Topological methods for exploring low-density states in biomolecular folding pathways

Yuan Yao,[1,a)] Jian Sun,[2,b)] Xuhui Huang,[3,c)] Gregory R. Bowman,[4,d)] Gurjeet Singh,[1,e)] Michael Lesnick,[5,f)] Leonidas J. Guibas,[2,g)] Vijay S. Pande,[6,h)] and Gunnar Carlsson[1,i)]
[1]*Department of Mathematics, Stanford University, Stanford, California 94305, USA*
[2]*Department of Computer Science, Stanford University, Stanford, California 94305, USA*
[3]*Department of Bioengineering, Stanford University, Stanford, California 94305, USA*
[4]*Biophysics Program, Stanford University, Stanford, California 94305, USA*
[5]*Insititute of Computational and Mathematical Engineering. Stanford University, Stanford, California 94305, USA*
[6]*Department of Chemistry, Stanford University, Stanford, California 94305, USA*

(Received 5 January 2009; accepted 3 March 2009; published online 14 April 2009)

Characterization of transient intermediate or transition states is crucial for the description of biomolecular folding pathways, which is, however, difficult in both experiments and computer simulations. Such transient states are typically of low population in simulation samples. Even for simple systems such as RNA hairpins, recently there are mounting debates over the existence of multiple intermediate states. In this paper, we develop a computational approach to explore the relatively low populated transition or intermediate states in biomolecular folding pathways, based on a topological data analysis tool, MAPPER, with simulation data from large-scale distributed computing. The method is inspired by the classical Morse theory in mathematics which characterizes the topology of high-dimensional shapes via some functional level sets. In this paper we exploit a conditional density filter which enables us to focus on the structures on pathways, followed by clustering analysis on its level sets, which helps separate low populated intermediates from high populated folded/unfolded structures. A successful application of this method is given on a motivating example, a RNA hairpin with GCAA tetraloop, where we are able to provide structural evidence from computer simulations on the multiple intermediate states and exhibit different pictures about unfolding and refolding pathways. The method is effective in dealing with high degree of heterogeneity in distribution, capturing structural features in multiple pathways, and being less sensitive to the distance metric than nonlinear dimensionality reduction or geometric embedding methods. The methodology described in this paper admits various implementations or extensions to incorporate more information and adapt to different settings, which thus provides a systematic tool to explore the low-density intermediate states in complex biomolecular folding systems. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3103496]

I. INTRODUCTION

The folding of biomolecules is a classic biophysical problem. Proteins and nucleic acids are synthesized as linear polymer chains. They must then spontaneously and rapidly fold into their three-dimensional native states. The folding process is determined by the underlying free energy landscape. These landscapes may have local minima on large-scale energy barriers corresponding to intermediates or misfolded states. Characterizing these states is critical for a full understanding of biomolecular folding. Experimental studies may point to the existence of such states but are usually unable to provide high resolution structural information due to the transience and/or heterogeneity of such states. Computer simulations have proven useful for sampling this complex high-dimensional space while yielding structures at full-atom resolution. However, these simulations tend to generate millions of configurations. The volume and high-dimensional nature of the output make it extremely difficult to discern the structure of the data.

One common approach to dealing with computer simulation results is to apply K-means clustering to the entire data set. However, K-means clustering suffers from a number of important limitations. First, it is limited by the need to specify the number of states from the beginning. Second, it tends to create spherical states. The relevant states of the free energy landscape, on the other hand, may be nonconvex. In this case, K-means clustering will tend to lump unrelated configurations together or split related configurations into separate states. This limitation may be overcome by splitting the configurations into many small states and grouping them together using various metrics that allow nonconvex states,

[a)] Electronic mail: yuany@stanford.edu.
[b)] Electronic mail: sunjian@stanford.edu.
[c)] Electronic mail: huangx@stanford.edu.
[d)] Electronic mail: gbowman@stanford.edu.
[e)] Electronic mail: gurjeet@stanford.edu.
[f)] Electronic mail: mlesnick@stanford.edu.
[g)] Electronic mail: guibas@cs.stanford.edu.
[h)] Electronic mail: pande@stanford.edu.
[i)] Electronic mail: gunnar@math.stanford.edu.

such as in Ref. 1. There is another widely used clustering method, single linkage, which may overcome these issues in K-means. Unfortunately, identifying sparsely populated intermediate states is still difficult. Simulation data tend to be heavily dominated by the most stable states, such as the folded and unfolded states, and single-linkage clustering of the entire data set tends to pick up densest states only and hardly distinguish the intermediates from noise.

Recently, geometric embedding techniques, such as nonlinear dimensionality reduction,[2–7] have been explored as a means to overcome the dimensionality hurdle in complex biomolecular systems. For example, ISOMAP (Ref. 2) has been applied to protein folding[8] and Laplacian eigenmap[4] has been applied to the dynamics of biological networks.[9] This class of techniques maps the data in high-dimensional spaces to a low-dimensional space by preserving some local/global metric relationship among neighboring data points. In this way, one can easily visualize data and possibly gain important insights. For instance, the new embedding coordinates may be biologically relevant reaction coordinates.[8] However, the performance of these geometric embedding techniques will suffer from the high degree of heterogeneity in distribution and be sensitive to the choice of the distance metric.

One efficient strategy to address these issues is to stratify the data into density level sets and study its topological features such as clustering which are less sensitive to the metric than geometric methods. High-density levels will contain the dominant states, such as the folded and unfolded states, while less populated states, such as intermediates, will occupy the low-density levels. Clustering on level sets of similar density will be less affected by the distributional heterogeneity and thus effectively disclose structural information about intermediates. This idea of stratification is reminiscent of Morse theory, which provides a general machinery for studying the topology of high-dimensional manifolds by looking at level sets of some nicely behaved function.[10] Inspired by Morse theory, Singh *et al.*[11] recently introduced MAPPER, a topological data analysis tool for high-dimensional data sets.

MAPPER is a way to visualize and cluster high-dimensional data. In its simple form, a filter function is used to decompose the data into overlapping level sets and clustering is then carried out in each of them. A graph is then generated by connecting clusters in neighboring level sets with an edge if they have nonempty overlapping. If an energy function is taken as the filter, the graph generated by MAPPER will provide the same kind of topological information as a disconnectivity graph of the energy landscape.[12] Moreover, in MAPPER, one can design other filter functions besides energy, so that this method can be applied to study a wide variety of data sets even including nonequilibrium simulation data. In its extended form, MAPPER can return a simplicial complex with high-dimensional topological information about the data. The method is computationally efficient and amenable to parallelization.

