THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

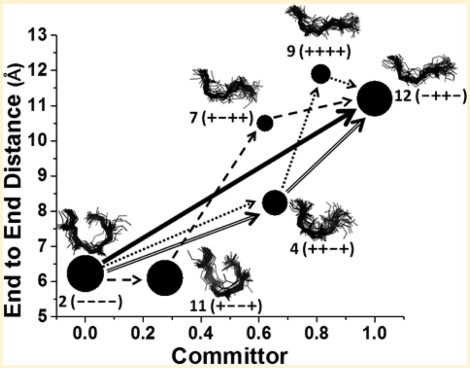
pubs.acs.org/JPCB

Transition Paths of Met-Enkephalin from Markov State Modeling of a Molecular Dynamics Trajectory

Rahul Banerjee and Robert I. Cukier*

Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States

ABSTRACT: Conformational states and their interconversion pathways of the zwitterionic form of the pentapeptide Met-enkephalin (MetEnk) are identified. An explicit solvent molecular dynamics (MD) trajectory is used to construct a Markov state model (MSM) based on dihedral space clustering of the trajectory, and transition path theory (TPT) is applied to identify pathways between open and closed conformers. In the MD trajectory, only four of the eight backbone dihedrals exhibit bistable behavior. Defining a conformer as the string XXXX with X = "+" or "−" denoting, respectively, positive or negative values of a given dihedral angle and obtaining the populations of these conformers shows that only four conformers are highly populated, implying a strong correlation among these dihedrals. Clustering in dihedral space to construct the MSM finds the same four bistable dihedral angles. These state populations are very similar to those found directly from the MD trajectory. TPT is used to obtain pathways, parametrized by committor values, in dihedral state space that are followed in transitioning from closed to open states. Pathway costs are estimated by introducing a kinetics-based procedure that orders pathways from least (shortest) to greater cost paths. The least costly pathways in dihedral space are found to only involve the same XXXX set of dihedral angles, and the conformers accessed in the closed to open transition pathways are identified. For these major pathways, a correlation between reaction path progress (committors) and the end-to-end distance is identified. A dihedral space principal component analysis of the MD trajectory shows that the first three modes capture most of the overall fluctuation, and pick out the same four dihedrals having essentially all the weight in those modes. A MSM based on root-mean-square backbone clustering was also carried out, with good agreement found with dihedral clustering for the static information, but with results that differ significantly for the pathway analysis.



1. INTRODUCTION

Atomistic simulations of proteins using molecular dynamics (MD) and Monte Carlo (MC) methods tend to stay around their initial configurations. Even for peptide simulations in explicit solvent, the equilibria between different substates may not be properly sampled unless simulations reach microsecond time scales. This generic sampling problem, where barriers are large compared with the thermal energy separate stable states, is a major concern in MD and MC simulations. Methods such as multicanonical ensemble,[1,2] simulated tempering,[3,4] and parallel tempering, also referred to as the replica exchange method (REM),[2,5−10] were designed to address this issue. Another approach is through the construction of Markov state models based on, e.g., MD trajectories, some with the use of multiple short trajectories[11−13] to enhance configuration space coverage. By focusing on metastable states identified by some configuration space clustering algorithm, it may be possible to construct a discrete state-space continuous time Markov process that could sample events on a longer time scale than the original trajectory time scale.[14−21] Software[22,23] for MSM construction and analysis is available to the research community. Replica exchange molecular dynamics trajectories have been used to construct a Markov state model for peptide folding studies.[24]

Because a MSM relies on an adequate definition of a state space, a great deal of effort has gone into providing a variety of clustering methods, and coordinates to which the clustering can be applied, such as clustering in RMSD (root-mean-square distance) and dihedral angle spaces.[25,26] With a set of metastable states defined, attention can focus on obtaining pathways that connect these states. One approach is via transition path theory (TPT)[13,27,28] that has found great utility in protein folding studies to identify pathways of folding to a native state,[29] for peptide conformation exploration,[24−26,30] for finding binding sites,[31] and for other purposes.[32] TPT, based on the idea of committor analysis,[33,34] permits construction of a graph of (one-way) fluxes that connect specified source and destination states (e.g., unfolded and folded states of a protein[29]) through a set of intermediate states. The various pathways connecting source and destination can then be obtained as a function of the committor values of the intermediate states. And, potential correlations between committor values and other putative reaction coordinates can be investigated.

Received: December 11, 2013
Revised: February 26, 2014
Published: February 26, 2014

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

In this work, the conformational space of Met-enkephalin (MetEnk), an opioid pentapeptide with sequence Tyr-Gly-Gly-Phe-Met, is explored on the basis of a long MD trajectory. MetEnk has been shown to exhibit great conformational plasticity by experiment[35−40] and computation.[7,41−52] Of particular interest is the zwitterionic form (protonated N-terminus and ionized C-terminus), which should predominate in polar media. The competition between stabilization due to salt-bridged conformers and charge-solvated (terminal peptide charges interacting with solvent dipoles) conformers provides stable closed and open conformers, respectively.[51] A study[51] of MetEnk in explicit solvent using both distance REM and Hamiltonian REM, found that MetEnk has a distinct salt-bridge form that is separated by a low barrier from a broad range of open forms, when the end-to-end distance is used as a PMF reaction coordinate, with a roughly equal mixture of closed and open form conformers. The Hamiltonian REM provided a more complete picture of the conformational states by revealing another significant reaction coordinate, corresponding to a correlated compensating transition of succeeding psi and phi dihedrals that permits the existence of two distinct salt-bridged, closed conformers. Explicit solvent zwitterionic[44−46,51] studies, where both open and closed conformations were found, show that the time scale for these transitions is in the multi-nanosecond range.

In the current work, MD is used on a microsecond time scale to obtain good configurational sampling. While the MD is long enough to sample the backbone configurations well, it is of interest to use available software[22,23] to build a Markov state model for the purpose of investigating pathways between the closed and open states of MetEnk based on transition path theory. For backbone conformation exploration, the psi/phi dihedral angles of the peptide are natural internal coordinates, as used in the MSM context before.[25,26]

Four views of the states are obtained in this work: (1) those obtained from directly monitoring the eight psi/phi backbone dihedrals from the MD trajectory, (2) those obtained from using dihedral space clustering to obtain states appropriate to a MSM, (3) those obtained from using RMSD (root-mean-square-distance) clustering to obtain dihedral states appropriate to a MSM, and (4) those obtained from carrying out a principal component analysis (PCA)[53] on the MD trajectory in dihedral space and investigating the first few modes. As will be shown here, there is good agreement among all these methods of defining the dihedral state space.

MetEnk can transit between predominant metastable dihedral states that correspond to closed (source) and open (destination) conformers, and TPT can be used to construct the graph of intermediates spanning these source and destination states. From the flux values connecting pairs of states, pathways between source and destination conformers can be obtained. Then, with an approach we introduce to enumerate pathway costs, a number of pathways that are ordered in terms of their "costs" can be found. Of interest is to identify which of the dihedral angles define the dihedral states that are involved in the major pathways through which MetEnk passes in the transition from closed to open conformations. In addition, it is of interest to see if there is a correlation between the intermediates' committor values and the end-to-end distance. In this way, a view of how MetEnk samples what is found to be a very restricted dihedral space is obtained, yet it is sufficient to provide facile closed to open pathways.

The rest of this manuscript is organized as follows. In section 2, the MD simulation setup is described, along with the analysis methods used on the trajectory. The MSM particulars are detailed that lead to the committor values and the graph of the state-to-state fluxes. An analysis of how the MSM relaxation times become independent of the coarse-graining time used in their construction is given, based on a two-state model. We introduce a method to obtain the shortest, least cost pathways from the fluxes based on a K-shortest path algorithm.[54] Section 3 analyzes the MetEnk MD trajectory to find which of the dihedral angles are responsible for the conformational states of MetEnk. MSM clustering is used to create dihedral states, along with their populations, committors, and flux graph. The main pathways spanning closed to open states are obtained, and the correlation between end-to-end distance and committor values investigated. The results on PCA in dihedral space are presented here. A comparison of the MSM and TPT analyses based on RMSD versus dihedral clustering is also carried out. Section 4 presents our main conclusions.

2. METHODS

2.1. MD Protocol. The protein molecular dynamics program CUKMODY[55] that uses the GROMOS96 force field[56] was used to generate the MetEnk trajectories. The trajectory was initiated from a structure picked from an NMR-generated ensemble[38] (PDB 1PLW) corresponding to an open form. In this configuration, the end-to-end distance (nitrogen of the N-terminus to carboxylate carbon of the C-terminus) is 10.5 Å; this distance in the ensemble of 80 lowest energy structures is ∼10−11 Å. The simulations were carried out in a cubic box with side 29.6 Å, initially with 864 waters, and 51 were removed to prevent overlap with MetEnk. The temperature was set to 303 K with a Berendsen thermostat[57] with a relaxation time of 0.2 ps.[58] SHAKE[58] was used to constrain bond distances permitting use of a 2 fs time step. For the evaluation of the electrostatic and attractive parts of the Lennard-Jones energies and forces, the PME method[59] was applied, with a direct-space cutoff of 8.56 Å, an Ewald coefficient of 0.45, and a 30 × 30 × 30 $Å^{-3}$ reciprocal space grid. After equilibration, a total of 2 μs of trajectory data was collected with protein coordinates written out every 1 ps.

