JCTC Journal of Chemical Theory and Computation

Article

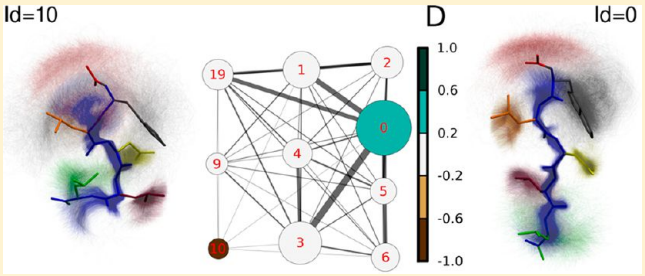
pubs.acs.org/JCTC

Atomistic Simulations of Wimley–White Pentapeptides: Sampling of Structure and Dynamics in Solution

Gurpreet Singh and D. Peter Tieleman*

Department of Biological Sciences and Institute for Biocomplexity and Informatics, University of Calgary, 2500 University Drive N.W., Calgary, Alberta T2N1N4, Canada

ABSTRACT: Wimley–White pentapeptides (Ac-WLXLL) can be used as a model system to study lipid–protein interactions as they bind to lipid/water interfaces, like many antimicrobial peptides, and thermodynamic experimental data on their interactions with lipids are available, making them useful for both force field and method testing and development. Here we present a detailed simulation study of Wimley–White (WW) peptides in bulk water to investigate sampling, conformations, and differences due to the different X residue with an eye to future simulations at the lipid/water interface where sampling problems so far have hindered free energy calculations to reproduce the experimental thermodynamic data. We investigate the conformational preferences and slowest relaxation time of WW peptides in bulk water by building Markov State Models (MSM) from Molecular Dynamics (MD) simulation data. We show that clustering based on binning of backbone ϕ, ψ dihedrals in combination with the community detection algorithm of Blondel et al. provides a quick way of building MSM from large data sets. Our results show that in some cases, implied times even in these small peptides range from 224 to 547 ns. The implications of these slow transitions on determining the potential of mean force profiles of peptide interactions with a lipid bilayer are discussed.



INTRODUCTION

Computer simulations are increasingly used as a tool to understand the mechanism of function of biomolecules at the atomistic level.[1,2] As computing power increases, computer simulations such as Molecular Dynamics (MD) simulations can be applied to study more and more complex biomolecular systems.[3−5] In classical simulations of biomolecules the two major sources of errors are inaccuracies in the force fields and inadequate sampling of the relevant degrees of freedom.[6−9] For this reason, relatively simple systems such as alanine dipeptide[10−13] or polypeptides[14,15] and well-studied small soluble proteins such as Trp-Cage miniprotein,[16−19] villin headpiece,[20,21] or other fast folding proteins[22] are essential tools in testing new sampling techniques and force field descriptors. One of the methods to test the accuracy of protein force fields is based on the hydration free energies of amino acid side chain analogs compared with available experimental data.[7,23]

However, thermodynamic data on lipid–protein interactions are relatively rare and reproducing these data in simulations technically challenging. A lipid bilayer can be seen as a heterogeneous environment with large gradients in density, polarity, and other physicochemical properties along the bilayer normal. This presents a difficulty for obtaining high spatial resolution thermodynamic data.

Some thermodynamic data on lipid–protein interactions are available in the form of hydrophobicity scales.[24] The Wimley–White (WW) hydrophobicity scale is one such scale, developed by measuring the partitioning of pentapeptides of sequence Ac-WLXLL, where X can be any of the twenty amino acids, at the 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC)/water interface.[25] Computer simulations performed on WW peptides include atomistic simulations at the cyclohexane/water, octanol/water, and DOPC/water interfaces.[26,27] The difficultly of obtaining converged sampling of WW peptides at interfaces was noted in these simulation studies. Due to these difficulties, only the partitioning free energies of amino acid side chain analogues have been compared to the WW hydrophobicity scale.[28,29] Only recently, using the coarse-grained MARTINI force field, we were able to compute the partitioning free energies of these peptides at POPC/water interface and directly compare it with the WW scale.[30] This work played an important role in testing and improving a new generation of MARTINI parameters,[31] but calculating, let alone reproducing, peptide-based hydrophobicity scales with atomistic force fields appears currently just out of reach. However, computational access to this hydrophobicity scale would be a powerful test of current force fields and a useful tool in improving force fields. In addition, the potential of mean force (PMF) profile of binding of the WW peptides to a POPC bilayer will be helpful in understanding the molecular mechanism of peptide adsorption at the bilayer interface, and, once validated, similar approaches could be used to study a wide range of peptides for which less experimental data are available.

Received: September 20, 2012
Published: February 19, 2013

dx.doi.org/10.1021/ct3008217 | J. Chem. Theory Comput. 2013, 9, 1657−1666

Free energy calculations are generally computationally intensive, and obtaining reliable free energy estimates of complex biomolecular systems in a reasonable amount of time is a challenging task.[32] Umbrella Sampling (US) is one of the most commonly used methods to calculate the PMF profile along a reaction coordinate.[33] Generally, in membrane systems the reaction coordinate is the vector orthogonal to the interface.[28,34,35] The PMF profile of a molecule is obtained by constraining it using a biasing potential (often harmonic) at various positions along this reaction coordinate. In order to obtain a reliable PMF, the degrees of freedom that are orthogonal to this reaction coordinate must be sampled sufficiently, i.e., the simulation length should be several orders of magnitude longer than the slowest relaxation time of the processes that are orthogonal to the reaction coordinate. However, the slowest degrees of freedom are generally not known a priori and, if not adequately sampled during the simulation, lead to errors in determination of PMF profiles.[36,37] Recently, it has been shown that long time dynamics of biomolecules can be represented by Markov State Models that were built using short MD simulations.[38,39] The approach does not require predefined reaction coordinates; instead the method relies on partitioning of the conformational state of the biomolecules into several small states that are then used to construct a Markov model.[40−43]

The main goal of this study is to investigate conformational preferences of the WW peptides in bulk water and attempt to identify the slowest transitions among the observed conformational states. We carry out multiple simulations of the peptides in bulk water using MD and Temperature Replica Exchange Molecular Dynamic (TREMD) simulations. We calculate end-to-end distance autocorrelations from multiple MD simulations of same peptide to show that single long simulations (600 ns) of the pentapeptides are not sufficient to represent the dynamics of the peptide. We calculate end-to-end distribution functions to show that the average behavior of these peptides can vary depending on amino acid substitution, even in these relatively simple peptides. In order to identify the slowest transitions, we represent the conformational transitions of the WW peptides using Markov state models. We judge the convergence of conformational sampling from MD simulations by comparing the peptide conformational populations with those obtained using temperature replica-exchange simulations and discuss the implication of such slow transition on computation of PMF profiles using umbrella sampling techniques and MD simulations.

