JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Ultrametricity in Protein Folding Dynamics

Riccardo Scalco and Amedeo Caflisch*

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

ABSTRACT: The free energy of the transition state (TS) between two nodes of an ergodic Markov state model (MSM) can be obtained from the minimum cut, which is the set of edges that has the smallest sum of the flow capacities among all the possible cuts separating the two nodes. Here, we first show that the free energy of the TS is an ultrametric distance. The ultrametric property offers a way to simplify the MSM in a small number of states and, as a consequence, meaningful rate constants (free energy barriers) for the simplified MSM can be defined. We also present a new definition of the cut-based free energy profile (cbFEP), which is useful to check for the existence of a state for which the equilibration is much faster than the time to escape from it. From our analysis, a parallelism emerges between the minimum cut (maximum flow), and transition state theory (TST) or Kramers' theory.

INTRODUCTION

An N-state kinetic process can be formalized with a system of N rate equations—namely, the master equations—describing the system as a random process governed by the exponential distribution.[1,2] The N^2 rate constants appearing in the equations are supposed to be known, and the physics behind them is described by mainly two theories: transition state theory (TST)[3] and Kramers' theory.[4] Both of them define the rate constants from assumptions regarding the dynamics of the process. Being not the same theory, TST and Kramers' theory use different definitions, but it is crucial for the objective of our work to note that the two theories share some assumptions and that, in both of them, the definition of rate constant satisfies certain properties. In particular, we focus our attention on the meaning of "state" and the formal definition of the rate constant. TST has its origin in statistical mechanics, and with "state", it means a region of configuration space, namely, a subset of spatial coordinates usually in the neighborhood of an energy potential minimum. For example, in a two-state process one state is the *native state*, the other is the so-called *unfolded state*.[1] The former is described as a set of configurations around a potential minimum, while the latter includes all of the remaining configurations. In Kramers' theory, since the original paper on the diffusion model of chemical reactions,[5] the model consists of a classical particle (namely, the reaction coordinate) trapped in a one-dimensional potential well and subjected to a frictional force. Kramers asked for the rate of escape of the particle from the well. Hence, in both these theories, the term "state" indicates a finite region of the configuration space in the neighborhood of an energy minimum.

Assuming to divide the system in two states A and B, where B contains all of the phase space not occupied by A (Figure 1), in both TST and Kramers' theory, the definition of the rate constant $k_{A,B}$, between the initial state A and the final state B, is defined as the ratio $k_{A,B} \equiv Z_{A,B}/Z_A$, where Z_A is the partition function of the initial state. The quantity $Z_{A,B}$ depends only on the boundary dividing the states A and B but not on the direction; that is, $Z_{A,B} = Z_{B,A}$. TST and Kramers' theory differ from each other in the physical interpretation of the quantity $Z_{A,B}$. TST introduces the existence of a *transition state* (TS) between states A and B, and defines $Z_{A,B}$ as the partition function of the TS. Kramers' theory instead characterizes the rate $k_{A,B}$ to escape from state A by the *flux* of particles that pass through the bottleneck separating A and B.[4,5] Many approaches have been developed to calculate the flux;[4] one of them involves the calculation of the average time that the system needs to leave the domain of attraction for the first time. The key points here are the general properties of $Z_{A,B}$, namely, the dependence only on the boundary region and its independence on the direction, and the physical interpretations of it.

In describing the kinetics of protein folding by means of ergodic Markov state models (MSMs),[6−12] a natural definition of "state" and of "rate constant" emerges, and it is the objective of the present work to show how these two concepts could be defined from a cut-based free-energy analysis of the MSM.

To divide the phase space in states, it is most appropriate to identify the cut that maximizes the free energy of the TS between two nodes. Such a choice is of wide range applicability, indeed proving to be useful in protein folding dynamics[13] and in spin models of magnetic domains.[14] We show here that the free energy of the TS defines an ultrametric distance, which results in an automatic procedure to reduce the MSM, usually consisting of thousands of nodes, into few states. The proposed procedure could be seen as a community algorithm optimized for networks describing the protein folding free-energy landscape, because it is based on the definition of the free energy of the TS, i.e., the kinetics of the process. The procedure results in a reduced MSM whose states are a collection of nodes belonging to the original MSM. The number of transitions between the states (of the reduced MSM) are then used to calculate the free energy of the TS and the activation free energy between them, similar to a previous approach based on transitions observed during molecular dynamics.[15] The main goal of the present paper is not to solve problems such as lack of sampling or time scales overlapping; instead, our purpose is

Received: January 5, 2012
Published: March 10, 2012

dx.doi.org/10.1021/ct3000052 | J. Chem. Theory Comput. 2012, 8, 1580−1588

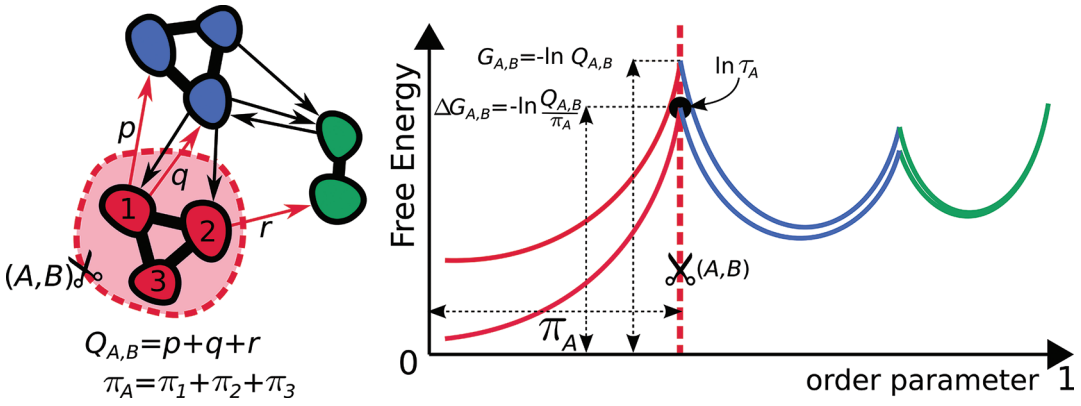

Figure 1. Schematic illustration of the main concepts. (Left) Schematic representation of a MSM consisting of eight nodes grouped in three states emphasized by different colors. The cut (A,B) (labeled by the scissor symbol) separates state A, of probability π_A, from the rest, and the cut value $Q_{A,B}$ is the flow through the cut (A,B). (Right) Graphical representation of the cut-based free energy profile (cbFEP) related to the MSM on the left. The cut (A,B) separates the profiles at the first peak, where the values $G_{A,B}$, $\Delta G_{A,B}$, and τ_A indicate the free energy of the TS, the activation free energy, and the mean time to escape from state A, respectively.

to extend the language of the mincut maxflow method applied to the kinetics of conformational transitions. In particular, we prove that, in the framework of finite Markov chains, the cut-based free energy of the TS defines an ultrametric distance on the set of states of a generic ergodic MSM that does not need to satisfy the detailed balance condition. Moreover, here, we present an alternative definition of the cut-based free-energy profile (cbFEP)[16] useful to check for the existence of a metastable state, being it easily comparable with other analysis of MSMs, such as mean exit time and mixing time.

