Graph representation of protein free energy landscape

Minghai Li,¹ Mojie Duan,¹ Jue Fan,^{1,a)} Li Han,² and Shuanghong Huo^{1,b)}

¹*Gustaf H. Carlson School of Chemistry and Biochemistry, Clark University, 950 Main Street, Worcester, Massachusetts 01610, USA*

²*Department of Mathematics and Computer Science, Clark University, 950 Main Street, Worcester, Massachusetts 01610, USA*

(Received 20 August 2013; accepted 28 October 2013; published online 14 November 2013)

The thermodynamics and kinetics of protein folding and protein conformational changes are governed by the underlying free energy landscape. However, the multidimensional nature of the free energy landscape makes it difficult to describe. We propose to use a weighted-graph approach to depict the free energy landscape with the nodes on the graph representing the conformational states and the edge weights reflecting the free energy barriers between the states. Our graph is constructed from a molecular dynamics trajectory and does not involve projecting the multi-dimensional free energy landscape onto a low-dimensional space defined by a few order parameters. The calculation of free energy barriers was based on transition-path theory using the MSMBuilder2 package. We compare our graph with the widely used transition disconnectivity graph (TRDG) which is constructed from the same trajectory and show that our approach gives more accurate description of the free energy landscape than the TRDG approach even though the latter can be organized into a simple tree representation. The weighted-graph is a general approach and can be used on any complex system.

© 2013 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4829768>]

I. INTRODUCTION

The thermodynamics and kinetics of protein folding is determined by the free energy landscape. With the growth of computer power, especially the development of GPUs¹ and the special-purpose supercomputer, Anton,² for the massive acceleration of molecular dynamics (MD) simulations, microsecond time scale simulations of proteins with explicit solvent becomes routine, even the millisecond time scale is approachable.³ With the enormous amount of collected protein conformations, the subsequent question is how to extract thermodynamic and kinetic information from the trajectories.

Network/graph-based approaches have been used to describe (free) energy landscape of peptides/proteins and their conformational changes.^{4–12} For example, the transition disconnectivity graph (TRDG) was first used to map the minima and transition states of a tetrapeptide by Czereminski and Elber,¹³ later, the relations between kinetic connectivity and spatial proximity of the same peptide were investigated by Becker and Karplus using TRDG and master equation.¹⁴ Subsequently, TRDG was successfully used to visualize the energy landscape of Lennard–Jones clusters, C₆₀, and water clusters,¹⁵ as well as of proteins with static pulling forces.¹⁶ Recently, the TRDG approach was extended to visualize free energy landscape by post-processing MD trajectories.^{17–19} Alternatively, the TRDG that depicts the free energy landscape can be constructed from kinetic transition networks based on geometry optimization and using harmonic densities of states to estimate the relative free energies of minima and transi-

tion states.²⁰ TRDG was also used as the “gold standard” to evaluate the calculation methods of configurational entropy²¹ and the quality of dimensionality reduction.²² However, the TRDG constructed from MD trajectories has its own limitations as shown in the section of Results, therefore, we propose to construct a weighted graph ($G(V, E)$) from the MD trajectories to depict the free energy landscape using discrete transition-path theory as a tool to analyze Markov state models.^{23–25} Hereinafter, TRDG refers to the one constructed from MD trajectories.

A graph contains a vertex set V (or node set) and a binary relation E (edge set) on V . The binary relation is whether or not there is an edge between two particular nodes. For a weighted graph, an edge weight is associated with each edge. To describe the free energy landscape, a free energy minimum or a collection of multi-minima (so-called superbasin) is represented by a node on the graph and the apparent barrier that separates each pair of nodes is associated with an edge weight. Fig. 1 gives an illustration on how to construct the graph. To calculate the apparent free energy barrier between a given pair of nodes, i.e., a and d , we first define a as the source (or reactant) and d as the sink (or product), then a flow network is constructed with the edge weights associated with the net flux (f_{ij}^+) contributing to the total flow from the source to the sink. The total net flow leaving the source and entering the sink is related to the apparent free energy barrier between them. The physical meaning of the total net flow from the source to the sink is the number of transitions from the reactant to the product per unit time. We then go through the same computational process for all of the pairs of nodes (of interest) to obtain the apparent free energy barriers and build the weighted graph that depicts the free energy landscape of the system (as illustrated in Fig. 1). The computational process of the total net

^{a)}Present address: Department of Plant Biology, Carnegie Institution for Science, 260 Panama St., Stanford, California 94305, USA.