Our approach is useful in identifying the lower bound of the MD simulation length that would be needed to get a reasonable estimate of conformational preferences or any other property that is dependent on peptide conformations in solution. Furthermore, a comparison between the WW peptides in bulk water reveals interesting trends among amino acids in terms of their effect on conformation preferences and transition networks. We expect to apply this to interfacial simulations in the future, where this information will also be used to identify additional reaction coordinates to enhance the sampling.

METHODS

Simulations. Peptides of sequence WLXLL, where X can be any of the 20 amino acids, were simulated in bulk water. All the atomistic simulations were performed using the OPLS force-field and SPCE water model.[44,45] The N-terminus of all peptides was capped with an acetyl group, while the C-terminus was deprotonated. The system was kept neutral by addition of potassium and chloride ions. All the simulations were performed with the Gromacs (version 4.0 and above) software.[46] A leapfrog integrator was used in the simulations.

Equilibrium MD Simulations. A single peptide was centered in a cubic box of length 4 nm and was solvated with pre-equilibrated SPCE water. For each system, 18 trajectories were generated. All trajectories were initiated from the same peptide configuration. For each particle, velocities were randomly assigned from a Maxwell−Boltzmann distribution at a temperature of 300 K. The system was equilibrated for 2 ns using Berendsen temperature and pressure coupling,[47] followed by a production run of 600 ns. The production runs were carried out using the stochastic velocity rescaling thermostat of Bussi et al.[48] as implemented in Gromacs, and the Parrinello-Rahman[49] barostat, with coupling constants of 1.0 and 2.5 ps, respectively. All systems were simulated at a temperature of 300 K and a pressure of 1 bar using periodic boundary conditions. The coordinates were saved every 2 ps. The initial 50 ns of each production run were omitted from the conformational analysis.

Temperature Replica Exchange Molecular Dynamic (TREMD) Simulations. In the TREMD method, multiple replicas of the system are simulated at different temperatures, and state−exchange moves are periodically attempted so that the two neighboring replicas exchange their thermodynamic state.[50−54] TREMD simulations can improve the sampling of the conformational space of biomolecular systems, because of their ability to overcome the kinetic trapping due to increased thermal energies at higher temperatures.[55,56] An extension of replica exchange for the isothermic-isobaric ensemble was proposed by Okabe et al.,[57] where the exchange probability between two replicas m and n is given as

$$P(m \leftrightarrow n) = min[1, \exp(-\Delta)] \tag{1}$$

$$\Delta = (\beta_m - \beta_n)(U(\boldsymbol{x}_n) - U(\boldsymbol{x}_m)) + (\beta_m P_m - \beta_n P_n)(V^n - V^m) \tag{2}$$

Here, β_m, $U(\boldsymbol{x}_m)$, P_m, and V_m are the inverse temperature, potential energy, pressure, and volume of the mth replica, respectively.

TREMD simulations were performed on systems containing peptides with X=D, F, G, L, R, S, and V. A stochastic velocity rescaling thermostat and Parrinello-Rahman barostat were used. The states between the two replicas were exchanged according to the standard Gromacs exchange scheme. In this scheme, the exchanges are attempted only between neighboring temperatures, and alternately between even and odd pairs. Twenty-nine temperatures, ranging from 300 to 400 K, were initially selected. Several short simulations of 0.4 ns duration were carried out, during which the temperature spacing was iteratively varied until the observed acceptance probability between adjacent replicas was within 0.2−0.3. Production runs of 300 ns were then initiated at temperatures 300.0, 303.3, 306.4, 309.4, 312.6, 315.9, 318.9, 322.2, 325.4 328.7, 332.1, 335.4, 338.8, 342.3, 345.8, 349.2, 352.9, 356.3, 360.2, 363.8, 367.4, 371.3, 375.0, 378.8, 382.8, 386.7, 390.6, 394.6, and 398.7 K, and exchanges were attempted every 1 ps, resulting in a total simulation time of 8.7 μs per peptide.

Analyses. *Assigning Conformational States.* Clustering of protein and peptide conformations is generally performed using either RMSD matrices or based on backbone dihedral angles.[58−61] In case of RMSD based methods, two

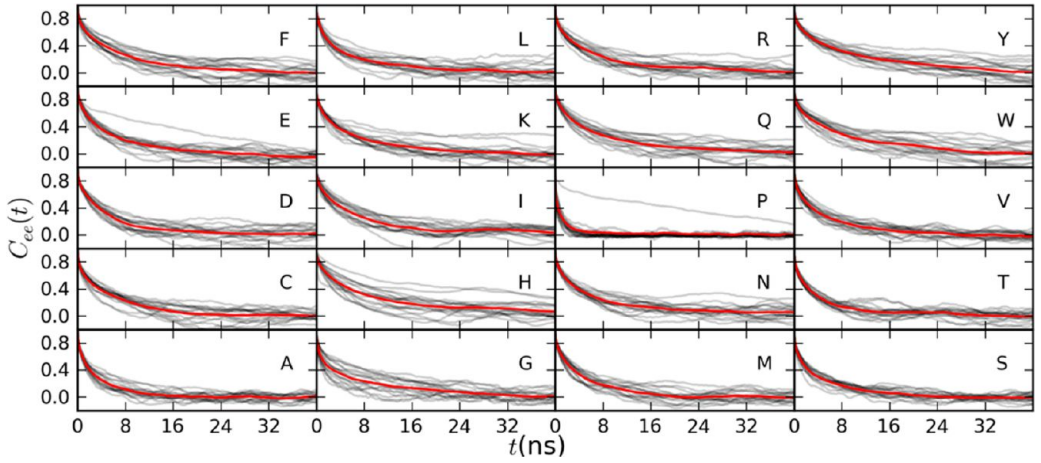


Figure 1. End-to-end distance autocorrelation $C_{ee}(t)$ for peptides in bulk water for each of the twenty peptides. The amino acid substituted at the third position in each pentapeptide is shown in the panel. The $C_{ee}(t)$ from the individual simulations are colored gray, and the average $C_{ee}(t)$ are colored red.

conformations are considered to belong to the same group if the RMSD between them is below an arbitrarily defined cutoff. In case of clustering based on backbone dihedral angles, the number of possible states is x^N, where x is the number of bins, and N is the number of ϕ and ψ dihedral angle pairs. The number of possible states rapidly becomes prohibitively large for proteins; therefore, this method is generally employed only on small peptides. The main advantage of this approach is that there is no need to build computationally expensive RMSD matrices and to employ clustering algorithms.

