PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Euclidean sections of protein conformation space and their implications in dimensionality reduction

Mojie Duan,[1] Minghai Li,[1] Li Han,[2]* and Shuanghong Huo[1]*

[1] Gustaf H. Carlson School of Chemistry and Biochemistry, Clark University, Worcester, Massachusetts 01610

[2] Department of Mathematics and Computer Science, Clark University, Worcester, Massachusetts 01610

ABSTRACT

Dimensionality reduction is widely used in searching for the intrinsic reaction coordinates for protein conformational changes. We find the dimensionality−reduction methods using the pairwise root−mean−square deviation (RMSD) as the local distance metric face a challenge. We use Isomap as an example to illustrate the problem. We believe that there is an implied assumption for the dimensionality-reduction approaches that aim to preserve the geometric relations between the objects: both the original space and the reduced space have the same kind of geometry, such as Euclidean geometry vs. Euclidean geometry or spherical geometry vs. spherical geometry. When the protein free energy landscape is mapped onto a 2D plane or 3D space, the reduced space is Euclidean, thus the original space should also be Euclidean. For a protein with *N* atoms, its conformation space is a subset of the *3N*-dimensional Euclidean space R^{3N}. We formally define the protein conformation space as the quotient space of R^{3N} by the equivalence relation of rigid motions. Whether the quotient space is Euclidean or not depends on how it is parameterized. When the pairwise RMSD is employed as the local distance metric, implicit representations are used for the protein conformation space, leading to no direct correspondence to a Euclidean set. We have demonstrated that an explicit Euclidean-based representation of protein conformation space and the local distance metric associated to it improve the quality of dimensionality reduction in the tetra-peptide and β-hairpin systems.

Proteins 2014; 82:2585–2596.

Key words: dimensionality reduction; protein conformation space; isomap; principal component analysis; free energy landscape.

INTRODUCTION

Protein free energy landscapes are of paramount importance in the study of protein thermodynamics and kinetics. To characterize the free energy landscapes with computational approaches, one usually samples the conformation space using molecular dynamics simulation or enhanced sampling methods, then extracts the thermodynamics and kinetics information by post-processing the collected conformations. For a protein of *N* atoms, its conformation space is defined to be the set of all its conformations and is studied as a subset of the *3N*-dimensional Euclidean space R^{3N}. However, not all the dimensions of the conformation space are essential in describing the global conformational changes of proteins: some motions are local and irrelevant to the global conformational change; some degrees of freedom are correlated or anticorrelated in protein folding/unfolding[1,2] and other large conformational changes;[3] and there are constraints due to covalent bonds, bond angles, and other steric factors. Therefore, the free energy can be mapped onto a reduced space with sufficient dimensions, but significantly less than the dimensions of the original conformation space, to describe the pathway of protein conformational changes, the free energy minima, and barriers. Such dimensions of the reduced space can be considered as intrinsic reaction coordinates.

It is nontrivial to search for a set of reaction coordinates to describe the protein conformational changes.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Institutes of Health; Grant number: R01-GM088326; Grant sponsor: NSF; Grant number: IIS-0713335.

*Correspondence to: Li Han, Department of Mathematics and Computer Science, Clark University, 950 Main Street, Worcester, MA 01610. E-mail: lhan@clarku.edu (or) Shuanghong Huo, Gustaf H. Carlson School of Chemistry and Biochemistry, Clark University, 950 Main Street, Worcester, MA 01610. E-mail: shuo@clarku.edu

Received 12 March 2014; Revised 6 May 2014; Accepted 30 May 2014
Published online 10 June 2014 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24622

