

Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states

Frank Noé

Computational Molecular Biophysics Group, Interdisciplinary Center for Scientific Computing (IWR), Im Neuenheimer Feld 368, 69120 Heidelberg, Germany

Illia Horenko and Christof Schütte

Scientific Computing Group, Free University of Berlin (FU), Arnimallee 2-6, 14195 Berlin-Dahlem, Germany

Jeremy C. Smith

Computational Molecular Biophysics Group, Interdisciplinary Center for Scientific Computing (IWR), Im Neuenheimer Feld 368, 69120 Heidelberg, Germany and Center for Molecular Biophysics, University of Tennessee/Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008, Oak Ridge, Tennessee 37831-6255

(Received 23 October 2006; accepted 13 February 2007; published online 19 April 2007)

Molecular dynamics simulation generates large quantities of data that must be interpreted using physically meaningful analysis. A common approach is to describe the system dynamics in terms of transitions between coarse partitions of conformational space. In contrast to previous work that partitions the space according to geometric proximity, the authors examine here clustering based on kinetics, merging configurational microstates together so as to identify long-lived, i.e., dynamically metastable, states. As test systems microsecond molecular dynamics simulations of the polyalanines Ala₈ and Ala₁₂ are analyzed. Both systems clearly exhibit metastability, with some kinetically distinct metastable states being geometrically very similar. Using the backbone torsion rotamer pattern to define the microstates, a definition is obtained of metastable states whose lifetimes considerably exceed the memory associated with interstate dynamics, thus allowing the kinetics to be described by a Markov model. This model is shown to be valid by comparison of its predictions with the kinetics obtained directly from the molecular dynamics simulations. In contrast, clustering based on the hydrogen-bonding pattern fails to identify long-lived metastable states or a reliable Markov model. Finally, an approach is proposed to generate a hierarchical model of networks, each having a different number of metastable states. The model hierarchy yields a qualitative understanding of the multiple time and length scales in the dynamics of biomolecules. © 2007 American Institute of Physics. [DOI: [10.1063/1.2714539](https://doi.org/10.1063/1.2714539)]

I. INTRODUCTION

Given the large amount of molecular dynamics (MD) simulation data available for macromolecules,^{1,2} models are needed to analyze the data that capture the essential dynamical features and can be physically interpreted. One approach is a “transition network,” which encodes the transitions of the system between a set of discrete states.³ In a transition network, a vertex corresponds to a discrete conformational state, and an edge corresponds to a transition between two states. Each edge may be weighted by the associated transition rate or probability. An equivalent description is that of a Markov state model,⁴ in which the interstate transitions are defined by a transfer matrix, $\mathbf{T}(\tau)$, whose elements $T_{ij}(\tau)$ specify the probability of undergoing a transition from state i to state j within time τ .

Transition networks and Markov state models are also useful in merging the combined dynamical information provided by ensemble dynamics, i.e., a large collection of relatively short MD trajectories.^{1,2} Using such a representation, master-equation dynamics allows the dynamics of processes to be recovered whose characteristic time scales are much

beyond the individual simulation times.^{5–16} Graph-theoretical algorithms can then be used to identify the best transition pathways connecting, and the transition state ensemble separating, a given pair of conformations.³ The information in the network can also be used to alleviate the ubiquitous sampling problem in molecular simulation by planning simulations so as to yield the properties of interest with a reduced amount of computing time.^{3,17,18}

A central question when generating interstate models is how best to partition state space into discrete states. It is often desirable that the partition reflects the dynamical behavior of the system. In particular, biomolecular function often depends on the ability to undergo transitions between long-lived or “metastable” states. These interstate transitions are rare events in which a molecular system stays for long periods of time within one state, before rapidly switching to another. From the geometrical point of view, a metastable state corresponds to a connected subset of the state space within which the system may vibrate or diffuse on a relatively fast time scale,¹⁹ surrounded by a boundary region across which transitions occur rarely. This is equivalent to

stating that a metastable state corresponds to a free energy basin surrounded by energy barriers which stabilize the system within the basin. Analysis methods, such as master-equation dynamics, usually require that the transitions between states are Markovian, i.e., the evolution of the system into the next state depends only on the current state, and there is no memory affecting the interstate transitions. The validity of this assumption requires that the states be chosen such that their mean lifetime is long compared to the length of the memory, which may be due, for example, to the presence of intrastate barriers. Therefore, a good definition is one that minimizes the flux between the states (and or maximizes the lifetimes of the states). Having defined metastable states in this way, a kinetic model for the biomolecule may be obtained by neglecting the fast motions within metastable states and only modeling the jump process between them.

It is a common practice to define distinct discrete states by clustering similar geometries that are visited in simulation. This may be done either manually by inspecting low-dimensional projections of the trajectory on essential, i.e., principal, coordinates,^{1,20,21} or automatically by a geometric clustering method.^{4,22–27} The geometric approach is useful in identifying dynamically metastable states only if the geometric and kinetic distances are similar for any given pair of states. Unfortunately this condition is not always met for biomolecules, as considerable free energy barriers may lie between geometrically close conformational states. This is the case, for example, in enzyme reactions, where the process of interest is very localized and, at physiological temperature, may lead to smaller geometric deviations than the vibrational motion of the rest of the system. Furthermore, geometrically distant conformations can be sometimes connected by relatively fast transitions, as can be the case for different unfolded states of a polypeptide.²⁸ Thus, in order to obtain a well-behaved kinetic model for the interstate transitions, a method for identifying discrete states should be based on “kinetic,” rather than geometric clustering.

An approach to kinetic clustering is known in the topological mapping literature where a transition network is defined based on the minima and transition states of an energy landscape. In this approach, basins are grouped together if they are connected by low free energy barriers.^{7,14,28,29} For the analysis of dynamical data, however, it is more straightforward to use an approach which directly makes use of the trajectory, without having to construct a free energy surface. Basic theory for such data-based approaches has been derived in the past few years,^{30,31} but applications have been as yet limited to small test systems.

In the present work, we use dynamical data-based kinetic clustering to decompose the dynamics of macromolecules undergoing complex conformational changes. We examine the relevance of metastability in the folding dynamics of polypeptides and whether metastability analysis is useful in providing insight into the thermodynamics and kinetics of such processes. The approach proposed here uses the improved Perron cluster cluster analysis (PCCA) method,^{32,33} which merges small partitions of conformational space using information from the Markov transition matrix so as to maximize the lifetimes of the clusters, which then represent ap-

proximations to the ideal metastable states. For this, we examine folding and unfolding processes in microsecond molecular dynamics simulations of 8- and 12-mer polyalanines with charged termini. Two coordinate systems which might in principle be useful for identifying metastable states are compared: the backbone torsion rotamer pattern and the hydrogen-bonding pattern. An approach for determining optimal numbers of metastable states is introduced. A metastable state decomposition is conducted for both rotamer and hydrogen-bond coordinate systems. For the test systems studied here, torsional rotamers prove to be much more useful for the identification of metastable states than is the hydrogen-bonding pattern.

Transition networks describing the interstate kinetics are constructed. For each given molecule and coordinate system, several transition networks are computed, using different numbers of metastable states, and a state hierarchy is thus obtained. For both polyalanines, a complex network of metastable states is identified. Use of only a few metastable states (e.g., 2, 3, etc.) leads to the distinction between the “main state” (the basin with the globally minimal energy and several “kinetic traps” which are thermally accessible and have half-lifetimes similar to, or greater than, that of the main state. By increasing the number of metastable states the main state is further decomposed into conformational subsets. An interesting feature of the networks is that some states which are structurally very similar are not kinetically contiguous, but rather interchange via structurally very different intermediates. This clearly demonstrates the usefulness of the kinetic clustering method in biomolecular simulation studies.

II. METHODS

A. Energy function and system setup

Equilibrium MD simulations for the peptides Ala₈ and Ala₁₂ with zwitterionic termini were conducted at a temperature of $T=300$ K using temperature rescaling. The CHARMM energy function³⁴ was used with parameter set 19 to model the intramolecular interactions and the solvent was modeled with the ACE2 implicit solvent model.³⁵ A switch function was used to smoothly truncate the nonbonded interaction energies between 8 and 12 Å. For each system, four independent simulations were performed in the following manner. A starting conformation was generated by setting the backbone ϕ/ψ angles to random values. This structure was energy minimized to a root mean square energy gradient of 10^{-3} kcal/mol Å in three subsequent stages: 1000 steps of steepest descent minimization, 1000 steps of conjugate gradient minimization, and, finally, 4000 steps of Newton-Raphson minimization. The system was then heated to $T=300$ K during 10 ps. A local equilibration (to relax the system within its local conformation) was conducted for 30 ps, followed by a nonconstrained production run of 1 μ s. An integration time step of 1 fs was used with a Verlet integrator. Thus, the total trajectory length was 4 μ s for each system. A coordinate set was recorded every $\tau=50$ fs and used for the subsequent analysis, corresponding to a total number of 8×10^7 structures for each system.

The simulations above were performed in implicit solvent for the pragmatic reason of quickly generating long-time trajectories with a large number of conformational transitions so as to test the analysis methods presented in this paper. In cases in which it is desired to make experimentally testable predictions on the actual biomolecular system studied, a more physically motivated setup should be used, such as an explicit solvent simulation with periodic boundary conditions and Ewald summation to treat the nonbonded interactions.

B. Definition of discrete microstates

In this first step of processing, the configurational space of the molecule is divided into a relatively fine discrete and

$$\mathbf{Tc}(\tau) = \begin{bmatrix} 2 \#(1 \rightarrow 1) & \cdots & \#(1 \rightarrow m) + \#(m \rightarrow 1) \\ \vdots & \ddots & \vdots \\ \#(m \rightarrow 1) + \#(1 \rightarrow m) & \cdots & 2 \#(m \rightarrow m) \end{bmatrix},$$

where $\#(a \rightarrow b)$ is the number of transitions from microstate a to microstate b in the trajectory. The entries of \mathbf{Tc} count both the transitions and back transitions between two microstates, assuming that detailed balance holds. This ensures that \mathbf{Tc} is exactly symmetric and thus has only real eigenvalues. In fact, for all the examples used in this study, \mathbf{Tc} was found to be nearly symmetric, even when transitions are counted in only one direction, indicating that the simulations are very close to equilibrium. From the transition count matrix we construct a transfer matrix \mathbf{T} by normalizing the entries of \mathbf{Tc} such that the sum over a row is unity: $T_{ab} = Tc_{ab} / \sum_{k=1}^m Tc_{ak}$. The resulting transfer matrix \mathbf{T} (also known as row stochastic transition matrix),

$$\mathbf{T}(\tau) = \begin{bmatrix} p(1 \rightarrow 1) & \cdots & p(1 \rightarrow m) \\ \vdots & \ddots & \vdots \\ p(m \rightarrow 1) & \cdots & p(m \rightarrow m) \end{bmatrix}, \quad (1)$$

contains, as element $p(a \rightarrow b)$, the conditional probability of moving from state a to state b within the time step τ for which \mathbf{T} was set up, i.e.,

$$T_{ab}(\tau) = p(a \rightarrow b) = p[i(t + \tau) = b | i(t) = a].$$

This microstate transfer matrix is used for subsequent kinetic clustering, explained below. Like \mathbf{Tc} , \mathbf{T} has purely real eigenvalues. If the transitions between microstates are Markovian (memoryless), then the time evolution of the probability distribution of microstates, $\mathbf{p}(t)$, can be expressed as

$$\mathbf{p}(t + \tau) = \mathbf{p}(t)\mathbf{T}(\tau). \quad (2)$$

The choice of coordinates used to obtain the discrete microspace is critical. These coordinates must be appropriate for characterizing the process of interest, i.e., they should

finite set of states, called here *micro-states*. This term is just used here as a convenient abbreviation for “small parts of configuration space” and should not be confused with the meaning of microstate as a point in phase space. Each microstate that is actually visited during the trajectories of a given system is assigned an integer index $i \in \{1, \dots, m\}$. Thus, each molecular configuration in the trajectory has an associated microstate index. We transform the trajectory of molecular configurations $[\mathbf{x}(0), \mathbf{x}(\tau), \mathbf{x}(2\tau), \dots]$ into a trajectory of discrete state indices $[i(0), i(\tau), i(2\tau), \dots]$. This discrete trajectory is used to construct a transition count matrix

distinguish between states that one wants to separate. In the following two sections, two alternative choices for coordinates in polypeptides will be described.

