JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Quantifying Hub-like Behavior in Protein Folding Networks

Alex Dickson[†] and Charles L. Brooks, III*[,‡]

[†]Department of Chemistry, The University of Michigan, Ann Arbor, Michigan

[‡]Department of Chemistry and Biophysics Program, The University of Michigan, Ann Arbor, Michigan

ABSTRACT: The free energy landscape of a protein is a function of many interdependent degrees of freedom. For this reason, conceptual constructs (e.g., funnels) have been useful to visualize these landscapes. One relatively new construct is the idea of a hub-like native state that is the final destination of many noninterconverting folding pathways. This is in contrast to the idea of a single predominant folding pathway connecting the native state to a rapidly interconverting ensemble of unfolded states. The key quantity to distinguish between these two ideas is the connectivity of the unfolded ensemble. We present a metric to determine this connectivity for a given network, which can be calculated either from continuous folding trajectories, or a Markov model. The metric determines how often a region of space is used as an intermediate on transition paths that connect two other regions of space, and we use it here to determine how often two parts of the unfolded ensemble are connected directly versus how often these transitions are mediated by the native state.

1. INTRODUCTION

The folding process by which protein molecules reach a single well-defined native structure has captured the attention of researchers for decades.[1−5] Predictive knowledge of the folding process is alluring, as it would allow for the determination of a three-dimensional (3D) structure for any arbitrary sequence. The folding process also sheds light on how proteins misfold and aggregate, a process that can initiate cellular dysfunction and, ultimately, diseases such as Parkinson's and Alzheimer's.[6] Unfortunately, study of the connection between sequence and structure is complicated by both the large number of degrees of freedom available to a folding protein, and the long time scale (with respect to molecular vibrations) on which folding occurs.

Recent progress in molecular dynamics has led to routine simulations of trajectories on the order of microseconds (μs), and specialized hardware and software has been developed that can run trajectories over 1 ms long.[7] This allows for the observation of many folding and unfolding events in a single trajectory. A distributed computing approach (Folding@home) has also been employed to study folding proteins, with aggregate simulation times of over 30 ms.[8,9] It is a considerable challenge to extract meaning from the vast amount of data generated by these simulations.

Network models have been used to analyze protein folding data sets for more than a decade.[10−14] In these models, millions of configurations are clustered into a smaller number of states, which are the nodes of the network, and edges are placed between nodes that are connected by transitions in the underlying trajectories. Using a set of discrete conformation "letters" for each residue that depend on the local secondary structure, states can be defined as the concatenation of all "letters" in the molecule.[12,13] Alternatively, configurations can be clustered together that are close in space, as measured by the root mean squared distance between the atoms after alignment.[11,14] If the transitions between states can be approximated as Markovian (i.e., a trajectory will stay in each state long enough that future transitions do not depend on past transitions), then these networks, also called Markov state models (MSMs), can make quantitative connections with experiments.

Network models differ drastically from experiments in that they are able to simultaneously describe motions along every degree of freedom in the system, whereas experiments are usually limited to monitoring one or two degrees of freedom. As such, folding pathways that can be distinguished in network models can be degenerate along experimental observables.[15] If, additionally, these pathways occur with similar rates, then the folding kinetics in experiment will appear to be single exponential, which points to two-state folding behavior along a single pathway. Consequently, there are two different pictures of folding that are congruent with single exponential folding kinetics: one, where the protein starts in a rapidly mixing unfolded state, and then folds along a single pathway; the other, where folding begins in a set of kinetically separated unfolded states, and proceeds along separate pathways to a single hub-like native state. The key quantity to distinguish between these two pictures is the connectivity of the unfolded ensemble: Are different unfolded states connected directly, or are transitions between them mediated by the native state?

This question has previously been answered by comparing mean first passage times (MFPTs) between regions in the network. If the average MFPT from one unfolded state to another is significantly larger than the average MFPT from an unfolded state to the native state, then this was taken as evidence that the unfolded states are kinetically partitioned and are connected primarily through a hub-like native state.[13,15] However, as we show here, the MFPT between unfolded states can be dominated by a small fraction of long trajectories that are mediated by the native state, especially when the time scale of unfolding is slow. As a result, large MFPTs between unfolded states, as compared to the MFPT from unfolded to native, are not a sufficient metric to determine the connectivity of the unfolded ensemble.

We introduce a method to directly determine the fraction of transitions between two regions of configuration space that are mediated by a third region, given a MSM that describes the system. We demonstrate using a simple model system how the MFPT comparison metric is affected by the rate of unfolding. A new metric, "hub scores", is then introduced that determines how often a given region of space acts as a mediator between two other regions of space. This metric is then applied to configuration space networks

Received: June 26, 2012
Published: August 7, 2012

   dx.doi.org/10.1021/ct300537s | *J. Chem. Theory Comput.* 2012, 8, 3044−3052

from two systems: a model folding network with tunable connectivity in the unfolded ensemble, and a Gō model representation of Protein A.

2. THEORY

2.1. Mean First Passage Times between Unfolded States Can Be Dominated by Unfolding Times. We demonstrate this phenomenon analytically using a simple model system of N unfolded states ($\{U_i\}$) that are connected to a single folded state (F). We make a set of assumptions that make the model as simple as possible: all unfolded states are connected to each other, and transitions occur between them with a rate $a = k_{uu}/(N-1)$; all unfolded states can also make transitions to the folded state with a rate $b = k_{uf}$, and the rate of transition from the folded state to each unfolded state is $c = k_{fu}/N$. All transitions $i \to j$ occur with probability $P(i,j) = k_{ij}/\sum_{j'} k_{ij'}$ and waiting time $t(i,j) = 1/\sum_{j'} k_{ij'}$.[16] The system is depicted in Figure 1.

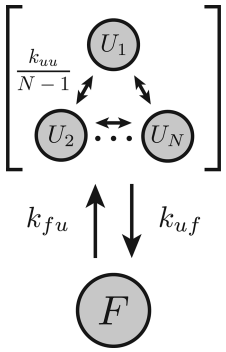

Figure 1. Schematic of the totally connected N-state folding model. All N unfolded states are connected and transition to another unfolded state with a rate $k_{uu}/(N-1)$ and to the folded state with a rate k_{uf}. The folded state is connected to all unfolded states. It makes transitions to an individual unfolded state with a rate k_{fu}/N, making the total rate of unfolding k_{fu}.