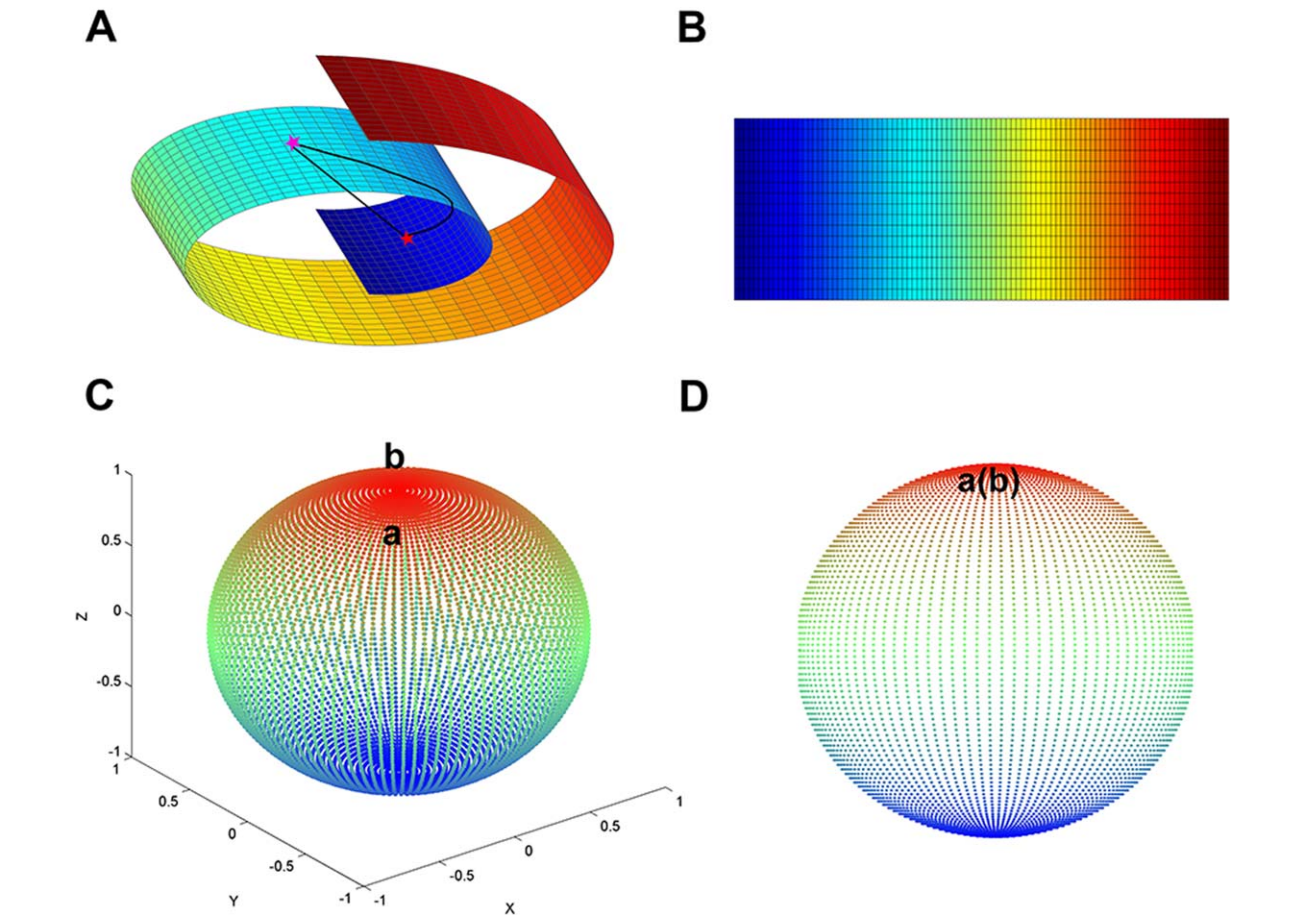

Figure 1

Schematic illustration of preserving geometric relations in dimensionality reduction. **A**: Swissroll model in the 3D space. The curved line shows the geodesic distance between two points on the surface. The straight line shows the Euclidean distance between these points. **B**: The swissroll model is reduced to the 2D space. The color illustrates the neighboring relation. According to the geodesic distances, the red region is next to the yellow region and is farthest from the blue region. The same pattern is seen in the 2D space. **C**: A sphere in 3D space. "a" and "b" are on the opposite side of the sphere. **D**: The sphere is reduced to the 2D space. "a" and "b" are reduced to the same point in the 2D space. Since the 3D sphere is spherical geometry and the 2D space is Euclidean, the dimensionality reduction cannot preserve the geometric relation.

Some intuitive approaches for the analysis of free energy surface rely on the projections of free energy onto some progress variables, such as native contacts, radius of gyration, and root-mean-square deviation (RMSD) from the native state. However, even when multiple projections on different progress variables are used, some possible intermediate states and the transitions are likely to be missed, let alone when only a single progress variable is used.[4] Dimensionality reduction methods provide a systematic approach to search for the reaction coordinates and have obtained promising results.[5–9] Mathematically, dimensionality reduction aims at mapping a higher dimensional space to a lower dimensional space while still preserving the properties of the original space. For example, when the "Swiss roll" data set in three-dimensional space is embedded onto two dimensions, the neighborhood relation is well preserved in the reduced space (Fig. 1). The reduced space has many advantages over the original space, such as feasibility of mathematical analysis, efficiency of computational studies, and visual inspections of geometrical properties, etc.

Dimensionality-reduction algorithms can be classified into two classes: linear methods, such as principal component analysis (PCA),[10] and nonlinear methods, such as Isomap,[11] locally linear embedding,[12] and diffusion maps.[13] Both PCA and Isomap aim at preserving geometric properties of the original space. For PCA, the mapping of the original data points onto the reduced space is the solution to an optimization problem that aims at maximizing data variance along the orthogonal directions in the reduced space. The principal components are orthogonal vectors among which those associated with the large variances capture the protein global conformational changes, while those associated with small variances correspond to the local motions. The computation of PCA is based on the Cartesian

coordinates of the points in the Euclidean space. Such a method is well suited for data points with a distribution close to a hyperplane.[14] In other words, if the sampled protein conformations lie on or near a linear subspace, then the projection of free energy surface onto the principal components associated with the large variance will reflect the true thermodynamic and kinetic relationships between the free energy states.

However, the underlying subspace on which the protein conformations lie is believed to be curved, for which the nonlinear methods, such as Isomap, are supposed to work better than PCA. One of the signature features of Isomap is to use geodesic distances to describe the geometric relations between points on a surface (Fig. 1), which should be preserved in the reduced space. The procedure of Isomap can be divided into three steps: construction of neighborhood graph, computation of geodesic distance, and embedding in the reduced space. Isomap has been shown to work well in mathematical model systems such as swissroll[11] and biomolecular systems such as the coarse-grained model of an SH3 domain.[6] In contrast, in our recent study on evaluating the results of several dimensionality reduction methods on a β-hairpin system, we found that the implementation of Isomap using the pairwise RMSD as the distance metric to construct the neighbor graph underperformed PCA.[15]

An underlying, but somewhat implicit, assumption for dimensionality-reduction approaches that aim at preserving the geometric relations in the original space is that the original space and the reduced space have the same kind of geometry (such as Euclidean geometry vs. Euclidean geometry or spherical geometry vs. spherical geometry) [Fig. 1(C,D)] and, when possible, with explicit representations. When the protein free energy landscape is mapped onto a 2D plane or 3D space, the reduced space is Euclidean, and thus the original space is supposed to be also Euclidean. An overlooked issue is whether the protein conformation space can be identified with a Euclidean set and thus well suited for dimensionality reduction to the low-dimensional Euclidean space.

The assumption of Euclidean to Euclidean mapping poses a challenge to all dimensionality reduction methods that use pairwise RMSD as a distance metric for neighbors because this metric uses implicit representation of the protein conformation space which does not correspond to a Euclidean geometry (see the Discussion sections). The poor performance of the pairwise RMSD-based Isomap on β-hairpin is likely to be attributed to the violation of this assumption. We have developed a scheme of explicit representation of protein conformation space, which forms a Euclidean section. We use Isomap as an example to demonstrate the impact of the Euclidean representation of protein conformations on the quality of dimensionality reduction.

METHODS AND MATERIALS

Model systems

We used alanine tetrapeptide and the second β-hairpin of the B1 domain of streptococcal protein G as benchmarks. These two systems are small to allow extensive sampling while contain sufficient complexities, especially the β-hairpin system. These kind peptide systems are commonly used as model systems to test new methods. A 1-μs Langevin dynamics simulation was performed for the tetra-alanine system using CHARMM[16] (version c35b2), where the ACE and CBX group are blocked at the two termini, respectively. The CHARMM 19 polar hydrogen potential function[17] and the ACE2 implicit solvation model[18] were employed. One million conformations were saved at the time interval of 1 ps and were used for the dimensionality reduction. MSMBuilder2[19] package was employed for clustering and lumping, resulted in 24 macrostates. The details of the simulation and clustering are published elsewhere.[20] The largest three macrostates, which include 93.5% of the total conformations, are used in the analysis of the embedding results.

The sequence of the second β-hairpin of the B1 domain of streptococcal protein G is G-E-W-T-Y-D-D-A-T-K-T-F-T-V-T-E without any blocking groups at the termini. A 4-μs equilibrium folding-unfolding simulation of this β-hairpin at 360 K using CHARMM (version c31b1) was published earlier.[21] The parm19 polar hydrogen potential function and EEF1 implicit solvation model[22] were employed. Snapshots were saved every 20 ps. A total of 200,000 conformations were used for dimensionality reduction.

