/kT$, where k is the Boltzmann constant, and T is the absolute temperature. $U(\mathbf{r})$ is the effective energy of the system for a given conformation $\{\mathbf{r}\}$, including the intramolecular interaction energy of the protein/peptide and the solvation free energy. Also in eq 1, $dl(\mathbf{r})$ is the line segment along the path. By minimizing the line integral of eq 1 using the self-penalty walk method,[7] one can obtain an optimal pathway corresponding to the fastest reaction rate.[30] This algorithm is called MaxFlux and has been successfully applied to study the conformational change of peptides.[32,33]

Based on the same idea of maximizing the flux, but using the differential equation derived by Berkowitz et al. instead of the integral equation,[3] Crehuet and Field[34] described an alternative MaxFlux by implementing the nudged-elastic-band method,[11] called MaxFlux-NEB. It has been demonstrated that MaxFlux-NEB works well for small molecules and model peptides.[34] However, both the original MaxFlux and MaxFlux-NEB rely on global minimization methods. They both start from an initial guess which is usually a linear interpolation between the reactant and the product. If the initial guess happens to be close to the optimal pathway, a local minimization method is good enough to find the optimal pathway.[30] However, when the initial guess is far away from the final path, computationally demanding global minimization methods are required, for example, simulated annealing or a molecular dynamics parallel tempering scheme.[35] Note that we are talking about searching for the global minimum in the path space, instead of the conformational space. The former is more time-consuming than the latter. One way to deal with the global minimization problem is to employ a stochastic sampling method proposed by Woolf and co-workers.[26] However, applications in biomolecular systems are still the major challenge for this method.

In this work, we present an alternative way to tackle the sampling issue. We propose to combine MaxFlux with the probabilistic roadmap (PRM) motion planning method that is originally developed for robotics motion planning and has been applied to study various problems.[36−38] We do not intend to give a complete review of PRM, rather, we mainly introduce Amato and co-workers's PRM application on protein folding landscape and kinetics[39] because their work is closely related to our present article. The details of the combined approach of MaxFlux and PRM are presented in the MaxFlux-PRM method section, and the applications are in Results and Discussion section.

II. The PRM Method

The basic idea of PRM is to extract the pathways from a roadmap. There are three stages of PRM as shown in Figure 1: (1) node generation; (2) roadmap connection; and (3) roadmap query. The objective of the node generation is to sufficiently sample the region of conformational space surrounding the final conformation or the product, e.g. the native state of protein. A conformation is called a node (q) in the roadmap. Biased sampling strategies in the (ϕ,ψ) space have been successfully applied for medium-size proteins (more than 100 amino acids).[38,39,42] A threshold can be set to remove the high energy nodes. The details of the techniques of node generation to ensure adequate coverage of the conformational space will be presented in the MaxFlux-PRM method section.

The second stage is to connect the generated nodes to obtain a roadmap or a graph. The objective of this stage is to construct a roadmap encoding the representative paths. For each generated node or conformation, its k nearest neighbors will be found. The neighbor can be defined using the root-mean-square (rms) distance or Euclidean distance. The value of k is adjustable. Previously, $k = 20$ was set.[38,39,42] The connection between a node (e.g. q_1) and its neighbor (e.g. q_2) is called an edge. A weight is assigned to each edge defined as

$$\omega(q_1, q_2) = -\ln(P_{\text{connection}}) \tag{2}$$

where

$$P_{\text{connection}} = \begin{cases} e^{-\Delta E/k_B T} & \text{if } \Delta E > 0,\ \Delta E = E(q_2) - E(q_1) \\ 1 & \text{if } \Delta E \le 0 \end{cases} \tag{3}$$

By connecting all the nodes, the roadmap is constructed and

appears like a net lying on the energy surface (Figure 1(b)). However, note that in eq 3, there is no difference between downhill movement and moving on a flat surface because when either $\Delta E < 0$ or $\Delta E = 0$, $P_{\text{connection}} = 1$. Moreover, the magnitude of downhill movement is not taken into account, for example, $\Delta E = -0.0001$ kcal/mol has the same value of $P_{\text{connection}}$ as $\Delta E = -1000$ kcal/mol.

The final stage is to query the roadmap to extract the optimal pathway from a given initial conformation to the final destination, such as the native state. Dijkstra's algorithm[43] is employed to find the smallest weight path between the initial and final conformation (Figure 1(c)). The optimal path from q_1 to q_n through $n-2$ nodes satisfies

$$w(q_1,q_2,.....q_n) = \min \{\sum_{i=1}^{n-1} - \ln(p_{i\text{connection}})\} \qquad (4)$$

where n is not fixed and may vary from path to path. If we plug eq 3 into eq 4, we can get an equivalent criterion of the optimal path as

$$w(q_1,q_2,.....q_n) = \min \left\{\frac{1}{kT}\sum_{\text{uphill}} \Delta E_i\right\} \quad \text{or}$$
$$w(q_1,q_2,.....q_n) = \min \{\sum_{\text{uphill}} \Delta E_i\} \qquad (5)$$

