JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Learning Kinetic Distance Metrics for Markov State Models of Protein Conformational Dynamics

Robert T. McGibbon and Vijay S. Pande*

Department of Chemistry, Stanford University, Stanford, California 94305-4401

ABSTRACT: Statistical modeling of long timescale dynamics with Markov state models (MSMs) has been shown to be an effective strategy for building quantitative and qualitative insight into protein folding processes. Existing methodologies, however, rely on geometric clustering using distance metrics such as root mean square deviation (RMSD), assuming that geometric similarity provides an adequate basis for the kinetic partitioning of phase space. Here, inspired by advances in the machine learning community, we introduce a new approach for learning a distance metric explicitly constructed to model kinetic similarity. This approach enables the construction of models, especially in the regime of high anisotropy in the diffusion constant, with fewer states than was previously possible. Application of this technique to the analysis of two ultralong molecular dynamics simulations of the FiP35 WW domain identifies discrete near-native relaxation dynamics in the millisecond regime that were not resolved in previous analyses.

1. INTRODUCTION

Molecular dynamics (MD) simulation is a powerful computational tool for probing complex molecular processes in atomistic detail, including vesicle fusion,[1] protein−ligand binding,[2] protein folding[3,4] and mis-folding,[5] conformational change,[6] and crystallization.[7] Advances in computing capabilities including commodity and specialized hardware,[8,9] faster simulation codes,[10−13] optimization for graphical processing units and game consoles,[14,15] and distributed and cloud computing frameworks[16] currently enable the collection of milliseconds of aggregate simulation data.

These systems are characterized by complex and high-dimensional conformational state spaces, with the possibility for multiple independent slow degrees of freedom. Without prior knowledge of the relevant conformational states, the objective and unbiased analysis of simulations of these systems is a significant challenge. Recently, Markov state models (MSMs) have emerged as an attractive, scalable approach to the analysis of these massive data sets.[17,18] Using clustering to divide the conformational space, MSMs model the system's dynamics as a memory-less jump process between discrete conformational states. This approach avoids the possibly deceptive projection of the system's dynamics onto a low-dimensional order parameter space[19,20] and enables the systematic combination of multiple short simulations for the prediction of long-time scale phenomena.[18,21,22] Furthermore, the same set of theoretical and computational tools can be used to build detailed models for quantitative prediction and course-grained models for insight.[23−26] In addition, MSM methodologies facilitate efficient allocation of compute resources during the sampling phase by adaptive exploration of conformational space.[27−30]

Despite the advantages of MSMs, a number of challenges remain. In particular, methods for discretizing the conformational space into a set of states using purely geometric criteria such as the Cartesian root-mean-square deviation (RMSD) suffer when there exist slow conformational transitions between geometrically proximate states, such as register shift dynamics in β topologies[25] or subtle but slow conformational change dynamics. The choice of distance metric with which to cluster conformations has recently received attention,[31,32] but a general framework for choosing the optimal distance metric for describing molecular kinetics is lacking. Here, we present a new approach. Motivated by recent work in semisupervised learning,[33,34] and specifically an algorithm by Shen et al.,[35] we present a novel algorithm for kinetic distance metric learning which constructs metrics explicitly designed to enable kinetic clustering of conformations from MD data sets.

Whereas here we consider MSMs built directly from molecular dynamics simulations, other approaches exist as well for producing network models for the dynamics of biomolecular systems. These include approaches that proceed by direct enumeration of potential energy basins.[36−39] In such methods, an effective clustering of configurations on the potential energy landscape is accomplished by considering the set of local energy minima and transition rates estimated by unimolecular rate theory. While these methods fit naturally into a familiar chemical kinetics framework, they suffer from an exponential scaling in complexity with respect to the system's number of degrees of freedom.

Received: February 19, 2013
Published: May 20, 2013

dx.doi.org/10.1021/ct400132h | *J. Chem. Theory Comput.* 2013, 9, 2900−2906

2. METHODS

MSM construction metholodologies based on the clustering of MD trajectories typically involve the use of some geometric criterion such as RMSD, grouping the sampled conformations into a large number of microstates. In addition to the choice of distance metric, considerable attention has been focused on the choice of clustering algorithm.[25,40,41] On this discrete state space, the model is parametrized by estimating transition probabilities or rates between the microstates from the available simulation data. The resulting transition or rate matrix can then be analyzed by various means to find transition pathways between states of interest and kinetically metastable macrostates.[23−26,42] The model can also be used to predict experimental observables such as those measured by temperature jump florescence and infrared spectroscopy[43,44] triplet−triplet energy transfer[45] and single molecule fluorescence resonance energy transfer (FRET).[46]