The expression for the MFPT from a state i to a state j (τ_{ij}) is given by

$$\tau_{ij} = P(i,j)t(i,j) + \sum_{j' \neq i,j} P(i,j')(t(i,j') + \tau_{j',j}) \tag{1}$$

which gives us a system of coupled equations for the τ variables. Due to the symmetry in our simple model, we have three equations and three unknowns:

$$\begin{aligned}
\tau_{UU} &= \frac{a}{b+(N-1)a}\left(\frac{1}{b+(N-1)a}\right) \\
&\quad + \frac{b}{b+(N-1)a}\left(\frac{1}{b+(N-1)a} + \tau_{FU}\right) \\
&\quad + \frac{(N-2)a}{b+(N-1)a}\left(\frac{1}{b+(N-1)a} + \tau_{UU}\right) \\
\tau_{FU} &= \frac{c}{Nc}\left(\frac{1}{Nc}\right) + \frac{(N-1)c}{Nc}\left(\frac{1}{Nc} + \tau_{UU}\right) \\
\tau_{UF} &= \frac{(N-1)a}{b+(N-1)a}\left(\frac{1}{b+(N-1)a} + \tau_{UF}\right) \\
&\quad + \frac{b}{b+(N-1)a}\left(\frac{1}{b+(N-1)a}\right)
\end{aligned} \tag{2}$$

The third equation is solved as $\tau_{UF} = {}^1/_b$, which is expected, since the rate constant for refolding is equal to $b = k_{uf}$ in every unfolded state. As can already be seen from the top equation, τ_{UU} depends, in part, on the time scale of unfolding, carried in the variable τ_{FU}. The top two equations can then be solved to yield

$$\tau_{UU} = \frac{1 + \frac{b}{Nc}}{\frac{b}{N} + a} \tag{3}$$

which can be compared to τ_{UF} using the ratio

$$\frac{\tau_{UU}}{\tau_{UF}} = \frac{N + N\phi}{1 + N\psi/(N-1)} \approx N\left(\frac{1+\phi}{1+\psi}\right) \tag{4}$$

where we introduce the variables $\phi = b/Nc = k_{uf}/k_{fu}$ and $\psi = (N-1)a/b = k_{uu}/k_{uf}$, and assumed N is large.

The variable ϕ is the ratio of folding to unfolding rates, and is hence the ratio of population in the folded and unfolded ensembles. For a typical protein folding system at physiological conditions, this ratio can be assumed to be much greater than 1. ψ is equal to the average number of transitions in the unfolded ensemble before folding. To put eq 4 into context, we consider the scenario where the folded state ensemble contains 90% of the probability in the system; therefore, $\phi \approx 10$. If the MFPTs from unfolded to unfolded states are on average 5 times longer than the MFPTs from unfolded to folded states ($\tau_{UU}/\tau_{UF} = 5$), then there is an average of $\psi \approx 2N$ transitions in the unfolded ensemble before folding. Furthermore, the fraction of unfolded to unfolded transitions that are mediated by the native state can be shown to be equal to $(N-1)/(2\psi + N - 1) \approx {}^1/_5$ (see Appendix A). So how can τ_{UF} be five times smaller than τ_{UU}, yet ${}^4/_5$ of the trajectories do not pass through F? This is possible because the contributions to the MFPT of the paths that go through the native state are weighted by their total time, which includes the time it takes for the protein to unfold.

2.2. Using "Hub Scores" to Describe the Connectivity of Unfolded States. To determine the connectivity of unfolded states in a network, we need to be able to determine how often the folded state mediates transitions between unfolded states. Because protein structure network models typically contain a large number of states (hundreds to thousands), we focus here on transitions between groups of states (macrostates), which capture large scale changes in the structure of the protein. We denote the transition probability from state i to state j as t_{ij}. The total transition probability from a macrostate A to another macrostate B is then

$$T_{AB} = \sum_{i \in A} \sum_{j \in B} \frac{w(i)}{W(A)} t_{ij} \tag{5}$$

where $w(i)$ is the weight (or equilibrium statistical probability) of state i, and $W(A) = \sum_{i \in A} w(i)$ is the weight of macrostate A. If a trajectory is in macrostate A, the probability of transitioning to B without returning to A is

$$P(A \to B) = \frac{\left(\sum_{i \notin A,B} q_i^{AB} T_{Ai}\right) + T_{AB}}{1 - T_{AA}} \tag{6}$$

where $T_{Aj} = \sum_{i \in A} w(i) t_{ij}/W(A)$, and q_i^{AB} is the committor function for state i between the basins A and B, in the forward direction. In other words, it is the fraction of trajectories that reach B before A, given that they are initialized in state i. The first term in the numerator describes the paths from A to B that occur

dx.doi.org/10.1021/ct300537s | J. Chem. Theory Comput. 2012, 8, 3044−3052

via some intermediate state i, and the second term describes paths that go directly from A to B.

Let us break the committor probability into two parts: $q_i^{AB} = q_i^{ABC^+} + q_i^{ABC^-}$ where the former part is the probability of reaching B before A, having gone through C, and the latter is the probability of reaching B before A, without having gone through C. We can now write the fraction of transition paths from A to B that are mediated by C as

$$h_C(A, B) = \frac{\sum_{i \notin A,B} q_i^{ABC^+} T_{Ai}}{\left(\sum_{i \notin A,B} q_i^{AB} T_{Ai}\right) + T_{AB}} \tag{7}$$

The average

$$H_C = \frac{1}{(N-1)(N-2)} \sum_{A \neq C} \sum_{B \neq A,C} h_C(A, B) \tag{8}$$

can be seen as a "hub score" for each C, since it determines how important C is to the transition path ensemble between other macrostates. The hub score takes on values in the range $[0,1]$, making it intuitive and easily comparable both across and between different networks.

When long, unbiased trajectories are available, the hub score for each macrostate can be computed simply by inspection. An algorithm to accomplish this is given in Appendix B. Such trajectories are not always available, especially when a configuration space network is built using enhanced sampling methods, or using a large number of short trajectories.[11] In this section, we show how to compute hub scores using transition rates between states of a Markov model. This necessitates calculation of the conditional committors $q_i^{ABC^-}$ and $q_i^{ABC^+}$ for every combination of macrostates A, B, and C, and every state $i \notin A, B, C$. Note that $q_i^{ABC^-} + q_i^{ABC^+} + q_i^{BAC^-} + q_i^{BAC^+} = 1$, since a trajectory starting in i can have four possible outcomes. For simplicity, we first describe a technique to calculate the committor function, q_i^{AB}, and then describe how we modify that technique to calculate the conditional committors, $q_i^{ABC^+}$ and $q_i^{ABC^-}$.