Dimensionality reduction

We applied PCA (*ptraj* module in AmberTools (v9.0)) and Isomap to the tetra-alanine and β-hairpin system. Two different distance metrics for neighbors were used in the implementation of Isomap: one is pairwise RMSD, the other one is common reference Euclidean (crEuclidean) distance. For the latter metric, we first chose one conformation described by *3N* Cartesian coordinates as the common reference to compute the best superimposed positioning for every other conformation to eliminate the translation and rotation, then computed the Euclidean distance between any two conformations. For the Isomap implementation, k ($= 20$) nearest neighbors of each conformation were picked using the two different distance metrics for neighbors. The β-strand conformation [backbone dihedral angles: (−100.2, 140.1), (−105.6, 122.2), (−83.4, 146.2), and (−95.3, 139.6)] of the tetra-peptide and the center of cluster 1 of β-hairpin were used as the references to remove translation and rotation for PCA and the crEuclidean distance-based Isomap method.

In the implementation of Isomap, each pair of neighboring conformations is connected by an edge whose weight is equal to the crEuclidean distance or the pairwise RMSD. For the non-neighboring conformations, a shortest path between them that corresponds to the minimum of the total edge weight along the path is found using the Dijkstra's algorithm.[23] The geodesic distance is defined as the sum of edge weights along the shortest path between the non-neighboring conformations or the edge weight of neighboring conformations. The goal of the Isomap method is to preserve the geometry of the data, which is encoded in the geodesic distance matrix ($\boldsymbol{D}$). In the low-dimensional space ($\boldsymbol{Y}$), the new set of coordinates $\vec{y}_i$ are chosen to minimize the objective function, $\Phi = |\tau(D) - \tau(D_Y)|$. The τ matrix is equal to $-HSH/2$, where $S_{ij} = D_{ij}^2$ and $H_{ij} = \delta_{ij} - 1/m$, respectively, with δ_{ij} as the Kronecker delta and m as the number of conformations. The τ operator uniquely characterizes the geometry of the data to allow an efficient optimization.

To further reduce the computational cost in the implementations of Isomap, the landmark-based approach is adopted. Only a small portion of conformations are selected as the landmarks and only the geodesic distances between each conformation and landmark conformations are preserved in the calculation. The details of this implement are described in Ref. 6. We chose a conformation every 800 ps as a landmark along the trajectory of β-hairpin, resulted in a total of 5000 landmarks. The procedure of connected-components[23] was used to find the largest connected component out of the 200,000 conformations. The conformations that do not belong to the largest connected component were removed, resulted in 4491 landmarks and 179,774 conformations and 179,690 conformations for the pairwise RMSD-based Isomap and crEuclidean distance-based Isomap, respectively. Following the same procedure, we found that all the conformations of the tetra-peptide belong to a single connected-component. One million conformations of the tetra-peptide and 400 landmarks were used in both implementations of Isomap.

Transition disconnectivity graph (TRDG) and the 2D free energy profile

The all-atom RMSD of 3 Å was used as the criterion to cluster β-hairpin conformations.[15] The TRDG for the β-hairpin using this cutoff was published before.[15] The free energy of each cluster i is defined as (see Ref. 29 for details):

$$F_i = -k_B T \ln n_i, \tag{1}$$

where k_B is the Boltzmann constant, T is the simulation temperature (T = 360 K in our simulation), and n_i is the number of conformations in cluster i. The free energy barrier between clusters i and j is calculated by:

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

where Z_{ij} is the partition function of the barrier and is related to the minimum cut value (n_{ij}) computed using the Gomory and Hu algorithm.[24]

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

where h is the Planck's constant and $\Delta t = 20$ ps which is the time interval between the collected conformations of β-hairpin. The 2D-grids were used 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 a given grid was calculated by Eq. (1).

Evaluation of embedding results

Residual variance defined as $1 - r^2(D_M, D_Y)$ is used to assess the error in preserving the pairwise "distances" for dimensionality-reduction methods. The value of residual variance is between 0 and 1 with zero indicating the perfect correlation and 1 showing no correlation between the "distances" defined in the original space/graph and the Euclidean distances in the reduced dimensions.[6,11,25] For PCA, D_M is the matrix of Euclidean distances between each pair of conformations after the removal of the rigid motions with respect to the reference conformation; for Isomap, D_M contains the geodesic distances between all conformations and the landmarks. D_Y is the Euclidean distance matrix in the reduced dimensions. The correlation coefficient, $r(\boldsymbol{D}_M, \boldsymbol{D}_Y)$, is computed for all the elements in the matrices.

We use sensitivity (Sn) and positive predictive value (PV^+) to evaluate the quality of neighborhood preserving for the β-hairpin system.[15] First, we chose the clusters with more than 50 conformations and used their cluster centers as representative conformations, for a total of 248 and 247 representative conformations for PCA and (the largest connected component of) Isomap, respectively. All the conformations that are within 3-Å RMSD cutoff of a given representative conformation are considered to form its neighborhood. We then searched for the maximal Euclidean distances between a given representative conformation and its neighborhood conformations in the reduced dimensions. We used one thousandth of this maximal Euclidean distance as an initial cutoff in the reduced dimensions as well as an incremental value. We subsequently calculated Sn and PV^+ at the initial cutoff, then recalculated Sn and PV^+ at the next incrementally increased cutoff until the cutoff reaches the maximum to include the embedding of all

the neighborhood conformations. For a given Euclidean distance cutoff with respect to a particular representative conformation, Sn and PV^+ are defined as follows.

$$Sn = \frac{N_{true\ positive}}{N_{true\ positive} + N_{false\ negative}} \times 100\%, \quad (4)$$

$$PV^+ = \frac{N_{true\ positive}}{N_{true\ positive} + N_{false\ positive}} \times 100\%. \quad (5)$$

If any conformation within the 3-Å RMSD neighborhood of a given representative conformation in the original space is embedded within the Euclidean distance cutoff to the representative conformation in the reduced dimensions, then the embedding of this pair of conformations is counted as a true positive, namely, the neighborhood relation of this pair of conformations is preserved, otherwise, it is called false negative. If any conformation outside the neighborhood of a given representative conformation in the original space is embedded outside the Euclidean distance cutoff in the reduced dimensions, then the embedding of this pair of conformations is counted as true negative. Otherwise, it is called false positive.

For a good dimensionality-reduction method, we expect that both false positives and false negatives are low, therefore, we consider the embedding of a given representative conformation and its neighbors is correct if PV^+ reaches 80% or greater when Sn = 80%. Then we define the overall accuracy of neighborhood preserving as

$$A = \frac{\sum_{i=1}^{n} (\delta_i \times NN_i)}{\sum_{i=1}^{n} NN_i} \times 100\% \quad (6)$$

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

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

RESULTS

We applied PCA and Isomap to the tetra-alanine and β-hairpin system. Two different distance metrics for neighbors were used in the implementation of Isomap: pairwise RMSD and crEuclidean distance. Here, we show the improved Isomap performance when the crEuclidean distance was used as a distance metric for neighbors. We used residual variance and the overall accuracy of neighborhood preserving as measures to evaluate the quality of dimensionality-reduction results in addition to the free energy surface projection.

