PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides

Hernán Stamati,[1] Cecilia Clementi,[2,3]* and Lydia E. Kavraki[1,3,4]*

[1] Department of Computer Science, Rice University, Houston, Texas 77005
[2] Department of Chemistry, Rice University, Houston, Texas 77005
[3] Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas 77030
[4] Department of Bioengineering, Rice University, Houston, Texas 77005

ABSTRACT

The automatic classification of the wealth of molecular configurations gathered in simulation in the form of a few coordinates that help to explain the main states and transitions of the system is a recurring problem in computational molecular biophysics. We use the recently proposed ScIMAP algorithm to automatically extract motion parameters from simulation data. The procedure uses only molecular shape similarity and topology information inferred directly from the simulated conformations, and is not biased by *a priori* known information. The automatically recovered coordinates prove as excellent reaction coordinates for the molecules studied and can be used to identify stable states and transitions, and as a basis to build free-energy surfaces. The coordinates provide a better description of the free energy landscape when compared with coordinates computed using principal components analysis, the most popular linear dimensionality reduction technique. The method is first validated on the analysis of the dynamics of an all-atom model of alanine dipeptide, where it successfully recover all previously known metastable states. When applied to characterize the simulated folding of a coarse-grained model of β-hairpin, in addition to the folded and unfolded states, two symmetric misfolding crossings of the hairpin strands are observed, together with the most likely transitions from one to the other.

Proteins 2010; 78:223–235.

Key words: nonlinear dimensionality reduction; reaction coordinates; free-energy landscapes; molecular dynamics simulations. distance measures; Isomap; ScIMAP.

INTRODUCTION

Biological processes at the molecular level usually involve motion and structural changes in biomolecules, such as proteins and peptides. Most *in silico* studies start by gathering large sets of conformational data through some form of simulation, for example molecular dynamics (MD).[1–3] When provided with a physical model of the molecule(s) to simulate, these techniques produce as output abundant conformational samples in the form of Cartesian (x,y,z) coordinates for each of the molecule's atoms. As the computational cost of MD simulations increases rapidly with atom count, significant research effort has been devoted to improve the time scales sampled for large molecular systems while retaining the most interesting simulation details, at the expense of some accuracy in the modeling. Toward this goal, several coarse-grained or multiresolution molecular models have been proposed (e.g., Refs. 4–29) that approximate the dynamics of a system by considering a reduced set of effective degrees of freedom, at least in part of the system, or part of the simulation time. Other methods to speed up simulations include taking adaptive and/or larger simulation steps,[30–32] using different forms of "accelerated" molecular dynamics,[33–36] replica exchange/parallel tempering,[37–40] implicit solvent models,[41] and large-scale distributed computing.[42]

Once abundant molecular samples of the studied process have been gathered through simulations, the data are subjected to analysis. Input data given as Cartesian coordinates is always high-dimensional, since there are three coordinates (or parameters) per atom. A central question is whether the configurational space spanned in the simulation can be described by using a few essential degrees of freedom reproducing the collective motion of the system. To address this question,

Grant sponsor: National Science Foundation; Grant numbers: CHE-0349303, CCF-0523908; Grant sponsor: National Institute of Health; Grant number: GM078988; Grant sponsor: Welch Foundation (Norman Hackermann Young Investigator award); Grant number: C-1570; Grant sponsor: Pharmacoinformatics Training Program of the Keck Center of the Gulf Coast Consortia; Grant sponsor: NIH; Grant numbers: 1 T90 DK070109-01, 1 R90 DK071505-01; Grant sponsor: CCF; Grant number: 0523908; Grant sponsor: NSF; Grant number: CNS-0421109; Grant sponsor: NSF; Grant number: CNS-0454333; Grant sponsors: Sloan Foundation, Rice University Funds.

*Correspondence to: Cecilia Clementi, Department of Chemistry, Rice University, 6100 Main Street, MS-60, Houston, TX 77005. E-mail: cecilia@rice.edu or Lydia E. Kavraki, Department of Computer Science, Rice University, 6100 Main Street, MS-132, Houston, TX 77005. E-mail: kavraki@rice.edu

Received 18 August 2008; Revised 29 May 2009; Accepted 24 June 2009
Published online 7 July 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22526

it is crucial to find ways to reduce the dimensionality of the simulated conformations to render the process understandable, without the need to visualize molecular trajectories directly.

For these purposes, so-called reaction coordinates have been devised over the years to succinctly describe molecular conformations, so they can be classified along a few, meaningful axes. Such coordinates are oftentimes used to construct free-energy landscapes and quantify the thermodynamics of the molecular process under consideration. Usually, reaction coordinates are either chosen from a pool of previously proposed ones, or empirically designed to suit a particular molecular system (e.g., Refs. 43, 44). Recently, methods have been proposed to automate the selection of reaction coordinates for a molecular process. In the last few years several groups have worked on the definition of the theoretical framework and computational approaches to extract a minimal number of reaction coordinates in high-dimensional systems.[45] Although it does not provide a complete view of all the relevant contributions emerging in this very active field, it is worth mentioning as significant examples the Transition Path Theory[46,47] and Transition Path Sampling,[48–51] the Markovian State Model,[52–54] Milestoning,[55–57] the Nudged Elastic Band Method,[58,59] and the String Method.[60–62] Other recent, relevant examples include an automated method proposed[63] for identifying an "optimal" set of reaction coordinates by using genetic neural networks to mine a database of known reaction coordinates and physical variables, and a combinatorial pattern discovery approach[64] that first turns each simulated conformation into a seven-dimensional vector of known reaction coordinates, then applies clustering to these vectors.

In this context, we have recently proposed the ScIMAP method[65] to automatically extract the essential parameters spanning the configurational landscape associated with a molecular motion, by using only the simulated conformations, without any bias by *a priori* information. We call such parameters structural reaction coordinates hereafter. The idea is to use dimensionality reduction methods[66,67] to describe the motion landscape with few, but meaningful, automatically recovered structural reaction coordinates. The purpose of dimensionality reduction is to extract the main features from a set of points, which are initially represented by a large set of redundant parameters. Most dimensionality reduction techniques produce as a result a lower-dimensional representation for each point that summarizes the variability of the (high-dimensional) original representation of the points. Ideally, one would like a low-dimensional Euclidean representation of the points that would serve as a "projection" or "map" of the input data. Such a low-dimensional map is easy to visualize and the process of interest can then be succinctly described by looking at this projection.

Dimensionality reduction is used in a plethora of fields, including classification problems, data mining, image analysis and recognition,[68–72] structural and computational biology,[73–75] economics, and language interpretation and analysis.[76] Molecular simulation data presents interesting challenges for dimensionality reduction. First of all, these data are highly nonlinear in nature. Also, data from molecular simulations tend to cluster around energy minima, producing uneven sampling that varies greatly in density throughout the input space.