2.2. Analysis of the MD Trajectory. The program ANALYZER[60] was used to obtain properties of the trajectory. Our previous simulations[51] of MetEnk found that the end-to-end distance between open and closed forms is sampled rather slowly. The 2 μs trajectory provided hundreds of such transitions, indicating adequate sampling. To check for convergence of these transitions, the data was split into equal records and the number of end-to-end transitions was found to be roughly the same in each record. As this end-to-end fluctuation is the slowest motion, as shown by a previous PCA analysis[51] and will be evident from the separation of relaxation times in the MSM analysis, the other modes of motion should also be well represented by this trajectory. The trajectories of the eight psi and phi backbone dihedral angles for the five residues were histogramed to find those that sample one conformation versus those that sample two conformations. Again to confirm that the simulation is of sufficient length, these phi and psi dihedral histograms were compared for the trajectory split into two parts with good agreement.

When the MSM clustering is carried out in dihedral space, which will be referred to as DIHED clustering, the trajectory for each state was analyzed as above to obtain the dihedral state

histograms along with the end-to-end distances of each dihedral state.

Principal component analysis[53,61,62] (PCA) was used to analyze the trajectory data. PCA diagonalizes the trajectory averaged ⟨...⟩ covariance matrix

$$\mathbf{C}_{\mathbf{XX}} = \langle \delta\mathbf{X}(t)\delta\mathbf{X}^{\mathrm{T}}(t)\rangle \tag{1}$$

of atom, or groups of atoms, fluctuations $\delta\mathbf{X}(t)$ from their trajectory-averaged values, using, e.g., the Cartesian components of the atoms. It decomposes the configuration point as

$$\delta\mathbf{X}(t) = \sum_{k=1}^{N_X} p_k(t)\mathbf{m}_k \tag{2}$$

where $\mathbf{m}_k$ are the (orthonormal) eigenvectors of the covariance matrix and the corresponding eigenvalues are denoted as λ_k^2. In the rotated Cartesian coordinate basis defined by $\mathbf{m}_k$ (k = 1, 2, ..., $3N$), the largest eigenvalue captures the largest fraction of the root-mean-square fluctuation (RMSF), the second largest the next largest fraction of the RMSF, etc. Ordering the eigenvalues from large to small leads, in favorable cases, to a small set of modes that capture most of the fluctuation. PCA is not restricted to harmonic motions; it can describe collective transitions between structures that differ greatly. It is well suited to identifying conformational states of peptides.[63] To carry out PCA in dihedral space, it needs to be formulated in a way that provides a Euclidian-space metric. One way to accomplish this is by use of the transformation

$$\mathbf{X}(t) = (\sin\vartheta_1(t), \cos\vartheta_1(t), ..., \sin\vartheta_{N_X}(t), \cos\vartheta_N(t)) \tag{3}$$

proposed by Stock and co-workers.[64,65] This "doubling" procedure eliminates the metric problems associated with dihedral angles as defined on $-\pi \le \vartheta_i(t) < \pi$. The method has been incorporated into ANALYZER.[60]

PCA can provide "participations";[52,66−68] these are defined as the proportions of the coordinates that contribute to each PCA eigenvector. In Cartesian coordinates, the participations are given by $(\mathbf{m}_k^{i,x})^2 + (\mathbf{m}_k^{i,y})^2 + (\mathbf{m}_k^{i,z})^2$ for the ith atom in the kth PCA eigenvector. The corresponding participation definition for the ith dihedral angle in the doubled dihedral space is $(\mathbf{m}_k^i)^2 + (\mathbf{m}_k^{i+1})^2$. A large participation value for the ith dihedral angle in the kth mode indicates that this mode is largely due to the ith dihedral's fluctuations.

2.3. MSM Analysis of a Two-State System. Markov state models in continuous time for a discrete state space can be formulated either by a rate equation perspective[67,68] or by a transfer matrix perspective.[17,22] They are of course formally equivalent. Here we use the latter approach because MSMBuilder[22] is formulated using transfer matrices. In this approach, the Markov process is characterized as

$$\mathbf{p}^{\mathrm{T}}(t+\tau) = \mathbf{p}^{\mathrm{T}}(t)\mathbf{T}(\tau) \tag{4}$$

with $\mathbf{p}^{\mathrm{T}}(t)$ being a row vector of state populations, $\mathbf{T}(\tau)$ a column stochastic matrix of state-to-state transition probabilities, and τ a coarse graining time characterizing an intermediate time scale between microscopic time scales (collision events) and a macroscopic time where relaxation to equilibrium has occurred. Relaxation times

$$t_i(\tau) = -\frac{\tau}{\ln\lambda_i(\tau)} \tag{5}$$

in the rate equation formulation can be related to the eigenvalues $\lambda_i(\tau)$ of $\mathbf{T}(\tau)$ from the transfer matrix approach. They provide the relaxation times in an eigenfunction−eigenvalue decomposition of the relaxation of the state populations to their equilibrium values, with the ordering 0 = $t_1(\tau) < t_2(\tau) < ...$, with $t_1(\tau) = 0$ to ensure a state of equilibrium is reached.

For the MSM procedure to provide a Markov process, the $t_i(\tau)$ must become independent of τ. Here we explore how this occurs for the simplest system composed of two states with equal equilibrium populations. It is then straightforward, noting the row-stochastic constraint of summing to unity, to obtain the transition probability matrix:

$$\mathbf{T} = \begin{pmatrix} \frac{1}{2}(1+\lambda_2(\tau)) & \frac{1}{2}(1-\lambda_2(\tau)) \\ \frac{1}{2}(1-\lambda_2(\tau)) & \frac{1}{2}(1+\lambda_2(\tau)) \end{pmatrix} \tag{6}$$

where

$$\lambda_2(\tau) = \mathrm{e}^{-\tau/t_2(\tau)} \tag{7}$$

The behavior with $t_2(\tau)$ with τ can then be obtained and a τ value picked to construct the MSM. The elements of $\mathbf{T}(\tau)$ are related to those of the time correlation functions, $\mathbf{C}(\tau)$, according to[17]

$$T_{ij}(\tau) = c_{ij}(\tau)/c_i \tag{8}$$

where the c_i are the equilibrium populations; here $c_1 = c_2 = 1/2$. For long τ, $c_{ij}(\tau) \to c_i c_j = (1/2)(1/2) = 1/4$, because the time correlation functions must eventually factorize, as is consistent with eq 6 at long time.

From the secular determinant that provides the eigenvalues $\mu_i(\tau)$ of $\mathbf{T}(\tau)$, one eigenvalue is $\mu_1(\tau) = 1$, the equilibrium eigenvalue, and the other is $\mu_2(\tau) = \lambda_2(\tau)$. As τ increases from zero, the diagonal elements of eq 6, noting eq 7, start at 1 and fall to 1/2.

Now consider the connection to the master equation perspective.[68,69] It is obtained from the Chapman−Kolmogorov equation

$$\mathbf{T}(\tau+\tau') = \mathbf{T}(\tau)\mathbf{T}(\tau') \tag{9}$$

along with the assumption

$$\mathbf{T}(\tau) = (1 - \tau\mathbf{K}\cdot\mathbf{I})\mathbf{I} + \tau\mathbf{K} \tag{10}$$

with $\mathbf{K}$ a constant matrix of transition rates. The master equation

$$\frac{\partial\mathbf{T}(\tau)}{\partial\tau} = \mathbf{K}\mathbf{T}(\tau) \tag{11}$$

follows from using eq 10 in eq 9 and taking the $\tau' \to 0$ limit. The solution of eq 11 is

$$\mathbf{T}(\tau) = \mathbf{I}\mathrm{e}^{\tau\mathbf{K}} \tag{12}$$

with $\mathbf{I}$ being the unit matrix, initial condition $\mathbf{T}(0) = \mathbf{I}$, and $\mathbf{K}$ here is

$$\mathbf{K} = \frac{-1}{2t_2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{13}$$

with t_2 constant. Analytically, $t_2(\tau)$ must increase from zero and reach a plateau value. The plateau value of $t_2(\tau)$ must be bounded as $t_{\mathrm{micro}} \ll t_2(\tau) \ll t_{\mathrm{macro}}$, where t_{micro} characterizes the

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

microscopic fluctuations and t_{macro} characterizes the slow transition time scale between the states. *Numerically*, $t_2(\tau)$ will increase from zero, may reach a plateau value, and then will become indeterminate.

To illustrate this behavior, a simple model that mimics a protein with a slow mode in some complex coordinate arising from a large number of coupled two-state degrees of freedom (e.g., from dihedral angles) was analyzed. It is composed of N particles in double well potentials that are coupled to each other. Stochastic trajectories are generated by solution of a Langevin equation (LE) in these coordinates as detailed elsewhere.[70] A slow mode, appropriate for Markov modeling, is the reaction coordinate, R, the average position of the particles. For appropriately chosen parameters for the double well potentials, their coupling strengths, the number of particles, and the temperature, the potential of mean force of the collective coordinate, PMF(R), will exhibit two equal population stable states separated by a free energy barrier. A sufficiently long trajectory was run to obtain an extremely accurate PMF.