Tetra-alanine

Since the peptide consists of only four residues, the Ramachandran plot of the second and third pair of backbone dihedral angles are employed to compare with the embedding results (Fig. 2). The (ϕ,φ) angles of the three largest macrostates are shown, which includes about 93.5% conformations in the trajectory. For the second pair of (ϕ,φ) angles [Fig. 2(A)], macrostate 2 is in the α helical region while macrostate 1 and 3 are predominantly distributed in the β and P_{II} region. The direct conformational transitions from macrostate 3 to macrostate 1 are more frequent than those from macrostate 3 to macrostate 2.

The free energy landscapes projected on the first two embedding dimensions by different methods are shown in Figure 3. Overall, the three methods result in similar free energy surface: macrostate 1 and 3 are projected onto the same basin separated from macrostate 2. However, the three methods give different heights of the barrier between macrostate 2 and the basin of macrostate 1 and 3. The pairwise RMSD-based Isomap shows a higher barrier than the other two methods. The results of residual variance (Fig. 4) indicate that the pairwise RMSD-based Isomap results in the worst "distance" preservation for the system.

β-Hairpin

We observed fifteen folding and unfolding transitions along the 4-μs trajectory of β-hairpin, given that cluster 1 is the native state and the conformations with the radius of gyration >10 Å are unfolded. Although the sampling of the high energy states are not converged, we do not expect substantial changes in the relative free energies of the 10 lowest free energy minima and the relative barriers between them even if longer simulations are performed. Therefore, in the evaluation of the free energy profile in the reduced dimensions, we focused on the ten lowest free energy minima and the relative barriers between them. The TRDG was employed to visualize the free energy landscape without any projection. The technique was first introduced to the biophysics community by Czerminski and Elber[26] and subsequently used by Becker and Karplus[27] to analyze the relations between kinetic connectivity and spatial proximity of a model peptide. Then it has been widely used in both biophysics[21,28–32] and material science[33] community.

To plot the TRDG, free energy minima are recursively partitioned into two disjoint subsets. Each pair of minima within a subset is connected by a low free energy barrier, but the system would overcome a high barrier to

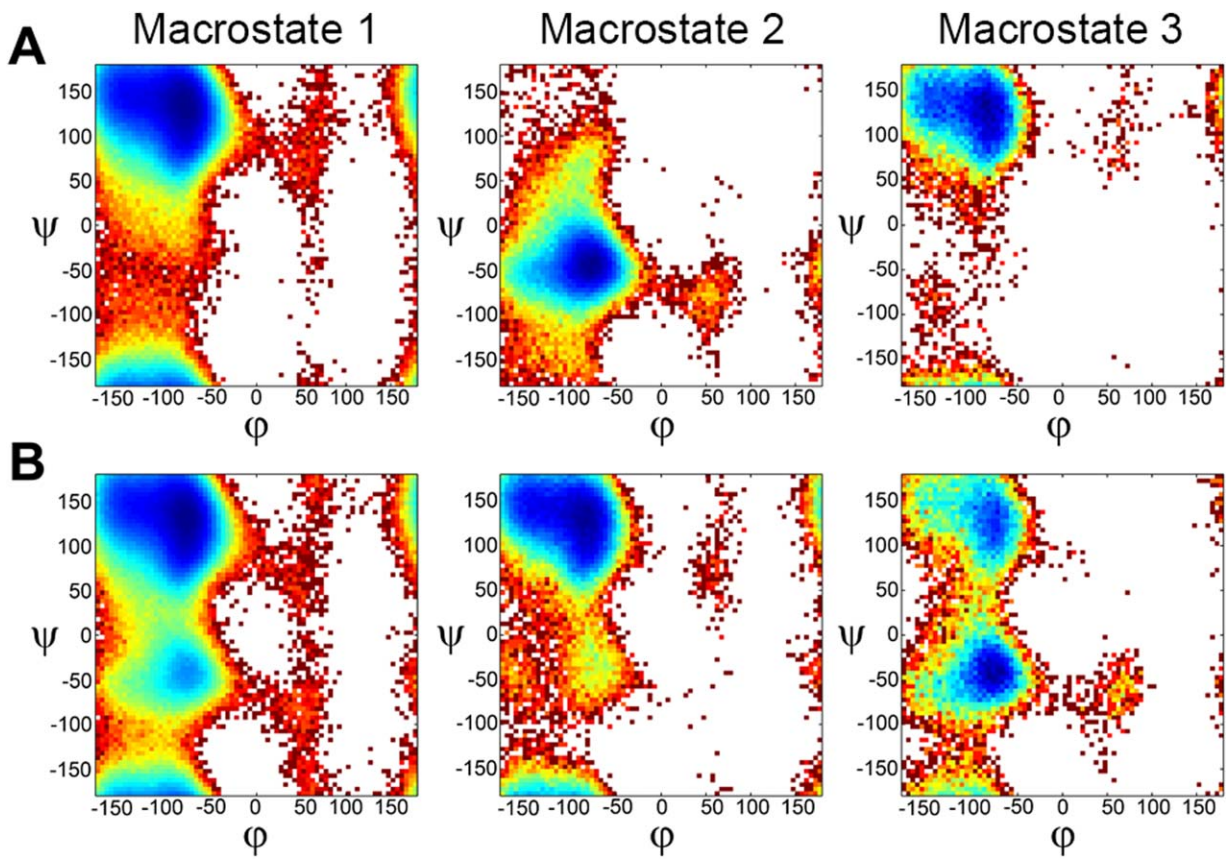

Figure 2

Ramachandran plot of the first three macrostates out of the 24 macrostates of tetra-alanine peptide. See Ref. 20 for the details of the clustering and lumping. **A**: The second pair of backbone dihedral angles of tetra-alanine peptide. **B**: The third pair of backbone dihedral angles of tetra-alanine peptide. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

reach a minimum in another subset. A TRDG is plotted on a vertical scale. All branches terminate at a node or a local minimum, which is a cluster. Two terminal nodes merge at an internal node, which represents the barrier between the two nodes. All the terminal nodes that connect directly or indirectly to an internal node are mutually accessible at the cost of the free energy barrier of that internal node. The connectivity patterns between the minima represent a mapping of the free energy landscape although the path information is sacrificed. The

