JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

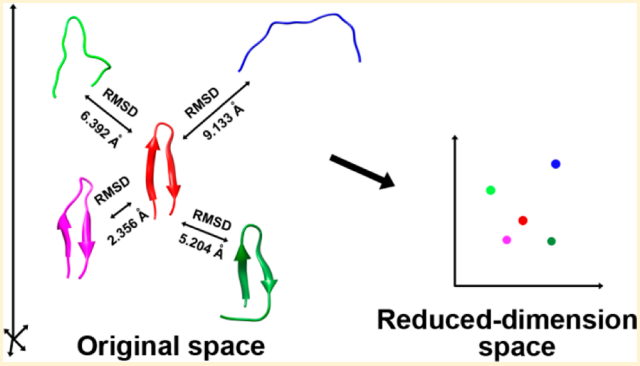
Evaluation of Dimensionality-Reduction Methods from Peptide Folding−Unfolding Simulations

Mojie Duan,[†] Jue Fan,[†] Minghai Li,[†] Li Han,*[,‡] and Shuanghong Huo*[,†]

[†]Gustaf H. Carlson School of Chemistry and Biochemistry and [‡]Department of Math and Computer Science, Clark University, Worcester, Massachusetts 01610, United States

Supporting Information

ABSTRACT: Dimensionality-reduction methods have been widely used to study the free energy landscapes and low-free-energy pathways of molecular systems. It was shown that the nonlinear dimensionality-reduction methods gave better embedding results than the linear methods, such as principal component analysis, in some simple systems. In this study, we have evaluated several nonlinear methods, locally linear embedding, Isomap, and diffusion maps, as well as principal component analysis from the equilibrium folding/unfolding trajectory of the second β-hairpin of the B1 domain of streptococcal protein G. The CHARMM parm19 polar hydrogen potential function was used. A series of criteria which reflects different aspects of the embedding qualities was employed in the evaluation. Our results show that principal component analysis is not worse than the nonlinear ones on this complex system. There is no clear winner in all aspects of the evaluation. Each dimensionality-reduction method has its limitations in a certain aspect. We emphasize that a fair, informative assessment of an embedding result requires a combination of multiple evaluation criteria rather than any single one. Caution should be used when dimensionality-reduction methods are employed, especially when only a few of the top embedding dimensions are used to describe the free energy landscape.



INTRODUCTION

The thermodynamics and kinetics of protein conformational changes and folding are determined by the free energy landscape of the system. For a given protein with N atoms, its degree of freedom is $3N$; however, the effective dimensionality of the free energy landscape is believed to be relatively small. The correlations/anticorrelations between degrees of freedom were found in the folding/unfolding processes.[1] It was also found that only a few degrees of freedom in a subspace are essential to global conformational changes,[2] but it is challenging to systematically choose the effective low-dimension coordinates. To describe the folding process, it is common to project the free energy surface onto some progress variables, such as the number of native contacts, radius of gyration, or root-mean-square deviation (RMSD). Nevertheless, the results are likely to lack the complexity compared to those in the original high dimensions, even when multiple projections onto different progress variables are used.[3] Thus, it is crucial to systematically search for intrinsic order parameters (or reaction coordinates) to characterize the progress of reaction that is the most relevant to the global conformational changes and separation of states.

In math, the mapping of the free energy landscape onto a few order parameters is a process of dimensionality reduction. Generally, a good dimensionality-reduction method should preserve the main features of the free energy landscape in the original dimensions, such as the number of low free energy minima and the relationships between them. Many dimensionality-reduction methods have been employed to characterize protein free energy landscape and conformational changes, such as principal component analysis (PCA),[4] local linear embedding (LLE),[5] Isomap,[6] and diffusion maps.[7] Broadly, dimensionality reduction can be classified into linear and nonlinear methods. PCA is a widely used linear method. The linear transformation matrix contains a set of orthogonal vectors which are called principal components and are ranked in the order of variance. The principal components associated with large variances capture the global conformational changes of a protein, while those associated with small variances correspond to fast motions which are irrelevant to the global conformational changes. It was shown that PCA can effectively describe large scale collective motions.[8] However, as a linear transformation method, PCA is well suited for the data points that lie on or close to a hyper-plane in the original space. For biomolecular systems, it is believed that the conformations lie on a curved hyper-surface, rather than a hyper-plane, of low dimensions ($d \ll 3N$) because of the complexity of the systems.[9]

Received: January 21, 2013
Published: March 11, 2013

 © 2013 American Chemical Society

It was shown that nonlinear dimensionality-reduction methods gave better embedding results than PCA.[10] LLE, Isomap, and diffusion maps all belong to the nonlinear dimensionality-reduction category. Unlike PCA, LLE approximates the conformations of a neighborhood, rather than the whole conformational ensemble, lying on a hyper-plane. Imagine that each neighborhood is a small patch in the high-dimension space. Conceptually, it is safe to approximate a small patch of the conformation space as a hyper-plane. The main ideas underlying LLE are as follows: (a) the characteristic of local geometry of each neighborhood in the high-dimension space is expected to be invariant after dimensionality reduction; (b) well preserved local geometry will lead to well preserved global geometry. LLE uses all the overlapping local hyper-planes to approximate the hyper-surface of the original conformation space and then searches for the optimal embedding in a lower dimensional space that preserves the geometry of these local hyper-planes and their relationship. It was shown that LLE is useful in the discovery of the conversion pathways of Trp cage (TC5b) from the mesostates to the native state.[11] Unlike LLE, Isomap utilizes global information that contains pairwise "distances" between all protein conformations on a graph. Here, a graph or a network is used as a discrete model of the conformation space, and the distance between two conformations is the shortest path between them on the graph/network. This "distance" is called the geodesic distance, which gives an indication of the easiness of conversion from one conformation to the other. The goal of Isomap is to preserve the geodesic distances after embedding. This technique has been applied to coarse-grained and all-atom molecular models to characterize the free energy landscape and pathways of conformational changes.[9,10,12] Instead of preserving the intrinsic geometry as Isomap and LLE, diffusion maps intend to preserve the dynamic proximity in the low dimensions.[13] The Euclidean distance between two conformations in the embedded dimensions reflects the ease with which the system can evolve from one conformation to the other. Diffusion maps have been used to extract global order parameters that describe the fundamental dynamical motions of the antimicrobial peptide microcin J25.[14] A variant of diffusion maps was employed to identify the folding pathways of a β-sheet mini-protein and the free energy landscape of the coarse-grained SH3 model.[15] Nonlinear dimensionality reduction is also useful as a guide for biased sampling.[16]

