JCTC Journal of Chemical Theory and Computation

Article

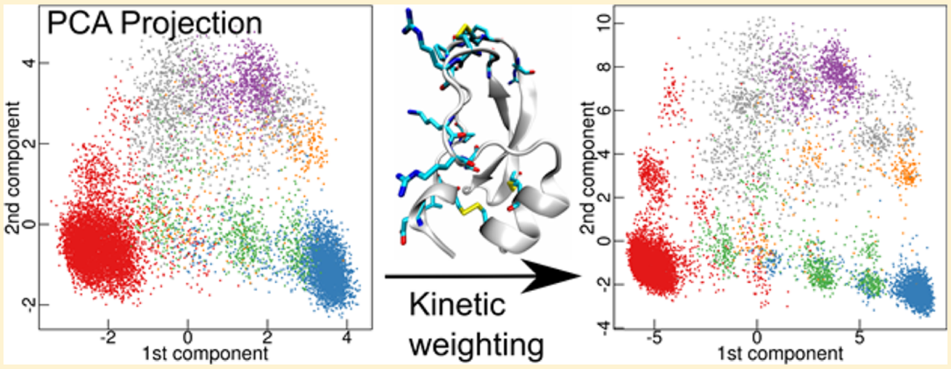
pubs.acs.org/JCTC

Weighted Distance Functions Improve Analysis of High-Dimensional Data: Application to Molecular Dynamics Simulations

Nicolas Blöchliger, Amedeo Caflisch, and Andreas Vitalis*

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

Ⓢ *Supporting Information*



ABSTRACT: Data mining techniques depend strongly on how the data are represented and how distance between samples is measured. High-dimensional data often contain a large number of irrelevant dimensions (features) for a given query. These features act as noise and obfuscate relevant information. Unsupervised approaches to mine such data require distance measures that can account for feature relevance. Molecular dynamics simulations produce high-dimensional data sets describing molecules observed in time. Here, we propose to globally or locally weight simulation features based on effective rates. This emphasizes, in a data-driven manner, slow degrees of freedom that often report on the metastable states sampled by the molecular system. We couple this idea to several unsupervised learning protocols. Our approach unmasks slow side chain dynamics within the native state of a miniprotein and reveals additional metastable conformations of a protein. The approach can be combined with most algorithms for clustering or dimensionality reduction.

1. INTRODUCTION

The analysis of high-dimensional data is susceptible to several pitfalls.[1−4] Most unsupervised learning methods, such as clustering or dimensionality reduction, require a notion of similarity or distance between individual observations or snapshots. If individual snapshots are vectors of high dimensionality, most functional forms measuring distance lack contrast, *i.e.*, for a given query point the nearest and farthest data points are almost equally far from it.[5,6] Additional problems arise because the data might contain a large number of irrelevant features (dimensions), and because the importance of features can differ for different data points or clusters.[7−9] As a consequence, the choice of a distance function offering sufficient contrast can be more important than the choice of learning method.[10−12] This calls for efficient protocols to derive similarity measures that do not suffer from lack of contrast and account for local feature relevance. These measures should be accessible without an intricate understanding of the system described by the data.

For high-dimensional data, it is common to select or generate features that are deemed informative. When performed manually, this process relies primarily on domain expertise. Measures of relevance, such as entropy or mutual information, can serve as guides to nonexpert users.[13] The term feature extraction is commonly associated with techniques of dimensionality reduction.[13] Many of these techniques try to generate new features that maximize a target property, *e.g.*, variance in principal component analysis.[14] Low-dimensional embeddings of high-dimensional data might be of limited use if these data contain many irrelevant features, and if the chosen distance function is unable to distinguish between similar and dissimilar points. It has been noted that feature selection prior to dimensionality reduction can improve the discriminatory power of the latter.[15] Lastly, the contrast level offered by a given distance function may also depend on the position of the two points in data space, and this is reflected in clustering algorithms with locally adaptive similarity measures.[8,9]

Here we focus on high-dimensional data from molecular dynamics (MD) simulations of biomolecules.[16] At its core, analysis of MD data is often concerned with identifying metastable conformations of the simulated system.[17−19] Unsupervised learning methods for this purpose include clustering and related techniques,[11,20−25] classical dimension-

Received: June 30, 2015
Published: October 27, 2015

DOI: 10.1021/acs.jctc.5b00618

ality reduction algorithms and modifications thereof,[14,26−33] as well as other approaches.[34,35] The success of all these methods depends on the careful selection of features or an informative distance function; however, a lot of trial-and-error is used in practice to improve results.

In this contribution, we present an efficient method to either globally or locally weight features according to a notion of relevance. Recognizing that features exhibiting slow modes are more likely to report on metastable states, we define weights based on effective rates. Global weights employ the autocorrelation function, while locally adaptive weights are a function of transition rates within a time window along the trajectory. We apply these approaches to an illustrative model system and two data sets generated by MD simulations. The first set of MD data originates from simulations of the reversible folding of Beta3S,[36] a 20-residue peptide adopting a three-stranded, antiparallel β-sheet fold. The second example is a very long explicit solvent simulation of the conformational dynamics of bovine pancreatic trypsin inhibitor (BPTI) within its native state.[37] Throughout, we discuss problems that can occur in conjunction with unmodified distance functions and show how weights address them. Where possible, we compare our results to analyses of the same data found in the literature. We show that a comprehensive description of the free energy surface can be extracted from MD trajectories of proteins by including degrees of freedom such as side chains, flexible loops, and terminal residues with appropriate weights. These features are often dismissed *a priori* as noisy and uninteresting, which entails the risk of losing important information.

2. METHODS

Weighted Distance Functions. Consider a set of N observations with each observation corresponding to a data vector of length D. The Euclidean distance between two observations $\mathbf{x}(t_k)$ and $\mathbf{x}(t_l)$ gives equal weight to all their D features:

$$d(\mathbf{x}(t_k), \mathbf{x}(t_l))^2 = D^{-1} \sum_{i=1}^{D} (x_i(t_k) - x_i(t_l))^2 \tag{1}$$

Conversely, the information content relevant for a given target application may differ between features. Given a notion of overall relevance expressed in a vector of weights, $\mathbf{w}$, a weighted Euclidean distance can take into account the heterogeneity of the features as follows:

$$d(\mathbf{x}(t_k), \mathbf{x}(t_l))^2 = \left(\sum_{i=1}^{D} w_i \right)^{-1} \sum_{i=1}^{D} w_i (x_i(t_k) - x_i(t_l))^2 \tag{2}$$

The elements of $\mathbf{w}$ used in eq 2 can represent any notion of importance. Here, we quantify the relevance of features by measurements of net rates obtained independently for each of them. Features associated with low rates are interesting as they are likely to report on metastable states.[38,39] It is expected that a subset of features is homogeneous on the same time scale as the life times of these states. In practice, for the weights in eq 2, we set $w_i = \max(R_i(\tau), 0)$, where $R_i(\tau)$ is the autocorrelation function of the i^{th} feature evaluated at a specific time lag τ. Note that this corresponds to scaling the data and is different from altering the metric itself, *e.g.*, by changing the Euclidean (L_2) to a rectilinear (L_1) norm. In the present work, we often use dihedral angles and represent them by sine and cosine terms. Rather than computing separate weights in this case, we simply keep the larger of the two values derived independently as the resultant weight.

Global weights as used in eq 2 cannot reflect that the importance of individual features might depend on where a given observation is situated in the overall data space. We use locally adaptive weights to account for this. Here, the notion of "local" is derived exclusively from proximity in time, which is a limitation. Unfortunately, the autocorrelation function computed over a data window of width Δ becomes misleading if transitions are absent. Instead, locally adaptive weights are derived by counting the number of times a feature crosses its global mean:

$$\begin{aligned} n_i^k(\Delta) &= \sum_{j=k-\Delta/2}^{j=k+\Delta/2} H(-(x_i(t_{j-1}) - \langle x_i \rangle_N)(x_i(t_j) - \langle x_i \rangle_N)) \\ w_i^k &= (\mathrm{n}_i^k(\Delta) + \alpha)^{-1} \end{aligned} \tag{3}$$

Here, H denotes the Heaviside step function, and α is a parameter required to be positive. The weights in eq 3 are expected to be low for features that sample unimodal distributions. If a feature differs between states, eq 3 rewards those features with locally small variances. False negatives can be obtained if the global data mean coincides with a specific peak position in a multimodal distribution. Distance is measured as

$$d(\mathbf{x}(t_k), \mathbf{x}(t_l))^2 = \left(\sum_{i=1}^{D} \sqrt{w_i^k w_i^l} \right)^{-1} \sum_{i=1}^{D} \sqrt{w_i^k w_i^l} (x_i(t_k) - x_i(t_l))^2 \tag{4}$$

We note that the function d does not necessarily satisfy the triangle inequality, *i.e.*, eq 4 no longer represents a metric. This may be undesirable. In the context of clustering algorithms, we might also require a measure of distance between an individual observation, $\mathbf{x}(t_k)$, and a group of observations (cluster). Representing the cluster by its unscaled centroid, $\mathbf{c}$, we have

$$d(\mathbf{x}(t_k), \mathbf{c})^2 = \left(\sum_{i=1}^{D} \sqrt{w_i^k w_i^c} \right)^{-1} \sum_{i=1}^{D} \sqrt{w_i^k w_i^c} (x_i(t_k) - c_i)^2 \tag{5}$$

In eq 5, $\mathbf{w}^c$ is the average weight vector across all observations that are part of the cluster with centroid $\mathbf{c}$.