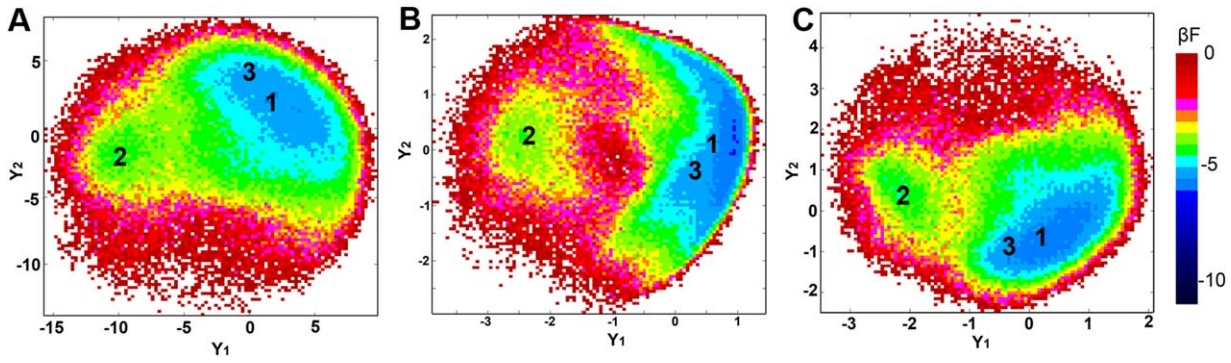

Figure 3

Free energy surface of tetra-alanine peptide projected on the first two reduced dimensions by (**A**) PCA; (**B**) Pairwise RMSD-based Isomap; (**C**) common reference Euclidean distance-based Isomap. The "1–3" label corresponds to the macrostate index in Figure 2. $\beta = 1/k_B T$, where $T = 500$ K. F is the free energy (kcal/mol).

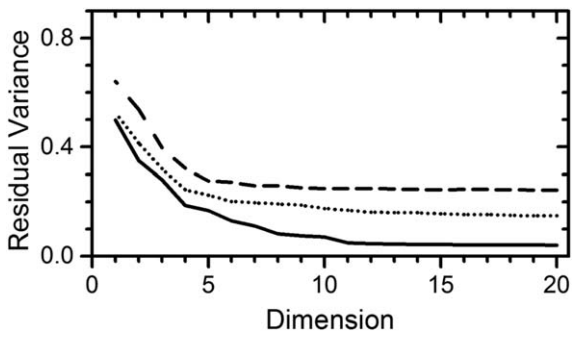

Figure 4

Residual variances of the dimensionality reduction of the tetra-alanine. The solid line represents PCA, the dotted line is for the common reference Euclidean distance-based Isomap, and the dashed line denotes the pairwise RMSD-based Isomap.

advantage of TRDG in representing the topology of the free energy landscape is that it is independent of the dimensionality of the system, whereas the projection of free energy onto reaction coordinates are restricted to one or two dimensions.

According to the connectivity pattern of the nodes on the TRDG of β-hairpin (Fig. 5), the transitions from cluster 4 to any of the other ten largest clusters need to cross a high barrier (∼2.0 kcal/mol marked by a circle in Fig. 5), while the rest nine largest clusters are mutually accessible below this free energy level. These nine clusters can be subsequently partitioned into three subsets: cluster 1 and 7 are in one subset; cluster 3, 5, and 6 belong to another subset; cluster 2, 9, 8, and 10 are in the third subset. The conformational transitions between the clusters within the same subset are to overcome a lower free energy barrier than those between the different subsets. For example, from cluster 7 to cluster 1, the system needs to overcome the barrier at the −2.5 kcal/mol mark of Figure 5, while for the transition from the subset of cluster 1 and 7 to the subset containing cluster 3, 5, and 6, the system needs to cross the barrier at the 1.75 kcal/mol mark of Figure 5.

The relationships between the free energy minima on the embedded landscape should be the same as shown in the TRDG, that is, two minima separated by a high barrier in TRDG are expected to be separated from each other and still have high barrier in the reduced space. The largest embedding error of the pairwise RMSD-based Isomap is reflected in the relationship between cluster 4 and the rest nine largest clusters (Fig. 6). Both PCA and crEuclidean distance-based Isomap show that cluster 4 is separated from the other nine clusters, which is consistent with the connectivity pattern shown in the TRDG, whereas the pairwise RMSD-based Isomap shows that cluster 4 is mixed with cluster 6, 8, and 10. This indicates that the pairwise RMSD-based Isomap is not able to preserve the relationship between cluster 4 and other 9 largest clusters in the first two dimensions.

Except for cluster 4, the three dimensionality-reduction methods show some consistencies in the distribution of the other nine largest clusters: cluster 1 and cluster 7 are in the same basin, which is in line with that of the TRDG connectivity pattern, and the other clusters are located in a superbasin with cluster 2 and 9 residing in one basin. The relation between cluster 2 and 9 also agrees with that shown in TRDG. The relation between cluster 6 and cluster 3/5 given by the crEuclidean distance-based Isomap is the most consistent with the TRDG pattern, where the barriers between these clusters are at the same height and higher than that between cluster 2 and 9. As to the relation between cluster 8 and 10 and the basin of cluster 2 and 9, pairwise RMSD-based Isomap gives consistent result with that of TRDG, which shows that cluster 8 and 10 are in one subset while cluster 2 and 9 belongs to another subset, and these two subsets are separated by a higher barrier than the nodes within the same subset of these two subsets.

While the visualization of free energy profiles is intuitive, they are limited to one or two dimensions. Numerical measures such as residual variance[6,11,25] and neighborhood preserving accuracy[15] are broadly applicable to all dimensions. As shown in Figure 7, the pairwise RMSD-based Isomap results in the largest error (residual variance = 0.35 in 2D) among the methods, while the crEuclidean distance-based Isomap performs the best. The residual variance only reflects the quality of the fitting between the "distances" in the original space/graph and the Euclidean distances in the reduced space.

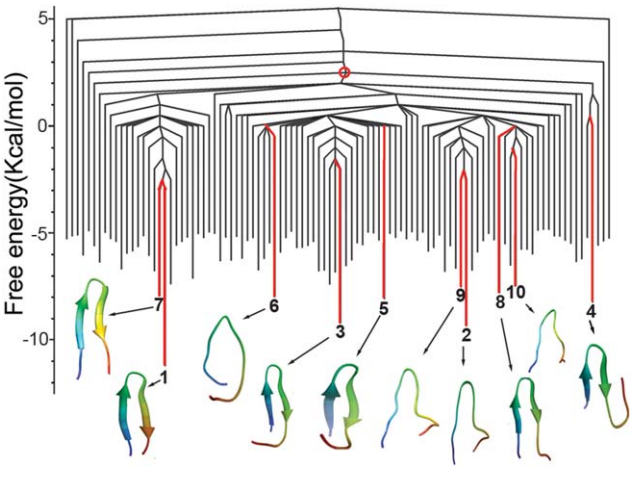

Figure 5

TRDG of β-hairpin. For clarity, only the 100 lowest free energy minima are shown. The circle highlights the free energy barrier between cluster 4 and the rest of the 10 largest clusters. The cluster centers of the 10 lowest free energy minima are depicted using chimera.[35] Reproduced with permission from Duan, J Chem Theory Comput, 2013, 9, 2490–2497, ©American Chemical Society. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