Several mathematical tools are available to perform dimensionality reduction automatically, and the output of these tools can be interpreted in the particular application domain. Linear methods, such as principal components analysis (PCA),[77] find a projection of the input data into the hyperplane best preserving the data variability. PCA has been largely used on molecular data, for example to analyze protein flexibility around equilibrium[74,78] and to capture essential dynamics.[75] However, linear methods such as PCA fail when the data distribution is highly nonlinear (as it is usually the case in large scale molecular motions). Nonlinear methods aim to recover an intrinsic parameterization for a data set that lies on a nonlinear (yet low-dimensional) surface, which is the case for most interesting molecular processes. Several nonlinear dimensionality reduction methods exist. Parametric methods augment linear methods with the notion of a kernel function and force the data to lie on this surface, for example as in kernel PCA.[79] Nonparametric methods, on the contrary, try to infer the nonlinearity of the data from the data itself. The most popular methods include Isomap[80] and locally linear embedding (LLE).[81] Here, we use the recently introduced ScIMAP method,[65] based on the Isomap algorithm. This procedure relies on the input data itself to infer its inherent topology using only a notion of similarity between the input points. Details are given in the next section.

In this work, we apply the ScIMAP method to compute structural reaction coordinates for two systems where a direct relationship between the computed parameters and structural properties of the molecules is easy to find. The results demonstrate the capabilities of the method and its value as a versatile and general analysis tool. The first system considered, an all-atom model of alanine dipeptide, has been studied thoroughly by simulation[82–84]; it is well known that with the force field used here (Amber99 with implicit water) only two parameters (namely, the dihedral angles ϕ, and ψ) are sufficient to accurately span the configurational space of the molecule at standard conditions. The application of ScIMAP correctly recovers these coordinates and identifies all the known states of the molecule. The second system studied, a coarse-grained β-hairpin, is used to show how the topological approach in the ScIMAP method identifies shapes and transitions that are not easy to distinguish with traditional reaction coordinates. For both cases, the automatically recovered parameters are computed using the same, unaltered ScIMAP method and molecular shape similarity as a basic operation, and serve as excellent structural reaction coordinates to characterize the molecular process.

MATERIALS AND METHODS

In this section, we describe the nonlinear dimensionality reduction technique we use to analyze MD trajectories. It is first presented in its pure mathematical form, and then adapted to work with molecular conformations. We also introduce the two molecular models and data sets to which we apply the method to automatically extract structural reaction coordinates.

The Isomap algorithm

The Isomap algorithm[80] is a nonparametric, nonlinear dimensionality reduction technique. It takes as input a set S of abstract "points," which are assumed to lie on a low-dimensional, nonlinear surface (or manifold), and a similarity measure between them, $d : S \times S \rightarrow \mathbb{R}$, so that $d(x_i, x_j)$ is the distance between points x_i and x_j. The Isomap algorithm requires as input the number of reduced dimensions to be considered and returns in output the error expected when the requested reduced dimensionality is used instead of the whole space. An "optimal" effective dimensionality for a given data set can therefore be estimated by considering the minimum number of dimensions providing a satisfactorily small error (see Refs. 65,80 for details).

Using the provided similarity measure, Isomap infers the inherent topology (or connectivity) of the manifold where the points reside to lie by connecting each point to its nearest neighbors according to the distance $d(x_i, x_j)$. For each input point, it then computes only a few coordinates such that the Euclidean distance between the points' low-dimensional coordinates best preserve the geodesic distance between all pairs of points. The geodesic distance between a pair of points is defined as the length of the shortest path between the points, when the path is confined to the surface where the points lie. By preserving all geodesic distances (rather than direct distances) Isomap "unrolls" the low-dimensional manifold into its intrinsic parameterization, as shown in Figure 1. A detailed implementation of the Isomap algorithm is provided in the original article.[80] The key point of the algorithm is the approximation of the geodesic distance as the shortest path on the nearest neighbors network on the data points. In practice, this is achieved through the following three steps:

1. Build a neighborhood graph, G: For each point, find the set of points that are nearest neighbors on the manifold using the distance measure $d(x_i, x_j)$. The typical approach is to select the k nearest neighbors to every point. Alternatively, a distance cutoff, ϵ, can be introduced, and all pairs of points closer than ϵ are considered nearest neighbors.
2. Compute the geodesics: The geodesics are approximated as the shortest paths on G for all pairs of points. Construct a matrix D where D_{ij} is the shortest path between x_i and x_j.
3. Compute the low-dimensional embedding: Use multidimensional scaling (MDS)[85] on the matrix D of estimated geodesic distances computed in step (2). This produces coordinates for each point that best preserve the geodesic distances. These coordinates, when plotted as Euclidean coordinates, have the effect of "unrolling" the nonlinear surface (see Fig. 1).

The advantages of the Isomap method over linear dimensionality reduction techniques stem from the fact that it deduces the topology of the input data by connecting nearby points, thus it "follows" the nonlinear process by computing coordinates that preserve global information based on the local similarity measure. The main disadvantage of Isomap is the computational cost of computing the neighborhood graph G, which is in general $O(n^2)$, and dependent on the cost of the distance measure $d(x_i, x_j)$. To alleviate the computational cost and memory requirements of steps (2) and (3) mentioned earlier, another version of Isomap, called Landmark Isomap,[86] was devised. It relies on the fact that if the data is truly low-dimensional, then it should suffice to preserve only a subset of the geodesic distances. In other words, instead of preserving all possible pairs of geodesic distances, it preserves only the geodesic distance from each point to a subset of landmark points, chosen among the original data set. Generally, as many landmark points as allowed by the computer system's memory are used to avoid the risk of underestimating the number needed. To adapt Isomap to work with molecular trajectories, an appropriate distance measure $d(x_i, x_j)$ is needed, when x_i and x_j are molecular conformations given as the Cartesian coordinates of the constituent atoms. A natural measure of similarity for different conformations of the same molecule is least-root-mean-squared-deviation (lRMSD).

ScIMAP: scalable isomap method

The ScIMAP method[65] includes some improvements over Isomap, such as an efficient parallelization of all three steps discussed earlier, and a method to map redundant points into the recovered coordinate space that is much less resource-demanding than including all the data points in the analysis. These improvements are crucial when working with big data sets of molecular conformations and allow the application of nonlinear dimensionality reduction to extract effective global coordinates from extensive configurational sampling. A detailed implementation of the algorithm and its testing has been described in Ref. 65, where ScIMAP has been used to characterize the folding process of a coarse-grained model of protein SH3, and ScIMAP coordinates have been shown to correctly locate the transition-state ensemble on the resulting free energy landscape.

However, the first applications of ScIMAP have not focused on the meaning of the coordinates or the effect of different molecular models on the results, which are the goals of the present article.

The computational complexity of the ScIMAP algorithm is dominated by the cost of computing the nearest-neighbors graph for the input data set. The efficient computation of nearest-neighbors remains an active research area; it has been shown that in high-dimensional spaces the scaling of the nearest-neighbors graph calculations is bounded by $d \times N^2$,[87] where N is the number of conformations in the data set and d is the dimensionality of the system. The ScIMAP implementation used for this work was based on the open-source package OOPSMP[88] to compute the nearest-neighbors graph. OOPSMP, which was originally developed for motion planning, contains robust and efficient implementations of nearest-neighbors algorithms, and it scales quadratically with N and linearly with d (as expected).

It is worth mentioning the recently proposed application of the distance projection onto Euclidean spaces (DPES) approximation to ScIMAP,[89] which significantly speeds up the neighborhood graph computation at the expense of a small approximation in the identification of neighboring points, as shown by the application of the DPES method originally applied in the robotic motion planning domain.[90] As a result, the overall method is general and efficient for large-scale data analysis. As the focus of this work is mainly on the interpretation of the recovered coordinates, the DPES approximation will not be used.

