Geometric Issues in Dimensionality Reduction and Protein Conformation Space

Mojie Duan*, Li Han†, Lee Rudolph† and Shuanghong Huo*
*Gustaf H. Carlson School of Chemistry and Biochemistry
†Department of Mathematics and Computer Science
Clark University, Worcester, MA01610
Email: [mduan,lhan,lrudolph,shuo]@clarku.edu

***Abstract*—Protein conformation space ($CSpace$) plays a fundamental role in the computational study of protein conformations. With the large dimensions and complicated structures of protein $CSpace$s, dimensionality reduction allows scientists to generate lower dimensional embedding of $CSpace$ data that can be used for further analysis and information extraction. In this paper, we discuss our puzzling dimensionally reduction results of β-hairpin where the linear method PCA performed better than nonlinear methods ISOMAP and LLE. We describe our findings and show that nonlinear surfaces without certain specified properties are not necessarily better suited for nonlinear dimensionality reduction methods than linear methods. We define protein $CSpace$ represented by atom Cartesian coordinates formally as the quotient space of its ambient Euclidean space by the group of 3D rigid motions. We explain that PCA essentially uses an explicit section (as defined mathematically) to represent protein $CSpace$, while the RMSD-based ISOMAP and LLE essentially use an implicit representation that poses additional challenges to dimensionality reduction. We present an ISOMAP variant that works like PCA on a section of protein $CSpace$ and achieves better results than the original ISOMAP for β-hairpin. We also describe open problems on fundamental issues for protein $CSpace$, motivated by dimensionality reduction but relevant to general computational study of protein conformations.**

I. Introduction

Conformation space($CSpace$) is to protein folding what configuration space ($CSpace$) to robotic motion planning. In the past two decades or so, robotics researchers have developed many innovative and successful methods for planning motions and analyzing $CSpace$s, as partly reflected in the books [15, 4, 16] and references therein. Among many families of planning methods, Probabilistic Roadmap Methods (PRM) [13] have had significant impacts in diverse fields; applications in computational structural biology include protein folding and drug design [20, 21, 18]. In protein folding studies, PRM uses protein conformation samples to build a discrete graph presentation of the inherently continuous protein $CSpace$, then extracts biologically meaningful information like protein folding pathways from the graph. Recently, PRM was also incorporated into ISOMAP-based dimensionality reduction study of protein conformation space [5, 22].

Mathematically, dimensionality reduction aims at mapping a set in a high-dimensional ambient space to a lower dimensional one while preserving the set's structure. It is widely used in fields dealing with data in high dimensional spaces. Reducing a data set to a lower dimensional space and working therein provides many advantages such as reduced complexity and improved efficiency for analysis and visualization. However, the usefulness of information extracted from the reduced space depends on whether the set has consistent properties in its two ambient spaces. For protein conformations, in theory the original data set is the continuous, high-dimensional protein conformation space, but in practice it is a discrete (finite) set of protein conformation samples. Dimensionality reduction methods are commonly used to generate one or more 2D or 3D embeddings of the sample set. Scientists then observe and analyze these embeddings to make inferences about properties of the theoretical $CSpace$, such as the number of free energy minima, their locations, and the energy barriers between them.

Dimensionality reduction methods can be classified as linear or nonlinear. Principle Component Analysis (PCA) [11], an important linear method, finds a low-dimensional embedding of the sample data points that best preserves their variance. It is well suited for data lying on or near a linear subspace of the original ambient space. Multi-Dimensional Scaling (MDS) [1] takes an input matrix giving distances (dissimilarities) between pairs of points and outputs a coordinate matrix whose configuration best preserves the distances. It is equivalent to PCA when the distance between two points is their Euclidean distance. ISOMAP (Isometric feature Mapping) is a nonlinear method that "builds on top of classical MDS and seeks to preserve the intrinsic geometry of the data, as captured in the geodesic manifold distances between all pairs of data points" [23]; in the spirit of PRM's roadmap, ISOMAP creates a graph reflecting the connectivity of $CSpace$, then uses the graph-based all-pair shortest distances as the estimate of their geodesic distances. Another nonlinear dimensionality reduction method, Locally Linear Embedding (LLE) [19], is for data sets forming manifolds. It exploits the fact that no matter how non-linear a manifold may be globally, it is covered by approximately linear local neighborhoods. In effect, LLE does a different linear dimensionality reduction near each point, then combines them all with minimal discrepancy.