Even with the successful applications of the nonlinear dimensionality-reduction methods in proteins and other molecular systems,[9−11,14] it is important to note that information in the effective dimension is useful only if the embedding truthfully reflects the properties of the system in the original dimensions. In this study, we evaluate a series of the dimensionality-reduction methods using the equilibrium folding/unfolding trajectory of the second β-hairpin of the B1 domain of streptococcal protein G (GB1). The adequate sampling allows this model system to be a benchmark for evaluation of any new dimensionality-reduction method. For the evaluation of embedding results, we present several criteria that reflect different aspects of the embedding quality and show that it is important to combine these criteria for evaluation instead of relying on any single one of them.

METHODS

Molecular Dynamics (MD) Simulation and the Transition Disconnectivity Graph. A 4-μs equilibrium folding−unfolding MD trajectory of the β-hairpin was collected at 360 K using CHARMM.[17] The parm19 polar hydrogen potential function[18] and EEF1 implicit solvation model[19] were employed. The conformations were saved every 20 ps, and 200 000 conformations in total were collected. We used the all-atom RMSD of 3 Å as the criterion to generate a neighbor list for each conformation. Hereafter, RMSD refers to all-atom RMSD. Then the conformation with the largest number of neighbors was identified and assigned along with all its neighbors to the first cluster. All the conformations of this cluster were then eliminated from the ensemble of conformations, and the neighbor list of all remaining conformations was updated. This process was repeated until the ensemble was empty. In this way, a series of nonoverlapping clusters of conformations is obtained.

The free energy of cluster i was computed as

$$F_i = -k_B T \ln(N_i) + c \tag{1}$$

where N_i is the number of conformations within this cluster, k_B is the Boltzmann constant, and c denotes a constant that is related to the total number of MD snapshots analyzed and that is inconsequential when free energy differences are the only concern. Cluster 1 has the lowest free energy. The free energy barrier between clusters i and j is calculated by

$$F_{ij} = -k_B T \ln Z_{ij} \tag{2}$$

where Z_{ij} is the partition function of the barrier and is related to the minimum cut value[20] (n_{ij}) by

$$Z_{ij} = \frac{1}{2} n_{ij} \times \frac{1}{k_B T} \times \frac{1}{\Delta t} \tag{3}$$

where h is the Planck's constant, T = 360 K, and Δt = 20 ps sampling interval. The factor of 1/2 in eq 3 is due to the definition of n_{ij} as the sum of numbers of transitions for $i \rightarrow j$ and $j \rightarrow i$. Note that the minimum cut value reflects the overall free energy barriers between clusters, including the contributions of all possible routes, direct or indirect. Thus, these free energy barriers are the apparent barriers that are comparable to experimental results, whereas the free energy barriers from the direct route alone cannot be compared to the experimental data. Using the information of free energy of each cluster and the barrier between the pairs of clusters, the transition disconnectivity graph (TRDG) was constructed. Using TRDG to describe the free energy landscape was first developed by Krivov and Karplus.[3] Our TRDG is very similar to that in ref 21, but with a different RMSD threshold (3 Å in this study).

Principal Component Analysis (PCA). If there are m snapshots for a protein that contains N atoms, the original matrix $\mathbf{X}$ is a $3N \times m$ matrix. The linear transformation matrix $\mathbf{P}$ transforms $\mathbf{X}$ to $\mathbf{Y}$ by $\mathbf{Y} = \mathbf{PX}$. The principal components are the eigenvectors of $(1/m)\mathbf{XX}^T$ which form the rows of $\mathbf{P}$. For the β-hairpin, $3N$ = 480 and m = 200 000 snapshots. We used the *ptraj* module in AmberTools (v9.0) to carry out the PCA. All conformations were superimposed to the center of cluster 1 to remove translation and rotation before the PCA analysis.

Locally Linear Embedding (LLE). In the conformation space, each conformation is represented by a vector $\vec{x}_i$ with $3N$ dimensions. Using a RMSD cutoff = 3.0 Å as a neighboring criterion, only the conformations that have greater than or equal to 20 neighbors were embedded, for a total of 179 629 conformations. The local geometry of each neighborhood is

characterized by the linear coefficients (w_{ij}) that are used to reconstruct each conformation from its neighboring conformations. The reconstruction errors are measured by $\varepsilon(w) = \sum_i \|\vec{x}_i - \sum_j w_{ij}\vec{x}_j\|^2$. The optimal set of reconstruction weights in the high dimensions was then used to reconstruct $\vec{y}_i$ in the low dimensions by choosing $\vec{y}_i$ to minimize the objective function, $\Phi = \sum_i |\vec{y}_i - \sum_j w_{ij}\vec{y}_j|^2$. Thereby, the intrinsic geometry of each conformation's neighborhood in the original space is preserved in the low dimensions.