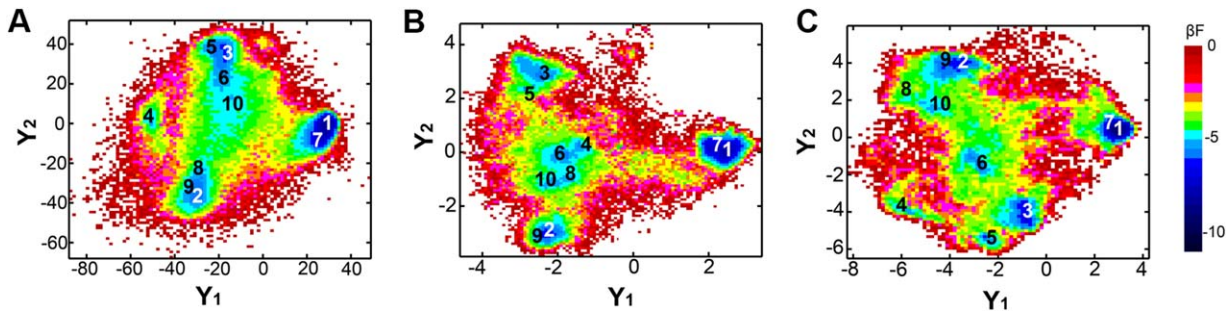

Figure 6

The free energy landscapes on the first two reduced dimensions (Y_1 and Y_2). **A**: PCA; (**B**) Pairwise RMSD−based Isomap; (**C**) crEuclidean distance-based Isomap. For PCA, 200,000 conformations were embedded. The first two embedded dimensions were divided into 90 × 90 grids. For the pairwise RMSD-based Isomap, the largest connected component is embedded, for a total of 179,774 conformations. For crEuclidean distance-based Isomap, 179,690 conformations were embedded. The embedding dimensions of both Isomap implementations were divided into 100 × 100 grids. The Arabian numbers 1–10 denote the ten largest clusters. $\beta = 1/k_B T$ where k_B is the Boltzmann constant and $T = 360$ K for this system. F is the free energy (kcal/mol). A and B are Reproduced with permission from Duan, J Chem Theory Comput, 2013, 9, 2490–2497, ©American Chemical Society.

However, if the distance metric is not good, the overall quality of dimensionality reduction will be poor even though the residaul variance is low. Therefore, other methods are needed to assess the embedding quality from different aspects.

We employ the Sn−PV^{+15} to assess the accuracy of neighborhood preserving. The basic criterion for the neighborhood preserving is that neighborhood conformations in the original space should remain in the neighborhood in the reduced dimensions, and vice versa. The crEuclidean distance-based Isomap gives the best overall neighborhood preserving accuracy up to 20 dimensions (with ∼95% accuracy; Fig. 8). However, after the first 20 dimensions, PCA shows the highest accuracy, which reaches 100% in 30 dimensions. The pairwise RMSD-based Isomap shows the lowest overall accuracy of neighborhood preserving. Even when the dimensions reach 100, its accuracy still remains around 80%.

In summary, the embedding quality of the pairwise RMSD−based Isomap for the tetra-alanine and β-hairpin is the lowest among the three tested methods according to the measures of residual variance and the overall accuracy of neighborhood preserving in addition to free energy surface project. This was unexpected since Isomap is a nonlinear method, which is supposed to work well for nonlinear systems, such as the protein conformation space.

DISCUSSION

Protein conformation space and coordinates in R^{3N}

In this section, we will first review relevant mathematical concepts and then present the underlying reasoning for the improved performance of the crEuclidean distance-based Isomap.

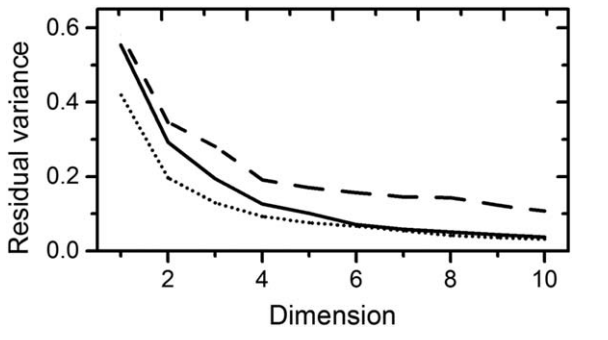

Figure 7

Residual variance as a function of dimension for β-hairpin. The solid line represents PCA, the dotted line is for the common reference Euclidean distance-based Isomap, and the dashed line denotes the pairwise RMSD-based Isomap.

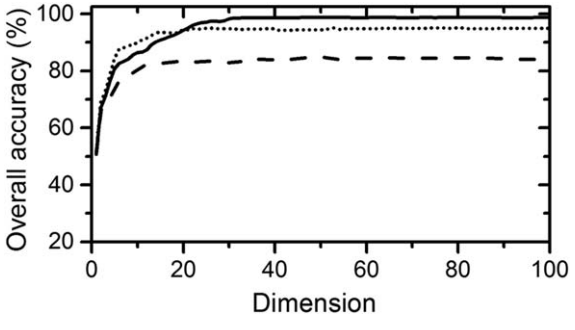

Figure 8

Overall neighborhood preserving accuracy as a function of dimension for the first 100 dimensions of β-hairpin. The solid line represents PCA, the dotted line is for the common reference Euclidean distance-based Isomap, and the dashed line denotes the pairwise RMSD-based Isomap.

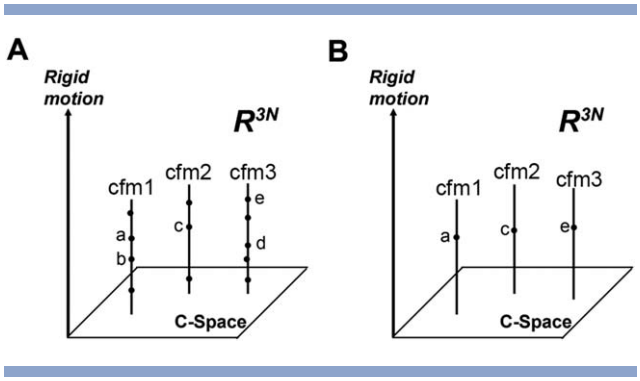

Figure 9

Schematic illustration of the representations of protein conformation space (CSpace). **A**: Implicit respresentation as used by the pairwise RMSD implementation. **B**: Explicit respresentation as used by PCA and crEuclidean distance-based Isomap.

Definition 1

A given binary relation $\sim$ on a set S is an **equivalence relation** if and only if the following properties hold for any a,b,c \in S,

- a$\sim$a, (reflexivity)
- if a$\sim$b, then b$\sim$a (symmetry)
- if a$\sim$b and b$\sim$c, then a$\sim$c (transitivity)

Definition 2

Given a set S and an equivalence relation $\sim$, an **equivalence class** having element a, commonly denoted by [a], consists of all elements equivalent to a. In other words, $[a]=\{b \in S | b \sim a\}$.