The querying process searches for the **least uphill** movement. As a result, this PRM method misses the intermediate states if any. In other words, it is more suitable to two-state transitions than multistate transitions.This shortcoming can be illustrated on model potentials as shown in the section of results. Furthermore, since $1/kT$ can be moved outside the sum in eq 5, the pathway does not change with temperature. As a result, the pathway found by PRM is not temperature dependent. However, the strength of PRM is its enhanced sampling and efficiency. It can map the potential energy landscape and generate sets of representative transition pathways from a single roadmap. The system can avoid being trapped in a local minimum, and no detailed simulations are needed plus it is initial-guess free. The PRM method has been applied to study protein folding pathways of 14 proteins up to the size of 110 amino acids.[39] The order of secondary structure formation along the PRM pathway has been found to be in good agreement with the experimental results.[39,42] Note that all of the proteins that have been studied by PRM have two-state folding behavior. PRM and MaxFlux naturally complement each other. By replacing the edge weight function of PRM in eqs 2−5 with the MaxFlux equation (eq 1), we can search for temperature-dependent transition pathways with enhanced sampling and the capability to study the multistate transition pathways. This idea falls in the same category as a recent effort in network models for kinetics,[44−46] while the differences are in node generation and the definition of edge weight. In MaxFlux-PRM, nodes are not generated by molecular dynamics or replica exchange as other network models.

III. The MaxFlux-PRM Methods

We have tested the MaxFlux-PRM method on two model potentials, Müller and 3-hole, and two peptides, alanine dipeptide (AD) and β-hairpin. For the Müller potential and 3-hole potential, we uniformly searched the potential energy surface to generate 5000 and 20 000 nodes. All of the generated nodes were accepted without setting the energy threshold. Five intermediate conformations were added between two nodes by linear interpolation. The edge weight between nodes, e.g. q_0 and q_1, can be defined as

$$w_i(c_0 = q_0, c_1, ..., c_m, c_{m+1} = q_1) = \sum_{i=0}^{m} e^{\beta U(c_i)} |r(c_{i+1}) - r(c_i)| \qquad (6)$$

where $m = 5$ and $|r(c_{i+1}) - r(c_i)|$ is the rms distance between conformations c_{i+1} and c_i. $\beta = 1/kT$, where k is the Boltzmann constant, and T is the absolute temperature. U is the energy of the system. For these model potentials, 5000 nodes can be considered a complete search; therefore, interpolation between nodes did not significantly improve the accuracy of the path but it made the path smoother.

For AD and β-hairpin, we describe the protein intramolecular interaction energy using CHARMM 19 polar hydrogen energy function[47] and an effective energy function (EEF1) for solvation.[48] We search for the path in the full $3n$-dimension Cartesian coordinate. For AD, a node was generated by assigning each backbone dihedral angle a random value in its allowable range ($[-\pi, \pi]$ in this simulation). Once the dihedral angles were generated, the conformation was minimized 100 steps using the adopted basis Newton−Raphson method with restraints on the backbone dihedral angles. Suppose that the thermal fluctuation of each dihedral angle is 10°, a search with more than 36×36 nodes can be considered complete for AD. We randomly generated 5000 nodes. Twenty nearest neighbors were found for each node. No linear interpolation between neighbors was applied.

Due to the high dimensionality of the protein conformation space, we adopt a focused sampling strategy[42] for β-hairpin based on the fact that the reactant and product conformations are known. The MaxFlux-PRM procedure was illustrated in the flowchart (Figure 2). In particular, we generate a set of Gaussian distributions around the backbone dihedral angles of the reactant and product with a set of standard deviations (STDs) of {3°, 5°, 10°, 15°, 20°, 25°, 30°, 40°, 50°, 60°, 80°, 100°, 130°, 160°}. The small STDs capture the detail around the reactant and product, while the larger STDs ensure adequate roadmap coverage of the conformation space. By random sampling from these distributions, we initially generated 200 000 nodes. The side chains were built using CHARMM and minimized (3000 steps or $\Delta E < 0.005$ kcal/mol whatever comes first) with the backbone dihedral angles restrained. This procedure helps to remove some bad contacts. To sample the side-chain conformations, 10^4 steps of MC were performed for each node by moving the side chain only using the Monte Carlo (MC) module in CHARMM.[49] By eliminating the nodes with a threshold, $E_{\max} = -400$ kcal/mol, we obtained 93 886 nodes, where $E_{\max}$ is adjustable and varies with the size of protein. The number of nodes is almost four times of that reported recently for the same system and the same force-field.[50] We used rms

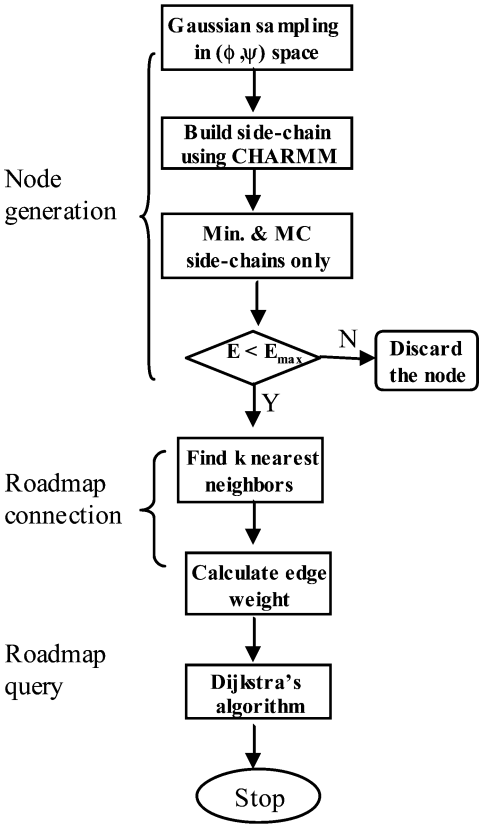

Figure 2. Flowchart for MaxFlux-PRM (β-hairpin).

deviation as the criterion to find 40 nearest neighbors for each node with the prerequisite that the node has no more than one backbone hydrogen bond difference with its neighbor. The nodes that do not satisfy the neighboring requirements were removed. In fact, this kind of node is rare. No interpolation between nodes was carried out.