Isomap. Isomap intends to preserve the pairwise geodesic distances between conformations on a graph. The MD generated conformations are nodes on the graph. The k (k = 20) nearest neighbors of each conformation were identified using the RMSD criterion. If the RMSD between a particular conformation and any of its nearest neighbors is greater than 3 Å, then this neighbor will be removed from the neighbor list. Each pair of neighboring conformations is connected by an edge whose weight was assigned to be equal to the RMSD between the pair of conformations, resulting in up to 20 edges for each conformation. Given two conformations on a graph, a path between them is a sequence of edges connecting these two conformations, and the path length is defined as the sum of the edge weights. The length of the shortest path is defined as the geodesic distance between these two conformations. The distance matrix, $\mathbf{D}$, contains all the pairwise geodesic distances. In the space of low dimension ($\mathbf{Y}$), the new set of coordinates $\vec{y}_i$ is chosen to minimize the objective function,$\Phi = \|\boldsymbol{\tau}(\mathbf{D}) - \boldsymbol{\tau}(\mathbf{D}_\mathbf{Y})\|$. The $\boldsymbol{\tau}$ matrix is equal to $-\mathbf{HSH}/2$ where $S_{ij} = D_{ij}^2$ and $H_{ij} = \delta_{ij} - (1/m)$, where δ_{ij} is the Kronecker delta and m is the number of conformations.

The calculation of all the pairwise geodesic distances is prohibitively expensive for a large data set. To reduce the computational cost, the implementation of Isomap actually preserves the geodesic distances between each conformation and landmark conformations. When the numbers of landmarks are substantially less than the number of conformations but sufficiently greater than the essential dimensions, preserving the geodesic distances to the landmarks is virtually preserving the geodesic distances between all the conformations. A recent adaptation of Isomap[9] further reduces the computational cost by reinsertion; specifically, the authors use only a fraction of conformations for embedding and then linearly reinsert the rest of the conformations into the low dimensions based on their neighborhood relations to the embedded point in the original dimensions. We adopted the landmark-based approach, but we did not use the reinsertion. Along the 4-μs trajectory, we chose a conformation every 800 ps as a landmark, resulting in a total of 5000 landmarks. The procedure of connected-components[22] was used to find the largest connected component out of the 200 000 conformations (or nodes). The conformations that do not belong to the largest connected component were removed. As a result, 179 774 conformations and 4491 landmarks remained and were used in the Isomap calculation.

Diffusion Maps. To approximate the dynamic proximity between all pairs of conformations, the structural similarity metric (pairwise RMSD) was employed. The element of matrix $\mathbf{A}$ is

$$A_{ij} = \exp\left(-\frac{(\mathrm{RMSD}_{ij})^2}{2\varepsilon}\right) \quad i, j = 1, 2, \ldots m \tag{4}$$

where m is the number of snapshots used in the calculation and ε = 10 (Figure S1 shows how the value of ε was chosen). The $\mathbf{M}$ matrix is defined as $\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}$, where the $\mathbf{D}$ matrix is diagonal with $D_{ii} = \sum_{j=1}^{m} A_{ij}$, i = 1, 2, ...m. The element of $\mathbf{M}$ (M_{ij}) was proposed to represent the transition probability from conformation i to conformation j in a finite time step corresponding to ε in eq 4. The eigenvectors of $\mathbf{M}$ are arranged in decreasing order of their corresponding eigenvalues. The first eigenvector is a trivial all-one vector and is not used as an embedding dimension. The top n nontrivial eigenvectors form the n embedding dimensions. It was shown that the diffusion distance

$$\sqrt{\sum_{k=1}^{m} \frac{(M_{ik} - M_{jk})^2}{D_{kk}}} \tag{5}$$

between conformation i and j in the original space was identical to the Euclidean distance in the reduced-dimension space when all the nontrivial eigenvectors were considered. To reduce the computational cost, we used 40 000 snapshots which were chosen every 100 ps from the 4-μs trajectory. After embedding, the rest of the conformations were inserted back to the low-dimension space according to the geometric relationship between these conformations and their neighbors.[9]

Evaluation of Embedding Results. In this paper, we evaluate the results of dimensionality reduction from various aspects, including free energy profiles, the quality of neighborhood preserving, and residual variance. Free energy profiles provide an intuitive, visual inspection of the embedding results but are generally limited to two and three dimensions. We used 2D grids to describe the free energy profile as a function of the first two embedding dimensions. The number of conformations in each grid was counted, and the free energy corresponds to each grid calculated by the same way as that of a cluster (eq 1). Since the free energy surface obtained by TRDG is unprojected, we use it as a "gold" standard to evaluate the projection of the free energy surface obtained by the dimensionality-reduction methods. The measure of neighborhood preserving quality and residual variance are numeric criteria and broadly applicable to different granularities (ranging from clusters to all conformations) and all dimensions.

The basic criterion for the neighborhood preserving is that nearby points in the reduced-dimension space after embedding are also nearby points in the original space, and vice versa. It is not necessary to evaluate every conformation and its neighborhood because for the conformations in the same cluster their neighborhoods overlap extensively. Therefore, we chose the centers of the clusters that have more than 50 conformations as representative conformations, and a RMSD cutoff of 3 Å was utilized to define the neighborhood for each of the conformations, resulting in 248 representative conformations (247 representative conformations for Isomap because the largest connected component contains only 247 such cluster centers). To evaluate the quality of neighborhood preserving, we first calculated the Euclidean distances between each representative conformation and every conformation in its neighborhood in the reduced-dimension space and found the largest distance. Then, we used one thousandth of the largest Euclidean distance as an initial cutoff in the reduced-dimension space as well as an incremental value. The cutoff in the reduced-dimension space defines a hyper-sphere with the representative conformation as its center and the cutoff as its radius. Then the evaluation of embedding results for the representative conformation and its neighborhood was conducted with respect to the sphere. We call such a sphere an evaluation sphere in the

following text. Finally, we employed sensitivity (Sn) and a positive predictive value (PV^+) for the evaluation. For the embedding of each representative conformation and its neighborhood, we first calculated Sn and PV^+ at the initial cutoff, then increased the cutoff by the incremental value, and recalculated Sn and PV^+ until the evaluation sphere reached its maximum to include the embedding of all the conformations in the neighborhood, for a total of 1000 cycles of Sn−PV^+ evaluation. For a given Euclidean distance cutoff with respect to a particular representative conformation, Sn is defined as

$$\mathrm{Sn} = \frac{\#\text{of true positive}}{\#\text{of true positive} + \#\text{of false negative}} \times 100\% \quad (6)$$

and the positive predictive value is defined as

$$\mathrm{PV}^+ = \frac{\#\text{of true positive}}{\#\text{of true positive} + \#\text{of false positive}} \times 100\% \quad (7)$$

