Extracting Physically Intuitive Reaction Coordinates from Transition Networks of a β-Sheet Miniprotein

Bo Qi,[†,‡,§] Stefanie Muff,[⊥] Amedeo Caflisch,*[,⊥] and Aaron R. Dinner*[,†,‡,§]

Department of Chemistry, James Franck Institute, and Insitutue for Biophysical Dynamics, The University of Chicago, Chicago, Illinois 60637, and Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Received: February 17, 2010; Revised Manuscript Received: April 13, 2010

Simulations are important for understanding complex reactions, but their interpretation is challenging owing to the large number of degrees of freedom typically involved. To address this issue, various means for relating the dynamics of a stochastic system to its structural and energetic features have been introduced. Here, we show how two leading approaches can be combined to advantage. We use the network of transitions observed in a reversible folding/unfolding simulation of a 20-residue three-stranded antiparallel β-sheet peptide (beta3s) to estimate the probabilities of committing to stable states (the native state and major nonnative states), and these then serve as the basis for an efficient statistical procedure for identifying physical variables that describe the dynamics. We find that a single coordinate that jointly characterizes the formation of the two native turns of beta3s can adequately describe the overall folding process, despite its complex nature. Additional features associated with major pathways leading from individual nonnative states are resolved; indeed, a key result is an improved understanding of the unfolded state. Connections to other methods for analyzing complex reactions are discussed.

1. Introduction

Reactions in the condensed phase, especially biomolecular ones, are typically complex, proceeding by multiple pathways consisting of many elementary steps. What makes characterizing these reactions challenging is that many degrees of freedom participate, and there is no obvious difference between relevant and irrelevant motions (i.e., ones that do or do not determine whether the reaction proceeds to completion). Understanding complex reactions requires experimentally or computationally probing the underlying free energy landscape, which accounts for changes in the balance between energy and entropy. However, any interpretable (and scalable) representation of the free energy landscape involves grouping, or equivalently projecting, states. How best to make this grouping remains an outstanding question in studies of reactions in the condensed phase. Here, we consider this question from a simulation perspective.

Work from many groups makes clear that the grouping should be informed by dynamics. One approach is to harvest many trajectories that contain the event(s) of interest and then to use them to identify coordinates that separate transition states from stable states.[1] To avoid biasing the results, the transition states and stable states are defined by their likelihoods of leading to products and reactants in additional simulations initiated with momenta drawn from a Maxwell−Boltzmann distribution (i.e., their basin commitment probabilities). To facilitate this analysis, several statistical methods have been introduced to correlate commitment probabilities and physical variables.[2−6] Surprisingly, the commitment probabilities of seemingly very complex reactions can be well predicted from only a small number (one to three) of physical variables; the resulting low-dimensional projections of the free energy onto those variables have the advantage that they are readily interpretable.

An alternative approach is to use an equilibrium molecular dynamics simulation to define a graph in which long-lived states are represented as nodes and transitions between them are represented by edges.[7−11] Here, the physical variables enter through the procedure that clusters structures into nodes. Many variables, including commitment probabilities[12,13] and others based on dynamics,[11] can be used for the grouping to minimize the loss of information. However, when high-dimensional, the graph can be difficult to visualize. Krivov and Karplus[12,13] have shown that there exists a one-dimensional barrier-preserving free energy projection. This projection provides a detailed description of the free energy basins and barriers according to the transitions in reactive trajectories at equilibrium, but the coordinate is a partition function, making it less readily interpretable than geometric or energetic variables.

In the present paper, we seek to relate these two approaches. To this end, we focus on the folding of a 20-residue antiparallel β-sheet miniprotein (beta3s). This reaction shares many features with others in the condensed phase in that the interplay of energy and entropy is important and the dynamics in the transition state are diffusive. Beta3s has been extensively studied because it folds reversibly in atomic-resolution molecular dynamics simulations and has a unique native state but heterogeneous nonnative states.[7,13−17] Attempts to identify geometric coordinates that describe its folding have failed in the past.[17] Our studies begin with construction of a conformational space network by a long molecular dynamics simulation. Commitment probabilities, relative partition functions, and free energy profiles are calculated on the basis of complex network analysis. Both enthalpic and entropic basins are identified within the network. Detailed

* To whom correspondence should be addressed. E-mails: (A.C.) caflisch@bioc.uzh.ch, (A.R.D.) dinner@uchicago.edu.
[†] Department of Chemistry, The University of Chicago.
[‡] James Franck Institute, The University of Chicago.
[§] Insitutue for Biophysical Dynamics, The University of Chicago.
[⊥] University of Zurich.

10.1021/jp101476g
Published on Web 05/03/2010

information about folding pathways and mechanisms are thus deduced from the studies.

Here, we exploit the graph to estimate commitment probabilities for the major stable states, and these serve as the input to a statistical procedure for identifying physical variables that describe the dynamics. Specifically, we employ the genetic neural network (GNN) approach, as in ref 18. We find that the sum of the distances of eight backbone hydrogen bonds is the most effective reaction coordinate in the overall beta3s folding/unfolding processes; different reaction coordinates better describe individual folding pathways, and these coordinates provide structural insights into the nature of basins within the unfolded state. The location of the folding free energy barriers is in agreement with the earlier studies. Although different folding pathways correspond to different shapes of free energy profiles, they share almost the same values of free energy barriers.

2. Methods

The goal of the present study is to show that commitment probabilities that are obtained by analysis of the network of transitions in an equilibrium molecular dynamics simulation can serve as the basis for statistical approaches for automatically identifying reaction coordinates. In this section, we first review the network analysis procedure used to estimate the commitment probabilities. Then we describe the specific method used to relate commitment probabilities to physical variables and the database used in the present study. Finally, we describe how we project the free energy and diffusion constant onto selected coordinates.

2.1. System and Network Analysis. The system we study is a 20-residue, three-stranded, antiparallel β-sheet peptide, named beta3s, in an aqueous solution. We represent beta3s with all its heavy atoms and its polar hydrogen atoms (those bound to nitrogen or oxygen atoms) and the solvent by the SASA (solvent accessible surface area) implicit solvation model.[14,19] There are 215 atoms in this system. We obtained a total of 20 μs of folding/unfolding trajectories at 330 K (close to the melting temperature) with the program CHARMM.[20] The SHAKE algorithm was used to fix the covalent bonds involving hydrogen atoms, and the integration time step was 2 fs. The structures were saved every 20 ps, for a total 10^6 structures.

The network of states visited during beta3s folding and unfolding has been analyzed extensively.[7,13−17] The nodes and links in the network represent configurations and transitions between them, respectively (Figure 4 of ref 13). To define the nodes and links, an all-atom rmsd cutoff of 2.5 Å is used. Starting from a comparison with a representative native structure, a structure is taken as a new node if its rmsd exceeds a given threshold (2.5 Å); each subsequent structure is compared with the contributing structures of existing nodes found so far and assigned to either an existing node or a new node if its rmsds with all the existing contributing structures exceed the threshold.[21] By this scheme, the 10^6 structures generated from the equilibrium trajectories are grouped into 34 671 nodes with 30 223 pairwise links within nodes and 121 596 links between different nodes. A link from node j to node i occurs when a snapshot in node i is followed by a snapshot in node j. As mentioned above, structures are saved every 20 ps, which makes the time interval of a link 20 ps. The weight of a node i is calculated as $w_i = n_i/N$, where n_i is the number of structures within node i and $N = 10^6$. A transition matrix with element n_{ij} equal to the number of links from node i to node j is calculated as well as the transition probability of node i to node j, $p_{ij} = n_{ij}/\Sigma_k n_{ik}$, which is to be used in the calculation of commitment probabilities.