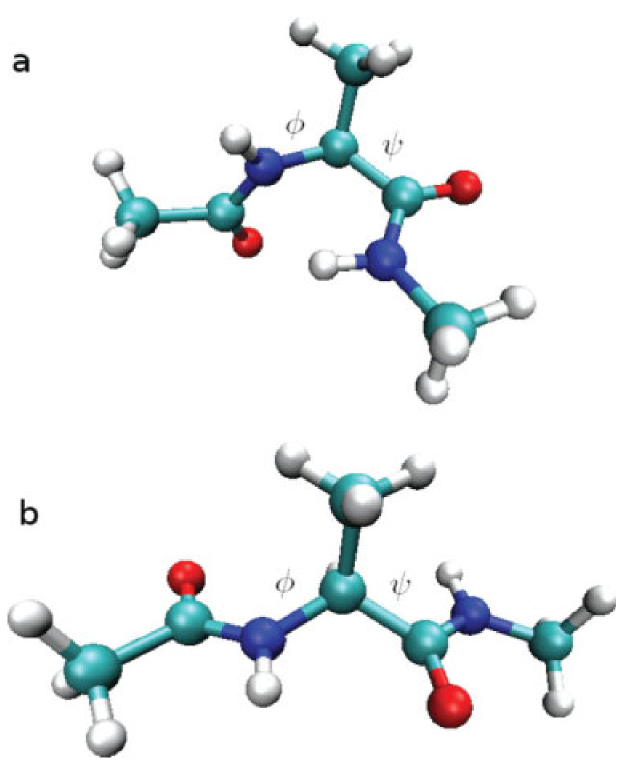

Figure 2

Alanine dipeptide most populated configurations: (a) Right-turn and (b) extended. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

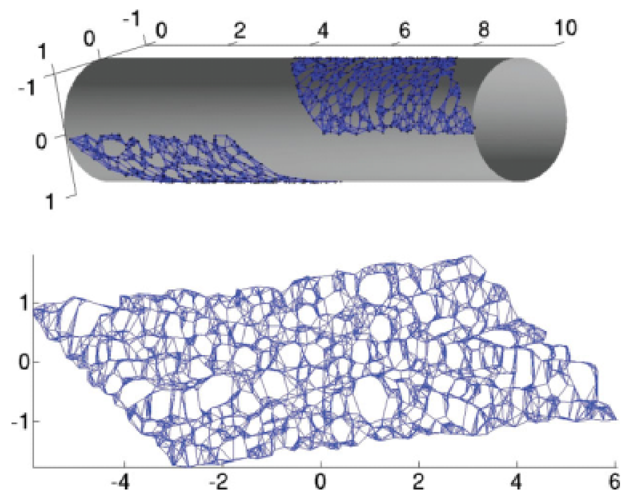

Figure 1

The Isomap algorithm. Top: An intrinsically nonlinear 2D data set, given as 3D data. The neighborhood graph is overlaid for illustration. Bottom: The resulting two-dimensional embedding coordinates for each point, as resulting from the application of the Isomap algorithm. The neighborhood graph is overlaid for comparison. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Alanine dipeptide model

We first present the application of ScIMAP to a small biomolecule that has been studied thoroughly in the past, the alanine dipeptide. We used an all-atom model consisting of 22 atoms, namely CH_3—CONH—$CHCH_3$—CONH—CH_3. This peptide is composed by a very short piece of backbone and one alanine side chain attached to it. Because the two peptide bonds present in the peptide are quite rigid, the configurational space of the molecule can be well approximated by using only the two torsions around the C_α atom, named ϕ and ψ, as shown in Figure 2.

A long MD simulation was performed using the Sander module of the AMBER 9.0 package[91] using an implicit water model to simulate the molecule in solution. The system was first randomized at 400 K, then equilibrated at room temperature (300 K). A sampling of 500,000 conformations was gathered from a molecular dynamics trajectory corresponding to 100 ns of simulated time, at 300 K.

Several computational studies[82–84,92,93] have used the backbone (ϕ, ψ) angles to explore the molecule's conformational landscape, both in vacuum and in solution. These two angles determine the overall shape of the peptide, and are sufficient to characterize its configurational landscape. It is worth mentioning that results obtained with different force fields and parameterization of the alanine dipeptide system have been reported in the literature; while simulations with different choice suggested that other degrees of freedom (besides the ϕ and ψ dihedral angles) participate to the dynamics,[†] in the solvated model used here as well as in explicit water simulation there was no significant motion of this angle other than a vibration around equilibrium.

Several different (meta)stable states of the peptide as a function of its (ϕ, ψ) angles have been characterized in previous studies. In particular, at standard conditions, there are two main conformations, clearly distinct from one other, as shown in Figure 2:

1. Extended: Also called "C_{7eq}", "C_5" or "β-like" in the literature, since it is the more extended shape [Fig. 2(b)].
2. Right-turn: Also called "α_R" or "α-like" since it resembles a tiny piece of a right-handed helix [Fig. 2(a)].

Figure 3 shows the free energy profile associated with the system, as measured on the sampled data, as a function of the (ϕ, ψ) dihedral angles. Free energy (potential of mean force) was computed using the WHAM method.[94–99] The C_5 state corresponds to the top-left corner of the free energy plot, whereas the α_R state corresponds to the other main minimum, southeast of the C_5 state. Several other less populated states of alanine dipeptide have been observed and characterized in the literature. The ones presented here are those corresponding roughly to the center of the local minima on the free energy plot:

1. Left-turn: Also called "α_L." This corresponds to the lonely minimum in the middle-right of the plot, and is extremely unlikely with the given model.
2. α_P: This is the minimum west of α_R, another helix-like conformation.
3. P_{II}: Also a less likely minimum. The physical reasons for this minimum have been studied in the literature.[84]

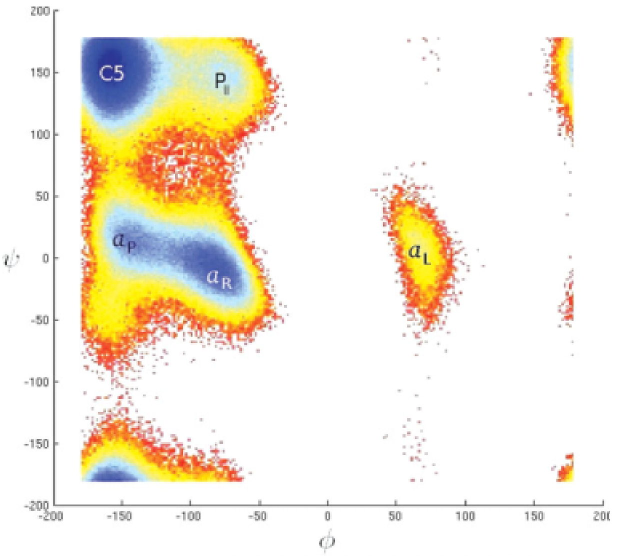

Figure 3

Alanine dipeptide free energy as a function of the (ϕ, ψ) backbone angles.