If a conformation in the 3 Å RMSD neighborhood of the original space is embedded within an evaluation sphere of a given cutoff, then this case is called true positive. If it is embedded outside the evaluation sphere, it is called false negative. If a conformation outside the RMSD cutoff in the original space remains outside the evaluation sphere, then the case is called true negative. Otherwise, it is called false positive.

Given a representative conformation and a cutoff distance in the reduced-dimension space, Sn and PV^+ reflect the quality of embedding in terms of how well the embedding of the original neighbors stays inside the evaluation sphere as well as how well the embedding of other neighborhoods remains outside of the boundary of the sphere. Sn represents the probability for a neighbor of a particular representative conformation to fall within the evaluation sphere of a given Euclidean distance cutoff. Reversely, PV^+ is the probability that a conformation within the evaluation sphere in the reduced-dimension space is indeed in the neighborhood of the representative conformation in the original space. Suppose that different conformations in the original space are embedded to different points, when the cutoff of Euclidean distance to the representative conformation in the reduced-dimension space is very small, the corresponding evaluation sphere is small, and the number of false negative cases will greatly exceed the number of true positives in eq 6. Thus, Sn will be close to 0%. Meanwhile, if different neighborhoods in the original space are still largely separated after embedding, the number of false positives will be very small, leading to PV^+ close to 100%. When the evaluation sphere is enlarged, the number of false negatives will drop. Consequently, the value of Sn will monotonically increase to 100% when the cutoff reaches the maximum. For a good dimensionality-reduction method, we expected that the embedding of different neighborhoods in the original space remains largely separated in the reduced-dimension space. This will keep the number of false positives small and PV^+ still close to 100% even when the evaluation sphere reaches its maximum value. For a method that mixes up the originally separated neighborhoods during embedding, the large number of false positives will drive the value of PV^+ down.

Residual variance is commonly used to evaluate how well the pairwise "distances" are preserved.[6,9,10b] Following the definition of Tenenbaum et al.,[6] the residual variance is computed as $1 - r^2(\mathbf{D_M}, \mathbf{D_Y})$, where $\mathbf{D_Y}$ is the matrix of Euclidean distances in the low-dimension embedding obtained by each method and the matrix $\mathbf{D_M}$ contains the distances that each algorithm is supposed to preserve. For PCA, $\mathbf{D_M}$ is the matrix of Euclidean distances between each pair of conformations in the Cartesian coordinates. All conformations were superimposed to the center of cluster 1 to remove translation and rotation; for LLE, the element of $\mathbf{D_M}$ is the RMSDs between conformation i and its 20 nearest neighbors. For Isomap, $\mathbf{D_M}$ contains the geodesic distances between all conformations and the landmarks. For diffusion maps, it is the diffusion distance (eq 5) between the pairs of conformations. In the calculation of residual variance, r is the correlation coefficient:

$$r(\mathbf{D_M}, \mathbf{D_Y}) = \frac{\sum_{i=1}^{N} (d_M^i - \overline{\mathbf{D_M}})(d_Y^i - \overline{\mathbf{D_Y}})}{\sqrt{\sum_{i=1}^{N} (d_M^i - \overline{\mathbf{D_M}})^2 \times \sum_{i=1}^{N} (d_Y^i - \overline{\mathbf{D_Y}})^2}}$$

where N is the number of elements in the distance matrices and d_M^i and d_Y^i represents the elements in $\mathbf{D_M}$ and $\mathbf{D_Y}$, respectively. $\overline{\mathbf{X}}$ is the mean value of all the elements in matrix $\mathbf{X}$.

RESULTS AND DISCUSSION

We use the 4-μs MD simulation of the second β-hairpin of the B1 domain of streptococcal protein G as a test case to evaluate the performance of the dimensionality-reduction methods. During the 4-μs simulation, 15 folding and unfolding events were observed when the unfolded state is defined as the conformations with a radius of gyration greater than 10 Å and the native state is cluster 1. The thorough sampling of the important portions of the free energy surface warrants a good estimate of the low free energy minima and the transitions between them using the TRDG. Figure 1 shows the TRDG of the β-hairpin peptide (only the 100 lowest free energy minima are shown for clarity). Cluster 1 corresponds to the global free energy minimum (also called native state) with 73 270 conformations. The 10 lowest free energy minima, which include 56% of the total conformations, are labeled on the TRDG, and they will be used for the comparison with the free

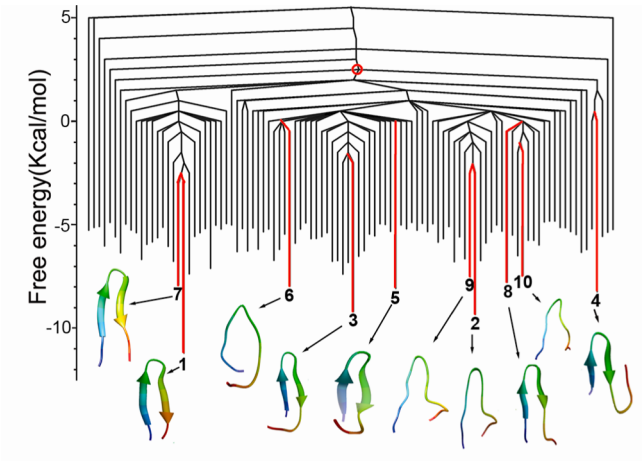


Figure 1. Transition disconnectivity graph of the second β-hairpin of the B1 domain of streptococcal protein G. Transition disconnectivity graph shows the 100 lowest free energy minima. The leaves that are associated with the 10 lowest free energy minima are colored in red. The red circle highlights the free energy barrier between cluster 4 and rest of the 10 lowest free energy minima. The cluster centers of the 10 lowest free energy minima are depicted using Chimera.[23] The N-terminus is in blue, and the C-terminus is in orange.