Given the minimum cut $(A,B|i,j)$ between nodes i and j, the cut-based quantity $Q_{A,B}$ (see below) naturally fulfills both the mathematical properties and the physical interpretations of the above-mentioned quantity $Z_{A,B}$. By definition, $Q_{A,B}$ is dependent only on the cut (A,B), and, thus, $Q_{A,B} = Q_{B,A}$ represents both the flow through the cut and the partition function used to calculate the *free energy of the TS* $G_{A,B}$. It is straightforward to define the rate constant as

$$k_{A,B} \equiv \frac{Q_{A,B}}{\pi_A}$$

where π_A is the probability of finding the system in state A ($\pi_A \equiv \sum_{i\in A}\pi_i$). We will show that such definition is meaningful only if the system satisfies restricted conditions, namely, the mean time to escape from a state must be much longer than the time needed to equilibrate inside the state, and we present a way to check the validity of this assumption. Then, in similarity with the Van't Hoff–Arrhenius law, the *activation free energy*, $\Delta G_{A,B}$, can be defined as

$$\Delta G_{A,B} \equiv -\ln k_{A,B} = G_{A,B} - G_A$$

where G_A is the free energy of state A (defined as $G_A \equiv -\ln \pi_A$, given in units of k_BT).

We observe here that, by means of the cut-based free-energy definition, two distinct observables relating to states kinetics are derived: the free energy of the TS between the reference state and the remaining part of the phase space, and the free energy of activation in order to escape from the reference state. Note that the two quantities are different, and they become closer as the probability π_A to find the system in state A increases toward unity.

In the following sections, we present the theory behind the cut-based analysis. We show that the free energy of the TS defines an ultrametric distance between nodes, and how this property motivates the partition of the entire phase space in a few number of states. We then present the standard procedure to derive a Markov process in continuous time from the reduced MSM, obtaining, in this way, the master equation of the system. The rate constants appearing in the master equation are then compared with the mean escape time from the states in the original MSM, giving a strong criterion in order to establish the degree of approximation involved in the reduction procedure. This comparison is indeed related to the assumption of separated time scales for the system to equilibrate inside a state and to escape from that state; this assumption is at the base of both TST and Kramers' theory. Finally, we guide the reader along the entire cut-based analysis of a propaedeutic example and we conclude with an application to a MSM generated by molecular dynamics of the reversible folding of a structured peptide.

CUT FREE ENERGY AS METRIC DISTANCE

Assumptions and Definitions. Let P be the transition matrix defining an ergodic Markov chain[17] and G be the associated directed graph ($G \equiv (V,E)$, where V and E are the set of nodes and edges, respectively). Given the steady state of the chain ($\pi = \pi P$), the rate probability between nodes i and j is

$$q_{ij} = \pi_i P_{ij}$$

Ergodicity implies that the conservation law $\sum_j q_{ij} = \sum_j q_{ji}$ holds for every node i:

$$\sum_j q_{ij} = \sum_j \pi_i P_{ij} = \pi_i \sum_j P_{ij} = \pi_i$$

$$\sum_j q_{ji} = \sum_j \pi_j P_{ji} = [\pi P]_i = \pi_i$$

Any partition of the node set V in two disjoint subsets A and B ($A \cup B = V$, $A \cap B = \varnothing$ and $A,B \neq \varnothing$) defines a cut $C \equiv (A,B)$ on the graph G,[18,19] the *cut-set* of which is the subset of edges $C_{A,B} \equiv \{(i,j) \in E | i \in A, j \in B\}$. Weighting every edge $(i,j) \in E$

dx.doi.org/10.1021/ct3000052 | J. Chem. Theory Comput. 2012, 8, 1580−1588

by the value q_{ij}, the cut-value $Q_{A,B}$ is defined as the sum of the weights:

$$Q_{A,B} \equiv \sum_{C_{A,B}} q_{ij}$$

In the following, we make use also of expressions as "$Q_{A,B}$ is the flow going from A to B" or "$Q_{A,B}$ is the flow through the cut (A,B)". The *free energy of the TS between sets A and B* is defined as $G_{A,B} \equiv -\ln Q_{A,B}$.[13] As proved below, for an ergodic system, the symmetry property holds, namely, for every cut (A,B), it is true that $Q_{A,B} = Q_{B,A}$ and so $G_{A,B} = G_{B,A}$. With the objective of interpreting the argument of the logarithm in the above definition as a probability, we may say that it is the probability to observe on the Markov chain P a transition from a node in A to a node in B at a certain time, without conditional knowledge about which subset the initial node belongs to:

$$\sum_{i\in A} \pi_i \sum_{j\in B} P_{ij} = \sum_{\substack{i\in A\\ j\in B}} q_{ij} \equiv Q_{A,B}$$

Given two nodes i and j, we denote with the symbol $(A,B|i,j)$ a generic cut such that $i \in A$ and $j \in B$. $Q_{A,B|i,j}$ then indicates the flow through the cut $(A,B|i,j)$. Many such cuts are possible (as many as the number of bipartitions of V, such that i and j are not in the same subset) and we indicate with $Q^{\dagger}_{A,B|i,j}$ the minimum of the associated cut values:

$$Q^{\dagger}_{A,B|i,j} \equiv \min\{Q_{A,B|i,j}\}$$

The *free energy of the TS between two nodes* is then defined as

$$G^{\dagger}_{A,B|i,j} \equiv -\ln Q^{\dagger}_{A,B|i,j}$$

with the additional convention that $G^{\dagger}_{A,B|i,i} \equiv 0$. For the sake of brevity, we choose not to indicate the bipartition in subsets A and B and simply write $G^{\dagger}_{ij} \equiv -\ln Q^{\dagger}_{ij}$. An observation here is appropriate: no subset of nodes of V represents the TS between nodes i and j.