In both cases, the conformations are clustered according to a given geometric criterion. However, the trajectories generated by MD simulation algorithms also contain information about transition probabilities between conformations. This information can be used to further enhance the clustering by grouping together those conformations that have high transition probabilities. Such kinetic clustering has been employed in simulation studies of Fs peptide, λ repressor protein, β harpin peptide, NTL9 and PinWW domain.[40,58,62−65]

In this study, we analyzed the peptide conformations using a similar protocol as implemented in the MSMbuilder package.[41,43] However, instead of using RMSDs, we clustered the peptide conformations based on the ϕ and ψ dihedral angles of the first four amino acid residues (WLXL). We divided the Ramachandran space into 36 equal sized squares. Each conformation in the trajectory was uniquely assigned to one of the 1679616 possible states, after which the interstate transition probability matrix was computed. Finally, using the community detection algorithm developed by Blondel et al.,[66] we partition the clusters further into communities. For this purpose, we represented the data as a graph, where the nodes correspond to the clusters obtained using the backbone dihedrals as clustering criteria, while the edges represent the number of transitions between the clusters. Two nodes are connected by an edge only if the number of transitions between those clusters is greater than zero.

A community detection algorithms attempts to identify groups of densely connected nodes. In a weighted network, community structure can be determined by maximizing the modularity Q of the partition, which is defined as

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i}{k_j}\right)\delta(c_i, c_j) \tag{3}$$

where $m = (1/2)\sum_{ij}A_{ij}$ is the number of edges in the graph; A_{ij} is the weight of the edge between i and j; k_i is the degree of vertex i that in a weighted network is computed as $k_i = \sum_j A_{ij}$; c_i is the community to which vertex i is assigned; the δ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise.[67] The Q is zero if the resulting partition has the same number of edges within the community as expected by random chance. Although the exact maximization of modularity is computationally hard, several algorithms have been developed to provide approximate solutions.[68] We used the modularity partition algorithm developed by Blondel et al., which is able to find high modularity partitions on large networks in relatively short times.[66] The algorithm provides decomposition of networks into communities for different level of organizations. In this study, we only consider the partitions obtained at the top level of organization.

End-to-End Distance. The end-to-end distance was defined as the distance between the first and the last C-α atom. The end-to-end distance autocorrelations $C_{ee}(t)$ were computed using Gromacs tools.

RESULTS

End-to-End Distances. The internal dynamics of short unstructured peptides can be probed by calculating end-to-end distance autocorrelations $C_{ee}(t)$, which are shown in Figure 1. All the peptides, except X=P, show a slow decay in $C_{ee}(t)$, with an average $C_{ee}(t)$ decaying to zero in a range of t between 30 and 50 ns. A high variability in correlation decay is seen among different simulations of the same peptide. For most of the peptides, two main peaks are seen in the probability densities of end-to-end distance $P(r)$. Based on the magnitude of these peaks, the $P(r)$ can be roughly divided into three categories, as shown in Figure 2 (top three panels). The majority of the peptides belong to the first category, where the peak at ∼1.24 nm (peak I) has a higher magnitude than the broad peak at ∼1.0 nm (peak II). The second category consists of peptides where X is S, T, or V. The two peaks have similar magnitude for these peptides. The third category consists of peptides (X is F, Y, and W) where peak II is higher in magnitude than peak I.

dx.doi.org/10.1021/ct3008217 | J. Chem. Theory Comput. 2013, 9, 1657−1666

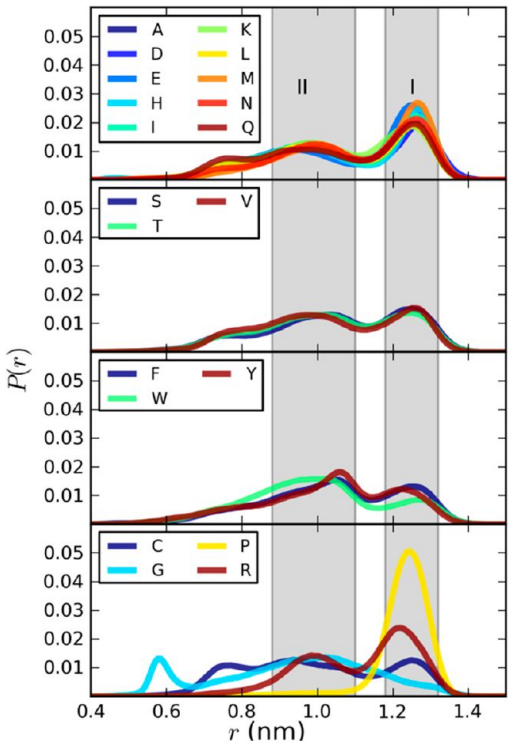


Figure 2. The probability densities of end-to-end distance $P(r)$ for various peptides. The peaks at ~1.24 and ~1.0 nm are labeled as peak 1 and peak II, respectively. The regions corresponding to the peaks are shaded to guide the eye. The legends are labeled according to the amino acid substitution at the third position in each pentapeptide.

The peptide with X=P shows only one prominent peak at ~1.24 nm, whereas X=G has a prominent peak at 0.59 nm, and no peak at 1.24 nm. The $P(r)$ of X=C is relatively flat, while for X=R the peak I is shifted to 1.21 nm.

Conformational States and Transitions. The division of the Ramachandran space into 36 equal sized squares for the four dihedrals results in a maximum of 36^4 clusters. The number of observed clusters in the MD simulations is in the range of 0.5% (X=P) to 2.7% (X=G) of the total possible clusters. The application of the community detection algorithm on the resulting graph partitions the nodes into communities. The number of communities at the top level of hierarchy ranged between 10 (X=P) to 23 (X=V or W). The top ten communities encompass 78% (X=G) to 100% (X=P) of the total observed population. The percentages of the total population in the top ten communities are shown in Figure 3. The largest community generally accounts for 16−36% of the total population, and the top five communities account for over 60% of the total conformations observed (except for X=G). For each community, a random subset of conformations was chosen and used to analyze the backbone dihedral propensity. Figure 5 shows for each community the percentage of conformations with ϕ, ψ dihedral pairs within a given region of Ramachandran plot. For the peptides with a dominant peak at $P(r) \sim 1.24$ nm, the most populated communities consist mainly of conformations where all the observed dihedral angles are either in polyproline II or strand regions.