To introduce the temperature effect, we define the edge weight ($\omega(q_0,q_1)$) as

$$w(q_0,q_1) = e^{\beta U(q_0)}|r(q_1) - r(q_0)| \tag{7}$$

where $|r(q_1) - r(q_0)|$ is the all-atom rms distance between conformation q_0 and its neighbor q_1 in the Cartesian coordinate. $\beta = 1/k_BT$, where k_B is the Boltzmann constant and T is equal to 300 K for AD and β hairpin. U is the effective energy of the system (protein plus solvent) including the intramolecular interaction energy of the protein and the solvation free energy. Once all the neighbors were identified and all the edge weights were calculated, the roadmap was constructed. To find the minimum weight path between the extended state and the native fold, that corresponds to the maximum flux path, Dijkstra's algorithm was used to query the roadmap. The minimum weight path between the initial and final states is defined in the discretized form of eq 1 as

$$\omega_{\min} = \min\{\sum_{i=0}^{n-1} e^{\beta U(q_i)}|\mathrm{r}(q_{i+1}) - \mathrm{r}(q_i)|\} \tag{8}$$

where the path connecting the initial and final states goes through $n-1$ nodes. When the samping is enough, either $U(q_i)$ or $U(q_{i+1})$ can be used in eq 8. Otherwise, the midpoint energy should be used. The weight of the path is the sum of the weights of its constituent edges. This is a well-defined single-pair shortest-path problem when all edges have nonnegative weight. As mentioned in the Introduction, the minimum weight path defined in eq 8 corresponds to the shortest mean first passage time or the fastest reaction rate.

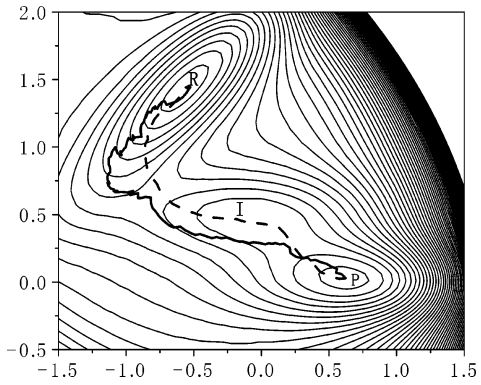

Figure 3. The transition pathway found for the Müller potential. The solid line is the PRM pathway, while the dashed line is the MaxFlux-PRM pathway. The pathways with 20 000 nodes are presented.

IV. Results and Discussion

Müller Potential. The MaxFlux-PRM method was applied on the two-dimensional model potential of Fukui, Kato, and Fujimoto[40] studied by Müller.[41] There are three minima in the Müller potential: the reactant, the intermediate (a shallow metastable minimum), and the product. As shown in Figure 3, the PRM method (solid line) misses the metastable intermediate state. If a pathway goes through an intermediate state, the system has to move downhill to visit the minimum and then move uphill to escape from the minimum. According to eq 5, in the PRM method, only the uphill transition contributes to the edge weight, while the downhill transition is not taken into account. As a result, the downhill-to-uphill edge weight to visit the intermediate state is larger than that obtained by avoiding the intermediate state and directly going down to the product well. When we replaced the PRM edge weight function (eq 5) with the MaxFlux criterion (eq 1) to search for the pathway with maximum reactive flux or minimum mean first passage time, we identified a new path that visited the intermediate state (dashed line). The MaxFlux-PRM path is an estimate of the best path for the reaction at nonzero temperature $1/\beta$ (β=0.06). This path is comparable to what would be defined by a minimum-energy path.[51]

3-Hole Potential. A more challenging test is the 3-hole potential developed by Huo and Straub[30] which has been used by other groups.[29,34] It is designed to isolate the globally optimal transition pathway. As shown in Figure 4, besides the reactant (lower left well) and product (lower right well) minima there is a third energy minimum (above center). There are two possible reaction pathways. The lower path moves roughly left-to-right between the reactant and product wells crossing a single high energy barrier. The upper path overcomes a low energy barrier, through a basin, and then overcomes another low energy barrier before reaching the product state. The optimal minimum-energy path is the lower path that has a high barrier but is shorter.[30] However, the

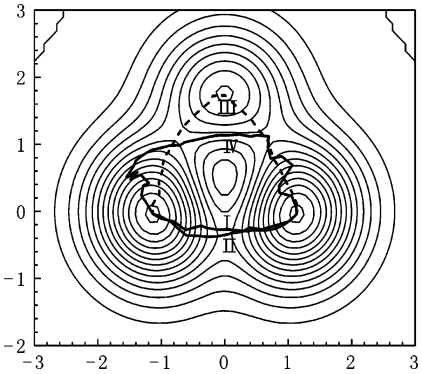

Figure 4. The transition pathway found for the 3-hole potential. MaxFlux-PRM has identified the lower reaction pathways (I and II) at high temperature ($\beta = 1$ or 2), while the upper pathway(III) is dominant at low temperature ($\beta = 3.3$) using MaxFlux-PRM. In contrast, PRM found either pathway IV or pathway II at whatever temperature. Pathways IV and II are identical based on the PRM criterion in eq 5. We found that 5000 nodes can yield a smooth path.