In this work we demonstrate the applicability of MAPPER in its simple form to the biomolecular folding problem. We begin with a discussion of MAPPER itself from a perspective of Morse theory and then present the details of a filtering function that is well suited for biomolecular folding problems, the conditional density filter. This filter puts important weights on conformations generated from simulations. If every conformation is weighed equally, the filter is an estimator of the density of sampled conformations. Furthermore, by weighing more heavily those conformations close to a state of interest, the filter facilitates the identification of intermediate states leading up to it. We then describe the use of single-linkage clustering within level sets to allow the identification of an unspecified number of nonconvex states. Finally, we discuss the application of MAPPER to the folding of a small RNA hairpin, which gives some structural evidence from computer simulations in support of the multistate hypothesis.[13] The biological implications of the MAPPER results are discussed in Ref. 14 elsewhere. We also briefly discuss the advantages of MAPPER over nonlinear dimensionality reduction techniques. In the future we hope to explore the combination of those geometric embedding techniques with MAPPER in order to take advantage of the strengths of both approaches.

II. MATERIALS AND METHODS

A. MAPPER: A tool for topological data analysis

One way to reduce the computational complexity in the study of massive data sets is to decompose the data by classifying the data into groups and doing analysis on each of the group individually instead of performing analysis on the whole. This strategy is amenable for parallel computation, which is particularly important for studies of biomolecular folding, where a great amount of configurations are normally generated.

Here we pursue this idea in the particular case where the decomposition is induced by the choice of some filter function on the data set, $h:\mathcal{X}\rightarrow\Omega$. In this paper, we will only consider filters that take values in the real line, although the MAPPER methodology is equally applicable for filter functions taking values in higher dimensional space, or even spheres, tori, or any other topological space. With this choice, we introduce MAPPER from a perspective of Morse theory, which differs from the original paper[11] but discloses a deeper inspiration.

Morse theory[10] tells us that when $h:\mathcal{X}\rightarrow\mathbb{R}$ is some nicely behaved function, topological information of $\mathcal{X}$ can be inferred from the level sets $h^{-1}(\omega)$. Such nice functions are called Morse functions, i.e., those smooth functions with only nondegenerate critical points; in other words, the Hessian at each critical point where the gradient vanishes has full rank. Morse functions are generic in the sense that they are dense in the space of smooth functions, as well of continuous functions. Hence every continuous function can be approximated arbitrarily well by Morse functions. Morse theory is an extremely powerful tool to analyze the topology of high-dimensional manifolds, which lies in the heart of proving the celebrated Poincare conjecture of dimension no less than 5.[15]

The simplest example in this spirit may be *Reeb graph*,[16] by contracting to points the connected components within level sets $h^{-1}(\omega)$, illustrated as Fig. 1(a). This simple scheme

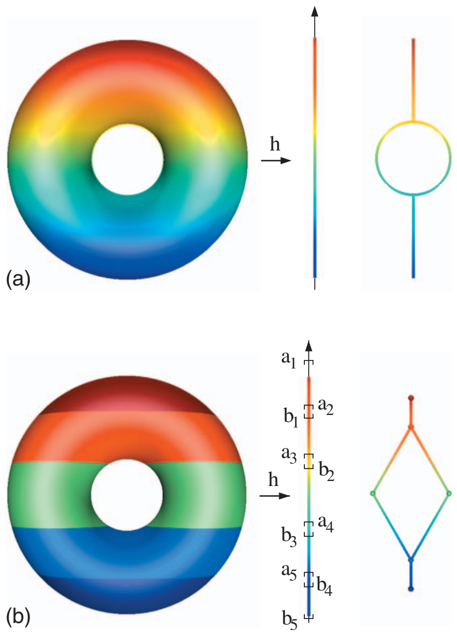

FIG. 1. (Color) (a) Construction of Reeb graph; (b) construction of MAPPER. h maps each point on torus to its height.

turns out to be useful in various fields under different names, e.g., contour trees in computational geometry[17] and cluster trees in statistics.[18–20]

MAPPER (Ref. 11) extends this construction to incorporate the discrete setting where $\mathcal{X}$ is a finite set of data points in high-dimensional spaces or metric spaces. First, instead of working with the level set of a single value which is difficult to capture in discrete settings, MAPPER considers the preimage of subinterval $h^{-1}([a,b])$. Second, it replaces by clustering the contraction of connected components in continuous settings. Specifically, the procedure of MAPPER used in this paper is as follows.

(1) *Level-set formation.* Cover the range of $h:\mathcal{X}\to\mathbb{R}$ by a set of subintervals which overlap in neighbors, i.e., $U_i=[a_i,b_i]$ with $U_i\cap U_{i+1}\neq\emptyset$ and $U_i\cap U_j\cap U_k=\emptyset$, and stratify $\mathcal{X}$ into level sets by taking inverse images $h^{-1}([a_i,b_i])$.

(2) *Clustering.* On each level set $h^{-1}([a_i,b_i])$, construct the connected components or point clusters.

(3) *Graph representation.* Represent each component or cluster by a node. Add an edge between a node pair whenever they have nonempty intersection.

MAPPER thus returns an undirected graph representing the connectivity information between data clusters across level sets $h^{-1}([a_i,b_i])$. See the example in Fig. 1(b). Note that those degree-one nodes lie in the intervals containing local minima/maxima and the branching (degree-three) nodes lie in the intervals with saddle points, a sort of critical points.

More generally, if the filter value range Ω takes some higher dimensional space or other topological spaces, MAPPER may return a simplicial complex which is, however, not pursued in this paper. This construction can easily yield a multiresolution structure by choosing subintervals of different granularities, which helps handle noise.

The key choice in MAPPER will be the filter map $h:\mathcal{X}\to\Omega$. In fact, the name, MAPPER, was coined to emphasize the importance of choosing such a map. There is no universal scheme for this choice, which may vary from application to application. In Ref. 11 some examples are presented with the choice of density function and a certain eccentricity function measuring data depth as filters. In the following, we will discuss it in detail in the setting of biomolecular folding problems, with a particular example in RNA hairpin folding.