PCA, ISOMAP and LLE have all been adapted for the study of protein conformations [10, 5, 14, 3, 22]. It is widely assumed that the protein conformation spaces are nonlinear and, thus, nonlinear dimensionality reductions methods can get better results. This is indeed the case in the solid work reported in [5, 14, 3, 22].

II. Puzzling Data

We recently applied dimensionality reduction in our study and reported in the paper [6] our evaluation results of various methods to the second β-hairpin of the B1 domain of streptococcal protein G, which has sequence G-E-W-T-Y-D-D-A-T-K-T-F-T-V-T-E without blocking groups at the termini. A total of 200,000 conformations were generated from molecular dynamics simulation, each represented by a 480-dimensional vector of atom Cartesian coordinates. We used the publicly available dimensionality reduction code, with RMSD as the distance metric, in computing the graph for ISOMAP and local neighborhood for LLE. Our results were very puzzling: *PCA performed better than LLE and rmsdISOMAP*, as measured in the residual variance (Fig. 1) as well as other measures such as free energy profile. Please refer to the paper [6] for more information. (Note that residual variance is one commonly used numerical measure for variance and similarity in inter-point distances between two embeddings. The lower the residual variance, the more similar the inter-point distances, and implicitly, the more similar the geometry of the embeddings.)

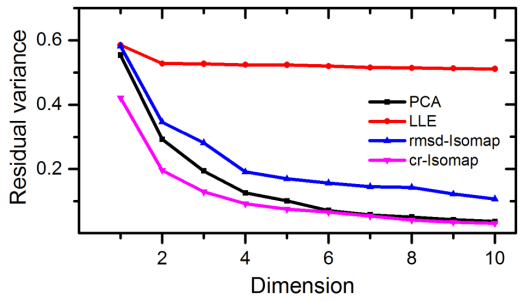

Fig. 1. Residual Variance of Embedding Results for β-Hairpin

Our β-hairpin results contradict the common understanding that nonlinear dimensionality reduction methods generally work better than linear methods for nonlinear manifolds, as supported by the solid prior work. So we have conducted a serious investigation into relevant issues, and developed an explicit representation of the conformation space and an ISOMAP variant called crISOMAP [7]. As shown in Fig. 1, with respect to β-hairpin and residual variance, crISOMAP performed better than rmsdISOMAP and LLE in the top 10 dimensions, and is better than PCA in dimensions 5 or lower and comparable in dimensions 6 or above. In this paper, we explain some subtle issues for dimensionality reduction and protein conformation space. We also present our recent theoretical findings and several open questions that aim at helping us develop a better understanding of the mathematical properties of protein conformation space and select as well as adapt dimensionality reduction methods.

III. Geometric Issues

As we investigated several possible explanations for the unexpected results for β-hairpin, we eventually rediscovered what is clearly stated in the ISOMAP paper but had somehow escaped our attention before.