Progress Index and SAPPHIRE Plots. Recently, we have developed an algorithm for the analysis of long MD trajectories.[34,35] The resulting SAPPHIRE (States And Pathways Projected with HIgh REsolution) plot is a comprehensive visualization of the thermodynamics and kinetics of the simulated system and is used here to study the performance of the distance functions introduced above.

We briefly describe the method next and refer the reader to the original publications for more details.[34,35] Specifically, all snapshots are assumed to form a complete graph, and the minimum spanning tree or an approximation to it is computed. From a given starting snapshot, the snapshot connected by the shortest available edge is added to a growing partition. The resulting sequence, the so-called progress index, proceeds through regions of high sampling density one after another and avoids overlap of distinct states.[34] The progress index can be annotated to yield a SAPPHIRE plot as described in recent work.[35] Here, we employ the following annotation functions to highlight and interpret the states along the progress index. First, we use a kinetic annotation function to localize the individual states on the progress index. Specifically, for every snapshot i along the progress index, we plot the average of the mean first-passage times between A_i and S_i, denoted τ_{MFP}, where A_i is the set of snapshots added to the progress index before i and S_i is the set of those added after i. The value of this annotation function is low within a state and high in transition regions, and barriers are highlighted reliably (although they cannot be interpreted quantitatively).[34] Second, we plot the actual sampling time of the individual snapshots to illustrate when and in which sequence the different states were sampled. Third, we characterize the states themselves by structural annotations. For Beta3S, we have used the secondary structure assignment according to the DSSP algorithm[40] and the χ_1 angle of Trp10. For BPTI, we show selected dihedral angles using binning with boundaries given in the Supporting Methods. The boundaries were obtained from direct inspection of the individual histograms for each angle. In addition, we show state assignments according to Shaw et al.[37] and Xue et al.[41]

The method is implemented in the CAMPARI simulation and analysis package (http://campari.sourceforge.net). Detailed parameter settings are given in the Supporting Methods. In contrast to previous work, we modify the underlying spanning tree before computing the progress index (Vitalis, manuscript submitted). In particular, we collapse the leaves into their parent vertex, which means that they are added to the progress index as soon as it encounters their parent vertex. This places snapshots from the fringe region around regions of high sampling density next to the snapshots from the closest state. The procedure can be repeated a number of times, and this is a controllable parameter. It is set by CAMPARI keyword FMCSC_CPROGMSTFOLD, which was 1 throughout except for Figure 2 (where it was 2).

Clustering and Cut-Based Free Energy Profiles. Besides SAPPHIRE plots, we employ clustering and cut-based free energy profiles[23] (cfeps) to study the influence of the distance function. Clustering according to a recent, tree-based algorithm partitions the data into tight clusters that have little overlap and are of controllable size.[20] Cfeps order the resultant clusters by their kinetic distance from a chosen reference state. The ordering is kinetically annotated with τ_{MFP}, defined as above. As for SAPPHIRE plots, the value of τ_{MFP} is expected to be low within a basin and high in transition regions. This is what allows an immediate partitioning into metastable states.

3. RESULTS

To illustrate the problems that occur when analyzing data without feature selection, we use a model system and two high-dimensional real-world data sets from MD simulations of the peptide Beta3S[36] and the protein BPTI[37] obtained in implicit and explicit solvent, respectively. We highlight the performance of the different similarity measures by employing a recently developed algorithm for the analysis of dynamical systems that uses a distance function as its only essential parameter.[34] The similarity (or better, dissimilarity) measures evaluated are the unweighted Euclidean distance (UW), the Euclidean distance weighted by the global autocorrelation function at fixed lag time per dimension (GW), and a locally adaptive distance defined by time-local transition rates (LAW). They are defined in eqs 1, 2, and 4, respectively (see Methods). We demonstrate that the weighted distance functions, GW and LAW, offer substantial benefits in all cases investigated. For brevity, we will repeatedly refer to the three dissimilarity measures as UW, GW, and LAW measures below.

Model System. Figure 1a schematically depicts a Markov model of 4 states and its associated transition matrix with the states identified by color throughout. A Markov chain (random walker) is used to generate a continuous trajectory of length 2 × 10^5 snapshots, which means that even the least likely (red) state is sampled sufficiently. To be able to meaningfully test

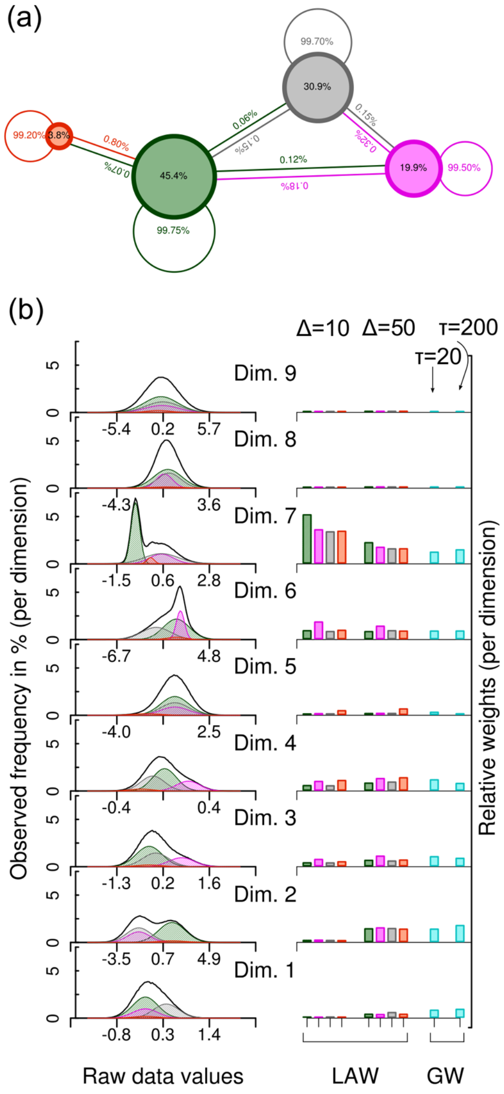


Figure 1. Model system and its representation. (a) Schematic description of the 4-state Markov model. The text within circles gives the steady-state population of each state. Nonzero elements of the transition matrix are shown as lines with the conditional probabilities indicated. Coloring of lines is by source state. (b) Each snapshot of the model is represented by 9 features. Features are generated by independent, memory-free, Gaussian processes with parameters that depend on the macrostate. On the left, we plot actual histograms (black lines) from a trajectory of 2 × 10^5 snapshots along with the generating functions scaled according to the steady-state population of each state (shaded areas). On the right, weights computed for the same trajectory are shown for each dimension for both LAW and GW measures. Locally adaptive weights are averaged separately for the true state the trajectory resided in and produced for two different window sizes, Δ = 10 and Δ = 50, with $\alpha = 0.01$ (see eq 4). Global weights are computed at two different lag times ($\tau = 20$ and $\tau = 200$).

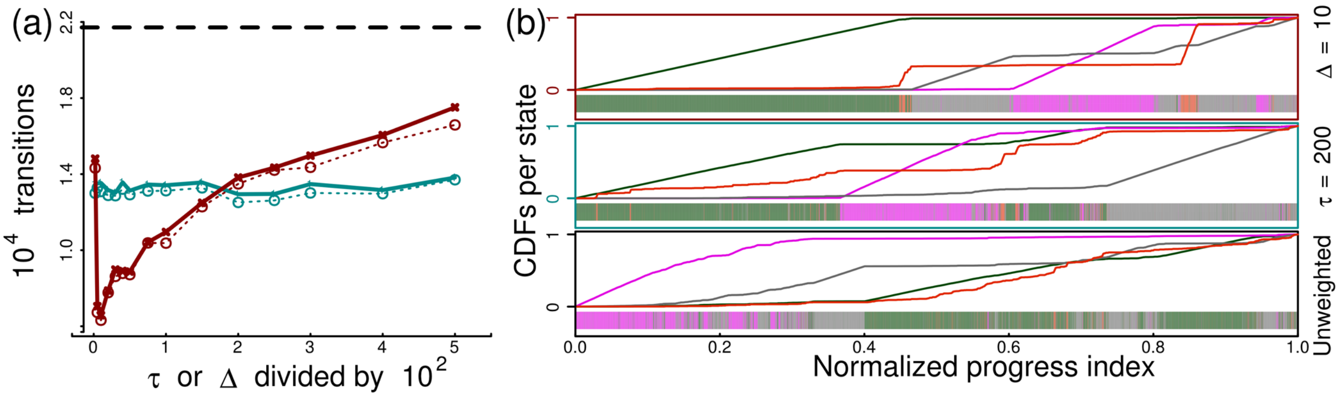


Figure 2. Evaluation of different distance functions for the model system in Figure 1. (a) The number of transitions between states in the progress index is shown. The construction of the progress index relies on preorganization via clustering, and we made use of a recent improvement to the algorithm (see Methods). Each condition for both types of weights was evaluated for 8 (GW, cyan lines) or 5 (LAW, dark red lines) different clustering settings, and the medians (solid lines) and minima (dotted lines) are plotted. The black dashed line is the minimum value for the unweighted case. (b) The exact state annotation (color bar) along the progress index is plotted for every 10th snapshot. Cumulative distribution functions were analyzed and normalized independently for each state. From bottom to top, we show the data for UW, GW, and LAW measures, respectively.

different distance measures for this system, it is represented by 9 data dimensions (features). Every feature is generated from a normal distribution whose parameters depend on the state the system currently resides in. As seen in Figure 1b (left-hand side), no feature is informative for all states. The overlap is generally large, and two features (#8 and #9) are completely uninformative. Despite the moderate dimensionality, this challenges the UW measure.