The correlation function matrix in eq 8 is readily evaluated from the reaction coordinate trajectory. Of course, a decision has to be made as to what constitutes a state, with some cutoff in the value of the reaction coordinate to eliminate values corresponding to the transition state. The correlation function matrix elements are shown in Figure 1. The diagonal (off-diagonal) elements start at 1/2 (0) and all approach 1/4 at large τ. The eigenvalue $\lambda_2(\tau)$ is obtained by constructing the matrix $\mathbf{T}(\tau)$ from the $\mathbf{C}(\tau)$ matrix, and diagonalizing it for each τ value. (The other, equilibrium eigenvalue is unity for all τ as must be.)

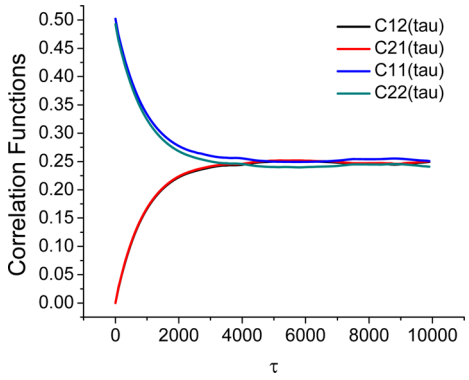


Figure 1. The elements of the correlation matrix that are related to the transition matrix $T_{ij}(\tau) = c_{ij}(\tau)/c_i$, with c_i being the equilibrium populations. The diagonal (off-diagonal) elements start at 1/2 (0) and all approach $c_{ij}(\tau) \rightarrow c_i c_j = (1/2)(1/2) = 1/4$ at large τ.

Using eq 7, the behavior of $t_2(\tau)$ is shown in Figure 2. It increases from zero to reach a plateau value for a range of τ values and then becomes erratic. It is easy to see why this happens from the $\mathbf{C}(\tau)$ plots. For $\tau \ll t_2(\tau)$, no transitions occur. For $\tau \sim t_2(\tau)$, $t_2(\tau)$ is well determined because $\mathbf{C}(\tau)$ varies substantially with τ. For $\tau \gg t_2(\tau)$, $t_2(\tau)$ cannot be well-determined because essentially the elements of $\mathbf{C}(\tau)$ (and $\mathbf{T}(\tau)$) approach the same limiting values. Then, for large τ, the eigenvalue $\lambda_2(\tau)$ must approach zero from above, and its numerical evaluation is ill-conditioned.

In more complex situations with multiple relaxation times and with data that is not so extensive, there will be a tendency for the longer relaxation time "constants" to not approach their plateau values and for the shorter relaxation time "constants" to, for a given value of τ, be ill-determined. The latter problem arises from correlation matrix elements that approach their limiting behavior. Note that, in addition to the potential numerical difficulties just addressed, another source of error in the evaluation of relaxation times arises from the finite length of simulated trajectories.

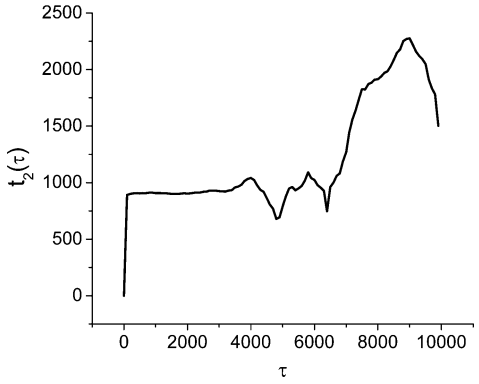


Figure 2. The behavior of $t_2(\tau)$ with τ. It starts at zero, reaches a plateau value, and for longer τ becomes unreliable, as is evident from the approach to the same limiting value of the correlation function matrix elements displayed in Figure 1.

2.4. MSM on the MetEnk MD Trajectory. The 2 μs MD trajectory for MetEnk was used to construct a Markov state model using the MSMBuilder[22] software package. The MSM is built by first clustering the atom coordinate trajectories with some criterion. Here, we cluster in phi/psi backbone dihedral space, which will be referred to as DIHED clustering (to distinguish this clustering from the direct way of histograming the dihedral angles directly from the MD trajectory, as discussed in section 2.2). For each snapshot, the sin and cos of the eight dihedrals form a vector, and the Euclidean distances among these vectors used in the clustering. We also cluster in backbone root-mean-square-displacement (RMSD) space where the RMSD is evaluated over all four backbone (C, CA, C, O) atoms of each of the five residues. To define states, a clustering algorithm parametrized by a cluster cutoff must be given. The "hybrid kcenter-kmedoid"[22] algorithm was used. For the DIHED clustering, a cutoff of 3.5° was used, and for RMSD clustering, a cutoff of 2.0 Å was used. Note that the DIHED clustering is a sum over all eight phi and psi dihedral angles, in analogy to the RMSD clustering. RMSD clustering does suffer from the requirement of first having to best fit the MD trajectory snapshots to, here, the initial backbone configuration. For a flexible peptide, that does introduce some imprecision that is not present for internal coordinates, such as the dihedral angles. These cutoffs produced 14 DIHED states and 15 RMSD states.

Once states are defined, the eigenvalues $\lambda_i(\tau)$ of $\mathbf{T}(\tau)$, the transition matrix, are constructed and related to the time scales according to eq 5 to obtain a coarse graining time τ. With this time set, the final transition matrix $\mathbf{T}$ is obtained with a routine that uses a maximum likelihood estimation method that leads to a symmetrized $\mathbf{T}$ that ensures detailed balance is numerically obeyed, in spite of finite data. The corresponding equilibrium probabilities, c_i (defined in section 2.3), are also obtained. These are the ingredients of the MSM. In principle, this model could be used to model events on longer time scales than the MD time scale. However, for MetEnk, the MD trajectory is sufficiently long to adequately explore the backbone configuration space.

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

2.5. Committors and Fluxes. With the transition matrix and equilibrium probabilities available, transition path theory can be used to investigate the sequence of intermediates that MetEnk follows in transiting between given end point states. Derivations of TPT have been outlined[13] and detailed elsewhere.[27,71,72] The theory is built up from the concept of a committor, originally used for finding dividing surfaces between stable states.[33,34] For the purposes of TPT, the committor is defined via the following probability

$$q_I^+ = P(\tau_B^+ < \tau_A^+) \tag{14}$$

where $\tau_A^+ = \min(\mathbf{X}(t) \in A)$ ($\tau_B^+ = \min(\mathbf{X}(t) \in B)$) are the times to arrive in $A(B)$ from intermediate I without first arriving in $B(A)$. The committors satisfy $0 \le q_I^+ \le 1$ where, naturally, $q_I^+ = 0 (I \in A)$ and $q_I^+ = 1 (I \in B)$. Thus, smaller (larger) values of q_I^+ indicate that the probability to first reach B before reaching A is lower (higher). The backward committor q_I^- for an equilibrium trajectory satisfies $q_I^- = 1 - q_I^+$. The committors can be evaluated from the elements of the transition matrix $\mathbf{T}$ by linear algebra,[13] as implemented in the MSMBuilder[22] software package. Filtering the transition probabilities with the committors q_I^+ and q_J^-, weighted with the equilibrium probabilities, c_I, to provide fluxes f_{IJ}^{AB} that contribute to $A \rightarrow B$ transitions according to

$$f_{IJ}^{AB} = c_I q_I^- T_{IJ} q_J^+ \tag{15}$$

and obtaining net fluxes as

$$(f_{IJ}^{AB})^+ = \max[0, f_{IJ}^{AB} - f_{JI}^{AB}] \tag{16}$$

provides a scheme to obtain pathways of intermediates in an $A \rightarrow B$ transition. Note that these fluxes are dimensionless, because they are defined in terms of the transition matrix $\mathbf{T}$. They can be related to dimensional (inverse time) fluxes using the connection $T_{ij}/\tau = K_{ij}$ $(i \neq j)$, obtained from eq 10. On average, these net fluxes provide ordered event sequences for the progress of reactive trajectory segments (those parts of the total trajectory that begin in state A and end in state B). For the purpose of obtaining a set of pathways ordered by their relative costs, we will use the dimensionless definition of eq 15.

2.6. First K Pathways. There are various ways of ordering the costs of pathways.[27,29,72] One strategy[29] is based on bottlenecks, where successive pathways are defined by eliminating the bottleneck, the rate limiting step, from the current pathway. Here, we suggest and will use another way to order the pathways based on a kinetic equation approach. The state-to-state fluxes form a linear reaction network, and the desired pathways through it can be viewed as corresponding to a set of consecutive reactions between given source A and destination B species.

For convenience, use $f_i \equiv (f_{IJ}^{AB})^+$ to denote the net flux between state pairs I and J in the $A \rightarrow B$ transition. The corresponding "cost" of this flux is $1/f_i$. Thus, if the sequence of net fluxes contributing to pathway α are indexed as f_i^α, then the overall reactive flux f^α for path α is given by

$$\frac{1}{f^\alpha} = \sum_{i=1}^{N_\alpha} \frac{1}{f_i^\alpha} \tag{17}$$

where i runs over the N_α net fluxes in pathway α spanning the source and destination states. The pathways are then to be ordered according to $f^1 > f^2 > \ldots > f^K$ for the K top flux pathways. As the simplest one-intermediate example, consider a pathway $A \rightarrow I \rightarrow B$ with probabilities and transition matrix elements about the same for both net fluxes. Then, $f \sim q_I^+ q_I^- = q_I^+(1 - q_I^+)$ and this flux will be maximal for committor values $q_I^+ = q_I^- = 1/2$.