The same conformational sampling used to produce the free energy plot of Figure 3 is used for the application of the the ScIMAP method. The results are presented in Results and Discussion section.

β-hairpin Model

A β-hairpin system is considered as a second example. In particular, the Honeycutt–Thirumalai C_α coarse-grained model.[100] This coarse-grained model allows for a faster sampling of larger scale motions and disregards individual atomic vibrations that do not contribute to the overall hairpin shape. The fact that this model has 22 effective "particles" yields the same computational cost as for the alanine dipeptide model for the basic shape similarity operation in the ScIMAP algorithm, and provides a good example of how the method can be applied independently from the molecular model used. The coarse-grained model for the hairpin considers three distinct amino acid types:

- P: Polar or hydrophilic residues.
- H: Hydrophobic residues.
- N: Neutral residues.

The sequence for the hairpin studied is:

$$PH_9(NP)_2NHPH_3PH$$

The folded conformation of the β-hairpin is shown in Figure 4, and consists of two strands, one two residues

[†] Another angle was found to have an important span, but only in vacuum.[83,84] This angle is defined as the torsion formed by the atoms O—C—N—C_α, before the ϕ angle.

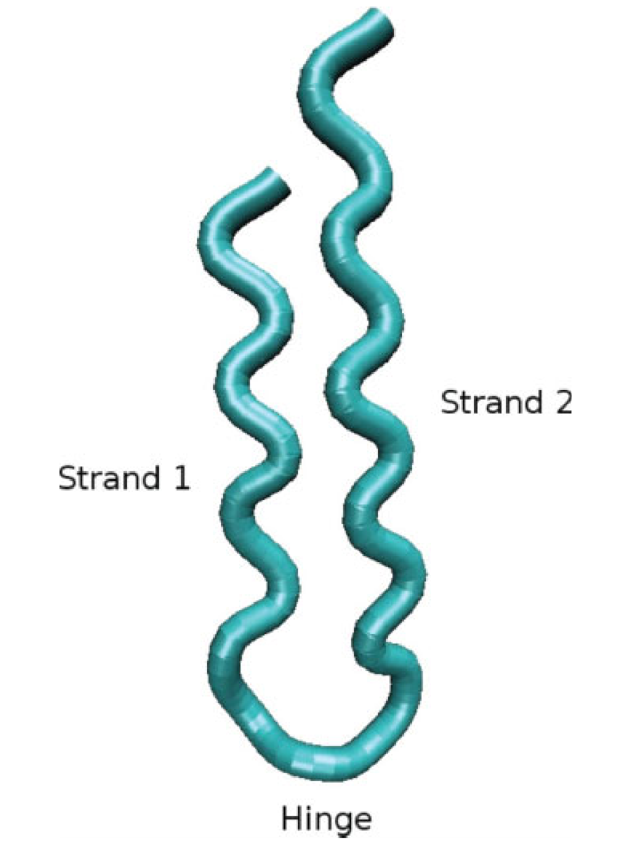

Figure 4

Model of a 22-residue β-hairpin rendered as a tubular representation, showing the "closed" (folded) conformation. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

longer than the other, connected by a hinge region (the $(NP)_2$ stretch of the sequence) and packed together. An exhaustive sampling of the conformational space was obtained by running Langevin dynamics simulation, around the folding temperature T_f ($T_f = 0.7$ in the system natural units).[101] The data sampling was obtained by running eight independent simulations starting from different random initial conditions. Each of the eight simulations gathered 45,000 conformations for a total of 360,000 conformations.

The coarse-grained energy function assigns a certain degree of rigidity to the strands, and more flexibility to the hinge residues.[100] However, the hairpin still exhibits bending and twisting of the strands. Previous computational studies[100,101] have shown that this hairpin model exhibits a two-state folding behavior where only a closed (folded) and an open (unfolded) state are significantly populated.

RESULTS AND DISCUSSION

We present the results obtained from the application of the ScIMAP algorithm to both data sets. We show that in both cases the first two coordinates are sufficient to completely characterize the conformational landscape spanned by MD.

Alanine dipeptide

The ScIMAP algorithm was applied to the 500,000 simulated conformations using lRMSD as the distance measure, as explained earlier. To build the neighborhood, several values of k (namely, 10, 15, 20, and 25) were used to build a neighborhood graph. There was no significant difference in the recovered low-dimensional landscape obtained with different values of k, other than a slight shift in the placement of the free-energy minima, attesting to the robustness of the method against varying neighborhood parameters, as previously shown.[65] Five thousand landmarks were chosen randomly from the trajectory.

Some care needs to be taken in the definition of the lRMSD metric regarding conformations associated with the same physical state of this system. In particular, the hydrogen atoms in the CH_3 groups of the alanine dipeptide present a C3 symmetry around the C atom. The lRMSD metric, which considers each atom individually, will classify all three 120° rotationally symmetric positions of the hydrogen atoms as different, when in fact they should be considered indistinguishable from a chemical perspective. To circumvent this problem, the lRMSD distance for this system is defined modulus 120° rotations of hydrogens around the C atom. This allows to consider the hydrogen atoms as indistinguishable while their vibrations are still sampled. The free-energy landscape as a function of the first few ScIMAP coordinates is shown in Figure 5. All the main states of the peptide described in Meterials and Methods section are correctly recovered as free-energy minima. In particular, the first ScIMAP coordinate clearly distinguishes between the "extended" and "helical" states; adding the second ScIMAP coordinate separates the two distinct routes connecting the minima.

Figure 3 shows that there are two regions where the ψ angle makes the transition from the helical to the extended shapes: around $\psi \approx 55°$ (near the top of the plot) and $\psi \approx -100°$ (wrapping around the vertical axis, near the bottom of the plot). In the low-dimensional embedding, the first two coordinates clearly capture the circular topology of ψ.

Since ϕ is not sampled in its full 360°, the use of one extra dimension (the third ScIMAP coordinate) more clearly separates the two main conformational states. The less likely α_L state is identified as a cluster of its own, separated from the other states.

The ScIMAP coordinates are ordered by data variance, so that the first coordinate explains the most data variability, the second coordinate adds the most variability after that, and so on. Figure 6 shows that the first two