In Figure 1b we also compare the resultant global and locally adaptive weights underlying the GW and LAW measures, respectively. The global weights obtained from the autocorrelation function at fixed lag time de-emphasize features #5, #8, and #9 irrespective of lag time. At τ = 20, all remaining dimensions have roughly equivalent weights, whereas at τ = 200 features #2 and #7 dominate. These correspond exactly to the histograms with the clearest peak separations. Since we know the correct state for each snapshot, the locally adaptive weights can be averaged separately for different states. These weights correspond to the inverse crossing rate of the global data mean for a given feature (eq 4). This is why they emphasize features that have low variance for a given state, *e.g.*, #6 is particularly important for the magenta state or #7 for the green state. Similarly, they also reflect whether a feature's value in a given state is far away from the global mean, *e.g.*, #5 is only relevant for the red state. Note that these synthetic data are memory-free, *i.e.*, time correlation comes exclusively from state persistence.

We scanned a wide range of possible lag times and window sizes, and the particular values shown in Figure 1b correspond to the top performing cases in the subsequent analysis, which was performed as follows. Using a recent algorithm,[34] we computed the progress index that corresponds to stepping through an approximation of the minimum spanning tree (see Methods for details). This procedure is very sensitive to the distance function in use. Ideally, it should arrange snapshots exactly by their underlying states assuming they are geometrically separable. The large overlap seen in Figure 1b makes this task challenging. As a measure of sorting quality, we simply count the number of times the state annotation changes in the progress index, and these data are shown Figure 2a (lower is better). It is clear that the UW measure is rigorously outperformed by both the GW and LAW measures irrespective of parameter settings. GW is inferior to LAW (in terms of peak performance), and its performance appears to change relatively little with lag time. The results for the LAW measure show a clear preference for window sizes that are considerably less than the average life times of states, which is ~300 steps. We picked the respective top-performing cases for the results obtained with the UW, GW, and LAW measures and visualize the progress index in Figure 2b. Cumulative distribution functions resolved by state highlight the successive decrease in state overlap when changing from UW (bottom row) to GW (middle row) and finally to LAW (top row) measures. We note the improvement of the localization of the red state in particular. With the UW measure, this state would certainly not have been identified as a metastable state of the system.

Beta3S. The first MD data set is taken from an implicit solvent simulation of the 20-residue antiparallel β-sheet peptide Beta3S.[36] Multiple folding and unfolding events are observed during the total sampling time of 20 μs. The unfolded state ensemble is characterized by the presence of several metastable states that are enthalpically stabilized. The data set consists of 10^6 snapshots saved at an interval of 20 ps. We represent the peptide via 99 dihedral angles. The rotation of 2-fold or 3-fold symmetric groups consisting entirely of hydrogen atoms and χ_2 and χ_3 angles of tyrosine were ignored. Dihedral angles enter as their sine and cosine values to avoid intricacies with circular variables.[26]

First, we investigate how the use of locally adaptive weights affects clustering. We clustered the data according to both UW and LAW measures using a recent, tree-based algorithm[20] with thresholds that yield a total number of clusters within 2% of one another. We identified two clusters in the region of highest sampling density (*i.e.*, in the native state) sharing the same centroid. For adjacent lower density regions (see Figure S1 in Supporting Information for details), we picked two clusters whose centroids differ but which are of similar size and distance from the native state. Because distance functions based on dihedral angles are putatively uninformative, we crosscheck cluster definition against the most common and intuitive distance function, *viz.*, the root-mean-square deviation (RMSD) computed over the Cartesian coordinates of all atoms after

DOI: 10.1021/acs.jctc.5b00618

pairwise alignment. For the native state, Figure 3 (left) reveals that the quality of the clusters obtained by both UW and LAW measures is comparable, which indicates that excellent sampling density may overcome weaknesses of the distance function. However, more overlaps of alternative conformations are obtained when omitting weights (recognizable most clearly for Trp10). This is confirmed by the RMSD histograms in Figure 4a that exhibit a distinct tail for the cluster based on the UW measure. Such a tail is absent when inspecting the histograms for the UW measure directly (Figure 4b).

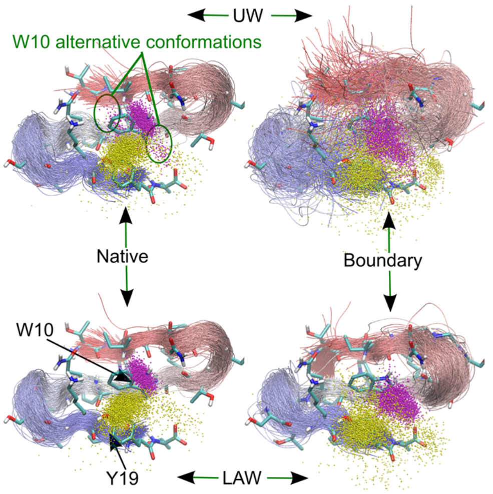


Figure 3. Illustrations of clusters corresponding to the native state and a lower density region, respectively. The clustering uses 99 nonsymmetric dihedral angles in conjunction with UW (eq 1) and LAW (eq 4, $\Delta = 2$ ns, $\alpha = 1$) measures. Further statistics are provided in the caption of Figure S1. For both measures, we identified clusters in the native basin and near its boundary. For the native state, two clusters could be obtained from UW (35215 snapshots) and LAW (33445 snapshots) measures, which share their centroid snapshot and overlap to 82% identity. For the boundary case, the two clusters shown were identified with the help of Figure S1 (UW: 4013 snapshots; LAW: 3883 snapshots). All cluster members were aligned to the native state centroid (displayed as sticks). For each case, ~500 snapshots are shown in ribbon representation (N-terminus is red). Magenta and yellow spheres document the positions of the NE1 and OH atoms of Trp10 and Tyr19, respectively. These data are shown for ~4000 cluster members. All graphics were rendered with VMD.[42]

We next focus on a region of lower sampling density, *viz.*, clusters situated in the boundary region of the native state (see Figure S1). Figure 3 (top right) demonstrates that the UW measure fails to produce an ensemble satisfying intuitive criteria for what a cluster is. This is rectified by applying the LAW measure, which produces a cluster ensemble that maintains native topology albeit with much increased fluctuations, and that has the side chain of Trp10 in a well-defined region distinct from that of the native state (bottom right). This result is quantified clearly by differences in the histograms of pairwise distances using the RMSD measure (Figure 4c). Obviously, the UW "cluster" contains a wide variety of structures with pairwise distances exceeding 8 Å. There is no difference between self-similarity and similarity to the native state. This is improved dramatically with the LAW cluster, for which the native state clearly is a more dissimilar conformation than the other cluster members.

Figure 4d suggests that the result in Figure 4c for the UW measure is likely due to dimensionality problems, *i.e.*, the distance distribution to a snapshot not part of the cluster is almost the same as the pairwise distribution within the cluster. This lack of contrast also holds when analyzing the distance distribution of all snapshots with respect to the native state. In fact, Figure 5a shows that the distribution remains nearly

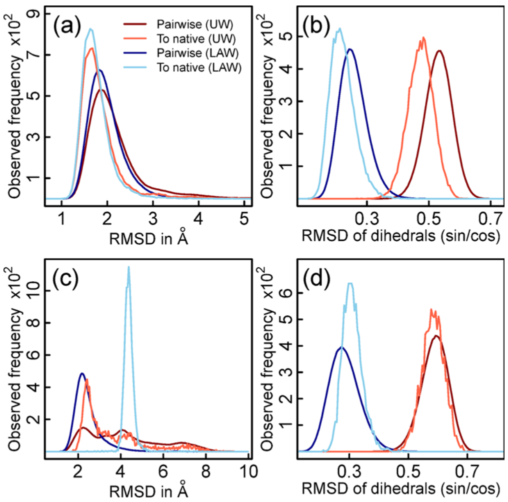


Figure 4. Distance histograms for the clusters in Figure 3. The clustering uses 99 nonsymmetric dihedral angles in conjunction with UW (eq 1) and LAW (eq 4, $\Delta = 2$ ns, $\alpha = 1$) measures. The legend in panel (a) applies to all panels. (a) For native-state clusters, a total of ~2×10^6 randomly selected and unique pairwise distances of the all-atom coordinate RMSD were computed, and histograms are shown along with complete histograms of distances to the native state centroid (bin size of 0.05 Å). (b) The same as (a) but using the actual UW and LAW measures to compute distances. (c) The same as (a) for the clusters from the boundary region of the native state. (d) The same as (c) but using the actual UW and LAW measures to compute distances.