> Just as PCA and MDS are guaranteed, given sufficient data, to recover the true structure of linear manifolds, Isomap is guaranteed asymptotically to recover the true dimensionality and geometric structure of a strictly larger class of nonlinear manifolds. Like the Swiss roll, these are manifolds whose **intrinsic geometry is that of a convex region of Euclidean space**, but whose ambient geometry in the high-dimensional input space may be highly folded, twisted, or curved. For non-Euclidean manifolds, such as a hemisphere or the surface of a doughnut, Isomap still produces a globally optimal low dimensional Euclidean representation, as measured by Eq. 1. [23, p. 2321; our boldface emphasis]

So what about manifolds whose intrinsic geometry is not a convex region of Euclidean space? How will the dimensionality reduction methods perform? To answer these questions and more, we applied dimensionality reduction methods to representative samples from several example surfaces; Figs. 2 and 3 illustrate our results for two variants of spherical surfaces. Note that the spherical surfaces are simple and can serve as good examples of manifolds since they don't satisfy the desired manifold properties for ISOMAP in various ways.

- A spherical surface, when using the Euclidean distance and considered as a subset in the Euclidean ambient space, is Euclidean but not convex. This is essentially how a spherical surface is treated by PCA.
- A spherical surface, when using the spherical distance and considered as a subset in the spherical ambient space, is non-Euclidean but convex. This is essentially how a spherical surface is treated by ISOMAP and LLE.
- A truncated sphere is not convex in Euclidean or spherical geometry.

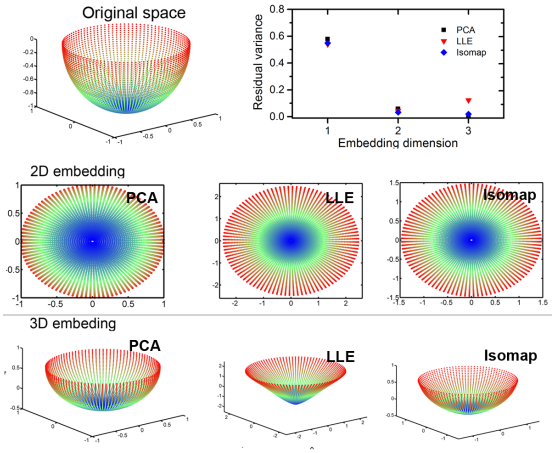

Fig. 2. Dimensionality Reduction Results of a Hemisphere by PCA, LLE and ISOMAP

As Figs. 2 and 3 show, PCA and ISOMAP achieve similar performances and better residual variance than LLE for both examples. Thus, **nonlinear surfaces without the properties specified for ISOMAP are not necessarily better suited for nonlinear dimensionality reduction methods than linear methods**.

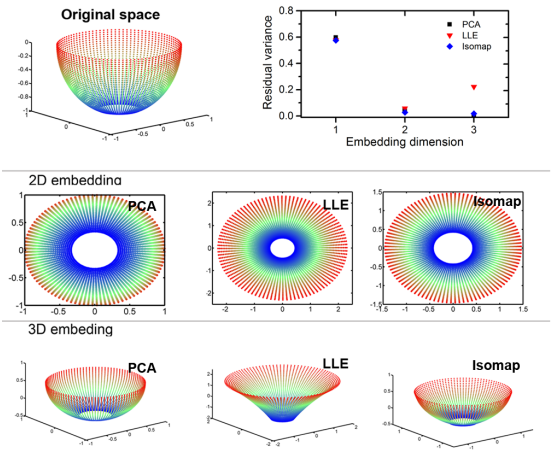

Fig. 3. Dimensionality Reduction Results of a Zone (Truncated Hemisphere) by PCA, LLE and ISOMAP

As for LLE, it has solid theoretical underpinnings in differential geometry, based on the property that a manifold resembles Euclidean space near each point. More precisely, each point of an n-dimensional manifold has a neighborhood homeomorphic to the Euclidean space of dimension n. Lines and circles, but not figure 8 or letter X, are 1-dimensional manifolds. Although near each point, a manifold resembles Euclidean space, globally a manifold might not, as in the case of spherical surfaces. The relatively poor performance of LLE in the two hemisphere examples seems to be related to the non-uniform distribution of points on the surface: the density of the point samples is higher at the polar area (the bottom) than at the equator. Also the region with more densely distributed points seems to have a smaller ratio between its area and the area of the overall surface after the reduction. This trend seems similar to the observation in [17] that, if a data set has multiple clusters of points connected by single straight lines between clusters, then LLE compresses each cluster to a single point. This suggests that LLE is better suited for embedding surfaces with one well-connected component instead of multiple components with fewer connections between them. We are continuing investigations of these issues.