^{b)}Author to whom correspondence should be addressed. Electronic mail: shuo@clarku.edu

FIG. 1. Schematic illustration of graph representation of the free energy landscape. (Left panel) Flow networks. The nodes are denoted as circles. The size of a node denotes the free energy of the state. The source (s , in red) and sink (t , in blue) are specified as a different pair of nodes in different flow networks. The edges have directions and their associated weights are the net flux (f_{ij}^+) contributing to the total flow from the source to the sink. Note that for different flow network with different source and sink, the values of f_{ij}^+ are different. The thickness of the edge denotes the edge weight. (Middle panel) The total net flow (F_{st}) leaving the source and entering the sink is related to the apparent free energy barrier ($F_{st}^\ddagger$). (Right panel) The final graph is built to describe the free energy landscape. The size of a node denotes the free energy of the state and the thickness of an edge is the free energy barrier between its two incident nodes. The thicker the edge, the lower the free energy barrier.

flow was used to study protein folding.²⁴ In this work, our interest is the free energy barrier between all pairs of nodes, not limited to the unfolded state and the folded state.

Our graph is different from TRDG because only the net flux (net transitions) between nodes (free energy minimum) i and j that contributes to the total flow (overall transitions) from the source (reactant) to the sink (product) is of interest in our graph, which were computed using the transition-path theory (TPT).²⁶ Consider a transition between minima i and j which are neither reactant nor product, there are four possible scenarios: (1) *reactant* $\rightarrow \dots \rightarrow i \rightarrow j \dots \rightarrow$ *product*; (2) *reactant* $\rightarrow \dots \rightarrow i \rightarrow j \dots \rightarrow$ *reactant*; (3) *product* $\rightarrow \dots \rightarrow i \rightarrow j \dots \rightarrow$ *reactant*; (4) *product* $\rightarrow \dots \rightarrow i \rightarrow j \dots \rightarrow$ *product*, among which only in the first scenario the i to j transition contributes to the reaction. TPT considers only the net flux between nodes that contributes to the reaction, while TRDG counts all the i to j transitions. Fig. 2 gives an extreme example: a pair of minima is separated by one barrier (Fig. 2(a)) or by two barriers of the same height (Fig. 2(b)). For the situation illustrated in Fig. 2(b), TRDG gives free energy barrier as $F_{ac}^\ddagger = \max(F_{ab}^\ddagger, F_{bc}^\ddagger)$ that is a good approximation only when there is a rate limiting barrier so that the recrossing can be ignored. However, when barriers ab and bc are comparable, the error introduced by the approximation

FIG. 2. Illustration of a pair of minima is separated by one barrier (a) or by two barriers (b) of the same height. Minima a and c are the reactant and product, respectively. TRDG gives the same barrier heights for the (a) and (b) situations.

is large. TRDG will show that for the two situations in Fig. 2 the system overcomes the same barrier height from minimum a to minimum c . In contrast to TRDG, the recrossing is taken into account by TPT.

II. METHOD

A. Molecular dynamics simulations and clustering

A 1- μ s Langevin dynamics simulation was performed for the peptide with the sequence of ACE-(ALA)₄-CBX, where ACE and CBX are the blocking groups at the termini. The simulations were carried out using CHARMM²⁷ (c35b2) with the CHARMM 19 polar hydrogen potential function²⁸ and the ACE2 implicit solvation model.²⁹ The friction coefficient of 64 ps⁻¹ was used and the simulation temperature is 500 K to facilitate the sampling. With the time step of 1 fs, the conformations were saved every 1 ps with a total of 1×10^6 conformations.

The MSMBuild2³⁰ package was used to analyze the 1- μ s trajectory. The clustering method in this package is a combination of k-centers clustering and k-medoids clustering. We chose the conformations every 100 ps along the trajectory for the initial clustering of the microstates, followed by assigning the rest of the conformations to the existing microstates. The radius of the microstate clusters was set to 1 Å because the range of 0 to 1 Å covers the 99.96% of the distribution of RMSD (root-mean-square deviation) between every pair of adjacent conformations along the trajectory (supplementary material: Fig. S1⁴⁰). We obtained 733 microstate clusters by the above protocol.