To implement this approach, Yen's algorithm[54] can be used to find the desired paths. In this algorithm, Dijkstra's[73] method is used to find successively the first K increasing length (cost) pathways between designated source and destination nodes of a graph. The algorithm scales linearly with K and is therefore efficient. A MATLAB implementation is available.[74] The input is the net cost graph. Here, this is the matrix formed from the inverses of the net fluxes that characterize the connectivity of the graph. A straightforward program was written to sort the net flux $(f_{IJ}^{AB})^+$ array, with the source A and destination B states specified, in descending order, and, for some kept number of fluxes, a $K \times K$ dimensional matrix of the inverses produced with K set to 20 for our purposes. This cost matrix network is input to the algorithm, and the list of the K shortest paths along with their pathway costs is returned.

3. RESULTS

3.1. Four out of Eight Dihedral Angles Define a Conformational Space. The five-residue Met-enkephalin with sequence tyrosine-glycine-glycine-phenylalanine-methionine (YGGFM) has eight phi/psi backbone dihedral angles that we will label as Psi_Y1, Psi_G2, Psi_G3, Psi_F4, Phi_G2, Phi_G3, Phi_F4, and Phi_M5. From the MD trajectory, histograms of these dihedral angles show that four are in essentially one conformation and four are in two conformations with close to equal populations (data not shown). Thus, the conformational sampling is limited to the four dihedrals Psi_G2, Psi_G3, Psi_F4, and Phi_G3. Of these angles, Psi_G2 and Phi_G3 form a compensating pair that is a known feature of peptides, whereby a psi(i) and phi($i + 1$)[75] residue sequence can undergo a crankshaft motion that leads to structures with essentially the same overall conformation. From a previous study,[51] two closed (small end-to-end distance), zwitterion-like conformers were found that reflect this psi/phi compensation.

The four, two-conformation dihedrals Psi_G2, Psi_G3, Psi_F4, and Phi_G3 will be denoted by the string XXXX with X = −, + indicating, respectively, negative or positive dihedral angles. It should be noted that peptides such as MetEnk exhibit well-defined but broad dihedral angle distributions with, e.g., standard deviations of ∼25°. With four, two-conformation dihedrals, there are, in principle, $2^4 = 16$ overall conformers. Filtering the trajectory data to construct the fractional populations of the XXXX overall conformers produces the population fractions displayed in Figure 3.

It is clear that, in spite of these four dihedral angles individually exhibiting essentially two equally populated conformers, there is a correlation among those dihedrals that leads to a strong propensity for a few overall conformers. Said otherwise, there must be a strong propensity to allow only certain correlated "flips" of these dihedral angles, as will be addressed in section 3.3.

3.2. MSM-Based DIHED Clustering Also Leads to States Defined by These Four Dihedral Angles. The MSMBuilder software was used to construct a Markov state model by clustering the MD trajectory in dihedral space. The hybrid kcenter-kmedoid[22] algorithm was used, as discussed in section 2.4. We will refer to this clustering as DIHED clustering, to keep it distinct from the direct clustering of the

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

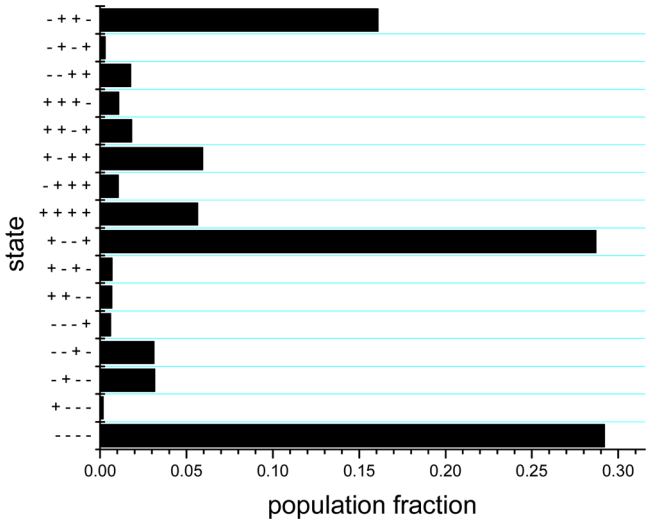


Figure 3. States and their population fractions from the MD trajectory for the four XXXX dihedral angles. The major populations agree well with those obtained from the MSM DIHED clustering algorithm (see Table 1).

trajectory employed in section 3.1, and the resulting cluster conformations will be referred to as states. For the DIHED clustering cutoff of 3.5°, 14 dihedral states were found. The relaxation times as a function of coarse graining time, τ (see section 2.4), are displayed in Figure 4.

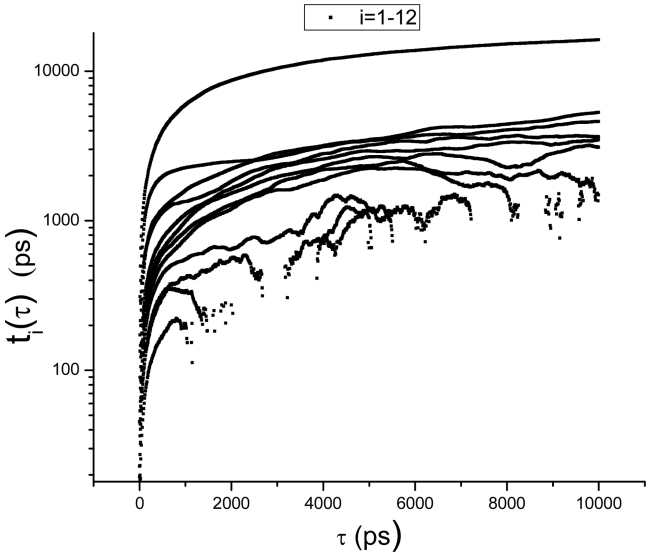


Figure 4. The first 12 relaxation times versus lag time, τ, for the MSM DIHED angle clustering. As τ increases, shorter relaxation times become ill-determined, as noted in section 2.3, though the finite trajectory length can also contribute to the irregular behavior with τ.

The plateau values of $t_i(\tau)$ are reached by $\tau \sim 600$ ps with some underestimation for the slowest modes but well before the faster modes become irregular. There is one slow mode, well separated from the rest, which should correspond to the slow end-to-end motion, as found previously by PCA in RMSD Cartesian coordinate space.[51]

The MD trajectory was used to associate snapshots with these DIHED clustered states and the histograms of the dihedral angles for each DIHED state obtained. For the six states 2, 11, 12, 4, 9, and 7, which are shown in descending order of population (see Table 1), these dihedral histogram plots are displayed in Figure 5. The histograms are generated from 10 000 samples for each state, to provide good statistics. The trajectory provides 2×10^6 samples; thus, there are more than 10 000 samples available for all these states.

Table 1 provides these state populations along with their end-to-end distances. States 2 and 11 are closed states (small end-to-end distance) and are related by the psi_G2/phi_G3 compensation mechanism, and state 12 is the highest-population open state. We focus on these six states, as they will be shown to be involved in the lowest cost pathways.

Examination of Figure 5 reveals the following interesting features: (1) The same four dihedral angles that were identified as exhibiting two conformations in section 3.1 also sample two conformations in these DIHED clustering based states, while the other four dihedral angles are here also confined to sampling one conformation. (2) There is a one-to-one correspondence between these DIHED states and the strings, XXXX, introduced in section 3.1. The mapping is given in Table 1. (3) The populations of the six DIHED states listed in Table 1 compared with the population fractions shown in Figure 3 show excellent agreement. Thus, the same correlation pattern of obtaining a limited number of states defined by the strings, XXXX, found by direct analysis of the MD trajectory, is picked up by the MSM DIHED cluster algorithm. The end-to-end distances in Table 1 show that states 4 and 7 are intermediate in terms of this distance and, in addition to the high population open state 12, state 9 is also open. Note that some of these states are quite broad, again due to the flexible nature of the peptide.

In summary, DIHED clustering of the MD trajectory finds a set of six highest population states that are unique in terms of the strings XXXX. These DIHED MSM states agree in kind and quite well in population with those found in section 3.1 by filtering the MD trajectory itself.

3.3. Pathway Analysis Shows the Four Dihedrals Are Involved in the High Flux Pathways. The transition matrix and equilibrium probabilities are the ingredients required for a transition pathway theory analysis. As discussed in section 2.5, forward committors, q_I^+, satisfying $0 \le q_I^+ \le 1$ are obtained that are the probabilities that intermediate state I between two chosen source (beginning) and destination (ending) states, A and B, respectively, reaches B without first visiting A. The forward and backward committors can be used to obtain net fluxes of a network that provide the connectivity spanning states A and B through the various intermediate states. The graph edge weights of the network are the net fluxes $(f_{IJ}^{AB})^+$. In section 2.6, we introduced a method to order the various pathway costs based on elementary kinetic arguments for consecutive reactions (or for series resistance networks), whereby the cost of pathway α is given by $(f^\alpha)^{-1} = \sum_{i=1}^{N_\alpha} 1/f_i^\alpha$, with $f_i^\alpha \equiv (f_{IJ}^{AB})^+$ and i denotes an IJ pair contributing to pathway α. Then, Yen's algorithm,[54] as implemented in a MATLAB program,[74] which provides the first K pathways ordered by increasing costs, thus higher to lower flux pathways, is used to obtain the desired significant pathways. As noted in section 2.5, because our focus is on comparing the costs of different pathways, the pathway costs, constructed from the dimensionless net fluxes defined by eqs 15 and 16, are presented without dimensions.