B. MAPPER design in biomolecular folding

Simulation data in biomolecular systems produce massive data in high-dimensional space and exhibit heterogeneity in distributions. The general procedure of MAPPER above is adapted toward such challenges. The first crucial design is to construct filters based on conditional density functions estimated from the data, which effectively enable us to focus on important local regions in configuration spaces and separate less populated pathways from the overwhelmed states. In clustering we choose the single-linkage method to capture possibly nonconvex clusters. Below we give a detailed description on these particular implementations.

1. Conditional density filters for MAPPER

Our key construction of filters here is based on conditional density functions estimated from data, conditioning on the states of interests. For example, in the study of folding process, we extract conformations from folding events and focus on the region close to folded states, while in unfolding process we draw samples from unfolding events and pay more attention to the zone around extended states. Simulation trajectories of those processes are often dominated by stochastic fluctuations around the initial states. It is near the target states that one may observe interesting structural information about pathways.

Although the simulation data of biomolecular systems often lie in a high-dimensional configuration space, the degree of freedom is much less due to the constraints and cooperation among atoms in folding process. It is often expected that the pathway samples are concentrated around some low-dimensional manifolds which can be described by a relatively small number of intrinsic reaction coordinates.[8] The existence of multiple pathways as in the example of this paper may lead to holes in such manifolds with nontrivial topology. Note that in the continuous case, the Reeb graph of a (unconditional) density function defined on the Euclidean space $\mathbb{R}^n$ turns out to be trivially a tree. However, conditional density functions adopted here may restrict on interesting regions where the loops in the Reeb graph might shed light on the hole structures. Reconstructing the low-dimensional topology of densely sampled regions, thus, may disclose the nature of multiple pathways. In theory, it is possible to efficiently recover the topology from samples of such low-dimensional manifolds.[21] In this paper, through conditional density filters we approach such manifolds via data level sets and extract some low-dimensional topological features which provide structural evidence on the existence of multiple pathways.

Here we describe a general approach to construct conditional density filters, which will be specialized in Sec. III with the application to RNA hairpin folding.

- Draw random samples $S \subseteq \mathcal{X}$ from the folding events. Choose importance weights on S, $w(x) \geq 0$, with higher values on interested states.
- Define the filter function by

  $$h(x) = -\log \frac{\Sigma_{y \in S} w(y) K(x,y)}{\Sigma_{y \in S} w(y)}, \quad (1)$$

  where the kernel function is defined by

  $$K(x,y) = e^{-d^{\beta}(x,y)/\alpha}, \quad (2)$$

  where $d(x,y)$ is some distance function between configurations x and y, $\alpha > 0$ is the bandwidth, and $\beta > 0$ is the exponent. For example, the Euclidean distance with $\beta = 2$ is used in the case of Gaussian kernels or the Hamming distance between structural contact maps with $\beta = 1$ is used later in this paper.
- Resample from S according to the new distribution,

  $$p(x) = \frac{w(x)}{\Sigma_x w(x)}.$$

To avoid the normalization in large data sets, we can use the rejection method or extended,[22] e.g., a sequential Bernoulli experiments where a new configuration x is accepted with probability $q(x) = w(x)/\max\{w(x)\}$.

These configurations, together with the filter function (1), will be the inputs of MAPPER procedure shown in the last section.

Filter (1) assumes a density function in Boltzman form $f(x) = (1/Z)e^{-h(x)}$, with partition function $Z = \Sigma_x e^{-h(x)}$. Thus up to a constant filter (1) approximates the free energy near the folded state. Since only order information of $h(x)$ will be used below, it leads to the same result choosing any monotone transform on h, e.g., $\Sigma_{y \in S} w(y) K(x,y)$. Our construction is equivalent to a kernel density estimator which can be replaced by other methods.[23]

2. Level-set formation in MAPPER

To increase the robustness of MAPPER allowing more errors in density estimation, we only use the order information of filter (1) to construct level sets.

Level-set formation. Order the samples according to values of $h(x)$ and classify the samples into m consecutive overlapping groups of equal or similar size, whose filter value ranges $[a_i, b_i]$ cover the range of h.

Up to an arbitrary small perturbation, a real valued function $h: \mathcal{X} \rightarrow \mathbb{R}$ induces a linear order on samples. Therefore any monotone transform on $h(x)$, such as $c_1 \exp c_2 h(x)$, leads to the same level sets.

3. Clustering in MAPPER

The graphical representation of MAPPER depends on the choice of clustering methods. MAPPER itself does not place any prerequisite on the clustering algorithm. In the study of biomolecular folding such as RNA hairpins, our purpose is to identify those connected components in free energy or density level sets, which might be of nonconvex shapes and whose numbers are unknown to us beforehand. Single-linkage clustering is the simplest choice to meet those two features.

- On each level set, construct a weighted graph, with nodes for configurations and edge weights as pairwise distances.
- Find a minimal spanning tree (MST) of such a graph.
- Find a threshold value for edges. We construct a histogram of MST edge weights with k bins. Once some empty bins are found from top bins containing p longest edges, we set the threshold to be the center of the first empty bin. Otherwise, set the threshold the maximal edge (diameter).
- Truncate the graph by breaking those edges greater than the threshold, dividing the graph into connected components.
- Prune those components of size no more than q.

Single linkage will separate those clusters where within each cluster two points can be joined by a path consisting of short edges, but relatively longer edges are required to merge the clusters. When we draw random samples from compact connected components in an Euclidean space, the distances between configurations within the same components will drop down to zero as the sample size grows. Hence the distances across components will be kept in the longest edges and can be separated from a large amount of short edges. Thresholding above tries to capture such a gap. Truncation may create several components/clusters of different sizes, where pruning helps reduce the noise and identify those dominant components.

In the continuous setting, single-linkage clustering will consistently locate those connected components when the samples are dense enough.[18] Such a feature makes it a desirable choice for MAPPER,[11] as well as density cluster trees.[19,20] However, in the latter part of this paper, we will meet a discrete configuration space, i.e., the space of contact maps as undirected graphs. Thus we need to explain in what sense we extend the "connected components" in such a discrete setting.