Consider the transition rate matrix $(\mathbb{R})$, whose elements r_{ij} are the rates of transitioning from state i to state j, and the diagonal elements are given by $r_{ii} = -\sum_j r_{ij}$, such that the sum of each column is equal to zero. Let $|X(t)\rangle$ be a normalized vector describing the population of each state at time t, so that the populations of each state at future times can be determined by $|X(t)\rangle = [\exp(\mathbb{R}t)]\,|X(0)\rangle$. Let $\mathbb{R}_{A0,B0}$ be a modified rate matrix where every column $j \in A$ and $k \in B$ are set to zero. This sets up probability traps at macrostates A and B, such that probability can enter, but it cannot leave. Diagonalization of $\mathbb{R}_{A0,B0}$ yields $\mathbb{A}\mathbb{V}\mathbb{A}^{-1}$, where $\mathbb{A}$ is a matrix of eigenvectors, $\mathbb{V}$ is a matrix with the eigenvalues on the diagonal, and the other elements equal to zero. Eigenvalues corresponding to columns $j \in A$ and $k \in B$, will be equal to zero, and the rest will be negative. At long times, only states $j \in A$ and $k \in B$ will be populated, and their relative populations are given by

$$\lim_{t\to\infty} |X(t)\rangle = \lim_{t\to\infty} \mathbb{A}e^{\mathbb{V}t}\mathbb{A}^{-1}|X(0)\rangle = \mathbb{A}\mathbb{I}_{A,B}\mathbb{A}^{-1}|X(0)\rangle \tag{9}$$

where $\mathbb{I}_{A,B}$ is the matrix with all diagonal elements i_{jj} and i_{kk} equal to one for all $j \in A$ and $k \in B$, and all others equal to zero. The committors are given by $q_i^{AB} = \sum_{k\in B}\langle k|\mathbb{A}\mathbb{I}_{A,B}\mathbb{A}^{-1}|i\rangle$. Note that all committors between A and B are thus obtained in a single matrix operation.

To distinguish between trajectories that have visited C and those that have not, we set up an extended state space where the states are denoted by two integers: $\vec{x} = (i,\alpha)$, with i being the state index used above and $\alpha = \{0,1\}$ denoting whether a trajectory has visited C $(\alpha = 1)$ or not $(\alpha = 0)$ (Figure 2). The extended rate

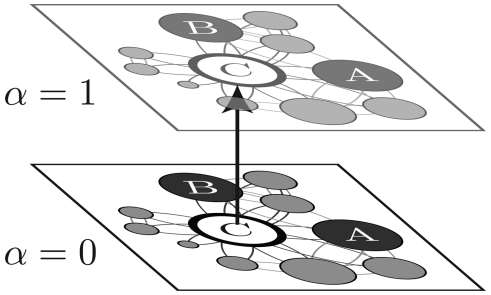

Figure 2. Schematic of the extended space used to calculate the conditional committor functions, $q_i^{ABC^+}$. Trajectories in the top level, $\alpha = 1$ have visited C, and trajectories in the bottom level $(\alpha = 0)$ have not.

matrix is $2n$ by $2n$, where n is the original number of states. The elements of the matrix are

$$r_{(i,\alpha)\to(j,\gamma)} = \begin{cases} r_{ij} & \text{if } \alpha = 0,\ \gamma = 0 \text{ and } j \notin C \\ r_{ij} & \text{if } \alpha = 0,\ \gamma = 1 \text{ and } j \in C \\ r_{ij} & \text{if } \alpha = 1 \text{ and } \gamma = 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

An extended space has been used previously to determine transition rates during enhanced sampling,[17,18] as well as to reproduce triplet−triplet energy transfer data.[14] We then set up probability sinks in $\mathbb{R}$ at A and B by setting all columns $\vec{i} = (i,\alpha)$, such that $i \in A$ or B, to zero, for both values of α. This matrix is diagonalized, and the corresponding matrix $\mathbb{A}\mathbb{I}_{\vec{A},\vec{B}}\mathbb{A}^{-1}$ reveals the conditional committors:

$$\begin{aligned} q_i^{ABC^+} &= \sum_{k\in B} \langle k, \alpha = 1|\mathbb{A}\mathbb{I}_{\vec{A},\vec{B}}\mathbb{A}^{-1}|i, \alpha = 0\rangle \\ q_i^{ABC^-} &= \sum_{k\in B} \langle k, \alpha = 0|\mathbb{A}\mathbb{I}_{\vec{A},\vec{B}}\mathbb{A}^{-1}|i, \alpha = 0\rangle \\ q_i^{BAC^+} &= \sum_{j\in A} \langle j, \alpha = 1|\mathbb{A}\mathbb{I}_{\vec{A},\vec{B}}\mathbb{A}^{-1}|i, \alpha = 0\rangle \\ q_i^{BAC^-} &= \sum_{j\in A} \langle j, \alpha = 0|\mathbb{A}\mathbb{I}_{\vec{A},\vec{B}}\mathbb{A}^{-1}|i, \alpha = 0\rangle \end{aligned} \tag{11}$$

To get hub scores for all N macrostates in the network requires $N(N-1)(N-2)/2$ matrix diagonalizations. We use small numbers of macrostates (ranging from 4 to 10), for networks that range in size from 349 to 4651. For larger numbers of macrostates, single values of $h_C(A,B)$ can be useful instead of full hub scores.