Definition 3

The quotient space, $S/\sim$, is defined to be the set of equivalence classes of elements of S:

$$S/\sim = \{[a] : a \in S\}.$$

When the above definitions are applied to a protein with N atoms, the R^{3N} space is the set (S) of Cartesian coordinates of the protein. Rigid motions define an equivalence relation on R^{3N}.

- Each positioning of the protein described by $3N$ Cartesian coordinates corresponds to a $3N$-vector in the R^{3N} space, and each positioning is related to itself through the identity motion.
- If positioning a can be transformed to positioning b through rigid motion g, then b can be transformed to a through g^{-1}, the inversion motion of g.
- If a can be transformed to b through g and b can be transformed to c through h, then a can be transformed to c through the composition of the motion g and h.

Thereby, each protein conformation corresponds to an equivalence class of Cartesian coordinates in R^{3N} that consists of all the positioning that can be related to each other through rigid motions. Now, the protein conformation space can be formally defined as the quotient space of R^{3N} by the equivalence relation of rigid motions:

$$\text{Protein conformation space} = R^{3N}/\text{3D-Rigid-Motion}.$$

After the quotient space undergoes dimensionality reduction, the free energy will be "projected" onto the reduced dimensions to depict the free energy landscape. As mentioned in Introduction, the implied assumption of dimensionality reduction is that the original space and the reduced space have the same kind geometry, such as Euclidean geometry to Euclidean geometry and spherical geometry to spherical geometry. Since the "projected" free energy landscape in the 2D space is Euclidean, the quotient space defined above should also be Euclidean. In fact, whether the quotient space is Euclidean or not depends on how it is parameterized.

Space representation

Isomap uses geodesic distances between points to capture the geometric properties of the original space and maps the data points to a lower dimensional Euclidean space. The usage of geodesic distances is mathematically sound for curved systems and is found to work well in some molecular systems.[6,34] However, when it is applied to protein conformations with pairwise RMSD as the distance metric for neighbors and edge weight, the protein conformation space only has an implicit, abstract representation. In the process of computing pairwise RMSDs, multiple positioning of one conformation may be used.

For example, consider a simple example of pairwise RMSD computation of three conformations, *cfm1*, *cfm2*, and *cfm3*, as shown in Figure 9(A): *cfm1* has positioning *a* and *b* (related through rigid motions), *cfm2* has positioning *c*, and *cfm3* has positioning *d* and *e*. Assume that

- *a* and *c* are the best superimposed positioning for *cfm1* and *cfm 2*,
- *d* and *c* are the best superimposed positioning for *cfm3* and *cfm2*, and
- *b* and *d* are the best superimposed positioning for *cfm1* and *cfm3*.

While *a* and *d* are the best superimposed positioning with respect to *c* when the pairwise RMSD between *cfm1* and *cfm2* as well as that between *cfm3* and *cfm2* are calculated, *a* and *d* are not necessarily the best superimposed positioning for the RMSD between *cfm1* and *cfm3*. Since positioning *b* in *cfm1* is assumed to be the best for the superposition with positioning *d* for the RMSD

calculation between *cfm1* and *cfm3*, then *cfm1* is represented by two positioning (*a* and *b*) in the calculation of pairwise RMSD, which is subsequently used as the distance metric for neighbors and edge weight on the graph to compute geodesic distances.

As shown in the above example, each conformation is implicitly represented by a cloud of points in R^{3N} in the computation of pairwise RMSD. After dimensionality reduction, the clouds of points in the original space are mapped to individual points in the reduced space. The quotient space is, therefore, not explicitly represented and does not directly correspond to a Euclidean set. This is inconsistent with the assumption that the Euclidean reduced-space should correspond to a Euclidean set in the original space. We believe that this problem is likely a contributing factor to the poor performance of the pairwise RMSD-based Isomap on tetra-alanine and β-hairpin systems.

In contrast, both PCA and the crEuclidean distance-based Isomap use one particular positioning to represent a conformation, or in mathematical term, use one section of the quotient space that consists of one representative positioning for each conformation. As illustrated in Figure 9(B), assume that *a* is the positioning of *cfm1* used as the reference to factor out the rigid motions, *c* and *e* are the best superimposed positioning in *cfms 2* and *3* with respect to *a*. With PCA, the positioning of *a, c, e* forms the original matrix for the subsequent linear transformation; and with crEuclidean distance-based Isomap, the positioning of *a*, *c*, and *e* in Figure 9(B) is used to compute the edge weight, to generate the neighbor lists, and then to calculate the geodesic distances between the conformations. For example, the crEuclidean distance between *cfms 2* and *3* are computed from the coordinates of *c* and *e* by $\sqrt{\frac{\sum_{i=1}^{3N}(x_i(c)-x_i(e))^2}{N}}$. Note that this crEuclidean distance maybe greater than the RMSD between the two conformations since *c* and *e* are not necessarily the best superimposed positioning of the corresponding conformations with respect to each other. Since the only difference in the two implementations of Isomap is the distance metric for neighbors, we credit the improved performance in the tetra-alanine and β-hairpin systems to the crEuclidean distance measure.

It was reported in prior work that PCA performs poorly in nonlinear model systems[11] and coarse-grained protein systems,[6,34] while we observed that PCA gives the highest overall accuracy in neighborhood preserving for β-hairpin in more than 20 dimensions and the least residual variance for tetra-alanine peptide. One possible explanation of the discrepancy is as follows. Comparing the polar hydrogen force field that we used with the coarse-grained models, our sampling of the high energy states are not converged as well as that of the coarse-grained model. Accordingly, our reported assessment of embedding quality, such as the results of residual variance and neighborhood preserving accuracy, is dominant by the conformations in the low energy states, which locate in a relatively small region with respect to the whole conformation space.

RMSD cutoff

RMSD is a common metric for clustering protein conformations. We used RMSD for clustering in the construction of TRDG. The RMSD cutoff determines the size of the cluster and accordingly the free energy of each node as well as the barriers between nodes in TRDG. It was found that pairs of β-hairpin conformations with large RMSDs are generally separated by high free energy barriers, while for the RMSDs <3 Å, the correlation between RMSD and free energy barrier is not present.[21] Therefore, the RMSD cutoff should not be >3 Å for the β-hairpin system. Yet, unnecessarily small RMSD cutoff results in many clusters with a few conformations. If one starts a molecular dynamics simulation from an NMR structure, it is reasonable to obtain a trajectory with about 3 Å all-atom RMSD relative to the native state. Therefore, the choice of 3 Å all-atom RMSD for the clustering cutoff is reasonable. We evaluated the influence of the clustering cutoff on the performance of dimensionality reduction methods. Comparing the free energy surface projected on the first two reduced dimensions with TRDG, the method that performs poorly when the large RMSD cutoff is used still does badly when the cutoff is reduced.[15] To be consistent with the clustering cutoff, the cutoff in the evaluation of neighborhood preserving accuracy should also be 3 Å all-atom RMSD. We show that when the neighborhood cutoff is reduced to 2 Å all-atom RMSD, the Sn-PV^+ plot shows little changes (Supporting Information Fig. S1) In short, as long as the RMSD cutoff is within a reasonable range, the performance of dimensionality reduction methods is not sensitive to the cutoff.