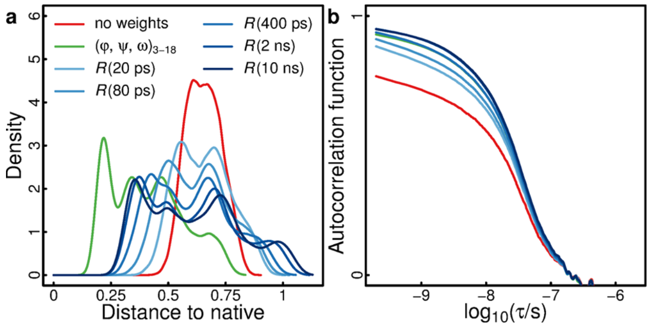


Figure 5. Weighted distance functions capture thermodynamics and kinetics of Beta3S better than an unweighted one. (a) Distance distributions for Beta3S with respect to a representative snapshot of the native basin. Using 103 dihedral angles (including χ_2 and χ_3 angles of tyrosine) and the UW measure, the distribution is essentially unimodal (red curve). With manual feature selection, the distribution has several distinct peaks that indicate coarse clusters in the data (green curve). Here we used the backbone dihedral angles of residues 3−18 as in previous work.[20] Increasing the time lag τ for the GW measure in this range leads to more and better separated peaks (blue curves). The GW measure with a time lag of $\tau = 2$ ns and the Euclidean measure with manual feature selection are correlated (Pearson's correlation coefficient $\rho = 0.978$). (b) Autocorrelation functions of the distance time series used in (a). For this figure, distances were only computed for every 10th snapshot in the trajectory.

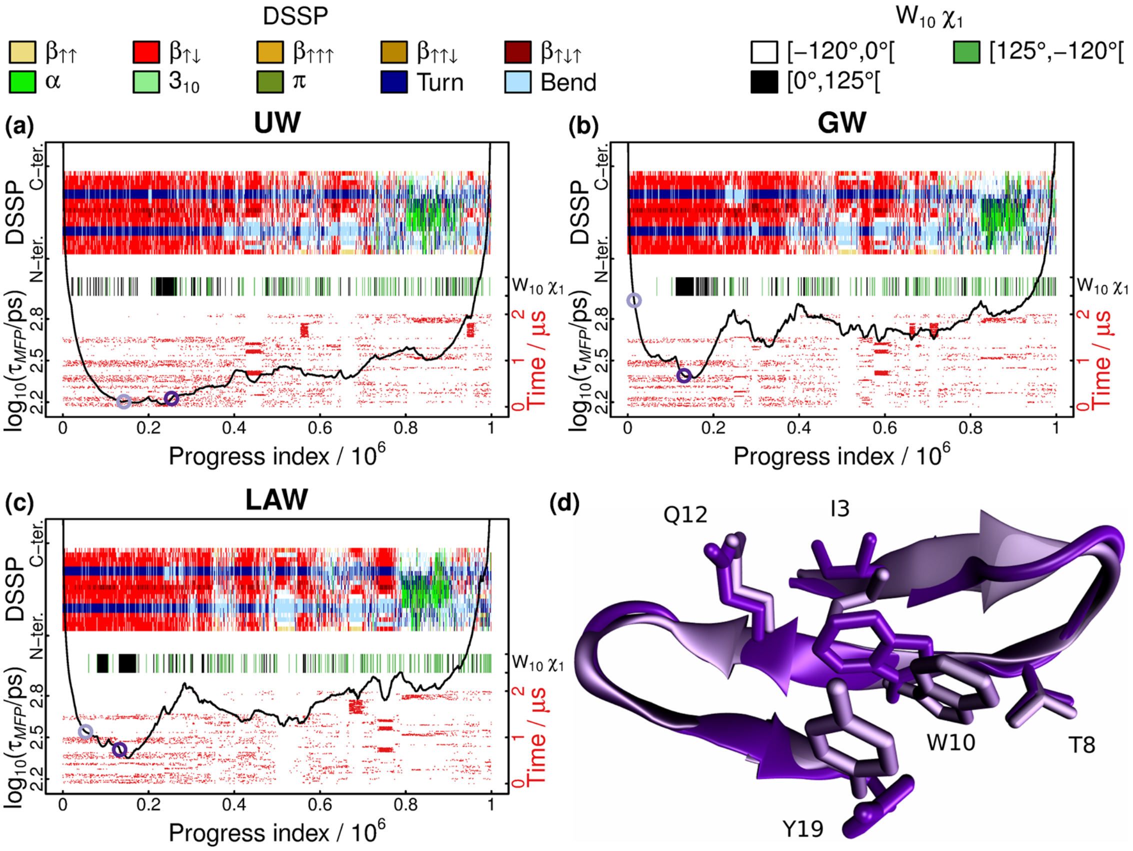


Figure 6. SAPPHIRE plots for Beta3S. (a) SAPPHIRE plot for Beta3S obtained with the UW measure. The peptide is represented by the sine and cosine values of 99 nonsymmetric dihedral angles. The progress index (x axis) represents a reordering of the trajectory snapshots that groups similar snapshots next to each other (see Methods). It is annotated with kinetic information (τ_{MFP}, a function whose value is low within states and high in transition regions, black profile in the bottom), sampling time (red dots, only shown for one out of 10 simulation runs), DSSP assignment[40] by residue (legend on top), and the χ_1 angle of Trp10 (legend on top). (b) The same as (a) for the GW measure with τ = 2 ns. (c) The same as (a) for the LAW measure with Δ = 2 ns and α = 1. All profiles in (a)−(c) start from the same snapshot. (d) Cartoon representations of two alternative native state conformations marked by color-coded circles in (a)−(c). Sticks highlight specific residues.

unimodal. This is due to the presence of many irrelevant and weakly coupled features. As a consequence, no threshold can be defined to approximately separate the native state from unfolded conformations, *i.e.*, nearest neighbor relations become meaningless.[1,10] Upon utilizing global weights, slow features have more influence, and the distribution has several distinct peaks that can be associated with native and unfolded conformations, respectively. We emphasize that a featureless distance spectrum is a fundamental and not merely a statistical problem, *i.e.*, it is not rectifiable by increasing the overall sampling density.

Figure 5b documents that on short time scales (<10 ns) the GW measure yields higher values for the autocorrelation of the corresponding distance time series than the UW measure. This result implies that kinetic proximity can be represented more accurately by weighted distance functions. The grouping or ordering of snapshots to reveal kinetically homogeneous states is precisely what Markov models,[24] diffusion maps,[31,32,43,44] cut-based free energy profiles[23] (see Figure S1), or SAPPHIRE plots[34] try to accomplish. The latter are an efficient tool for the analysis and visualization of long MD trajectories. SAPPHIRE plots offer an intuitive illustration of the states and sequence of events encountered during the simulation (see Methods), and we have previously used SAPPHIRE plots to analyze data from MD simulations of protein folding,[34,35] the conformational dynamics of proteins,[35,45] and the binding of a peptide to a protein domain.[46] We next use the method to further evaluate the discriminatory ability of the UW, GW, and LAW measures.

Figure 6 shows SAPPHIRE plots based on all 3 measures. The time lag for the GW measure was set to τ = 2 ns, and we used Δ = 2 ns and α = 1 for the LAW measure. All profiles start from a snapshot in the native basin of Beta3S (see Supporting Methods for further details). The UW measure is unable to discriminate between kinetically similar and dissimilar snapshots, which leads to a relatively featureless profile (Figure 6a). The low height of the folding barrier at a progress index value of 4 × 10^5 indicates that the cutting surface does not delineate metastable states accurately. With weights, higher barriers are obtained everywhere, and several metastable states can be detected besides the native state (Figures 6b and 6c). We use a secondary structure annotation resolved by residue that is based on the DSSP algorithm[40] to confirm that the individual basins correspond to distinct conformations of the peptide. For weighted distance functions, the kinetic annotation and the sampling time reveal substructure in the native state of Beta3S, some of which is the result of the dynamics of the χ_1 angle of Trp10. This side chain samples two distinct conformations within the native state as shown in Figure 6d. Previous analyses did not capture this partitioning of the native basin because the relevant features were omitted or because their effective weight was too low.[20,34,36,47−49] We show in Figure S1 that a backbone-centric RMSD distance places both of the conformations in Figure 6d in exactly the same basin.

In summary, Figure 5a illustrates why the inclusion of all features with equal weight is generally infeasible. We provide evidence for this in the context of three different unsupervised learning protocols (Figures 3, 4, 6, and S1). Our observations also point to the risks incurred by manual feature selection. Specifically for the mining of MD data, the primary risk lies in lumping kinetically separable states together as has happened for the native state of Beta3S in prior analyses. We believe that our approach of weighting the individual features according to kinetic information is a suitable compromise between these two extremes.