In order to identify the slowest transitions between the clusters, we modeled the kinetics of peptides using Markov state models. For the construction of the Markov state model, the count matrix $\mathbf{C}(\tau)$ was constructed from MD simulation data. The lag time τ corresponds to the evolution of the system by a certain number of time steps. In order to build $\mathbf{C}(\tau)$, all the trajectory frames were uniquely assigned to a particular community, and matrix element $\mathbf{C}_{i,j}(\tau)$ was computed as

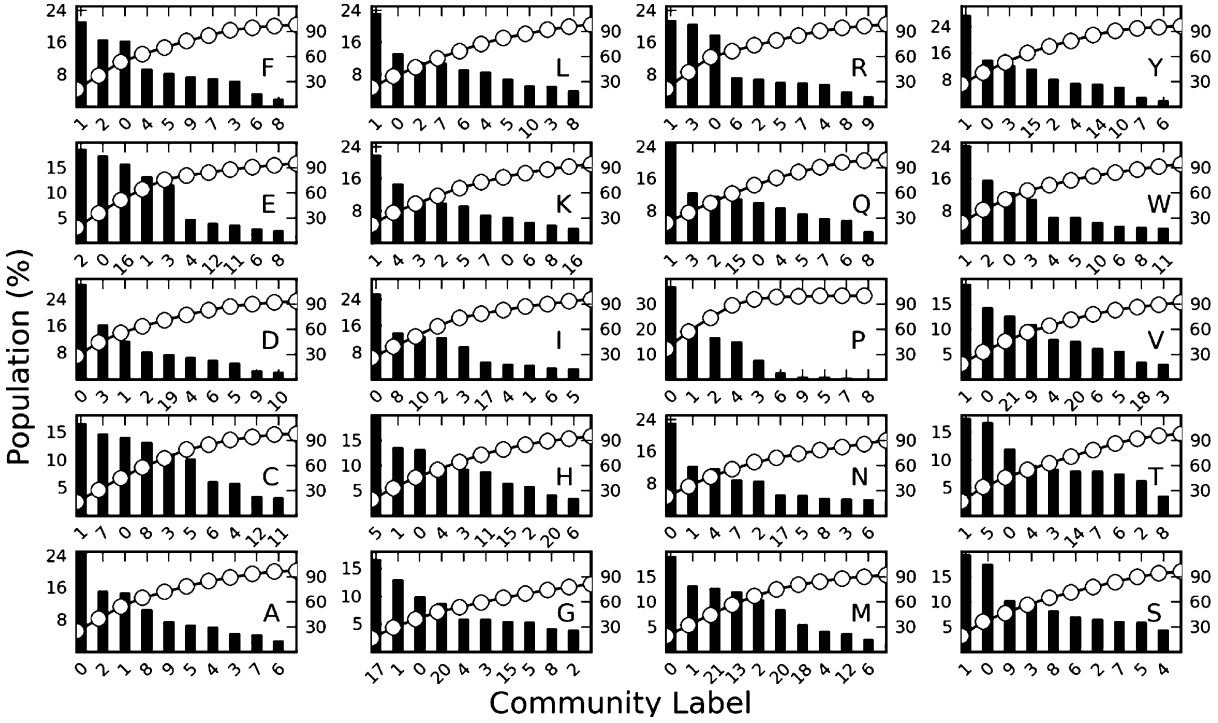


Figure 3. Community populations in the top ten communities for the twenty peptides. The line with circles shows the cumulative population. For each panel, the letter label indicates the amino acid substitution. The left Y axis gives the percent population and the right Y axis gives the cumulative population, and the X axis labels correspond to the label by which each community is identified (community ids).

dx.doi.org/10.1021/ct3008217 | J. Chem. Theory Comput. 2013, 9, 1657−1666

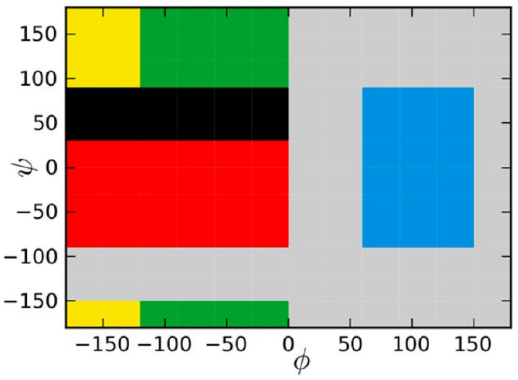


Figure 4. Ramachandran plot showing the regions that were used for color coding dihedral angle combinations. The yellow, green, red, and cyan colors correspond to the polyproline II, sheet, helix, and turn regions of the Ramachandran plot, respectively.

$$\mathbf{C}_{j,i}(\tau) = \sum_{t=1}^{N-\tau} \chi_i(\boldsymbol{x}_t)\chi_j(\boldsymbol{x}_{t+\tau}) \tag{4}$$

where $\chi_i(\boldsymbol{x})$ is defined as a membership function that gives the probability of configuration $\boldsymbol{x}$ to belong to community i such that $\sum_{i=1}^{n}\chi_i(\boldsymbol{x}) = 1$, here n is the total number of communities (see refs 43 and 42). Due to finite sampling, the $\mathbf{C}(\tau)$ is generally not symmetric. A symmetric transition metric $\hat{\mathbf{T}}(\tau)$ is estimated by methods involving symmetrizing the matrix by addition of its transpose,[69] using maximum likelihood estimator[42,43] or a method based on Bayesian inferences.[70] The data in this study are gathered from several long simulations; it can be assumed that the populations of the conformational states obtained from MD simulations data will be close to an equilibrium population. For data sets that are close to equilibrium, it has been shown that all the above-mentioned methods lead to similar results.[69] We symmetrized the count matrix by adding its transpose $\hat{\mathbf{C}}(\tau) = (\mathbf{C}(\tau) + \mathbf{C}^{\mathrm{T}}(\tau))/2$ and calculated the transition matrix elements $\mathbf{T}_{i,j}(\tau) = \hat{\mathbf{C}}_{i,j}(\tau)/\sum_{i=1}^{n}\hat{\mathbf{C}}_{i,j}(\tau)$. The eigenvalues $\mu_k(\tau)$ and eigenvectors $\boldsymbol{u}_k(\tau)$ of the transition matrix $\boldsymbol{T}(\tau)$ were used to identify the implied time scale $\tau_k = -\tau/[\ln \mu_k(\tau)]$ and the associated aggregate transitions, respectively.[58,71] The implied times are the exponential decay constants of the various relaxation processes in the system. The components of the eigenvectors provide information about the states undergoing exchanges