Ultrametricity. Here, we show that the free energy of the TS between any two nodes i and j ($G^{\dagger}_{ij}$) is an ultrametric distance on the set of nodes V. In other words, we must show that $G^{\dagger}_{ij}$ satisfies the following three conditions ($\forall i,j \in V$):

(1)

$$G^{\dagger}_{ij} = 0 \qquad \text{if and only if} \qquad i = j$$

To prove this property, it suffices that $G^{\dagger}_{ij} \neq 0$ if $i \neq j$ ($G^{\dagger}_{ii} = 0$ is true by definition). This is implied by the fact that for every cut (A,B) performed on a graph associated to a Markov chain, it is true that $C_{A,B} \neq E$, and so $Q_{A,B} \neq 1$.

(2)

$$G^{\dagger}_{ij} = G^{\dagger}_{ji}$$

The symmetry property is a consequence of the conservation law $\forall i \in V$: $\sum_j q_{ij} = \sum_j q_{ji}$, and of the fact that $G^{\dagger}_{ij}$ corresponds to the cut with the minimum flow. First, we show that, for every cut (A,B), it is true that $Q_{A,B} = Q_{B,A}$, where $Q_{B,A}$ is the cut value of (B,A). Remembering that a cut is a bipartition of V, we have

$$\sum_{i\in A} \sum_j q_{ij} = \sum_{i\in A} \left(\sum_{j\in A} q_{ij} + \sum_{j\in B} q_{ij}\right) = \sum_{\substack{i\in A\\ j\in A}} q_{ij} + \sum_{\substack{i\in A\\ j\in B}} q_{ij}$$

$$\sum_{i\in A} \sum_j q_{ji} = \sum_{i\in A} \left(\sum_{j\in A} q_{ji} + \sum_{j\in B} q_{ji}\right) = \sum_{\substack{i\in A\\ j\in A}} q_{ji} + \sum_{\substack{i\in A\\ j\in B}} q_{ji}$$

From the conservation law, it follows that $\sum_{i\in A}\sum_j q_{ij} = \sum_{i\in A}\sum_j q_{ji}$, which directly implies

$$Q_{A,B} \equiv \sum_{\substack{i\in A\\ j\in B}} q_{ij} = \sum_{\substack{i\in A\\ j\in B}} q_{ji} \equiv Q_{B,A}$$

because

$$\sum_{\substack{i\in A\\ j\in A}} q_{ij} = \sum_{\substack{i\in A\\ j\in A}} q_{ji}$$

Finally, from the above equality, if $(A,B|i,j)$ is the cut with the minimum flow from A to B, then the cut $(B,A|j,i)$ has the minimum flow in the opposite direction.

(3)

$$G^{\dagger}_{ij} \leq \max\left\{G^{\dagger}_{ik}, G^{\dagger}_{kj}\right\}$$

The strong triangle inequality is a consequence of the fact that, in defining $G^{\dagger}_{ij}$, we make use of the minimum cut value, $Q^{\dagger}_{A,B|i,j}$, in the set of all the possible ones. By the properties of the logarithm, the triangle inequality becomes the inequality $Q^{\dagger}_{ij} \geq \min\{Q^{\dagger}_{ik}, Q^{\dagger}_{kj}\}$. Given the cut $C = (A,B|i,j)$ associated with $Q^{\dagger}_{A,B|i,j}$, there are two mutually exclusive possibilities for the third node k: $k \in A$ or $k \in B$. In the case that $k \in A$, the cut C is also a cut $(A,B|k,j)$ and the associated cut value satisfies at

$$Q^{\dagger}_{ij} = Q_{A,B|k,j} \geq \min\{Q_{T,S|k,j}\} \equiv Q^{\dagger}_{kj}$$

That being so, in the case $k \in A$, the triangular inequality is equivalent to the logical function

$$Q^{\dagger}_{ik} \geq Q^{\dagger}_{kj} \qquad \text{OR} \qquad Q^{\dagger}_{ij} \geq Q^{\dagger}_{ik}$$

In order to show that the logical disjunction is true, it suffices that the arguments are not both false. If, by hypothesis, they are both false (namely, $Q^{\dagger}_{ik} < Q^{\dagger}_{kj}$ and $Q^{\dagger}_{ij} < Q^{\dagger}_{ik}$ are both true), then we have a reduction ad absurdum: $Q^{\dagger}_{ij} < Q^{\dagger}_{kj}$ negates the inequality $Q^{\dagger}_{ij} \geq Q^{\dagger}_{kj}$ already proven. This ensures the strong triangle inequality stated above in the case $k \in A$. In the case $k \in B$, the reasoning is entirely similar, and we omit the proof.

The second and third properties, namely, $G^{\dagger}_{ij} = G^{\dagger}_{ji}$ and $G^{\dagger}_{ij} \leq \max\{G^{\dagger}_{ik}, G^{\dagger}_{kj}\}$, are the formalizations of the following two observations. First, the free energy of the TS between two nodes is not dependent on the system direction. However the system goes from i to j, or from j to i, it must overcome the same free energy of the TS. Note that such property assumes

ergodicity but does not assume the detailed balance (even if detailed balance is expected to be fulfilled in equilibrium molecular dynamics[20]). Second, forcing the system to go from i to j, passing through k, cannot result in a lower free energy of the TS, regardless of the node k. In other words, the strong triangle inequality ensures that the only possible triangles are either isosceles with a small base or equilateral. This property could be better understood by remembering that, in the mathematical framework that we are using here, the TS is always a phase space region that acts like a ring (the cut) dividing the entire conformational space into two disjointed subsets. It is straightforward that the TS cannot be a *simply* connected region, and the choice to force the system to go from i to j passing through k cannot avoid crossing the minimum cut between i and j.

Applications: Disconnectivity Graphs, Reduced Markov Chains, and Escape Time. Ultrametricity is a relatively new concept in physics as well as in biology. A detailed review of its applications[21] shows how, despite their abstractness, ultrametric distances are of wide-range usability. Here, we are interested in the possibility of subdividing the entire network in subsets of nodes, called states, such that the simplified picture still depicts the original kinetics quantitatively. In particular, for free-energy projections that preserve the barriers,[16] it becomes apparent that such subdivision is not only possible, but also is very useful in order to derive a chemical master equation that is comparable with the experimental analysis.[22,23] Here, we use a well-known application of the ultrametricity, namely, the fact that, from any ultrametric set, a dendrogram can be unambiguously built. There is indeed a one-to-one relationship between an indexed hierarchy, a dendrogram with positive real values defined at each divergence, and an ultrametric set. Dendrograms based on potential energy or free energy (disconnectivity graphs) have been introduced in the past 15 years to characterize the shape of the multidimensional (free) energy surface.[13,24−27]

Yet, the ultrametric nature of the cut-based free energy of the TS has not been reported in previous works. In this context, it is important to note that, different from previous studies,[26,28] ergodicity is a fundamental assumption, whereas detailed balance is not required. Moreover, the definition of an ultrametric distance on the set of nodes is a different way to state the important properties of the minimum cut method, and it is potentially useful for further theoretical investigations in Markov theory, as well as in all the applications that use finite regular Markov chains models.