Equipped with a metric, e.g., Hamming distance, the discrete configuration set can be viewed as a weighted complete graph, where each node represents a structure and the weight of an edge is the distance between its end points. Single-linkage clustering first builds up a MST of this graph and then truncates the MST by keeping the edges with the length less than a given threshold, which breaks the MST into several connected components or clusters. In this way, single linkage computes the components where two nodes within a component are joined by a path consisting of the short edges, but relatively longer edges are required to merge different components.

One may also consider other clustering schemes, such as k-means, which is widely used in clustering the configura-

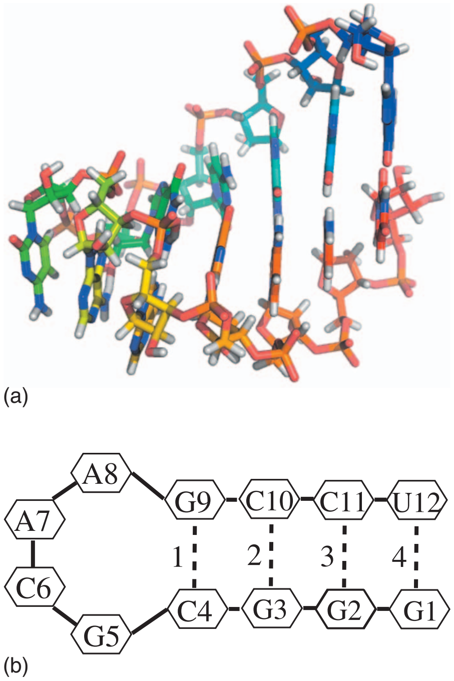

FIG. 2. (Color) (a) NMR structure of the GCAA tetraloop. (b) Contact map for the native state. Bases are numbered from 1 to 12 and native basepair contacts (dotted lines) are numbered 1–4.

tions in the biomolecular folding simulations. In contrast to single linkage, k-means attempts to find the clusters such that within cluster *any* two nodes are connected by a short edge, rather than by a path made up of short edges. Therefore, roughly speaking, k-means attempts to find *spherical-shape* clusters while single linkage can discover *snake-shape* clusters. Both may provide useful but different kinds of information in biomolecular folding problems. However, k-means needs one to specify the number of clusters *a priori* while the single linkage does not. This is a shortcoming for k-means since we do not have such information in advance. So in this paper, single linkage is chosen as the basic scheme and k-means is only used in comparative studies, when we already know the number of clusters from single linkage. Other choice of clustering methods includes average linkage, complete linkage, spectral clustering,[24,25] etc., which are, however, not pursued in this paper.

III. RESULTS AND DISCUSSIONS

Recently, Ref. 14 performed serial replica exchange molecular dynamics[26,27] (SREMD) simulations of the GCAA tetraloop (5′-GGGCGCAAGCCU-3′) on the Folding@home distributed computing platform. The hairpin motif consists of a primarily Watson–Crick base-paired stem capped with a loop of unpaired or non-Watson–Crick base-paired nucleotides, as shown in Fig. 2(a). Despite their simple structures, there is some debate over whether or not there are intermediate states in the folding of hairpins, e.g., see Ref. 13.

With the technique developed in this paper, we are able to disclose the structures of multiple intermediate states on the folding pathways, which in the first time provides structural evidence from computer simulations about RNA hairpin folding pathways. The biological implications of this discovery are discussed in detail by Bowman *et al.*[14] Here, we only focus on details of data analysis.

The RNA molecule examined here has 389 atoms. Including the solvent there are about $N=12\,000$ atoms in the system, yielding $3N=36\,000$ parameters. To reduce the dimensionality of this large space, we chose to represent each configuration with a contact map. Contact maps can faithfully describe the base-pair interactions in the stem, which provides important structural information of RNA hairpin folding. A contact map is a bit string specifying pairs of contacting residues that are not immediately adjacent in the sequence. Following Bowman *et al.*,[14] we define the *native state* as any conformation with all four stem base-pair contacts formed. Each of these base-pair contacts is referred to as a native contact. For example, Fig. 2(a) shows a native state whose contact map model is illustrated in Fig. 2(b). An *unfolding* event is defined as the set of conformations between the first point with no contacts between any two residues on opposite sides of the stem and the first preceding point with four native contacts. A *refolding* event is defined as the set of conformations between the first point with no contacts between any two residues on opposite sides of the stem and the first subsequent point where the number of native contacts is 4.

A. Structural analysis by MAPPER

MAPPER is an ideal tool for such a problem due to the enormous size of the simulation data set, the high probability of nonconvex states, and the need to identify folding intermediates with low populations relative to the folded and unfolded states. Application of MAPPER to this data set revealed a number of intermediate states.

The data generated from SREMD simulations are normally dominated by the folded and unfolded structures. For example, a typical refolding trajectory starts from an unfolded state, undergoing a significant period of stochastic fluctuation around that, then proceeds gradually to the folded state. It is in the neighborhood of folded states that interesting structural information about folding pathways are exhibited. Therefore, in the construction of the conditional density filters, we treat folding and unfolding separately. In the study of folding pathways, we take configurations from refolding events, and then weight heavily a neighborhood around the native states. However, in the study of unfolding pathways, we sample from unfolding events, and focus on a neighborhood of the unfolded states.

The following parameters are used to produce the results in Fig. 3. We use the Hamming distance $d_H(x,y)$ between a pair of contact maps in the conditional density function [Eq. (1)] and choose $\alpha=\beta=1$ in kernel (2). For simplicity, the important weights are set to one within the neighborhood of the state of interest and zero otherwise. In refolding events, we choose a neighborhood within seven-bit Hamming distance from the native state in Fig. 2. In unfolding events, a neighborhood of the extended state is chosen as the set of configurations with no more than six nonadjacent contacts formed. In the level-set formation, the filter is divided into