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

Table 1. Pathway 2−12 Dihedral States of Highest Population Obtained from DIHED Clustering with Their Corresponding Committor Values and End-to-End Distances

DIHED state	2	11	7	4	9	12
committors	0	0.275	0.621	0.655	0.814	1
populations	0.244	0.230	0.050	0.118	0.0688	0.218
EtoE[a]	6.23/1.22	6.10/1.04	10.5/1.84	8.24/2.24	11.9/1.43	11.2/1.53
diheds[b]	− − − −	+ − − +	+ + − +	+ − + +	+ + + +	− + + −

[a]End-to-end distance average and standard deviation. [b]XXXX denotes dihedrals Psi_G2, Psi_G3, Psi_F4, and Phi_G3, with negative − and positive + dihedral angle values.

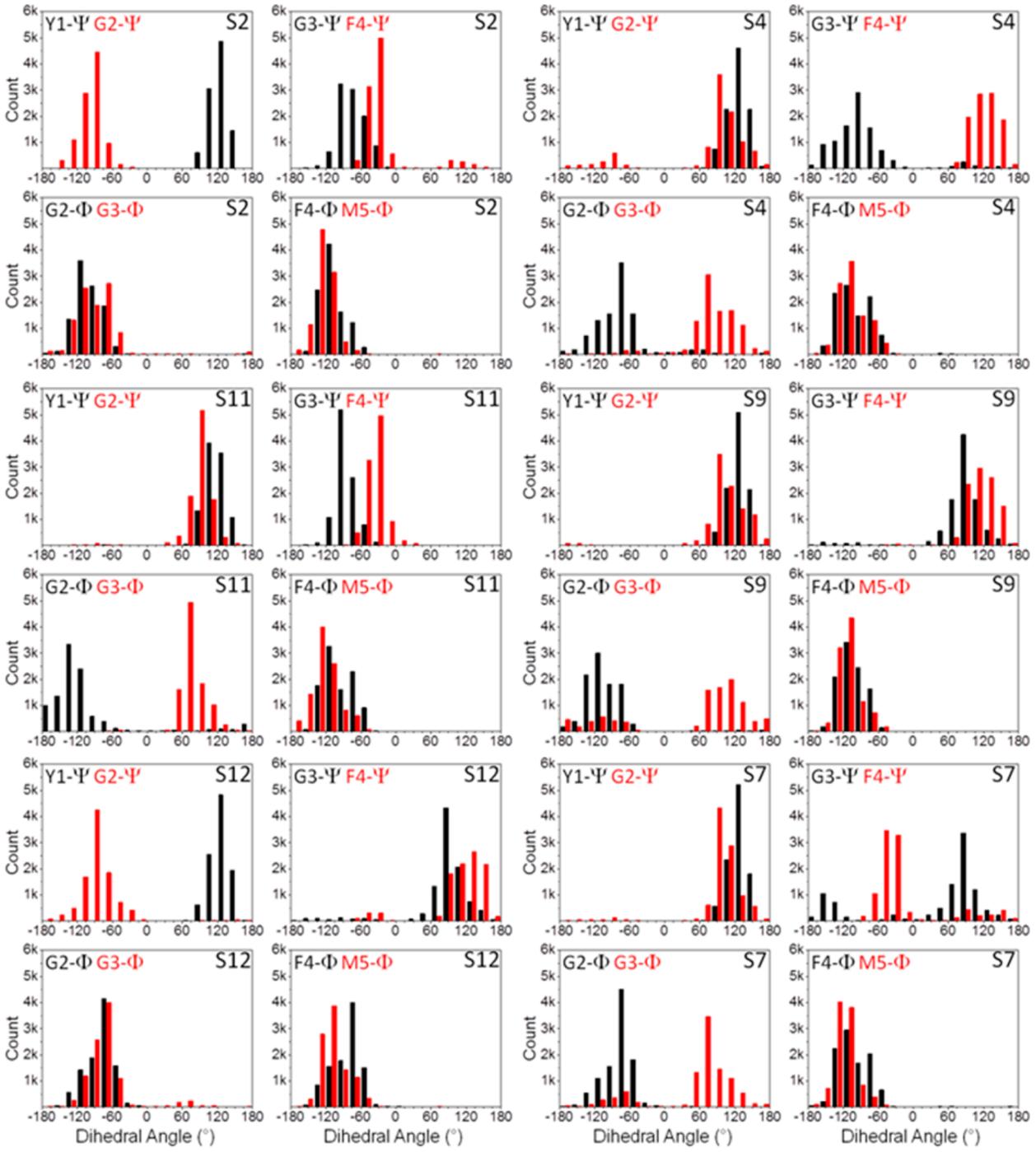


Figure 5. DIHED clustered MSM states 2, 11, 12, 4, 9, and 7 that are the most highly populated states, as listed in Table 1. States 2 and 12 correspond to closed states (small end-to-end distance), while state 11 is an open state. The histogram for each state is constructed from 10 000 samples.

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

For the *A* and *B* states, we take closed state 2 and open state 12 (the highest population open state listed in Table 1), respectively, and denote it as network 2−12. Because state 11 is also a closed state, related to state 2 by the psi/phi compensation, we also consider *A* to be state 11 and again use state 12 as *B*, and denote this as the 11−12 network. All 14 states are included in the pathway cost analysis.

Table 1 has the committor values for the 2−12 pathway that are the basis for the construction of the 2−12 network. The analysis of this network produces pathways with their respective costs listed in Table 2, with only the first eight listed. A key observation is that only intermediate states 4, 7, 9, and 11 are present in these first eight pathways. These, along with states 2 and 12, are the only states that are characterized as having bistable dihedrals of the same four XXXX angles (cf. Figure 5).

Table 2. Pathway 2−12 Costs of First Eight Pathways for State *A* = 2 (Closed) and State *B* = 12 (Open)

pathway	pathway cost
2 → 12	1.00
2 → 4 → 12	3.13
2 → 11 → 7 → 12	7.19
2 → 4 → 9 → 12	7.52
2 → 9 → 12	9.32
2 → 11 → 4 → 12	9.60
2 → 11 → 12	9.71
2 → 7 → 12	10.21

Table 3 lists the first four pathways, designated as P1−P4, along with the conformational (+ or −) designations of these four dihedrals. For P1, there is a direct transition that just requires the Psi_G3 and Psi_F4 dihedrals to flip from − to +. It is interesting that this direct path is the lowest cost. A feature of dihedral angle changes is that they can cause large Cartesian coordinate changes, here, a transition between a closed and an open conformation. P2 uses the Psi_G2/Phi_G3 compensation that is relatively easy because it does not produce a large Cartesian coordinate change, along with one of the other dihedrals to go through one intermediate, state 4.

In P3, the psi/phi compensators flip, and then flip back, interrupted by Psi_G3 flipping in the middle and Psi_F4 flipping in the last transition. In P4, again the psi/phi compensators flip and flip back, interrupted by a Psi_F4 flip, now in the first transition, and a Psi_G3 flip in the middle. Clearly, the repeated use of the low free energy psi/phi compensation mechanism is key to the P2−P4 low cost pathways.

The other pathway, closed state 11 to open state 12, has its committors listed in Table 4. (The remaining data is the same of course, as they are independent of the chosen source and destination states.) Note that the smallest nonzero committor is for state 2 that is connected with state 11 by the psi/phi compensation mechanism, as is also the case for the 2−12 pathway (see Table 1).

Table 5 lists the costs of the first eight pathways, and Table 6 indicates the dihedral angle transitions of the first five pathways. Again, the major pathways only involve the same set of DIHED states as in the 2−12 pathways, with of course state 2 replaced by state 11 here as the source state. Now, a direct path linking this closed to open state is postponed to P5, in contrast with the 2−12 pathways.

The P1 path first goes to the other closed state, 2, and then proceeds to the open state by a concerted Psi_G3/Psi_F4 flip. In all the other pathways listed, there is always at least one psi/phi compensation step. In particular, in P4 there are three such compensating steps and in P5 there is one compensation along with the other two dihedrals flipping that is required to produce a direct transition. The next two paths, schematized as P6, + − − + → + + − + → + + + + → − + + −, and P7, + − − + → + − + + → + + + + → − + + −, involve only one Psi/Phi compensation and may therefore be of lower flux than previous paths.

3.4. Committors and Pathways Are Correlated with the End-to-End Distance. It is natural to inquire if there is a correlation between the committor values and the end-to-end distances, although there is no *a priori* reason for there to be such a correlation. The committor values can be viewed as a "dynamic" reaction coordinate because they report on the sequence of events that span the transition from the source to destination state. The end-to-end distance is a natural spatial reaction coordinate because the two stable states of MetEnk should be a closed, zwitterionic state and an open state where the (charged) N and C termini are stabilized by ion−dipole and other electrostatic interactions with the water molecules. A fit of the 2−12 pathway committors versus end-to-end distances for all states shows that there is a correlation (Pearson correlation coefficient ∼0.7), with a similar result for the 11−12 pathway.

For the states involved in the major pathways, Figures 6 and 7 summarize this committor versus end-to-end distance correlation, along with the state designations, populations, and ensembles of representative backbone configurations. The pathways, listed in order of increasing cost, do show a tendency for MetEnk to keep increasing its end-to-end distance in each pathway that does involve at least one intermediate. The backbone conformations sampled are clearly distinct, though each samples a broad conformational ensemble, especially for the end dihedrals, consonant with peptide configuration exploration.