2.2. Estimating Commitment Probabilities. We characterize the dynamics of structures with their commitment probabilities. Typically, for a two-state reaction system in which A and B represent the two stable states, the commitment probability p_B is defined as the likelihood that a structure taken from a reactive trajectory will commit to basin B prior to basin A in molecular dynamics simulations initiated with momenta drawn isotropically from a Maxwell−Boltzmann distribution. From the above coarse-grained network, we see that the system has a unique native state but various different nonnative states. Owing to the heterogeneous nature of the nonnative states, here, the commitment probability of reaching the folded state, denoted p_{fold}, is modified to be the likelihood that dynamics trajectories of 10 ns all starting from one node end with a structure within 2.5 Å all-atom rmsd of the native structure, in keeping with earlier work.[17] A commitment time of 10 ns is chosen because it is far less than the folding time (0.1 μs) but long enough for local relaxation. Furthermore, $p_{unfold,i}$, the probability for a structure to commit to a nonnative state, node i, is calculated in the same manner with the native state replaced by a nonnative state, node i, in the above scheme. p_{fold} and $p_{unfold,i}$ represent the probabilities for a structure to undertake folding and unfolding reactions, respectively. Structures within the same node are assumed to have the same values of p_{fold} and $p_{unfold,i}$. To calculate p_{fold} and $p_{unfold,i}$, we solve the equations $p_i = \sum_j p_{ij}(p_j - p'_j)$ for all i iteratively with the boundary condition $p_B = 1$ (where B represents either the native state or node i), where p_i is the commitment probability for node i, p_{ij} is the transition matrix element, and p'_j is the probability for node j to commit to state B at exactly the commitment time (10 ns) later. In the calculation, we divide the commitment time, 10 ns, into 500 steps (because structures are saved every 20 ps), so that the commitment probability, p_i (the probability of node i to commit to basin B within 500 steps), is equivalent to the sum of the product of p_{ij}, the probability of node i to transit to an arbitrary node j in the first step, times $(p_j - p'_j)$, the probability of node j to commit to basin B within 499 steps; p'_j is the probability for node j to commit to basin B at exactly 500 steps. Solution involves multiplications of the transition matrix and the weights of the nodes. The Supporting Information of ref 13 provides more details.

To assess the quality of the commitment probability values calculated by the above procedure, 340 structures (30, 110, and 200 in the folded state, denatured state, and transition state region, respectively) were selected and their values of p_{fold} were calculated by the conventional shooting method. A scatter plot comparing p_{fold} obtained by the conventional method and the complex network analysis is shown in Figure 1. The comparison shows that the correlation between p_{fold} from these two methods is 0.95.[17] Therefore, the commitment probability values calculated by the network procedure are expected to be almost as accurate as routine direct evaluations.

2.3. Identifying Reaction Coordinates. The idea of statistical approaches for automatically identifying reaction coordinates is that a set of representative configurations (or phase space points) can be used to correlate measures of dynamic behavior (e.g., commitment probabilities) with physical variables (geometric and energetic features of the system).[2] There are now a number of such procedures that differ in detail (see the Discussion). We employ the genetic neural network approach,[2,18] which was originally developed for elucidating quantitative structure−activity relationships in medicinal chemistry.[22,23] In this approach, artificial neural networks are used to determine the functional dependence of the commitment probabilities on

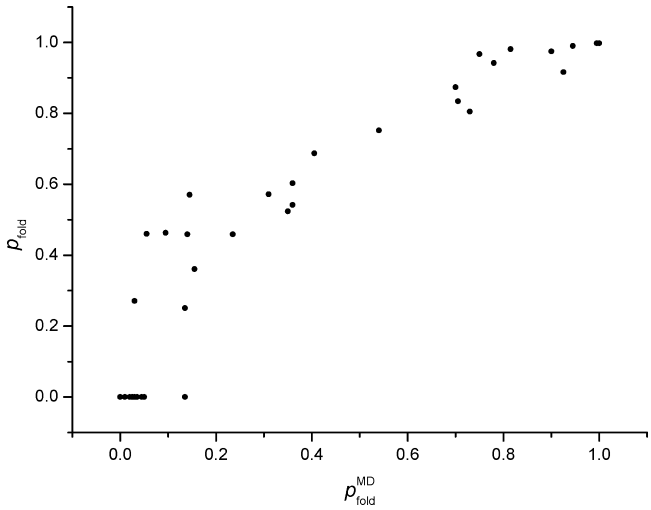

Figure 1. Comparison of the p_{fold} values calculated by the conventional shooting procedure shown on the horizontal axis and the complex network analysis on the vertical axis.

combinations of physical variables, and then a genetic algorithm is used to search the input combinations that enable the best fits (Figure 1 of ref 2). GNN details are given in refs 2 and 18; we employ the implementation in the program CHARMM (version 36a1). Default parameters are used, except that in the genetic function approximation (GFA) model, the number of generations to reproduce and the number of individual models in the production pool are set to be 20 and 10, respectively.

The input for the GNN procedure is a database that consists of commitment probabilities and physical properties for a set of representative structures. We manually identified as many structural and energetic variables as we could to characterize the overall, main chain, and side chain conformations; in total, there are 593 descriptors (Table 1), which can be categorized as energy terms, distances, angles, dihedral angles, rmsds from the representative native structure, solvent accessible surface areas, and fractions of native contacts. The descriptors along with the p_{fold} and $p_{unfold,i}$ data are used to identify the reaction coordinates for the overall folding reaction and individual folding pathways, respectively. The original network analysis was based on 10^6 structures, with the distribution of p_{fold} shown with the histogram in Figure 2. From these structures, we selected 26 720 with a roughly uniform distribution of p_{fold}, shown with the horizontal line in Figure 2, to avoid biasing the neural network toward the stable states. We put 24 720 structures in the training set and the remaining 2000 in the test set. Both sets maintain uniform distributions of p_{fold} to avoid biasing the fit to particular structures. For $p_{unfold,i}$, the number of structures remaining after structures are removed to obtain a uniform distribution of commitment probabilities is small (less than 1000). Thus structures were kept in a single set, and Jackknife cross-validation was used. In both cases, for identifying only one descriptor as a reaction coordinate, exhaustive enumeration was used to select descriptors; for the identification of two or three descriptors, the genetic function approximation was used because it would be prohibitive to test all 175 528 or 34 579 016 possible models, respectively.

2.4. Calculating Mean First Passage Times and Free Energy Projections. We obtain free energy projections onto the reaction coordinates identified for both the overall reaction and the individual folding pathways. To this end, a new

TABLE 1: Physical Variables Forming the GNN Database[a]

variable category	specific descriptions	no. of variables
energy terms	solvation energy, VDW energy, electrostatic energy	194
	interactions between hydrogen bond atoms of the backbone	
	interactions between residues	
	interactions between side chains	
distances	distances between hydrogen bonding backbone O and H atoms	70
	$C_\beta-C_\beta$ distances	
	$C_\alpha-C_\alpha$ distances	
	$C_\alpha-C_\beta$ distances	
	distances between side chains (geometrical centers)	
	distances between residues (geometrical centers)	
angles	angles between backbone atoms of different residues	17
dihedral angles	ϕ	195
	ψ	
	χ_1	
	χ_2	
	ω	
	other dihedrals between atoms of different residues	
rmsd to the native structure	rmsd of the whole protein	6
	rmsd of side chains	
	rmsd of backbone	
solvent-accessible surface area (ASA)	total ASA	41
	ASA per residue	
	ASA per side chain	
DSSP	angle $C\alpha_{i-2}-C\alpha_i-C\alpha_{i+2}$	63
	dihedral angle $C\alpha_{i-1}-C\alpha_i-C\alpha_{i+1}-C\alpha_{i+2}$	
	cosine of angle between C═O of residue i and C═O of residue $i-1$	
fraction of native contacts[14]	hydrogen bond distances + side chain distances	7

[a] ϕ, ψ, χ_1, χ_2, and ω are dihedral angles $C_{i-1}-N_i-C\alpha_i-C_i$, $N_i-C\alpha_i-C_i-N_{i+1}$, $N_i-C\alpha_i-C\beta_i-C\gamma_i$, $C\alpha_i-C\beta_i-C\gamma_i-C\delta_i$, and $C\alpha_i-C_i-N_{i+1}-C\alpha_{i+1}$. DSSP stands for definition of secondary structure of protein. We use the DSSP program ((a) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637. (b) Andersen, C. A. F.; Palmer, A. G.; Brunak S.; Rost, B. *Structure* **2002**, *10*, 175−184)) to obtain a quantitative description of the beta3s secondary structure. In the text, the DSSP variables are categorized into angles and dihedral angles. The variables in the fraction of native contacts are taken from ref 14.