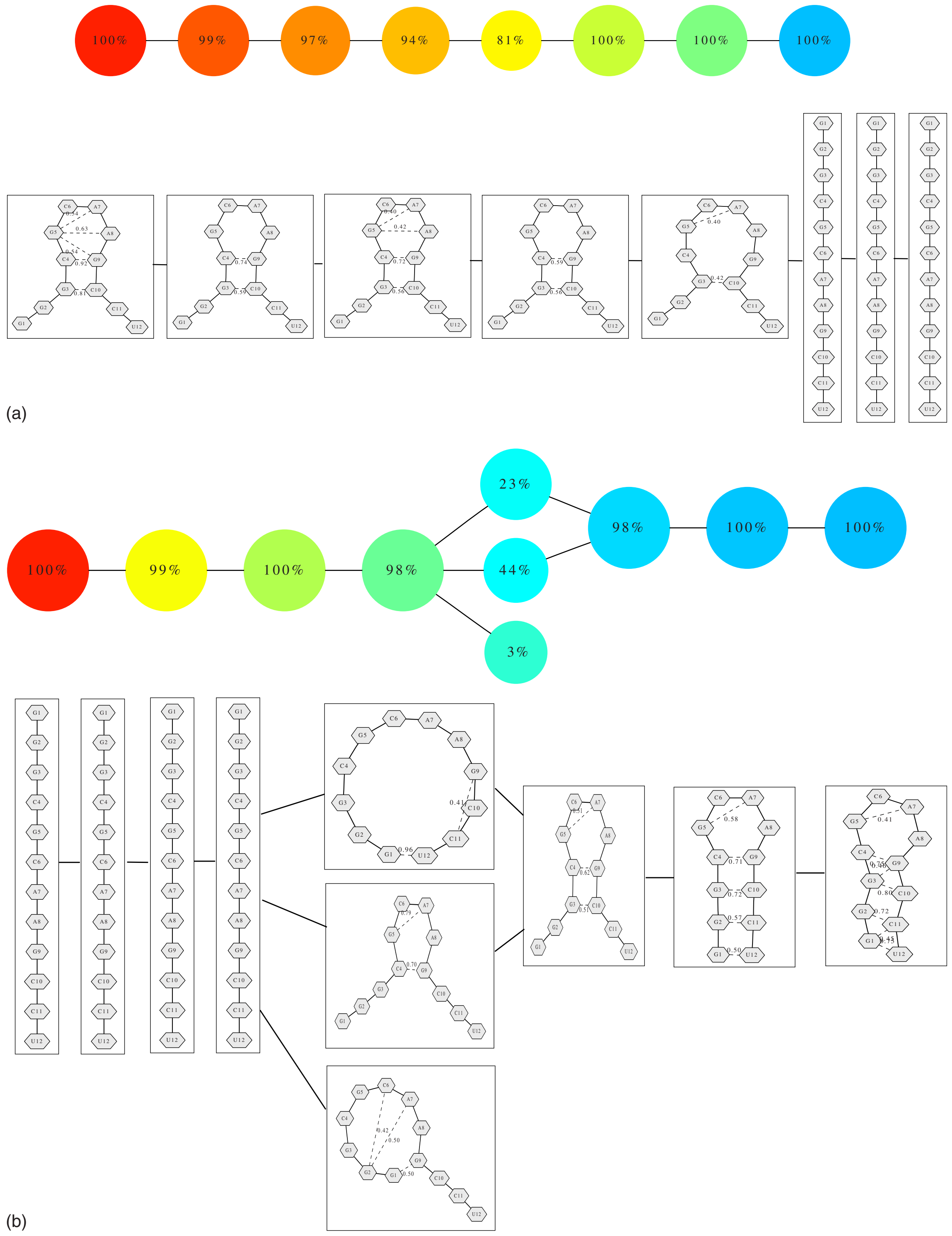

FIG. 3. (Color online) Graphical representation of pathways by MAPPER. (a) Unfolding pathway. (b) Folding pathway. In both cases, the top row graphs are the outputs from MAPPER, while the bottom row depicts the mean contact maps of the corresponding clusters. For clarity in mean contact maps, we drop those mean contacts lower than 0.4. The node colors from red to blue indicate the density from high to low, and the labels (e.g., 100%) show the percentage of configurations of the same level included in the cluster corresponding to the node. We dropped all the clusters of size smaller than 3% of the level size. (a) shows that unfolding has a single dominant pathway characterized by unzipping from the end base pair. (b) shows that folding process has two dominant pathways, passing through either the formation of the closing base pair or the end base pair. A noisy cluster consisting 3% of the level size was also shown in (b), which accounts for reptation, i.e., sliding of the two strands of the stem.

eight levels of equal size with 25% overlap. In the clustering, a histogram with five bins is used, with thresholding from top bins consisting largest $p=20\%$ edges and the cluster pruning size $q=2\%$ of the level sample size. More details on parameter tuning will be provided in Appendix B.

The graphical output of MAPPER with such parameters shows distinct pictures about folding and unfolding pathways. Unfolding has a single dominant pathway characterized by unzipping from the end base pair [Fig. 3(a)], while folding process has two dominant pathways, passing through

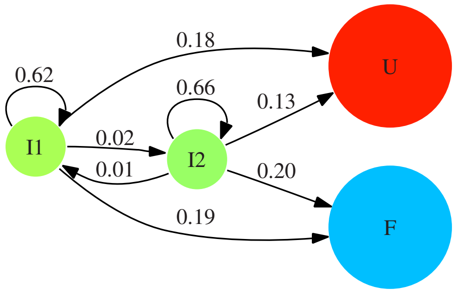

FIG. 4. (Color online) Transition probability from two intermediate states. Lag time is 2 ps. The left four nodes as extended structures [Fig. 3(b)] are merged into node U, and the right three nodes as folded structures are collected in node F. The two intermediate states on pathways are denoted by I1 and I2, respectively. The transition probability from I1 and I2 to other states are noted as numbers on the arrows. One can see that I1 and I2 are kinetically separated.

either the formation of the closing base pair or the end base pair [Fig. 3(b)]. Such an observation reveals a number of intermediate states in the folding process, which supports the multistate hypothesis. It is interesting to notice in Fig. 3 that conditional density filters seem good indicators of reaction coordinates, suggesting that the folding/unfolding processes start from the densest zone and become sparser as the reactions proceed.

B. Verifying the kinetic separation of the two pathways

Are the two pathways in refolding [Fig. 3(b)] are truly separate pathways or just the artifact of noise? This question can be answered from the kinetic information of simulation trajectories by computing the transition probability. Note that our purpose here is not to create a Markov model[1] for metastable states, but investigate how the two intermediate states in refolding pathways are kinetically connected. Therefore the shortest lag time, 2 ps, is chosen which provides the finest resolution in simulation trajectories.