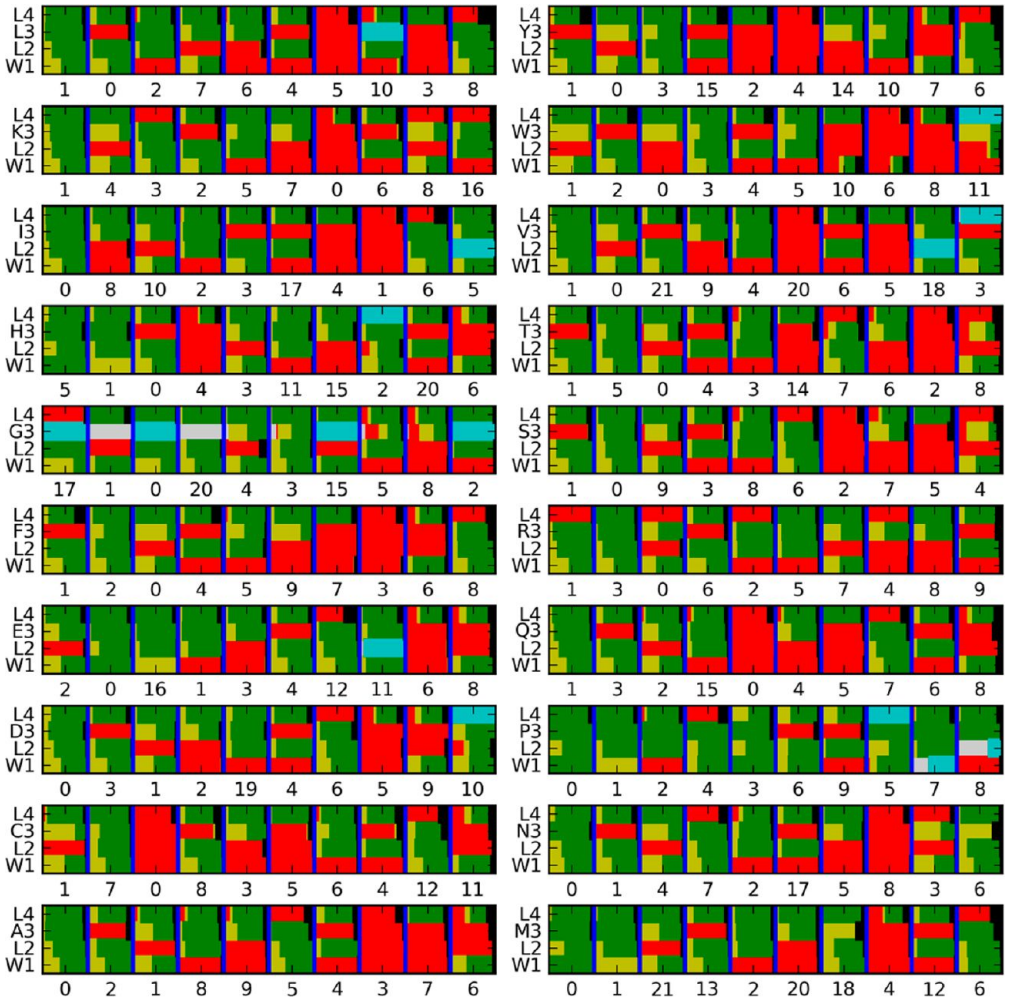


Figure 5. The percent populations of the backbone dihedrals belonging to various regions of Ramachandran plot are shown for the top ten communities. Twenty panels corresponding to twenty peptides are shown. In each panel, the community labels are shown on the X axis, the amino acid label are shown on the Y axis, individual communities are separated by blue lines, and each backbone ϕ, ψ combination is colored according to Figure 4. The width of the color is proportional to its percent population.

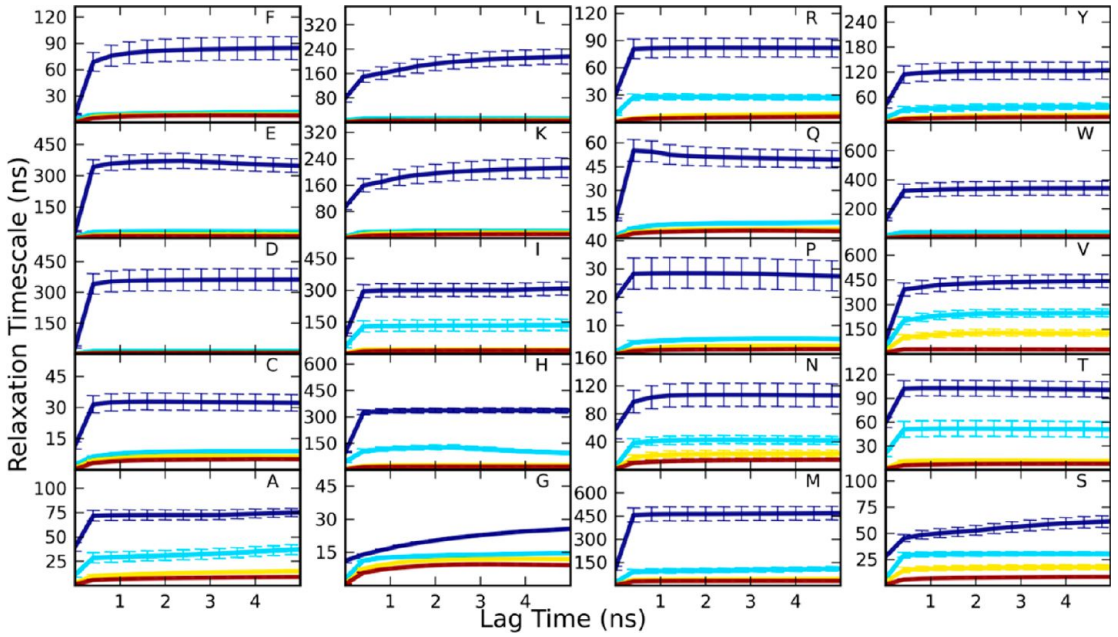


Figure 6. Implied time scales as a function of the lagtime for the top four slowest processes as computed from the Markov models of MD simulation data of various peptides. The amino acid substitution at the third position in each pentapeptide is shown in the panel. The error bars represent 95% confidence interval as determined by bootstrapping using 100 replicates.

where the degree of participation is given by the magnitude of the components.[42,72] The eigenvector associated with the largest eigenvalue (unity) gives the stationary probability distribution of the states.

The statistical uncertainties in the implied time scale were determined by the bootstrapping method.[73] From the simulation data set, 100 bootstrap sets were generated by randomly selecting trajectories with replacement. The implied time scales were calculated for each bootstrap set, and 95% confidence intervals centered on the sample mean were estimated. The implied times as a function of lag times are shown for all the peptides in Figure 6. The slowest implied time can be in the range of 30 ns (X=C) to 450 ns (X=M). As the slowest transitions among the most populated communities are likely to have the highest impact on the determination of equilibrium population from the MD simulations, we also determined the slowest relaxation rate among the top ten communities (Table 1). The extent of participation of each of the top ten communities in the slowest transition can be visualized using network graphs. Figure 7 shows the top ten communities for six peptides, where the nodes are colored according to the magnitude of the associated eigenvector. In many cases, the slowest transition is between the largest community and the community having the peptide conformations where one of the dihedrals is in the turn region of the Ramachandran plot. Figure 8 shows several randomly chosen conformations from such communities for the three peptides, where X is D, P, or V. In case of X=D or P, the slowest transition involves a leucine residue at the fourth position, whereas for X=V a leucine residue at the second position is involved. In both cases, the dihedral transitions between the polyproline and the turn regions of the Ramachandran plot.