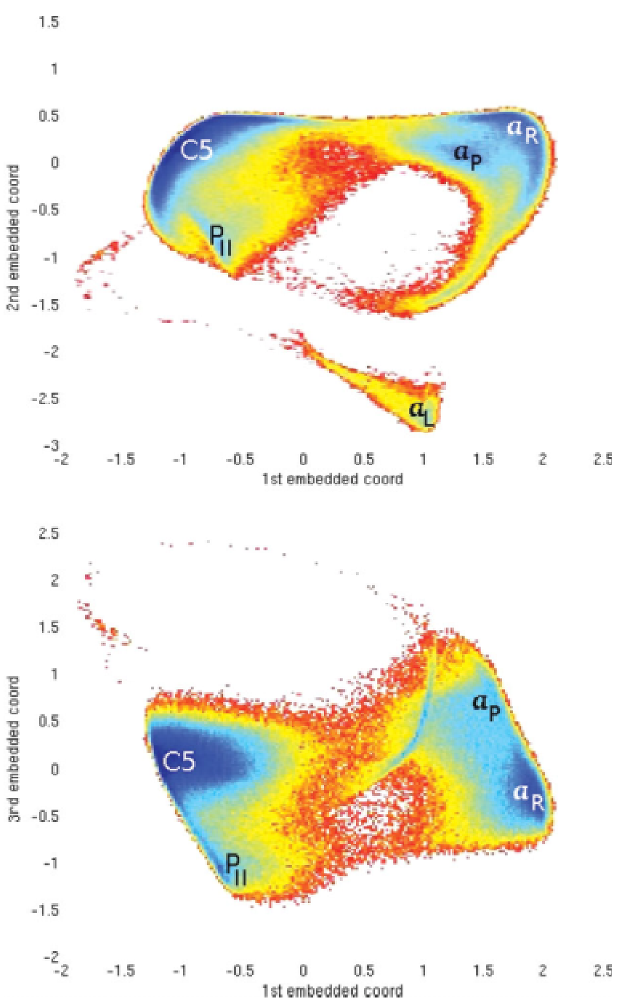

Figure 5

Free energy versus the first three ScIMAP coordinates for the indistinguishable hydrogen model of the alanine dipeptide. First to second coordinates (top), and first to third coordinates (bottom).

coordinates provide a good representation of the data, as the residual variance for coordinates higher than two is <0.05. Residual variance for each dimension d is computed as $1 - R_d$, where R_d is the squared correlation coefficient between the original (geodesic) distances and the corresponding Euclidean distances for the same pair of conformations, using the first d low-dimensional coordinates. Some features of interest may still be captured in dimensions higher than the third coordinate, although they would correspond to more localized events. In particular, a careful analysis of the fourth, fifth, and sixth ScIMAP coordinates provides a classification of the different configurations of the capping CH_3 dihedral angle (data not shown).

For comparison, results obtained from the application of PCA are shown in Figure 6. Note that for this system, the PCA projection resembles the ScIMAP results, in the sense that the conformations are projected into a barrel-like shape. PCA was applied to the coordinates after re-positioning the hydrogen atoms to take into account of their indistiguishability, as explained earlier. Because of the small size of the system and the fact that the atoms do not move far from their equilibrium positions, PCA can capture the main motions and the rotation around the ψ and ϕ angles. However, the linear projection done by PCA mixes the cluster boundaries and does not provide the clean separation (and transitions) of the conformational states that ScIMAP does. In other words, even though four main minima are observed (labeled in Fig. 6), the correspondence with the four main states is not as clear as with the ScIMAP coordinates. In addition, the α_L state is mixed together with the other two helix-like states, and cannot be distinguished in Figure 6; clearly, a PCA projection of coordinates cannot capture the difference between a right- and a left-turn of the alanine dipeptide; consistently, the residual variance comparison reflects the lower accuracy of PCA with respect to ScIMAP. Residual variance is generally considered the measure of choice to estimate the error in dimensionality reduction,[80] and it has been previously used to compare Isomap and PCA coordinates in protein dynamics.[65]

β-hairpin

The ScIMAP results presented here for the analysis of the β-hairpin configurational data were obtained using lRMSD as distance measure and $k = 12$ neighbors. ScIMAP embedding for $k = 15$, 10, 5, and 3 were also performed to check the robustness of the procedure, and yield almost identical results. It is worth noting that the application of the ScIMAP method to this system is no more expensive than for the all-atom alanine dipeptide model presented earlier: in both cases the data consist of three-dimensional configurations with 22 “atoms.” Figure 7 shows the free-energy profile for the β-hairpin model, as a function of the first three ScIMAP coordinates. The first ScIMAP coordinate clearly distinguishes between the “closed” (free energy minimum on the left side in Fig. 7), and the “open” (free energy minimum on the right side) hairpin conformations, accounting for the main direction of the folding reaction. The second ScIMAP coordinate reveals additional features on the folding process; the free energy plot as a function of the first two ScIMAP coordinates exhibits a symmetry along the second coordinate, which can be explained by looking at representative hairpin shapes placed by ScIMAP in the two local free-energy minima symmetrically located, above and below the deeper minimum corresponding to the closed state, as illustrated in Figure 7. Conformations labeled as **M1** and **M2** in Figure 7 (representatives from the top and bottom minima, respectively) correspond to two stereo-chemically different partially misfolded twists of the hairpin. It is worth stressing that the Isomap’s

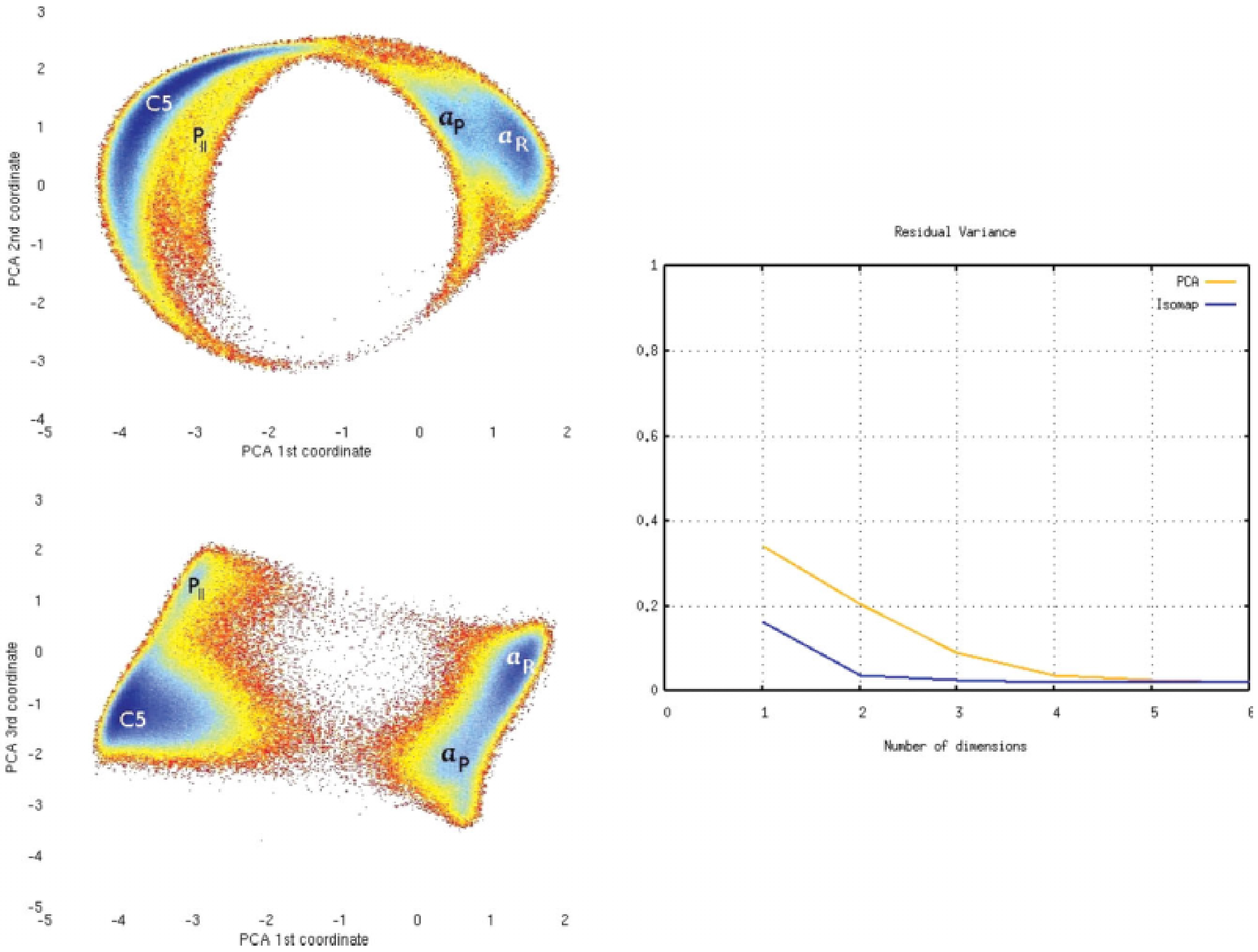

Figure 6

Free energy as a function of the first three PCA coordinates for the β-hairpin model (left). Residual variance of PCA compared with ScIMAP (right).