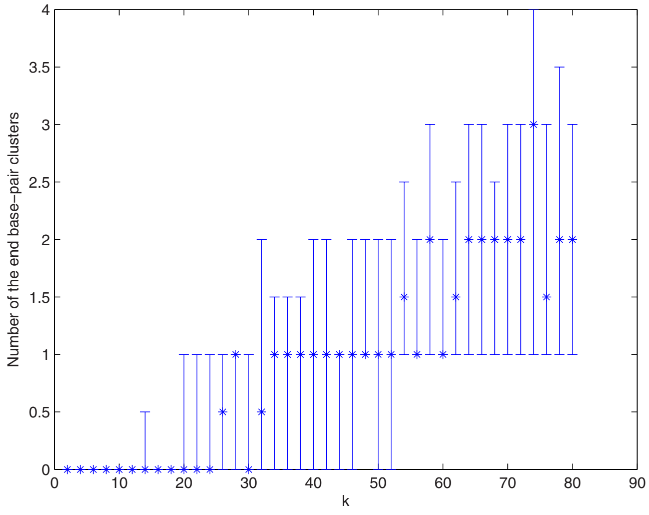

FIG. 5. (Color online) The number of end base-pair clusters found by K-means. Here k ranges from 2 to 80 with step 2. For each k, 20 experiments are repeated with K-means clustering. The number of clusters with end base pair formed are recorded. The star is the median of such numbers and the bar delimits the distribution range from 10% to 90%. Starting from around $k=25$, such clusters appear with at least 1/2 probability. Around $k=55$, such clusters begin to split. The instability of K-means clusters is increasing as k grows, indicated by the expanding ranges.

To simplify the result, we merge the four nodes with extended structures as a single unfolded state U and collapse the three blue nodes with folded structures as folded state F leaving alone the two intermediates, I1 and I2. This does not change the topology of MAPPER graph, but highlights the dynamics associated with intermediate states. Configurations in simulations are mapped to such four-node states by nearest neighbor method. One-step (2 ps) transition probability is then computed among the four states.

The result is shown in Fig. 4. It can be seen that the two intermediates, I1 and I2, are kinetically well separated on folding pathways. Once the simulation climbs up the energy barrier I1 and I2, the majority will either proceed to F or withdraw to U, while an ignorable minority will cross the intermediates from I1 to I2. Moreover, we note that since our SREMD simulations perform a random walk in the temperature space, we are not able to extract rates of the folding or unfolding reactions at a certain temperature of interest.

C. Importance of conditional density filters

Conditional density filters play a crucial role here, without which clustering methods such as K-means or single linkage tend to split the sparse intermediates and lump them with densest clusters.

To see this, we make a comparison between MAPPER clusters found in Fig. 3(b) and K-means clustering on the same data set. Since the number of K-means clusters is not unknown *a priori*, we performed a series of experiments with k varying from 1 to 80, each of which has 20 repeated experiments. Our first purpose is to locate the value of k around which the MAPPER cluster with end base pair formed becomes identifiable. Hence for each K-means experiment, we count the number of the end base-pair clusters, defined as the clusters containing more than 75% configurations with native contact 4 [Fig. 2(b)] formed and less than 25% for any other native contact. Figure 5 plots a rough distribution of the numbers of end base-pair clusters against the growth of k. It can be seen that around $k=25$ this intermediate state becomes identifiable, in the sense that with more than 1/2 probability such clusters are found indicated by nonzero medians. Notice that as k grows, the variation range ($10\% \sim 90\%$) of such cluster numbers expands, showing a trend of increasing instability. Particularly around $k=55$, such a state begins to split into several K-means clusters.

We can further see how K-means clusters might split the intermediate states and lump them toward densest clusters. Figure 6 illustrates this when $k=30$ for K-means clustering, on the same data set for the construction of MAPPER clusters on refolding pathways.

D. Comparative studies on single linkage versus *k*-means

Single-linkage clustering is motivated by its ability to identify possibly nonconvex clusters of unknown number. It is also interesting to explore other clustering methods such as

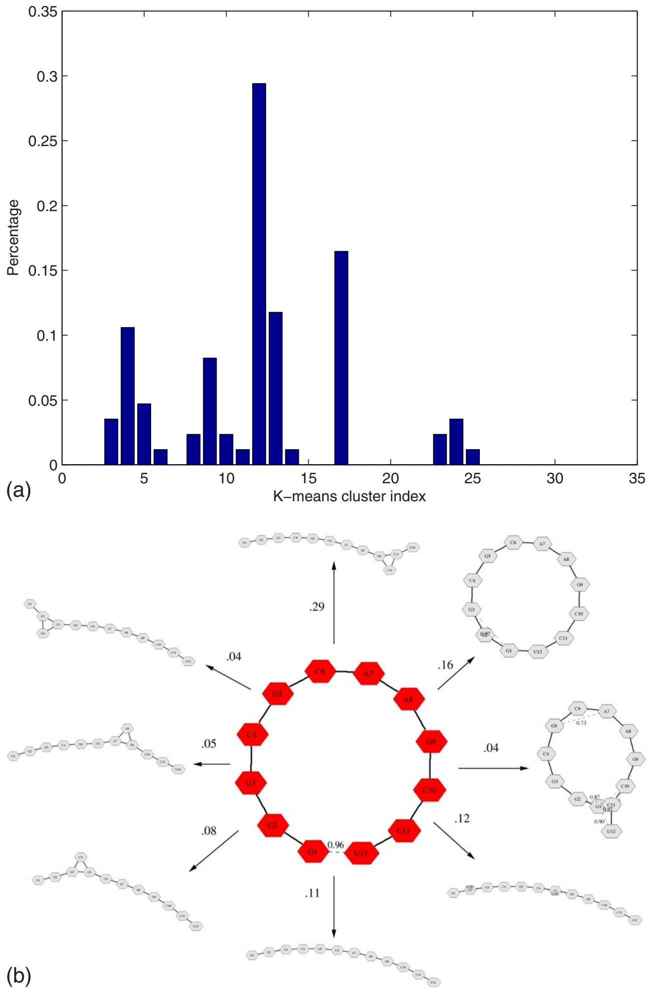

FIG. 6. (Color online) K-means clustering fails to capture the low-density intermediate states with one end base pair formed. The illustration here chooses $k=30$ for K-means clustering. (a) shows how end base-pair formed structures are distributed in different k-means clusters; (b) illustrates the mean structures of the top eight K-means clusters (gray) which contain base-pair formed structures. K-means splits the MAPPER cluster and lumps them with densest clusters.