We also want to point out that the intrinsic dimension and the embedding dimension of a manifold are not necessarily the same. For the Swiss roll, a benchmark surface widely used in the evaluation of nonlinear dimensionality reduction methods, these two dimensions are the same. On the other hand, a spherical surface has intrinsic dimension 2 but can only be embedded in Euclidean space of dimension 3 or larger; the same is true for other surfaces like the torus and Klein bottle. One well-known theorem in differential topology, the so-called Strong Whitney Embedding Theorem, states that any smooth real m-dimensional manifold (required also to be Hausdorff and second-countable) can be smoothly embedded in the real $2m$-space R^{2m}. The practical implication is that, to achieve a "good" embedding with "low" residual variance value, the m-dimensional conformation space of a protein might need to be embedded in an ambient Euclidean space higher than m. This raises a question: are the intrinsic and embedding dimensions of protein conformation space biologically meaningful?

IV. Protein Conformation Space

In the common setting for dimensionality reduction, there is a data set S and an ambient Euclidean space E, generally with each point in S represented by one coordinate vector in E. So the data set S starts with an explicit representation in E. However, this is not the case for protein conformations represented by atom Cartesian coordinates. For a protein with n atoms, the Cartesian coordinate based approach uses $3n$-dimensional vectors to represent conformations and thus puts the protein conformation space in an ambient Euclidean space R^{3n}. But each conformation is represented by all the different $3n$ coordinates that are related to it through rigid motions of R^3, the ambient physical space of the protein. Unless one and only one $3n$-vector for each conformation is explicitly used, conformation space is only implicitly represented as a subset in R^{3n}. This implicit nature of the Cartesian coordinate representation of protein conformations and the wide use of $3n$-vectors in conformation-related computation makes the protein conformation space fundamentally different from the explicit point representation of data sets used in the common setting of dimensionality reduction, and may blur the difference between protein conformation space and the ambient space R^{3n}. To clarify this and related points, we introduce some useful mathematical formalism.

A. Formal Mathematical Definition

Given a set S and an equivalence relation $\equiv$, the *quotient set* $S/\equiv$ consists of equivalence classes in S with respect to the equivalence relation $\equiv$. In other words, each element s of S determines one element of $S/\equiv$, namely, the set of all points in S equivalent by $\equiv$ to s. If S has an additional structure, such as a topology or a geometry, $S/\equiv$ can usually be equipped with a structure of the same kind.

For protein conformations, S is the set of possible placements of the given protein in physical space R^3. We call such a placement a *pose* of the protein. Two poses are equivalent if a rigid motion of R^3 carries one to the other. A conformation of the protein is an equivalence class of poses. In these terms, the protein conformation space $CSpace$ is formally defined as the quotient space of pose space by 3D rigid motion equivalence.

$$\text{protein } CSpace = PoseSpace/\text{3D}RigidMotion \qquad (1)$$

This is conceptually identical to the definition of robot deformation space [9]: in both cases, rigid motions are factored out from the (protein or robot) pose space, creating a quotient space that is the natural arena for formulating and analyzing those properties of (protein or robot) configurations which are intrinsic, i.e., independent of rigid motions.

To apply dimensional reduction methods, protein $CSpace$ must be equipped with a metric (which then gives it a topology and a geometry). **There is more than one way to do this.** For example, so-called internal coordinates represent protein $CSpace$ as a subset of $S_{k,\ell,m} = R^k \times (S^1)^\ell \times (S^1)^m$, where S^1 is a 1-dimensional circle and a conformation is fully specified

by k bond lengths, ℓ bond angles, and m torsion angles; similar coordinates are used in robotics to represent $CSpace$ of an open or closed chain with spherical joints (see [24]). The metric that $CSpace$ inherits from $S_{k,\ell,m}$ generally has globally curved geometry. By contrast, an alternative set of "interpoint distance coordinates" that has proved useful in robotics [8] often gives $CSpace$ a global metric that is piecewise flat (these coordinates have not yet been applied to proteins).