BPTI. We next analyzed the simulated dynamics of the 58-residue protein BPTI as reported in a very long MD trajectory containing 41250 snapshots saved every 25 ns.[37] In these data, BPTI explores several distinct, native-like states interconverting on the μs time scale. Compared to Beta3S, the data are of higher dimensionality, yet the overall variance is smaller.

To illustrate that angles decay on a wide range of time scales, Figure 7a plots the autocorrelation functions of selected dihedral angles. The time series of the slow χ_1 and ψ angles of Cys14 and Arg42, respectively, show that these angles likely report on metastable states, *i.e.*, jumps in these dihedral angles coincide with jumps in the RMSD time series (Figure S2). In contrast, no such conclusion is obtained for the time series of a fast angle, *e.g.*, the χ_1 angle of Leu6. These observations corroborate our hypothesis that slow degrees of freedom are more relevant than fast ones. In Figure 7b, we plot the global weights required for the GW measure. The data confirm that the slow dynamics are generally governed by the anchor points of the Cys14-Cys38 disulfide bond as well as their immediate surroundings and by the N-terminal helix.[37,50] Interestingly, the ω angle between Cys14 and Lys15 includes a component on the high μs time scale even though the peptide bond does not isomerize during the runs (Figure S2). A cartoon illustration of BPTI highlighting the slowest residues is given in Figure 7c.

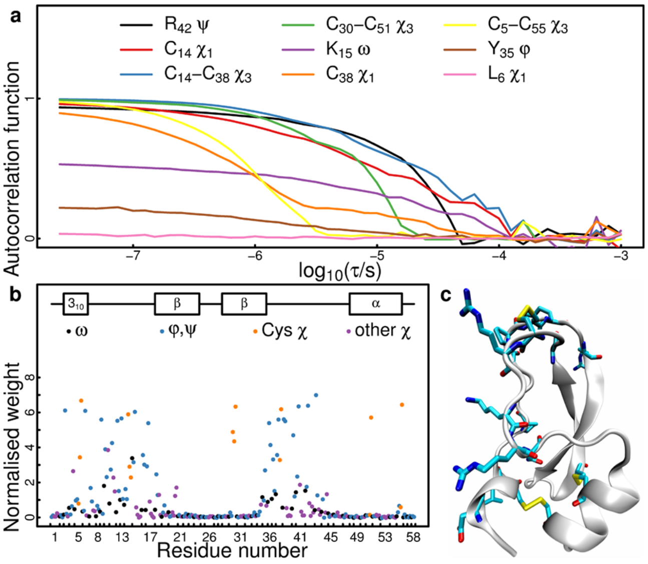


Figure 7. Autocorrelation functions and derived weights for BPTI. (a) Autocorrelation functions of selected dihedral angles. For each dihedral angle, the autocorrelation function was computed as the maximum of the autocorrelation functions of its sine and cosine values. (b) Weights $w_i = \max(R_i(1\ \mu\text{s}), 0)$ for 271 nonsymmetric dihedral angles including χ_2 and χ_3 angles of cysteines. Here, R_i is the autocorrelation function of the i^{th} dihedral angle as in (a). The weights were normalized such that their average is one and are ordered first by residue and then by type (ω, φ, ψ, χ_1, ..., χ_n) from left to right. Weights pertaining to χ_3 angles of disulfide bonds are assigned to the cysteine with lower residue number. The weights are colored according to the type of the corresponding dihedral angle. Secondary structure elements found in the crystal structure (PDB ID 5PTI)[51] are indicated on top. (c) Cartoon illustration of the crystal structure of BPTI. The residues having at least one dihedral angle with a normalized weight above 5 are shown in a stick-like representation. The illustration was rendered with VMD.[42]

Without weights (UW measure), we anticipate that a distance function based on dihedral angles is unable to reveal the conformational states of BPTI. Irrelevant features such as the χ_1 angle of Leu6 are expected to outweigh the impact of important features such as the ψ angle of Arg42 (Figure 7a). The SAPPHIRE plot[35] shown in Figure 8a confirms this prediction. The kinetic profile lacks significant barriers. With the help of the structural and sampling time annotations, 2 major and possibly 1 to 3 minor states can be identified. About 30% of the data seem to correspond to a heterogeneous ensemble. Figure 8b demonstrates that this interpretation is erroneous. The GW measure with $\tau = 1\ \mu$s allows the kinetic and time series annotations to unmask several metastable states that are structurally distinct. The notion of a heterogeneous state with rapid interconversion is lost. The most populated basin ranges from progress index values of about 1 to 16500. The second most populated basin, found at progress index values of about 24000 to 32500, contains those snapshots most similar to the crystal structure (PDB ID 5PTI).[51] Both major basins are observed directly in NMR experiments, albeit with different weights.[37,52]

The structural annotation in Figure 8b highlights that the two major states and the conformations located between progress index values of about 21500 and 24000 all exhibit different arrangements of the Cys14−Cys38 disulfide bond. In fact, this disulfide bond has been the focal point of several studies of the native state dynamics of BPTI.[41,52,53] In particular, Xue et al. used the same MD simulation data in order to improve interpretation of data from NMR relaxation dispersion measurements.[41] They defined conformational states based on the side chain dihedral angles of Cys14 and Cys38, thus neglecting all other degrees of freedom. Comparison with their state decomposition as shown in Figure 8b demonstrates that the conformational space of BPTI is captured surprisingly well with these dihedral angles alone. The five well-defined states do not overlap significantly. This annotation also highlights the poor performance of the UW measure as seen in Figure 8a. It is of course expected that states differing in other parts of the protein are likely to be missed by the manual feature selection of Xue et al. For example, the distinct basins between progress index values of about 18500 and 21500 would be annotated as either the most populated state (red) or as unclassified (gray) by Xue et al. Similarly, the small basin at progress index values of around 33500 would be annotated as crystal-like (blue), yet it differs from the crystal structure in the orientation of the Cys30−Cys51 disulfide bond. We note that the analysis of Shaw et al.[37] also failed to separate these minor states despite selecting features that are putatively sensitive to them (separate annotation in Figure 8).

In Figure 8c we study the same data set using the LAW measure ($\Delta = 1\ \mu$s, $\alpha = 1$). The resultant picture is very similar to the one based on the GW measure. We hypothesize that the