lower path is not always the path of maximum transition rate at nonzero temperature. According to the definition of the optimal path with maximum flux in eq 1, the lower reaction path will be the pathway of maximum flux only at high temperatures. At low enough temperatures the upper pathway, which crosses two lower energy saddle points, is dominant.[30]

Independent runs of PRM with different seeds of random number generator have identified either the pathway IV or the lower path II (Figure 4) at any of the tested temperatures because these paths are identical according to eq 5 and the method is temperature-independent. Interestingly, pathway IV is similar to the "last" stable path in the upper region identified by MaxFlux-NEB at high temperature.[34] Once again, this test has illustrated that the PRM method is more suitable to two-state transitions than multistate transitions. It fails to identify the intermediate state due to the flaw of the edge weight function (eqs 2−5). In contrast, the MaxFlux-PRM method has identified different pathways at different temperatures (Figure 4), consistent with the original MaxFlux results[30] and MaxFlux-NEB.[34]

Alanine Dipeptide. We have applied the MaxFlux-PRM method to alanine dipeptide (AD). The potential energy surface and the pathways of conformational transition of this system have been studied extensively.[7,30,34,52] Although the molecule is small, it has emerged as a standard test for reaction path algorithms. The reactant conformation is $C7_{eq}$ at ($\phi = -86°$, $\psi = 79°$) with effective energy equal to −28.74 kcal/mol, and the product conformation is $C7_{ax}$ ($\phi = 76°$, $\psi = -55°$) whose effective energy is −30.17 kcal/mol. For the pathway obtained at 300 K, $E_{barrier} - E_{reactant}$ is equal to 4.98 kcal/mol and $E_{barrier} - E_{product}$ is 6.40 kcal/mol, where $E_{barrier}$ is the maximum effective energy of the conformation along the optimal path. The effective energy barrier is ca. 10 times of RT at room temperature; therefore, the conformation transition pathway at 300 K is similar to the minimum-energy path.

β-Hairpin Formation. Investigating the folding of key secondary structural elements is proving to be useful for understanding the thermodynamics and kinetics of large protein folding. Therefore, the folding mechanism of β-hairpin has been studied extensively by experimental[53−64] and computational approaches.[45,46,50,65−84] Herein, we demonstrate the MaxFlux-PRM method can be applied to β-hairpin to locate optimal folding pathways from the extended state to the folded state (pdb entry: 3GB1,[85] residues 41−56). The folding path was analyzed using the all-atom rms deviation from the native state, radius of gyration (R_g) for the hydrophobic core (W43, Y45, F52, and V54 in Figure 6), and the number of native contacts (N_{NC}) as shown in Figure

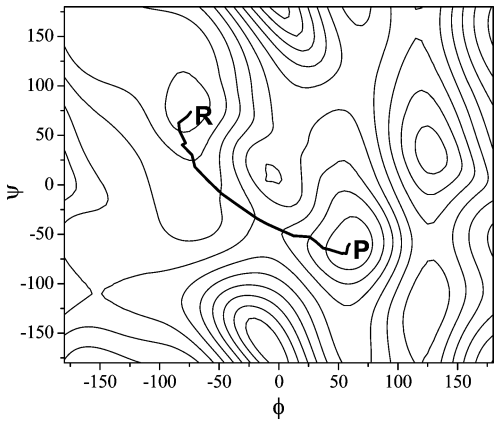

Figure 5. The reaction path of alanine dipeptide on the (ϕ,φ) potential energy map. The solid line connecting $C7_{eq}$($\phi = -86°$, $\varphi = 79°$) and $C7_{ax}$($\phi = 76°$, $\varphi = -55°$) is the transition pathway identified by the MaxFlux-PRM method.

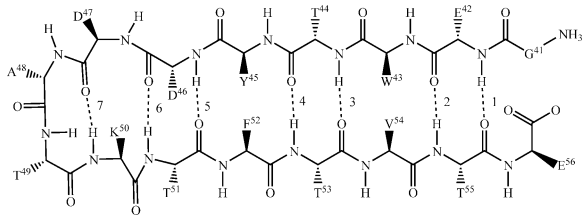

Figure 6. Schematic representation of the second β-hairpin of the B1 domain of streptococcal protein G (GB1). The main-chain hydrogen bonds are numbered from tail to the turn region.

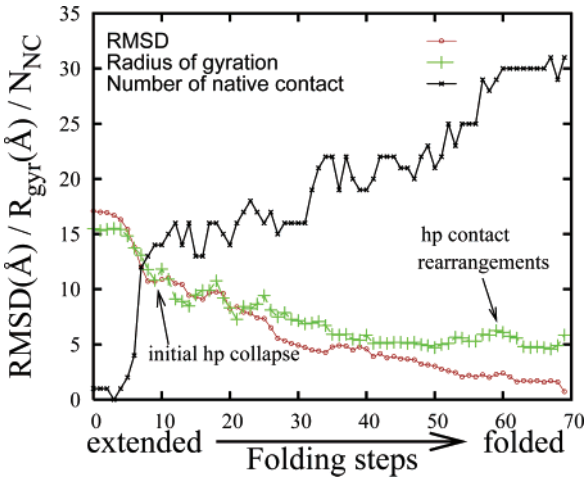

Figure 7. RMSD from the native state, radius of gyration for the hydrophobic core, and number of native contacts are computed along the path.