1. Torsion-rotamer-pattern microstates and independence of coordinates

Bond and angle degrees of freedom are rather stiff. Thus, the conformation of a polypeptide is defined mainly by its torsional angles. For polyalanines, which are used as the application in this study, there are no side-chain torsions, so the ϕ and ψ backbone torsions are sufficient. Before discretizing these coordinates to obtain the microstates, it must be determined whether they are statistically dependent or not. For statistically independent coordinates one may use independent projections of the trajectory on the individual coordinates, while for mutually dependent coordinates, a projection on the joint coordinate space must be used for the discretization. The statistical dependencies can be checked by computing normalized mutual information coefficients between all pairs of coordinates. Here, between torsion angles x and y we compute the measure proposed in Ref. 36,

$$\text{NMI}_{xy} = \frac{\text{MI}_{xy}}{\min\{H_x, H_y\}},$$

where the mutual information, MI_{xy} , is defined as

$$\text{MI}_{xy} = \int_x \int_y p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} dy dx,$$

where $p(x, y)$ is the joint distribution, and $p(x)$ and $p(y)$ are the marginal distributions of x and y . The information-theoretic entropies H_x , H_y are defined as

$$H_x = - \int_x p(x) \ln p(x) dx.$$

Here, the integrals are approximated by central summation over 20 equally sized bins in the dihedral range $[-180, 180]$. There are several different notions of normalized mutual information^{36,37} and the concept of a generalized correlation coefficient,³⁸ which all have the property that they are bounded by $[0, 1]$, where a value of 0 means that the random variables x and y are uncorrelated while a value of 1 means that one is determined by the other. However, these different correlation coefficients differ in the meaning of intermediate values. Here, the off-diagonal elements of NMI_{xy} were all below 0.1 for both Ala_8 and Ala_{12} , which gives confidence that the torsion angle distributions are largely statistically independent of each other in the simulation. In this case, it is possible to perform a kinetic clustering for each of these angles individually, resulting in a set of rotamers for each dihedral. The accessible rotamer patterns then define the microstate space for the whole molecule.

Here, each torsion angle, α , was discretized into 36 boxes having a range of 10° each. By counting transitions between boxes in consecutive time steps $(0, \tau, 2\tau, \dots)$, a corresponding transfer matrix $\mathbf{T}_\alpha(\tau) \in \mathbb{R}^{36 \times 36}$ was set up, each entry $T_{\alpha,uv}(\tau)$ containing the conditional probability of a transition from box u to box v in time τ . Using the PCCA method described in Sec. II C, $\mathbf{T}_\alpha(\tau)$ was clustered so as to yield a block-diagonal form. The number of clusters (and thus rotamers) for each torsion angle was chosen such that the residence probability, i.e., the probability of staying within a rotamer, was at least 0.99 for $\tau=50$ fs. This produces n_α rotamers for each torsion angle, where n_α turned out to be either 1 or 2. In most cases, the result of this first stage of clustering is identical to defining the rotamers by splitting the angle range at the local probability maxima for all angles. It is therefore not necessary to use PCCA in this first stage of clustering, yet we did so to remain consistent with the rest of the procedure.

Having identified n_α rotamers for each torsion angle, and having R torsion angles in the molecule define a torsion rotamer pattern space with up to $\prod_{\alpha=1}^R n_\alpha$ states. Out of these, only a small number of states are actually visited during the finite trajectory. For Ala_8 , 12 ϕ/ψ angles had two rotamers each, thus yielding up to $2^{12}=4096$ rotamer states. Out of these, only 457 were visited during the $4 \mu\text{s}$ dynamics trajectory. Therefore each structure along the trajectory was assigned to one of these 457 states, yielding a trajectory of microstate indices $[i(0), i(\tau), i(2\tau), \dots]$, from which a transfer matrix [Eq. (1)] was constructed. Likewise, for Ala_{12} , 20 ϕ/ψ angles had two rotamers each, thus yielding $2^{20} \approx 1 \times 10^6$ rotamer states, out of which only 3437 were visited.

2. Hydrogen-bonding-pattern microstates and transition state hysteresis

An alternative coordinate system that may be appropriate for characterizing the overall configuration of a polypeptide is the hydrogen-bonding pattern. This pattern forms a discrete space if each hydrogen bond is considered to be

either formed or not formed for each acceptor-hydrogen-donor triplet. Two practical issues need to be dealt with here.

Firstly, even stable hydrogen bonds occasionally break and quickly reform and sometimes very unlikely hydrogen bonds occasionally form for short times. These short “re-crossing” events of individual H bonds do not affect the overall structure and should thus be removed. For this, we define a transition state hysteresis as follows. An acceptor-hydrogen-donor triplet is characterized as follows: (a) as a *strong* H-bond if the donor-acceptor-hydrogen angle is less than 45° and the acceptor-hydrogen distance is less than 2.1 \AA ; (b) as a *weak* H bond if the donor-acceptor-hydrogen angle is less than 90° and the acceptor-hydrogen distance is less than 2.5 \AA ; and (c) as *no* H bond, otherwise. A contact is “switched on” only when a strong H bond is formed, and “switched off” only if no H bond is present. Transitions into the weak H-bond region are ignored until this transition region is crossed completely. This creates a transition state hysteresis which avoids taking rapid recrossings into account.

Secondly, the number of acceptor-hydrogen-donor triplets is very large. To avoid obtaining too many microstates, the number needs to be reduced to a “relevant” set of triplets used for the definition of the microstate space. Only stable hydrogen bonds contribute to relevant metastabilities. To reduce the set of contacts to a feasible size, we use only those contacts which form hydrogen bonds at least 1% of the time. For Ala_8 , this involves 11 out of 72 bonds, giving a theoretical state space size of $2^{11} \approx 2048$, out of which only 295 states were visited. For Ala_{12} , 32 out of 156 bonds are involved, giving up to $2^{32} \approx 4.3 \times 10^9$ states, from which 42 159 states were visited.

C. Kinetic clustering to obtain metastable states

A metastable state is a set of microstates in which the system stays for a long time before leaving it. Following this idea, a straightforward definition for a partition into metastable states would be the following: Given a discrete trajectory $[i(0), i(\tau), i(2\tau), \dots]$ in space of microstates $\{1, \dots, m\}$, find a partition of microstates into C metastable sets, such that the number of transitions within sets is maximized. Since the total number of transitions in the trajectory is given, this is equivalent to minimizing the number of transitions between metastable sets. This formulation of the problem is known in graph theory as a C -min-cut problem but is, unfortunately, NP hard, i.e., the computational cost to solving it increases exponentially with m .³⁹ A similar partition, however, can be obtained very efficiently with the PCCA method which was first introduced in Ref. 40. The method is illustrated in Fig. 1. Its basic idea is the following: Recall that Markov time evolution of the probability distribution, $\mathbf{p}(t)$, is given by $\mathbf{p}(T+\tau) = \mathbf{p}(t)\mathbf{T}(\tau)$. This equation can also be expressed in terms of an expansion in left eigenvectors of \mathbf{T} , \mathbf{q}_i , and corresponding eigenvalues, λ_i .⁴¹

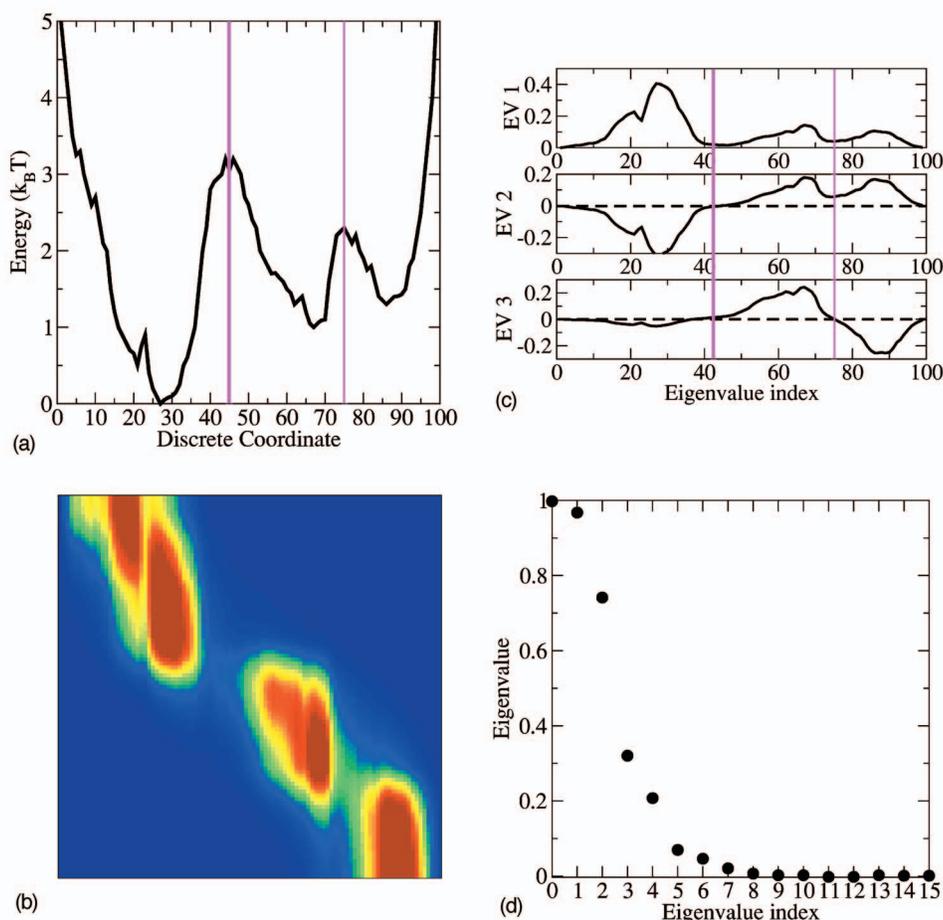


FIG. 1. (Color) Illustration of the PCCA clustering procedure on a sample potential. (a) Potential in units of $k_B T$, defined over a discrete coordinate with 100 boxes. (b) Transition matrix, $\mathbf{T}(\tau)$, for a Markov chain sampling the potential, using a lag time of $\tau=200$ steps. Each matrix entry T_{ij} represents the transition probability from cell i to cell j within time τ (blue: $T_{ij}=0$, red: $T_{ij} \geq 0.1$). The Markov chain was generated by using a Metropolis Monte Carlo where in each step only jumps to the current and the adjacent boxes were considered. The nearly block-diagonal structure of \mathbf{T} is apparent. (c) Left eigenvectors of \mathbf{T} used to identify metastable states. The first eigenvector gives the stationary distribution. The sign structure of the second eigenvector decomposes the state space into two metastable states (thick magenta line). The sign structure of the third eigenvector further splits the right metastable state (thin magenta line), obtaining three metastable states. (d) The eigenvalue spectrum of \mathbf{T} , indicating how many states are metastable. There are clear gaps after two and three eigenvalues.

$$\mathbf{p}(t + \tau) = \sum_{i=1 \dots m} c_i \lambda_i \mathbf{q}_i,$$

where c_i are appropriate constants which depend on the starting condition, $\mathbf{p}(t)$. For longer time increments, $n\tau$, this equation becomes

$$\mathbf{p}(t + n\tau) = \sum_{i=1 \dots m} c_i \lambda_i^n \mathbf{q}_i. \quad (3)$$