3.5. PCA in Dihedral Space Shows the Four Dihedrals Dominate the Main Modes. The MD trajectory was analyzed using PCA in backbone dihedral space with the dihedral doubling method that provides a (dimensionless) metric,[64,65] as discussed in section 2.2. The total mean square

Table 3. Pathway 2−12 Dihedral Transitions in the First Four State 2 (closed) to State 12 (open) Pathways with — Denoting No Transition and ⇒ Denoting Transition of a Particular Angle in the Indicated Transition

dihedral	P1 (2 → 12)	P2 (2 → 4 → 12)	P3 (2 → 11 → 7 → 12)	P4 (2 → 4 → 9 → 12)
Psi_G2	− — −	− ⇒ + ⇒ −	− ⇒ + → + ⇒ −	− ⇒ + — + ⇒ −
Psi_G3	− ⇒ +	− — − ⇒ +	− — − ⇒ + — +	− — − — − ⇒ +
Psi_F4	− ⇒ +	− ⇒ + — +	− — − — − ⇒ +	− ⇒ + — + — +
Phi_G3	− — −	− ⇒ + ⇒ −	− ⇒ + — + ⇒ −	− ⇒ + — + ⇒ −
pathway cost	1.0	3.13	7.19	7.52

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

Table 4. Pathway 11−12 Dihedral States of Highest Population Obtained from DIHED Clustering with Their Corresponding Committor Values and End-to-End Distances

dihedral state	11	2	7	4	9	12
committors	0	0.391	0.576	0.714	0.830	1
populations	0.230	0.244	0.050	0.118	0.0688	0.218
EtoE[a]	6.10/1.04	6.23/1.22	10.5/1.84	8.24/2.24	11.9/1.43	11.2/1.53
diheds[b]	+ − − +	− − − −	+ + − +	+ − + +	+ + + +	− + + −

[a]End-to-end distance average and standard deviation. [b]XXXX denotes dihedrals Psi_G2, Psi_G3, Psi_F4, and Phi_G3, with negative − and positive + dihedral angle values.

Table 5. Pathway 11−12 Costs of the First Eight Pathways for State *A* = 11 (Closed) and State *B* = 12 (Open)

pathway	pathway cost
11 → 2 → 12	2.95
11 → 7 → 12	5.38
11 → 4 → 12	6.65
11 → 2 → 4 → 12	6.89
11 → 12	7.32
11 → 7 → 9 → 12	11.23
11 → 4 → 9 → 12	13.51
11 → 2 → 4 → 9 → 12	13.75

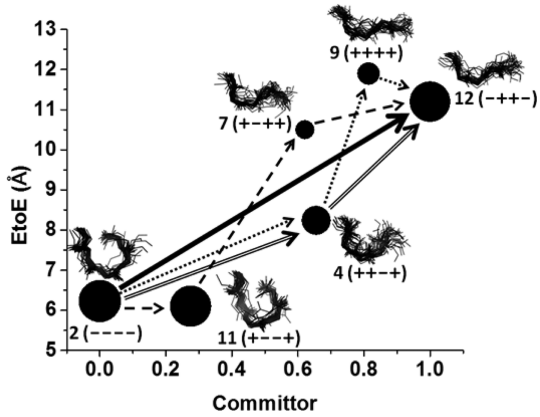


Figure 6. Plot of committor versus end-to-end (EtoE) distance for the 2−12 DIHED pathway. The first four pathways are indicated: P1, solid line; P2, double line; P3, dashed line; P4, dotted line. The sizes of the circles indicate the state populations (see Table 4). Also displayed are ensembles of backbone structures for the various states indicating the broad yet distinct conformations sampled.

fluctuation (MSF) is 4.45. The first three modes, with relative eigenvalues 0.365, 0.240, and 0.073, respectively, account for ∼2/3 of the total MSF. Most importantly, these three main modes are completely dominated by the fluctuations in the same four dihedral angles, Psi_G2, Psi_G3, Psi_F4, and Phi_G3. Table 7 lists the participations—the fractional contributions of each dihedral angle to a particular mode—of these four dihedral angles in the first three modes.

In each mode, these dihedrals essentially exhaust the mode's MSF. Note that mode 1 corresponds exclusively to the Psi_G2/Phi_G3 compensation that occurs repeatedly in the various low cost pathways. The other two modes involve the two other angles, Psi_G3 and Psi_F4, that are required for all the major pathways. Thus, dihedral space PCA on the MD trajectory succeeds in finding the dihedrals that are involved in the MetEnk conformational transitions, in excellent agreement with the direct analysis of the MD trajectory and with DIHED clustering.

3.6. RMSD Clustering Picks out the Same Four Dominant Dihedrals but Produces Some Different Pathways. While clustering in dihedral space is indicated for peptide conformations, it still is of interest to investigate whether clustering in RMSD produces similar results. In contrast to internal coordinates, clustering in RMSD does suffer from having to first best fit the snapshots in Cartesian coordinates that is an especially serious issue for flexible peptides. Using a 2.0 Å cutoff for the cluster algorithm (see section 2.4) leads to 15 RMSD-based states. From the end-to-end distances and the state populations of the seven highest population states listed in Table 8, states 6 and 12 are closed states and state 14 is the highest-population open state.

That the RMSD clustering picks up two closed states indicates that it is capable of finding the psi/phi compensated conformers. The high population states can, once again, be characterized by the same four dihedral bistable angles as in the DIHED clustering. However, Table 8 also shows that the RMSD clustering does not provide unique XXXX states. In particular, states 4 and 8 are the same in terms of dihedral designation but differ in their end-to-end distances, while states 11 and 7 could be lumped together to form a coarser-grained representation.

These seven states are involved in the major pathways. For the pathways with source state 6 (closed, with conformation + − − +) and destination state 14 (open), Table 9 lists the first seven pathways and their costs. The corresponding transitions of the four dihedral angles are given in Table 10, and a concordance between these RMSD clustered pathways and those based on DIHED clustering is given in Table 11. The comparison of the RMSD results is made with the 11−12

Table 6. Dihedral Transitions in the First Five State 11 (closed) to State 12 (open) Pathways with — Denoting No Transition and ⇒ Denoting Transition of a Particular Angle in the Indicated Transition

dihedral	P1(11 → 2 → 12)	P2(11 → 7 → 12)	P3(11 → 4 → 12)	P4(11 → 2 → 4 → 12)	P5(11 → 12)
Psi_G2	+ ⇒ − — −	+ — + ⇒ −	+ — + ⇒ −	+ ⇒ − ⇒ + ⇒ −	+ ⇒ −
Psi_G3	− — − ⇒ +	− ⇒ + — +	− — − ⇒ +	− — − — − ⇒ +	− ⇒ +
Psi_F4	− — − ⇒ +	− — − ⇒ +	− ⇒ + — +	− — − ⇒ + — +	− ⇒ +
Phi_G3	+ ⇒ − — −	+ — + ⇒ −	+ — + ⇒ −	+ ⇒ − ⇒ + ⇒ −	+ ⇒ −
pathway cost	2.95	5.38	6.65	6.89	7.32

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

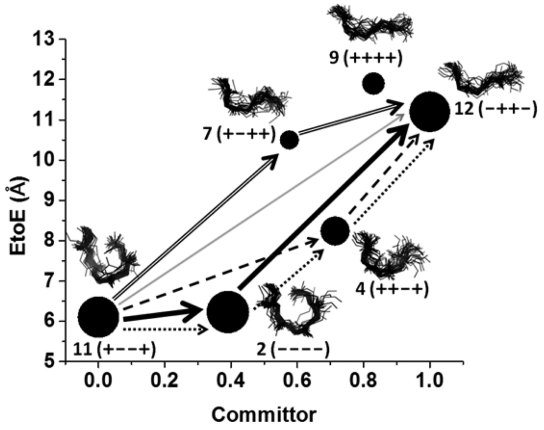


Figure 7. Plot of committor versus end-to-end (EtoE) distance for the 11−12 DIHED pathway. The first five pathways are indicated: P1, solid line; P2, double line; P3, dashed line; P4, dotted line; P5, solid, light line. The sizes of the circles indicate the relative state populations (see Table 4). Also displayed are ensembles of backbone structures for the various states indicating the broad yet distinct conformations sampled.

Table 7. The (Fractional) Mode Participations for the First Three Dihedral PCA Modes

mode	Psi_G2	Psi_G3	Psi_F4	Phi_G3
1	0.475	0.042	0.010	0.447
2	0.028	0.389	0.514	0.021
3	0.028	0.513	0.400	0.004

Table 8. RMSD Clustering Based Committors, Populations, End-to-End Distance (EtoE) Averages/Standard Deviations, and Dihedral Conformations in the Four Dihedral Angle Space for States of the Seven Lowest Cost Pathways

state	committor	population	EtoE ave/std dev	dihed[a]
6	0	0.205	5.96/0.582	+ − − +
12	0.211	0.228	5.97/0.659	− − − −
11	0.682	0.044	8.81/1.09	+ − + +[b]
7	0.685	0.048	8.40/1.25	+ − + +[b]
4	0.810	0.048	9.90/1.04	+ + + +
8	0.870	0.061	12.2/0.80	+ + + +[b]
14	1	0.157	11.1/0.97	− + + −

[a]XXXX denotes, respectively, dihedrals: Psi_G2, Psi_G3, Psi_F4, and Phi_G3. [b]Phi_G3 is found somewhat in two conformations.