In order to define the dendrogram, the free energy of the TS between any pair of nodes must be calculated. This can be done by means of the isomorphism between the rate probabilities q_{ij} of the Markov chain and the capacities c_{ij} defined in a flow network. The task to find the minimum cut is then solved with standard methods such as the Ford−Fulkerson algorithm,[19] in order to find the minimum cut between two nodes, and the Gomory−Hu algorithm,[28] which is useful to deduce all the $V(V-1)/2$ minimum cuts after only $V-1$ flow problems have been computed. Different from what Gomory and Hu assumed, here, we do not assume detailed balance ($c_{ij} = c_{ji}$); nevertheless, the method still holds, because the symmetry property $Q_{A,B} = Q_{B,A}$ is true for every cut (A,B) in an ergodic chain. Note also that the strong triangle inequality, written in the form $Q^{\dagger}_{ij} \geq \min\{Q^{\dagger}_{ik}, Q^{\dagger}_{kj}\}$, is a necessary and sufficient condition for a generic matrix $Q^{\dagger}$ to be realizable by some flow network, as proven in ref 28. Moreover, the strong triangle inequality is easily understood considering its graphical interpretation on the dendrogram (see, for example, Figure 2). The free energy of the

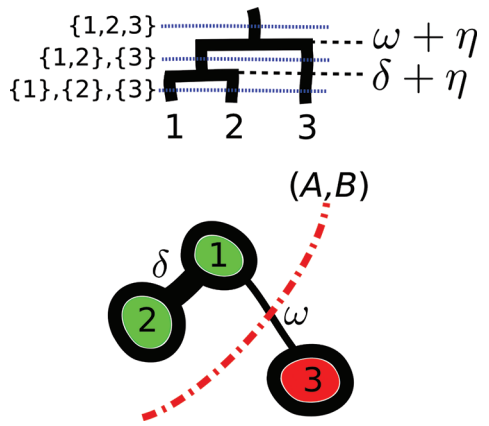

Figure 2. Depiction of a degenerate state. The top image shows the dendrogram resulting from the ultrametric distance matrix $G^{\dagger}$. Note that the free energy of the TS between two nodes is the value at the divergence between them in the hierarchical tree. The strong triangle inequality—that is, the fact that the free energy of the TS between nodes i and j cannot decrease imposing the passage through a third node k—is a consequence of the tree structure. Here, for example, it is true that $G^{\dagger}_{1,3} \leq \max\{G^{\dagger}_{1,2}, G^{\dagger}_{2,3}\}$ and $G^{\dagger}_{1,2} \leq \max\{G^{\dagger}_{1,3}, G^{\dagger}_{3,2}\}$. The bottom image shows a schematic picture of the original Markov chain subdivided into two states.

TS between two nodes i and j is the value at the divergence between them in the hierarchical tree. Therefore, the third node k must be either a descendant of the same divergence, in which case the free energy of the TS between nodes i and j passing through k does not change, or it is located outside the subtree containing nodes i and j, in which case the free energy of the TS between nodes i (or j) and k is greater than the one between i and j.

Once the dendrogram is known, a natural clustering procedure is defined by proceeding from the bottom to the top. In this way, nodes are merged into states, according to the hierarchy, and the free energy of the TS between two states corresponds to the one calculated between a node in one state and a node in the other, with the choice of these two nodes being not important.

Once the MSM is clustered in states, a reduced MSM is defined in the following way. Let x and y be two states (namely, two disjointed subsets of nodes), the rate probability between them is defined as $q_{xy} \equiv \sum_{i\in x, j\in y} q_{ij}$, and the transition probability is then calculated as $P_{xy} \equiv q_{xy}/\sum_y q_{xy}$. The procedure presented here to reduce a Markov chain is not free of issues; in particular, important questions emerge from the analysis of kinetic observables. Here, we focus on the following question: Is the mean escape time from a state in the reduced chain equal to the one calculated in the original chain? Generally, the answer is negative; we will determine how to check its validity using analytical calculations.

MEAN ESCAPE TIME

From a Markov chain in discrete time, it is possible to derive the corresponding Markov process in continuous time.[29,30] The Markov process is described by the master equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_x(t) = \sum_y \left(k_{yx}\pi_y(t) - k_{xy}\pi_x(t)\right)$$

dx.doi.org/10.1021/ct3000052 | J. Chem. Theory Comput. 2012, 8, 1580−1588

where the rate constants k_{xy} are the transition probabilities P_{xy} of the starting discrete time Markov chain, and the mean time to escape from a state x is equal to the inverse of the probability $\sum_{y\neq x}P_{xy} = 1 - P_{xx}$ to move from x to another state. A comparison between the mean escape time from state x of the reduced chain and the mean escape time from the subset of nodes denoting the same state x in the original (not reduced) chain is an interesting criterion in order to establish the quality of the approximation, where a good approximation requires similar values. As we will see, such a request is a necessary condition for a well-known assumption made in TST and Kramers' theory, namely, the assumption that a probability distribution of configurations belonging to a state maintain a local equilibrium form at all times. In other words, the ratio between the probabilities associated with two different configurations belonging to the same state does not change over time. By looking at the dynamic of the process, such an assumption is equivalent to ask that the content of a state must relax to equilibrium much faster than the mean time of leaving that region.[3] The role of this assumption is to neglect every deviation from thermal equilibrium distribution (namely, the Boltzmann distribution). A similar assumption is at the base of Kramers' theory: from the nonlinear dynamics of the model, a time scale separation emerges for values of the barrier height much greater than thermal energy k_BT. In that case, the random frictional force is acting as a small perturbation and the particle will have the time to equilibrate on minima of the potential well before the accumulated action of the random force will drive it over the barrier into a neighboring state. If there is no separation of time scales (that is, when the barrier height is of the order of k_BT), a rate description is not suitable.[4] Both assumptions consist of a separation between the time scale for the system to equilibrate inside a state, and the time scale to escape from it. The justification of such an assumption is generally contingent on a good partitioning of the configuration space in states.[31−34]