For systems fulfilling detailed balance there is only a single eigenvalue $\lambda_1=1$, while $\lambda_i < 1$ for all $i > 1$. Thus, for $n \rightarrow \infty$ (after infinity time), all terms but the first one vanish and Eq. (3) converges to

$$\mathbf{p}(\infty) = c_1 \mathbf{q}_1,$$

where c_1 is simply a normalization factor. Thus, the left eigenvector with the largest eigenvalue, $\lambda_1=1$, specifies the stationary distribution (and has therefore only non-negative entries). All other eigenvectors, $\mathbf{q}_2 \dots \mathbf{q}_m$, are associated with decaying processes, their speed of decay being determined by the magnitudes of the associated eigenvalues. As apparent from Eq. (3), eigenvectors associated with eigenvalues close to 1 correspond to processes which decay very slowly and thus are, by definition, related to transitions between metastable states. Let \mathbf{q}_2 be one such eigenvector with $\lambda \approx 1$, then \mathbf{q}_2 has both positive and negative entries, and its sign structure can be used to identify metastable states as follows: the microstates belonging to one metastable state have positive,

and the others negative signs in \mathbf{q}_2 . The sign structure of the next eigenvector, \mathbf{q}_3 , may be used to further split the state space into three metastable states, and so on. For systems exhibiting metastability there is a time scale τ , for which there is typically at least one clear gap separating the C eigenvalues close to 1, $\mathbf{q}_1, \dots, \mathbf{q}_C$, from the rest of the spectrum, $\mathbf{q}_{C+1}, \dots, \mathbf{q}_m$, thus allowing for an appropriate choice of C to be made. An alternative way of choosing C is introduced in Sec. II D, below. In the original PCCA method, metastable states are identified as sketched above: each pair of microstates which shares the same sign structure in the C eigenvectors with the largest eigenvalues belongs to the same metastable state. Although theoretically correct, this way of assigning microstates to metastable states has proven to be numerically unstable because the eigenvector components close to 0 are very sensitive to insufficient sampling.^{33,42} An improved PCCA method^{32,33} instead identifies C representative microstates with maximum pairwise distances in the coordinates of the first C eigenvectors. Each microstate can then be expressed as a convex combination of these representative microstates and thus be assigned a degree of similarity with each representative microstate according to the convex coordinates. Therefore, clustering is made by assigning each microstate to the cluster containing the representant with which it has the strongest similarity. This method has shown to be more robust than standard PCCA,³³ and is used in the present study.

After the clustering, each microstate is assigned to one metastable state. Thus, each structure along the trajectory can also be assigned to one metastable state $\in \{1, \dots, C\}$. From the trajectory of metastable states, two properties are computed.

- (1) A stationary probability distribution, $\mathbf{p}^{(\infty)}$, containing the probability of finding the trajectory in each metastable state.
- (2) A transfer matrix, \mathbf{Tm} , containing the conditional transition probabilities between pairs of metastable states.

D. Choosing the number of metastable states

A direct way to choose the number of clusters is to count the number of eigenvalues of the transfer matrix that are close to 1 (see Sec. II C above). This method has the implicit assumption that the dynamics is truly Markovian on time scale τ , because otherwise the relative order of eigenvalues may change with the time τ for which the transfer matrix $\mathbf{T}(\tau)$ is constructed and hence the spectral gap between “slow” and “fast” processes would then be not well defined. Here, we introduce an alternative method for selecting the number, C , of metastable states, which does not require the Markovian assumption and produces selections of C which agree remarkably well with the spectral gaps in ranges of τ within which the dynamics is Markovian. An indicator of the quality of a partition into states is the probability of transitions between metastable states, p_{inter} . p_{inter} is given by the sum over the individual transition probabilities between pairs of different metastable states,

$$p_{\text{inter}} = \sum_{i \neq j} p_i^{(\infty)} \cdot T_{Cij}.$$

The corresponding mean transition time between metastable states, $L(C)$, is then given by

$$L(C) = \frac{\tau}{p_{\text{inter}}},$$

where τ is the time step length. In contrast to the eigenvalues, $L(C)$ is no longer strictly monotonic with respect to C (see Fig. 2). Nevertheless $L(C)$ has a direct physical interpretation and exhibits very clear gaps. Moreover, $L(C)$ is proportional to the total number of transitions which should be high for a partition into metastable states (see Sec. II C, above). In the present case, we chose the numbers, C , of metastable states, such that there is a large gap between the mean transition times corresponding to partitions into C and $C+1$ states. For torsion rotamer coordinates, these are $C=4,6,10$ for Ala_8 and $C=4,6,8$ for Ala_{12} . For hydrogen-bond coordinates, these are $C=2,3,5$ for both Ala_8 and Ala_{12} .

E. Quantitative characterization of metastable states

A number of quantitative properties of the metastable states can be computed from the trajectory after clustering. Of particular interest are the probability or free energy, the

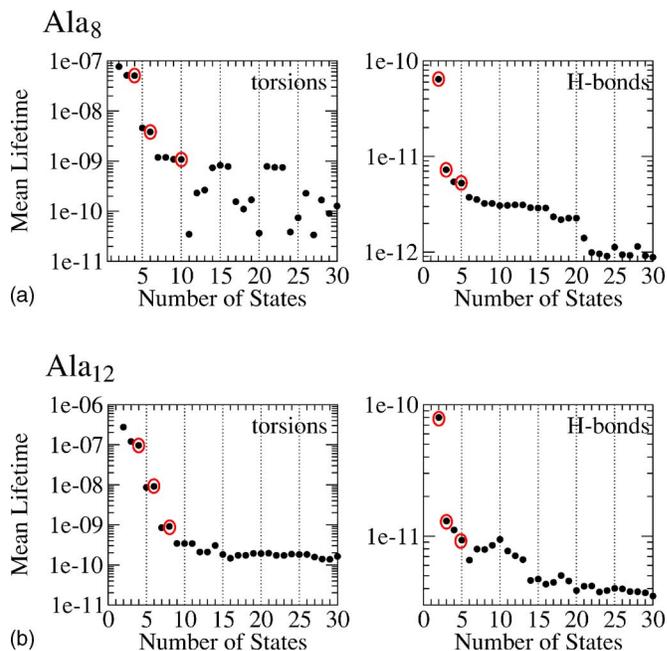


FIG. 2. Mean transition time for state decompositions using the Ala_8 and Ala_{12} dynamics.

entropy, and the mean lifetime of each metastable state. Here we describe how these properties and their uncertainties due to insufficient sampling can be estimated.

1. Probability, free energy, and entropy

The probability, p_i , of being in metastable state i is estimated by

$$p_i = \frac{N_i}{N},$$

where N_i is the number of time steps in which metastable state i is visited, and N the total number of time steps. The distribution of p_i measured from different simulations (or different subsets of the same simulation) is nonsymmetric because p_i is bounded by 0 and 1. Thus, it is nontrivial to estimate uncertainties in p_i . However, instead of the probability we can equivalently specify the free energy difference, ΔA_i , of state i with respect to some arbitrary reference state j (which is chosen here to be the most visited state in the trajectory of metastable states),

$$\Delta A_i = -k_B T \ln \frac{p_i}{p_j},$$

where k_B is Boltzmann's constant and T is the temperature. ΔA_i is not bounded and therefore more likely to have a symmetric distribution. To estimate the uncertainties in ΔA_i , we must consider that due to the metastabilities in our system, subsequent time steps are correlated. An estimate of the correlation time for thermodynamic variables is the slowest relaxation time, τ_{corr} , in the system. τ_{corr} is approximately the mean lifetime of a metastable state for the case where the system is partitioned into two states, and thus estimates the global equilibration time of the system. In the present test simulations this slowest relaxation time is for two metastable

states of torsion rotamers in Ala₁₂: $\tau_{\text{corr}}=274$ ns (see Fig 2). Therefore, we divided the trajectories into $n_{\text{slabs}}=10$ blocks of 400 ns each and computed an estimate $\Delta A_{i,s}$ for each block s . The standard error on ΔA_i , (ΔA_i), is estimated using the standard deviation of the set of individual estimates as follows:

$$\sigma(\Delta A_i) \approx \frac{1}{\sqrt{n_{\text{slabs}}}} \sigma\{\Delta A_{i,1}, \dots, \Delta A_{i,n_{\text{slabs}}}\}.$$

The potential energies from the simulation are averaged over all members of a metastable state, giving rise to a mean potential energy difference, $\Delta \bar{U}_i$, calculated relative to the free energy minimum for which we define $\Delta \bar{U}_i=0$. The uncertainties of the mean potential energies may be computed the same way as the uncertainties of the free energies. For the small systems studied here, however, they are negligible compared to uncertainties in the free energies, and so they are not specified. Given both the free and the mean potential energies, the entropy, $T\Delta S_i$, of a metastable state can directly be computed as

$$T\Delta S_i = \Delta \bar{U}_i - \Delta H_i.$$

2. Mean lifetime

If the escape from a state, i , behaves like a first-order transition and no return to the state is considered, then the probability of being in the state for a time, t , is given by

$$p_i(t) = \exp(-k_{\text{escape},i}t) = \exp(-t/\bar{L}_i), \quad (4)$$

where $k_{\text{escape},i}$ is the escape rate and \bar{L}_i the mean lifetime of state i . For short time steps τ , a good discrete approximation of $p_i(t)$, $p_i(n\tau)$, can be generated for each state i from the simulations given the sorted series of lifetimes $(L_{i,1}, L_{i,2}, \dots, L_{i,m})$ with $L_{i,1} < L_{i,2} < \dots < L_{i,m}$.

$$p_i(n\tau) = 1 - \frac{1}{m} \sum_{j=1}^m s(L_{i,j}, n\tau), \quad j=0, \quad (5)$$

where $s(t, t_0)$ is a step function defined by

$$s(t, t_0) = \begin{cases} 0, & t < t_0 \\ 1, & t \geq t_0. \end{cases}$$

For the metastable states in the dynamical trajectories analyzed here, the simple escape kinetics suggested by Eq. (4) is indeed a good model, after a nonexponential initial phase (see Fig. 3 for some examples). For all metastable states used in this study, the nonexponential part dies after 10% of the longest lifetimes observed for the state. We thus ignore all visits shorter than these first 10% and determine $\tau_{\text{life},i}$ by linear regression to the rest of the curve. The errors are estimated from the standard deviation of a series of decay times $\tau_{\text{life},i,k}$ from $\sqrt{|\tau|}$ partitions of the lifetime series τ ,

$$\sigma(\tau_{\text{life},i}) \approx |\tau|^{-1/4} \sigma\{\tau_{\text{life},i,1}, \dots, \tau_{\text{life},i,\sqrt{|\tau|}}\}.$$

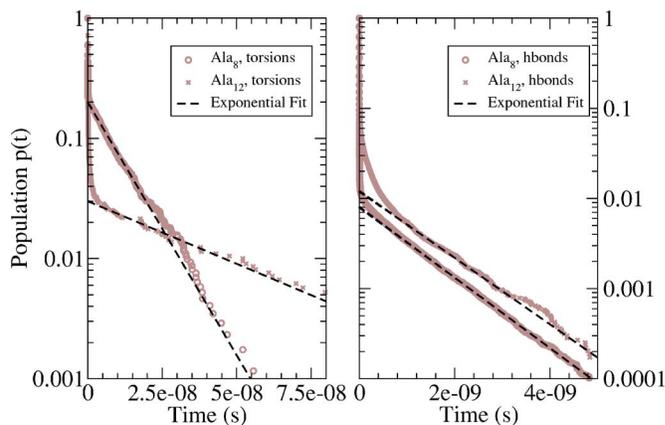


FIG. 3. Examples of decay curves of metastable state populations. The brown circles and crosses show the fraction of the initial population remaining in a given state for a given time. The selected states are the global energy minima of following metastable state partitions: Ala₈: ten states for torsion microstates and five states for hydrogen-bond microstates; Ala₁₂: eight states for torsion microstates and five states for hydrogen-bond microstates.