TREMD Simulations. In the case of the TREMD simulations, the percentage population in each community at a temperature of interest was obtained by combining data from all the temperatures using the MBAR method as implemented in pyMBAR.[74] In brief, for K temperatures, each containing N_k samples, resulting in a total of N samples, $N = \sum_{k=1}^{K} N_k$, we computed the $N \times K$ matrix of weights $\mathbf{W}$ where each element W_{ni} is calculated as

$$W_{ni} = \hat{c}_i^{-1} \frac{e^{u_i(x_n)}}{\sum_{k=1}^{K} N_k \hat{c}_k^{-1} e^{u_k(x_n)}} \quad (5)$$

$$\hat{c}_i = \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{e^{u_i(x_{jn})}}{\sum_{k=1}^{K} N_k \hat{c}_k^{-1} e^{u_k(x_{jn})}} \quad (6)$$

Table 1. Slowest Implied Time Involving Transitions to/from Top Ten Communities Obtained from Transition Matrix T(τ) with τ = 4 ns

amino acid	implied time (ns)
A	14
C	7
D	547
E	406
F	9
G	25
H	353
I	216
K	22
L	225
M	12
N	8
P	57
Q	11
R	37
S	10
T	18
V	362
W	283
Y	56

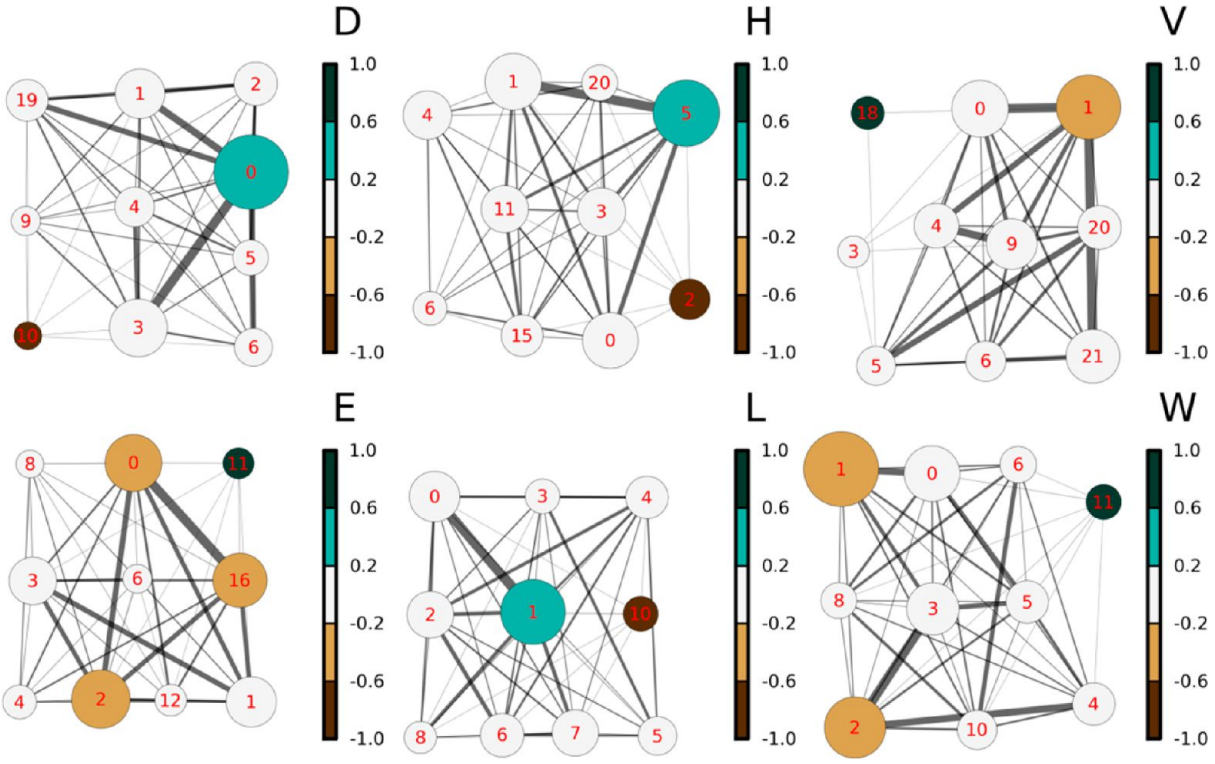


Figure 7. Undirected graphs with the nodes representing the communities and the edges representing transition probabilities, for six representative peptides represented by their letter code. The sizes of the nodes are proportional to the equilibrium population of respective communities, and the widths of the edges are proportional to transition counts. Only the top ten communities are shown. Each node is colored according to the magnitude of the eigenvector associated with the transition, using the color scale shown in the colorbars.

Here, $u_i(\boldsymbol{x})$ is a reduced potential function $u_i(\boldsymbol{x}) = \beta_i U_i(\boldsymbol{x})$ where $\boldsymbol{x}$ denotes the configuration of the system, and $\hat{c}_i$ is computed by solving eq 6 self-consistently.[74]

In order to obtain the percentage population of each community for comparison with standard MD simulations, we clustered the peptide conformations from individual temperatures using the same backbone dihedral angle criteria as used for the standard MD simulations. Furthermore, we partitioned the REMD peptide clusters into communities using the community mappings that were obtained from the standard MD simulations. As $\sum_{n=1}^{N} W_{ni} = 1$ for all $i = 1, ..., K$, the percentage population of communities at temperature of interest was obtained by summing the weights of configurations assigned to particular communities.

The community populations observed in MD and TREMD and those obtained from the eigenvalue decomposition of the transition matrix with a lag time of 4 ns are compared in Figure 9. In most of the cases, the population of the largest community is overestimated in MD simulations as compared to REMD simulations. Overall the agreement between MD and TREMD is good. The community population observed MD and that obtained from the MSM model built with a 4-ns lag time are in agreement. The error bars in the observed population are the standard deviations that were obtained using the bootstrapping procedure described previously.

DISCUSSION

We investigated the conformational dynamics of WW peptides by performing multiple MD simulations. We first inspected multiple MD simulations of the same peptide by comparing the end-to-end distance autocorrelations $C_{ee}(t)$. On average, the $C_{ee}(t)$ can extend well beyond 40 ns for most of the peptides. Several simulations have an extremely slow $C_{ee}(t)$ decay (see Figure 1 P, E, and H) compared to the rest. A marked variability in different replicas of the same simulations suggests that the dynamics of the peptide are quite sensitive to the initial conditions and that a single MD simulation of 0.6 μs duration is inadequate for sampling the dynamics of the peptide.

Among the amino acids, proline has a remarkable effect on backbone flexibility and dynamics. A WW peptide with X=P has comparatively few populated states. The $P(r)$ of X=P has only one peak at 1.24 nm, and the states with end-to-end distance less than 1.0 nm are practically nonexistent. It exists predominantly in polyproline-II type conformations; 60% of the total conformations sampled have all the backbone dihedrals either in polyproline-II or strand regions.

The majority of the peptides (except X=P) have two peaked end-to-end probability densities. The peak at ~1.24 nm (peak I) is due to the peptide backbone adopting conformations where all the dihedrals are in polyproline II or strand region. The broad peak at ~1.0 nm (peak II) is the result of flipping of one or more dihedral to the helical regions of the Ramachandran plot. The communities for most of the peptides consist of conformations that have the same dihedral in either polyproline II or strand conformation. This indicates that the transitions between the two states are fairly rapid.