Two major sources of error characterize the MSM approach. The first is error introduced by the Markov approximation, the assumption that the future evolution of the system is independent of its history given its present state. Although Hamiltonian dynamics are Markovian in the full $6N$-dimensional space of the atomic positions and their conjugate momenta,[47] the dynamics between discrete conformational states are not, as the projection formally introduces a nontrivial memory kernel into the equations of motion.[48] The neglect of these memory effects in the MSM formalism is a source of systematic error. Although memory effects can be modeled by building higher-order or variable-order Markov models,[49,50] the number of parameters to be estimated increases exponentially with the order of the Markov chain, limiting the utility of this approach to systems with large or high-dimensional state spaces. Although this systematic error due to non-Markovity can be systematically reduced by increasing the number of microstates,[51] this bias reduction is countered by an increase in the second major source of error, statistical uncertainty in the estimation of the pairwise transition probabilities. As the microstates become smaller and more numerous, they become less likely to contain internal free energy barriers and more Markovian. Unfortunately, due to the quadratic scaling of the number of transition probabilities or rates that require estimation with respect to the number of states, these estimations become noisy and liable to overfitting if the number of states is increased without bound.

Other avenues for mitigating the systematic bias due to non-Markovian projected dynamics include methods such as the weighted ensemble (WE).[52−54] Similar to MSMs, the WE also utilizes a discrete partitioning of phase space to define reaction rates. In contrast to MSMs, which are generally applied as a postprocessing technique to analyze an ensemble of short-time trajectories, WE methodologies leverage the state decomposition during the sampling to stop and start trajectories as they explore phase space carrying different probabilistic weights. This need for state definitions before the WE production simulation leads to a substantially increased computational cost.

To reduce the error due to non-Markovian memory effects within the MSM framework without increasing the size of the state space, we seek a kinetic clustering at the microstate level, such that conformations that can kinetically interconvert quickly are grouped together in contrast to conformations separated by longer time scales. For this purpose, we propose that a gold-standard distance metric for such a kinetic clustering would be the *commute time*, the mean first passage time for the round trip commute between infinitesimal voxels in phase space surrounding two structures. Such a metric would measure the mutual kinetic accessibility of two structures, such that low distances are associated with rapid interconversion, while maintaining the proper symmetry with respect to exchange. Unfortunately, with N frames of MD simulation data, the estimation of N^2 commute times between all pairs of sampled conformations is not achievable without performing a considerable amount of further simulation.

Instead, we introduce a new distance metric for clustering conformations obtained by MD simulation into microstates with the explicit inclusion of kinetic information. Adapting an algorithm from Shen et al. for distance metric learning in the large-margin machine learning framework,[35] we attempt to learn a Mahalanobis distance metric from molecular dynamics trajectories, which, operating on the structural features of a pair of conformations, is able to predict whether these conformations are kinetically close or kinetically distant. Using this approach, we find that it is possible build more Markovian MSMs with fewer states than is possible with existing methods and identify previously hidden slow conformational transitions.

2.1. Algorithm. Our goal is to learn a geometric distance metric that maximizes the margin of separation between kinetically close and kinetically distant conformations, which we call "kinetically discrimitory metric learning" (KDML). We take as input a collection of N triplets of structures, (a, b, c), such that conformations a and b are subsequent frames from a single MD trajectory separated by a short lag time, τ_{close}, whereas structure c is selected from further down the trajectory, at a longer delay time, τ_{far}, from a. The structures must be projected into a suitable vector space such as the space of all heavy-atom heavy-atom pairwise distances or the space of backbone dihedral angles.

We then look for a distance metric that can distinguish the fact that in each triplet, conformations a and b are likely "kinetically close" whereas a and c are "kinetically distant". We choose to restrict our attention to squared Mahalanobis distance metrics, which generalize squared weighted euclidean metrics. The distance metric is an inner product.

$$d^X(\vec{a}, \vec{b}) = (\vec{a} - \vec{b})^T X(\vec{a} - \vec{b}) \tag{1}$$

$$=(\vec{a} - \vec{b})^T Q^T \Lambda Q(\vec{a} - \vec{b}) \tag{2}$$

$$=[Q(\vec{a} - \vec{b})]^T \Lambda [Q(\vec{a} - \vec{b})] \tag{3}$$

The Mahalanobis matrix, $\boldsymbol{X}$, is required to be symmetrical and positive semidefinite, which guarantees that d^X is symmetric and that $d^X(\vec{a},\vec{b}) \geq 0$. (Two special cases of $\mathbf{X}$ are of interest. When $\mathbf{X}$ is diagonal, we have a squared weighted euclidean metric, and when $\mathbf{X} = \mathbf{I}$, the distance metric is simply the squared euclidean distance. In the general case, as demonstrated by eq 3, the squared Mahalanobis distance can be viewed as a squared weighted euclidean distance after projection onto the eigenbasis of $\mathbf{X}$.

Ideally, the data would permit a metric capable of correctly classifying all of the training examples, yielding a positive margin, ρ_i, on each training example i, where $\rho_i = d^{\mathbf{X}}(\vec{a}_i,\vec{c}_i) - d^{\mathbf{X}}(\vec{a}_i,\vec{b}_i)$.