K-means which tries to group data in *spherical* clusters and is widely used in the studies of biomolecular folding simulations. Given the cluster number returned by single linkage, comparisons with K-means of similar number of clusters on the same level sets might disclose how far the intermediate states deviate from spherical shapes. For this purpose, we perform K-means clustering on the same data set for refolding pathways in Fig. 3(b). We use the same number of clusters returned by single linkage, and especially on level 5 we set $k=2$. It turns out that K-means finds two clusters on level 5 with similar structural features to single linkage, i.e., one with closing base pair formed and the other with the end base pair. However K-means has different partition: 48% versus 52%, in contrast to 23% versus 44% in single linkage. Clearly to form spherical clusters, K-means clusters mix more configurations from different single-linkage clusters, which can be shown by the percentage dropping of dominant end base pair from 96% to 65% in the smaller cluster. However, the structural similarity in both methods suggests that single-linkage clusters are not very far from spherical shapes.

E. On nonlinear dimensionality reduction

Although a biomolecular system is typically described by a high-dimensional configuration space, it is expected that those configurations often visited in a folding process may concentrate around some low-dimensional manifolds which might be described by a much smaller number of reaction coordinates. Recently Das *et al.*[8] shows that Isomap can be applied to recover such reaction coordinates in simple folding processes with a single pathway. Isomap tries to preserve both the local and global geodesic distances between configurations defined as shortest path distance on a neighborhood graph. However, Isomap might not work in complex problems where multiple pathways exist. Isomap requires that the data manifold are globally isometric to a convex domain of low-dimensional space.[2,5] The existence of more than two pathways connecting two metastates may lead to holes in sampled regions which fails the convex domain assumption. Moreover, Isomap is too sensitive to the metric in choice. In this paper we use a coarse metric as Hamming distance for contact maps, where the geodesic distance between configurations does not reflect the distance in folding process. Moreover, The high heterogeneity in distribution is also a hurdle for Isomap technique to identify useful intermediates.

The last two issues also challenge other techniques for nonlinear dimensionality reduction, such as Locally Linear Embedding,[3] Laplacian eigenmap,[4] Hessian eigenmap,[5] diffusion map,[6] etc. These geometric embedding techniques map the data in high-dimensional spaces to a low-dimensional space by preserving some local metric relations among neighbors of data points, e.g., see Ref. 7. They are thus sensitive to the metric in choice and heterogeneous distribution might distort local metrics. In applications to complex biomolecular systems, successful examples are only found in simple settings such as with a single protein folding pathway[8] or quasisteady state in dynamics of signal transduction networks.[9]

However, as a topological tool MAPPER with density filters is shown efficient in dealing with heterogeneous distributions and less sensitive to the metric in choice. In this paper even with such a coarse metric as Hamming distance, it efficiently discloses structural information in pathways which are difficult to other geometric embedding techniques. Thus, one of our ongoing directions is to combine the topological tool MAPPER with those geometric embedding techniques, such as applying nonlinear dimensionality reduction separately on components or clusters discovered by MAPPER.

IV. CONCLUSIONS

In this paper we develop MAPPER, a topological data analysis tool, in the analysis of simulation data for biomolecular folding pathways. As an application, in the first time we are able to obtain structural evidence from computer simulations in support that RNA hairpin folding has two dominant pathways with multiple intermediate states. We have also incorporated the temporal information from simulation trajectories to verify that the twofolding pathways are kinetically separated. It is thus a promising direction to explore with MAPPER such structural information in biomolecular folding problems.

We have shown that with proper designs of conditional density filters and clustering schemes, MAPPER can address the heterogeneity issue in distribution, deal with multiple pathway data with nontrivial topology, and be less sensitive to the metric in choice. These features can be used to enhance traditional nonlinear dimensionality reduction methods, such as Isomap, Laplacian eigenmap, diffusion maps, etc. One of our ongoing direction is to explore the combinations of the topological tool MAPPER with those geometric tools for better characterizations of biomolecular systems.

As clustering method plays a fundamental role toward building up many important models such as Markov state models. MAPPER as a methodology adds a new perspective to existing clustering tools. One of our future direction is to build up more sophisticated dynamical models based on MAPPER which incorporate intermediate states and can be reduced to traditional Markov state models describing merely basin-to-basin transitions.

ACKNOWLEDGMENTS

Y.Y. would like to thank Wing-Hung Wong, Nancy Zhang, and Qing Zhou for helpful discussions, X.H. thanks Michael Levitt for his support. Y.Y., J.S., G.S., L.G., and G.C. are funded by DARPA Grant No. HR0011-05-1-0007; X.H. by NIH Roadmap for Medical Research Grant No. U54 GM072970; G.B. by the NSF Graduate Research Fellowship Program; M.L. by a NDSEG fellowship. Y.Y. and G.C. are also supported by NSF Grant No. DMS 0354543, and LG by NSF Grant No. FRG-0354543 and NIH Grant Mo. GM-072970. This work was also supported by NIH Grant No. P01 GM066275. Computing resources were provided by the Folding@home users and NSF Award No. CNS-0619926.

APPENDIX A: RNA HAIRPIN FOLDING SIMULATIONS

Our simulations used the AMBER 94 potential. 2800 SREMD simulations with an aggregate simulation time of 54.6 μs were performed, see Ref. 14 for details. Even with this amount of simulation, reversible folding was not achieved and we have not reached equilibrium sampling.[14] However, among 2800 SREMD simulations, we obtain 760 trajectories with a complete unfolding event and 550 trajectories with a complete refolding event. Therefore, we have sufficient data to define the dominant conformational states in the folding and unfolding pathways. Note that an unfolding event defined above only contains one unfolded state as the end point, whose density is thus too low in samples. Therefore, among such trajectories, we randomly choose 149 extended unfolding events and 23 extended refolding events, which includes $m=10$ more points after the end point of each event. In this way we obtain about 100 000 samples for either class of events.

Note that contact maps are used as a discrete representation of structures, whence different configuration samples might have the same contact map representation. Such repetitions should be kept for density estimation, but can be compressed into unique structures for clustering analysis. In fact, those samples contain 49 332 and 56 118 unique contact maps, for unfolding and refolding extended events, respectively. They are sufficient for the analysis by MAPPER.