CONCLUSION

The conformation space of a protein with *N* atoms is considered to be *3N-6* dimensions, where *3N* corresponds to the Cartesian coordinates of the atoms and 6 is the dimension of the rigid motions in the space. However, to the best of our knowledge, there was no formal definition for protein conformation space or discussion on the relations between protein conformation space and the Cartesian coordinate space R^{3N}. In this article, we formally define the protein conformation space to be the quotient space of R^{3N} over rigid motions in the space. We focus on the representation issue of protein conformation space and its impact on the dimensionality.

For dimensionality-reduction approaches to preserve the geometric relations between the objects, an implied assumption is that both the original space and the

reduced space have the same kind of geometry, such as Euclidean geometry versus Euclidean geometry or spherical geometry vs. spherical geometry. For a protein with N atoms, while its Cartesian coordinate space R^{3N} is Euclidean, its conformation space, which is the quotient space as defined above is not necessarily Euclidean because its geometrical properties depend on how it is parameterized. The computation of pairwise RMSD among a set of protein conformations generally involves multiple positioning of each conformation, but in the reduced space, one conformation may only be at one point. When the pairwise RMSD is used as the local distance metric, the protein conformation space is not explicitly represented and does not directly correspond to a Euclidean set, which is inconsistent with the assumption of the (Euclidean geometry to Euclidean geometry or spherical geometry to spherical geometry) correspondence relation between the original space and the reduced space. We show that when the protein conformation space is represented as a Euclidean section, the embedding results are significantly improved over those using the implicit representation of protein conformations. Even though only the Isomap implementation is tested, the mathematical concept is generally applicable to other dimensionality-reduction methods. Furthermore, the application of the formal definition of protein conformation space as the quotient space of R^{3N} over rigid motions in the space can go beyond dimensionality reduction.

ACKNOWLEDGMENTS

The authors acknowledge the Scientific Computing and Visualization group at Boston University and the National Center for Supercomputing Applications for providing part of the computational resources.

REFERENCES

1. Yang M, Lei M, Brüschweiler R, Huo S. Initial conformational changes of human transthyretin under partially denaturing conditions. Biophys J 2005;89:433–443.
2. Yang M, Yordanov B, Levy Y, Brüschweiler R, Huo S. The sequence-dependent unfolding pathway plays a critical role in the amyloidogenicity of transthyretin. Biochemistry 2006;45:11992–12002.
3. Harte WE, Jr., Swaminathan S, Mansuri MM, Martin JC, Rosenberg IE, Beveridge DL. Domain communication in the dynamical structure of human immunodeficiency virus 1 protease. Proc Natl Acad Sci USA 1990;87:8864–8868.
4. Caflisch A. Network and graph analyses of folding free energy surfaces. Curr Opin Struct Biol 2006;16:71–78.
5. Ichiye T, Karplus M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins 1991;11:205–217.
6. Das P, Moll M, Stamati H, Kavraki LE, Clementi C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proc Natl Acad Sci USA 2006;103:9885–9890.
7. Kentsis A, Gindin T, Mezei M, Osman R. Calculation of the free energy and cooperativity of protein folding. PLoS One 2007;2:e446.
8. Ferguson AL, Zhang S, Dikiy I, Panagiotopoulos AZ, Debenedetti PG, James Link A. An experimental and computational investigation of spontaneous lasso formation in microcin J25. Biophys J 2010;99:3056–3065.
9. Garcia AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett 1992;68:2696–2699.
10. Jolliffe IT. Principal Components Analysis. New York: Springer; 1986.
11. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290:2319–2323.
12. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290:2323–2326.
13. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci USA 2005;102:7426–7431.
14. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. London: Academic Press; 1979.
15. Duan M, Fan J, Li M, Han L, Huo S. Evaluation of dimensionality-reduction methods from peptide folding-unfolding simulations. J Chem Theory Comput 2013;9:2490–2497.
16. Brooks BR, Brooks CL, III, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. J Comput Chem 2009;30:1545–1614.
17. Neria E, Fischer S, Karplus M. Simulation of activation free energies in molecular systems. J Chem Phys 1996;105:1902–1921.
18. Schaefer M, Bartels C, Karplus M. Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. J Mol Biol 1998;284:835–848.
19. Beauchamp KA, Bowman GR, Lane TJ, Maibaum L, Haque IS, Pande VS. MSMBuilder2: modeling conformational dynamics at the picosecond to millisecond scale. J Chem Theory Comput 2011;7:3412–3419.
20. Li M, Duan M, Fan J, Han L, Huo S. Graph representation of protein free energy landscape. J Chem Phys 2013;139:185101.
21. Li DW, Khanlarzadeh M, Wang J, Huo S, Bruschweiler R. Evaluation of configurational entropy methods from peptide folding-unfolding simulation. J Phys Chem B 2007;111:13807–13813.
22. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
23. Cormen TH, Leiserson CE, Rivest RL. Introduction to Algorithms. Cambridge, MA: MIT Press; 1992.
24. Gomory RE, Hu TC. Multi-terminal network flows. SIAM J Appl Math 1961;9:551.
25. Brown WM, Martin S, Pollock SN, Coutsias EA, Watson JP. Algorithmic dimensionality reduction for molecular structure analysis. J Chem Phys 2008;129:064118.
26. Czerminski R, Elber R. Reaction path study of conformational transitions in flexible systems: applications to peptides. J Chem Phys 1990;92:5580–5601.
27. Becker OM, Karplus M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. J Chem Phys 1997;106:1495–1517.
28. Li DW, Han L, Huo S. Structural and pathway complexity of beta-strand reorganization within aggregates of human transthyretin(105–115) peptide. J Phys Chem B 2007;111:5425–5433.

29. Krivov SV, Karplus M. Free energy disconnectivity graphs: application to peptide models. J Chem Phys 2002;117:10894–10903.
30. Krivov SV, Karplus M. Hidden complexity of free energy surfaces for peptide (protein) folding. Proc Natl Acad Sci USA 2004;101:14766–14770.
31. Carr JM, Wales DJ. Global optimization and folding pathways of selected alpha-helical proteins. J Chem Phys 2005;123:234901.
32. Evans DA, Wales DJ. Free energy landscapes of model peptides and proteins. J Chem Phys 2003;118:3891–3897.
33. Wales DJ, Miller MA, Walsh TR. Archetypal energy landscapes. Nature 1998;394:758–760.
34. Stamati H, Clementi C, Kavraki LE. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. Proteins 2010;78:223–235.
35. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera-a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–1612.