However, in general, it will not be possible to satisfy all of these constraints simultaneously. Instead, we seek a metric that will admit as large margin as possible on as many of the training examples as possible.

Following Shen,[35] we define our objective function f, which we seek to maximize with respect to $\mathbf{X}$ and ρ.

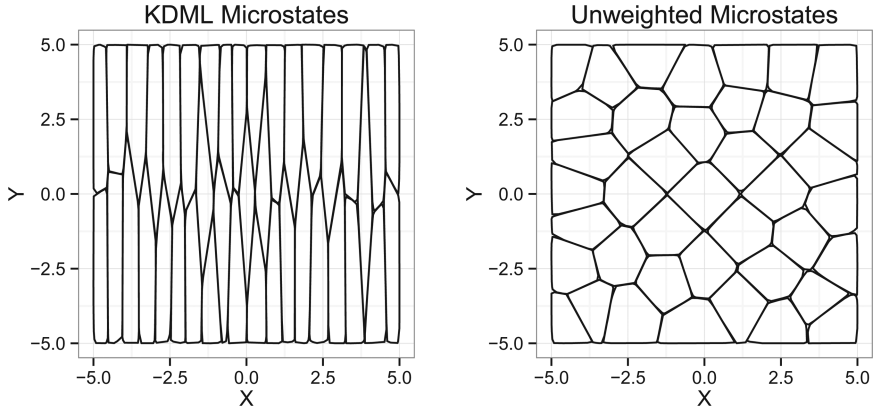

Figure 1. Voronoi decomposition of a two-dimensional space into 40 microstates using both the KDML and unweighted euclidean distance metric with the hybrid K-medoids clustering algorithm.[41] Using the kinetic distance metric, the microstates are geometric elongated in the direction of fast motion and are less likely to contain internal free energy barriers, leading to fewer non-Markovian memory effects without increasing statistical error.

$$f(X, \rho) = \alpha\cdot\rho - \frac{1}{N}\sum_{i=1}^{N} \lambda(d^X(\vec{a}_i, \vec{c}_i) - d^X(\vec{a}_i, \vec{b}_i) - \rho) \tag{4}$$

Here, ρ represents the "target" margin, and $\lambda(\cdot)$ is a smooth hinge loss function that penalizes margins on the training data less than the target margin ρ. The parameter α controls the balance between the desire to maximize the margin and to minimize the losses. We find $\mathbf{X}$ relatively insensitive to α for $0 < \alpha < 1$.

The objective function, f, is maximized subject to the constraints that $\mathrm{Tr}(\mathbf{X}) = 1$, $\rho \geq 0$, and $\mathbf{X}$ is positive semidefinite. The constraint on the trace of $\mathbf{X}$ removes the scale ambiguity of the Mahalanobis matrix. For efficiency with gradient based optimization techniques, we employ the Huber loss function λ_{Huber}, a differential extensions of the hinge loss.

$$\lambda_{\mathrm{Huber}}(\xi) = \begin{cases} 0 & \text{if } \xi \geq h \\ \dfrac{(h-\xi)^2}{4h^2} & \text{if } -h < \xi < h \\ -\xi & \text{if } \xi \leq -h \end{cases} \tag{5}$$

For the special case that $\mathbf{X}$ is diagonal, a change of variables allows the optimization to be performed without constraints. This yields an efficient solution via Broyden−Fletcher−Goldfarb−Shanno (BFGS) method. For the general $\mathbf{X}$, we employ a specialized gradient descent algorithm[35] that outperforms general purpose semidefinite programming solvers by taking advantage of the fact that a trace-one positive semidefinite matrix can be written as a convex linear combination of rank-one trace-one matrices[55] to construct gradient descent steps that naturally preserve the power spectral density (PSD) property.

The algorithms described above have been implemented as plugins for the MSMBuilder package (version 2.6 and later), which is released under the GNU General Public License and available at https://simtk.org/home/msmbuilder. This code itself is available at https://github.com/SimTk/KDML and is released under the same terms.

3. RESULTS AND DISCUSSION

3.1. Toy System. We first demonstrate the algorithm for a simple toy system: two-dimensional Brownian motion with an anisotropic diffusion constant. Despite its extreme simplicity, this model captures one essential feature of biomolecular conformational dynamics: the time scales corresponding to orthogonal degrees of freedom can be vastly different. An effective kinetic clustering of such systems requires that the characteristic length scale of the clusters in the directions with the slow diffusion be lower than those in the direction of fast diffusion. Such a clustering, with states that are elongated in the direction of fast motion, is statistically efficient in the sense that it uses no more states, and thus parameters, than are necessary to capture the intrastate dynamics.