APPENDIX B: PARAMETER CHOICE IN MAPPER

Conditional density filter. The data generated from SREMD simulations are normally dominated by the folded and unfolded structures. For example, a typical refolding trajectory starts from an unfolded state, undergoing a significant period of stochastic fluctuation around that, then proceeds gradually to the folded state. It is in the neighborhood of folded states that interesting structural information about folding pathways are exhibited. Therefore, in the construction of the conditional density filters, we treat folding and unfolding separately. In the study of folding pathways, we take configurations from refolding events, and then weight heavily a neighborhood around the native states. However, in the study of unfolding pathways, we sample from unfolding events and focus on a neighborhood of the unfolded states.

To be specific, in the study of unfolding, we extract 4330 configurations around the extended states with no more six nonadjacent contacts formed. On the other hand, in the study of refolding, we extract 2952 configurations of no more that seven-bit Hamming distance away from the native state to avoid the highly populated extended states. This is equivalent to the choice of a weight function $w(x)$ which is a constant in a neighborhood of the extended states (no more than six nonadjacent contacts) or the native state (no more than seven-bit Hamming distance) and zero otherwise.

Since the space of contact maps is discrete, we use Hamming distance and choose $\alpha=\beta=1$ in kernel density estimation (2), which is equivalent to the Gaussian kernel with the standard Euclidean distance in $\mathbb{R}^{55}$.

We note that in a range of $1\leq\alpha\leq 8$ MAPPER returns qualitatively similar results. In fact, smoothing the density filter without changing the order leads to the same result in MAPPER. However, decreasing α, even to 0.9, causes the disappearance of the smaller passway. A small choice of α creates a rugged density filter, which alters the results of MAPPER. Our experiments show that $\alpha=1$ is close to this bifurcation point.

Level sets. We divide the range of the density filters into n overlapped intervals, where each interval contains the same number of samples. In other words, we order the samples according to the filter value, then divide the sample into overlapped bins of equal size. It is also possible to consider division by equal filter value intervals,[11] but the former has at least two advantages. First the former method only takes into account the order information about the filters to stratify the data, whence any monotone transformation on $h(x)$, such as $-\log h(x)$, leads to the same result. This makes the result from MAPPER relatively more robust to the error in density estimation. Second it is more convenient to control the computational cost where each level has similar running time due to the same sample size, which is suitable for parallel computations.

We have tested the choice of n among 4, 6, 8, 10, 12, 14, so that each level contains around several hundred configurations. Overlap percentage can be chosen from 15% to 75%. All of them give qualitatively similar results although smaller number of levels and larger overlap cause longer computation. The results presented in this paper are generated under the choice of eight intervals with 25% overlap.

1. Single-linkage clustering

To determine an appropriate threshold in single linkage, we build up a histogram based on the edge weights in the MST using $k=5$ bins. We only focus on those bins containing the largest $p=20\%$ edges. The threshold value is chosen to be the center of the first empty bin among them, defined as less than $q=2\%$ samples in the largest bin. The reason we do so is that the empty bin with short edges often appear due to undersampling which does not tell us information about gaps among components. If there is no such short bin, take the entire level set as one cluster. The results of MAPPER will be sensitive to the choice of such k. Generally speaking, increasing k will increase the number of clusters and vice versa. Considering the fact that the diameter of the data is about 14, we normally choose k an integer between 5 and 10, which all gave qualitatively the same results as $k=5$. We note that although the threshold found by histogram method is sensitive to the bin number k, the threshold leading to the two clusters on level five in Fig. 3(b) is always two-bit Hamming distance, which is very robust in different choices of k.

[1] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).

[2] J. Tenenbaum, V. de Silva, and J. Langford, Science **290**, 2319 (2000).

[3] S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).

[4] M. Belkin and P. Niyogi, Neural Comput. **15**, 1373 (2003).

[5] D. Donoho and C. Grimes, Proc. Natl. Acad. Sci. U.S.A. **100**, 5591 (2003).

[6] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U.S.A. **102**, 7426 (2005).

[7] P. W. Jones, M. Maggioni, and R. Schul, Proc. Natl. Acad. Sci. U.S.A. **105**, 1803 (2008).

[8] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **103**, 9885 (2006).

[9] P. Barbano, M. Spivak, M. Flajolet, A. C. Nairn, P. Greengard, and L. Greengard, Proc. Natl. Acad. Sci. U.S.A. **104**, 19169 (2007).

[10] J. Milnor, *Morse Theory* (Princeton University Press, Princeton, NJ, 1963).

[11] G. Singh, F. Mémoli, and G. Carlsson, Eurographics Symposium on Point-Based Graphics, 2007 (unpublished).

[12] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).

[13] H. Ma, C. Wan, A. Wu, and A. H. Zewail, Proc. Natl. Acad. Sci. U.S.A. **104**, 712 (2007).

[14] G. R. Bowman, X. Huang, Y. Yao, J. Sun, G. Carlsson, L. J. Guibas, and V. S. Pande, J. Am. Chem. Soc. **103**, 9676 (2008).

[15] S. Smale, Ann. Math. **74**, 391 (1961).

[16] G. Reeb, Comptes Rendus Acad. Science Paris **222**, 847 (1946).

[17] M. van Kreveld, R. van Oostrum, C. Bajaj, V. Pascucci, and D. Schikore, Proceedings of the 13th Annual ACM Symposium on Computational Geometry, 1997 (unpublished), pp. 212–220.

[18] J. A. Hartigan, J. Am. Stat. Assoc. **76**, 388 (1981).

[19] W. Stuetzle, J. Classif., **20**, 25 (2003).

[20] Q. Zhou and W.-H. Wong, The Annals of Applied Statistics **2**, 1307 (2008).

[21] P. Niyogi, S. Smale, and S. Weinberger, Discrete and Computational Geometry **39**, 419 (2008).

[22] J. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2004).

[23] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London/CRC, Boca Raton, FL, 1986).

[24] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer-Verlag, Berlin, 2001).

[25] C. Ding and H. Zha, *Spectral Clustering, Ordering and Ranking–Statistical Learning with Matrix Factorizations* (Springer, New York, 2007).

[26] X. Huang, G. R. Bowman, and V. S. Pande, J. Chem. Phys. **128**, 205106 (2008).

[27] M. Menger, F. Eckstein, and D. Porschke, Biochemistry **39**, 4500 (2000).