In this paper we focus on the atom Cartesian coordinate representation of poses, so that for a protein with n atoms, pose space is R^{3n} and

$$protein\,CSpace = R^{3n}/3\mathrm{D}RigidMotion \tag{2}$$

as illustrated schematically in Fig. 4. In subsection IV-C we address the question of equipping protein $CSpace$, in this representation, with a metric.

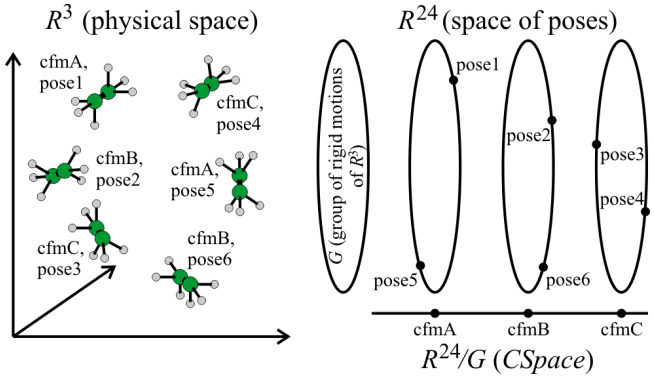

Fig. 4. (a) Six Poses Representing Three Conformations. (b) Schematic representation of $CSpace$ as R^{3n}/G.

B. Dimensionality Reduction Issues

We briefly describe how we applied the dimensionality reduction methods to our model system β-hairpin and then discuss challenging issues with dimensionality reduction of protein conformation space. Please refer to papers [6, 7] for more information including discussions of the choice and effects of parameter values.

[Conformation Data] CHARMM [2] (version c31b1) was used in the 4 μs equilibrium folding-unfolding simulation of β-haripin at $360K$, with parm19 polar hydrogen potential function and EEF1 implicit solvation model22. Snapshots were saved every 20 ps and a total of 200,000 conformations were generated. The atom coordinate vectors have 480 dimensions.

[PCA] For each atom coordinate vector, we computed its best alignment with the native fold and used all such best aligned coordinate vectors to form a matrix of the original points and submit the matrix to the standard PCA computation (ptraj module in AmberTools (v9.0)).

[rsmdISOMAP] We computed $RMSD$ between every pair of atom coordinate vectors and chose 20 closest neighbors for each conformation to build a graph, with $RMSD$ values as edge weights. To reduce the computation amount, we used the landmark idea given in [5] and picked 5000 landmarks, one in every 800 ps of the simulation trajectory. We also computed the largest connected-component of the graph, and removed all conformations not in this component. We then computed the shortest path lengths from each of the remaining 179,774 conformations to all the remaining 4491 landmarks, as an approximation of the geodesic distances, and used the resulting distance matrix in the ISOMAP computation.

[LLE] We used a RMSD cutoff value of $3.0\mathring{A}$ as a neighboring criterion and only kept the conformations having at least 20 neighbors for embedding. This led to a total of 179,629 conformations for LLE, and the computation was done by the publicly available LLE code.

[Issues] We now elaborate on the issues related to the implicit representation of protein $CSpace$ and challenges for dimensionality reduction. With pairwise RMSD-based computation, each conformation is generally represented by multiple poses of the same conformation. Consider the example shown in Fig. 4, and assume that the best alignments for conformation pairs (cfmA, cfmB), (cfmB, cfmC) and (cfmC, cfmA) are (pose1, pose2), (pose6, pose4) and (pose4, pose1) respectively. Then cfmB already uses two poses labeled pose 2 and pose6 in this small example. Intuitively, when using $RMSD$, each conformation corresponds to a cloud of points in the ambient space R^{3n}. However, with dimensionality reduction, the clouds of points in the ambient space need to be mapped to individual points in the reduced space. This problem of having multiple coordinates for each point is fundamentally different from the original problem setting of having one coordinate for each point setting. It also raises questions about whether the protein conformation space is Euclidean, convex, or even a manifold.