F. Testing for the Markov property

Many analysis methods for transition networks rely on the assumption that the interstate dynamics has no memory, i.e., they exhibit the Markov property. It is therefore desirable to find a state definition which exhibits the Markov property as clearly as possible. Here we review an indicator for the Markov property that has been introduced in Refs. 2 and 43. It is used in the Results section to compare the different metastable state decompositions. Note that other approaches to testing the Markov property exist.^{23,44}

For a Markov process, the population of the states at some time $t+n\tau$ can be expressed by applying the transfer matrix n times to a population at a previous time t , which leads to the Chapman-Kolmogorov equation⁴¹

$$\mathbf{p}(t+n\tau) = \mathbf{p}(t)[\mathbf{T}(\tau)]^n = \mathbf{p}(t)\mathbf{S}(n\tau),$$

where we have defined $\mathbf{S}(n\tau) := [\mathbf{T}(\tau)]^n$. Using the eigenvalue-eigenvector equation,

$$\mathbf{Q}\mathbf{\Lambda} = \mathbf{T}(\tau)\mathbf{Q}, \quad \mathbf{Q}\mathbf{\Gamma}(n) = \mathbf{S}(n\tau)\mathbf{Q},$$

where \mathbf{Q} is the matrix of eigenvectors $\mathbf{Q}=[\mathbf{q}_1 \cdots \mathbf{q}_m]$ and $\mathbf{\Lambda}$ and $\mathbf{\Gamma}(n)$ are diagonal matrices with eigenvalues $\lambda_1, \dots, \lambda_m$ and $\gamma_1(n), \dots, \gamma_m(n)$ on the diagonal, respectively. We can write

$$[\mathbf{T}(\tau)]^n = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1} = \mathbf{Q}\mathbf{\Gamma}(n)\mathbf{Q}^{-1} = \mathbf{S}(n\tau),$$

and hence the eigenvalues λ_i of $\mathbf{T}(\tau)$ are related to the eigenvalues $\gamma_i(n)$ of $\mathbf{S}(n\tau)$ by

$$\gamma_i(n) = \lambda_i^n.$$

This relation can equivalently be transformed to

$$\tau_i^* := -\frac{n\tau}{\ln \gamma_i(n)} = -\frac{\tau}{\ln \lambda_i}, \quad (6)$$

where τ_i^* is the characteristic time scale corresponding to the decay of eigenmode i . According to this equation, if the process associated with eigenmode i is Markovian, then τ_i^* is

constant and thus independent of n . For many processes, this is true only for some minimum time scale $n\tau > \tau_{\text{mem}}$, after which the memory pertaining to the interstate dynamics has disappeared. In this case, the Markov model of the system must employ the Markov matrix $\mathbf{S}(n\tau)$ instead of $\mathbf{T}(\tau)$. The time $n\tau$ which is used to set up the Markov matrix is referred to as lag time. It is desirable to find a definition of states such that the lag time is minimal, i.e., such that there is minimal memory in the system.

Note that τ_i^* being constant is a necessary but not a sufficient condition for Markov behavior, although it seems to be a rather reliable indicator and helps us to identify a range of lag times for which Markovian behavior may be present. Ultimately, the best test for the model is to compare the molecular dynamics simulations to the predictions of the Markov dynamics produced with the transition network model.⁴⁵ Here, this is done as follows: First, a lag time $n\tau$ for which the τ_i^* are approximately constant is selected. Then the Markov matrix $\mathbf{S}(n\tau)$ is computed. A Markov dynamics trajectory is computed for each of the C metastable states as follows: In each dynamics run, the initial population is chosen such that one of the metastable states has population 1 and the others have population 0. Then, the dynamics trajectory of a total length of $T = mn\tau$ is computed by successively applying

$$\mathbf{p}(t + n\tau) = \mathbf{p}(t)\mathbf{S}(n\tau),$$

$m-1$ times. This will relax the initial population towards equilibrium. The relaxation curve of the state k which was initiated with population 1, $p_{\text{Markov},k}(t)$, is of particular interest. For comparison, a relaxation dynamics can be obtained from the molecular dynamics simulation as follows: After defining the C metastable states, the trajectory can be expressed in terms of a series of indices of metastable states, say, $[s(0), s(\tau), \dots, s(N\tau)]$. From this trajectory, windows of length T are sampled each and with a time resolution of $n\tau$, such that $n_w = N/n - m$ windows are obtained, window i being $W_i = [s(i), s(i+n\tau), \dots, s(i+mn\tau)]$. The relaxation curve is computed for state k , $p_{\text{MD},i}(t)$, by average over all n_k windows which start at state k ,

$$p_{\text{MD},k}(t) = \frac{1}{n_k} \sum_{i=1 \dots n_k} \begin{cases} 1 & \text{for } W_i(t) = s \\ 0 & \text{otherwise,} \end{cases}$$

$$W_{i0} = k.$$

It is then compared whether

$$p_{\text{Markov},k}(t) \approx p_{\text{MD},k}(t),$$

i.e., whether the relaxations of the population of the individual metastable states k predicted by the Markov model resemble those observed in the molecular dynamics simulations.

G. Hierarchical transition network analysis

A transition network consists of vertices connected by edges. The vertices correspond to the metastable states resulting from the kinetic clustering. An edge (u, v) is present in the network if there is at least one transition $u \rightarrow v$ or $v \rightarrow u$

in the trajectory of metastable states. For simulations that are close to equilibrium and short time steps τ , the resulting connectivity of the network represents neighborhoods in configuration space. A network is said to be connected if each metastable state can be reached from each other metastable state by some pathway. It is said to be fully connected if all vertices are directly connected by an edge.

For each system (Ala₈ and Ala₁₂) and each microstate system (torsional rotamers and hydrogen bonds), three kinetic clusterings with different numbers of metastable states were computed (see Sec. II D). These three networks define a three-level hierarchy of the conformational dynamics of the system. For example, the Ala₈ torsion rotamer network, and the $C=4$ and $C=6$ networks are related in that state a for $C=4$ splits into states a, f , and g for $C=6$ (see Fig. 4). Such a split is not always “perfect,” i.e., in some cases, the boundaries of the higher-order network do not remain the same when going to the lower-order network because some of the microstates have been traded between the split vertex and its neighbors. This is because the kinetic clustering is performed independently for each value of C and does not use the results of the higher-level clustering for a lower value of C . To visualize the relationships between different order networks of the same class, in the figures presented here an arrow is drawn to each vertex from the main contributor in the higher-order network.

III. RESULTS

In this section, no attempt is made to reproduce experimental results or to predict the structure of oligoalanines but rather the systems are used to examine the utility of the methodological approach and to test whether the state decomposition can be used to construct a kinetic model that itself reproduces the observations from the molecular dynamics simulations. However, some background on the known conformational preferences of polyanines is useful. Polyanines have been frequently studied both experimentally and theoretically (Ref. 12 and references therein). Alanine is believed to be the amino acid with the highest helical propensity.⁴⁶ Nevertheless the identity of the most stable state of polyanines in aqueous environment is not fully agreed upon, as experiments are rendered difficult by the fact that polyanines tend to aggregate. Simulation results differ.^{9,47} However, most studies do indicate that solvated polyanines longer than some minimum length prefer the α -helical structure. This propensity may be modified by mutating individual sites, modifying the terminal charges, or, in simulation, changing the solvent representation.¹² Most simulations involve capped (neutral) termini. In the present study, we use charged (NH_3^+ and COO^-) termini instead, because (1) it is generally accepted that polypeptides in solvent exist in a zwitterionic form and (2) the presence of oppositely charged termini may lead to richer (and thus more “interesting”) metastability. Transition networks and Markov state models between metastable states for Ala₈ and Ala₁₂ were constructed as described in Sec. II and are analyzed in the present section.

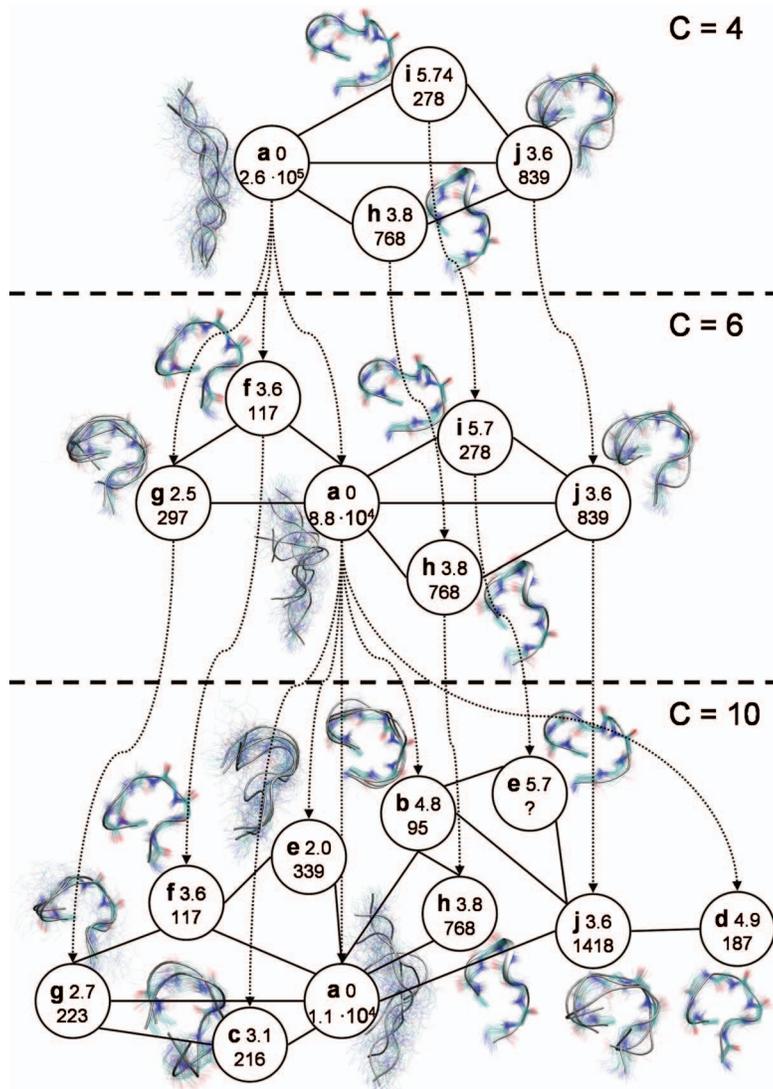


FIG. 4. (Color) Hierarchical transition network analysis for the Ala₈ peptide based on backbone torsion rotamer microstates. Each bullet and the structure next to it corresponds to one metastable set of backbone torsion rotamer patterns. The bullets contain the state name (a letter), the free energy in kcal/mol (upper number), and the mean lifetime in picoseconds (lower number). Each structure is shown by a few representative tubes and an overlay of 100 examples randomly drawn from the ensemble of structures of each state, shown as line representations of the backbone. A pair of states is connected if at least one transition between these states was observed in the trajectory. The hierarchical relationship between the three networks for C=4, 6, and 10 metastable states is indicated by the dotted arrows. Each arrow starts at the metastable state in the higher-order network which contains the majority of microstates in the state the arrow points to. For example, the microstates of state *a* in the C=4 network are split into three substates, *a*, *f*, and *g*, in the C=6 network.

TABLE I. Quantitative description of the metastable states found by grouping backbone torsion rotamers in Ala₈. *s*, name of the state. ΔA , free energy difference with respect to the free energy minimum (state *a*) and its error. $\Delta \bar{U}$, potential energy difference. $T\Delta S = \Delta \bar{U} - \Delta A$ is the free energy difference due to entropy. All energies are given in kcal/mol, the uncertainties in $\Delta \bar{U}$ are negligible, and thus the uncertainties in $T\Delta S$ are identical to the uncertainties ΔA . \bar{L} , mean lifetime of the state in picoseconds and its error. *n*, number of visits in the state long enough to be useful to fitting the mean lifetime. The lines are ordered so as to indicate how states split when going to a network with more metastable states, e.g., state *a* for C=4 splits up into *a*, *f*, and *g* for C=6.