3. APPLICATIONS

3.1. Variably Connected *N*-State Folding Model. To begin to explore these concepts on more realistic networks, we consider here a generalization of the totally connected model studied analytically above. We introduce a connectivity parameter c_{uu} that controls the number of connections between

dx.doi.org/10.1021/ct300537s | J. Chem. Theory Comput. 2012, 8, 3044−3052

unfolded states and varies between $c_{uu} = 0$ (zero connectivity) and $c_{uu} = 1$ (total connectivity). This allows us to observe how the hub scores of the folded and unfolded states depend on the connectivity of the unfolded ensemble. We build a number of random graphs for each value of c_{uu}, and determine hub scores for each node, as well as mean first passage times between the nodes. The graphs are built as follows. First, all unfolded states are connected to the native state. Second, the total number of connections between unfolded states is determined as $N_{con} = c_{uu}N(N-1)/2$, rounded down to the nearest integer. Connections are then placed between unfolded states in a random fashion, until a total of N_{con} connections are made. The nondiagonal elements of the transition rate matrix $\mathbb{R}$ are set as follows:

$$r_{ij} = \begin{cases} k_{fu}/N & \text{if } i = 0 \text{ and } j \neq 0 \\ k_{uf} & \text{if } i \neq 0 \text{ and } j = 0 \\ k_{uu}/c(i) & \text{if } i, j \neq 0, \text{ and } i \text{ and } j \text{ are connected} \end{cases} \quad (12)$$

where $i = 0$ is the folded state, and $c(i)$ is the number of connections from state i to other unfolded states. The diagonal elements of the rate matrix are again defined as the opposite of the sum of the nondiagonal elements. We use $N = 10$ unfolded states, with each state defining its own macrostate.

An ensemble of 20 graphs were computed for many different values of c_{uu}. We choose values of k_{uu}, k_{uf}, and k_{fu} to be 0.92625, 0.025, and 0.001. For these values, unfolding is the slowest process. Folding is the second slowest process, occurring 25 times faster than unfolding, and transitions between unfolded states occur with a total rate that is ∼1000 times faster than unfolding. Figure 3 shows example graphs with low and high

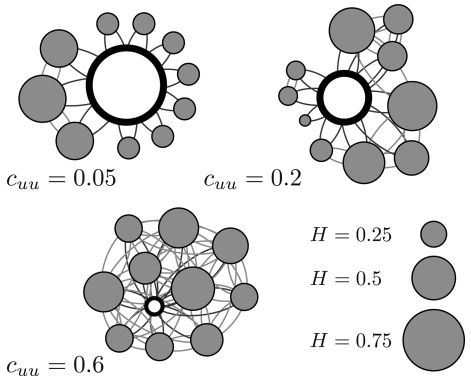

Figure 3. Examples of networks with three different connectivity values. In each network, the folded state is shown in white with thick black outline, and the unfolded states are shown in gray. The diameters of the circles representing the states are proportional to the hub scores for each state. Each network is composed of 10 unfolded states and a single folded state. The graph representation were created using Gephi.[19]

values of the connectivity parameter c_{uu}. For $c_{uu} = 0.05$, the unfolded states are mostly unconnected, and the average hub score for the native state ($\langle H_0 \rangle$) is 0.95, indicating that 95% of unfolded transitions go through the native state. As c_{uu} grows to 0.2, most of the unfolded states agglomerate into a single sparsely connected component, and $\langle H_0 \rangle = 0.43$, which is slightly above the average hub score in the network. For $c_{uu} = 0.6$, the unfolded states are densely connected, and $\langle H_0 \rangle = 0.11$, which is much lower than the average hub score of the network.

Figure 4 shows the average hub scores for the folded and unfolded states as a function of c_{uu}. For $c_{uu} < 0.2$, the hub score for

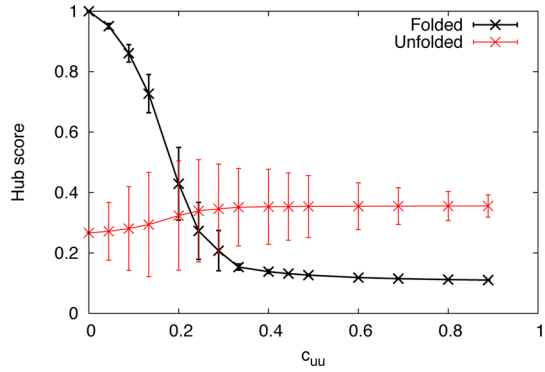

Figure 4. Hub scores of the variably connected N-state folding model as a function of c_{uu}. The error bars show the standard deviation of the hub scores upon different realizations of the graph structures at constant c_{uu}. These deviations are the largest at intermediate values of c_{uu} (around 0.2), where there is greater variability in graph structure, since a large connected component can coexist with unfolded nodes that are only connected to the folded state.

the folded state is significantly larger than the average hub score of the unfolded states. For $c_{uu} > 0.3$, the folded state hub score is less than the average hub score of the unfolded states. Using this metric, we can distinguish between hub-like ($c_{uu} < 0.2$) and non-hub-like ($c_{uu} > 0.3$) behaviors of the variably connected N-state folding model, for this set of rate constants. For contrast, the average MFPT between unfolded states, divided by the MFPT for folding is shown in Figure 5. The mean first passage times are

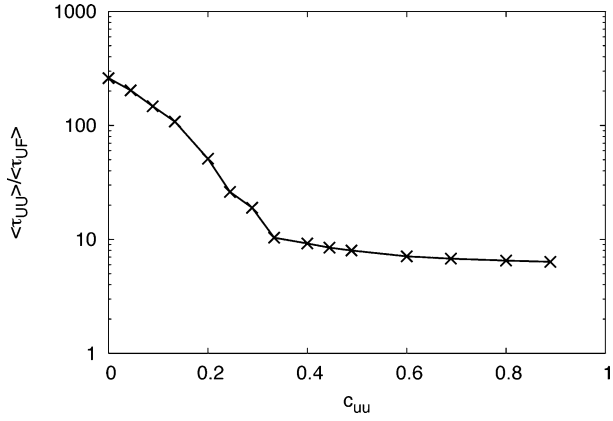

Figure 5. MFPT for transitions between unfolded states divided by the MFPT for transitions to the folded state in the variably connected N-state folding model.

computed in a similar fashion to the hub scores (see Appendix C for details). Using these parameters, we find that $\tau_{UU} > 5\tau_{UF}$ for all values of c_{uu}. This emphasizes the point that mean first passage times between unfolded states can have difficulties determining hub-like behavior of the folded state, particularly when the unfolding rate is slow.

The folding model presented here was developed to illustrate the difference between these two metrics (hub scores and MFPT ratios). There are a number of important differences between this model and realistic protein structure networks. First, in this model, all unfolded states are connected to the folded state, making it possible to reach the folded state in one step. Second,

dx.doi.org/10.1021/ct300537s | J. Chem. Theory Comput. 2012, 8, 3044−3052

the connections between unfolded states are determined randomly, and there is no heterogeneity in transition rates between unfolded states. In real protein structure networks, connections between unfolded states are governed by the structure of free energy barriers in phase space. For this reason, we apply this analysis to more realistic protein structure networks in the next section.