B. Markov state model

The direct transitions between every pair of microstate clusters during a certain lag time (τ) are stored in the count

matrix ($\mathbf{C}$). The direct transitions were counted using the sliding window method.³¹ A transition matrix ($\mathbf{T}$) is generated from the count matrix with $T_{ij} = C_{ij} / \sum_j C_{ij}$. To reduce the error in estimating the transition matrix required for satisfying detailed balance, we employed the Tarjan's algorithm³² to restrict the data to the maximal ergodic subgraph, followed by using the maximum likelihood estimator.³⁰ The element of the resulting transition matrix gives the probability of transition from the microstate cluster i to j during τ . The transition matrix was then diagonalized and the eigenvalues (λ s) of the transition matrix are related to the relaxation time of the master equation by $-\tau / \ln \lambda_i$.

The dynamics of a system becomes approximately Markovian when the relaxation time scale starts to be lag time independent. To determine such a lag time, a series of count matrices and transition matrices were generated with the lag time ranging from 1 ps to 100 ps. The relaxation time scale as a function of lag time is present in Fig. 3(a). It appears that the first tens slowest relaxation time scales level out after 40 ps, therefore, the lag time of 40 ps was used for the construction of transition matrix. Ideally, there is a strong separation between the slow and fast relaxation time scales in this kind plot.³³ The microstate clusters are then lumped into macrostates, and the number of macrostates is equal to the number of states with slow relaxation time scales (in the plot) plus one.³³ As a result, the kinetics of the system can be described effectively using a small number of macrostates rather than a large number of microstates. In the non-ideal cases as shown in Fig. 3(a), there is no strong separation between the slow and fast relaxation time scale.

We tried to use the MSMBuild2 package to lump the microstates into 16, 24, 48, and 64 macrostates. We found that when the number of macrostates increases, the lumping algorithm always divided the first three largest macrostates into smaller clusters. Since Macrostate 1 should not be too large and it is not convenient to illustrate many macrostates, we picked 24 macrostates for lumping using the MSMBuild2 package. Perron cluster cluster analysis (PCCA) was used to lump kinetically related microstates into macrostates.^{33,34} Note that the clustering is based on geometric similarity while the lumping is based on kinetic relations between the clusters. Although the algorithm and implementation of the PCCA lumping is different from the regrouping scheme proposed by Carr and Wales,^{35–37} both approaches are based on the idea of grouping kinetic neighbors into a macrostate.

We checked the distribution of the macrostates on the Ramachandran plot and found that some macrostates overlap significantly in all four pairs of (ϕ, φ) angles. We then further lumped these redundant macrostates, resulting in 15 macrostates eventually. We then used these 15 macrostates as nodes in the final graph and TRDG. The second stage of lumping based on the backbone dihedral angles are reasonable because we found that the total net flow between the macrostates that overlap in the Ramachandran plots of all the (ϕ, φ) pairs are large (data are not shown). The choice of number of macrostates and the lag time are validated using the Chapman–Kolmogorov test following the protocol presented in Ref. 31 (Figs. 3(b) and 3(c)). If the equation of $[T(\tau)]^k = T(k\tau)$ holds within statistical uncertainty, then the dynamics

FIG. 3. Relaxation time scale and Chapman–Kolmogorov test. (a) Relaxation time scale as a function of lag time (τ). The 50 largest eigenvalues (λ) are used in the calculation of the relaxation time scale ($= -\tau / \ln \lambda$). (b) Chapman–Kolmogorov test for 15 macrostates using 20 ps and 40 ps lag time, respectively. (c) Chapman–Kolmogorov test for 15 and 48 macrostates using 40 ps lag time. This is to test whether the equation of $T(k\tau) = [T(\tau)]^k$ is held within the statistical uncertainty for a certain lag time and the choice of macrostates. Macrostate 1 is the largest macrostate. The errors of $T(k\tau)$ are shown as error bars.

that was projected to the discrete state space satisfies Markov model. Here, $T(\tau)$ is the transition matrix estimated from the MD trajectory using lag time τ , and $T(k\tau)$ is the transition matrix estimated from the same data set using a longer lag time $k\tau$. As shown in Fig. 3(b), the test results of 40 ps lag time for the discretization of 15 macrostates indicate a reasonable approximation of Markov model. The discretization of 15 macrostates and 48 macrostates shows no substantial difference in the test results (Fig. 3(c)). The choice of 15 macrostates is for the ease of visualization of the final graph.