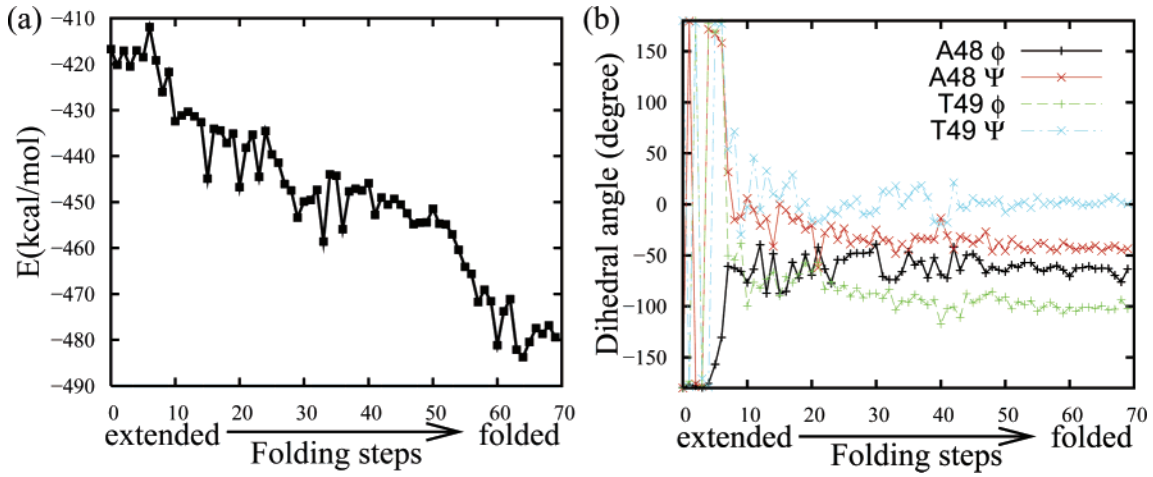

Figure 8. (a) Effective energy (molecular mechanics potential plus solvation energy) along the folding path. (b) Backbone dihedral angles of A48 and T49 in the turn region calculated along the folding path.

7. Two residues are considered to contact with each other if any heavy atom (C, N, O, S) of one residue is within 4.5 Å from any heavy atom of the other residue (the nearest neighbor in sequence is not included).[86] As shown in Figure 7, the very first step in the folding of β-hairpin from the extended state is hydrophobic (hp) collapse, as reflected in the initial drops in R_{gyr} for the hydrophobic core from 15.5 to 10.8 Å (at folding step #10). This initial hydrophobic collapse has been observed by a number of simulations[68,71,77,79] and experiments[63] on β-hairpin. Concomitantly, N_{NC} jumps to 14 among which almost all are the contacts between residues i and $i+2$. The side chains including hydrophobic contacts rearrange in the later stage to form the native packing (Figure 7, a local maximum in R_{gyr} at folding step #60). The effective energy spans from ca. −410 kcal/mol to ca. −480 kcal/mol along the folding path (Figure 8a). It is obvious that the native state (at folding step 70) is not the global minimum because there are two minima along the path which have even lower effective energy (Figure 8a). It is possible that multiple conformations coexist because experimentally estimated population of the β hairpin is only ca. 30%,[54,60,87] and the presence of multiple conformers with effective energy lower than native state energy was also reported by other computational work.[22]

To describe the turn formation, we used the dihedral angles of the central residues (($\phi_{A48} = -60°$, $\psi_{A48} = -30°$) and ($\phi_{T49} = -90°$, $\psi_{T49} = 0°$)).[88] If the four dihedral angles fall within ±30° of these values, we consider that the turn is formed. According to this criterion, the turn is first formed at the folding step #13 (Figure 8b), that is after the initial hydrophobic collapse (at folding step #10 (Figure 7)). The highest energy (at folding step #7, Figure 8a) along the folding pathway corresponds to the turn initiation when the dihedral angles (($\phi_{A48} = -61°$, $\psi_{A48} = 32°$) and ($\phi_{T49} = -50$, and $\psi_{T49} = 54$)) of the central residues started the transition from the extended state region to the β turn region, consistent with the experimental results which have suggested that the rate-limiting step is the turn formation.[61,63,64,89]

To investigate the role of the native hydrogen bonds and the hydrophobic cluster in the folding process, we present the native hydrogen bond formation and the accessible

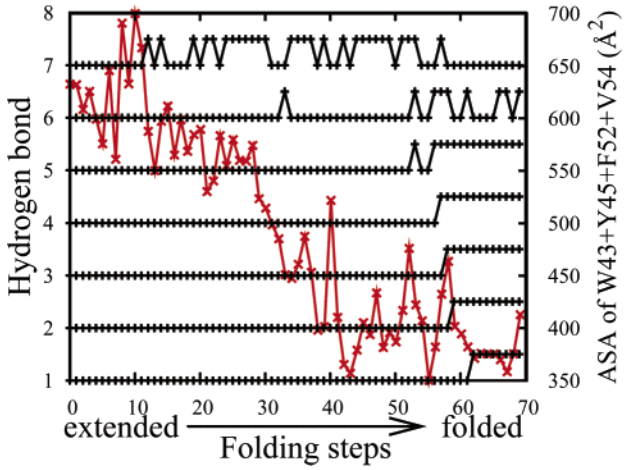

Figure 9. Interstrand main-chain hydrogen bond formation along the folding path. The hydrogen bonds are numbered in the same way as Figure 6. Hydrogen bond #7 was first formed at folding step 13. The accessible surface area (ASA) is denoted by the red line. A 1.4 Å probe was used.