The prevalence of polyproline II conformations among unstructured peptides has been reported in the literature.[75] The polyproline II conformations are considered to allow optimal hydrogen bonding between the peptide backbone and water molecules.[76,77] In the case of peptides with X=C, G, S, T, V, F, W or Y, peak II is either higher or similar in magnitude to peak

dx.doi.org/10.1021/ct3008217 | J. Chem. Theory Comput. 2013, 9, 1657−1666

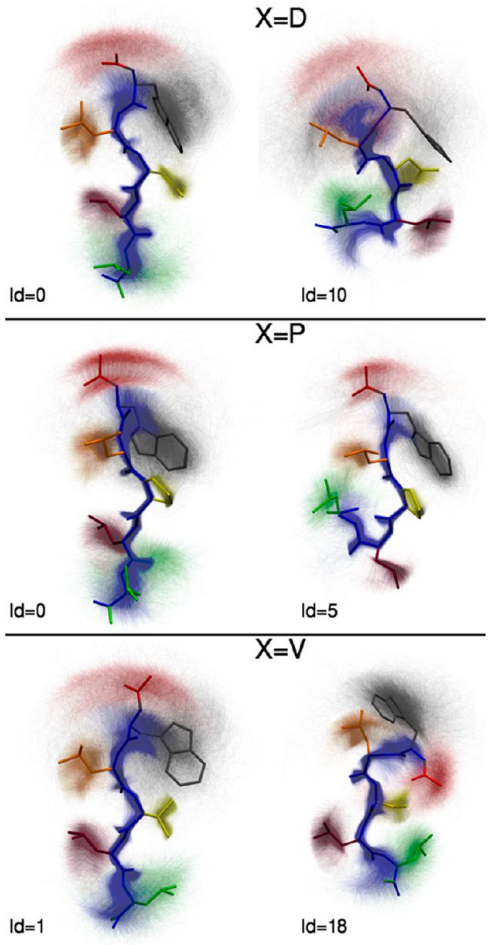


Figure 8. Representative conformations from the two clusters that are involved in the slowest transitions for X=D (community ids 0 and 10, top panel), X=P (community ids 0 and 5, middle panel), and X=V (community ids 1 and 18, bottom panel). The conformations were aligned using backbone atoms of residue numbers 2, 3, and 4. The template conformation on which all conformations were aligned is shown as Licorice. The backbone atoms are colored blue, and each side chain residue (including N terminal) is colored using different colors. In order to visualize the conformational distribution, the conformations are shown using material with high transparency. All hydrogen atoms have been omitted from all the representations. The figure was generated using VMD software.[78]

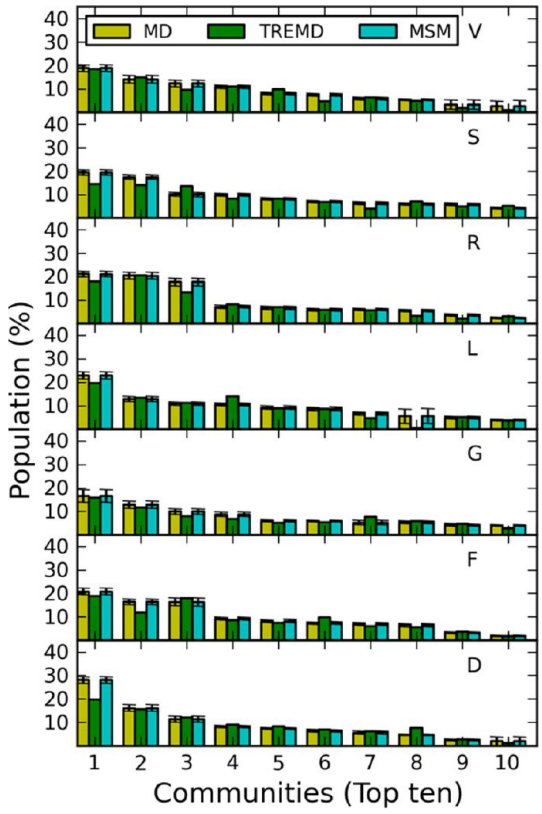


Figure 9. Percent population in the top ten communities for the six peptides. The populations obtained from MD simulations, TREMD simulations, and from MSM models are shown with yellow, green, and cyan bars, respectively. The error bars for MD and MSM population represents standard deviation obtained by bootstrapping.

I. Comparatively, these peptides do not prefer extended conformations.

Overall, even though the WW peptides do not have a single predominant conformation in one of the well-known secondary structure(s), the majority of the observed conformations can be clustered into a handful of states. The single amino acid substitution can have a significant impact on average behavior of the peptides of this length scale indicating that prediction of mean end-to-end distances and end-to-end distance distributions for small, unstructured peptides without simulations will be difficult.

The main advantage of the MSM building approach presented in this study is its computational efficiency. Clustering protein conformations based on backbone dihedral angles eliminates the need to build RMSD matrices. Even though the total number of possible conformational states increase as x^N where N is number of dihedral pairs, the observed population remains tractable, for example, the number of conformational states observed in MD simulations of these peptides is around 1−2% of the total possible states. Therefore, it should be practically feasible to apply clustering based on backbone dihedrals to larger peptides and proteins. This initial clustering, combined with community detection algorithm of Blondel et al., provides a quick way to generate MSM models for a large amount of data.

The implied time remains almost flat after the lag time of 2 ns for most of the peptides. The implied times provide information about the various relaxation rates of the system toward the equilibrium. While in some cases the slowest relaxation rate involves a transition to/from a cluster that has a very low equilibrium population, in many cases (X=D, E, H, I, L, V, and W), the slowest implied times among the top ten clusters are between 224 and 547 ns. In these cases, the transition involves a cluster where one of the dihedral is in the turn region.

It can be expected that, for REMD simulations, the choice of the initial conditions will introduce less bias in determining the equilibrium population of peptide conformations. The populations of communities obtained from MD simulations and those from replica exchange simulations are in good agreement, providing further confidence that the data collected from multiple MD simulations are adequate and that additional sampling will not drastically affect the observations made in this study.

dx.doi.org/10.1021/ct3008217 | J. Chem. Theory Comput. 2013, 9, 1657−1666

CONCLUSION

One goal of this study was to estimate the simulation length that might be required for the computation of PMF profiles of Wimley–White peptide binding to a lipid bilayer using Umbrella sampling and MD simulations. Assuming that for a reliable estimate of a PMF, the simulation length at each point along the reaction coordinate should be at least an order of magnitude longer than the slowest relaxation of the system, we estimate that MD simulations of approximately 2–5 μs each, along the reaction coordinate, would be needed. However, along the reaction coordinate, as the peptide approaches and interacts with the lipid bilayer, the conformational dynamics of peptide are likely to slow down even further; hence, even longer simulations are likely to be required. This also means that calculations of properties like binding affinities for antimicrobial peptides to membranes are unlikely to be accurately calculated with even microsecond-scale free-energy style windows and improved sampling methods are essential. Further studies are being carried out to investigate the conformational dynamics of the WW peptides at the bilayer interface.