C. Macrostate free energies and apparent free energy barriers based on TPT

The number of conformations, n_a , of a given macrostate is related to the Helmholtz free energy F_a of the state through (see Ref. 18 for details)

$$F_a = -k_B T \ln n_a, \quad (1)$$

where k_B is the Boltzmann constant, T is the simulation temperature ($T = 500$ K in this work). For a given pair of macrostates a and b , we specify a as the reactant (source) and b as the product (sink). Note that the macrostate a and b are collections of multiple microstates.

Now, we construct a flow network of which each node represents a microstate cluster for a total of 733 nodes. All the microstates that belong to macrostate a are considered as the source and the collection of the microstates within macrostate b is the sink. The DoTPT module of the MSMBuild2 package is employed to compute the net flux (f_{ij}^+) between every pair of nodes that contributes to the total flow from the reactant to the product,

$$\begin{aligned} f_{ij}^+ &= \max\{0, f_{ij} - f_{ji}\}, \\ f_{ij} &= \pi_i q_i^- T_{ij} q_j^+, \end{aligned} \quad (2)$$

where i and j are nodes (or microstates), π_i is the Boltzmann distribution of microstate i , T_{ij} is the transition probability from i to j , q_j^+ and q_i^- are the forward-committor probability and the backward-committor probability, respectively. The details of the TPT calculation were described in Ref. 24. Here, f_{ij}^+ is the weight of the edge that connects microstate i and j . It is also the “flow” carried by the edge, as defined in graph theory, because f_{ij}^+ satisfies the flow conservation requirement,²⁴ namely, the total amount of flow entering any node that is neither the source nor the sink is equal to the total amount of flow leaving this node.

The total net flow leaving the source (s) and entering the sink (t) is calculated as

$$F_{st} = \sum_{i \in s, j \notin s} f_{ij}^+, \quad (3)$$

The physical meaning of F_{st} is the effective transitions from the source (reactant) to the sink (product) during the lag time τ . Note that the total net flow remains unchanged when the source and sink is swapped. The apparent free energy barrier between the source and sink ($F_{st}^\ddagger$) is related to F_{st} by (see supplementary material for the derivation⁴⁰):

$$F_{st}^\ddagger(\text{TPT}) = -k_B T \ln \left(\frac{F_{st}}{\tau} \frac{h}{k_B T} N_{\text{total}} \right), \quad (4)$$

where h is the Planck’s constant, $T = 500$ K, $\tau = 40$ ps, and N_{total} is the total number of conformations which is 1×10^6 in our work. The total free energy of the system is defined as $-k_B T \ln N_{\text{total}}$. The reference point of the free energy of a macrostate and the free energy barrier between any two macrostates is zero kcal/mol. Equation (4) is based on the transition-state theory, given that the reactant is in equilibrium with the transition-state complex. This is a simplified treatment, while Kramers formulation or more accurate rate theory may be used as discussed in Ref. 38. The construction

of flow network is repeated for every pair of macrostates being defined as the source and sink, totally $15 \times 14/2$ flow networks, from which the pairwise free energy barriers between all the macrostates are obtained. The final graph is built using the macrostates as nodes and the pairwise free energy barriers as the edge weights.

D. Apparent free energy barriers based on the min-cut approach used in TRDG

The underlying graph of TRDG was constructed using the macrostates as nodes. The edge weight between any pair of nodes is defined as the number of direct transitions between the nodes. The free energy barrier between nodes a and b is calculated by

$$F_{ab}^\ddagger(\text{TRDG}) = -k_B T \ln Z_{ab}, \quad (5)$$

where Z_{ab} is the partition function of the barrier and is related to the minimum cut value (n_{ab}) between source a and sink b computed using the Gomory and Hu algorithm³⁹ by

$$Z_{ab} = \frac{1}{2} n_{ab} \times \frac{h}{k_B T} \times \frac{1}{\Delta t}, \quad (6)$$

where $\Delta t = 1$ ps is the sampling interval (for details, see Refs. 17 and 21). The free energies of the macrostates and the free energy barriers between the macrostates calculated using TRDG have the same reference point as those computed by TPT.