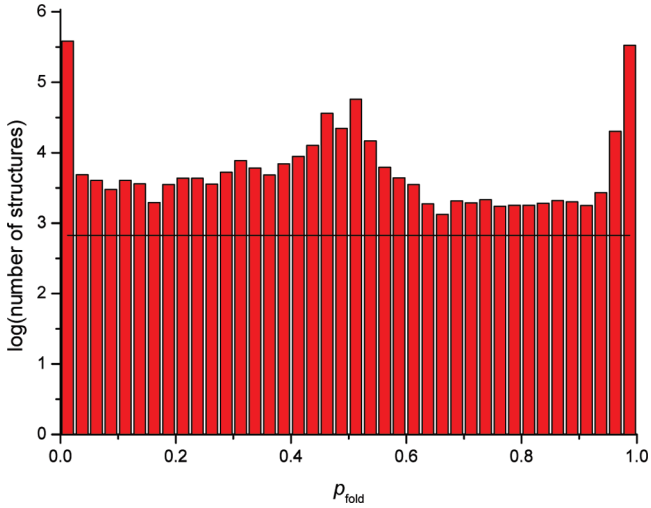

Figure 2. Distribution of p_{fold} for the original rmsd-based network. The line indicates the number of structures used to represent each bin width of 0.025 in p_{fold} in the GNN database (training and test sets together), which is uniform in p_{fold} by construction.

conformational space network is generated by grouping the 10^6 structures according to the identified reaction coordinate instead of the rmsd. The mean first passage time of each node, denoted as mfpt_c, is calculated by iteratively solving the equation: $\text{mfpt}_i = \Delta t + \Sigma(p_{ij} \times \text{mfpt}_j)$ with initial boundary condition $\text{mfpt}_B = 0$, where B represents the node of the native state. Δt is the saving interval of 20 ps. The mfpt of a node can be calculated as the sum of one time step plus the weighted average of the mfpt values of its adjacent nodes.

Given the new network, the program WORDOM[24] is used to perform the free energy calculations. The main idea is that the free energy profile can be obtained by concatenating information about transitions between nodes in the network ordered to reflect the kinetics. Specifically, at each point along the free energy profile, the nodes of the network are partitioned into two groups: A with $\text{mfpt} < \text{mfpt}_c$ and B with $\text{mfpt} > \text{mfpt}_c$. The partition function for group A, $Z_A = \Sigma_{i\in A} Z_i$, where $Z_i = \Sigma_j c_{ij}$ is the contribution for node i and c_{ij} is the number of transitions between nodes i and j after detailed balance, is imposed: $c_{ij} = (n_{ij} + n_{ji})/2$. Given $Z_{AB} = \Sigma_{i\in A, j\in B} c_{ij}$, the free energy of the barrier at mfpt_c between the two groups A and B thus can be calculated as $-kT \log(Z_{AB}/Z)$, where $Z = \Sigma_i Z_i$ is the partition function for the full network (Figure 1 of ref 13). Each structure with a specific value of the identified reaction coordinate corresponds to a specific mfpt_c and Z_A/Z and, thus, corresponds to a specific free energy value. Therefore, we can plot the free energy as a function of either the selected coordinate or the partition-function-like coordinate Z_A/Z.

2.5. Calculating Diffusion Constants. The diffusion constant in the selected coordinates can indicate the extent to which the free energy profile provides insights into the kinetic behavior. Because we did not have analytical derivatives for all the coordinates, we applied the approach put forward by Im and Roux,[25] which is discussed in detail in ref 26. Given q as the identified reaction coordinate, its diffusion constant is

$$D = \frac{\langle [\Delta q(t) - \langle \Delta q(t) \rangle]^2 \rangle}{2\tau}$$

for $\Delta q(t) = q(t + \tau) - q(t)$. If the displacement were driven purely by Brownian motion, the time interval, τ, could be chosen

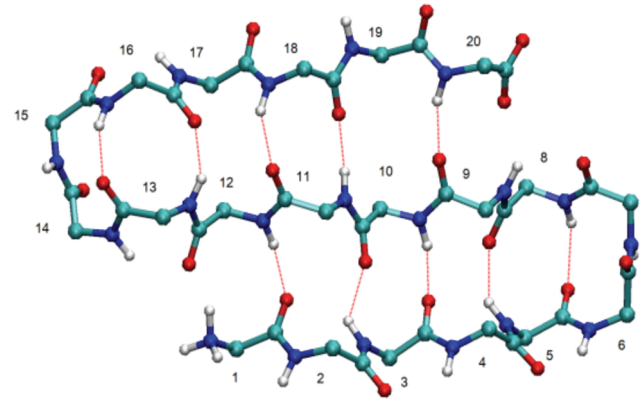

Figure 3. Native structure of beta3s with native backbone hydrogen bonds indicated.

arbitrarily. In practice, τ must be sufficiently large to remove memory effects but not too large, since the analysis leading to the above expression is based on the assumption of a small time step. We first recluster the 10^6 structures according to the identified reaction coordinate and assume that structures within the same node have the same diffusion constant. Then the diffusion constants can be calculated for each node via the above expression. We set τ to be 20 (the smallest choice which equals the time interval of structure saving), 40, 60, 200, and 2000 ps to assess how varying this parameter affects the estimated diffusion constant. We found that the diffusion constant was largest at $\tau = 20$ ps. Thus, we obtain a diffusion constant profile for the folding reaction at $\tau = 20$ ps.

3. Results

The goals of the present study are (1) to examine the extent to which small numbers of physical variables can be used to capture the dynamics of complex, multipathway reactions and (2) to illustrate the use of network analysis for estimating commitment probabilities that serve as inputs to statistical procedures for identifying reaction coordinates. We consider commitment probabilities for the overall folding reaction (p_{fold}) as well as ones for reaching major nonnative states identified in the equilibrium transition network analysis (p_{unfold}). In each case, we construct a database that consists of the commitment probabilities of interest and 593 physical variables, as detailed in the Methods section; we then use the GNN procedure to identify combinations of one to three descriptors that best predict the commitment probabilities (as quantified by the rms error for the test set structures in p_{fold} and for all the structures in p_{unfold}). We compare projections of the free energy onto the partition-function-like progress coordinate (Z_A/Z) on the basis of two networks: one coarse-grained by the selected coordinates and the other, by the previous rmsd scheme. The rmsd-based free energy profile has been argued to be able to preserve basins and barriers for the folding reaction.[13] Free-energy profiles and diffusion constants projected onto the selected coordinates are also generated.

3.1. Coordinates for the Overall Folding Reaction. The NMR structure of beta3s[27] is shown in Figure 3, and we take it as our native reference. Trajectories that reach the native state (with rmsd from the native structure less than 2.5 Å) within 10 ns contribute positively to p_{fold}. The single coordinate that is most effective in predicting p_{fold} is the sum of distances between the H and O atoms of the backbone hydrogen bonds between residues 3 and 10, residues 5 and 8, residues 11 and 18, and residues 13 and 16. The rms error over the 2000

TABLE 2: GNN Results for the Overall Reaction Coordinates

database	selected coordinate[a]	rms error over 2000 structures
one variable	d_{HB} of 3−10, 5−8, 11−18, 13−16	0.1768
	d_{HB} of 3−10, 5−8, 11−18, 13−16, 9−20, 1−12	0.1835
	d_{HB} of 5−8, 13−16	0.1839
	E_{HB}^{elec} of 3−10, 5−8, 11−18, 13−16	0.1890
	E_{HB} of 3−10, 5−8, 11−18, 13−16, 1−12, 9−20	0.1892
	E_{HB}^{elec} of 3−10, 5−8, 11−18, 13−16, 1−12, 9−20	0.1894
	d_{HB} of 3−10, 13−16	0.1908
	d_{HB} of 5−8, 13−16, 1−12, 9−20	0.1910
combination of two variables	$E_{side\text{-}chain}^{VDW}$ of 3−10, 5−8, 11−18, 13−16, 1−12, 9−20 + E_{HB} of 11−18	0.1834
	$d_{side\text{-}chain}$ of 3−12, 10−17 + dihedral angle N_4−$C\alpha_4$−$C\beta_4$−$C\gamma_4$	0.2158
combination of three variables	$E_{side\text{-}chain}^{VDW}$ of 2−11, 10−19 + d_{HB} of 3−10, 13−16 + sum of angle $C\alpha_4$−$C\alpha_6$−$C\alpha_8$ and angle $C\alpha_{13}$−$C\alpha_{15}$−$C\alpha_{17}$	0.1858
	E_{HB}^{VDW} of 10−3 + $E_{side\text{-}chain}^{VDW}$ of 2−11, 10−19 + d_{HB} of 3−10, 13−16	0.1887

[a] d_{HB} and $d_{side\text{-}chain}$ denote the sum of distances between hydrogen bonding backbone O and H atoms and between the geometric centers of side chains, respectively. E_{HB}, E_{HB}^{elec}, $E_{side\text{-}chain}^{VDW}$, and E_{HB}^{VDW} are energy terms. HB and side chain subscripts denote hydrogen bond and side chain interactions, respectively; elec and VDW superscripts denote the electrostatic and van der Waals parts of the energy function.