energy profiles obtained by the dimensionality-reduction methods. The pairwise RMSD of the cluster centers of the 10 largest clusters is listed in Table S1, among which the largest RMSD is 7.77 Å between the centers of cluster 1 and cluster 4, and the smallest RMSD is 3.01 Å between the centers of cluster 1 and cluster 7. We define basins as sets of minima separated by low barriers. The lowest 10 free energy minima can be assigned into four basins: Basin 1 contains clusters 1 and 7. Basin 2 includes clusters 3, 5, and 6. Clusters 2, 9, 8, and 10 belong to basin 3, although the barrier between clusters 2 and 9 is lower than that from them to either cluster 8 or cluster 10. Cluster 4 is separated from the rest of the 10 minima by the highest barrier (at the 2.0 kcal/mol mark of Figure 1) among those between any pair of the 10 clusters. Within a basin, the sampling is expected to approach equilibrium on a relatively short time scale, while the transition between basins occurs in a longer time. For example, from cluster 7 to the native state, the system needs to overcome the barrier at the −2.5 kcal/mol mark of Figure 1, while for the transition from the native basin to the basin where cluster 3 is located, the system needs to cross the barrier at the 1.75 kcal/mol mark of Figure 1.

We compare TRDG with the free energy profiles obtained by each dimensionality-reduction method (Figure 2A–D). We expect that a good dimensionality-reduction method is able to preserve the low free energy minima and the relationship among them. All the methods give multiple free energy minima on the first two embedding dimensions. PCA gives four basins on the profile: Clusters 1 and 7 are in one basin. Clusters 3, 5, and 6 belong to another basin. Clusters 2, 8, and 9 are in a third basin. Cluster 4 is separated from the rest into another basin. However, cluster 10 was embedded close to clusters 3, 5, and 6, inconsistent with TRDG. LLE "squeezes" clusters 2, 3, 5, 6, 8, 9, and 10 into one area while it separates the native state basin and cluster 4 from the rest of the clusters. The free energy profile mapped by LLE is harder to visualize than those obtained by other methods. Isomap groups clusters 4, 6, 8, and 10 into one basin. According to TRDG, the barrier between cluster 4 and the rest of the 10 lowest free energy minima is higher than any of the other barriers that separate the pairs of the 10 clusters. This indicates that Isomap is not able to preserve the relationship between cluster 4 and the other nine clusters in the first two dimensions. The β-hairpin system with the parm19 force field has greater complexity than the test systems using the coarse-grained protein models; therefore, it is not surprising that the embeddings of different neighborhoods are well separated in the coarse-grained models,[9,10] but not in our system. The free energy profiles generated by PCA and diffusion maps share a common feature: both methods embed clusters 2, 3, 5, 6, 8, 9, and 10 into a superbasin, which is separated from the native state basin and cluster 4. The barrier between the native basin and the superbasin is lower than that between cluster 4 and the superbasin. But none of them gives the exact same relationships between the pairs of clusters within the superbasin as TRDG, for example, the barrier between cluster 10 and cluster 3 relative to that between cluster 10 and cluster 9 in the diffusion maps.

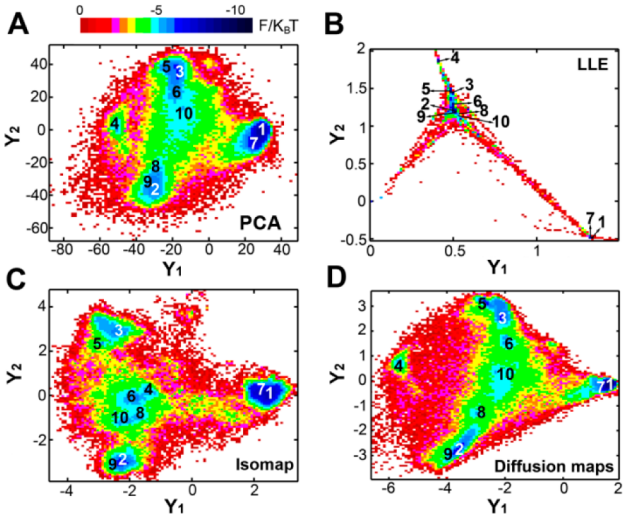


Figure 2. The free energy profiles in the embedding dimensions. (A–D) Free energy profile in the first two embedding dimensions (Y_1 and Y_2) by PCA, LLE, Isomap, and diffusion maps, respectively. For PCA, 200 000 conformations were embedded. The first two embedding dimensions were divided into 90 × 90 grids. For LLE, the conformations that have greater than or equal to 20 neighbors (neighboring cutoff = 3.0 Å) were embedded, for a total of 179 629 conformations on a 100 × 100 grid of the first two embedding dimensions. For Isomap, the largest connected component is embedded, for a total of 179 774 conformations. The embedding dimensions were divided into 100 × 100 grids. For diffusion maps, 200 000 conformations were embedded on the first two dimensions of 180 × 180 grids. For the sake of space, only the area that contains the low free energy minima is shown. The whole free energy profile of the 180 × 180 grids is shown in Figure S3. The first and second embedding dimensions are on the order of 10^{-5}. The Arabian numbers 1–10 are the cluster indexes. The free energy of each grid (F_i) is calculated using eq 1. The color schemes of B–D are the same as those of A.