In this section, we present how to analyze such a separation of time scales by means of cut-based free energy. As mentioned above, in Kramers' theory, one way to calculate the rate $k_{A,B}$ consists of evaluation of the mean time τ_A to escape from state A. The equivalence of mean escape time and Kramers' rate could be easily motivated by the following reasoning. For a given ergodic Markov chain of transition matrix P, a subset A of its nodes ($B \equiv V - A$) and a probability vector v defined on nodes belonging to A, the mean time τ_A to escape from A is calculated as described in ref 17. Let P_A be the submatrix of P containing the transition probabilities of the nodes inside A, and N be the fundamental matrix of the associated absorbing Markov chain ($N \equiv (\mathrm{II} - P_A)^{-1}$); then the mean time $\tau_A(v)$ to escape from A, starting from the distribution v, is

$$\tau_A(v) = \sum_{i\in A} v_i \sum_{j\in A} N_{ij}\varepsilon_j$$

where ε is the column vector with all entries equal to 1. The vector $Z_i(v)$ ($Z_i(v) = \sum_{j\in A} v_j N_{ji}$) gives the mean number of times that the process is in state $i \in A$ before leaving region A and is therefore proportional to the steady distribution in A related to the absorbing process starting with the initial distribution v. Noting that $\tau_A(v) = \sum_{i\in A} Z_i(v)$ and $\sum_{i\in A, j\in B} Z_i(v)P_{ij} = 1$,[17] we easily recognize that

$$\frac{1}{\tau_A(v)} = \frac{\sum_{\substack{i\in A\\ j\in B}} Z_i(v)P_{ij}}{\sum_{i\in A} Z_i(v)} \equiv k_{A,B}(v)$$

where the last equivalence is established noting that the mathematical form of the central term is equivalent to the Kramers' rate constant, defined as the net flux out of A normalized by the population inside A.[4,35] A more detailed and general proof of the equivalence between mean escape time and Kramers' rate is presented in ref 35.

The assumption of a separation between the time scales of equilibration inside state A and escaping from it is formalized here assuming that the quantities $Z_i(v)$ do not depend on the starting distribution v and the ratio between any two of them is equal to the ratio between the steady-state probabilities π_i corresponding to the same nodes, in formula

$$\frac{Z_i}{Z_j} = \frac{\pi_i}{\pi_j} \qquad \forall\ i, j$$

From this assumption, we recognize that the ratio $\gamma = \pi_j/Z_j$ does not depend on node j and we have

$$k_{A,B} = \frac{\gamma \sum_{\substack{i\in A\\ j\in B}} Z_iP_{ij}}{\gamma \sum_{i\in A} Z_i} = \frac{\sum_{\substack{i\in A\\ j\in B}} \pi_iP_{ij}}{\sum_{i\in A} \pi_i} = \frac{Q_{A,B}}{\pi_A}$$

The escape time τ_A could be compared with the value $\pi_A/Q_{A,B}$, if the difference

$$\ln\left(\frac{Q_{A,B}}{\pi_A}\right) - \ln\left(\frac{1}{\tau_A}\right)$$

does not approach zero, then the assumption $Z_i/Z_j = \pi_i/\pi_j$ does not hold; hence, we discard the hypothesis of separated time scales. Note that the equality $-\ln(Q_{A,B}/\pi_A) = \ln \tau_A$ is necessary but not sufficient for the separated time scales condition: there are cases where the equality is true but the time scales are not separated, and cases where the equality holds only under a limit operation (see the Examples section).

EXAMPLES

A Degenerate State. Here, we present an example of the arguments introduced in the last section. We define an ergodic Markov chain with a set of three states; we then show how, depending on the transition probabilities, it is possible to face a situation in which the equality $-\ln(Q_{A,B}/\pi_A) = \ln \tau_A$ holds without regard for the separated time scales condition, or a situation such that the above equality holds only under a limit condition (the same condition with which we have the time scales separation).

Let define the set of states $V \equiv \{1,2,3\}$ and the transition matrix

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1-\delta-\omega & \delta & \omega \\ \delta & 1-\delta-\eta & \eta \\ \omega & \eta & 1-\omega-\eta \end{pmatrix} \end{array}$$

where $\delta > \omega > \eta$. The Markov chain described by P is irreducible and aperiodic; it is easy to see that the unique steady state is $\pi = {}^1/_3(1, 1, 1)$ and that the detailed balance condition

$\pi_i P_{ij} = \pi_j P_{ji}$ holds. The matrix of rate probabilities $Q_{ij} \equiv \pi_i P_{ij}$ is then $Q = {}^1/_3 P$.

Because of the small size of the problem, the minimum cut matrix $Q^\dagger$ is easily calculated by examining the values of all possible cuts, which results in

$$Q^\dagger = \begin{pmatrix} 0 & \delta+\eta & \eta+\omega \\ \delta+\eta & 0 & \eta+\omega \\ \eta+\omega & \eta+\omega & 0 \end{pmatrix}$$

From $Q^\dagger$, the free energy of the TS between any couple of nodes is defined as $G^\dagger = -\ln Q^\dagger$ (where, here, $-\ln 0$ is defined as 0). From the dendrogram associated with $G^\dagger$ (see Figure 2, top), it is possible to see the three possible ways to subdivide the chain in states according to the kinetics of the system. Here, we consider the partition scheme $\{\{1,2\},\{3\}\}$ and we define the cut (A,B), where $A \equiv \{1,2\}$ and $B \equiv \{3\}$. The cut value is $Q_{A,B} = {}^1/_3(\omega + \eta)$; we then have

$$\frac{Q_{A,B}}{\pi_A} = \frac{\omega+\eta}{2}$$

The mean escape time τ_A from region A is defined as $\tau_A \equiv \nu N \varepsilon$, where ν is the probability distribution of the starting nodes (here, we choose $\nu \equiv {}^1/_2(1, 1)$ for nodes in A), ε is the column vector with all entries 1, and N is the fundamental matrix of the absorbing Markov chain:

$$N \equiv (\mathrm{II} - P_A)^{-1} = \frac{1}{(\delta+\omega)(\delta+\eta) - \delta^2}\begin{pmatrix} \delta+\eta & \delta \\ \delta & \delta+\omega \end{pmatrix}$$

The resulting mean escape time is $\tau_A = (4\delta + \eta + \omega)/(2\delta\eta + 2\delta\omega + 2\eta\omega)$.

Some observations are needed. The probabilities ω and η are responsible for the transitions between A and B, so different values of them cause different scenarios. For example, in the case $\omega = \eta$, we obtain $Q_{A,B}/\pi_A = {\tau_A}^{-1} = \omega$, regardless of the value of δ, which means that there could be no separation between the time to equilibrate inside A and the time to escape from it. As already mentioned, the equality $Q_{A,B}/\pi_A = {\tau_A}^{-1}$ is not sufficient for the condition of separated time scales.