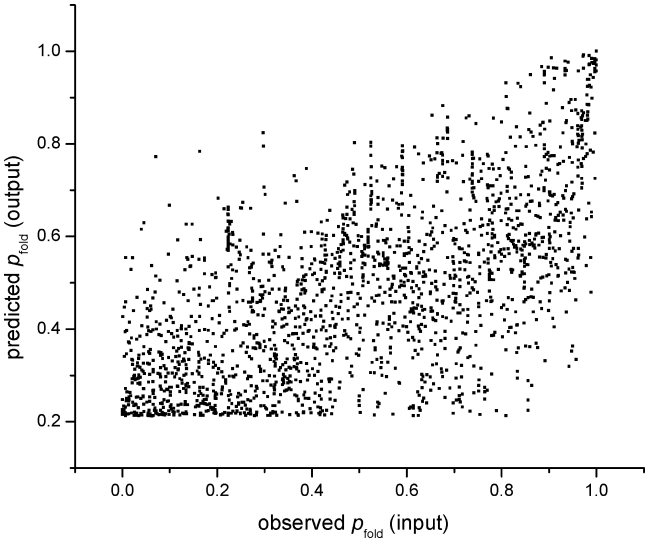

Figure 4. Comparison of the p_{fold} values input to and output from the GNN for the 2000 structures in the test set for the highest ranked model based on d_{HB} of 3−10, 5−8, 11−18, 13−16, the sum of distances between the H and O atoms of the backbone hydrogen bonds between residues 3 and 10, residues 5 and 8, residues 11 and 18, and residues 13 and 16.

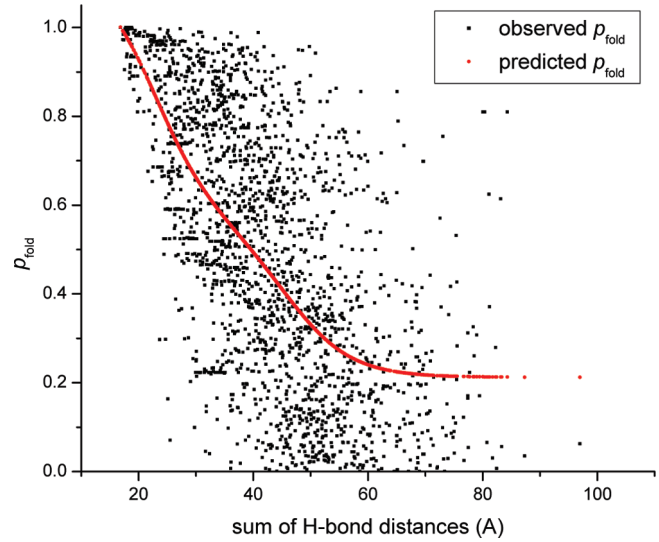

Figure 5. Dependence of p_{fold} on the coordinate yielding the best prediction, d_{HB} of 3−10, 5−8, 11−18, 13−16 (see text and caption to Figure 4).

structures in the test set is 0.177 (Table 2 and Figures 4 and 5); no significant improvement was obtained with additional descriptors (Table 2). This error is somewhat higher than that obtained in previous studies[2,18] but is reasonable given the fact that the commitment probabilities are estimated from the network rather than calculated directly. The actual error in the commitment probabilities (relative to what would be obtained in an infinite number of shooting simulations) sets a limit on the error in predicting the observed values. Figure 4 is a graphical comparison of the p_{fold} values input to and output from the GNN. Figure 5 shows the correlation between the reaction coordinate and the input/output p_{fold} values.

Since each pair of the aforementioned residues has two hydrogen bonds, the reaction coordinate is actually the sum of eight OH distances. We interpret the choice of this coordinate to mean that the transition states for folding involve formation of one of the two hairpins. This result is consistent with the earlier observation[7] that there are two main folding pathways: one with a transition state ensemble characterized by a native-like C-terminal hairpin and an unstructured N-terminal hairpin and the other with a native-like N-terminal hairpin and an unstructured C-terminal hairpin. Indeed, Figure 6 shows representative structures with both predicted and observed p_{fold} close to 0.5 (for TS1, $p_{fold(observed)} = 0.5080$, $p_{fold(predicted)} = 0.5034$; for TS2, $p_{fold(observed)} = 0.5190$, $p_{fold(predicted)} = 0.5002$); representatives of the native state and various unfolding states (with both predicted and observed p_{fold} close to 0) are also shown. Statistics of this reaction coordinate and the contributing OH distances for the representative structures in Figure 6 are given in Tables 3 and 4. Corroborating the importance of forming these native backbone hydrogen bonds, the eight top-ranked single-descriptor models are all related variables (Table 2). Although side chain descriptors were included in the database, none were selected.

The cut-based free energy profile as a function of the sum of the eight OH distances is obtained from the equilibrium transition network as detailed in the Methods section (Figure 7a). We recluster the 10^6 structures by grouping them into 8169 bins (thus, 8169 nodes) with the bin width 0.01 Å according to their OH distance sums. There are 1094 pairwise links within the same nodes and 861 131 pairwise links between different nodes in the reclustered network. We then can define an analogous partition-function-like variable, Z_A/Z, and perform the free energy calculation (see section 2.4, Calculating Mean First Passage Time and Free Energy Projections). Each Z_A/Z

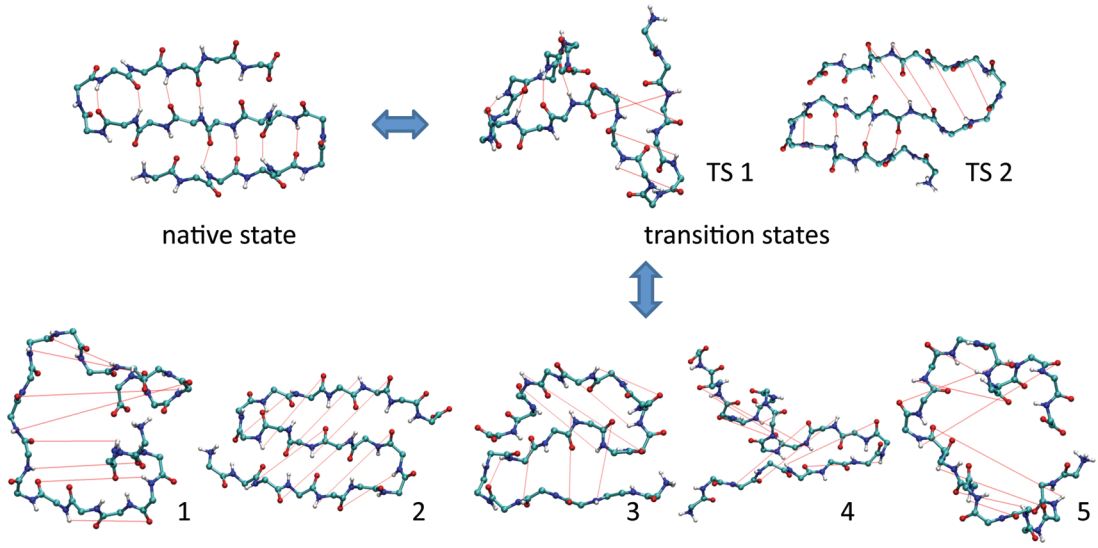

Figure 6. Representative structures for the overall folding/unfolding reaction. Non-native states 1, 4, and 5 are extended states; nonnative states 2 and 3 are misfolded states.