surface area (ASA) of the hydrophobic side chains (W43, Y45, F52, and V54) along the folding path in Figure 9. We used the criteria of hydrogen bonds defined by Dinner et al.,[68] the distance between the corresponding heavy atoms is 3.4 Å and the out-of-line angle is less than 70°, who used the same force field on the same system. Even though the hydrogen bond in the turn region formed first, this hydrogen bond is not stable and even not present in the minimized structure because the distance between K50(N) and D47(O) is 3.49 Å. The hydrogen bond next to K50(N)−D47(O) is not stable either (Figure 9). The remaining hydrogen bonds are formed in the late stage of folding (after folding step #55). This kind stability is consistent with the analysis on the temperature-dependent hydrogen bond stability.[68,74,83] However, as demonstrated by Zhang et al.[84] in their simulations of trpzip2 the hydrogen bond with the highest stability is not necessarily important in the folding process. The hydrogen bond in the turn region with the lowest stability forms first in the transition state in their simulations,[84] which is in line with our results. After the turn formation, yet before the hydrogen bonds to continue to propagate, the ASA drops

Table 1. Computational Cost (β-Hairpin)

process	node generation minimization/MC[a] (h)	roadmap connection[b] (h)	roadmap query (min)
running time	18/82	74	10

[a] There are 200 000 nodes were generated. The minimization and MC were for the side chain only. [b] Ca. 3.6×10^6 edges were connected, and their edge weights were calculated.

to 456 Å^2 (close to the native state value 422 Å^2) at folding step #35 and remains below this value for the majority of the conformers along the rest of the path, strongly suggesting that the hydrophobic assembly is formed before the propagation of the hydrogen bonds from the turn to the tail. Figure 9 is in generally good agreement with Figure 3c of ref 68.

In general, our mechanism is a mixture of zipper model[54,65−67] and hydrophobic core-centric model.[68,69,71,79] The sequence of events along the folding pathway is as follows: (1) initial hydrophobic collapse; (2) formation of the β turn; (3) occurrence of hydrophobic cluster; and (4) repacking of the hydrophobic core accompanied by the propagation of the hydrogen bonds beyond the turn region (in the late stage of folding). Our mechanism is supported by the fact that the mutation in the turn region affects the folding rate, while the stronger hydrophobic cluster only decreases the unfolding rate,[64] suggesting that the turn forms in the transition state while the hydrophobic cluster formed after the transition state.

Advantages and Limitations of MaxFlux-PRM. To better assess the computational cost, we list the running time of each step of path searching in Table 1. All of the calculations are on a single-CPU of AMD Opteron (1.4 GHz). The two most time-consuming tasks are the MC in node generation and the pairwise rms deviation calculation during roadmap connection, as seen in Table 1. This computational cost of pairwise rms deviation, which scales up as N^2 (N = number of nodes), can be reduced by the two-step divide-and-conquer approach employed by Brooks and co-workers.[22] To save running time, for node generation, minimization, MC, and roadmap connection, the job can be readily partitioned into multiple CPUs.

As demonstrated on the Müller potential and 3-hole potential, the edge weight of MaxFlux-PRM is more physical than the original PRM which must miss the intermediate state due to the flaw in the edge weight definition. The advantage of MaxFlux-PRM over the original MaxFlux[30] is the enhanced sampling and efficiency. Huo and Straub started from the initial guess which is a straight line connecting the reactant and the product on the 3-hole potential. To identify the upper path at low temperature, they applied simulated annealing to minimize the path, which took hours on an SGI supercomputer. For the MaxFlux-PRM method, it only took a few minutes on a desktop PC. Furthermore, the original MaxFlux method requires several restraints on the path that are implemented by setting several parameters, for example, (1) a pseudobond restraint between nearest neighbor intermediate structures to encourage the mean-square distances between adjacent conformations along the path to be approximately equal and (2) a repulsion interaction between any non-nearest neighbor intermediate structures along the path. These restraints are implemented through three adjustable parameters. Additionally, one more parameter is needed to control the cooling rate for the simulated annealing. Even though the authors provided some rules of thumb to set the parameters,[30] it takes a huge amount of human time to adjust these parameters for each system. On the contrary, there are only a few adjustable parameters used in MaxFlux-PRM: n (the number of nodes), k (the number of nearest neighbors of each node), and E_{max} which vary with the size of the system but must be larger than the effective energy of the extended state.

For other available temperature-dependent reaction path algorithms such as the transition path sampling (TPS) method[24,25] and its analogue, discrete path sampling method,[50] the quality of the paths relies on a proper definition of the stable states by order parameters. To choose proper order parameters, one needs to carry out numerous trail simulations before production runs. In addition, TPS can tackle only one free-energy barrier at a time. Consequently, one needs to define a number of (meta)stable states. One limitation of the MaxFlux-PRM method is that it employs implicit solvation. The reason we used an implicit solvation model is that in Berkowitz's description of reactive flux,[3] which is based on Smoluchowski equation of stochastic processes, $U(\mathbf{r})$ is the potential of mean force of the system. The potential of mean force for a given conformation of a solvated macromolecule is the free energy of the system consisting of the macromolecule and the solvent with an average over all solvent degrees of freedom at a given temperature. The solvation free energy ($\Delta G_{solvation}$) can be calculated with the implicit solvation model, while the explicit water model is not practical to give $\Delta G_{solvation}$; nevertheless, the water expulsion in protein folding cannot be addressed with such implicit solvation. To compensate for this, one can use MaxFlux-PRM to get the initial path to define the (meta)stable states and employ TPS with explicit water to search for the ensemble of pathways. As a result, the definition of the (meta)stable states will be more reasonable than that obtained from an unfolding trajectory[28] because the (meta)stable states sampled by the unfolding simulation are not necessarily on the folding pathway.