There are also cases where the above equality holds only under a limit operation (for example, in the case of $\eta = 0$ and $\omega \to 0$). In this situation ($\eta = 0$, $\omega > 0$, as depicted in the lower part of Figure 2), we have

$$-\ln\left(\frac{Q_{A,B}}{\pi_A}\right) = -\ln\left(\frac{\omega}{2}\right)$$

$$-\ln\left(\frac{1}{\tau_A}\right) = \ln\left(\frac{4\delta+\omega}{2\delta\omega}\right)$$

$$\Delta \equiv \ln\left(\frac{Q_{A,B}}{\pi_A}\right) - \ln\left(\frac{1}{\tau_A}\right) = \ln\left(\frac{Q_{A,B}}{\pi_A}\tau_A\right) = \ln\left(1 + \frac{\omega}{4\delta}\right)$$

The above equalities show that the ratio between $Q_{A,B}/\pi_A$ and ${\tau_A}^{-1}$ goes to 1 for $\omega \to 0$; in the same limit, we obtain the time scales separation. It is easy to see that, with $\delta = 0.5$ and $\Delta < 0.05$, we have $\omega < 0.1$ and $-\ln Q_{A,B}/\pi_A > 2.97$ (in units of $k_B T$). Hence, within this system, which could be considered to be composed of a degenerate state (made of two states) and another state, the approximation $Q_{A,B}/\pi_A \simeq {\tau_A}^{-1}$ (the difference is Δ) holds for a barrier $\Delta G_{A,B} > 3k_B T$. Under such conditions ($\delta = 0.5$ and $\omega = 0.1$), the chain could be reduced in the two-state system described by the master equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_A(t) = \frac{\omega}{2}\pi_A(t) - \omega\pi_B(t)$$

and is governed by the exponential distribution $P_{A,B}(t) = (1 - \exp(-\gamma t/\tau_A)$, denoting the probability to see a jump from A to B in the interval of time $(0,t)$. Note that γ has units of inverse of time and it defines the time scale between the number of steps in the Markov chain and the corresponding time t for the Markov process: $n = \gamma t$.

A Complex MSM. Here, we apply the cut-based analysis on the MSM describing the reversible folding of the 20-residue three-stranded antiparallel β-sheet peptide studied in ref 22. The molecular dynamics sampling has been clustered according to backbone dihedral angles values, by means of a hierarchical algorithm[36] implemented in the molecular modeling package CAMPARI.[37] The resulting network has $V = 157\,380$ nodes and $E = 329\,011$ edges; it has been analyzed with PYKOV (a Markov chain Python module).[38] Because of the size of the network, instead of calculating the V^2 mincuts (which is a problem of complexity $O(V^4)$),[39] we selected the reference state by means of the cbFEP method[16] with ordering of the nodes according to mean first passage time to the folded state. This procedure is motivated by the fact that the cbFEP offers an approximated solution for the problem to collect nodes in states, it is an approximation because there is no guarantee to find the minimum cut and so the free energy of the TS calculated by cbFEP is lower or equal to the free energy of the TS derived from the minimum cut ($Q^\dagger_{A,B|i,j}$). At the cut (A,B), located at $\pi_A \simeq 0.33$ and separating the reference state A from all the rest B (see top of Figure 3), the TS and the activation free energies have values (in units of $k_B T$) of $G_{A,B} \simeq 5.69$ and $\Delta G_{A,B} = G_{A,B} - G_A \simeq 4.57$, respectively. Let ν be the restriction of π on the state A,

$$\nu_i = \frac{\pi_i}{\pi_A} \qquad \forall\, i \in A$$

then the mean escape time from state A, calculated by

$$\tau_A(\nu) = \sum_{i\in A} \nu_i \sum_{j\in A} N_{ij}\varepsilon_j$$

is $\tau_A \simeq 104$ steps (around 2 ns as the saving frequency was 20 ps), and its logarithm ($\ln \tau_A \simeq 4.64$) is similar to the activation free energy $\Delta G_{A,B}$, since the difference is much smaller than their absolute values. Moreover the distance $|Z(\nu)/\tau_A - \nu|$ (see below) is <0.02, where the vector $Z_i(\nu) = \sum_{j\in A} \nu_j N_{ji}$ indicates the mean number of times that the system is in node i before escaping from A. These results suggest that the system spends much less time to equilibrate inside state A than the time needed to escape from it. This observation can be further validated by means of the mixing time of state A, which is calculated by the following procedure. The subnetwork related to state A is extracted from the entire network. Since this subnetwork is not ergodic we then used its largest strongly connected component[40] as the network representing state A, which covers the 99% of the extracted network and, upon normalization, defines the MSM $\tilde{P}$ related to state A. We then

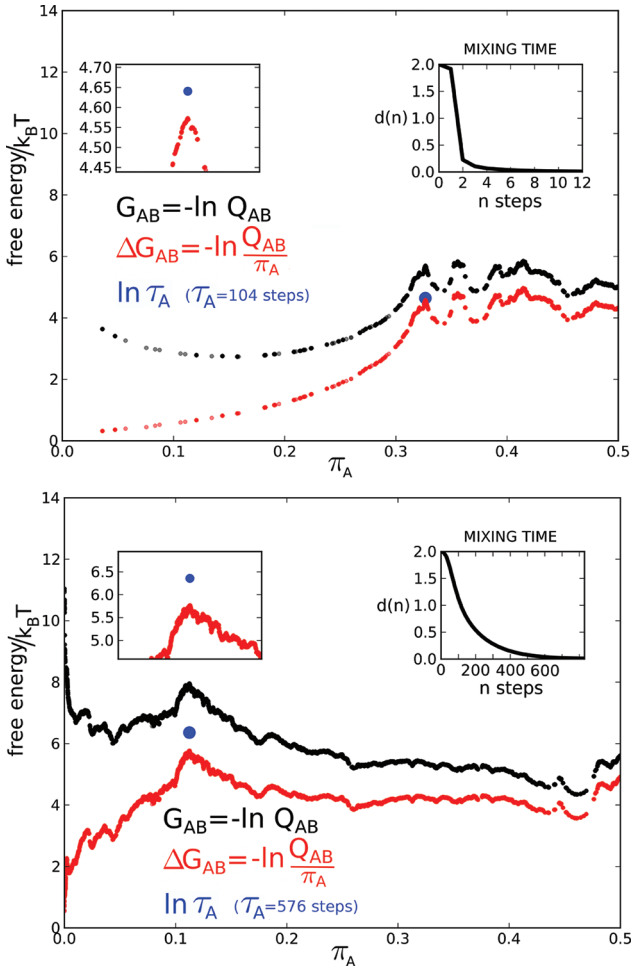

Figure 3. Application to a MSM of peptide folding. (Top) The folded state of a 20-residue β-sheet peptide was determined by the cbFEP.[22] The logarithm of the mean time to exit from the native state (blue dot) is essentially identical to the height of the activation free energy (red profile), suggesting that the system spends much less time to equilibrate inside the state than the time needed to escape from it. The time to mix within the native state (inset on the right) is much smaller than the mean exit time τ_A, as expected. (Bottom) Cut (A,B) of a non-native region of the phase space that has helical secondary structure content and is stabilized entropically.[22] The logarithm of the mean escape time from A does not overlap with the peak of the activation free energy, which implies that the system escapes before reaching equilibrium inside A. The mixing time (inset on the right) is not significantly shorter but rather similar to the mean escape time τ_A.