3.2. Gō Model of Protein A. The recombinant B domain of staphylococcus protein A (PDB id: 1bdc[20]) has 60 residues that form three helices. We use a coarse-grained representation (a Gō model[21]) in which each residue is described by a single sphere, and there are attractive interactions between residues that are close together in the native state. The model is constructed using the Gō model builder offered by the MMTSB toolset.[22,23] We build networks of states in configuration space by running a series of long trajectories that contain many folding and unfolding events. Dynamics are performed using CHARMM.[24] A Langevin integrator is used with a 10 fs time step, and a friction coefficient of 5 ps^{-1}. A 10−12 van der Waals potential is used. The SHAKE algorithm is used to constrain all bonds, with a tolerance of 10^{-6}. Configurations are saved every 100 ps. Clustering and construction of the Markov state model (MSM) is performed with MSMBuilder2,[25] using the hybrid k-centers and k-medoid clustering algorithm. We choose a lag time based on the relaxation time scales of the model and find 2 ns to be sufficiently long that the slowest three relaxation time scales are approximately constant as a function of lag time.[25] We symmetrize the resulting transition count matrix as $t_{ij}^s = (t_{ij} + t_{ji})/2$; this is equivalent to including the reverse of each trajectory as well. We found a clustering radius of 8 Å to be appropriate for our model and find that MSMs built using clustering radii of 6 Å and 7 Å yield similar results to those presented here.

Our goal is to quantify the connectivity of the unfolded state and to see if this result is robust to changes in temperature, as well as changes in the definition of the macrostates. To build configuration space networks, we run 4 trajectories, each 10 μs in length, at each of the temperatures: 340, 360, and 380 K. These temperatures are in the neighborhood of the folding temperature obtained from the measured heat capacity profile from all-atom MD simulations (362 K).[26] The total numbers of unfolding transitions that are observed in these trajectory sets are 2, 20, and 47, for T = 340, 360, and 380 K, respectively. We define an unfolding transition as the RMSD from native (measured over the helical regions: residues 11 to 56) reaching at least 12 Å, and subsequently returning to below 1 Å. The average folding times measured were 0.32 μs (0.03), 0.32 μs (0.07) and 0.53 μs (0.09) with increasing temperature, where the numbers in parentheses are standard errors. These fast folding times are consistent with what we would expect from Gō model simulations.

To compare hub scores across networks built at different temperatures, we need to define our macrostates in a way that does not depend on a specific network representation. We define our macrostates using two distances: d_{12} is defined as the distance between residues 15 and 30, which are in the middle of helices 1 and 2, respectively. Similarly, d_{23} is defined as the distance between residues 30 and 46, which are in the middle of helices 2 and 3. The macrostate index ($M(i)$) of a given node i in our network is then determined by

$$M(i) = \begin{cases} 1 & \text{if } d_{12} < 17 \text{ Å and } d_{23} < 14 \text{ Å} \\ 2 & \text{if } d_{12} > 17 \text{ Å and } d_{23} < 14 \text{ Å} \\ 3 & \text{if } d_{12} < 17 \text{ Å and } d_{23} > 14 \text{ Å} \\ 4 & \text{if } d_{12} > 17 \text{ Å and } d_{23} > 14 \text{ Å} \end{cases} \quad (13)$$

where the cutoffs are chosen by hand to maximize separation of the macrostates.

The resulting networks for each temperature are shown in Figure 6. The size of each node is proportional to its statistical weight, and the color shows the macrostate assignment given by eq 13. As expected, the number of states visited (N) grows as temperature is increased: N = 115, 349, and 577 for T = 340, 360, and 380 K. The total population in the folded basin (RMSD to native < 8 Å) is equal to 98%, 81%, and 33% for T = 340, 360, and 380 K, where the RMSD is again measured using residues 11 to 56.

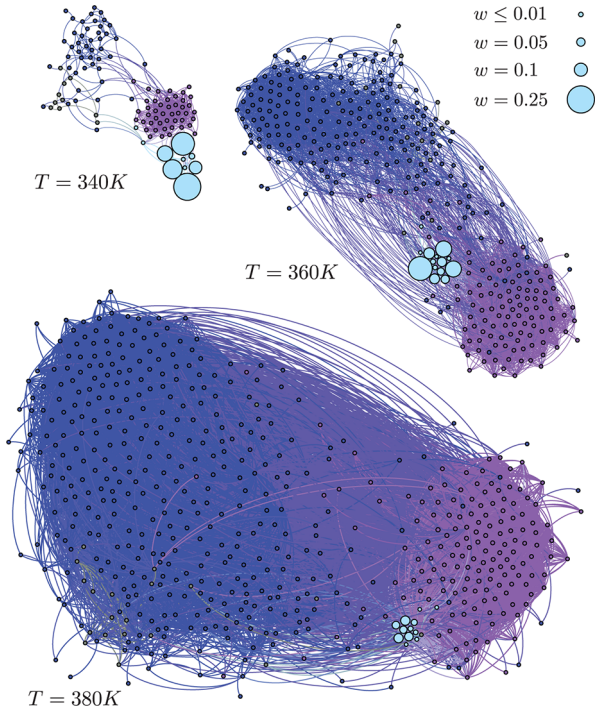

Figure 6. Configuration space networks for the Gō model of Protein A obtained at three different temperatures. The size shows the statistical weight of each node, and the color shows the macrostate (light blue = 1, gray = 2, purple = 3, and blue = 4). The macrostate assignments are determined using eq 13 with values of d_{12} and d_{23} that are averages over five randomly sampled configurations belonging to each node. To aid in visualization, an edge between nodes i and j is only shown if the transition between i and j is recorded at least twice, and following this procedure, any node that is not connected to the giant component by an edge is discarded. The legend in the upper right shows weight values for various node sizes. The sizes of each node are adjusted appropriately so that they can be compared across graphs. The network representations were created using Gephi.[19]

Figure 7 shows representative structures in each of the four regions, as well as their populations as a function of temperature. From Figures 6 and 7, we can see that macrostate 3 has a greater population than macrostate 2 at all temperatures studied here. Therefore, these results predict that the dominant folding pathway is where helices 1 and 2 come into contact first, and helix 3 joins subsequently. This is in agreement with previous work on Protein A.[27]