III. ILLUSTRATIVE APPLICATIONS

Macrostate 1 is the largest state with 84.7% population. Macrostates 2 and 3 have 10.8% and 1.6% population, respectively. The distribution of the macrostates on the Ramachandran plot of the second and third pair of (ϕ, φ) angles is presented in Figure 4. Even though the backbone dihedral angles are crude parameters for the projection of free energy landscapes, they still provide some insights into the discretization of the conformation space. The distribution of the first three macrostates show significant overlap in the Ramachandran plot of the third pair of (ϕ, φ) angles, whereas there are some distinctions between the macrostates in the second pair of (ϕ, φ) angles, where Macrostate 2 is mainly in the α helical region while Macrostate 1 is predominantly distributed in the β and P_{II} region. Compared to the first three macrostates, Macrostate 4 is located in the different region of the Ramachandran plot of the second pair of (ϕ, φ) angle.

The final graph (Fig. 5) gives free energy barriers between the 15 macrostates. This graph should not be confused with the flow network where a particular pair of macrostates is specified as source and sink. The edge weights on this graph are the free energy barriers between the incident nodes. The graph summarizes the results of $15 \times 14/2$ flow networks with every pair of macrostates is successively specified as a source and a sink. Macrostates 1, 2, 3 and 4 form a low-barrier area on the graph. Macrostates 1 and 2 are separated by the lowest barrier on the graph. Our weighted graph cannot be further visualized using a tree representation because the computation of the total net flow based on TPT is conducted for each pair

FIG. 4. Ramachandran plot of the second (a) and third pairs (b) of (ϕ, ψ) angles of the 15 macrostates. One conformation is denoted by a dot on the plot. Macrostates 1–15 are ordered from left to right along the rows.

of nodes individually, and the free energy barriers are different for different pairs as shown in Table I.

In contrast to our approach, the free energy barriers resulted from TRDG can be naturally organized into a tree structure (Fig. 6), but it is important to realize that TRDG gives a less accurate picture than our approach because it

counts every transition between two nodes instead of the effective transitions that contribute to the reaction as described in the Introduction. As a result, all the barrier heights given by TRDG are underestimated to a certain extent (up to 2.24 kcal/mol) for the system. Based on Eq. (1), the free energy of macrostates on the TRDG (Fig. 6) has the same value

FIG. 5. Graph that depicts the free energy landscape of the tetrapeptide. Each node denotes a macrostate. The Arabian numbers are the indices of the macrostates. The size of each node represents the free energy of the macrostate. Macrostate 1 has the lowest free energy. Node 13–15 are not shown because they are separated by high barriers from the rest of the nodes. The thickness of an edge represents the free energy barrier between its incident nodes. The thicker the edge, the lower the barrier. The legend indicates the edge thickness that corresponds to the free energy (–4 kcal/mol to 0 kcal/mol). The edges that represent barriers higher than 0.63 kcal/mol are not shown on the graph. The color scheme helps to estimate the relative free energies. The free energy barriers are calculated using TPT.

as those on our graph (represented by the size of the nodes of Fig. 5). Macrostate 1 is the global free energy minimum.

Intrinsically, TRDG can be represented by a tree structure using Gomery-Hu algorithm³⁹ that recursively partitions a set of nodes into two non-overlapping subsets, as a result, any pair of the nodes with one from each subset has the same max-flow value. Therefore, the nodes in the two subsets belong to two different branches of the tree, and the free energy barrier between any pair of the nodes, each from the two different subsets, is represented by the same free-energy-barrier node. Iteratively, all the nodes can be organized into a tree. For example, Macrostates 1–8 and 10 belong to one subset, while Macrostate 9 is from another subset, accordingly, the barrier

FIG. 6. TRDG of tetrapeptide. The 15 macrostates and the barriers calculated between them are organized into a tree structure. The Arabian numbers are the indices of the macrostates. Macrostate 1 is the global minimum.