geodesic distance formulation, used in ScIMAP, automatically captures the topological difference between M1 and M2, while the lRMSD distance alone would have classified both configurational states as very similar. The transitions between the M, F, and U states clearly show that the hairpin cannot directly "jump" from an M1 conformation to an M2 conformation without separating the strands first. This can happen in two ways: either the hairpin folds into the F state, which leaves both strands in antiparallel position allowing their re-positioning, or the hairpin unfolds into the U state from which the re-positioning can also occur. The free-energy surface as a function of the first two coordinates clearly shows these possible transitions as saddle points.

In Figure 7, the third ScIMAP coordinate is also shown to provide a more exhaustive analysis. This dimension adds to explaining the data variability but does not introduce new minima or transitions. A free energy landscape computed as a function of the first three PCA coordinates is shown in Figure 8 for comparison. Since the hairpin's conformational landscape is relatively simpler than a full-size protein,[65] PCA can roughly separate open and closed states, and has a third minimum to the right, also roughly corresponding to a single semi-open state. However, the separation is less clear. A representative ensemble of conformations picked from the F and M regions includes many misclassified shapes that fall into both minima. Obviously, the M1 and M2 states explained earlier, which ScIMAP separates, and can be accessed through the F or U states, cannot be clearly distinguished by PCA. The residual variance of ScIMAP and PCA as a function of the number of dimensions used is also shown. Quantitatively, ScIMAP clearly classifies the data variance better with just one coordinate. Qualitatively, the nonlinear nature of ScIMAP captures more interesting features than PCA, beyond the residual variance comparison. The clear separation of the two stereo-chemically symmetric states illustrates

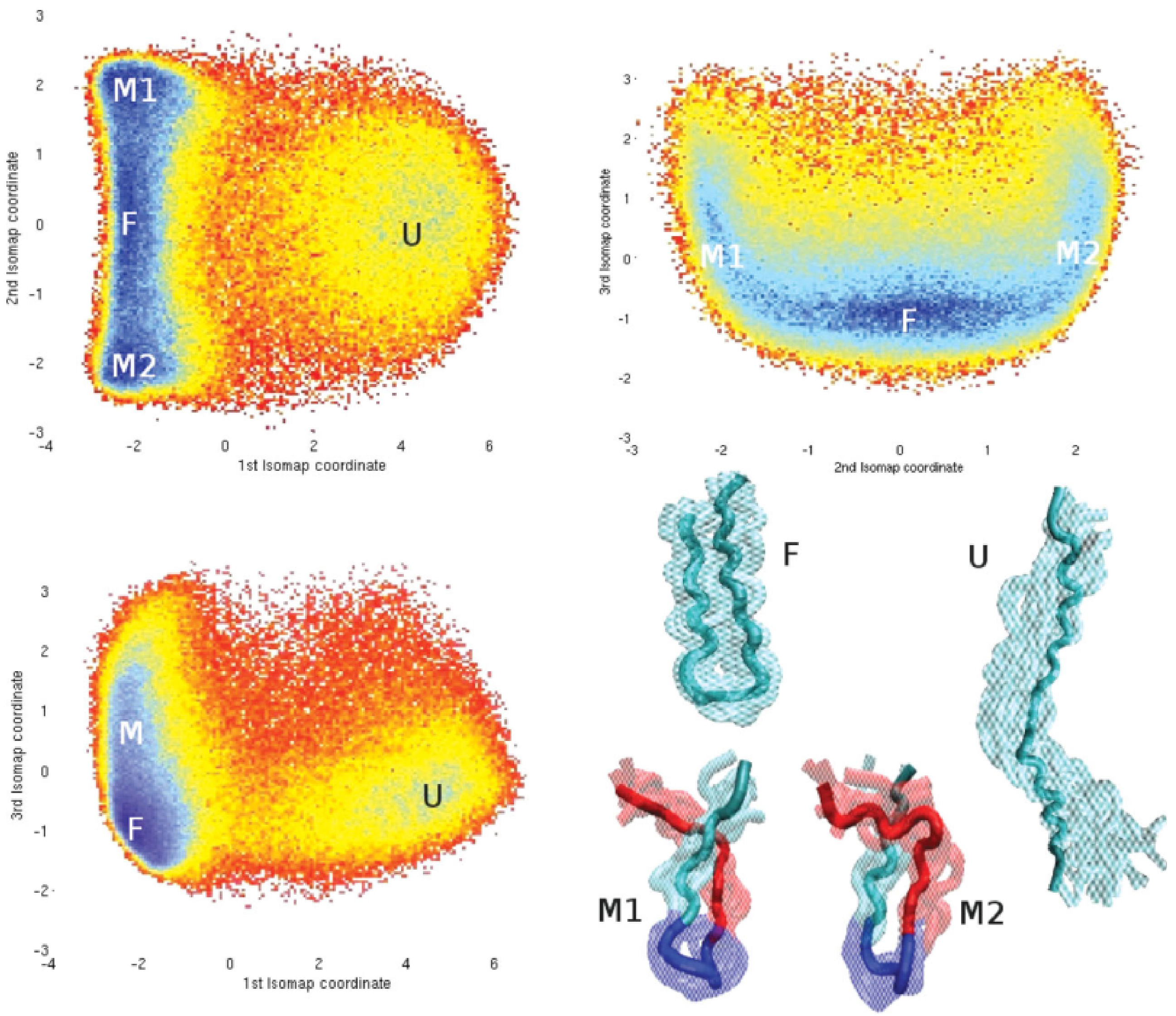

Figure 7

Free energy plots using the first three ScIMAP coordinates. Representative ensemble pictures for the four main minima are shown in the bottom right panel.

ScIMAP's superior performance in identifying states and transitions.