In summary, as demonstrated in the applications of these four systems, we have successfully identified the temperature-dependent transition pathways at atomic level with improved sampling and efficiency by a new method, MaxFlux-PRM. The system can avoid being trapped in a local minimum, and no detailed simulations are needed plus initial-guess free**.** This method can be employed to study conformational transition of biomolecular systems not involving bond breaking or formation. The node-generation and roadmap connection are naturally parallel computing processes. The computational efficiency is expected to be dramatically increased by employing parallel processing techniques.[90]

Acknowledgment. This work was supported by NIH (1R15 AG025023-01 to S.H.). We are also thankful for support from the National Science Foundation Major Research Instrumentation Fund (DBI-0320875). S. Huo is very

grateful to Prof. Aaron Dinner for his help with the MC module in CHARMM.

References

(1) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151−6.

(2) Berkov, D. V. *J. Magn. Magn. Mater.* **1998**, *186*, 199−213.

(3) Berkowitz, M.; Morgan, J. D.; McCammon, J. A.; Northrup, S. H. *J. Chem. Phys.* **1983**, *79*, 5563−5565.

(4) Cho, A. E.; Doll, J. D.; Freeman, D. L. *Chem. Phys. Lett.* **1994**, *229*, 218−224.

(5) Choi, C.; Elber, R. *J. Chem. Phys.* **1991**, *94*, 751−760.

(6) Chu, J.-W.; Trout, B. L.; Brooks, B. R. *J. Chem. Phys.* **2003**, *119*, 12708−12717.

(7) Czerminski, R.; Elber, R. *Int. J. Quantum Chem.* **1990**, *Suppl. 24*, 167−186.

(8) Elber, R.; Karplus, M. *Chem. Phys. Lett.* **1987**, *139*, 375−380.

(9) Elber, R.; Meller, J.; Olender, R. *J. Phys. Chem. B.* **1999**, *103*, 899−911.

(10) Gillilan, R. E.; Wilson, K. R. *J. Chem. Phys.* **1992**, *97*, 1757−1772.

(11) Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998.

(12) Olender, R.; Elber, R. *J. Chem. Phys.* **1996**, *105*, 9299−9315.

(13) Olender, R.; Elber, R. *J. Mol. Struct.: THEOCHEM* **1997**, *398*−*399*, 63−71.

(14) Passerone, D.; Ceccarelli, M.; Parrinello, M. *J. Chem. Phys.* **2003**, *118*, 2025−2032.

(15) Passerone, D.; Parrinello, M. *Phys. Rev. Lett.* **2001**, *87*, 108302/1−108302/4.

(16) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2004**, *120*, 7877−7886.

(17) Smart, O. S. *Chem. Phys. Lett.* **1994**, *222*, 503−512.

(18) Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *120*, 2082−2094.

(19) Ulitsky, A.; Elber, R. *J. Chem. Phys.* **1990**, *92*, 1510−1511.

(20) Woodcock, H. L.; Hodoscek, M.; Sherwood, P.; Lee, Y. S.; Schaefer, H. F. I.; Brooks, B. R. *Theor. Chem. Acc.* **2003**, *109*, 140−148.

(21) Zaloj, V.; Elber, R. *Comput. Phys. Commun.* **2000**, *128*, 118−127.

(22) Khavrutskii, I. V.; Byrd, R. H.; Brooks, C. L. *J. Chem. Phys.* **2006**, *124*, 194903/1−194903/14.

(23) Northrup, S. H.; McCammon, J. A. *J. Chem. Phys.* **1983**, *78*, 987−989.

(24) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.

(25) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964−1977.

(26) Zuckerman, D. M.; Woolf, T. B. *J. Chem. Phys.* **1999**, *111*, 9475−9484.

(27) Zuckerman, D. M.; Woolf, T. B. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2000**, *111*, 016702−016711.

(28) Bolhuis, P. G. *Biophys. J.* **2005**, *88*, 50−61.

(29) Elber, R.; Shalloway, D. *J. Chem. Phys.* **2000**, *112*, 5539−5545.

(30) Huo, S.; Straub, J. E. *J. Chem. Phys.* **1997**, *107*, 5000−5006.

(31) Gardiner, C. W. *Handbook of stochastic methods for physics, chemistry, and the natural sciences*; Springer-Verlag: Berlin, New York, 2001.

(32) Huo, S.; Straub, J. E. *Proteins* **1999**, *36*, 249−61.

(33) Straub, J. E.; Guevara, J.; Huo, S.; Lee, J. P. *Acc. Chem. Res.* **2002**, *35*, 473−81.

(34) Crehuet, R.; Field, M. J. *J. Chem. Phys.* **2003**, *118*, 9563−9571.