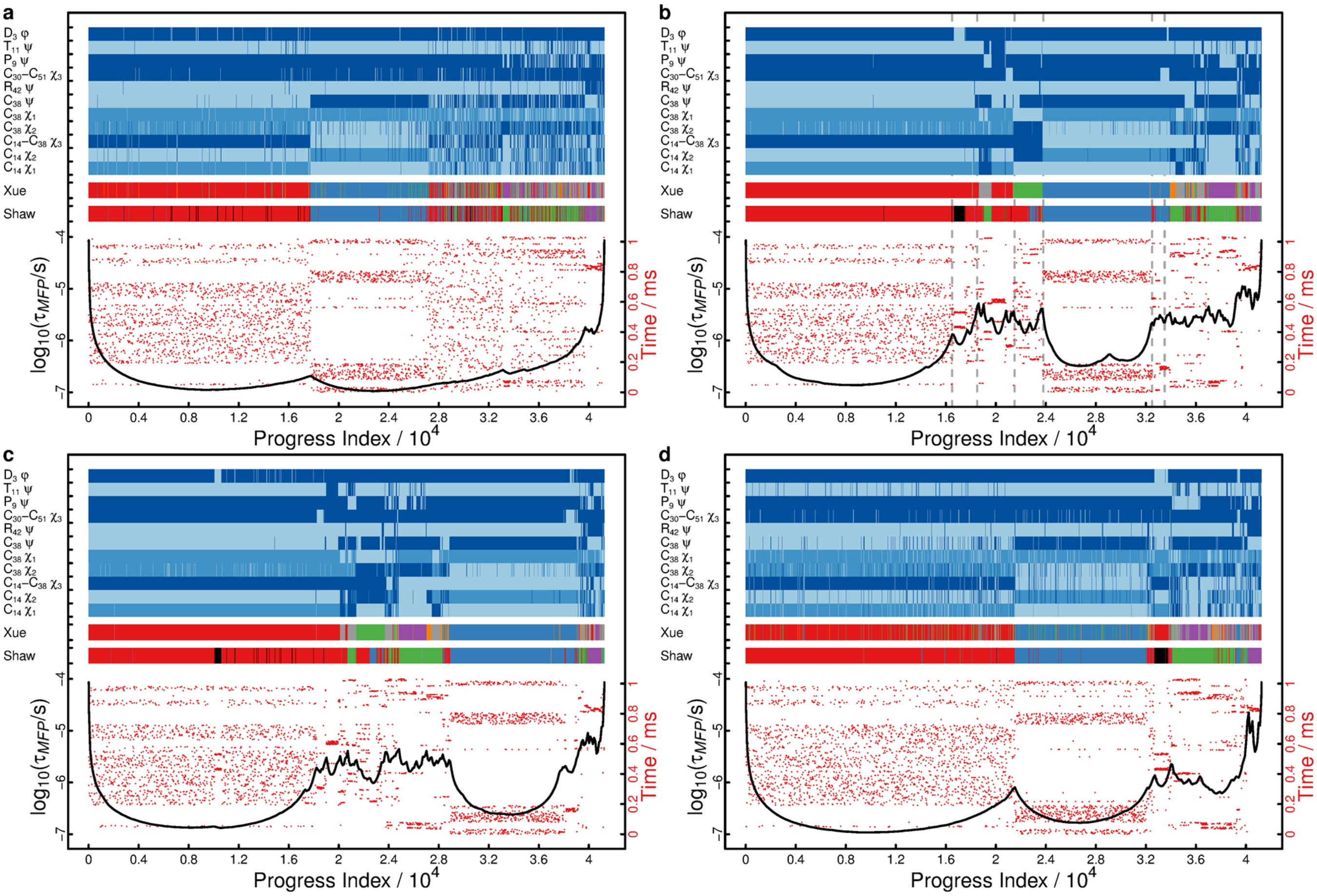


Figure 8. SAPPHIRE plots for folded BPTI. We use the same algorithm[34] as in Figure 6. (a) The (unweighted) Euclidean distance of the sine and cosine values of 271 dihedral angles (UW measure) is used to generate the progress index (x axis), which is annotated with kinetic information (black curve), sampling time (red dots), and structural information (color annotation on top). We extend this SAPPHIRE plot[35] by color-coded state assignments according to Shaw et al.[37] (red, blue, green, magenta, and black for states 0 to 5) and Xue et al.[41] (M1 - blue, M2 - orange, M3 - magenta, m_{C14} - red, m_{C38} - green, and other states - gray). The color-coded structural information uses binning of selected dihedral angles for clarity (see Methods). All annotations except the kinetic one are plotted every 4th snapshot to keep the size of the original vector image manageable. (b) The same as (a) for the GW measure with $\tau = 1$ μs. Dashed, gray lines indicate features of the plot discussed in the text. (c) The same as (a) for the LAW measure with $\Delta = 1$ μs and $\alpha = 1$. (d) The same as (a) using the RMSD of all 699 nonsymmetric atoms as the distance function.

relevance of individual features stays roughly the same throughout the conformational space sampled, which is composed of metastable states with high mutual similarity. Consequently, a locally adaptive distance function is not essential for this particular system. Figure 8d demonstrates that a SAPPHIRE plot employing the RMSD of all nonsymmetric atoms as the distance functions captures most states. However, it does not capture the m_{C38} (green) state in the model of Xue et al., a conformation for which there is experimental evidence.[41,52]

An obvious question to ask regards the dependency of our approach on the time domain parameters, τ (GW) and Δ (LAW), and the accuracy of the derived weights. It is well-known that estimators of second moments or related quantities have poor convergence properties with the numbers of samples. For time correlation measures, this is exacerbated in cases where the raw data do not sample the span of the underlying distribution recurrently. Surprisingly, Figure S3 demonstrates that the results obtained with the GW measures are largely preserved even with radically different choices for the parameter τ. This indicates that the main benefit of the GW measure for BPTI lies in reducing the influence of fast dihedral angles (compare Figures 7a and S2). Conversely, it appears to be less important how the slow dihedral angles are weighted with respect to one another. This is critical since it means that an accurate estimation of the true autocorrelation function at fixed lag time is not needed, thus preserving applicability of the method to cases where sampling is still poor (several smaller states in both Figures 6 and 8 are visited only once as indicated by the time series annotations).

We can take this point further by considering a very slow backbone dihedral angle that undergoes no significant transition for a given finite data set. Due to slow modes not actually being sampled, the autocorrelation function at large enough lag time would in all likelihood be close to zero, thereby giving a very slow degree of freedom a negligible weight. However, this seemingly misleading result is beneficial for the analysis as it reduces the weight of a feature containing no useful information (lack of variance). This example illustrates that the weights in the GW measure are data-driven, *i.e.*, they respond meaningfully to the finite samples available and need not be informed by the true distribution in a hypothetical limit. The same argument can be extended to the LAW measure using the same example. The data-driven origin of weights also implies that simulations using biased Hamiltonians, *e.g.*, umbrella sampling,[54] can be analyzed in the same way as shown here. The caveat that the true dynamics are unlikely to be represented faithfully by the data does not concern the analysis *per se*. Other simulation approaches yield ensembles of short trajectories at a given condition, *e.g.*, the replica exchange method.[55] Trajectory ensembles mean that limited amounts of data are available for inferring the time correlations underlying the GW and LAW measures, which is exacerbated for large lag times (τ) and window sizes (Δ). We are currently investigating the use of these measures with small values for τ and Δ in the context of a recent trajectory ensemble sampling method.[56]

To summarize, Figure 8 provides evidence that the weighted GW and LAW measures provide a richer picture of the conformational space of BPTI than two reference approaches, *viz.*, the use of (nearly) complete sets of features with equivalent weights for either dihedral angles or coordinate RMSD. We show that both of the latter bear the risk of lumping distinct states together. We characterize these states as distinct because they are structurally and kinetically homogeneous as highlighted by the annotations in Figure 8. An accurate definition of states is required to appropriately study state-dependent processes such as the exchange of the internal water molecules of BPTI.[57]

4. DISCUSSION

We propose to weight features in the evaluation of the distance between high-dimensional vectors, *e.g.*, dihedral angles recorded along MD trajectories. Specifically, the two approaches introduced are as follows. For the GW measure (eq 2), we globally weight (*i.e.*, scale) individual features according to the autocorrelation function at fixed lag time. For the LAW measure (eq 4), we count transitions across the global mean of a feature in a time-local window. Both approaches are designed to enhance the influence of slowly varying degrees of freedom. We have provided evidence that both the GW and LAW measures improve the quality of information that can be extracted from large sets of MD snapshots. The weighted distance functions have been tested on a 9-dimensional model system and on two data sets from MD simulations in conjunction with different unsupervised learning methods. The feature weighting method has unmasked slow dynamics of side chain packing within the native state of Beta3S (Figure 6) and revealed metastable conformations of BPTI that were not resolved in previous analyses (Figure 8).

A significant advantage of our method is that it is predominantly data-driven, *i.e.*, little prior knowledge about the system is needed, and potential sources of human bias are eliminated. The weighted distance functions reduce the impact of features lacking or failing to sample slow modes. The weights do not correct for heterogeneous variances and cross-correlation effects. The use of dihedral angles may be advantageous in both regards. The GW measure is expected to define a metric space offering increased contrast between similar and dissimilar data points for data of this type. The same holds for the LAW measure with the caveat that the rigorous notion of a metric is lost (see Methods). The method is easy to implement, and the weights can be computed in linear time with respect to the number of snapshots. Evaluating the resulting distance function scales linearly with the dimensionality of the data. For the LAW measure, the major limitation is given by the saving frequency, which has to be high enough to resolve state-specific fluctuations over a time window that does not exceed lifetimes of the states of interest. This limitation is a result of using time (rather than geometric) locality to derive locally adaptive weights.

Related work on distance learning for MD data ranges from manual and application-specific feature selection to defining new functional forms[58,59] and modifying classical methods for dimensionality reduction. Among the latter, sketch-map is a version of multidimensional scaling[33] focusing on matching intermediate distances.[29] If the distance function used in the original high-dimensional space lacks contrast, such an intermediate distance separating similar from dissimilar data points cannot be defined. Locally scaled diffusion map is an extension of diffusion map[30,31,43] using a Gaussian kernel with data point-dependent local scales.[32] At present, both methods give equal weight to all the original features no matter how noisy or irrelevant they are, which might reduce their effectiveness in capturing the kinetics of the system if no additional feature selection is performed.

The so-called time structure-based independent component analysis (TICA)[60,61] provides a linear (but not orthonormal) transformation of the input data to yield components with maximal autocorrelation function at a given lag time. This method is conceptually similar to our GW measure, and we provide a direct comparison in Figure 9 for an attempted embedding in 2 dimensions. Comparison of Figure 9a with Figure 9b highlights that standard signal processing tools such as variance normalization can hurt rather than help the resolution of a projection based on principal component analysis (PCA). The GW measure is representable in PCA by scaling the input features according to factors of $\sqrt{w_i}$ (eq 2, τ =

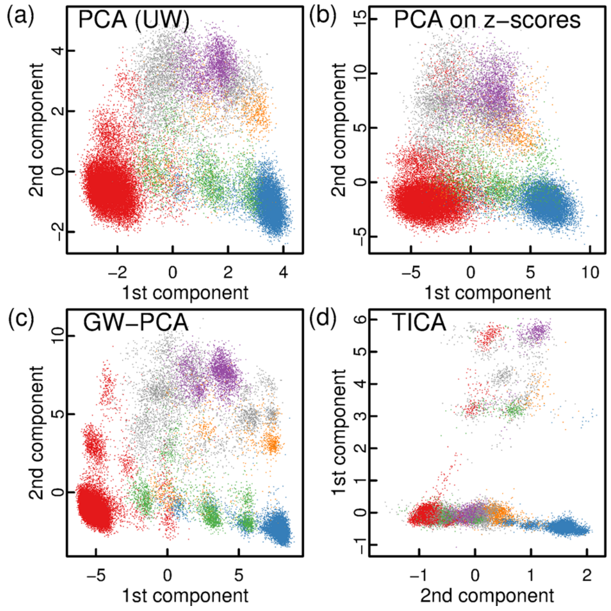


Figure 9. Two-dimensional embeddings for folded BPTI. All data points in all panels are colored according to the model introduced by Xue et al.,[41] *i.e.*, M1 - blue, M2 - orange, M3 - magenta, m_{C14} - red, m_{C38} - green, and other states - gray. (a) Projection of the data (sine and cosine values of 271 nonsymmetric dihedral angles including χ_2 and χ_3 angles of cysteines) onto the first two principal components without prior scaling of the input features. (b) The same as (a) for input features scaled to have unit variance. (c) The same as (a) for GW-PCA and a lag time of τ = 1 μs. (d) Projection of the same data onto the first two time structure-based independent components using a lag time of τ = 1 μs. Note that components are swapped to highlight similarity to other panels.