Resource utilization

The computation of ScIMAP coordinates is more computationally expensive than traditional structural reaction coordinates, since global information is being preserved. This is reflected in the topology-preserving mapping, which is based on the neighborhood around each input point. The computation of the nearest-neighbors graph remains the bottleneck of the procedure. For both models presented here, the ScIMAP calculations were performed on a cluster of 50 processors (AMD Opteron 275, at 2.2 GHz). The running times and memory usage are summarized in Table 1. Wall time refers to the actual time elapsed (as opposed to the time spent only on computation-exclusive CPU cycles). Wall time is typically used to report performance of parallel algorithms since it includes time spent on communication between processors. Table 1 clearly shows that the neighbor computation stage takes significantly longer than the other two stages. However, this is the stage that requires the least amount of memory. On the other hand, building a matrix of geodesic distances of size $n \times n_l$, where n is the number of points and n_l the number of landmarks, requires almost all of the memory available to each processor. The amount of memory used per processor can be reduced

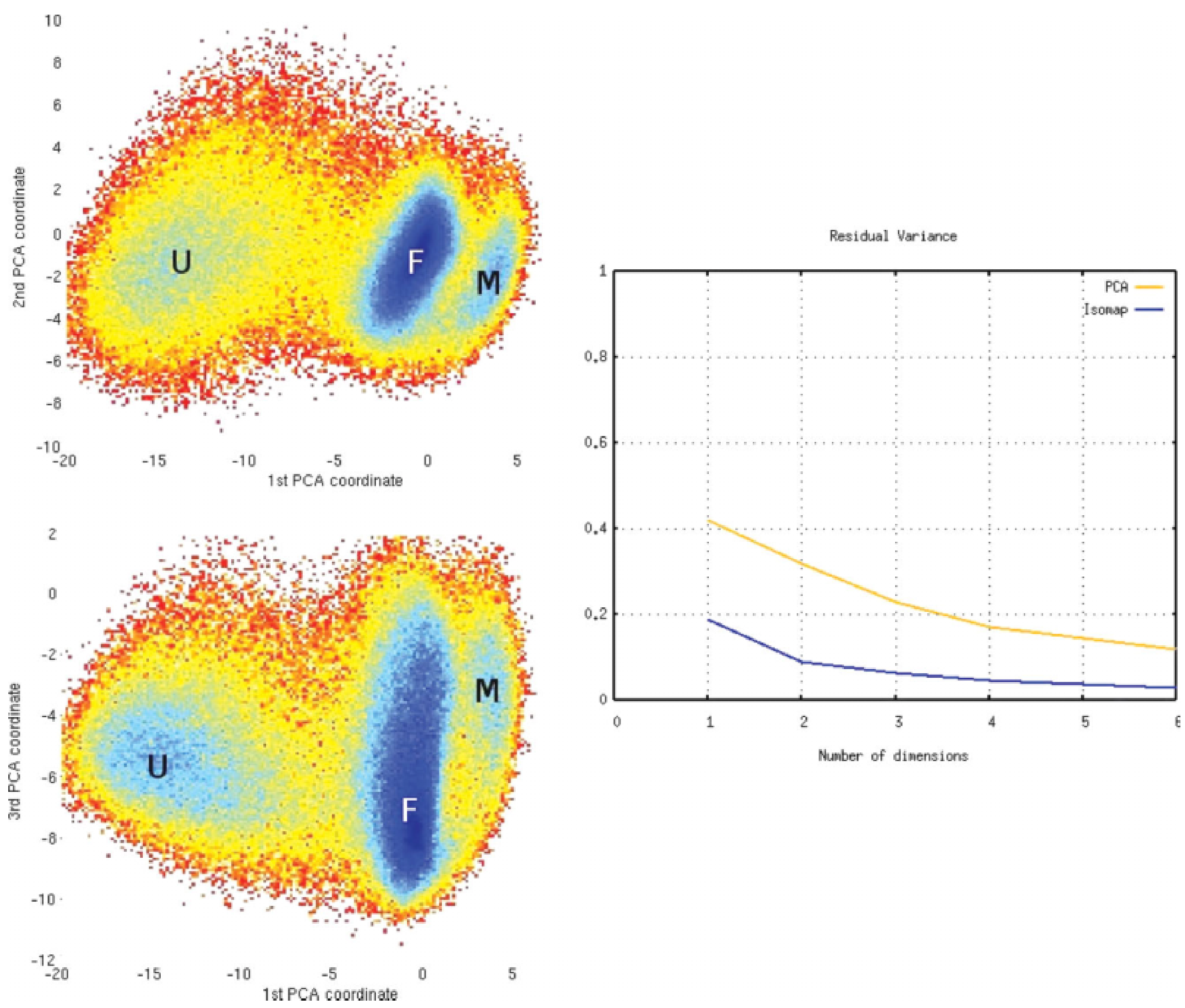

Figure 8

Free-energy plots using the first three PCA coordinates. The U, F, and M states only roughly correspond to those in Figure 7. The residual variance versus number of computed dimensions for both ScIMAP and PCA is shown to the right.

by using more processors and/or fewer landmarks, but maximizing the number of landmarks results in higher quality coordinates for a given number of processors.[65] Computing the final coordinates requires the computation of the top eigenvalues and corresponding eigenvectors of a similarly sized matrix.

CONCLUSION

We have presented the results obtained in the application of the ScIMAP algorithm to analyze a large configurational sampling of an all-atom model of alanine dipeptide, and a coarse-grained β-hairpin model. We have shown that the low-dimensional representation

Table I

Resource Utilization of ScIMAP

ScIMAP stage	Wall time	Memory
(a) Alanine dipeptide (500,000 conformations)		
Neighbor finding ($k = 20$)	7 h	70 MB
Geodesics (5000 landmarks)	10 min	1800 MB
Embedding coordinates	12 min	2200 MB
(b) β-Hairpin (360,000 conformations)		
Neighbor finding ($k = 12$)	5 h	50 MB
Geodesics (5000 landmarks)	4 min	1200 MB
Embedding coordinates	3 min	1500 MB

Computational resource utilization of ScIMAP for the systems studied. Wall time indicates the actual time used by 50 processors in parallel. The memory usage shown in per-node.

obtained by ScIMAP, using shape similarity as a basic operation and no information on the actual degrees of freedom, successfully classifies the samples along axes that have a high correspondence with *a priori* known parameters. In the case of alanine dipeptide, the ϕ and ψ backbone angles are recovered as the most important coordinates of the system and the first automatically recovered coordinate differentiates between the two main shapes of the peptide: extended and helical. In the case of the coarse-grained model of a β-hairpin the first ScIMAP coordinate follows the main folding/unfolding reaction, whereas the second coordinate distinguishes two stereo-chemically symmetric partially misfolded states. This example shows the power of the geodesic formulation of ScIMAP, separating geometrically similar states that cannot be reached directly one from the other, and the possible routes connecting them.

This work illustrates the robustness of the ScIMAP method against different models and further validates its usefulness to automatically extract structural reaction coordinates from simulation data, in an unbiased way. Even though computing these coordinates is computationally more expensive than computing most reaction coordinates used to date, in many cases it may eliminate the need of devising custom, empirically designed, reaction coordinates.

An intriguing question that remains to be answered is whether a physical interpretation can be generally associated to the reaction coordinates obtained by the ScIMAP algorithm. At this stage of development, the method does not provide a straightforward way to interpret the resulting coordinates, nor if/what particular features is missed when a reduced number of variables is used. In the systems discussed here, we could *a posteriori* interpret the extracted coordinates by comparing them with *a priori* known physical observables (e.g., specific dihedral angles, opening of the angle between the hairpin strands); work toward a more general understanding of the meaning (if any) of automatically extracted reaction coordinates is ongoing.