C. (Mathematical Defined) Section of Protein CSpace

In contrast to rmsdISOMAP and LLE, our PCA process for β-hairpin computed the best alignments for all conformations with respect to the native fold and used the best aligned conformations in the embedding computation. More precisely, the native fold conformation was represented by one of its poses (arbitrary, but fixed), and for each other conformation that one of its poses which is best aligned to the fixed pose of the native fold. Mathematically, selecting one point from each equivalence class defines a *section* for the quotient space. So, for PCA with a reference pose r, its computation is really done in the corresponding section of $CSpace$, which we call $CSection(r)$ for short.

$$\begin{aligned} &\text{protein } CSection(r): \\ &\text{protein } CSpace \text{ represented by a section wrt ref. cfm } r \\ &= \{p \in R^{3n} \mid \text{ best alignment matrix } (p, r) = I\} \end{aligned} \tag{3}$$

In words, the last line of definition (3) states that a $3n$-vector p is the best aligned pose of the corresponding conformation with respect to the reference pose r of the reference conformation if and only if the best alignment matrix for p and r is the identity matrix I.

Since PCA worked well for β-hairpin and the $CSection$ representation follows the common setting for dimensionality reduction (with one and only one coordinate vector for each

conformation), we also adapted the ISOAMP method for $CSection$ as follows and named the corresponding method crISOMAP, short for common reference ISOMAP.

[crISOMAP] The process for crISOMAP was very similar to that for rmsdISOMAP. The main difference was that we initially computed the best alignments of all 200k conformations with respect to the native fold and then used the pairwise Euclidean distances of these best aligned conformations as their distances to identify 20 neighbors for each conformation. We then built a graph with edge weights being the Euclidean distances between neighboring conformations. We also just kept the conformations and landmarks in the largest connected component for the embedding computation.

As shown in Fig. 1 for residual variance of the embedding results of β-hairpin, crISOMAP did better than rmsdISOMAP in the first 10 embedding dimensions, better than PCA in the first five (5) dimensions and comparably for dimensions 6 through 10. Recall that, from the computational perspective, the only difference between crISOMAP and rmsdISOMAP is that crISOMAP uses one $3n$-vector for each conformation and thus treats protein $CSpace$ as an explicit subset in the ambient space R^{3n}. This difference is very likely a significant factor leading to the improved ISOMAP performance, but thorough investigation needs to be done to develop a better understanding of the protein conformation space and dimensionality reduction methods. Here we present one result on the property of the set of best aligned poses with respect to one reference conformation.

Theorem 1: Given a point set of n points in d dimensions and one reference pose r of a reference conformation, the best aligned poses of other conformations that have the same chirality as r form a linear set. In other words, if p and q have the same chirality as r and are both best aligned wrt r, then $sp + tq$, with $s, t \in R$, is also best aligned wrt r.

Proof: Given a point set, aligning poses of conformations based on point coordinates has a translation part and a rotation part. The translation part is to put the centroid at the origin of the coordinate system by subtracting the coordinates of the respective centroid from the point coordinates. It is easy to see that this is a linear constraint.

For the rotation part, denote the reference pose r as well as the best aligned poses p and q as matrices of dimension $d \times n$. The Kabsch Algorithm [12] computes the optimal rotation matrix for p and r as follows.

Define the covariance matrix $A = pr'$, where r' is the matrix transpose of r. Denote its singular value decomposition as $A = V_A S W'_A$, where V_A and W_A are $d \times d$ orthogonal matrices satisfying $V_A V'_A = I_d$, etc.