The hub score components $h_C(A,B)$ from eq 7 are shown for each temperature in Figure 8. These were computed using statistics from the Markov model. We also computed these values directly from the trajectories and found the two sets to agree to within 1.2%, 2.0%, and 11% for T = 380, 360, and 340 K (see Figure 9). We attribute the larger error for T = 340 K to the

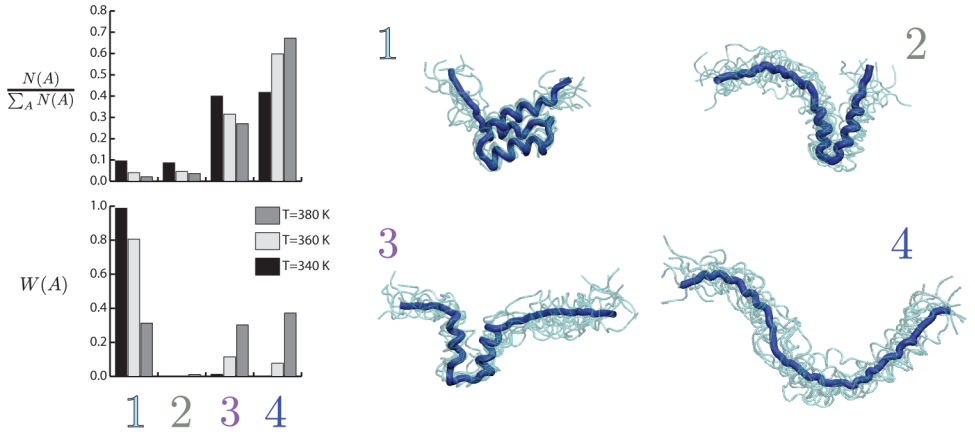

Figure 7. Four macrostates based on distances d_{12} and d_{23}. Top left: the fraction of nodes belonging to each macrostate, for each of the temperatures studied. Bottom left: the statistical weight of each macrostate. Right: For each macrostate, an ensemble of 10 randomly chosen structures are shown in transparent light blue, and their average structure is shown in dark blue. These structures were taken from the T = 360 K ensemble. Each structure is shown with the N terminus on the left. The colors of the macrostate labels correspond to the colors of the macrostates in Figure 6.

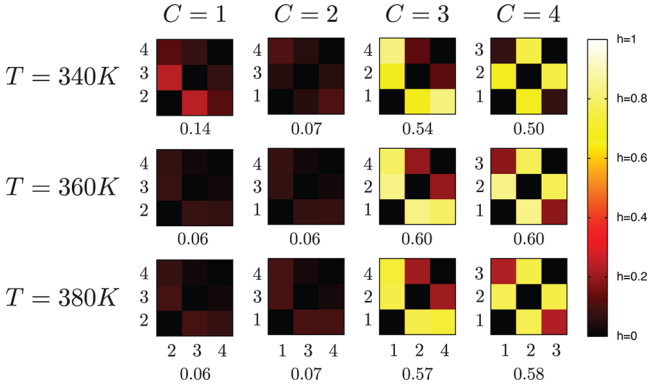

Figure 8. Hub score components $h_C(A,B)$ for each temperature. Each matrix shows $h_C(A,B)$ values for all A and B not equal to C, the A and B indices are shown along the left and bottom. Note that since $h_C(A,B) = h_C(B,A)$, these matrices are symmetric about the diagonal. The hub scores H_C are shown below each matrix.

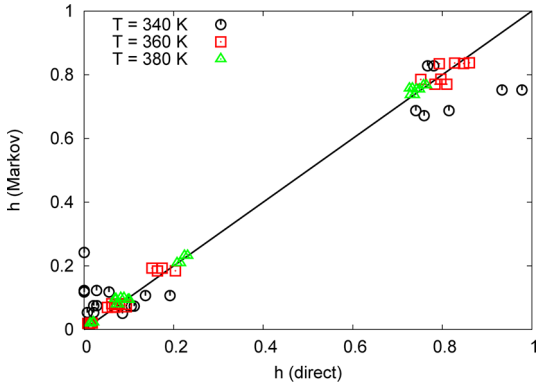

Figure 9. Comparison of hub score components ($h_C(A,B)$) calculated directly from trajectories ("direct") and from the Markov models ("Markov") for each temperature for the Gō model of Protein A. The black line marks perfect agreement.

comparatively few unfolding events observed in trajectories at that temperature. For the range of temperatures studied here, the overall results are consistent: macrostates 1 and 2 show poor hub-like behavior, and macrostates 3 and 4 show strong hub-like behavior. Macrostate 3 is on an average of 78% of 1 ↔ 4 pathways and 76% of 1 ↔ 2 pathways. $h_3(2,4)$ shows a dependence on temperature, increasing from 0.12 to 0.21 for T = 340 K to T = 380 K. Macrostate 4 is on an average of 76% of 1 ↔ 2 paths and 76% of 2 ↔ 3 paths, and similarly, its lowest contribution $h_4(1,3)$ shows a temperature dependence, increasing from 0.07 to 0.23 from T = 340 K to T = 380 K. The largest hub score component for the folded macrostate is $h_1(2,3)$ = 0.24 for T = 340 K, and this decreases to $h_1(2,3)$ = 0.09 at T = 380 K. This makes sense, as 2 and 3 are both partially folded but in different parts of the molecule; so by construction, they can either be connected by fully folding (via macrostate 1) or fully unfolding (via macrostate 4). It is intuitive that the 2 ↔ 1 ↔ 3 pathway would become more likely as the temperature is lowered, since 1 is lower in entropy, which is the behavior we observe here.

Using hub scores, we can see that at all three temperatures examined here, the folded basin does not act as a hub. This result is expected for trajectories based on a Gō model system, as we can expect a high degree of connectivity in the unfolded ensemble due to the absence of nonnative interactions. For comparison, we also compute the mean first passage time between macrostates for each temperature. We find that the ratio $\langle\tau_{UU}\rangle/\langle\tau_{UF}\rangle$ has a strong temperature dependence, and takes on the values 71, 1.8, and 0.33 for T = 340, 360, and 380 K, respectively. This is primarily due to the increase in mean folding time as T increases. The T = 340 K results clearly reveal the difference between using hub scores and using the MFPT ratio to determine the hub like behavior of the folded state: $\langle\tau_{UU}\rangle/\langle\tau_{UF}\rangle$ = 71 predicts hub-like behavior, while the hub score of the folded region (H_1 = 0.14) does not. At higher temperatures, these two metrics come into agreement, in the sense that they both predict that the folded state does not display hub-like behavior.