Clearly, the precise feature of the free energy landscape, such as the height of each free energy barrier and the free energy value of each minimum, depends on a number of parameters and choices, such as the similarity measure (RMSD in our case), the cluster size (cutoff), and the clustering algorithm. RMSD is a common metric for clustering protein conformations and is widely used as a similarity measure in the construction of a neighborhood graph.[10a] It was found that pairs of conformations with large RMSDs are generally separated by high free energy barriers, while for the RMSDs less than 3 Å, the correlation between RMSD and free energy barrier is not present.[21] To evaluate the effect of the clustering cutoff, we reduced the RMSD threshold to 2.0 Å (Figure S2). The smaller cutoff decreases the cluster sizes but otherwise does not affect the performance of dimensionality reduction (Figure S2).

While the visualization of the free energy profiles provides valuable insights into the performance of each method, it is limited up to 3D at a time. It is desirable to have means to evaluate embedding results at higher dimensions for individual free energy minima as well as for the overall data set. We now provide evaluations based on the quantitative measures described in the Methods: sensitivity (Sn), positive predictive value (PV^+), and residual variance. These measures are general and applicable to the evaluation of all dimensionality-reduction methods in any dimension.

We first used Sn and PV^+ to evaluate whether a dimensionality-reduction method can preserve the relationship between a given conformation and its neighboring conformations (defined by 3 Å RMSD). According to the TRDG in Figure 1, cluster 4 has the highest barrier to the other nine clusters among the top 10 largest clusters; thus the embedding

of the neighborhood of the center of cluster 4 should be well separated from other neighborhoods. We, thereby, chose the center of cluster 4 as a representative conformation to test the accuracy of its neighborhood preserving. As explained in the Methods section, a good dimensionality-reduction result is expected to have high Sn and PV^+ values for this test. Figure 3A

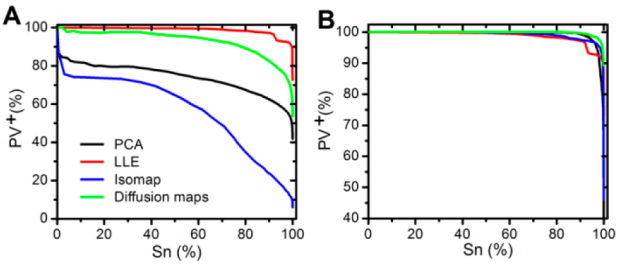


Figure 3. Evaluation of the embedding of the neighborhood of the center of cluster 4. The center of cluster 4 was used as a representative conformation. Embedding in the top two (A) and five dimensions (B). The same scheme of line color is used for A and B.

shows that PV^+ is about 30% when Sn is 80% in the first two embedding dimensions of Isomap. The data indicate that when the evaluation sphere is chosen to contain 80% of the embedding of the original neighbors, 70% of the embedding within the evaluation sphere comes from the non-neighbor conformations (with RMSD >3 Å away from the center of cluster 4 in the original space). Clearly, there is significant mix-up of the embedding of the neighborhood of the center of cluster 4 and other neighborhoods, consistent with the result of the free energy profile (Figure 2C). Other methods perform better than Isomap in the first two dimensions. However, when the embedding dimension is increased to five, the performance of every method is nearly perfect, leading to $PV^+ > 95\%$ when Sn = 80% (Figure 3B). This means that only 5% of the embedding comes from other neighborhoods when the evaluation sphere is set to contain 80% of the embedding of the original neighbors.

We then used Sn and PV^+ to evaluate the overall embedding accuracy as well as to investigate the sufficient number of embedding coordinates. Sn and PV^+ were calculated for each representative conformation and its neighbors. To reiterate, the representative conformations are cluster centers as stated in the Methods. If the PV^+ value reaches 80% or greater when Sn = 80% for the embedding of a given representative conformation and its neighbors, then the embedding is considered to be correct. We define the overall accuracy of neighborhood preserving as

$$A = \frac{\sum_{i=1}^{n} (\delta_i \times NN_i)}{\sum_{i=1}^{n} NN_i} \times 100\% \quad (8)$$

where n is the number of representative conformations (n = 247 for Isomap and n = 248 for all other methods), NN_i is the number of conformations within the 3-Å RMSD cutoff of the representative conformation in the original space, and

$$\delta_i = \begin{cases} 1 & \text{if } \mathbf{PV}^+(\mathbf{i}) \geq 80\% \text{ when } \mathbf{Sn}(\mathbf{i}) = 80\% \\ 0 & \text{otherwise} \end{cases}$$

This criterion evaluates whether a particular method can separate the embedding of neighbors from that of the non-neighbors for all the representative conformations. For LLE, Isomap, and diffusion maps, no substantial improvement is seen if more than 10 embedding coordinates are included (Figure 4).

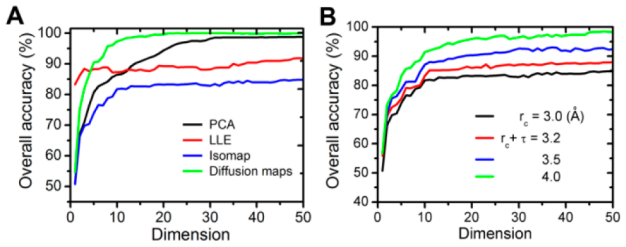


Figure 4. Overall accuracy of neighborhood preserving (eq 8) as a function of embedding dimensions. (A) Overall accuracy for PCA, LLE, Isomap, and diffusion maps. (B) Overall accuracy for Isomap using different tolerances (τ). r_c denotes the RMSD cutoff.

On the contrary, PCA improves its embedding accuracy slowly with the increase of embedding dimension and reaches the plateau region of 97% accuracy when 30 dimensions are included. Isomap shows the lowest overall accuracy of neighborhood preserving even when 50 dimensions are included.