AUTHOR INFORMATION

Corresponding Author

*E-mail: tieleman@ucalgary.ca.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council (Canada). G.S. is an Alberta Innovates Health Solutions (AIHS) postdoctoral fellow. D.P.T. is an AIHS Scientist and Alberta Innovates Technology Futures Strategic Chair in (Bio)Molecular Simulation. Calculations were carried out in part on WestGrid/Compute Canada facilities.

REFERENCES

(1) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–92.
(2) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 211–43.
(3) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. *Structure* **2006**, *14*, 437–49.
(4) Kasson, P. M.; Lindahl, E.; Pande, V. S. *PLoS Comput. Biol.* **2010**, *6*, e1000829.
(5) Sanbonmatsu, K. Y.; Joseph, S.; Tung, C.-S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15854–9.
(6) Romo, T. D.; Grossfield, A. *J. Chem. Theory Comput.* **2011**, *7*, 2464–2472.
(7) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.
(8) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLoS One* **2012**, *7*, e32131.
(9) Hayre, N. R.; Singh, R. R. P.; Cox, D. L. *J. Chem. Phys.* **2011**, *134*, 035103.
(10) Tobias, D. J.; Brooks, C. L. *J. Phys. Chem.* **1992**, *96*, 3864–3870.
(11) Cooke, B.; Schmidler, S. C. *J. Chem. Phys.* **2008**, *129*, 164112.
(12) Anderson, A. G.; Hermans, J. *Proteins* **1988**, *3*, 262–5.
(13) Smith, P. E. *J. Chem. Phys.* **1999**, *111*, 5568.
(14) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. *J. Am. Chem. Soc.* **2007**, *129*, 1179–89.
(15) Best, R. B.; Buchete, N.-V.; Hummer, G. *Biophys. J.* **2008**, *95*, L07–9.
(16) Chowdhury, S.; Lee, M. C.; Duan, Y. *J. Phys. Chem. B* **2004**, *108*, 13855–13865.
(17) Paschek, D.; Hempel, S.; García, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17754–9.
(18) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587–92.
(19) Snow, C. D.; Zagrovic, B.; Pande, V. S. *J. Am. Chem. Soc.* **2002**, *124*, 14548–14549.
(20) Duan, Y. *Science (80-)* **1998**, *282*, 740–744.
(21) Freddolino, P. L.; Schulten, K. *Biophys. J.* **2009**, *97*, 2338–47.
(22) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517–20.
(23) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.
(24) MacCallum, J. L.; Tieleman, D. P. *Trends Biochem. Sci.* **2011**, *36*, 653–62.
(25) Wimley, W. C.; White, S. H. *Nat. Struct. Biol.* **1996**, *3*, 842–848.
(26) Aliste, M. P.; Tieleman, D. P. *BMC Biochem.* **2005**, *6*, 30.
(27) Aliste, M. P.; MacCallum, J. L.; Tieleman, D. P. *Biochemistry.* **2003**, *42*, 8976–8987.
(28) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. *Biophys. J.* **2008**, *94*, 3393–3404.
(29) MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. *J. Gen. Physiol.* **2007**, *129*, 371–377.
(30) Singh, G.; Tieleman, D. P. *J. Chem. Theory Comput.* **2011**, *7*, 2316–2324.
(31) De Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schäfer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2013**, *9*, 687–697.
(32) *Free Energy Calculations. Theory and Applications in Chemistry and Biology*; Chipot, C., Pohorille, A., Eds.; Springer: 2007; Vol. *86*, pp 199–247.
(33) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
(34) Vorobyov, I.; Bennett, W. F. D.; Tieleman, D. P.; Allen, T. W.; Noskov, S. *J. Chem. Theory Comput.* **2012**, *8*, 618–628.
(35) Paloncýová, M.; Berka, K.; Otyepka, M. *J. Chem. Theory Comput.* **2012**, *8*, 1200–1211.
(36) Crouzy, S.; Woolf, T. B.; Roux, B. *Biophys. J.* **1994**, *67*, 1370–86.
(37) Neale, C.; Bennett, W. F. D.; Tieleman, D. P.; Pomès, R. *J. Chem. Theory Comput.* **2011**, *7*, 4175–4188.
(38) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. *Multiscale Model. Simul.* **2006**, *5*, 1214–1226.
(39) Noé, F. *J. Chem. Phys.* **2008**, *128*, 244103.
(40) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *J. Phys. Chem. B* **2004**, *108*, 6582–6594.
(41) Bowman, G. R.; Huang, X.; Pande, V. S. *Methods* **2009**, *49*, 197–201.
(42) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(43) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.
(44) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
(45) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
(46) Hess, B.; Kutzner, C.; Van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
(47) Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
(48) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.
(49) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.
(50) Swendsen, R.; Wang, J. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.
(51) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451–458.
(52) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140–150.
(53) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
(54) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96–123.

(55) García, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345−54.
(56) Trebst, S.; Troyer, M.; Hansmann, U. H. E. *J. Chem. Phys.* **2006**, *124*, 174903.
(57) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001**, *335*, 435−439.
(58) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(59) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 236−240.
(60) Lyman, E.; Zuckerman, D. M. *Biophys. J.* **2006**, *91*, 164−72.
(61) Yang, S.; Banavali, N. K.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 3776−81.
(62) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 664−7.
(63) Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19765−9.
(64) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011−6.
(65) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526−8.
(66) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. *J. Stat. Mech.: Theory Exp.* **2008**, *2008*, P10008.
(67) Newman, M. *Phys. Rev. E* **2004**, *70*, 056131.
(68) Brandes, U.; Delling, D.; Gaertler, M.; Goerke, R.; Hoefer, M.; Nikoloski, Z.; Wagner, D. 2006, ArXiv ID:physics/0608255.
(69) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
(70) Bacallado, S.; Chodera, J. D.; Pande, V. *J. Chem. Phys.* **2009**, *131*, 045106.
(71) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.
(72) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.
(73) Efron, B. *Ann. Stat.* **1979**, *7*, 1−26.
(74) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
(75) Shi, Z.; Woody, R. W.; Kallenbach, N. R. *Adv. Protein Chem.* **2002**, *62*, 163−240.
(76) Rath, A.; Davidson, A. R.; Deber, C. M. *Biopolymers* **2005**, *80*, 179−85.
(77) Pappu, R. V.; Rose, G. D. *Protein Sci.* **2002**, *11*, 2437−55.
(78) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 27−28, 33−38.