that separates Macrostate 9 from any of these macrostates is –3.04 kcal/mol (Fig. 6). Our graph gives different barrier heights from Macrostate 9 to these macrostates, ranging from –0.67 kcal/mol to –0.01 kcal/mol (Table I). Comparing the ranking of the barriers between Macrostate 1 and each of the other macrostates, the TRDG result (Fig. 6) is in line with that of our graph (column 1 of Table I): $F_{15,1}^\ddagger > F_{14,1}^\ddagger > F_{13,1}^\ddagger > F_{11,1}^\ddagger > F_{12,1}^\ddagger > F_{9,1}^\ddagger > F_{10,1}^\ddagger > F_{7,1}^\ddagger > F_{8,1}^\ddagger > F_{6,1}^\ddagger \cong F_{5,1}^\ddagger > F_{4,1}^\ddagger > F_{3,1}^\ddagger > F_{2,1}^\ddagger$. The trend of the barrier heights for the transitions from Macrostate 6–10 to Macrostate 5 shows the greatest inconsistency between the results of TRDG and our graph: $F_{9,5}^\ddagger > F_{10,5}^\ddagger > F_{7,5}^\ddagger > F_{8,5}^\ddagger > F_{6,5}^\ddagger$ (Fig. 6); $F_{10,5}^\ddagger \cong F_{9,5}^\ddagger > F_{8,5}^\ddagger > F_{6,5}^\ddagger > F_{7,5}^\ddagger$ (Table I).

For the tetrapeptide system, there are 15 macrostates and 14 barriers shown in TRDG, while there are 105 barrier values listed in Table I. Multiple barrier heights on our graph correspond to a single barrier value calculated by the TRDG because of the intrinsic nature of TRDG though the correlation

TABLE I. Apparent free energy barriers (kcal/mol) between the 15 macrostates computed by TPT.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
S1	...	–4.11	–2.51	–2.27	–1.29	–1.28	–1.01	–1.09	–0.67	–0.73	0.32	0.17	0.98	1.63	2.24
S2	–4.11	...	–2.35	–2.27	–1.24	–1.22	–0.99	–1.12	–0.73	–0.70	0.33	0.14	0.95	1.62	2.26
S3	–2.51	–2.35	...	–1.70	–1.02	–1.29	–0.80	–0.95	–0.53	–0.86	0.38	0.24	1.01	1.64	2.27
S4	–2.27	–2.27	–1.70	...	–0.96	–0.97	–0.74	–0.87	–0.58	–0.54	0.39	0.16	1.00	1.59	2.26
S5	–1.29	–1.24	–1.02	–0.96	...	–0.57	–0.98	–0.48	–0.27	–0.27	0.50	0.37	1.07	1.68	2.29
S6	–1.28	–1.22	–1.29	–0.97	–0.57	...	–0.42	–0.53	–0.24	–0.54	0.50	0.38	1.08	1.68	2.29
S7	–1.01	–0.99	–0.80	–0.74	–0.98	–0.42	...	–0.35	–0.16	–0.16	0.55	0.43	1.10	1.69	2.29
S8	–1.09	–1.12	–0.95	–0.87	–0.48	–0.53	–0.35	...	–0.22	–0.25	0.54	0.39	1.09	1.69	2.28
S9	–0.67	–0.73	–0.53	–0.58	–0.27	–0.24	–0.16	–0.22	...	–0.01	0.63	0.27	1.13	1.71	2.31
S10	–0.73	–0.70	–0.86	–0.54	–0.27	–0.54	–0.16	–0.25	–0.01	...	0.62	0.51	1.14	1.72	2.31
S11	0.32	0.33	0.38	0.39	0.50	0.50	0.55	0.54	0.63	0.62	...	0.94	1.39	1.86	2.39
S12	0.17	0.14	0.24	0.16	0.37	0.38	0.43	0.39	0.27	0.51	0.94	...	1.32	1.81	2.37
S13	0.98	0.95	1.01	1.00	1.07	1.08	1.10	1.09	1.13	1.14	1.39	1.32	...	1.94	2.50
S14	1.63	1.62	1.64	1.59	1.68	1.68	1.69	1.69	1.71	1.72	1.86	1.81	1.94	...	2.68
S15	2.24	2.26	2.27	2.26	2.29	2.29	2.29	2.28	2.31	2.31	2.39	2.37	2.50	2.68	...