To explore how hub scores depend on the particular definition of the macrostates, we use an alternative clustering method: Perron Cluster Cluster Analysis (PCCA),[28] implemented in MSMBuilder2.[25] PCCA defines macrostates using the kinetic connectivity of the states and divides a Markov model into any number of N macrostates. We define three sets of macrostates for the T = 360 K trajectories, using N = 4, 6, and 8 (Figure 10). The N = 4 set is similar to that determined using eq 13, except macrostate 2 now contains many nodes that previously belonged to macrostate 4. We also used PCCA+,[29] an improved version of PCCA that has been shown to be more robust, and obtained sets of macrostates similar to those shown in Figure 10.

dx.doi.org/10.1021/ct300537s | J. Chem. Theory Comput. 2012, 8, 3044−3052

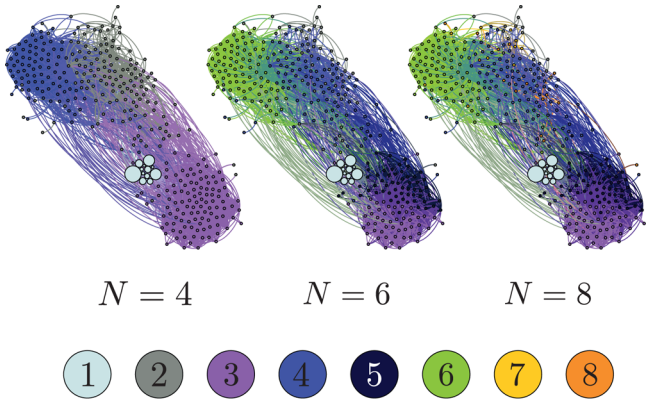

Figure 10. Configuration space networks for the Gō model of Protein A determined by PCCA. N is the number of macrostates. The colored circles along the bottom show the corresponding numerical indices for each macrostate, to use as a reference for Figure 11.

Figure 11 shows the hub scores for these macrostates, which are consistent in their prediction that the folded state does not act as a hub. The hub scores for the folded state also show reasonable quantitative agreement with the $H_1 = 0.06$ result from the $T = 360$ K ensemble using the eq 13 macrostates. Quantitative agreement of hub scores between models that use network-specific clustering methods like PCCA is not expected, since the macrostate assignments are not based on an constant partitioning, as in eq 13. However, these results are encouraging, since they demonstrate that the main conclusion is insensitive to a particular choice of macrostates.

4. DISCUSSION

We presented hub scores, a quantitative description of mediation probabilities in a set of macrostates of a configuration space network. We found in both model networks and a Gō model system that MFPT ratios and hub scores differed in their predictions of hub-like behavior, with MFPT ratios predicting stronger hub-like behavior of the native state. We expect the formalism introduced here will be a useful tool to characterize connectivity in a broad range of protein configuration space networks.

As shown, the nature of a protein structure network is temperature dependent. The temperature affects not only the relative population of each state, but also local connectivity (between nodes) and global connectivity (between macrostates). For the Gō model of protein A, we found that the hub score of the native state decreased with increasing temperature, which is consistent with the idea that, at high temperature, high entropy pathways would be preferred over low entropy pathways that involve the native state. We expect this trend to continue as the temperature is lowered further, although it becomes more difficult to obtain enough unfolding trajectories as the temperature lowers. It will be interesting to see if this trend generalizes to all-atom systems as well.

The absence of nonnative interactions in a Gō model prohibits the extension of these results to more realistic, all-atom systems. It has been previously shown by Pande[30] using a model Hamiltonian that nonnative interactions are the genesis of hub-like behavior of the native state in protein structure networks. Our results agree, in that for our Gō model system we observe weak hub-like behavior of the native state. Preliminary results on an all-atom system, the villin headpiece, indeed reveal much higher hub scores for the native state (~ 0.8) than observed here for the Gō model. However, they are sensitive to the number of macrostates used, as well as the method used to generate them (PCCA, or PCCA+). It will be interesting to apply the tool developed here to a broad range of protein structure networks at different simulation conditions to see if hub-like behavior of the native state is a general property of all proteins.

APPENDIX A

Derivation of the Fraction of Mediated Trajectories in the Totally Connected Model

For the totally connected model, the fraction of paths between two unfolded states that are mediated by the folded state is given by eq 7:

$$h_F(i,j) = \frac{q_k^{ijF^+}\left[\frac{(N-2)a}{(N-1)a+b}\right] + q_F^{ij}\left[\frac{b}{(N-1)a+b}\right]}{q_k^{ij}\left[\frac{(N-2)a}{(N-1)a+b}\right] + q_F^{ij}\left[\frac{b}{(N-1)a+b}\right] + \frac{a}{(N-1)a+b}} \tag{14}$$

where i, j, and k denote particular unfolded states. The conditional committors can be determined as follows.

$$q_k^{ijF^-} = \frac{a}{(N-1)a+b} + \frac{(N-3)a}{(N-1)a+b} q_k^{ijF^-} \tag{15}$$

yields $q_k^{ijF^-} = a/(2a+b)$. Similarly,

$$q_k^{ijF^+} = \frac{b}{(N-1)a+b} q_F^{ij} + \frac{(N-3)a}{(N-1)a+b} q_k^{ijF^+} \tag{16}$$

where $q_F^{ij} = {}^1/_2$, by symmetry. This equation yields $q_k^{ijF^+} = b/(4a+2b)$. Note that $q_k^{ij} = q_k^{ijF^+} + q_k^{ijF^-} = {}^1/_2$, which is required by symmetry.