To investigate the source of error of Isomap shown in the evaluation of overall accuracy, we added a tolerance in the calculation of PV^+. Originally, the embedding of any non-neighbor conformation inside the evaluation sphere was considered a false positive. We relaxed the definition of true positive. When Sn is set to 80%, within the evaluation sphere the embedding of the conformations outside the original neighborhood but within the 3 Å RMSD neighborhood cutoff plus the RMSD tolerance is also counted as true positive in eq 7. When the tolerance is varied from 0.2 Å to 1.0 Å, the recalculated overall accuracy of Isomap increases from 75% to 84% in the first five dimensions (Figure 4B). Therefore, the error of Isomap reflected in the evaluation of overall accuracy is mainly caused by the embedding of the non-neighbors with the RMSD slightly or moderately greater than the 3 Å RMSD cutoff. From another point of view, some of the original neighbors were placed further away from the representative conformation after embedding. Figure 4B also implies that most of the far away non-neighbors (e.g., RMSD > 4 Å) remain far away from the representative conformation after embedding. Therefore, the value of PV^+ is not affected by those conformations.

Residual variance is commonly used to estimate the error in distance preserving.[6,9,10] The possible value of residual variance ranges from 0 to 1, corresponding to a high accuracy of distance preserving to a low accuracy. LLE gives large error (residual variance $\cong$ 0.55), while other methods show small errors in distance preserving (residual variance < 0.2) when four or more dimensions are taken into account (Figure 5A). The low performance of LLE on the test of residual variance versus its high performance in the evaluation of Sn−PV^+ (Figure 4) is reasonable because Sn−PV^+ and residual variance evaluate different aspects. Sn−PV^+ intends to test whether LLE can separate the embedding of the representative conformation and its neighboring conformations from that of other neighborhoods. As long as there is no mixing of embedding of different neighborhoods within a given evaluation sphere, the PV^+ value is high. The residual variance test shows that LLE does not preserve the distances between every conformation and its k (= 20) nearest neighbors well. For all of the other methods,

dx.doi.org/10.1021/ct400052y | J. Chem. Theory Comput. 2013, 9, 2490−2497

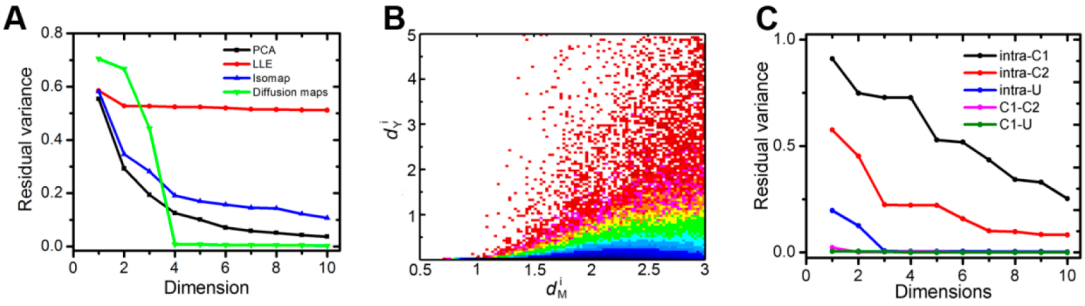


Figure 5. Residual variance. (A) Residual variance as a function of embedding dimensions. (B) Histogram of d_M^i and d_Y^i for LLE. The histogram value is the number of conformation or embedding pairs. Blue to red represents the order of large histogram values to small values. (C) Intra- and intercluster residual variance for diffusion maps. The unfolded state (U) consists of the conformations with a radius of gyration greater than 10.0 Å. C1 and C2 denote cluster 1 and cluster 2, respectively.

residual variance includes the distances between conformations across neighborhoods. In this sense, residual variance evaluates different kinds of distance preserving for different methods. To further illustrate the quality of distance preserving of LLE, we plot the histogram of d_M^i (RMSD between a pair of conformations) and d_Y^i (Euclidean distance in the reduced-dimension space) in Figure 5B. The histogram value is the number of conformation or embedding pairs. The figure clearly shows that for a large number of conformation pairs, far away neighbors (with large d_M^i values) were embedded into nearby regions (with small d_Y^i values). It was reported that LLE worked well for the conformations of complete sampling, while its performance is poor for MD trajectories.[10b] Diffusion maps appear to have a problem of distance preserving in low dimensions. We compare the intra- and intercluster residual variance for the conformations in clusters 1 and 2 and the unfolded state for diffusion maps (Figure 5C). Large embedding errors are seen for the conformations in the same cluster when the dimensions are less than four, while smaller errors are seen for the conformations in the different clusters.

It should be noticed that the residual variance only reflects the error of embedding of the original distance matrix used by a particular method. A desired small variance indicates that the original distances are preserved well in the reduced-dimension space. But if the original distance matrix used by a particular method is not a good descriptor of the relationships between the points in the original space, the overall performance of the method will still be poor even with a small residual variance.

SUMMARY

We used the equilibrium folding/unfolding MD trajectory of a β-hairpin as a benchmark to evaluate the embedding accuracy of several widely used dimensionality-reduction methods. The free energy landscape of this hairpin represented by the transition disconnectivity graph shows substantial complexity. We compared the free energy profiles mapped on the top embedding dimensions with the transition disconnectivity graph and visualized the separation of low free energy minima. We emphasize that it is indispensable to use a complex system with multiple free energy minima and barriers for the evaluation of the quality of embedding. We also utilized sensitivity, positive predictive value, and residual variance to quantitatively evaluate the embedding accuracy. Different evaluation criteria evaluate different aspects of embedding quality: sensitivity and positive predictive value reflect the quality of neighborhood preserving, while residual variance measures the quality of distance preserving. In general, a fair, informative assessment of an embedding result requires a combination of multiple evaluation criteria rather than any single one. For our test system, overall, the linear dimensionality-reduction method, PCA, is not worse than the nonlinear ones. LLE successfully separates cluster 4 and the native basin from the rest of the 10 lowest free energy minima even though it preserves the distance relation between a given conformation and its k nearest neighbors poorly, as shown in residual variance. Diffusion maps have problems in preserving the distance relation in low dimensions but perform very well when more (>4) dimensions are used in the reduced-dimension space. Isomap shows large errors in the test of sensitivity and positive predictive value, which stems from the embedding of nearby non-neighbors. Caution should be used when dimensionality-reduction methods are employed, especially when only a few of the top embedding dimensions are used to describe the free energy landscape. We call for benchmarks of more complex biomolecular systems for further assessment of the applications of dimensionality reduction.