TABLE 3: Statistics of d_{HB} of 3−10, 5−8, 11−18, 13−16[a]

sum of H-bond distances (Å)	no. of structures	av	SD
native state ensemble	288029	19.597	2.511
transition state ensemble	6413	43.502	9.489
nonnative state ensemble	381694	65.165	10.983

[a] Structures with $p_{fold} < 0.01$, $p_{fold} \in (0.495, 0.505)$, and $p_{fold} > 0.99$ based on the rmsd coarse-grained network are selected to form ensembles of the nonnative state, transition state, and native state.

corresponds to a specific node that is defined by its value of the reaction coordinate; we thus obtain the free energy projection onto the sum of the eight OH distances. We see that a broad, unfolded basin and a narrow, folded basin are separated by a barrier of ∼1 kcal/mol. The position-dependent diffusion constant along this coordinate varies over a range of ∼1 Å^2/ps (Figure 7b), which suggests that the kinetics can be meaningfully interpreted in terms of the free energy profile.

The earlier network analysis centered on Z_A/Z based on the rmsd coarse-grained network. For reference, the rms error for a neural network fit of p_{fold} to the rmsd (all referring to the representative native structure) yields a rms error of 0.195. To facilitate comparison with the earlier work,[13] we also reorder the network according to the sum of the eight OH distances and then obtain the free energy projection onto the OH-distances-based Z_A/Z (Figure 8). The free energy barrier to folding appears at almost the same place; that is, at $Z_A/Z = 0.38$, which indicates that the native basin is identified correctly. In contrast, the height of the barrier is ∼1.0 kcal/mol lower using the OH distances sum for coarse-graining than the result in ref 13. It is important to note that differences in the cut-based FEP after the first barrier (at $Z_A/Z = 0.38$ in Figure 8) are expected because of the overlap of structures after the first barrier; that is, different structures with the same kinetic distance from the folded state[12,13] that originate from parallel folding pathways. From Table 3, we can see that although the value of the reaction coordinate for the native state is at around 20 Å, the unfolded state covers the range from 54 to 76 Å. These numbers are in agreement with the range of the plateau in Figure 7.

Finally, with regard to the network reclustered by the sum of the OH distances, we examined whether the reclustered network could be used to improve the p_{fold} estimates. However, we found that the majority of structures were assigned $0.2 < p_{fold} < 0.4$ with the commitment time set to be 10 ns (data not shown). This narrow range results from the fact that the selected coordinate groups different transition states and different unfolded states together, so there is a loss in information relative to the rmsd-based network, consistent with the smoothing of the free energy profile in Figure 8.

Given the apparent importance of the eight OH distances in the turns, we also trained an artificial neural network (with an 8−2−1 input−hidden−output topology) to predict p_{fold} from these data directly. The resulting rms error was 0.162, the best accuracy that we achieved throughout the study. We also reclustered the 10^6 structures such that each node contained

TABLE 4: Changes of Key Hydrogen Bond Distances during the Reaction

				representative nonnative states[a]				
distance (Å)	native	TS 1	TS 2	1	2	3	4	5
3H−10O	2.293	8.790	1.847	7.928	9.607	9.356	10.746	13.236
3O−10H	1.960	7.459	1.833	9.114	10.052	10.899	11.301	10.461
5H−8O	2.118	8.292	2.247	9.977	10.854	9.898	11.005	6.784
5O−8H	2.278	6.522	5.630	6.945	6.759	6.580	7.898	5.941
11H−18O	1.970	1.972	7.457	15.840	10.340	9.182	13.385	14.026
11O−18H	2.186	2.264	7.736	14.188	10.237	12.686	7.948	10.390
13H−16O	1.999	1.969	7.416	10.237	9.593	10.608	8.156	8.604
13O−16H	2.095	2.118	5.422	6.561	7.421	7.556	4.118	5.652
sum	16.899	39.386	39.588	80.790	74.863	76.765	74.557	75.094

[a] The representative nonnative states are shown in Figure 6.

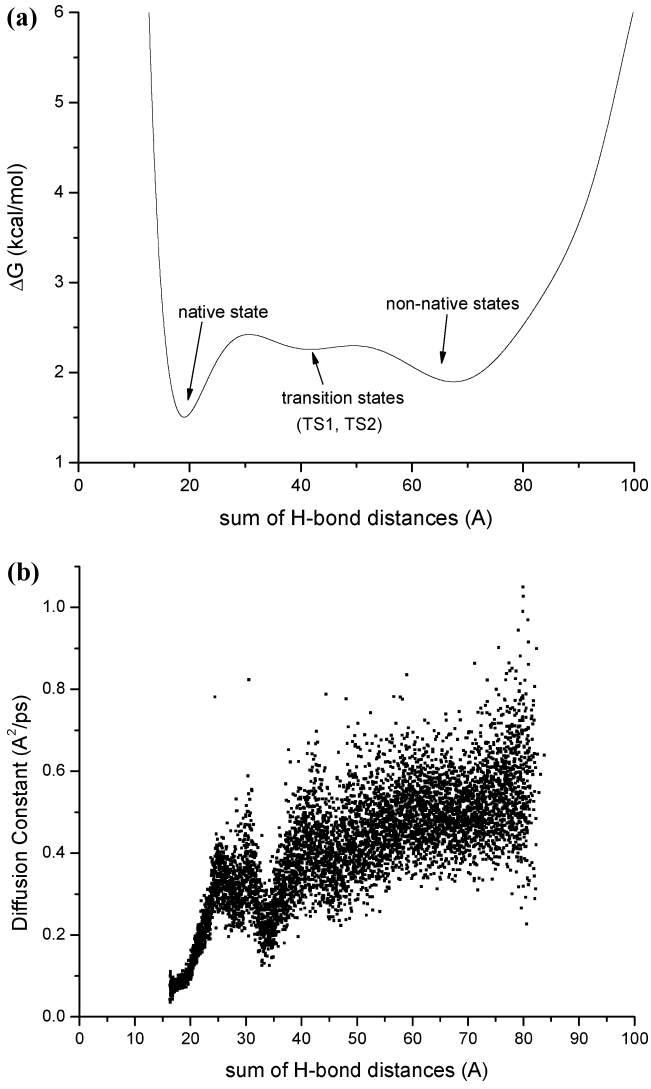

Figure 7. Thermodynamics and kinetics as a function of d_{HB} of 3−10, 5−8, 11−18, 13−16. (a) Free energy. Polynomial fitting was used to reduce statistical variation. (b) Position-dependent diffusion constant evaluated with $\tau = 20$ ps (see text).

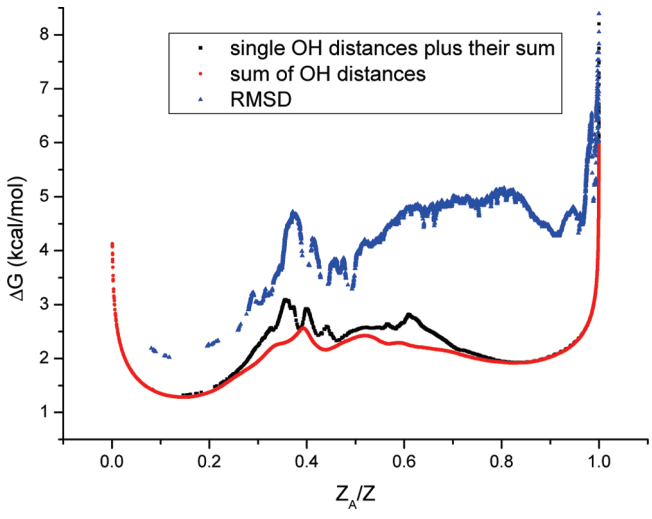

Figure 8. Comparison of cut-based free energy projections obtained from networks in which structures are grouped by rmsd (blue triangles), d_{HB} of 3−10, 5−8, 11−18, 13−16 (red circles), and combination of individual OH distances and their sum (black squares).