FIG. 7. Correlation between the TPT-based results and the results of TRDG. TRDG gives 14 barriers between the 15 macrostates, while the TPT-based approach gives 105 barrier values as shown in Table I. The correlation coefficient is 0.98.

coefficient between the two results is 0.98 (Fig. 7). In general, if the range of multiple barrier heights given by the TPT-based approach that corresponds to the single barrier in TRDG is small and the correlation between these two kind barriers is high, the TRDG provides an overall good illustration of the relationship among the conformational states although the barriers are underestimated. If such range is great (>1 kcal/mol) or the correlation is low, cautions are needed when the TRDG is utilized.

IV. CONCLUDING DISCUSSION

We propose to use a weighted graph constructed from MD trajectories to describe free energy landscape with the nodes on the graph representing the conformational states and the edge weights reflecting the free energy barriers between the states. Some intuitive and widely used approaches for the analysis of free energy landscape rely on the methods of dimensionality reduction by projecting the high-dimension data onto some progress variables, such as the number of native contacts, the radius of gyration, secondary structure content, and principal components. These approaches, however, might miss some possible intermediate states and the transitions involved, even when multiple projections along different progress variables are used. Our graph depicts the free energy landscape without such a projection. The calculation of free energy barriers was based on TPT using the MSM-Builder2 package, which computes the net reactive transitions. We compare our approach with the MD-based TRDG approach which uses a min-cut algorithm³⁹ to compute the max-flow between pairs of nodes. These max-flow values provide the upper bounds of the net reactive transitions and the corresponding free energy barriers are the lower bounds. Note that the TPT calculations are based on microstates, while the macrostates are for the purpose of visualization and description of the graph only. The MD-based TRDG can be used on either the microstate or macrostate level, depending on one's interest. Alternatively, TRDG can be constructed using geometry optimization and unimolecular rate theory by which the barrier-recrossing effect is taken into account (see Ref. 10, and references therein). The focus of our work is to compare the weighted graph constructed from MD trajectories with the

MD-based TRDG because both approaches are based on the same set of clusters.

Clearly, the calculations of free energy and free energy barriers depend on how the conformation space is discretized into states, which is, in practice, the clustering algorithm and clustering parameters. In this study, the discretization is based on a purely geometric criterion, RMSD. It was shown that the Markov state model is very robust with respect to changes of the clustering method and metric within a significant range.³¹ For a small peptide, it was shown that three different clustering methods and different number of clusters gave out the similar slowest relaxation time scale.³¹ Besides the error in partitioning the conformation space into discrete states, the statistical error of the transition matrix also affects the final results. For the tetrapeptide system, the total simulation time is $1\ \mu\text{s}$ which is orders of magnitude longer than the slowest relaxation time scale (~ 100 ps) shown in Fig. 3(a). Therefore, the statistic error due to the finite simulation data is small in our system.

The computational cost of the construction of our graph increases with the number of macrostates because our method needs to construct the flow networks for every pair of macrostates that are specified as source and sink. Generally, there are not many macrostates, usually in the order of tens. In some cases, it is even not necessary to calculate all the pairwise free energy barriers for the macrostates if only a few of the states are of interest. The computational expense of TPT calculations is minimal with respect to the molecular dynamics simulations. Therefore, our weighted-graph approach is computationally affordable.

ACKNOWLEDGMENTS

This work is funded by National Institutes of Health (R01-GM088326). We thank Dr. Da-wei Li for helpful discussions. We are very grateful to Professor David Wales for providing the programs of creating and manipulating disconnected graphs.

- ¹M. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. LeGrand, A. Beberg, D. Ensign, C. Bruns, and V. S. Pande, *J. Comput. Chem.* **30**, 864–872 (2009).
- ²D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles, in *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis (SC09)* (ACM, New York, NY, 2009).
- ³K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, *Science* **334**, 517–520 (2011).
- ⁴A. Berezhkovskii, G. Hummer, and A. Szabo, *J. Chem. Phys.* **130**, 205102 (2009).
- ⁵A. Caflisch, *Curr. Opin. Struct. Biol.* **16**, 71–78 (2006).
- ⁶F. Rao and A. Caflisch, *J. Mol. Biol.* **342**, 299–306 (2004).
- ⁷F. Noe and S. Fischer, *Curr. Opin. Struct. Biol.* **18**, 154–162 (2008).
- ⁸G. S. Buchner, R. D. Murphy, N. V. Buchete, and J. Kubelka, *Biochim. Biophys. Acta* **1814**, 1001–1020 (2011).
- ⁹M. Andreac, A. K. Felts, E. Gallicchio, and R. M. Levy, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6801–6806 (2005).
- ¹⁰D. J. Wales, *Curr. Opin. Struct. Biol.* **20**, 3–10 (2010).
- ¹¹S. M. Kreuzer, R. Elber, and T. J. Moon, *J. Phys. Chem. B* **116**, 8662–8691 (2012).
- ¹²C. B. Li, H. Yang, and T. Komatsuzaki, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 536–541 (2008).