Four clusters						Six clusters						Ten clusters						
<i>s</i>	ΔA	$\Delta \bar{U}$	$T\Delta S$	\bar{L}	<i>n</i>	<i>s</i>	ΔA	$\Delta \bar{U}$	$T\Delta S$	\bar{L}	<i>n</i>	<i>s</i>	ΔA	$\Delta \bar{U}$	$T\Delta S$	\bar{L}	<i>n</i>	
<i>a</i>	0	0	0	$(2.6 \pm 0.9) \times 10^5$	10	<i>a</i>	0	0	0	$(8.8 \pm 1.8) \times 10^4$	29	<i>a</i>	0	0	0	$(1.2 \pm 0.1) \times 10^4$	105	
												<i>b</i>	4.8 ± 0.5	-2.7	-7.6	95 ± 18	10	
												<i>c</i>	3.1 ± 0.7	-4.9	-8.0	216 ± 20	45	
												<i>d</i>	$4.9 \pm ?$	-4.4	-9.3	187 ± 26	4	
												<i>e</i>	2.0 ± 0.1	-1.9	-3.9	339 ± 50	148	
							<i>f</i>	3.6 ± 0.9	-5.8	-9.5	117 ± 7	19	<i>f</i>	3.6 ± 0.8	-5.9	-9.5	117 ± 7	19
							<i>g</i>	2.5 ± 0.6	-4.9	-7.4	297 ± 40	59	<i>g</i>	2.7 ± 0.5	-3.9	-6.6	223 ± 20	89
<i>h</i>	3.8 ± 0.9	-5.8	-9.6	768 ± 135	7	<i>h</i>	3.8 ± 0.9	-5.8	-9.6	768 ± 135	7	<i>h</i>	3.8 ± 0.9	-5.8	-9.6	768 ± 135	7	
<i>i</i>	$5.7 \pm ?$	-4.9	-10.6	$278 \pm ?$	1	<i>i</i>	$5.7 \pm ?$	-4.9	-10.6	$278 \pm ?$	1	<i>i</i>	$5.7 \pm ?$	-5.0	-10.7	$278 \pm ?$	1	
<i>j</i>	3.6 ± 1	-4.0	-7.4	839 ± 164	7	<i>j</i>	3.6 ± 1	-4.0	-7.7	839 ± 164	7	<i>j</i>	3.7 ± 0.8	-4.1	-7.8	1418 ± 219	4	

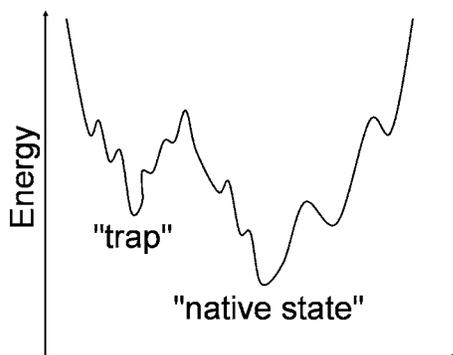


FIG. 5. Scheme illustrating the concept of a kinetic trap in the absence of a predefined native structure. The free energy minimum is understood as the “native state” while high-energy minima that are separated from the main basins with high barriers are defined as “kinetic traps.” Such traps are rarely visited, but when they are visited, they are stable for a long time.

A. Octaalanine

The Ala₈ rotamer microstates were clustered into numbers, C , of four, six, and ten metastable states (Fig. 4 and Table I). The global free energy minimum, a , is a highly entropic state consisting of mostly unfolded conformations the mean lifetime of which ranges from more than 250 ns (for $C=4$) down to around 10 ns (for $C=10$). For $C=4$, three metastable states, h , i , and j , are distinguished from state a . The structures of h , i , and j are coils in which the negatively charged C terminus is in contact with three or four backbone amide hydrogens, thus bending the chain. As this type of structure is quite common in the present simulations, we name it ρ loop, due to its similarity with the Greek letter. i and h are structurally well-defined low-entropy ρ loops in which the C terminus interacts with different positions of the chain and which interchange either via the main state, a , or the less-ordered intermediate, j . h , i , and j have free energies of 3.6 kcal/mol or more above the main state, a . h and j are visited only seven times each and have an estimated mean lifetime of over 700 ps. i was visited only once and was stable for nearly 300 ps. We therefore define h , i , and j as kinetic traps (see Fig. 5 for an illustration of the concept).

For higher numbers of metastable states, further ρ loops and similar coil structures are split off from the main state a , all of which are at least 2 kcal/mol above state a and have lifetimes between 95 ps and 1.5 ns. Most of these loops have relatively well-defined structures and very low entropies (state e being the only exception, which has an entropic contribution of $T\Delta S = -3.9$ kcal/mol with respect to a), whereas the main state a is structurally rather disordered even for $C=10$, consistent with its large entropy.

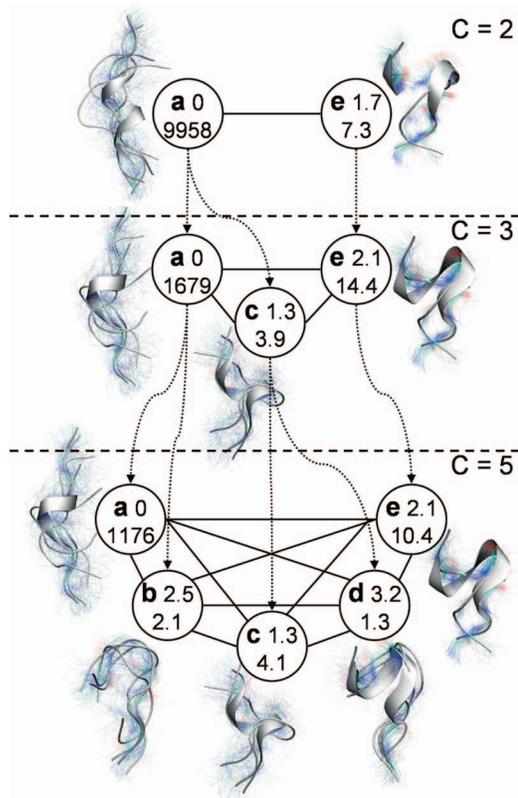


FIG. 6. (Color) Hierarchical transition network analysis for the Ala₈ peptide for two, three, and five metastable sets containing hydrogen-bonding pattern microstates. See caption of Fig. 4 for a complete description.

The topology of the networks is quite nonuniform. While the global minimum, a , is the vertex with the most connections, some vertices serve as intermediates on some pathway, thus having only two neighbors. State d is even a dead end, having only one neighbor. The network hierarchy indicates the presence of a single free energy basin, as for increasing C , states are always split off from the main state a which gets subsequently smaller.

For the hydrogen-bonding pattern dynamics of the same system, a rather different picture arises (Fig. 6 and Table II). In the hierarchical transition network analysis for $C=2$, 3, and 5 metastable states, there is again a global free energy minimum, a , from which the other metastable states split off in the lower-order networks. State a is very disordered for $C=2$, comprising many different structures. In general, the structures are quite dissimilar from the structures in the torsion angle networks. In contrast to the structurally well-defined ρ loops in the torsional rotamer network, the meta-

TABLE II. Quantitative description of the metastable states found by grouping hydrogen-bonding patterns in Ala₈. See Table I for a complete description.

Two clusters					Three clusters					Five clusters							
s	ΔA	$\Delta \bar{U}$	$T\Delta S$	T	n	s	ΔA	$\Delta \bar{U}$	$T\Delta S$	T	n	s	ΔA	$\Delta \bar{U}$	$T\Delta S$	T	n
a	0			9958±921	180	a	0	0	0	1680±129	487	a	0	0	0	1176±65	935
												b	2.5±0.1	-2.5	-5.0	2.2±0.05	6 261
						c	1.3±0.1	-2.6	-4.0	4.0±0.1	17 398	c	1.4±0.1	-2.6	-4.0	4.1±0.05	16 811
												d	3.2±0.1	-2.3	-5.6	1.3±0.03	4 001
												e	2.1±0.1	-4.2	-6.3	10.4±0.3	2 750
e	1.7±0.1	-4.1	-5.8	7.3±0.2	3627	e	2.1±0.1	-4.1	-6.3	10.4±0.3	2 750	e	2.1±0.1	-4.2	-6.3	10.4±0.3	2 750

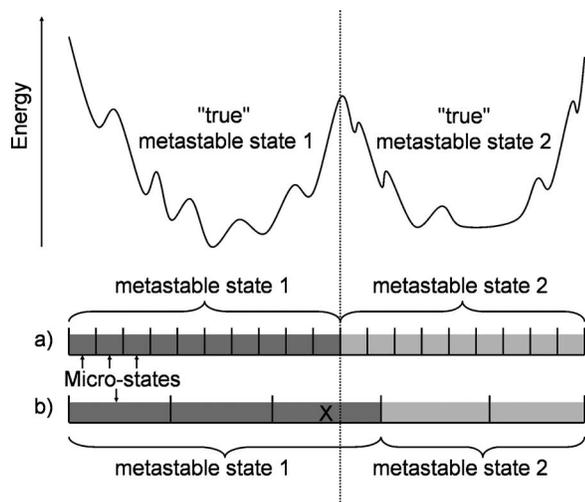


FIG. 7. Scheme illustrating why some definitions of microstates are better than others. Set (a) is optimal because, when partitioned into metastable sets, it splits exactly on the transition state. Set (b) is inappropriate because one of its states reaches across the transition state, including states from both basins. Such a definition yields shorter lifetimes because transitions into and out of cell x are frequent from both basins, even though no actual transition may occur. This also produces apparent connections between metastable states that are actually not directly connected.

stable states split off from the global minimum are structurally more disordered, and are characterized by a few stable hydrogen bonds. The most striking difference is perhaps that the lifetimes in the hydrogen-bond metastable states are two orders of magnitude smaller than those of the torsion rotamer metastable states. As it is desired to obtain metastable states with a maximum lifetime, this result shows that for Ala_8 torsion rotamer coordinates are more useful for detecting metastable states than hydrogen bonds.

Another indicator for the fact that hydrogen-bond coordinates are less able to distinguish kinetically separated states is the network connectivity: The hydrogen-bond network for Ala_8 is fully connected, while this is not the case for the torsion rotamer network. The hydrogen-bond coordinates are thus not able to satisfactorily resolve the boundaries between metastable states. This indicates that the hydrogen-bond coordinates are “coarser,” i.e., different rotamer patterns generate the same hydrogen-bonding pattern, which is also reflected by the fact that 457 rotamer microstates were used, but only 295 hydrogen-bond microstates. Figure 7 illustrates this problem. For Ala_8 , the dynamics appears thus to be dominated by switching dynamics between rotamer states, which is due to a combination of torsional potentials and nonbonded interactions between neighboring residues, rather than by formation of hydrogen bonds.

Figures 11(a) and 11(b) show the time scales τ_i^* implied by the eigenvalues of the transfer matrices set up for various lag times, $n\tau$ [see Sec. II F and especially Eq. (6)]. Most of the τ_i^* for the torsion angle coordinates level off around 100τ – 200τ , corresponding to a lag time of 5–10 ps, indicating that near-Markovian behavior may be found from that time scale and longer. In contrast, none of the implied time scales computed for the states defined by hydrogen bonds converges within 1000τ (50 ps), showing that these coordinates are much less useful for obtaining a good Markov state

model. This finding is confirmed by Figs. 12(a) and 12(b) which show the comparison of the relaxation dynamics for individual metastable states predicted by the Markov model with the observations from the molecular dynamics simulations (see Sec. II F), constructed for a lag time of 1000τ (50 ps). While the Markov state model for the hydrogen-bond states produces dynamics that are completely different from the molecular dynamics simulation, the Markov state model based on the torsion angle states agrees reasonably well with the molecular dynamics simulations for large enough populations and for states that are frequently visited. The deviations observed at low populations and for those states that are rarely visited are most likely due to insufficient sampling in the molecular dynamics simulations, and much longer simulations would be required to obtain approximate agreement for the relaxation dynamics of these states.

B. Dodecaalanine

The Ala_{12} rotamer dynamics was clustered into four, six, and eight metastable states (Fig. 8 and Table III). As in Ala_8 , the global free energy minimum, a , corresponds to a large set of conformations, containing different secondary structures, yet a is not the state with maximum entropy. State a has a mean lifetime of 456 ns for $C=4$. For $C=6$ and $C=8$, the mean lifetime is subsequently reduced as other metastable states are split off, with a mean lifetime of 87 ns for $C=8$. For $C=4$, the states f , g , and h are visited only three, two and one times, respectively, but are stable for nanoseconds, thus representing kinetic traps. Their free energies are more than 3 kcal/mol above the global minimum. f and g are ρ loops similar to those found for Ala_8 . The metastable states split off from the main state for $C=6$ and 8 contain significant segments of secondary structure, involving both α helices and β hairpins. As in the Ala_8 torsion rotamer network, the hierarchy between $C=4$ and $C=8$ indicates a single free energy basin, with state a as a central state from which metastable states are subsequently split off.