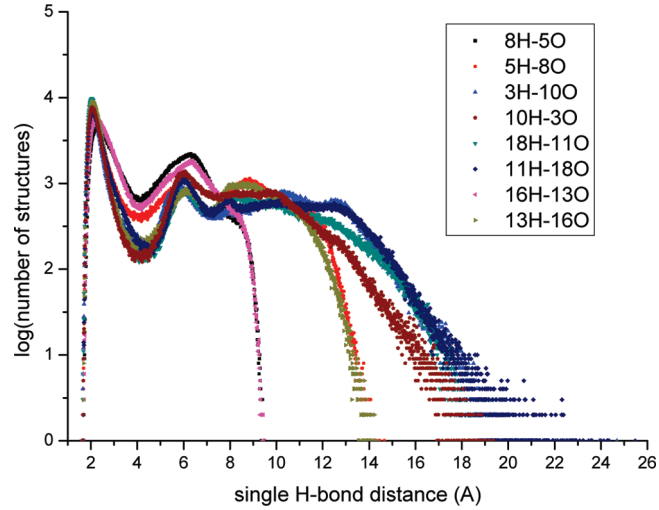

Figure 9. Distributions of each OH distance of 3−10, 5−8, 11−18, 13−16.

structures in which each of the eight distances were all within a single bin of width 3 Å (62 601 nodes). This tended to obscure the barriers to folding (data not shown), which suggests that transition states and stable states were grouped together by this procedure. Unfortunately, there were insufficient statistics to make the bin width significantly smaller, and thus, the homogeneity of structures within nodes could not be ensured.

We then designed another scheme to recluster the 10^6 structures that coarse-grained both the eight OH distances and their sum. Figure 9 shows the distributions of each OH distance. For each distribution, there is a trough at around 4.2 Å after the native-state peak. Setting the width of the trough to be 2 Å, we could then divide each distribution into three sections: before the trough (as in the native state), within the trough (as in the transition state), and after the trough (as in the unfolded state). We reclustered the structures according to these sections of the eight OH distributions and then grouped the structures into bins of width 1 Å according to their OH distance sums. A total of 23 217 nodes are generated by this hybrid scheme, and the free energy projection is again obtained (Figure 8). We see that this free energy projection provides richer information as compared with that obtained from the network reclustered using only the sum of OH distances. It is also in good agreement with the rmsd coarse-grained network in that it presents similar shapes and heights of free energy barriers around the transition state region. We were thus able to exploit the higher information content of the individual distances for procedures that coarse grained both the single OH distances and their sums.

To put the work in context, as previously suggested, no single hydrogen bond distance or sum of distances within only one hairpin is the most effective reaction coordinate. Also consistent with previous results,[17] Q_N and Q_C, the fractions of native contacts within the N hairpin and C-terminal hairpins, yield relatively large rms errors of 0.203 and 0.228, respectively. However, the GNN procedure was able to identify the sum of the eight OH distances in the turns as a reasonable descriptor of the dynamics. To further verify the effectiveness of this coordinate, we calculated p_{fold}^{MSM}, the folding probability corresponding to a Markov state model, as defined in ref 17. As discussed in that study, p_{fold}^{MSM} is very sensitive to the definition of the unfolded state. Here, we took the folded state to be structures with sums less than 20 Å and the unfolded state to be structures with sums more than 60 Å. The resulting distribution of p_{fold}^{MSM} shown in Figure 10 reflects the barrier

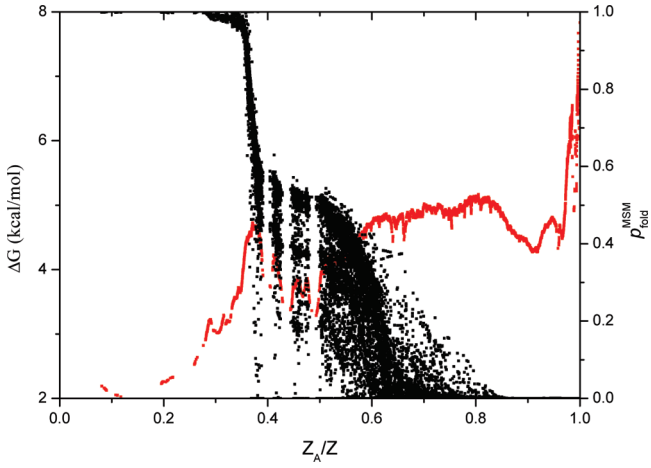

Figure 10. Distribution of p_{fold}^{MSM} (black symbols, right axis) calculated with $d_{HB} < 20$ Å for the folded state and $d_{HB} > 60$ Å for the unfolded state. The distribution is superimposed on the cut-based free energy projection (red symbols, left axis) from the original complex network. The distribution of intermediate values of p_{fold}^{MSM} is narrower than in Figure 6 of ref 17, indicating that the use of d_{HB} of 3−10, 5−8, 11−18, 13−16 to define the stable states is an improvement over the use of the number of native contacts.

region better than the distribution in Figure 7 of ref 17, which was obtained with basin definitions based on the fraction of native contacts (which gives a rms error of 0.216 for commitment probability predictions). This suggests that the sum of eight OH distances in the turns serves as a good coordinate for describing the overall reaction, at least in part because it is robust to the heterogeneity of the unfolded state.

3.2. Reaction Coordinates for Individual Folding Pathways. In addition to the native state, there are nine nonnative states identified in the earlier network analysis of beta3s folding and unfolding (see Figure 7 of ref 13). Direct transitions between these nine nodes of the network are rare as compared with transitions connecting each node to the native state. It is thus inferred that different reaction pathways exist in the beta3s system, distinguished by their nonnative states. In this study, we select the three most populated pathways for a detailed analysis; for the rest of the pathways, there are not enough data to generate a uniform distribution of commitment probabilities for the GNN procedure.

The unfolded structures of the three pathways are shown in Figure 11. All of them are states with one of the two hairpins misfolded. The transition states in pathways 2 and 3 belong to the TS1 ensemble in Figure 6; the transition state in pathway 1 belongs to the TS2 ensemble. A set of $p_{unfold,i}$ ($i = 1, 2, 3$), the probabilities for each structure to commit to each unfolding state, are calculated. Ideally, if a structure is within a pathway i, it must satisfy the condition $p_{fold} + p_{unfold,i} = 1$. However, because the p_{fold} and p_{unfold} calculations are performed independently on the basis of statistics of the network, the sum of these two commitment probabilities can exceed 1, and we group snapshots with pathway i ($i = 1, 2, 3$) when $(p_{fold} + p_{unfold,i}) \in (0.8, 1.2)$. We then select structures to form roughly uniform distributions of p_{fold} and $p_{unfold,i}$ which lead to three databases of 924, 436, and 872 structures for nodes 1, 2, and 3, respectively; the physical variables are the same as above.

The descriptors selected are listed in Table 5. Two sets of GNN calculations are performed: one with $p_{unfold,i}$ as a single target parameter; the other with both $p_{unfold,i}$ and p_{fold} as target parameters. A graphical comparison of the $p_{unfold,i}$ values input to and output from the GNN for the single-target-parameter set is shown in Figure 12. Since many fewer structures are used in the GNN procedure, the rms errors are larger than those obtained for the overall pathway. The results illustrate that different pathways are best described by different coordinates. For nodes 1 and 3, dihedral angles are selected; for node 2, the most highly ranked descriptor monitors interactions within a hairpin. To interpret these choices, we show representative transition states in Figure 11 and free energy projections in Figure 13. The dihedral angles in pathways 1 and 3 are located near the turns of hairpins 2 and 1, respectively. They change from native structures to nonnative ones as the hairpins misfold. Pathway 2 involves the change of hydrogen bond distances and side chain