For our toy system, we set the ratio of the diffusion constants in the x and y dimensions at 10:1. We apply the diagonal KDML learning algorithm described, training on $N = 4981$ triplets of structures sampled from 100 trajectories of length 500 steps. The time lag between "kinetically close" structures was taken to be $\tau_{\mathrm{close}} = 1$ time step, with $\tau_{\mathrm{far}} = 10$. The resulting Mahalanobis matrix is

$$X_{10} = \begin{pmatrix} 0.9915 & 0.0 \\ 0.0 & 0.0085 \end{pmatrix} \tag{6}$$

The Mahalanobis matrix shows that in the kinetic distance metric, motion in the x direction is up-weighted, since motion along this degree of freedom is slower. When we use this kinetic distance metric to run clustering with the hybrid K-medoids algorithm, the clusters produced are geometrically elongated in the y direction. A comparison of the clusters produced using both kinetic metric as well as a standard euclidean distance is shown in Figure 1.

In two dimensions, the effect of the distance metric on the state decomposition are obvious—the ratio of the weights on each degree of freedom are translated roughly into the geometrical aspect ratio of the clusters. Since the mean time to displace by one unit of distance in the x direction is longer than the mean time to displace by the same amount in the y direction, a successful clustering than minimizes the sum of the within-state mean first passage times is one that, similar to the clustering produced by the kinetic distance metric, is not geometrically spherical but instead is elongated in the y direction.

To benchmark the performance of this distance metric, we compute the longest two implied time scales for MSMs built with both an unweighted metric and the KDML metrics, with a variety of numbers of states. The implied time scales, which are computed from the eigenvalues of the transition matrix as $\tau_{\mathrm{lag}}/\ln \lambda_i$ describe the time scales for characteristic dynamical modes of the kinetic network. If system is Markovian, the implied time

scales are invariant to changes in the lag time; internal consistency demands that models built at a lag time of τ be equivalent to those built with a lag time of 2τ, provided that for every two steps propagated in the first model we take only one time step in the second model. In practice, non-Markovian behavior is generally manifest as erroneously fast time scales that increase with lag time.[56] As shown in Figure 2, when compared to

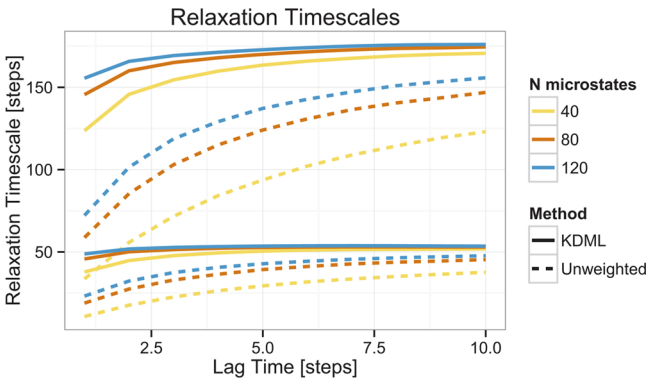

Figure 2. Longest two implied time scales for the MSMs produced by different procedures. Dashed curves are for models built using the standard euclidean distance metric, whereas solid lines are for those built using KDML. Models built using KDML are slower and are more Markovian compared to those built using the euclidean metric, even with fewer states.

models built using the standard distance metric, those built using the kinetic distance metric show longer time scales that converge more quickly with respect to lag time, indicating the quicker onset of Markovian behavior. Furthermore, the longest implied time scale for models built using the standard distance metric with 120 states shows behavior similar to that of models built with only 80 states using the kinetic metric. Because the number of fit parameters in the model goes as the square of the number of states, this corresponds to similar performance with less than half the number of adjustable parameters. In situations of comparative data sparsity, this enables a better balance between systematic and statistical error.

3.2. FiP35 WW. Next, we apply our approach to a real data set via a reanalysis of two one hundred microsecond simulations of the FiP35 WW domain, performed by D. E. Shaw et al.[57] A previous analysis of this data set using Markov State Model methods revealed the existence of two folding pathways.[43] However, in order to achieve a structural resolution in the microstate clustering of greater than 4.5, it proved necessary in that study to use more than 26 000 microstates. Parameterizing this model required formally estimating more than 6.7×10^8 pairwise transition probabilities. Because of the obvious potential for overfitting, we ask if it is possible to use KDML to recover more structural and kinetic detail with fewer states.

To construct our model, we begin by representing each conformation in the data set by a vector encoding the sine and cosine of its backbone Φ and Ψ dihedral angles. This has the effect of vectorizing the conformations, as well as breaking the periodic symmetry of the angle measurement. Using these vectors, we apply the KDML procedure with $k = 20\,000$ triplets sampled with $\tau_{close} = 2$ ns and $\tau_{far} = 20$ ns. One hundred rounds of optimization lead to a distance metric that reweights structural features in the protein coordinate space based on their kinetic relevance.