- ¹³R. Czerminski and R. Elber, *J. Chem. Phys.* **92**, 5580–5601 (1990).
- ¹⁴O. M. Becker and M. Karplus, *J. Chem. Phys.* **106**, 1495–1517 (1997).
- ¹⁵D. J. Wales, M. A. Miller, and T. R. Walsh, *Nature (London)* **394**, 758–760 (1998).
- ¹⁶D. J. Wales and T. Head-Gordon, *J. Phys. Chem. B* **116**, 8394–8411 (2012).
- ¹⁷D. W. Li, L. Han, and S. Huo, *J. Phys. Chem. B* **111**, 5425–5433 (2007).
- ¹⁸S. V. Krivov and M. Karplus, *J. Chem. Phys.* **117**, 10894–10903 (2002).
- ¹⁹S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14766–14770 (2004).
- ²⁰D. A. Evans and D. J. Wales, *J. Chem. Phys.* **118**, 3891–3897 (2003).
- ²¹D. W. Li, M. Khanlarzadeh, J. Wang, S. Huo, and R. Bruschweiler, *J. Phys. Chem. B* **111**, 13807–13813 (2007).
- ²²M. Duan, J. Fan, M. Li, L. Han, and S. Huo, *J. Chem. Theory. Comput.* **9**, 2490–2497 (2013).
- ²³W. E and E. Vanden-Eijnden, *Annu. Rev. Phys. Chem.* **61**, 391–420 (2010).
- ²⁴F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19011–19016 (2009).
- ²⁵V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).
- ²⁶W. E and E. Vanden-Eijnden, *J. Stat. Phys.* **123**, 503–523 (2006).
- ²⁷B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caffisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, *J. Comput. Chem.* **30**, 1545–1614 (2009).
- ²⁸E. Neria, S. Fischer, and M. Karplus, *J. Chem. Phys.* **105**, 1902–1921 (1996).
- ²⁹M. Schaefer, C. Bartels, and M. Karplus, *J. Mol. Biol.* **284**, 835–848 (1998).
- ³⁰K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, *J. Chem. Theory Comput.* **7**, 3412–3419 (2011).
- ³¹J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schutte, and F. Noe, *J. Chem. Phys.* **134**, 174105 (2011).
- ³²R. Tarjan, *SIAM J. Comput.* **1**, 146–160 (1972).
- ³³G. R. Bowman, X. Huang, and V. S. Pande, *Methods* **49**, 197–201 (2009).
- ³⁴P. Deuffhard, W. Huisinga, A. Fischer, and C. Schütte, *Linear Algebra Appl.* **315**, 39–59 (2000).
- ³⁵J. M. Carr and D. J. Wales, *J. Chem. Phys.* **123**, 234901 (2005).
- ³⁶J. M. Carr and D. J. Wales, *J. Phys. Chem. B* **112**, 8760–8769 (2008).
- ³⁷J. M. Carr and D. J. Wales, *Phys. Chem. Chem. Phys.* **11**, 3341–3354 (2009).
- ³⁸C. L. Brooks III, M. Karplus, and B. M. Pettitt, in *Proteins: A Theoretical Perspective of Dynamics, Structure & Thermodynamics*, edited by I. Prigogine and S. Rice (Wiley-Interscience, New York, 1988), Vol. LXXI.
- ³⁹R. E. Gomory and T. C. Hu, *SIAM J. Appl. Math.* **9**, 551 (1961).
- ⁴⁰See supplementary material at <http://dx.doi.org/10.1063/1.4829768> for the derivation of Eq. (4) and Fig. S1: Distribution of RMSD between every pair of adjacent conformations along the 1- μ s trajectory.