Table 9. First Seven Lowest Cost Pathways Obtained from RMSD Clustering

pathway	pathway cost
6 → 4 → 14	4.50
6 → 11 → 14	5.73
6 → 12 → 4 → 14	8.06
6 → 4 → 8 → 14	8.51
6 → 7 → 14	11.55
6 → 12 → 11 → 14	11.70
6 → 11 → 4 → 14	12.03

DIHED source and destination states, because the DIHED state 11 also corresponds to the + − − + state.

As Table 11 indicates, there are some pathways that use the same sequences of dihedral transitions, but the ordering is certainly not one to one. Thus, while RMSD clustering is successful in identifying the same four dihedral angles as the only ones participating in the major transitions from closed to open conformers, the specific pathways followed by these transitions are not identical with those from DIHED clustering.

Table 10. RMSD Clustering Based Pathways of the First Seven Lowest Cost Paths in the Four Dihedral Space

path	pathway
P1	+ − − + → + + + + → − + + −
P2	+ − − + → + − + +[a] → − + + −
P3	+ − − + → − − − − → + + + + → − + + −
P4	+ − − + → + + + + → + + + +[a] → − + + −
P5	+ − − + → + − + +[a] → − + + −
P6	+ − − + → − − − − → + − + +[a] → − + + −
P7	+ − − + → + − + +[a] → + + + + → − + + −

[a]Phi_G3 is found somewhat in two conformations.

Table 11. Comparison of RMSD 6−14 and DIHED 11−12 Cluster-Based Pathways

RMSD clustering	DIHED clustering
P2	P3
P5	P3
P6	P4
P7	P7

4. CONCLUDING REMARKS

In this work, we have investigated the conformational space that the backbone of MetEnk samples based on a long MD trajectory aided by a transition path theory analysis of a Markov State Model constructed from the trajectory. As sensible for a peptide, the dihedral psi and phi angles are a useful set of internal coordinates that have the virtue of not suffering from trajectory best fit imprecision, as do Cartesian coordinates. Dihedral angle changes can lead to large changes in Cartesian coordinates, with the exception of the psi(i)−phi(i + 1) compensation mechanism, as is clearly important to MetEnk.

Three views of dihedral-based conformations in MetEnk were investigated via the following: (1) histograms of the dihedral angles in the MD trajectory, (2) a dihedral-space clustering algorithm applied to the MD trajectory, and (3) modes obtained by PCA. All three methods come to the same conclusion regarding which dihedral angles are involved in the conformational exploration. In particular, there are four dihedral angles, Psi_G2, Psi_G3, Psi_F4, and Phi_G3, that are bistable and are therefore responsible for the conformational space of MetEnk. The XXXX (X = −, +) conformations that are defined from the values that these dihedrals take on map in a one-to-one manner onto the 14 DIHED states, those obtained from dihedral space clustering. There is a strong population ordering (Table 1) of these DIHED states that agrees well with the population ordering found by histograming the MD trajectory. The population ordering shows that, of the 2^4 conformations that could be formed from the XXXX (X = −, +) individual dihedral angles, only a limited set are present in substantial population, indicating a correlation among the values taken on by the four dihedral angles in XXXX.

As found here from the MD trajectory, and in previous work,[51] MetEnk has two closed, high-population zwitterionic forms (small end-to-end distance) that are connected by a Psi_G2/Phi_G3 compensation mechanism, and an open form (large end-to-end distance) of high population. The TPT

dx.doi.org/10.1021/jp412130d | J. Phys. Chem. B 2014, 118, 2883−2895

analysis based on committors and fluxes obtained from the DIHED clustered states permits construction of pathways between these closed and open states. By introducing a cost-of-pathway method based on consecutive reaction network reasoning, a set of pathways in increasing order of cost is produced. Among the 14 DIHED states, there are only four intermediate states that are involved in the major pathways in terms of increasing pathway costs, spanning either of the closed states to a high-population open state. The four states that appear in the major pathways do so from a combination of their population and state-to-state flux values as weighted in a complex fashion in the network of connected states.

Figures 6 and 7 summarize the major pathways in a manner that highlights the correlation between the committor values, which can be viewed as a progress variable spanning a source (one of the closed states) and destination (the open state), and the end-to-end distance. The end-to-end distance is a reasonable reaction coordinate indicating the progression through intermediates spanning the closed to open states. There is an interesting distinction between the pathways starting from the closed source state with conformation − − − − versus those starting from + − − +, these states differing by their Psi_G2/Phi_G3 compensating dihedral angles. In the former, the least cost pathway is a direct transition (Table 3) from this closed to open conformation, while in the latter (Table 6) the least cost pathway first requires a Psi_G2/Phi_G3 compensating flip from the + − − + to − − − − conformation, and then the other two dihedrals flip.

The cost ordering of the major pathways shown in Figures 6 and 7 can be qualitatively understood from the differing roles of Psi_G2/Phi_G3 compensating transitions and the conformational transitions of Psi_G3 and Psi_F4. Compensating transitions should be facile because the minimal Cartesian coordinate motion does not require much solvent displacement. In contrast, the Psi_G3 and/or Psi_F4 dihedral flips lead to large Cartesian coordinate displacements, and contribute to the changes in end-to-end distance. The ordering of the 2−12 pathway costs in Figure 6 that are detailed in Table 3 shows that P1 occurs by a direct path where both Psi_G3 and Psi_F4 flip. P2 goes through one intermediate, state 4, and uses two compensating transitions, and separate Psi_G3 and Psi_F4 flips, that should correspond to a more costly pathway than P1. The two other major pathways involve more intermediates, and are of essentially the same cost. For the 11−12 pathways, shown in Figure 7 and detailed in Table 6, as noted above, in P1 first a (facile) compensating transition occurs followed by both Psi_G3 and Psi_F4 flipping to produce the open state. In P2, Psi_G3 and Psi_F4 flip separately and then also invoke a compensating transition. P3 and P4 use, respectively, one and two intermediates. These three paths are of similar cost. Pathway P5 is direct, no intermediates, but it does require a concerted transition of all four dihedrals, with a modestly higher cost.

Finally, it should be stressed that the superior results obtained by clustering in dihedral versus Cartesian coordinate space as applied here to a peptide do not mean that for other circumstances other coordinate set choices would not perform better. For example, states for proteins that undergo transitions spanning ligand free to ligand bound conformations may be better described by some restricted set of Cartesian coordinates. Deciding which coordinate set to use in a given circumstance is key to providing states that are appropriate to creating a Markov state model.

AUTHOR INFORMATION

Corresponding Author

*E-mail: cukier@chemistry.msu.edu. Phone: 517-355-9715 x 263.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Dirk Colbry of the High Performance Computer Center at Michigan State University for technical support with MSMBuilder and MATLAB.

REFERENCES

(1) Berg, B. A.; Neuhaus, T. Multicanonical Algorithms for 1st Order Phase-Transitions. *Phys. Lett. B* **1991**, *267*, 249−253.

(2) Hansmann, U. H. E. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.* **1997**, *281*, 140−150.

(3) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. New Approach to Monte-Carlo Calculation of the Free-Energy - Method of Expanded Ensembles. *J. Chem. Phys.* **1992**, *96*, 1776−1783.

(4) Lyubartsev, A.; Laaksonen, A. Parallel Molecular Dynamics Simulations of Biomolecular Systems. *Lect. Notes Comput. Sci.* **1998**, *1541*, 296−303.

(5) Swendsen, R. H.; Wang, J. S. Replica Monte-Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.

(6) Geyer, C. J. Markov Chain Monte Carlo Maximum Likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*; Keramidas, E. M., Ed.; Interface Foundation: Fairfax Station, VA, 1991.

(7) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(8) Wang, J. S.; Swendsen, R. H. Replica Monte Carlo Simulation (Revisited). *Prog. Theor. Phys. Suppl.* **2005**, 317−323.

(9) Hukushima, K.; Nemoto, K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.

(10) Lei, H.; Duan, Y. Improved Sampling Methods for Molecular Simulation. *Curr. Opin. Struct. Biol.* **2007**, *17*, 187−191.

(11) Yang, S.; Banavali, N. K.; Roux, B. Mapping the Conformational Transition in Src Activation by Cumulating the Information from Multiple Molecular Dynamics Trajectories. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 3776−3781.

(12) Yang, S.; Roux, B. Src Kinase Conformational Activation: Thermodynamics, Pathways, and Mechanisms. *PLoS Comput. Biol.* **2008**, *4*, e1000047.

(13) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011−19016.

(14) Bowman, G. R.; Huang, X. H.; Pande, V. S. Using Generalized Ensemble Simulations and Markov State Models to Identify Conformational States. *Methods* **2009**, *49*, 197−201.

(15) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know About Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52*, 99−105.

(16) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(17) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(18) Swope, W. C.; Pitera, J. W.; Suits, F. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.