(35) Brumer, Y.; Golosov, A. A.; Chen, Z. D.; Reichman, D. R. *J. Chem. Phys.* **2002**, *116*, 8376−8383.

(36) Kavraki, L.; Svestka, P.; Latombe, J. C.; Overmars, M. *IEEE Trans. Robot. Automat.* **1996**, *12*, 566−580.

(37) Apaydin, M. S.; Brutlag, D. L.; Guestrin, C.; Hsu, D.; Latombe, J. C.; Varma, C. *J. Comput. Biol.* **2003**, *10*, 257−81.

(38) Song, G.; Amato, N. M. *IEEE Trans. Robot. Automat.* **2004**, *20*, 60−71.

(39) Amato, N. M.; Dill, K. A.; Song, G. *J. Comput. Biol.* **2003**, *10*, 239−55.

(40) Fukui, K.; Kato, S.; Fujimoto, H. *J. Am. Chem. Soc.* **1975**, *97*, 1−7.

(41) Muller, K.; Brown, L. D. *Theor. Chim. Acta* **1979**, *53*, 75−93.

(42) Amato, N. M.; Song, G. *J. Comput. Biol.* **2002**, *9*, 149−68.

(43) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; MIT Press: Cambridge, MA, 1992.

(44) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299−306.

(45) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766−70.

(46) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801−6.

(47) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.

(48) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133−52.

(49) Hu, J.; Ma, A.; Dinner, A. R. *J. Comput. Chem.* **2006**, *27*, 203−16.

(50) Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *121*, 1080−90.

(51) Fischer, S.; Karplus, M. *Chem. Phys. Lett.* **1992**, *194*, 252−261.

(52) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877−82.

(53) Kobayashi, N.; Endo, S.; Munekata, E. In *Conformational study on the IgG binding domain of protein G*, *Proc. Jpn. Symp, 1992*; Yanaihara, N., Ed.; ESCOM: Leiden, The Netherlands, 1992; pp 278−80.

(54) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196−9.

(55) Blanco, F.; Ramirez-Alvarado, M.; Serrano, L. *Curr. Opin. Struct. Biol.* **1998**, *8*, 107−11.

(56) Kobayashi, N.; Honda, S.; Yoshii, H.; Munekata, E. *Biochemistry* **2000**, *39*, 6564−71.

(57) Honda, S.; Kobayashi, N.; Munekata, E. *J. Mol. Biol.* **2000**, *295*, 269−78.

(58) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578−83.

(59) Espinosa, J. F.; Munoz, V.; Gellman, S. H. *J. Mol. Biol.* **2001**, *306*, 397−402.

(60) Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238−43.

(61) Dyer, R. B.; Maness, S. J.; Peterson, E. S.; Franzen, S.; Fesinmeyer, R. M.; Andersen, N. H. *Biochemistry* **2004**, *43*, 11560−6.

(62) Yang, W. Y.; Gruebele, M. *J. Am. Chem. Soc.* **2004**, *126*, 7758−9.

(63) Du, D.; Tucker, M. J.; Gai, F. *Biochemistry* **2006**, *45*, 2668−78.

(64) Du, D.; Zhu, Y.; Huang, C. Y.; Gai, F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15915−20.

(65) Munoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5872−9.

(66) Kolinski, A.; Ilkowski, B.; Skolnick, J. *Biophys. J.* **1999**, *77*, 2942−52.

(67) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544−9.

(68) Dinner, A. R.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068−73.

(69) Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9062−7.

(70) Ma, B.; Nussinov, R. *J. Mol. Biol.* **2000**, *296*, 1091−104.

(71) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345−54.

(72) Paschek, D.; Garcia, A. E. *Phys. Rev. Lett.* **2004**, *93*, 238105.

(73) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151−69.

(74) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931−6.

(75) Wu, X.; Wang, S.; Brooks, B. R. *J. Am. Chem. Soc.* **2002**, *124*, 5282−3.

(76) Tsai, J.; Levitt, M. *Biophys. Chem.* **2002**, *101−102*, 187−201.

(77) Zhou, Y.; Linhananta, A. *Proteins* **2002**, *47*, 154−62.

(78) Zhou, Y.; Zhang, C.; Stell, G.; Wang, J. *J. Am. Chem. Soc.* **2003**, *125*, 6300−5.

(79) Bolhuis, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12129−34.

(80) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *J. Phys. Chem. B* **2004**, *108*, 6582−6594.

(81) Wei, G.; Mousseau, N.; Derreumaux, P. *Proteins* **2004**, *56*, 464−74.

(82) Snow, C. D.; Qiu, L.; Du, D.; Gai, F.; Hagen, S. J.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4077−82.

(83) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C. K.; Li, M. S. *Proteins* **2005**, *61*, 795−808.

(84) Zhang, J.; Qin, M.; Wang, W. *Proteins* **2006**, *62*, 672−85.

(85) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.

(86) Daggett, V.; Levitt, M. *J. Mol. Biol.* **1993**, *232*, 600−619.

(87) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584−90.

(88) Creighton, T. E. *Proteins Structures and Molecular Properties*, 2nd ed.; W. H. Freeman and Company: New York, 1993; p 507.

(89) McCallister, E. L.; Alm, E.; Baker, D. *Nat. Struct. Biol.* **2000**, *7*, 669−673.

(90) Amato, N. M.; Dale, L. K. *Proceedings of the 1999 IEEE International Conference on Robotics and Automation (ICRA '99)*; 1999; pp 688−694.

CT0502054