Comparing models built using the KDML procedure or an unweighted euclidean distance metric, we find two striking results (Figure 3). First, the folding time scale is unchanged by

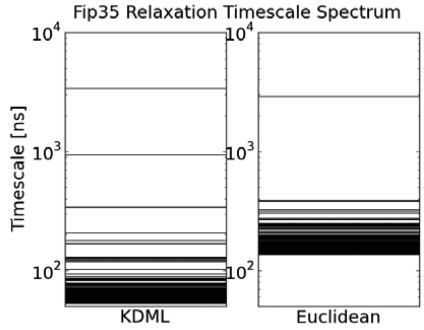

Figure 3. Two Markov state models for the FiP35 WW Domain. The model built using the KDML procedure (left) resolves a number of discrete, slow eigenprocesses that were suppressed using the unweighted euclidean distance metric. Both models contain 5000 microstates, a 75 ns lag time, and were built using the same clustering procedure.[41]

the introduction of the KDML method. In fact, given trends in the estimates of the folding time scale with respect to changes in the number of states (data not shown), we estimate that this analysis significantly underestimates the folding time scale. However, more strikingly, our analysis shows the emergence of new discrete time scales in the relaxation spectrum that were not observed with the euclidean distance metrics. While faster than the folding time scale, the dynamical processes in the hundreds of nanoseconds to microsecond regime are of significant interest (Figure 4).

To probe the structural dynamics corresponding to these new time scales, we construct a four macrostate model directly from our five thousand microstate model, which seeks to optimally capture the slowest three dynamical processes using the PCCA+ algorithm.[24] Visual analysis of the four states reveals that they correspond to the expected folded and unfolded states in addition to two near-native states characterized by rearrangements in the hydrogen bonding structure in the two loop regions. Specifically, one of the states (Figure 5) shows a reorganization of THR25 forming a nonnative hydrogen bond with HIS23. In the other loop, our final macrostate shows a reorganization of the hydrogen bonding network around ASP15 in which a native hydrogen bond between ASP15 and SER13 is broken and the inflection of the chain at the loop is altered to instead create a set of hydrogen bonds across the loop, including between ARG17O and ARG14N, as shown in Figure 6.

Transitions involving the folding and unfolding of the near native THR flip state occur faster than unfolding of the native state and dominate the second time scale detected in our analysis with a relaxation time of 750−1000 ns. Dynamics between the native and ASP flip state on the other hand occur faster and dominate the third discrete time scale on the order of 300 ns.

We believe that these discrete metastable but near native states were missed by conventional RMSD-based clustering analysis because despite interconverting with the native state slowly, the structural distinctions are subtle. Folding, on the other hand, may be robust to changes in the distance metric because the conformation difference between the folded and unfolded states is so vast.

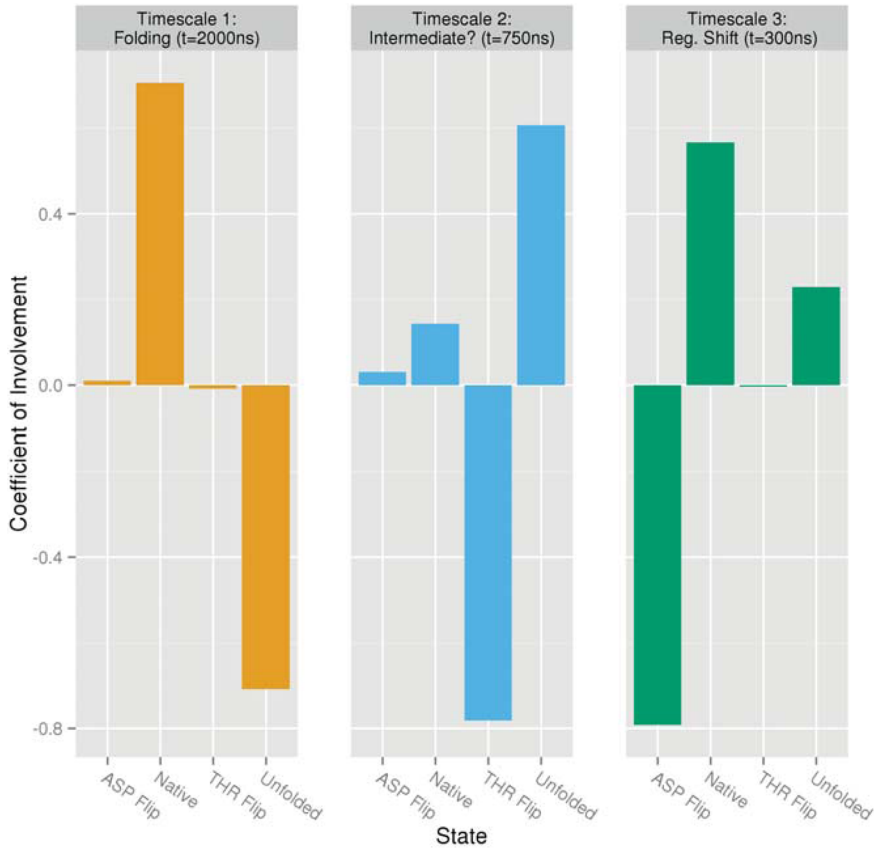

Figure 4. Four state macrostate model is able to capture the three slowest dynamic processes in the KDML model of FiP35. The slowest dynamical process represents the direct folding pathway. The second and third dynamical processes are largely associated with two states not previously identified.