Then determine whether we need to modulate the rotation matrix to deal with chirality (handedness) of point sets by computing $c = sign(\det(W_A V'_A))$, and finally, compute the optimal rotation matrix U_A as

$$U_A = W_A \times diag(1, \ldots, 1, c) \times V'_A$$

where $diag(1, \ldots, 1, c)$ is the diagonal matrix with all diagonal elements except the last equal to 1, the last equal to c.

Since p and r are assumed to have the same chirality (c is 1), the diagonal matrix in the equation above becomes the identity. Further, since p is assumed to be best aligned wrt r, this optimal rotation matrix is I_d itself. So we have

$$U_A = W_A \times V'_A = I_d$$

From here, it is easy to prove that $U_A = I_d$ if and only if $W_A = V_A$ and A is symmetric. The same property holds for the best aligned coordinate sets q and r, with their covariance matrix $B = qr'$ being symmetric. Thus the linear combination $sp + tq$ also has a symmetric covariance matrix with r and an identify matrix for rotation alignment. In other words, $sp + tq$ is also best aligned with respect to r. ■

Note that Theorem1 is applicable to point sets in an arbitrary dimension. In terms of protein conformations, it clears establishes that, if two poses p and q are best aligned with respect to r, their convex combination $\lambda p + (1 - \lambda)q$, with $\lambda \in [0, 1]$, is also best aligned. This means that **protein $CSection(r)$ as defined in equation 3 is convex in its ambient space** R^{3n}, so long as we disregard constraints on bond lengths, bond angles and torsion angles. These bond constraints are not necessarily satisfied since we don't know that, given p and q satisfying these constraints, whether $\lambda p + (1-\lambda)q$ also satisfies the constraints. In other words, it is not clear that, given two valid conformations best aligned to a common reference, their convex combinations produce valid conformations. If this turns our to be true, it would give the nice property of convexity to protein $CSection(r)$ in its ambient space R^{3n}, even with the bond constraints considered. Such a property would be useful for the analysis and sampling of protein conformation space.

While the properties of $CSection(r)$ in the ambient space might contribute to the better than expected performance of PCA over the nonlinear methods, we need to have a better understanding of the intrinsic geometry of protein conformation space and its section representation. This is important for computational study of protein conformations in general, including the applicability and adaptation of nonlinear dimensionality reduction methods, since these methods mainly rely on the intrinsic geometry, or more precisely, discrete approximation of the intrinsic geometry, in their computation. Questions related to the intrinsic topology and geometry of protein conformation space remain largely open.

V. SUMMARY

Protein conformation spaces plays a similar role for protein folding as robot configuration space for motion planning. Mathematically, when described in atom Cartesian coordinates, the $CSpace$ of a protein with n atoms is the quotient space of the ambient space R^{3n} by the group of 3D rigid motions, which is similar to the robot deformation space. This quotient space definition formally and explicitly captures the notion that the set of all $3n$-vectors related to each other through rigid motions corresponds to one conformation. It clearly distinguishes the protein $CSpace$ from its ambient Euclidean space and raises important questions about protein $CSpace$ on its representation, topological and geomet-

rical properties. Considering the fundamental roles of protein $CSpace$, the formal definition and progresses on related mathematical issues can be influential in a broad range of topics arising in the computational study of protein conformations.

In this paper, we focused on the dimensionality reduction of protein conformation space. We briefly described our puzzling results with the linear dimensionality reduction method PCA achieving better performance for our model system β-hairpin than the nonlinear methods rmsdISOMAP and LLE. These results are not consistent with the common belief that the conformation space of a protein is generally a nonlinear hypersurface and nonlinear dimensionality reduction methods are expected to do better than linear methods. We reported our recent findings and showed that nonlinear surfaces without certain specified properties are not necessarily better suited for nonlinear dimensionality reduction methods than linear methods. For protein conformation space, we explained that PCA essentially uses an explicit section to represent the protein $CSpace$, while the RMSD-based ISOMAP and LLE essentially use an implicit representation, which poses additional challenges to dimensionality reduction. We also described an ISOMAP variant that works on a section of protein $CSpace$ like PCA and achieved better results than the original rmsdISOMAP for β-hairpin.