calculated the mixing time of $\tilde{P}$ in the following way: given the steady state $\tilde{\pi}$ of $\tilde{P}$, we define the initial v as the probability vector having value 1 at the least-probable node of $\tilde{\pi}$ and zeros for all the others. We then iterate vector v, $v(n) \equiv v\tilde{P}^n$, and for every n, we calculate the distance from the steady state:

$$d(n) \equiv |v(n) - \tilde{\pi}| \equiv \sum_i |v(n)_i - \tilde{\pi}_i|$$

($d(n)$ is weakly monotonically decreasing in n: $|v(n) - \tilde{\pi}| \geq |v(n+1) - \tilde{\pi}|$). As shown in Figure 3 top, $d(n)$ reaches zero in few steps, and defining the mixing time τ_{mix} as the smaller n such that $d(n) < 0.25$, we have that the time to mix inside state A is much smaller than the mean time to exit from it: $\tau_{mix} \simeq 2$ steps $\ll \tau_A$. Note that state A is the native state, but we refer to ref 22 for a description of the structures according to the position of the cut and for a detailed characterization of the free energy surface of folding of this β-sheet peptide.

In the bottom portion of Figure 3, the same analysis is performed on a different cut (A,B) concerning the helical state, a non-native region of the phase space, which is stabilized mainly by entropy.[22] In this case, the logarithm of the mean escape time from region A is not comparable with the activation free energy, so we do not expect a mixing time much smaller than the escape time. The two time scales are indeed of the same order. It also emerges that the distance $|Z(v)/\tau_A - v| \simeq 0.2$, which provides further evidence that the time to equilibrate within this non-native region is not negligible.

Cut-Based Free Energy Profile (cbFEP) of the Free Energy of Activation. The most significative information derived from the cbFEP analysis is contained in the peak coordinates of the first barrier. The x-value is the probability π_A of the state A delimited by the barrier, while the y-value represents either the free energy of the TS ($G_{A,B}$) or the free energy of activation ($\Delta G_{A,B}$) to exit from state A. In the previous example of a complex MSM, we showed how the cbFEP analysis of the free energy of activation ($\Delta G_{A,B}$), compared to the calculation of the mean exit time and mixing time of state A, is a useful tool to check for the existence of a metastable state. In fact, such comparison is able to evaluate the separation of time scales for equilibration within and exit from state A.

The cbFEP defined by $\Delta G_{A,B}$ is also advantageous, with respect to the profile of $G_{A,B}$, for another reason. While an exhaustive sampling of the phase space is required for the values of π_A and $G_{A,B}$ to be meaningful, this is not necessary for $\Delta G_{A,B}$. Indeed, being the free energy of activation the logarithm of the conditional probability $Q_{A,B}/\pi_A$, the sampling of the phase space far from the region of interest (i.e., far from the state A) is not required (see Figure 4). We assume here that the

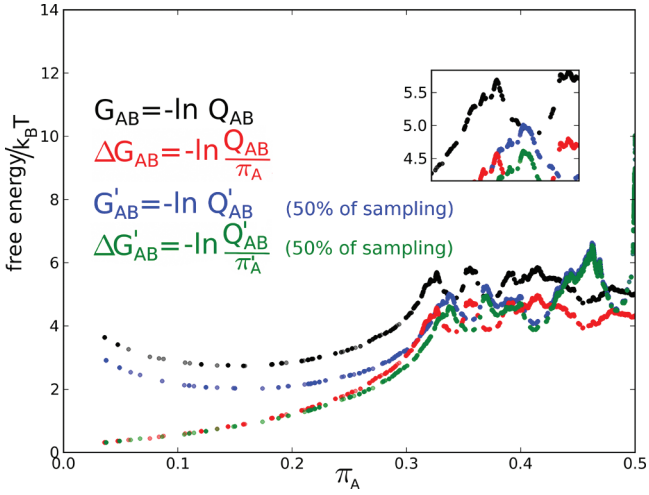

Figure 4. The free energy of activation is not affected by an incomplete sampling of the phase space. All the nodes located before $\pi_A = 0.5$ in the red (or black) profile define a subnetwork with half of the sampling of the original MSM. The subnetwork contains the native state and its barrier, but it does not contain the non-native region considered in the bottom of Figure 3. The quantity $Q'_{A,B}$ is the flow through the cut (A,B) calculated on the subnetwork (blue and green profiles). The figure shows that, at the peak of the first and most relevant barrier (inset on the figure), $G'_{A,B}$ is significantly different from $G_{A,B}$, whereas $\Delta G'_{A,B}$ is similar to $\Delta G_{A,B}$. The x-axis range of $G'_{A,B}$ and $\Delta G'_{A,B}$ is rescaled by a factor of 0.5, to better compare the cbFEPs.

dx.doi.org/10.1021/ct3000052 | J. Chem. Theory Comput. 2012, 8, 1580−1588

lack of a complete sampling entails that the calculated probabilities π_A and $Q_{A,B}$ can be approximated by the correct ones by a rescaling factor α, i.e., we are assuming the equalities $\pi_A = \alpha\pi'_A$ and $Q_{A,B} = \alpha Q'_{A,B}$. This assumption is justified by the fact that the maximum likelihood probability $Q_{A,B}$, as well as π_A, is defined as the ratio $N_{A,B}/N$ between the number $N_{A,B}$ of observed transitions through the cut (A,B) and the total number N of transitions.[40] Whereas N is affected by a lack of sampling of noninteresting phase space regions, because the exploration of a restricted region needs less sampling, the amount $N_{A,B}$ of observed transitions through the cut must remain constant. In formulas, we have that $Q_{A,B} \equiv N_{A,B}/N = N_{A,B}/(N' - k)$, where N' indicates the total number of observed transitions for a sampling of a larger phase space region and k is the difference $N' - N$. The above assumption is now easily recovered, noting that $N_{A,B}/(N' - k) = \alpha Q'_{A,B}$, with $\alpha = N'/(N' - k)$. This heuristic reasoning suggests that, while $G_{A,B}$ and π_A are each individually affected by the lack of sampling, under the above assumption, $\Delta G_{A,B}$ is not influenced, because the α factors at the numerator and denominator of the $Q_{A,B}/\pi_A$ quotient cancel out.