ASSOCIATED CONTENT

Supporting Information

Table of the pairwise RMSD between the cluster centers of clusters 1−10, the log−log plot to estimate the Gaussian kernel bandwidth (ε) in eq 4, the free energy profile in the embedding dimensions by Isomap with a clustering threshold of 2-Å RMSD, and the free energy profile including the embedding of all the 200 000 conformations by the diffusion maps. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: shuo@clarku.edu, lhan@clarku.edu.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

This work is funded by National Institutes of Health (R01-GM088326).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the Scientific Computing and Visualization group at Boston University and the National Center for Supercomputing Applications for providing part of the computational resources. We thank Prof. Dominik Reinhold for helpful discussion.

■ ABBREVIATIONS

RMSD, root-mean-square deviation; PCA, principal component analysis; LLE, local linear embedding; TRDG, transition disconnectivity graph; MD, molecular dynamics; Sn, sensitivity; PV^+, positive predictive value

■ REFERENCES

(1) (a) Yang, M.; Lei, M.; Brüschweiler, R.; Huo, S. Initial conformational changes of human transthyretin under partially denaturing conditions. *Biophys. J.* **2005**, *89* (1), 433−443. (b) Yang, M.; Yordanov, B.; Levy, Y.; Bruschweiler, R.; Huo, S. The Sequence-Dependent Unfolding Pathway Plays a Critical Role in the Amyloidogenicity of Transthyretin. *Biochemistry* **2006**, *45*, 11992−12002.

(2) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential dynamics of proteins. *Proteins: Struct., Funct., Bioinf.* **1993**, *17* (4), 412−425.

(3) Krivov, S. V.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (41), 14766−70.

(4) Jolliffe, I. T. *Principal Components Analysis*; Springer: New York, 1986.

(5) Roweis, S. T.; Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290* (5500), 2323−2326.

(6) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290* (5500), 2319−2323.

(7) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (21), 7426−7431.

(8) Ichiye, T.; Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Struct., Funct., Bioinf.* **1991**, *11* (3), 205−217.

(9) 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* (26), 9885−9890.

(10) (a) Stamati, H.; Clementi, C.; Kavraki, L. E. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins: Struct., Funct., Bioinf.* **2010**, *78* (2), 223−235. (b) Brown, W. M.; Martin, S.; Pollock, S. N.; Coutsias, E. A.; Watson, J. P. Algorithmic dimensionality reduction for molecular structure analysis. *J. Chem. Phys.* **2008**, *129* (6), 064118.

(11) Kentsis, A.; Gindin, T.; Mezei, M.; Osman, R. Calculation of the free energy and cooperativity of protein folding. *PLoS One* **2007**, *2* (5), e446.

(12) (a) Plaku, E.; Stamati, H.; Clementi, C.; Kavraki, L. E. Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction. *Proteins: Struct., Funct., Bioinf.* **2007**, *67* (4), 897−907. (b) Eitrich, T.; Mohanty, S.; Xiao, X.; Hansmann, U. H. E. Dimensionality Reduction Techniques for Protein Folding Trajectories. *From Computational Biophysics to Systems Biology (CBSB07)*; Hansmann, U. H. E., Meinke, J., Mohanty, S., Zimmermann, O., Eds.; John von Neumann Institute for Computing: Jülich, Germany, 2007; NIC Series, pp 99−102.

(13) (a) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (31), 13597−13602. (b) Ferguson, A. L.; Panagiotopoulos, A. Z.; Kevrekidis, I. G.; Debenedetti, P. G. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chem. Phys. Lett.* **2011**, *509*, 1−11.

(14) Ferguson, A. L.; Zhang, S.; Dikiy, I.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; James Link, A. An experimental and computational investigation of spontaneous lasso formation in microcin J25. *Biophys. J.* **2010**, *99* (9), 3056−3065.

(15) (a) Zheng, W.; Qi, B.; Rohrdanz, M. A.; Caflisch, A.; Dinner, A. R.; Clementi, C. Delineation of folding pathways of a beta-sheet miniprotein. *J. Phys. Chem. B* **2011**, *115* (44), 13065−13074. (b) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.* **2011**, *134* (12), 124116.

(16) (a) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Integrating diffusion maps with umbrella sampling: application to alanine dipeptide. *J. Chem. Phys.* **2011**, *134* (13), 135103. (b) Spiwok, V. e.; Králová, B. Metadynamics in the conformational space nonlinearly dimensionally reduced by Isomap. *J. Chem. Phys.* **2011**, *135*, 224504.

(17) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(18) Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **1996**, *105*, 1902−1921.

(19) Lazaridis, T.; Karplus, M. Effective energy function for proteins in solution. *Proteins: Struct., Funct., Bioinf.* **1999**, *35* (2), 133−152.

(20) Gomory, R. E.; Hu, T. C. Multi-Terminal Network Flows. *SIAM J. Appl. Math.* **1961**, *9*, 551.

(21) Li, D. W.; Khanlarzadeh, M.; Wang, J.; Huo, S.; Brüschweiler, R. Evaluation of configurational entropy methods from peptide folding-unfolding simulation. *J. Phys. Chem. B* **2007**, *111* (49), 13807−13813.

(22) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; MIT Press: Cambridge, MA, 1992.

(23) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605−1612.