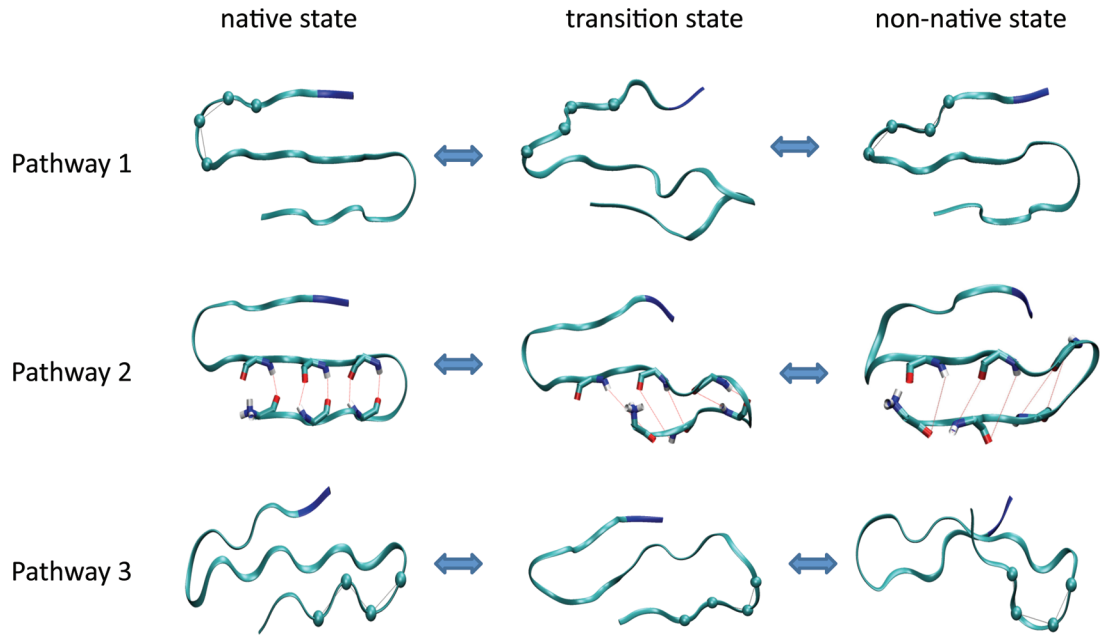

Figure 11. Representative backbone structures from folding/unfolding pathways leading between the native state and three major nonnative nodes. The beads shown in the structures of pathways 1 and 3 indicate the atoms contributing to the dihedral angles that are selected by the GNN (see Table 4). For pathway 2, the native backbone hydrogen bonds that contribute to the GNN-selected coordinate are marked.

TABLE 5: GNN Results for Unfolding to Major Nonnative Nodes[a]

pathway	p_{fold} as target	rms error	$p_{unfold,i}$ and p_{fold} as targets	rms error
1	$C\alpha_{14}-C\alpha_{15}-C\alpha_{16}-C\alpha_{17}$	0.1851	$C\alpha_{14}-C\alpha_{15}-C\alpha_{16}-C\alpha_{17}$	0.2384
	$C\beta_{15}-C\alpha_{15}-C\alpha_{16}-C\beta_{16}$	0.1883	$C\beta_{15}-C\alpha_{15}-C\alpha_{16}-C\beta_{16}$	0.2412
	$N_{15}-C_{15}-C\alpha_{15}-N_{16}$	0.1922	$N_{15}-C_{15}-C\alpha_{15}-N_{16}$	0.2437
2	q_{12}	0.1815	q_{12}	0.2294
	d_{HB} of 5−8, 11−18	0.1895	d_{HB} of 5−8, 11−18	0.2444
	$C\alpha_4-C\alpha_5-C\alpha_6-C\alpha_7$	0.1898	$C\alpha_4-C\alpha_5-C\alpha_6-C\alpha_7$	0.2457
3	$C\alpha_3-C\alpha_4-C\alpha_5-C\alpha_6$	0.1842	$d_{side\text{-}chain}$ of 4−9	0.2278
	$C\beta_4-C\alpha_4-C\alpha_5-C\beta_5$	0.1874	$E^{VDW}_{side\text{-}chain}$ of 4−9	0.2312
	$d_{side\text{-}chain}$ of 4−9	0.1901	$E_{side\text{-}chain}$ of 4−9	0.2324

[a] $C\alpha_{14}-C\alpha_{15}-C\alpha_{16}-C\alpha_{17}$ denotes the dihedral angle between the Cα atoms of residue 14, 15, 16, and 17. Other dihedral angles are denoted in the same way. q_{12} denotes the fraction of native contacts within the N-terminal hairpin.

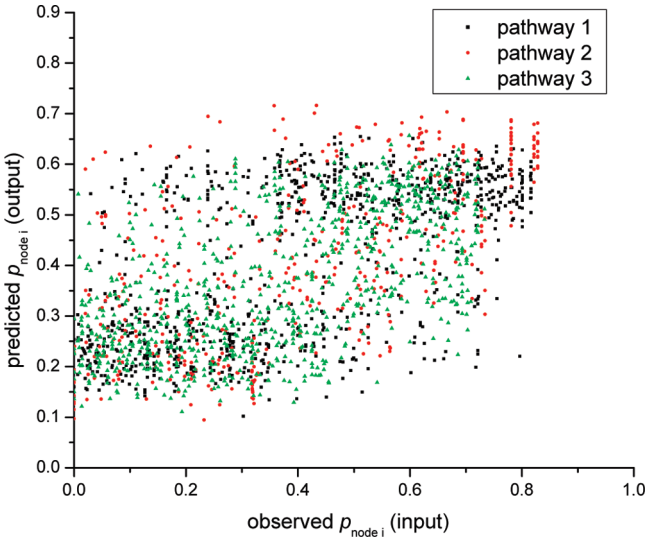

Figure 12. Comparison of the p_{fold} values input to and output from the GNN for the coordinates identified as describing unfolding to major nonnative nodes (see Table 5 and Figure 11).

distances within the entire hairpin 1. All the unfolding pathways involve the breaking of native hydrogen bonds, which confirms the importance of hairpin zipping. The fact that the barriers are comparable in the different pathways follows directly from the fact that they are all observed during folding and unfolding. Indeed, it is the competition between these pathways that makes the beta3s folding reaction complex.

4. Discussion

We have combined complex network analysis with a statistical procedure that relates commitment probabilities to physical variables to identify reaction coordinates for the beta3s folding reaction. This is a challenging system for description in that the dynamically defined transition states are relatively heterogeneous in nature. One might thus expect that interpretable reaction coordinates could not be found.[21] However, we obtain good prediction of commitment probabilities with a coordinate that tracks the formation of eight backbone hydrogen bonds, which are close to the turns in beta3s. Analysis of the contributing hydrogen bonds indicates that there are two structurally distinct transition state ensembles, each of which corresponds to nucleation of the folding reaction from one of the turns. This result is consistent with previous studies.[14] We have performed analogous studies for separate pathways identified during the complex network analysis. Different pathways are found to be best described by different coordinates, but the barriers to folding along the pathways are all comparable, as one would expect from the fact that all these pathways contribute significantly to the dynamics. The studies of individual pathways were facilitated by the fact that we were able to obtain the probabilities for commitment to different nodes from the same equilibrium molecular dynamics data; in this sense, the method employed here for estimating commitment probabilities has a significant advantage over the conventional procedure.

Complex networks[13] and free energy disconnectivity graphs (FE DG)[28] are useful representations of dynamics in that they preserve free energy barriers. The essential element of these approaches is that free energy basins and barriers are elucidated from transitions observed at equilibrium rather than from geometrical features. However, structures must be clustered to obtain transition statistics. Rao and co-workers[8] have shown that the optimal partition of the network into free energy basins is not obvious; in fact, different clustering algorithms detect different free energy basins. In our study, root-mean-square deviations (rmsd) between pairs of structures were used as the basis for grouping structures into nodes. Although this variable itself correlates poorly with commitment probabilities, it appears adequate for the purpose of clustering beta3s conformations in that most of the resulting nodes have small ranges in the physical coordinates ultimately identified as important. At the same time, large ranges were observed for a fraction of nodes such that we would expect the structures within these nodes to have heterogeneous commitment probabilities. As shown, schemes that incorporate the selected coordinates to recluster the simulation data can lead to improvements.

The idea of using statistics from an equilibrium molecular dynamics simulation for identifying coordinates capable of distinguishing dynamically defined transition states was first suggested by Best and Hummer.[3] They used a Bayesian approach to relate the equilibrium probability of observing particular values for a candidate physical variable to the probability of being on a transition path given those values and sought to maximize the peak height of a Gaussian fit to the latter. Independently, Ma and Dinner[2] suggested that one could efficiently search for combinations of coordinates that gave good prediction of commitment probabilities by statistical analysis of a database of candidate physical variables and precomputed commitment probabilities for a set of representative structures. Their key insight was to decouple the evaluation of the commitment probabilities from the testing of candidate variables. Their procedure thus obviated the traditional histogram test in which one harvested new putative transition states for each candidate set of coordinates and evaluated whether the distribution of their commitment probabilities was peaked at a value of one-half.