CONCLUSION

Markov state models (MSMs) offer a relatively easy and powerful mathematical framework within which to define and analyze the kinetics of a complex system. The cut-based analysis of a MSM is based on the evaluation of the minimum cut between two nodes i and j, calculated as the cut (A,B) separating node $i \in A$ from node $j \in B$ with minimum flow through it. The cut-based analysis has, as the main objective, the study of the free energy surface to derive kinetic observables. We have shown here that the cut-based free energy of the transition state (TS) defines an ultrametric distance on the set of nodes of an ergodic MSM, without the assumption of the detailed balance condition. This property offers a way to collect nodes in states, which is a simplification procedure motivated by the kinetics of the system. The time scale to equilibrate inside a state is expected to be much smaller than the time to escape from it, and such difference can be checked by analytical calculations on the MSM. Kinetic observables like the free energy of the TS, the activation free energy, and the rate constants are directly derived from the cut-based analysis. In the same direction, here, we have proposed a novel definition of the cut-based free energy profile (cbFEP) that allows one to check for the separation of time scales for equilibration within and exit from a state. In the framework of protein dynamics, the final target of such analysis is to compare simulation results with the corresponding experimental observables, and, in this sense, it was our intention to make explicit the existence of a parallelism between cut-based quantities (TS and activation free energies) and the concepts at the base of transition state theory (TST) and Kramers' theory.

Lastly, it is essential to mention the wide applicability of the ultrametric distance defined here; given that it is defined on the set of nodes of an ergodic Markov chain, its application is not dependent on the physical system under study and could prove to be useful in fields far from protein folding, e.g., material sciences and bioinformatics.

AUTHOR INFORMATION

Corresponding Author

*E-mail: caflisch@bioc.uzh.ch.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Andreas Vitalis for interesting discussions. This work was supported by a grant of the Swiss National Science Foundation to author A.C.

REFERENCES

(1) Zwanzig, R. *Proc. Natl. Acad. Sci., U.S.A.* **1997**, *94*, 148−150.
(2) Zwanzig, R. *Proc. Natl. Acad. Sci., U.S.A.* **1995**, *92*, 9801−9804.
(3) Zwanzig, R. *Nonequilibrium Statistical Mechanics*; Oxford University Press: New York: 2001.
(4) Hänggi, P.; Talkner, P.; Borkovec, M. *Rev. Mod. Phys.* **1990**, *62*, 251−341.
(5) Kramers, H. A. *Physica* **1940**, *7*, 284.
(6) Pande, V. S. *Phys. Rev. Lett.* **2010**, *105*, 198101.
(7) 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.
(8) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(9) Keller, B.; Hünenberger, P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2011**, *7*, 1032−1044.
(10) Buchete, N.-V.; Hummer, G. *Phys. Rev. E* **2008**, *77*, 030902.
(11) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.
(12) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99−105.
(13) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci., U.S.A.* **2004**, *101*, 14766−14770.
(14) McGarrity, E. S.; Duxbury, P. M.; Holm, E. A. *Phys. Rev. E* **2005**, *71*, 026102.
(15) Gfeller, D.; De Los Rios, P.; Caflisch, A.; Rao, F. *Proc. Natl. Acad. Sci., U.S.A.* **2007**, *104*, 1817−1822.
(16) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689−12698.
(17) Kemeny, J. G.; Snell, J. L. *Finite Markov Chains*; Springer−Verlag: Berlin, Germany, 1976.
(18) Jungnickel, D. *Graphs, Networks and Algorithms*; Springer−Verlag: Berlin, Germany, 2005.
(19) Ford, L.; Fulkerson, D. R. *Can. J. Math.* **1956**, *8*, 399−404.
(20) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(21) Rammal, R.; Toulouse, G.; Virasoro, M. A. *Rev. Mod. Phys.* **1986**, *58*, 765−788.
(22) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701−8714.
(23) Schuetz, P.; Wuttke, R.; Schuler, B.; Caflisch, A. *J. Phys. Chem. B* **2010**, *114*, 15227−15235.
(24) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495−1517.
(25) Wales, D.; Miller, M.; Walsh, T. *Nature* **1998**, *394*, 758−760.
(26) Krivov, S. V.; Karplus, M. *J. Chem. Phys.* **2002**, *117*, 10894−10903.
(27) Rylance, G. J.; Johnston, R. L.; Matsunaga, Y.; Li, C.-B.; Baba, A.; Komatsuzaki, T. *Proc. Natl. Acad. Sci., U.S.A.* **2006**, *103*, 18551−18555.
(28) Gomory, R.; Hu, T. *J. Soc. Ind. Appl. Math.* **1961**, *9*, 551−570.
(29) Liggett, T. M. *Continuous Time Markov Processes: An Introduction*; Cox, D., Krantz, S., Mazzeo, R., Scharlemann, M., Eds.; Graduate Studies in Mathematics, Vol. 113; American Mathematical Society: Providence, RI, 2010.
(30) Rozanov, Y. *Introduction to Random Processes*; Springer−Verlag: Berlin, Germany, 1982.
(31) Bovier, A.; Eckhoff, M.; Gayrard, V.; Klein, M. *Commun. Math. Phys.* **2002**, *228*, 219−255.
(32) Larralde, H.; Leyvraz, F. *Phys. Rev. Lett.* **2005**, *94*, 160201.

(33) Avetisov, V.; Bikulov, A. *Tr. Math. Inst. Steklova* **2009**, *265*, 82−89.
(34) Beltrán, J.; Landim, C. *Stoch. Proc. Appl.* **2011**, *121*, 1633−1677.
(35) Reimann, P.; Schmid, G. J.; Hänggi, P. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1999**, *60*, R1−R4.
(36) Vitalis, A.; Caflisch, A. *J. Chem. Theory Comput.* **2012**, *8*, 1108−1120 (DOI:10.1021/ct200801b).
(37) Vitalis, A.; Pappu, R. *Ann. Rep. Comput. Chem.* **2009**, *5*, 49−76.
(38) Scalco, R. page http://riccardoscalco.github.com/Pykov/ (accessed February 20, 2012).
(39) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; Stein, C. *Introduction to Algorithms*; MIT Press: Cambridge, MA, 2001.
(40) Scalco, R.; Caflisch, A. *J. Phys. Chem. B* **2011**, *115*, 6358−6365.