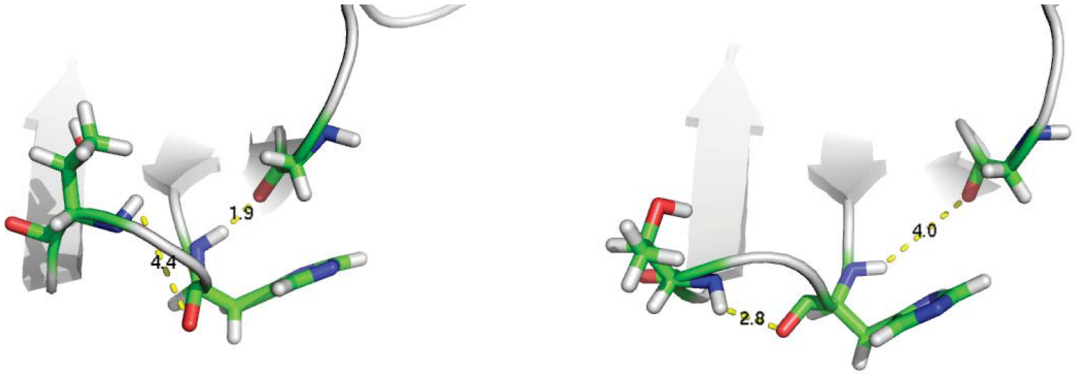

Figure 5. Representative structure from the simulation's native basin (left) and THR Flip macrostate (right). The subtle structural difference between the two conformational basins, which corresponds to a repacking of the hydrogen bond network in the second hairpin, is identified only via the KDML algorithm.

4. CONCLUSION

By optimizing the distance metric, we have shown that it is possible to build more accurate Markov state models from finite simulation data. An optimal distance metric is one that, while grouping structures geometrically, incorporates a measure of kinetic proximity. In this work, drawing on recent advances in the machine learning community, we approach that goal within a large-margin Mahalanobis distance metric learning framework. These learned KDML distance metrics allow the construction of Markov state models with fewer states and thus less statistical error than those built using conventional techniques without the bias in the model due to the increase prevalence of non-Markovian effects in this regime. Furthermore, such distance metrics may enable further accuracy improvements within other frameworks for the analysis of biomolecular conformational dynamics, which leverage a discrete-state partitioning, such as weighted ensemble methods.

Reanalyzing two-hunded microseconds of atomistic MD simulation of the FiP35 WW domain by Shaw et al.,[57] we show that the KDML metric permits the identification of key metastable states and discrete relaxation time scales, which were obscured in previous analyses. These metastable near-native states involve subtle structural changes, making them difficult to identify with conventional distance metrics. Such slow dynamics, which nonetheless involve only small changes in gross structural distance metrics such as RMSD, may be a common feature of

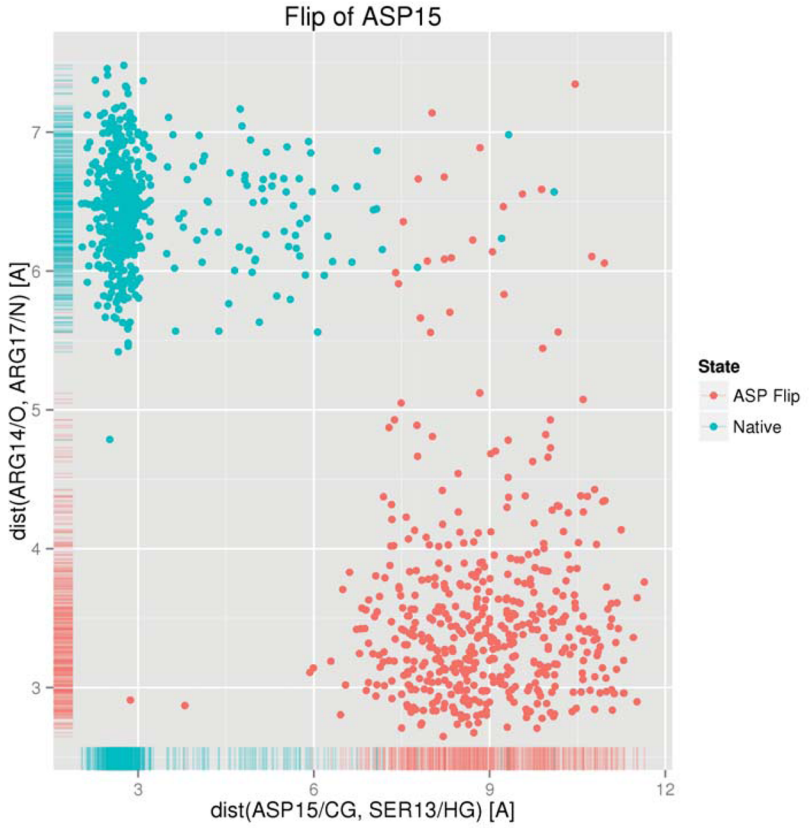

Figure 6. Flip of ASP15 and the hydrogen bonding structure in the first loop of Fip35 distinguishes the native state from a kinetically distinct near native state.

large proteins, especially those with functionally relevant conformation changes.