(19) Keller, B.; Hunenberger, P.; van Gunsteren, W. F. An Analysis of the Validity of Markov State Models for Emulating the Dynamics of Classical Molecular Systems and Ensembles. *J. Chem. Theory Comput.* **2011**, *7*, 1032−1044.
(20) Deng, N. J.; Dai, W.; Levy, R. M. How Kinetics within the Unfolded State Affects Protein Folding: An Analysis Based on Markov State Models and an Ultra-Long Md Trajectory. *J. Phys. Chem. B* **2013**, *117*, 12787−12799.
(21) Xia, J. C.; Deng, N. J.; Levy, R. M. Nmr Relaxation in Proteins with Fast Internal Motions and Slow Conformational Exchange: Model-Free Framework and Markov State Simulations. *J. Phys. Chem. B* **2013**, *117*, 6625−6634.
(22) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMbuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(23) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noé, F. Emma: A Software Package for Markov Model Building and Analysis. *J. Chem. Theory Comput.* **2012**, *8*, 2223−2238.
(24) Buchete, N. V.; Hummer, G. Peptide Folding Kinetics from Replica Exchange Molecular Dynamics. *Phys. Rev. E* **2008**, *77*, 030902(R).
(25) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States. *J. Chem. Phys.* **2007**, *126*, 155102.
(26) Singh, G.; Tieleman, D. P. Atomistic Simulations of Wimley-White Pentapeptides: Sampling of Structure and Dynamics in Solution. *J. Chem. Theory Comput.* **2013**, *9*, 1657−1666.
(27) Metzner, P.; Schutte, C.; Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Model. Simul.* **2009**, *7*, 1192−1219.
(28) Noe, F.; Fischer, S. Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.
(29) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder Ntl9(1−39). *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.
(30) Buchete, N. V.; Hummer, G. Coarse Master Equations for Peptide Folding Dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.
(31) Bowman, G. R.; Geissler, P. L. Equilibrium Fluctuations of a Single Folded Protein Reveal a Multitude of Potential Cryptic Allosteric Sites. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 11681−11686.
(32) Sadiq, S. K.; Noe, F.; De Fabritiis, G. Kinetic Characterization of the Critical Step in HIV-1 Protease Maturation. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 20449−20454.
(33) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition Path Sampling: Throwing Ropes over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.
(34) Hummer, G. From Transition Paths to Transition States and Rate Coefficients. *J. Chem. Phys.* **2004**, *120*, 516−523.
(35) DAlagni, M.; Delfini, M.; DiNola, A.; Eisenberg, M.; Paci, M.; Roda, L. G.; Veglia, G. Conformational Study of [Met5]Enkephalin-Arg-Phe in the Presence of Phosphatidylserine Vesicles. *Eur. J. Biochem.* **1996**, *240*, 540−549.
(36) Graham, W. H.; Carter, E. S.; Hicks, R. P. Conformational-Analysis of Met-Enkephalin in Both Aqueous-Solution and in the Presence of Sodium Dodecyl-Sulfate Micelles Using Multidimensional NMR and Molecular Modeling. *Biopolymers* **1992**, *32*, 1755−1764.
(37) Higashijima, T.; Kobayashi, J.; Nagai, U.; Miyazawa, T. NMR-Study on Met-Enkephalin and Met-Enkephalinamide - Molecular Association and Conformation. *Eur. J. Biochem.* **1979**, *97*, 43−57.
(38) Marcotte, I.; Separovic, F.; Auger, M.; Gagne, S. M. A Multidimensional H-1 NMR Investigation of the Conformation of Methionine-Enkephalin in Fast-Tumbling Bicelles. *Biophys. J.* **2004**, *86*, 1587−1600.
(39) Surewicz, W. K.; Mantsch, H. H. Solution and Membrane-Structure of Enkephalins as Studied by Infrared-Spectroscopy. *Biochem. Biophys. Res. Commun.* **1988**, *150*, 245−251.
(40) Takeuchi, H.; Ohtsuka, Y.; Harada, I. Ultraviolet Resonance Raman-Study on the Binding Mode of Enkephalin to Phospholipid-Membranes. *J. Am. Chem. Soc.* **1992**, *114*, 5321−5328.
(41) Hansmann, U. H. E. Protein Folding Simulations in a Deformed Energy Landscape. *Eur. Phys. J. B* **1999**, *12*, 607−611.
(42) Hansmann, U. H. E.; Okamoto, Y. Numerical Comparisons of Three Recently Proposed Algorithms in the Protein Folding Problem. *J. Comput. Chem.* **1997**, *18*, 920−933.
(43) Hansmann, U. H. E.; Okamoto, Y.; Eisenmenger, F. Molecular Dynamics, Langevin and Hybrid Monte Carlo Simulations in a Multicanonical Ensemble. *Chem. Phys. Lett.* **1996**, *259*, 321−330.
(44) vanderSpoel, D.; Berendsen, H. J. C. Molecular Dynamics Simulations of Leu-Enkephalin in Water and Dmso. *Biophys. J.* **1997**, *72*, 2032−2041.
(45) Aburi, M.; Smith, P. E. A Conformational Analysis of Leucine Enkephalin as a Function of pH. *Biopolymers* **2002**, *64*, 177−188.
(46) Nielsen, B. G.; Jensen, M. O.; Bohr, H. G. The Probability Distribution of Side-Chain Conformations in [Leu] and [Met]-Enkephalin Determines the Potency and Selectivity to Mu and Delta Opiate Receptors. *Biopolymers* **2003**, *71*, 577−592.
(47) Shen, M. Y.; Freed, K. F. Long Time Dynamics of Met-Enkephalin: Comparison of Explicit and Implicit Solvent Models. *Biophys. J.* **2002**, *82*, 1791−1808.
(48) Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F. Computer Simulation of Met-Enkephalin Using Explicit Atom and United Atom Potentials: Similarities, Differences, and Suggestions for Improvement. *J. Phys. Chem. B* **2003**, *107*, 1685−1691.
(49) Karvounis, G.; Nerukh, D.; Glen, R. C. Water Network Dynamics at the Critical Moment of a Peptide's Beta-Turn Formation: A Molecular Dynamics Study. *J. Chem. Phys.* **2004**, *121*, 4925−4935.
(50) Sanbonmatsu, K. Y.; Garcia, A. E. Structure of Met-Enkephalin in Explicit Aqueous Solution Using Replica Exchange Molecular Dynamics. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 225−234.
(51) Su, L.; Cukier, R. I. Hamiltonian and Distance Replica Exchange Method Studies of Met-Enkephalin. *J. Phys. Chem. B* **2007**, *111*, 12310−12321.
(52) Cukier, R. I. Ferreting out Correlations from Trajectory Data. *J. Chem. Phys.* **2011**, *135*, 225103.
(53) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer Science: New York, 2004.
(54) Yen, J. Y. Finding the K Shortest Loopless Paths in a Network. *Manage. Sci.* **1971**, *17*, 712−716.
(55) Lou, H. F.; Cukier, R. I. Molecular Dynamics of Apo-Adenylate Kinase: A Distance Replica Exchange Method for the Free Energy of Conformational Fluctuations. *J. Phys. Chem. B* **2006**, *110*, 24121−24137.
(56) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P. *Biomolecular Simulation: The Gromos96 Manual and User Guide*; Vdf hochschulverlag AG an der ETH: Zürich, Switzerland, 1996.
(57) Berendsen, H. H. C.; Postma, J. P. M.; Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(58) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(59) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G.; Smooth, A. Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(60) Lou, H.; Cukier, R. I. *Analyzer*, 2.0; East Lansing, MI, 2008.
(61) García, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696−2699.
(62) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412−425.
(63) Murphy, R. D.; Conlon, J.; Mansoor, T.; Luca, S.; Vaiana, S. M.; Buchete, N.-V. Conformational Dynamics of Human IAPP Monomers. *Biophys. Chem.* **2012**, *167*, 1−7.

(64) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations. *J. Chem. Phys.* **2007**, *126*, 244111.
(65) Mu, Y. G.; Nguyen, P. H.; Stock, G. Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 45−52.
(66) Gittins, R. *Canonical Analysis: A Review with Applications in Ecology*; Springer-Verlag: Berlin, 1984.
(67) Risken, H. *The Fokker-Planck Equation: Methods of Solution and Applications*; Springer-Verlag: Berlin, 1984.
(68) van Kampen, N. *Stochastic Processes in Physics and Chemistry*; Elsevier: Amsterdam, The Netherlands, 1981.
(69) Cox, D. R.; Miller, H. D. *The Theory of Stochastic Processes*; John Wiley & Sons Inc: New York, 1965.
(70) Cukier, R. I. Variance of a Potential of Mean Force Obtained Using the Weighted Histogram Analysis Method. *J. Phys. Chem. B* **2013**, *117*, 14785−14796.
(71) E, W.; Vanden-Eijnden, E. Towards a Theory of Transition Paths. *J. Stat. Phys.* **2006**, *123*, 503−523.
(72) Berezhkovskii, A.; Hummer, G.; Szabo, A. Reactive Flux and Folding Pathways in Network Models of Coarse-Grained Protein Dynamics. *J. Chem. Phys.* **2009**, *130*, 205102.
(73) Weiss, M. A. *Data Structures and Algorithm Analysis*, 2nd ed.; Benjamin/Cummings: Redwood City, CA, 1995.
(74) Shirazipour, M. http://www.mathworks.com/matlabcentral/fileexchange/32513-k-shortest-path-yens-algorithm, 2011.
(75) Korn, A. P.; Rose, D. R. Torsion Angle Differences as a Means of Pinpointing Local Polypeptide Chain Trajectory Changes for Identical Proteins in Different Conformational States. *Protein Eng.* **1994**, *7*, 961−967.