ACKNOWLEDGMENTS

This work was partially funded by the grants National Institutes of Health R01-GM088326 and NSF IIS-0713335.

REFERENCES

[1] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[2] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[3] W. M. Brown, S. Martin, S. N. Pollock, E. A. Coutsias, and J. P. Watson. Algorithmic dimensionality reduction for molecular structure analysis. *J. Chem. Phys.*, 129(6): 064118–1–13, 2008.

[4] H. Choset, W. Burgard, S. Hutchinson, G. Kantor, L. E. Kavraki, K. Lynch, and S. Thrun. *Principles of Robot Motion: Theory, Algorithms, and Implementation*. MIT Press, 2005.

[5] P. Das, M. Moll, H. Stamati, L.E. Kavraki, and C. Clementi. Low-dimensional free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci.*, 103(26):9885–9890, 2006.

[6] M. Duan, J. Fan, M. Li, L. Han, and S. Huo. Evaluation of dimensionality reduction methods from peptide folding-unfolding simulations. *J. Chem. Theory Comput.*, 9(5):2490–2497, 2013. doi: 10.1021/ct400052y.

[7] M. Duan, M. Li, L. Han, and S. Huo. Euclidean sections of protein conformation space and their implications in dimensionality reduction. *Proteins: Struct., Funct.,, Bioinf.*, 2014. doi: 10.1002/prot.24622.

[8] L. Han and L. Rudolph. A unified geometric approach to inverse kinematics of a spatial chain with spherical joints. In *Proc. Int. Conf. Robotics Automation*, pages 4420–4427, 2007. doi: 10.1109/ROBOT.2007.364160.

[9] L. Han, L. Rudolph, J. Blumenthal, and I. Valodzin. Convexly stratified deformation space and efficient path planning for a planar closed chain with revolute joints. *Int. J. Robot. Res.*, 27:1189–1212, 2008.

[10] T. Ichiye and M. Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Struct., Funct., Bioinf.*, 11(3):205–217, 1991.

[11] I. T. Jolliffe. *Principal Components Analysis*. Springer, New York, NY, 1986.

[12] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, 32(922): 827–828, 1978.

[13] L. Kavraki, P. Svestka, J. C. Latombe, and M. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.

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

[15] J. C. Latombe. *Robot Motion Planning*. Kluwer Academic Publishers, Boston, MA, 1991.

[16] S. M. LaValle. *Planning Algorithms*. Cambridge University Press, Cambridge, U.K., 2006.

[17] N. Melchior. Manifold Learning: ISOMAP and LLE. *Presentation for Carnegie Mellon 16-721: Advanced Machine Perception, Spring 2006*.

[18] M. Moll, D. Schwarz, and L. Kavraki. Roadmap methods for protein folding. *Methods in Molecular Biology*, 413: 219239, 2008.

[19] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

[20] A.P. Singh, J.C. Latombe, and D.L. Brutlag. A motion planning approach to flexible ligand binding. In *Proc Int Conf Intelligent Syst for Molecular Biology (ISMB)*, page 252261, 1999.

[21] G. Song and N. M. Amato. Using motion planning to study protein folding pathways. In *Proc. Int. Conf. Comput. Molecular Biology*, pages 287–296, 2001.

[22] H. Stamati, C. Clementi, and L. E. Kavraki. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins*, 78 (2):223–235, Jul 2010. doi: 10.1002/prot.22526.

[23] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[24] J.C. Trinkle and R.J. Milgram. Complete path planning for closed kinematic chains with spherical joints. *Int. J. Robot. Res.*, 21(9):773–789, 2002.