The specific statistical approach that Ma and Dinner employed, the genetic neural network, which was also employed here, was introduced originally for elucidating quantitative

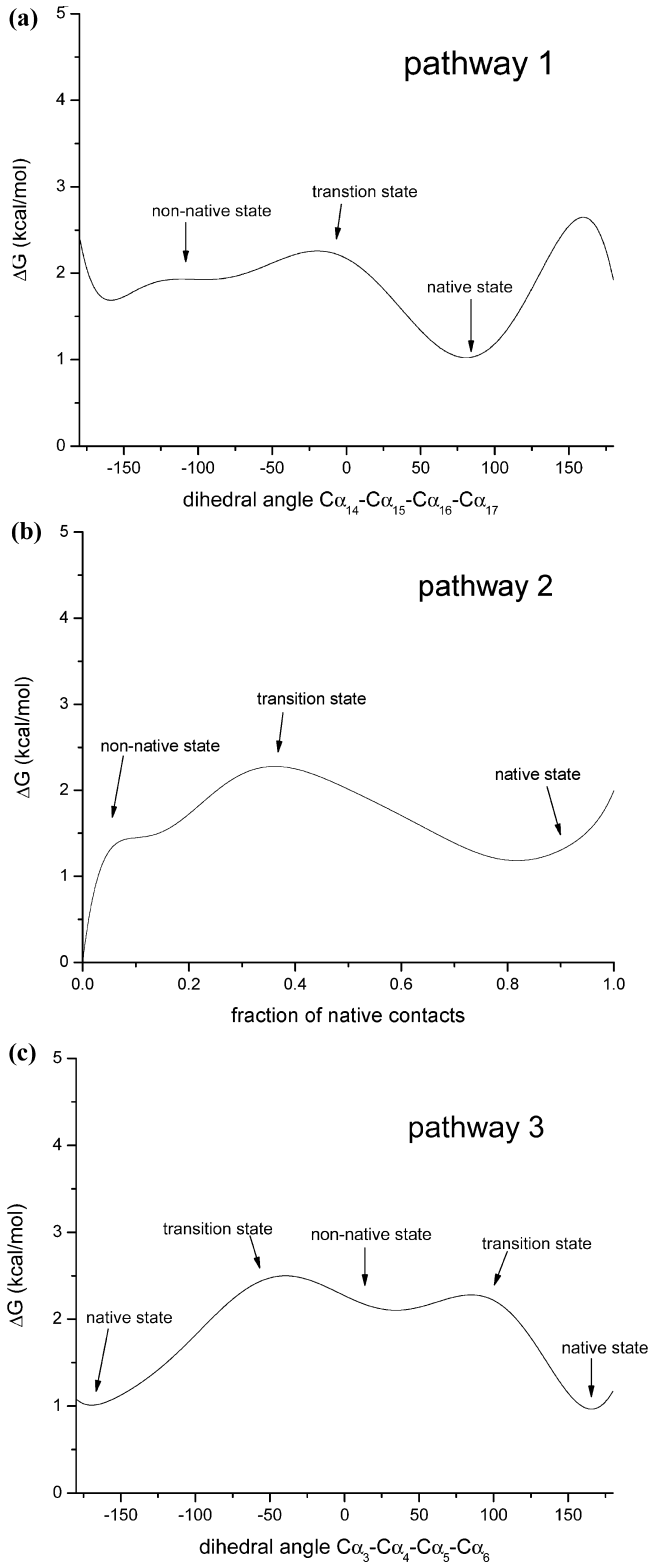

Figure 13. Free energies projected onto the coordinates identified as describing unfolding to major nonnative nodes (see Table 5 and Figure 11). Polynomial fitting was used to reduce statistical variation; continuity was not enforced across the periodic boundary of the dihedral angles in pathways 1 and 3, resulting in somewhat different derivatives at ±180°.

structure−property relationships.[22,23] In the GNN approach, artificial neural networks are used to fit commitment probabilities in terms of combinations of descriptors, and a genetic algorithm is then used to select the combinations that enable the best fit. Peters and Trout subsequently suggested likelihood maximization[4] as an alternative to the GNN (see discussion of the relation of the two methods in ref 5 and the Methods section of ref 18).

Likelihood maximization has the advantage that the measures of statistical significance are better defined, but the relative performance of the two methods in practical applications remains open to debate. Recently, Antoniou and Schwartz[6] presented a new approach to the identification of reaction coordinates by locating the transition state ensemble, defined as the stochastic separatrix, and examining the distributions of candidate coordinates on the separatrix. Their approach is conceptually transparent in that coordinates that are components of the reaction coordinate should have a significantly smaller variation along the separatrix, but it is restricted to systems without diffusive transition states. The GNN is the only tested approach that permits importance sampling (rather than exhaustive enumeration) of combinations of descriptors and nonlinear commitment probability dependencies, which appear to be quite common.

Peters and Trout additionally introduced aimless shooting, which enables binary estimates of commitment probabilities to be obtained during a transition path sampling simulation.[4] In systems that permit shooting and have only a single dominant reaction pathway,[29] aimless shooting is likely to provide significant computational savings. For the strongly diffusive system considered in the present study, we do not expect that to be the case. In particular, limiting the range of points for shooting could slow exploration of the space of trajectories in directions orthogonal to a reaction pathway. For beta3s, the existing equilibrium data[13] obviated the need for further molecular dynamics simulations and permitted reasonably accurate estimates of commitment probabilities to be obtained. It is important to stress that the means of constructing the database of commitment probabilities and the statistical method used to analyze it are separate choices, despite the fact that they have been touched on jointly by Ma and Dinner[2] and Peters and Trout,[4] as well as in the present study. The complementarity between complex network analysis and the GNN makes combining them a very promising approach for studying complex reactions.

Acknowledgment. A.C. thanks Dr. Andreas Vitalis for interesting discussions. This work was supported by a Swiss National Science Foundation grant to A.C. and a National Science Foundation CAREER Award to A.R.D. (MCB-0547854).

References and Notes

(1) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334–350.
(2) Ma, A.; Dinner, A. R. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.
(3) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6732–6737.
(4) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.
(5) Peters, B.; Beckham, G. T.; Trout, B. L. *J. Chem. Phys.* **2007**, *127*, 034109.
(6) Antoniou, D.; Schwartz, S. D. *J. Chem. Phys.* **2009**, *130*, 151103.
(7) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.
(8) Gfeller, D.; De Los Rios, P.; Caflisch, A.; Rao, F. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1817–1822.
(9) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415–425.
(10) Metzner, P.; Schutte, C.; Vanden-Eijnden, E. *J. Chem. Phys.* **2006**, *125*, 084110.
(11) Hummer, G.; Kevrekidis, I. G. *J. Chem. Phys.* **2003**, *118*, 10762–10773.
(12) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.

(13) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.
(14) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10780–10785.
(15) Muff, S.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1185–1195.
(16) Muff, S.; Caflisch, A. *J. Phys. Chem. B* **2009**, *113*, 3218–3226.
(17) Muff, S.; Caflisch, A. *J. Chem. Phys.* **2009**, *130*, 125104.
(18) Hu, J.; Ma, A.; Dinner, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4615–4620.
(19) Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133–152.
(20) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
(21) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.
(22) So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521–1530.
(23) So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 5246–5256.
(24) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625–2627.
(25) Im, W.; Roux, B. *J. Mol. Biol.* **2002**, *319*, 1177–1197.
(26) Ma, A.; Nag, A.; Dinner, A. R. *J. Chem. Phys.* **2006**, *124*, 144911.
(27) De Alba, E.; Santoro, J.; Rico, M.; Jimenez, M. A. *Protein Sci.* **1999**, *8*, 854–865.
(28) Krivov, S. V.; Karplus, M. *J. Chem. Phys.* **2002**, *117*, 10894–10903.
(29) Dellago, C.; Bolhuis, P. G.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 9236–9245.

JP101476G