The results for the hydrogen-bonding pattern dynamics of the same system are not as different from the torsion angle dynamics as was the case with Ala_8 (Fig. 9 and Table IV). The ρ loop in metastable state e is similar that in metastable state f of the torsion network. However, the relative free energy of e is lower and its lifetime shorter than those of f in the torsion network. This indicates again that state e is not cleanly separated from the main state in a kinetic sense but is actually a mixture of the two different metastable states. In the $C=3$ network, α and β secondary structures separate out into two different branches of the network.

The difference in lifetimes between the Ala_8 torsion angle and hydrogen bonding networks is still considerable, though not as large as in the Ala_8 case. This indicates that using the hydrogen bonds to indicate the metastabilities of the Ala_{12} dynamics is still not as good as using the torsion angles, but much better than in the Ala_8 case. As for Ala_8 , the Ala_{12} hydrogen-bonding network is fully connected, while the torsion network is not.

Finally, we computed another transition network for

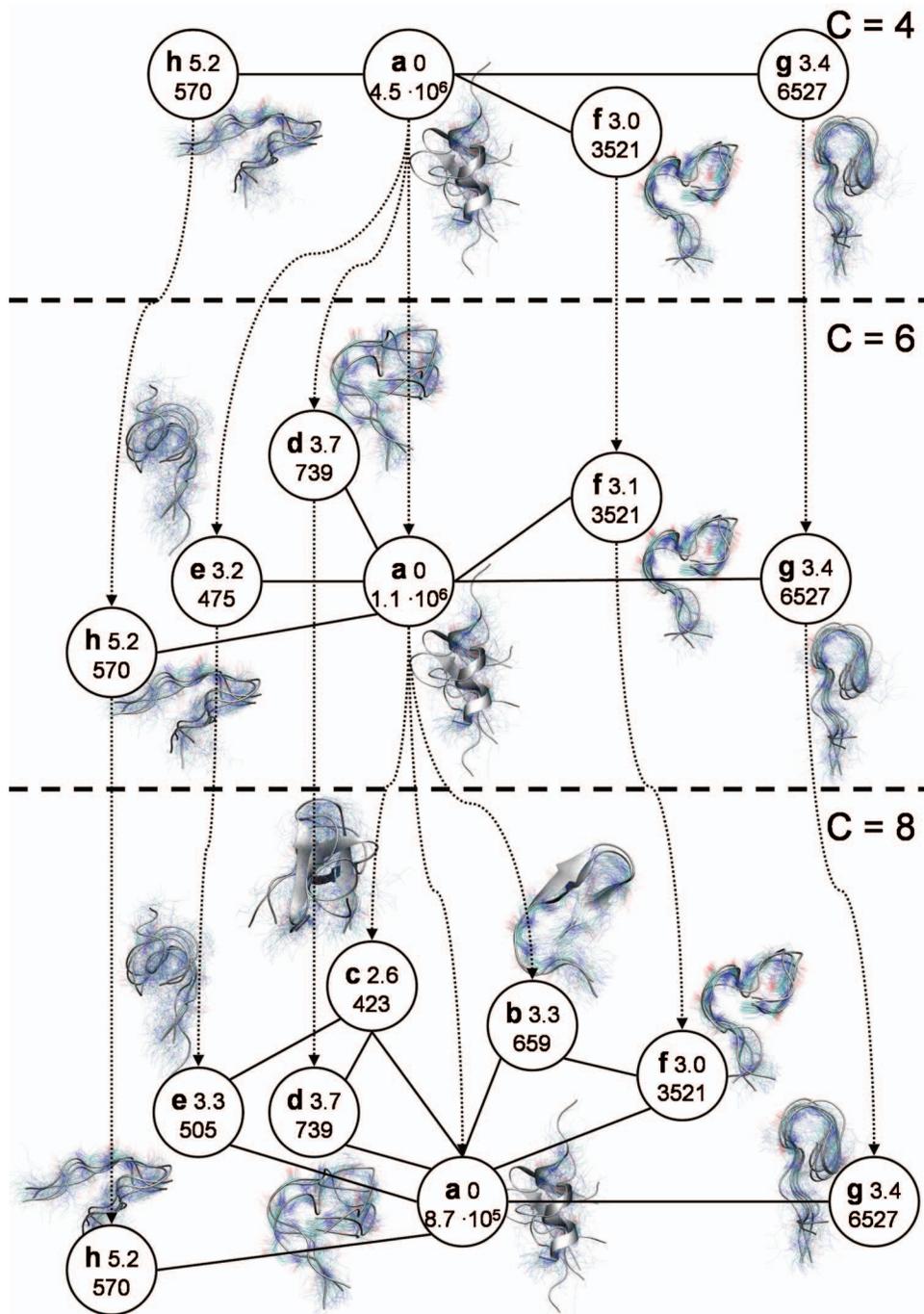


FIG. 8. (Color) Hierarchical transition network analysis for the Ala₁₂ peptide for two, six, and eight metastable sets containing backbone torsion rotamer microstates. See caption of Fig. 4 for a complete description.

TABLE III. Quantitative description of the metastable states found by grouping torsion rotamers in Ala₁₂. See Table I for a complete description.

Four clusters						Six clusters						Eight clusters					
<i>s</i>	ΔA	$\Delta \bar{U}$	$T\Delta S$	<i>T</i>	<i>n</i>	<i>s</i>	ΔA	$\Delta \bar{U}$	$T\Delta S$	<i>T</i>	<i>n</i>	<i>s</i>	ΔA	$\Delta \bar{U}$	$T\Delta S$	<i>T</i>	<i>n</i>
<i>a</i>	0	0	0	$(4.5 \pm 1.6) \times 10^5$	6	<i>a</i>	0	0	0	$(1.2 \pm 0.4) \times 10^5$	20	<i>a</i>	0	0	0	$(8.7 \pm 2.2) \times 10^4$	27
												<i>b</i>	3.3 ± 0.7	7.2	3.9	659 ± 192	18
												<i>c</i>	2.6 ± 0.3	4.0	1.3	423 ± 48	30
						<i>d</i>	3.7 ± 0.9	10.1	6.5	739 ± 216	6	<i>d</i>	3.7 ± 0.9	10.2	6.6	739 ± 216	6
						<i>e</i>	3.2 ± 0.6	5.6	2.4	475 ± 69	27	<i>e</i>	3.3 ± 0.7	5.4	2.2	505 ± 83	23
<i>f</i>	$3.0 \pm ?$	9.4	6.3	$3521 \pm ?$	3	<i>f</i>	$3.1 \pm ?$	9.4	6.4	$3521 \pm ?$	3	<i>f</i>	$3.0 \pm ?$	9.5	6.5	$3521 \pm ?$	3
<i>g</i>	3.4 ± 1.3	7.8	4.4	$6527 \pm ?$	2	<i>g</i>	3.4 ± 1.3	7.9	4.5	$6527 \pm ?$	2	<i>g</i>	3.4 ± 1.3	8.0	4.5	$6527 \pm ?$	2
<i>h</i>	5.2 ± 0.6	2.3	-3.0	$570 \pm ?$	1	<i>h</i>	5.2 ± 1.2	2.3	-3.0	$570 \pm ?$	1	<i>h</i>	5.2 ± 1.3	2.4	-2.8	$570 \pm ?$	1

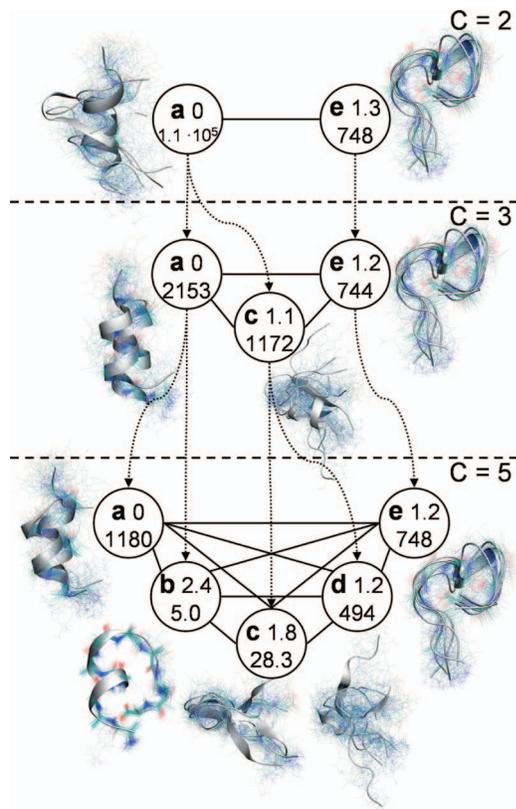


FIG. 9. (Color) Hierarchical transition network analysis for the Ala₁₂ peptide for two, three, and five metastable sets containing hydrogen-bonding pattern microstates. See caption of Fig. 4 for a complete description.

Ala₁₂ using torsion coordinates and $C=20$ (see Fig. 10). This state partitioning can be thought of dissecting the bottom of the main free energy basin, as it bifurcates into two main branches when going from $C=8$ to $C=20$: State c splits into $k, l, m, n,$ and c and state a splits into $c, o, p, q, r, s, a,$ and t .

Several states that are structurally similar are not directly kinetically connected, but rather via a structurally distant intermediate. Examples of these are the ρ loops, $f, p,$ and $e,$ which are connected via the global minimum $a,$ and $p, b,$ and $o,$ which are connected via the α -helix state, t . It is possible that there are high-energy pathways which directly connect these ρ loops and were not sampled in the 4 μ s trajectory. However, the fact that these states are distinguished with the present clustering method displays the strength of the method over geometric clustering, with which it would be difficult to separate these structurally very similar states.

Like for Ala₈, Figs. 11(c) and 11(d) and Figs. 12(c) and

12(d) confirm that the torsion angle states are a better choice for constructing a Markov state model for the Ala₁₂ dynamics used here than the hydrogen-bond states. Most of the τ_i^* for torsion angle coordinates level off at a lag time of around 100τ (5 ps), while they do not converge within 1000τ for the hydrogen-bond based states. The relaxation dynamics for individual torsion angle metastable states predicted by the Markov model constructed for a lag time of 1000τ agrees well with the observations from the molecular dynamics simulations for large populations and frequently visited states. In contrast, the Markov state model based on hydrogen-bond states produces relaxation dynamics which strongly deviate from the MD results for all states, indicated that this state definition was not successful in producing a valid Markov state model.

IV. CONCLUSIONS

Metastability analysis is useful for understanding the thermodynamics, kinetics, and topology of complex molecular free energy surfaces. The PCCA approach employed here allows a predefined number of conformational states to be identified which are maximally kinetically separated. This kinetic separation is not achieved by methods that rely on geometric clustering. There is no “correct” or “incorrect” number of conformational states, and this number has to be chosen by the user, depending on how much “resolution” of the process is desired. However, a reasonable choice can be made by selecting numbers of states, C , such that there is a large gap between the mean transition times corresponding to partitions into C and $C+1$ states. The mean transition time is computed as the inverse of the interstate transition flux for each possible state decomposition. The rationale for choosing the number C in this way is that, at the time scale of the mean transition time corresponding to C , the dynamics decomposes into C states.

By computing the free energies, energies, entropies, and mean lifetimes of the metastable states, and connecting pairs of states between which transitions exist, one obtains an informative transition network between metastable states. Such a network yields a comprehensive picture of the interstate dynamics at any given resolution of states. For a long enough simulation and a short observation interval τ , the connectivity of the network resembles the neighborhoods of the metastable states in configurational space. Insights into the hierarchical structure of the free energy surface are gained when networks with different numbers of metastable states for the same system are connected in a network hierarchy. For meta-

TABLE IV. Quantitative description of the metastable states found by grouping hydrogen-bonding patterns in Ala₁₂. See Table I for a complete description.