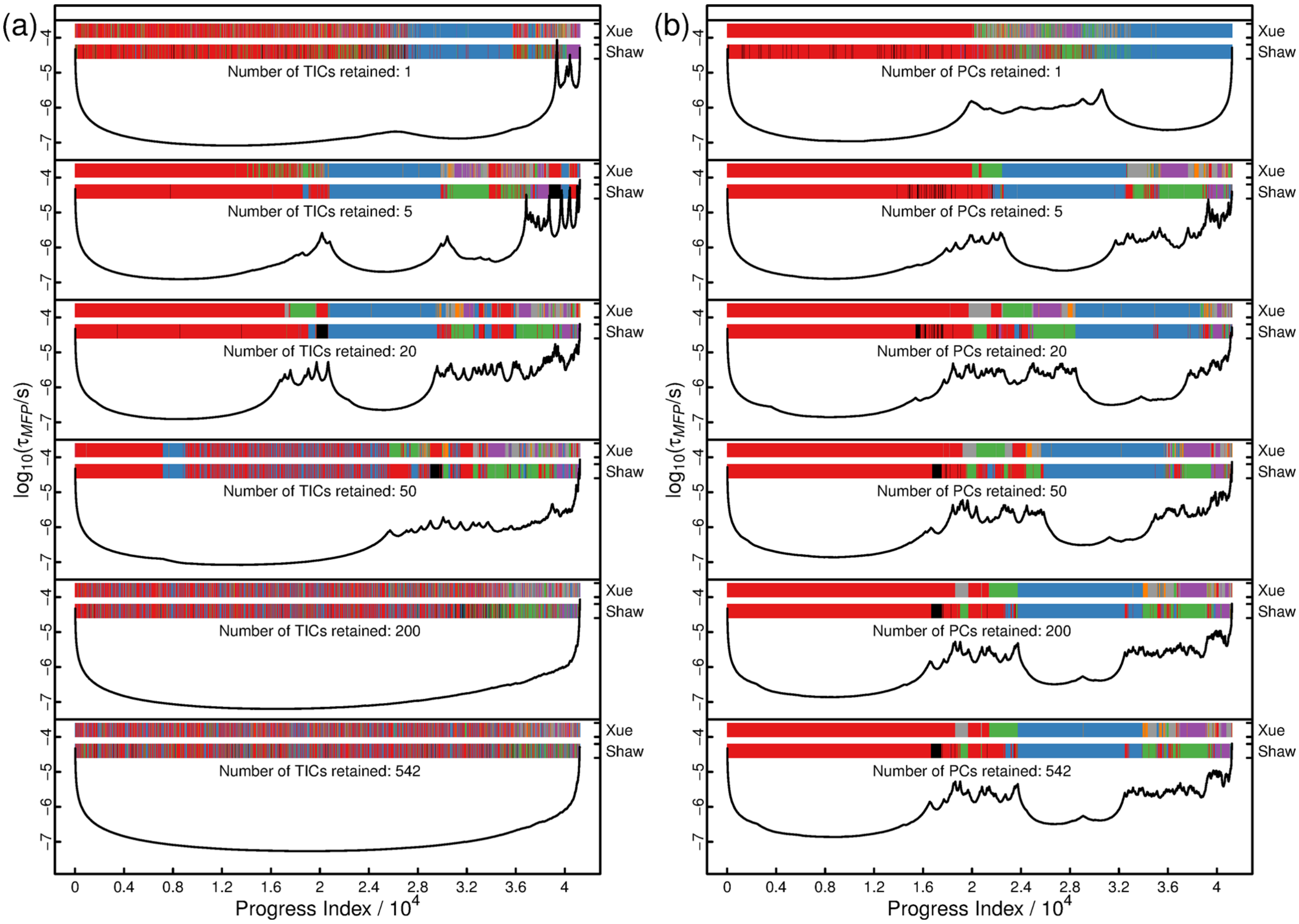


Figure 10. Comparison of TICA and GW-PCA for BPTI. Simplified SAPPHIRE plots are shown as in Figure 8 with only two annotations and the kinetic cut function plotted. (a) TICA eigenvectors and eigenvalues were computed for the raw data (Euclidean distance of the sine and cosine values of 271 dihedral angles) using a lag time of $\tau = 1$ μs. Components were ordered by eigenvalue (value of the autocorrelation function). Data were then transformed and different numbers of those features with the largest eigenvalues were retained as indicated. (b) The same for GW-PCA. PCA was applied to data scaled by the global weights as defined for the GW measure and a lag time of $\tau = 1$ μs. When retaining all 542 features, the progress index becomes identical to that in Figure 8b because PCA yields an orthonormal transformation.

1 μs). The resultant GW-PCA approach (Figure 9c) separates the data much better, and numerous states emerge. In contrast, the TICA approach (Figure 9d)[61] appears to overemphasize a particular slow mode leading to excellent separation of a small subpopulation but dramatic overlap of everything else. As an additional point, Figure 9 emphasizes that the usefulness of the weights is not specific to the analysis methods employed in Figures 2, 3, 6, and 8.

To test this result further, we recomputed the progress index for BPTI using both GW-PCA and TICA with a wide range of retained dimensionalities. It emerges that an appropriate choice of dimensionality is critical in TICA but not in GW-PCA (Figure 10). The best-performing TICA case retains 20 features, which correlates well with the eigenvalue spectrum (not shown). However, even this case is not obviously adding to the information provided by GW-PCA, which appears robust for most of the tested dimensionalities. Importantly, the full-dimensional GW-PCA case is equivalent to the original GW measure in Figure 8b and performs as well or better than any TICA example shown. We note that TICA fails dramatically at high dimensionality and provides even less information than the UW measure (Figure 8a).

Based on Figure 10, it is clear why recent TICA applications resort to a low-dimensional embedding of the transformed data.[38,39] A common limitation of TICA and GW-PCA is their inherent linearity although kernel-based extensions might capture nonlinear structure in the data.[62] A method similar to the LAW measure is that by Singer et al.[63,64] who propose the semimetric

$$d(\mathbf{x}(t_k), \mathbf{x}(t_l)) = (\mathbf{x}(t_k) + \mathbf{x}(t_l))^T (\Sigma_k^{-1} + \Sigma_l^{-1})(\mathbf{x}(t_k) + \mathbf{x}(t_l))$$

where Σ_k is a local covariance matrix associated with $\mathbf{x}(t_k)$. It can be determined by running short stochastic simulations starting from $\mathbf{x}(t_k)$,[63,64] which is not feasible for the large MD data sets considered in the present study, or from the data within a short time window along the trajectory around $\mathbf{x}(t_k)$,[65] similar to what we have proposed here (eqs 3 and 4). Other approaches directly take advantage of kinetic information to determine relevant features before applying any unsupervised learning protocol. The method of McGibbon and Pande learns a distance function that tends to return low values for pairs of data points that are close in time and large values for those pairs that are far in time along the trajectory.[66] This task is formulated as a complex optimization problem depending on several parameters. When the approach was applied to MD data of the protein Fip35, side chain and peptide bond dihedral angles were discarded *a priori*, suggesting that manual feature selection is still needed.

All the algorithms used here have been implemented in the free software package CAMPARI (http://campari.sourceforge.net), and the current development version is available upon request (campari.software@gmail.com). Ongoing work is focused on combining the framework of weighted distances with the RMSD metric in order to study processes involving multiple molecules such as ligand binding to receptors. However, external motion complicates the definition of weights for atoms.

In conclusion, we have developed a data-driven method for feature weighting to improve the contrast of distance functions. Our method reveals metastable states in the reversible folding of Beta3s and the native state of BPTI, which were not resolved in previous studies of the same data sets.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00618.

> Supporting Methods with values for auxiliary parameters and further data on Beta3S (Figure S1) and BPTI (Figures S2 and S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: a.vitalis@bioc.uzh.ch.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank David E. Shaw for sharing the trajectory data and the state annotation for BPTI (color code in Figures 8 and 10). This work was supported by a grant from the Swiss National Science Foundation to A.C. and a fellowship from the Holcim Stiftung Wissen to A.V.