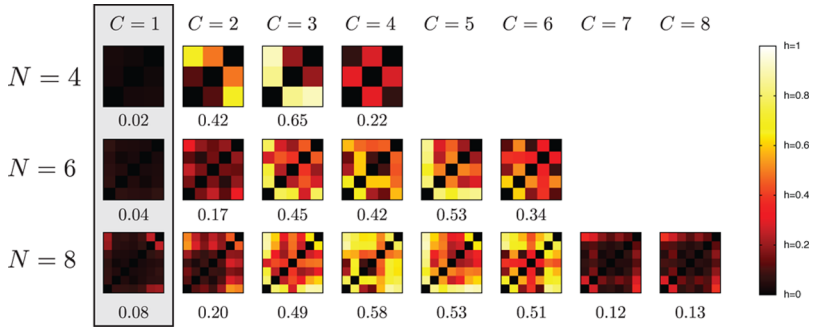

Figure 11. Hub scores for macrostates determined by PCCA. The individual matrix elements are arranged as in Figure 8: in numerical order, skipping the value of C, which is the column index. The gray rectangle shows the macrostates that correspond to the folded basin.

dx.doi.org/10.1021/ct300537s | J. Chem. Theory Comput. 2012, 8, 3044−3052

Substitution of these values into the equation for $h_F(i,j)$ gives

$$h_F(i,j) = \frac{\frac{b}{4a+2b}(N-2)a + \frac{b}{2}}{\frac{(N-2)a}{2} + \frac{b}{2} + a} = \frac{b}{2a+b} = \frac{N-1}{2\psi + N - 1} \quad (17)$$

APPENDIX B

Determination of Hub Scores by Inspection of Continuous Folding Trajectories

When continuous folding trajectories are available, hub scores can be determined directly. This method is ideal since it does not involve a Markovian assumption, and it does not involve diagonalization of many transition state matrices. An efficient algorithm to accomplish this is as follows. We will use the array $h(A,B,C)$ to count the number of trajectory segments going from A to B that pass through C, and the array $t(A,B)$ to count the total number of trajectory segments going from A to B. The array $r(i)$ stores the macrostate index assigned to each state i.

The input to the algorithm is simply the sequence of states visited along a trajectory, and the array "latest(A)" stores the latest time point that macrostate A was visited.

```
set all h(A,B,C), latest(i) = 0

set all t(A,B) = 0


time = 1

while (i = nextstate())
      B = r(i)
      oldlatest = latest(B)
      latest(B) = time
      foreach A != B
            if latest(A) > oldlatest
                  t(A,B)++
                  foreach C != A and C !=B
                        if latest(C) > latest(A)
                              h(A,B,C)++
                        end if
                  end foreach
            end if
      end foreach
      time++
end while
```

At the end, the hub score for each node can be computed by eq 8, where $h_C(A,B) = h(A,B,C)/t(A,B)$.

APPENDIX C

Calculation of Mean First Passage Times Using Transition Count Matrices

Here, we describe how we determine the mean first passage time from every macrostate A to macrostates B, given the transition count matrix of a system. A rate matrix with every column $j \in B$ set to zero is constructed ($\mathbb{R}_{B0}$) and diagonalized such that $\mathbb{R}_{B0} = \mathbb{A}\mathbb{V}\mathbb{A}^{-1}$. We then determine the matrix $\mathbb{A}[\exp(\mathbb{V}t)]\mathbb{A}^{-1}$ for many values of t. Here we space these t values geometrically, with a factor of 1.2 separating adjacent values of t. We then obtain the fraction of trajectories, starting in A that will arrive in state B before a time t as

$$P_{AB}(t) = \sum_{i\in A}\sum_{j\in B} \frac{w_i}{W_A} \langle i|\mathbb{A}e^{\mathbb{V}t}\mathbb{A}^{-1}|j\rangle \quad (18)$$

The mean first passage time from A to B is then computed using a discretization of the integral:

$$\tau_{AB} = \int_0^\infty \frac{dP_{AB}(t)}{dt} t\, dt \quad (19)$$

AUTHOR INFORMATION

Corresponding Author

*E-mail: brookscl@umich.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to Vijay Pande and Kyle Beauchamp for sharing Markov state models for the villin headpiece and for help with MSMBuilder2. We thank the NIH through grant GM037554 for funding.

REFERENCES

(1) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199−203.
(2) Dill, K. A. *Biochemistry* **1990**, *29*, 7133−7155.
(3) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins* **1995**, *21*, 167−195.
(4) Dill, K. A.; Chan, H. S. *Nat. Struct. Biol.* **1997**, *4*, 10−19.
(5) Dobson, C. M. *Nature* **2003**, *426*, 884−890.
(6) Selkoe, D. J. *Nature* **2003**, *426*, 900−904.
(7) Shaw, D. E.; et al. *Commun. ACM* **2008**, *51*, 91−97.
(8) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903−1904.
(9) Voelz, V. A.; Jäger, M.; Yao, S.; Zhu, L.; Waldauer, S. A.; Bowman, G. R.; Friedrichs, M.; Bakalin, O.; Lapidus, L. J.; Shimon, W.; Pande, V. S. *J. Am. Chem. Soc.* **2012**, *134* (30), 12565−12577.
(10) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740−744.
(11) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011−19016.
(12) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299−306.
(13) Muff, S.; Caflisch, A. *Proteins* **2008**, *70*, 1185−1195.
(14) Beauchamp, K. A.; Ensign, D. L.; Das, R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12734−12739.
(15) Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *107*, 10890−10895.
(16) Gillespie, D. T. *J. Phys. Chem.* **1977**, *81*, 2340−2361.
(17) Dickson, A.; Warmflash, A.; Dinner, A. R. *J. Chem. Phys.* **2009**, *131*, 154104.
(18) Dickson, A.; Maienschein-Cline, M.; Tovo-Dwyer, A.; Hammond, J. R.; Dinner, A. R. *J. Chem. Theory Comput.* **2011**, *7*, 2710−2720.
(19) Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *International AAAI Conference on Weblogs and Social Media*, California, May 17−20, 2009; Hamilton, M., Ed.; AAAI Press: Palo Alto, CA, 2009.

(20) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. *Biochemistry* **1992**, *31*, 9665−9672.
(21) Abe, H.; Go, N. *Biopolymers* **1981**, *20*, 1013−1031.
(22) Karanicolas, J.; Brooks, C. L., III *Protein Sci.* **2002**, *11*, 2351−2361.
(23) Karanicolas, J.; Brooks, C. L., III *J. Mol. Biol.* **2003**, *334*, 309−325.
(24) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(25) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(26) Lei, H.; Wu, C.; Wang, Z.; Zhou, Y.; Duan, Y. *J. Chem. Phys.* **2008**, *128*, 235105.
(27) Boczko, E. M.; Brooks, C. L., III *Science* **1995**, *269*, 393−396.
(28) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear Algebra Appl.* **2000**, *315*, 39−59.
(29) Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161−184.
(30) Pande, V. S. *Phys. Rev. Lett.* **2010**, *105*, 198101.