Two clusters						Three clusters						Five clusters					
s	ΔA	$\Delta \bar{U}$	$T \Delta S$	T	n	s	ΔA	$\Delta \bar{U}$	$T \Delta S$	T	n	s	ΔA	$\Delta \bar{U}$	$T \Delta S$	T	n
a	0	0	0	113 807±32 879	21	a	0	0	0	2153±125	412	a	0	0	0	1180±56	764
												b	2.4±0.5	11.0	8.6	5.0±0.1	3743
						c	1.1±0.5	5.5	4.3	1172±133	143	c	1.8±0.6	5.4	3.6	28.3±1.7	822
												d	1.2±0.4	5.6	4.3	494±45	124
e	1.3±0.8	9.0	7.7	748±138	152	e	1.2±0.9	9.8	8.6	744±138	151	e	1.2±0.9	10.0	8.8	748±138	152

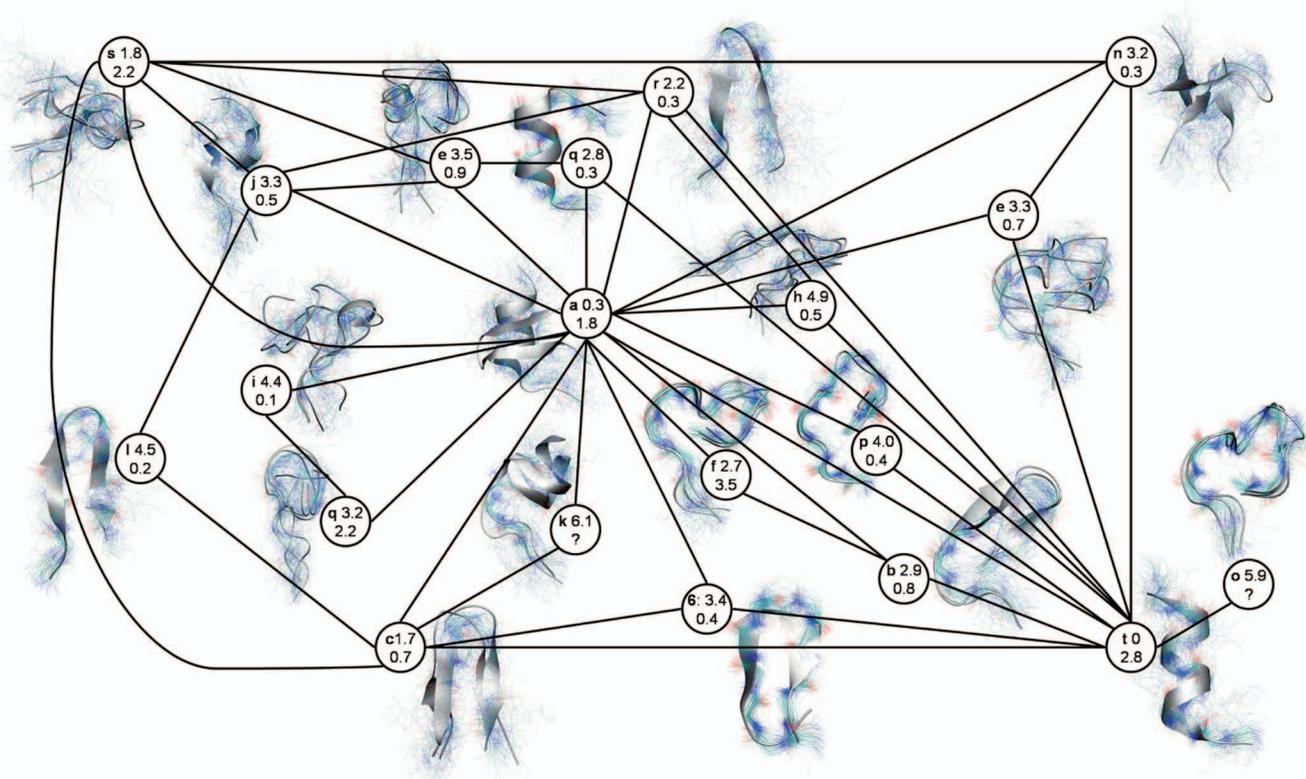


FIG. 10. (Color) Transition network analysis for the Ala₁₂ peptide for 20 metastable sets containing backbone torsion rotamer microstates. See caption of Fig. 4 for a complete description.

stable states based on torsion angle coordinates we have constructed a Markov state model which is successful in reproducing the dynamics from molecular dynamics simulations at least for large populations and frequently visited states. The deviations at low populations and rarely visited states are most likely due to insufficient sampling in the molecular dynamics simulation, giving us confidence that successful Markov state models can be built based on the techniques introduced here.

In the present peptide torsion rotamer networks, the conversion from networks with few metastable states to those with more is mostly characterized by clean splits of existing metastable states. This means that, in going from a network with C metastable states to a network with $C+1$ metastable states, in most cases a single state is split into two substates while the boundaries of all other states remain the same. This is an indicator for the existence of a hierarchy of clear barriers in the system. In other words, it is possible to find meaningful metastable states in the simulations of biomolecules studied here. Moreover, these states have relatively long lifetimes, on the order of hundreds of picoseconds to nanoseconds, a fact that is not apparent when visualizing the trajectory graphically, as no stable states are clearly seen. The method is successful in distinguishing states that are kinetically separated even when they are structurally similar, provided that the microstates are chosen finely enough to distinguish them, which is a clear advantage over geometry-based clustering methods.

Another feature found for both systems, Ala₈ and Ala₁₂, is that they have a number of kinetic trap states which are

rarely visited, but very long lived. Typically, these traps are structurally well defined, i.e., they are characterized by low entropy. In free energy surface plots these traps are likely to be invisible, because they are not strongly populated. In agreement with another recent study⁴⁸ this finding indicates that the thermodynamics and kinetics of peptide folding (and possibly also of proteins) may often be much more complex than free energy projections on one or two order parameters, such as principal components, number of native contacts, or root-mean-square deviation to native structure, suggest.

Clearly, the force field, the solvent model, and the thermostat being used to produce the molecular dynamics simulations will affect the thermodynamics, kinetics, and topology of the resulting transition networks. To obtain an accurate model, the simulation setup should be validated by comparing experimental observables computed from the model with experimental measurements, if available. However, there may still be a number of different models that fit well to the experimental findings but perform differently from a theoretical point of view. For this, the treatment of the solvent is an important factor. In the present study, an implicit solvent model that does not incorporate solvent viscosity was employed. While this setup will lead to errors in the system kinetics, it is not clear how modeling the viscosity (by a Langevin thermostat or by explicit solvent) would change the quality of the Markov state model. Inclusion of friction would probably reduce transitions between metastable states (due to both the presence of friction and the increased computational costs leading to shorter trajectories). This would increase the uncertainties in the transfer matrix

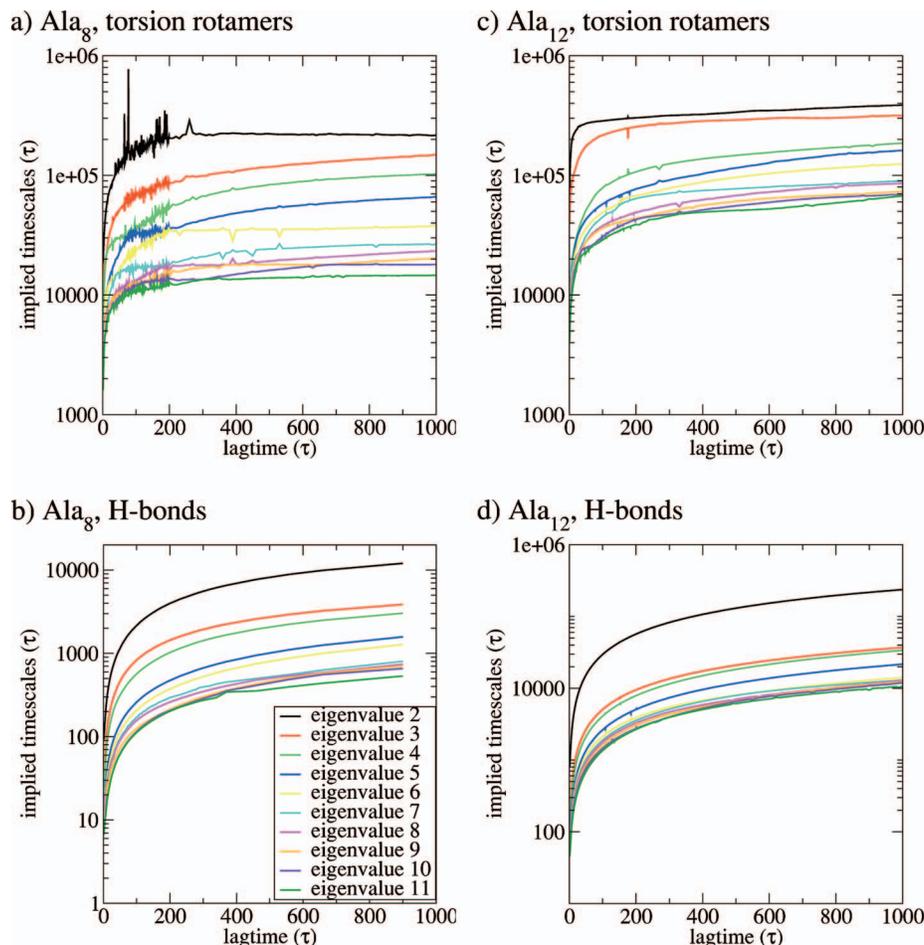


FIG. 11. (Color) The implied time scales of the processes associated with individual eigenvectors, depending on the lag time $n\tau$, computed as $\tau_i^*(n\tau) = n\tau / \ln[\lambda_i(n\tau)]$, where λ_i is the i th eigenvalue. The implied time scales for the torsion rotamer states, [(a) and (c)] become flat for most processes at around $100\tau = 5$ ps, which is much below the lifetimes of the metastable states, indicating that the interstate transitions are Markovian. This is not the case for the hydrogen-bond state definition, [(b) and (d)] whose time scales do not converge within 1000τ .

and hence decrease the quality of the model. On the other hand, the presence of solvent viscosity is likely to reduce the memory of the system and thus increase the Markovian nature at a given time scale τ . The study of the effects of the solvent model or other setup parameters on transition networks and Markov state models is an interesting problem for the near future.

One of the main findings in the present study is that the choice of microstates that are used to be grouped into metastable states is crucial. We have compared the use of the backbone torsion rotamer states versus the pattern of the most frequently formed hydrogen bonds for both systems, Ala₈ and Ala₁₂. The results indicate that the backbone torsion rotamers are considerably more successful in separating kinetically separated states. This is indicated by (1) the longer lifetimes for the metastable states constructed from torsion rotamers than those constructed from hydrogen-bonding patterns, (2) the fact that the torsion rotamer networks are rather sparsely connected while the hydrogen-bonding networks are fully connected, and (3) the fact that a Markov state model constructed with metastable states based on torsion angles reproduces the dynamics observed in the molecular dynamics simulations much better than a Markov state model constructed with the hydrogen bond metastable state does. This relative failure of hydrogen-bonds is consistent with the results of Ref. 1 where macrostates defined for a β peptide by hydrogen-bonding patterns were not successful in producing Markovian behavior. However, one cannot conclude that hy-

drogen bonds are generally inappropriate in defining metastable states. The clear votum in the present study is certainly due to the fact that there are few long-lived hydrogen bonds in the simulated dynamics of these short peptides. The results for Ala₁₂, in which the hydrogen-bond based analysis fares better than for Ala₈, support the idea that hydrogen bonds become a better indicator for metastability as the system lengthens and contains more stable hydrogen-bond structures. Moreover the hysteresis used here to switch between hydrogen-bonded and hydrogen-nonbonded states introduces some artificial memory into the dynamics which may contribute to the relative failure of metastable states defined by the hydrogen-bonding pattern in the Markovity test (3).

For the dynamics of a protein in its native state, hydrogen bonds might be a better choice than torsion rotamers, as, for example, shear motions of secondary structure elements relative to each other may happen without a significant change in torsion angle space, while being clearly related to changes in the hydrogen bonds on the interface. Moreover, unlike in the present case, torsions in proteins are typically not statistically independent but move rather in a concerted fashion so as to preserve the backbone chain closure. In this case, sets of mutually correlated angles must not be clustered independently, but rather using, for example, a multidimensional grid in the configuration space spanned by these angles, which makes the use of torsional coordinates more difficult.