■ REFERENCES

(1) Clarke, R.; Ressom, H. W.; Wang, A.; Xuan, J.; Liu, M. C.; Gehan, E. A.; Wang, Y. The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nat. Rev. Cancer* **2008**, *8*, 37−49.

(2) Allison, D. B.; Cui, X.; Page, G. P.; Sabripour, M. Microarray data analysis: From disarray to consolidation and consensus. *Nat. Rev. Genet.* **2006**, *7*, 55−65.

(3) Hilario, M.; Kalousis, A. Approaches to dimensionality reduction in proteomic biomarker studies. *Briefings Bioinf.* **2008**, *9*, 102−118.

(4) Bhat, P. C. Multivariate analysis methods in particle physics. *Annu. Rev. Nucl. Part. Sci.* **2011**, *61*, 281−309.

(5) Beyer, K. S.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When is "nearest neighbor" meaningful? In *ICDT '99, Proceedings of the 7th International Conference on Database Theory*, Jerusalem, Israel, January 10−12, 1999; Beeri, C., Buneman, P., Eds.; Springer: Berlin, 1999; pp 217−235.

(6) Hinneburg, A.; Aggarwal, C. C.; Keim, D. A. What is the nearest neighbor in high dimensional spaces? In *VLDB 2000, Proceedings of the 26th International Conference on Very Large Data Bases*, Cairo, Egypt, September 10−14, 2000; El Abbadi, A., Brodie, M. L., Chakravarthy, S., Dayal, U., Kamel, N., Schlageter, G., Whang, K.-Y., Eds.; Morgan Kaufmann: Orlando, USA, 2000; pp 506−515.

(7) Kriegel, H. P.; Kröger, P.; Zimek, A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discovery Data* **2009**, *3*, 1.

(8) Domeniconi, C.; Gunopulos, D.; Ma, S.; Yan, B.; Al-Razgan, M.; Papadopoulos, D. Locally adaptive metrics for clustering high dimensional data. *Data Min. Knowl. Discovery* **2007**, *14*, 63−97.

(9) Domeniconi, C.; Papadopoulos, D.; Gunopulos, D.; Ma, S. Subspace clustering of high dimensional data. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, Lake Buena Vista, FL, USA, April 22−24, 2004; Berry, M. W., Dayal, U., Kamath, C., Skillicorn, D. B., Eds.; SIAM: 2004; pp 517−521.

(10) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, 2009; pp 485−698.

(11) Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651−666.

(12) Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78−87.

(13) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.

(14) Jolliffe, I. T. *Principal Component Analysis*; Springer: New York, 2002; pp 1−405.

(15) Xu, R.; Damelin, S.; Nadler, B.; Wunsch, D. C. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. *Artif. Intell. Med.* **2010**, *48*, 91−98.

(16) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064−4092.

(17) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in protein-folding simulations. *Nat. Phys.* **2010**, *6*, 751−758.

(18) Buchner, G. S.; Murphy, R. D.; Buchete, N. V.; Kubelka, J. Dynamics of protein folding: Probing the kinetic network of folding-unfolding transitions with experiment and theory. *Biochim. Biophys. Acta, Proteins Proteomics* **2011**, *1814*, 1001−1020.

(19) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58−65.

(20) Vitalis, A.; Caflisch, A. Efficient construction of mesostate networks from molecular dynamics trajectories. *J. Chem. Theory Comput.* **2012**, *8*, 1108−1120.

(21) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495−1517.

(22) Rao, F.; Caflisch, A. The protein folding network. *J. Mol. Biol.* **2004**, *342*, 299−306.

(23) Krivov, S. V.; Karplus, M. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B* **2006**, *110*, 12689−12698.

(24) Noé, F.; Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.

(25) Buchete, N. V.; Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.

(26) Mu, Y.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 45−52.

(27) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319−2323.

(28) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 9885−9890.

(29) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023−13028.

(30) Coifman, R. R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5−30.

(31) Coifman, R. R.; Kevrekidis, I. G.; Lafon, S.; Maggioni, M.; Nadler, B. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Model. Simul.* **2008**, *7*, 842−864.

(32) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.* **2011**, *134*, 124116.

(33) Cox, M. A. A.; Cox, T. F. Multidimensional scaling. In *Handbook of Data Visualization*; Chen, C.-h., Härdle, W. K., Unwin, A., Eds.; Springer: Berlin, 2008; pp 315−347.

(34) Blöchliger, N.; Vitalis, A.; Caflisch, A. A scalable algorithm to order and annotate continuous observations reveals the metastable states visited by dynamical systems. *Comput. Phys. Commun.* **2013**, *184*, 2446−2453.

(35) Blöchliger, N.; Vitalis, A.; Caflisch, A. High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations. *Sci. Rep.* **2014**, *4*, 6264.

(36) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: New insights into the folding process. *J. Phys. Chem. B* **2008**, *112*, 8701−8714.

(37) 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.

(38) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.

(39) 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.

(40) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

(41) Xue, Y.; Ward, J. M.; Yuwen, T.; Podkorytov, I. S.; Skrynnikov, N. R. Microsecond time-scale conformational exchange in proteins: Using long molecular dynamics trajectory to simulate NMR relaxation dispersion data. *J. Am. Chem. Soc.* **2012**, *134*, 2555−2562.

(42) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(43) Nadler, B.; Lafon, S.; Coifman, R. R.; Kevrekidis, I. G. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 113−127.

(44) Xue, Y.; Ludovice, P. J.; Grover, M. A.; Nedialkova, L. V.; Dsilva, C. J.; Kevrekidis, I. G. State reduction in molecular simulations. *Comput. Chem. Eng.* **2013**, *51*, 102−110.

(45) Huang, D.; Caflisch, A. Evolutionary conserved Tyr169 stabilizes the $\beta 2$-$\alpha 2$ loop of the prion protein. *J. Am. Chem. Soc.* **2015**, *137*, 2948−2957.

(46) Blöchliger, N.; Xu, M.; Caflisch, A. Peptide binding to a PDZ domain by electrostatic steering via nonnative salt bridges. *Biophys. J.* **2015**, *108*, 2362−2370.

(47) Qi, B.; Muff, S.; Caflisch, A.; Dinner, A. R. Extracting physically intuitive reaction coordinates from transition networks of a β-sheet miniprotein. *J. Phys. Chem. B* **2010**, *114*, 6979−6989.

(48) Zheng, W.; Qi, B.; Rohrdanz, M. A.; Caflisch, A.; Dinner, A. R.; Clementi, C. Delineation of folding pathways of a β-sheet miniprotein. *J. Phys. Chem. B* **2011**, *115*, 13065−13074.

(49) Kalgin, I. V.; Caflisch, A.; Chekmarev, S. F.; Karplus, M. New insights into the folding of a β-sheet miniprotein in a reduced space of collective hydrogen bond variables: Application to a hydrodynamic analysis of the folding flow. *J. Phys. Chem. B* **2013**, *117*, 6092−6105.

(50) Long, D.; Brüschweiler, R. Atomistic kinetic model for population shift and allostery in biomolecules. *J. Am. Chem. Soc.* **2011**, *133*, 18999−19005.

(51) Wlodawer, A.; Walter, J.; Huber, R.; Sjölin, L. Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. *J. Mol. Biol.* **1984**, *180*, 301−329.

(52) Grey, M. J.; Wang, C.; Palmer, A. G., III Disulfide bond isomerization in basic pancreatic trypsin inhibitor: Multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling. *J. Am. Chem. Soc.* **2003**, *125*, 14324−14335.

(53) Otting, G.; Liepinsh, E.; Wüthrich, K. Disulfide bond isomerization in BPTI and BPTI(G36S): An NMR study of correlated mobility in proteins. *Biochemistry* **1993**, *32*, 3571−3582.

(54) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(55) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(56) Bacci, M.; Vitalis, A.; Caflisch, A. A molecular simulation protocol to avoid sampling redundancy and discover new states. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850*, 889−902.

(57) Persson, F.; Halle, B. Transient access to the protein interior: Simulation versus NMR. *J. Am. Chem. Soc.* **2013**, *135*, 8735−8748.

(58) 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.

(59) Zhou, T.; Caflisch, A. Distribution of reciprocal of interatomic distances: A fast structural metric. *J. Chem. Theory Comput.* **2012**, *8*, 2930−2937.

(60) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634−3637.

(61) Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.* **2011**, *134*, 065101.

(62) Schwantes, C. R.; Pande, V. S. Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.* **2015**, *11*, 600−608.

(63) Singer, A.; Coifman, R. R. Non-linear independent component analysis with diffusion maps. *Appl. Comput. Harmon. Anal.* **2008**, *25*, 226−239.

(64) Singer, A.; Erban, R.; Kevrekidis, I. G.; Coifman, R. R. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 16090−16095.

(65) Dsilva, C. J.; Talmon, R.; Rabin, N.; Coifman, R. R.; Kevrekidis, I. G. Nonlinear intrinsic variables and state reconstruction in multiscale simulations. *J. Chem. Phys.* **2013**, *139*, 184109.

(66) McGibbon, R. T.; Pande, V. S. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 2900−2906.

DOI: 10.1021/acs.jctc.5b00618