Remaining challenges in the analysis of ultralarge molecular dynamics data sets include the enhancement of the connection between the identification of metastable states and the direct search for slow structural degrees of freedom, the direct incorporation of dynamical information into clustering procedures, and further efforts to automate and systematize the Markov state model construction process from MD, for example by optimal selection of parameters within a likelihood maximization framework.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Kasson, P. M.; Lindahl, E.; Pande, V. S. Atomic−resolution simulations predict a transition state for vesicle fusion defined by contact of a few lipid tails. *PLoS Comput. Biol.* **2010**, *6*, e1000829.

(2) Hansson, T.; Oostenbrink, C.; van Gunsteren, W. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2002**, *12*, 190−196.

(3) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120−127.

(4) Bowman, G. R.; Voelz, V. A.; Pande, V. S. Taming the complexity of protein folding. *Curr. Opin. Struct. Biol.* **2011**, *21*, 4−−11.

(5) Lin, Y.; Bowman, G.; Beauchamp, K.; Pande, V. Investigating how peptide length and a pathogenic mutation modify the structural ensemble of amyloid β monomer. *Biophys. J.* **2012**, *102*, 315−24.

(6) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.* **2002**, *9*, 646−652.

(7) Matsumoto, M.; Saito, S.; Ohmine, I. Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing. *Nature* **2002**, *416*, 409−413.

(8) Allen, F.; et al. A vision for protein science using a petaflop supercomputer. *IBM Syst. J.* **2001**, *40*, 310−327.

(9) Shaw, D. E.; et al. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **2008**, *51*, 91−97.

(10) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(11) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.

(12) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. ACM/IEEE Conf. Supercomput. (SC06)* **2006**, 43−43.

(13) Pearlman, D.; Case, D.; Caldwell, J.; Ross, W.; Cheatham, T.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. A. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1−41.

(14) Luttmann, E.; Ensign, D. L.; Vaidyanathan, V.; Houston, M.; Rimon, N.; Øland, J.; Jayachandran, G.; Friedrichs, M.; Pande, V. S. Accelerating molecular dynamic simulation on the cell processor and Playstation 3. *J. Comput. Chem.* **2009**, *30*, 268−274.

(15) Friedrichs, M.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A.; Ensign, D. L.; Bruns, C.; Pande, V. S. Accelerating molecular dynamic simulation on graphics processing units. *J. Comput. Chem.* **2009**, *30*, 864−872.