An appropriate discretization of distance-based coordi-

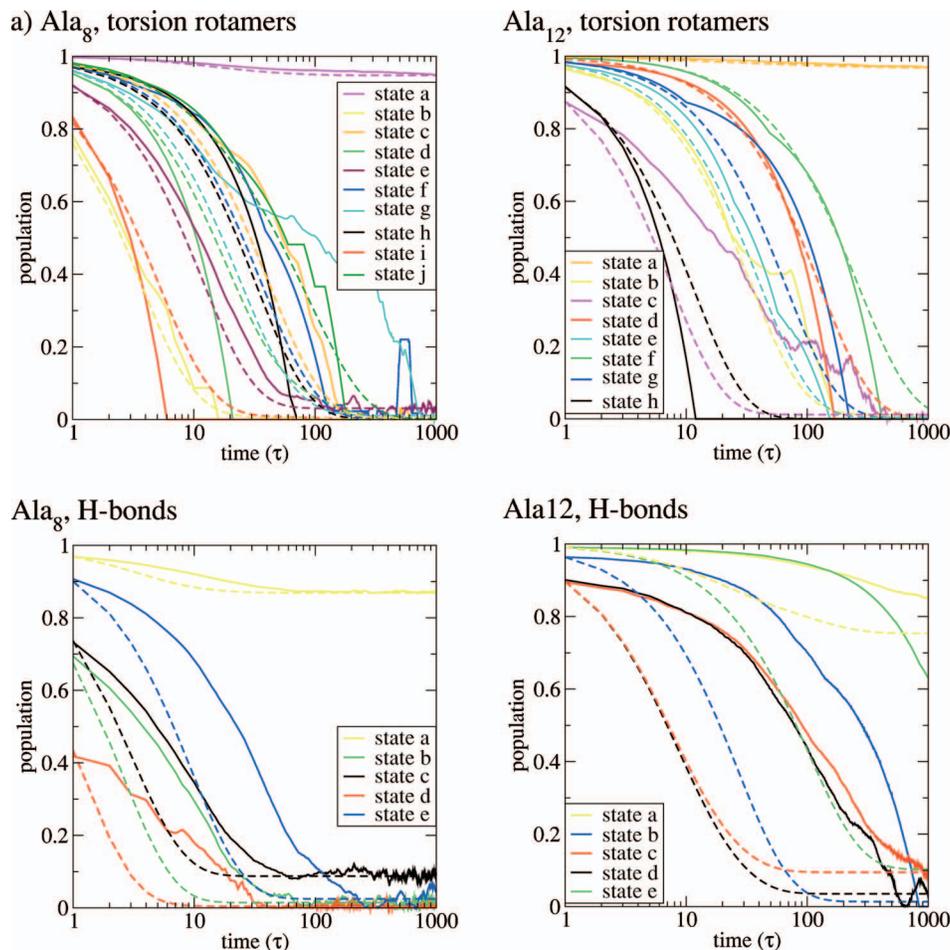


FIG. 12. (Color) Comparison of the molecular dynamics simulations (solid lines) with the Markov dynamics produced by the model network of metastable states (dashed lines) for a lag time of 1000τ . The Markov dynamics is run C times, each time initializing a single state with population 1 and the remaining states with population 0. The dashed curves show the relaxation of the selected state's population towards equilibrium. For comparison, the relaxation dynamics directly computed from the molecular dynamics trajectory are shown. While the Markov model based H-bond definition fails for all states, the one based on the torsion rotamer definition fits well for many states and for all states in the high population range. The mismatch for some states in the low-population range [states i and d for Ala_8 and states d, f, g , and h for Ala_{12} are likely to be due to insufficient sampling in the molecular dynamics trajectory (see number of long-time visits in Tables I and III)]. The mismatch for states g (Ala_8) and c (Ala_{12}) which are frequently visited indicates the presence of considerable internal energy barriers, which means that these states should be further decomposed into substates (by increasing the number of metastable states in the model) in order to improve the fit.

nates (which may require states intermediate to “close” and “far” like in hydrogen-bonding patterns) may be generally useful in identifying the metastabilities of the system, included for cases studied here. A first approach to this is made in the accompanying paper by Chodera *et al.* in this issue.⁴⁹

Apart from the selection of an appropriate coordinate set, the actual choice of which states to distinguish is, for complex systems, likely to be related to the process or quantities of interest. Proteins, for example, contain dynamical processes with a large range of time scales, ranging from fast methyl rotations to internal side-chain repacking processes which may last longer than the lifetime of the folded state.⁵⁰ The time scale of interest is determined by the process or quantities of interest which then determines how many metastable states may be distinguished at this time scale. However, it may not be necessary to actually distinguish all these metastable states in the model, because some slow processes may not be coupled to the process of interest. Consider the example of side-chain flips or repacking in proteins, which may occur on any time scale from picoseconds to the lifetime of the native state. If the dynamics of some ligand release is the process of interest, then there may be side-chain motions in regions that are distant from the binding site which are not significantly coupled to the product release process. Thus a good kinetic model for the product release process would not require the explicit treatment of these uncoupled processes. In fact, treating them explicitly is undesirable because each uncoupled process will at least double

the number of states to be distinguished in the model and therefore would lead to worse statistics in each individual state. It may thus be generally desirable to distinguish states according not only to their lifetimes but also to their correlation with some user-defined target properties. The development of such an adaptive analysis method is one of the methodological challenges for the future.

ACKNOWLEDGMENTS

The authors wish to thank John D. Chodera (UC San Francisco) for carefully and critically reading the manuscript and suggesting many helpful improvements. Furthermore, the authors gratefully acknowledge the funding provided by Center for Modeling and Simulation in the Biosciences, Heidelberg [to one of the authors (F.N.)] and the DFG research program “SFB450: Analysis and Control of ultrafast photoinduced reactions” [to another author (I.H.)].

¹W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, and M. Eleftheriou, *J. Phys. Chem. B* **108**, 6582 (2004).

²S. P. Elmer, S. Park, and V. S. Pande, *J. Chem. Phys.* **123**, 114902 (2005).

³F. Noé, D. Krachtus, J. C. Smith, and S. Fischer, *J. Chem. Theory Comput.* **2**, 840 (2006).

⁴N. Singhal, C. Snow, and V. S. Pande, *J. Chem. Phys.* **121**, 415 (2004).

⁵R. Czerminski and R. Elber, *J. Chem. Phys.* **92**, 5580 (1990).

⁶R. S. Berry and R. Breitengraser-Kunz, *Phys. Rev. Lett.* **74**, 3951 (1995).

⁷O. M. Becker and M. Karplus, *J. Chem. Phys.* **106**, 1495 (1996).

⁸Y. Levy and O. M. Becker, *J. Chem. Phys.* **114**, 993 (2001).

⁹Y. Levy, J. Jortner, and O. M. Becker, *Proc. Natl. Acad. Sci. U.S.A.* **98**,

- 2188 (2001).
- ¹⁰ P. N. Mortenson and D. J. Wales, *J. Chem. Phys.* **114**, 6443 (2001).
- ¹¹ Y. Levy, J. Jortner, and O. M. Becker, *J. Chem. Phys.* **115**, 10533 (2001).
- ¹² P. N. Mortenson, D. A. Evans, and D. J. Wales, *J. Chem. Phys.* **117**, 1363 (2002).
- ¹³ D. A. Evans and D. J. Wales, *J. Chem. Phys.* **118**, 3891 (2003).
- ¹⁴ D. A. Evans and D. J. Wales, *J. Chem. Phys.* **119**, 9947 (2003).
- ¹⁵ D. A. Evans and D. J. Wales, *J. Chem. Phys.* **121**, 1080 (2004).
- ¹⁶ F. Despa, D. J. Wales, and R. S. Berry, *J. Chem. Phys.* **122**, 024103 (2005).
- ¹⁷ N. Singhal and V. S. Pande, *J. Chem. Phys.* **123**, 204909 (2005).
- ¹⁸ F. Noé, M. Oswald, G. Reinelt, S. Fischer, and J. C. Smith, *Multiscale Model. Simul.* **5**, 393 (2006).
- ¹⁹ C. Schütte and W. Huisinga, *Handbook of Numerical Analysis, Computational Chemistry Vol. X*, edited by P. G. Ciaret and J.-L. Lions (North-Holland, Amsterdam, 2003), pp. 699–744.
- ²⁰ H. Hotelling, *J. Educ. Psychol.* **24**, 417 (1933).
- ²¹ L. S. D. Caves, J. D. Evanseck, and M. Karplus, *Protein Sci.* **7**, 649 (1998).
- ²² M. E. Karpen, D. T. Tobias, and C. L. Brooks III, *Biochemistry* **32**, 412 (1993).
- ²³ B. de Groot, X. Daura, A. Mark, and H. Grubmüller, *J. Mol. Biol.* **301**, 299 (2001).
- ²⁴ G. Moraitakis and J. M. Goodfellow, *Biophys. J.* **84**, 2149 (2003).
- ²⁵ E. J. Sorin and V. S. Pande, *Biophys. J.* **88**, 2472 (2005).
- ²⁶ S. P. Elmer, S. Park, and V. S. Pande, *J. Chem. Phys.* **123**, 114903 (2005).
- ²⁷ M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6801 (2005).
- ²⁸ O. M. Becker, *Proteins* **27**, 213 (1997).
- ²⁹ F. A. Hamprecht, C. Peter, X. Daura, W. Thiel, and W. F. van Gunsteren, *J. Chem. Phys.* **114**, 2079 (2001).
- ³⁰ I. Horenko, E. Dittmer, A. Fischer, and Ch. Schütte, *Multiscale Model. Simul.* **5**, 802 (2006).
- ³¹ V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, *J. Chem. Theory Comput.* **1**, 515 (2005).
- ³² M. Weber, ZIB Report No. 03–04, 2003 (unpublished).
- ³³ P. Deuffhard and M. Weber, ZIB Report No. 03–09, 2003 (unpublished).
- ³⁴ B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.* **4**, 187 (1983).
- ³⁵ M. Schaefer and M. Karplus, *J. Chem. Phys.* **100**, 1578 (1996).
- ³⁶ A. Strehl and J. Ghosh, *J. Mach. Learn. Res.* **3**, 583 (2002).
- ³⁷ Y. Y. Yao, *Entropy Measures, Maximum Entropy Principle and Emerging Applications* (Springer, New York, 2003), Chap. 6, pp. 115–136.
- ³⁸ O. F. Lange and H. Grubmüller, *Proteins* **62**, 1053 (2006).
- ³⁹ M. Weber, W. Rungtarityotin, and A. Schliep, ZIB Report No. 04–39 (2004) (unpublished).
- ⁴⁰ C. Schütte, A. Fischer, W. Huisinga, and P. Deuffhard, *J. Comput. Phys.* **151**, 146 (1999).
- ⁴¹ N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 4th ed. (Elsevier, Amsterdam, 2006).
- ⁴² P. Deuffhard, W. Huisinga, A. Fischer, and C. Schütte, *Linear Algebr. Appl.* **315**, 39 (2000).
- ⁴³ W. C. Swope, J. W. Pitera, and F. Suits, *J. Phys. Chem. B* **108**, 6571 (2004).
- ⁴⁴ S. Park and V. S. Pande, *J. Chem. Phys.* **124**, 054118 (2006).
- ⁴⁵ J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, *Multiscale Model. Simul.* **5**, 1214 (2006).
- ⁴⁶ A. Chakrabarty, T. Kortemme, and R. L. Baldwin, *Protein Sci.* **3**, 843 (1994).
- ⁴⁷ M. A. Moret, P. G. Bascutti, P. M. Bisch, and K. C. Mundim, *J. Comput. Chem.* **19**, 647 (1998).
- ⁴⁸ S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14766 (2004).
- ⁴⁹ J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, *J. Chem. Phys.* **126**, 155101 (2007).
- ⁵⁰ J. D. Honeycutt and D. Thirumalai, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 3526 (1990).