(16) Shirts, M.; Pande, V. S. Screen savers of the world unite! *Science* **2000**, *290*, 1903−1904.
(17) 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.
(18) Noé, Frank; Schütte, Christof; Vanden-Eijnden, Eric; Reich, Lothar; Weikl, Thomas R. Constructing the equilibrium ensemble of folding pathways from short offequilibrium simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011−19016.
(19) Krivov, S. V.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766−14770.
(20) Muff, S.; Caflisch, A. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins Struct. Funct. Bioinf.* **2008**, *70*, 1185−1195.
(21) Chodera, J.; Swope, W.; Pitera, J.; Dill, K. Long-time protein folding dynamics from short time molecular dynamics simulations. *Multiscale Model. Simul.* **2006**, *5*, 1214−1226.
(22) 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.
(23) Deuflhard, P.; Huisinga, W.; Fischer, A.; Scütte, C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.* **2000**, *315*, 39−59.
(24) Deuflhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* **2005**, *398*, 161−184.
(25) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17807−17813.
(26) Bowman, G. R. Improved coarse-graining of Markov state models via explicit consideration of statistical uncertainty. *J. Chem. Phys.* **2012**, *137*, 134111.
(27) Hinrichs, N. S.; Pande, V. S. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J. Chem. Phys.* **2007**, *126*, 244101.
(28) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced modeling via network theory: adaptive sampling of Markov state models. *J. Chem. Theory Comput.* **2010**, *6*, 787−794.
(29) Pronk, S.; Larsson, P.; Pouya, I.; Bowman, G.; Haque, I.; Beauchamp, K.; Hess, B.; Pande, V.; Kasson, P.; Lindahl, E. Copernicus: A new paradigm for parallel adaptive molecular dynamics. *Int. Conf. High Perform. Comput., Network., Storage Anal. (SC11)* **2011**, 1−10.
(30) Weber, J. K.; Pande, V. S. Characterization and rapid sampling of protein folding Markov state model topologies. *J. Chem. Theory Comput.* **2011**, *7*, 3405−3411.
(31) Cossio, P.; Laio, A.; Pietrucci, F. Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? *Phys. Chem. Chem. Phys.* **2011**, *13*, 10421−10425.
(32) Zhou, T.; Caflisch, A. Distribution of reciprocal of interatomic distances: A fast structural metric. *J. Chem. Theory Comput.* **2012**, *8*, 2930−2937.
(33) Xing, E. P.; Ng, A. Y.; Jordan, M. I.; Russell, S. Distance metric learning, with application to clustering with side-information. *Adv. Neural Inf. Process. Syst. (NIPS)* **2002**, *15*, 505−512.
(34) Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. Constrained K-means clustering with background knowledge. *Proc. 18th Int. Conf. Mach. Learn.* **2001**, *18*, 577−584.
(35) Shen, C.; Kim, J.; Wang, L. Scalable large-margin Mahalanobis distance metric learning. *IEEE Trans. Neural Networks* **2010**, *21*, 1524−1530.
(36) Noé, F.; Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opp. Struct. Biol.* **2008**, *18*, 54−162.
(37) Carr, J. M.; Wales, D. J. Globals optimization and folding pathways of selected α-helical proteins. *J. Chem. Phys.* **2005**, *123*, 234901.
(38) Carr, J. M.; Wales, D. J. Folding pathways and rates for the three-stranded β-sheet peptide β3s using discrete path sampling. *J. Phys. Chem. B* **2008**, *112*, 8760−8769.
(39) Prada-Gracia, D.; Gómez-Gardeñes, J.; Echenique, P.; Falo, F. Exploring the free energy landscape: from dynamics to networks and back. *PLoS Comput. Biol.* **2009**, *5*, e1000415.
(40) Keller, B.; Daura, X.; van Gunsteren, W. F. Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* **2010**, *132*, 074110.
(41) 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.
(42) Huang, X.; Yao, Y.; Bowman, G. R.; Sun, J.; Guibas, L. J.; Carlsson, G.; Pande, V. S. Constructing multi-resolution Markov state models (msms) to elucidate RNA hairpin folding mechanisms. *Pac. Symp. Biocomput.* **2010**, *15*, 228−239.
(43) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.
(44) Zhuang, W.; Cui, R. Z.; Silva, D.-A.; Huang, X. Simulating the T-jump-triggered unfolding dynamics of trpzip2 peptide and its time-resolved IR and two-dimensional IR signals using the Markov state model approach. *J. Phys. Chem. B* **2011**, *115*, 5415−5424.
(45) Beauchamp, K. A.; Ensign, D. L.; Das, R.; Pande, V. S. Quantitative comparison of villin headpiece subdomain simulations and triplet−triplet energy transfer experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12734−12739.
(46) Voelz, V.; J Aager, M.; Yao, S.; Chen, Y.; Zhu, L.; Waldauer, S. A.; Bowman, G. R.; Friedrichs, M.; Bakajin, O.; Lapidus, L. J.; Weiss, S.; Pande, V. S. Slow unfolded-state structuring in Acyl-CoA binding protein folding revealed by simulation and experiment. *J. Am. Chem. Soc.* **2012**, *134*, 12565−12577.
(47) Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*, 3rd ed.; Elsevier: Amsterdam, 2007.
(48) Zwanzig, R. Memory effects in irreversible thermodynamics. *Phys. Rev.* **1961**, *124*, 983−992.
(49) Bacallado, S. Bayesian analysis of variable-order, reversible Markov chains. *Ann. Stat.* **2011**, *39*, 838−864.
(50) Park, S.; Pande, V. S. Validation of Markov state models using Shannon's entropy. *J. Chem. Phys.* **2006**, *124*, 054118.
(51) Sarich, M.; Noé, F.; Schütte, C. On the approximation quality of Markov state models. *Multiscale Model. Simul.* **2010**, *8*, 1154−1177.
(52) Feng, H.; Costaouec, E.; Darve, E.; Izaguirre, J. A. A comparison of weighted ensemble and Markov state model methodologies. *J. Chem. Theory Comput.*, under review.
(53) Huber, G. A.; Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.* **1996**, *70*, 97−110.
(54) Bhatt, D.; Zuckerman, D. M. Heterogeneous path ensembles for conformational transitions in semiatomistic models of adenylate kinase. *J. Chem. Theory Comput.* **2010**, *6*, 3527−3539.
(55) Shen, C.; Welsh, A.; Wang, L. PSDBoost: Matrix-generation linear programming for positive semidefinite matrices learning. *Adv. Neural Inf. Process. Syst. (NIPS)* **2008**, *21*, 1473−1480.
(56) Djurdjevac, N.; Sarich, M.; Schütte, C. Estimating the eigenvalue error of Markov state models. *Multiscale Model. Simul.* **2012**, *10*, 61−81.
(57) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341−346.