ACKNOWLEDGMENTS

The authors thank Lorenzo Mazzoni for his help with the β-hairpin simulation, Amarda Shehu for her help in the data analysis, and Paul Ledbetter for his comments on the manuscript.

REFERENCES

1. Hansson T, Oostenbrink C, van Gunsteren WF. Molecular dynamics simulations. Curr Opin Struct Biol 2002;12:190–196.
2. Tai K. Conformational sampling for the impatient. Biophys Chem 2004;107:213–220.
3. Karplus M, Kuriyan J. Molecular dynamics and protein function. Proc Natl Acad Sci 2005;102:6679–6685.
4. Liwo A, Khalili M, Scheraga HA. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptied chains. Proc Natl Acad Sci USA 2005; 102:2362–2367.
5. Izvekov S, Voth GA. A multiscale coarse-graining method for biomolecular systems. J Phys Chem Lett B 2005;109:2469–2473.
6. Chu JW, Ayton GS, Izvekov S, Voth GA. Emerging methods for multiscale simulation of biomolecular systems. Mol Phys 2007; 105:167–175.
7. Tama F, Charles L Brooks I. Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. Ann Rev Biophy Biomol Struct 2006;35:115–133.
8. Tozzini V, Trylska J, En Chang C, McCammon JA. Flap opening dynamics in hiv-1 protease explored with a coarse-grained model. J Struct Biol 2007;157:606–615.
9. Doruker P, Jernigan RL, Bahar I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. J Comput Chem 2002;23:119–127.
10. Kevrekidis IG, Gear CW, Hummer G. Equation-free: The computer-aided analysis of complex multiscale systems. AIChE J 2004; 50:1346–1355.
11. Clementi C, Nymeyer H, Onuchic J. Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298:937–953.
12. Nielsen SO, Lopez CF, Srinivas G, Klein ML. Coarse grain models and the computer simulation of soft materials. J Phys Condens Matter 2004;16:481–512.
13. Das P, Matysiak S, Clementi C. Balancing energy and entropy: a minimalist model for the characterization of protein folding landscapes. Proc Natl Acad Sci USA 2005;102:10141–10146.
14. Levy Y, Cho SS, Onuchic JN, Wolynes PG. A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. J Mol Biol 2005; 346:1121–1145.
15. Praprotnik M, Matysiak S, Delle Site L, Kremer K, Clementi C. Adaptive resolution simulation of liquid water. J Phys: Condens Matter 2007;19:292201.
16. Head-Gordon T, Brown S. Minimalist models for protein folding and design. Curr Opin Struct Biol 2003;13:160–167.
17. Karanicolas J, Brooks C, III. Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. J Mol Biol 2003;334:309–325.
18. Ayton G, Noid W, Voth G. Multiscale modeling of biomolecular systems: in serial and in parallel. Curr Opin Struct Biol 2007; 17:192–198.
19. Christen M, van Gunsteren W. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. J Chem Phys 2006;124:154106.
20. Dokholyan NV. Studies of folding and misfolding using simplified models. Curr Opin Struct Biol 2006;16:79–85.
21. De Mori G, Colombo G, Micheletti C. Study of the villin headpiece folding dynamics by combining coarse-grained monte carlo evolution and all-atom molecular dynamics. Proteins 2005;58:459–471.
22. Lyman E, Ytreberg F, Zuckerman D. Resolution exchange simulation. Phys Rev Lett 2006;96:028105.
23. Das P, Wilson C, Fossati G, Wittung-Stafshede P, Matthews K, Clementi C. Characterization of the folding landscape of monomeric lactose repressor: quantitative comparison of theory and experiment. Proc Natl Acad Sci USA 2005;102:14569–14574.
24. Matysiak S, Clementi C. Minimalist protein model as a diagnostic tool for misfolding and aggregation. J Mol Biol 2006;363:297–308.
25. Marrink S, Risselada H, Yefimov S, Tieleman D, de Vries A. The MARTINI forcefield: coarse grained model for biomolecular simulations. J Phys Chem B 2007;111:7812–7824.
26. Clementi C. Coarse-grained models of protein folding: toy-models or predictive tools? Curr Opin Struct Biol 2008;18:10–15.

27. Kwak W, Hansmann UH. Efficient sampling of protein structures by model hopping. Phys Rev Lett 2005;95:138102.
28. Neri M, Anselmi C, Cascella M, Maritan A, Carloni P. Coarse-grained model of proteins incorporating atomistic detail of the active site. Phys Rev Lett 2005;95:218102.
29. Praprotnik M, Delle Site L, Kremer K. Adaptive resolution molecular-dynamics simulation: changing the degrees of freedom on the fly. J Chem Phys 2005;123:224106.
30. Olender R, Elber R. Calculation of classical trajectories with a very large time step: formalism and numerical examples. J Chem Phys 1996;105:9299–9315.
31. Xu L, Henkelman G. Adaptive kinetic monte carlo for first-principles accelerated dynamics. J Chem Phys 2008;129:114104.
32. Yang LJ, Gao YQ. An approximate method in using molecular mechanics simulations to study slow protein conformational changes. J Phys Chem B 2007;111:2969–2975.
33. Henkelman G, Jonsson H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. J Chem Phys 1999;111:7010–7022.
34. Andricioaei I, Straub JE, Voter AF. Smart darting monte carlo. J Chem Phys 2001;114:6994–7000.
35. Shim Y, Amar JG, Uberuaga BP, Voter AF. Reaching extended length scales and time scales in atomistic simulations via spatially parallel temperature-accelerated dynamics. Phys Rev B 2007;76: 205439.
36. Rensen MRS, Voter AF. Temperature-accelerated dynamics for simulation of infrequent events. J Chem Phys 2000;112:9599–9606.
37. Marinari E, Parisi G. Simulated tempering—a new monte-carlo scheme. Europhys Lett 1992;19:451–458.
38. Hansmann U, Okamoto Y. Generalized-ensemble monte carlo method for systems with rough energy landscape. Phys Rev E 1997; 56:2228–2233.
39. Whitfield TW, Bu L, Straub JE. Generalized parallel sampling. Phys A: Stati Mech Appl 2002;305:157–171.
40. Kim J, Keyes T, Straub JE. Replica exchange statistical temperature monte carlo. J Chem Phys 2009;130:124112.
41. Chen JH, Brooks CL, Khandogin J. Recent advances in implicit solvent-based methods for biomolecular simulations. Curr Opinion Struct Biol 2008;18:140–148.
42. Jayachandran G, Vishal V, Pande VS. Using massively parallel simulation and markovian models to study protein folding: Examining the dynamics of the villin headpiece. J Chem Phys 2006;124: 164902.
43. Baumketner A, Shea JE, Hiwatari Y. Improved theoretical description of protein folding kinetics from rotations in the phase space of relevant order parameters. J Chem Phys 2004;121:1114–1120.
44. Cho S, Levy Y, Wolynes P. P versus Q: structural reaction coordinates capture protein folding on smooth landscapes. Proc Natl Acad Sci USA 2006;103:586–591.
45. Elber R. Long-timescale simulation methods. Curr Opin Struct Biol 2005;15:151–156.
46. Metzner P, Schuette C, Vanden-Eijnden E. Illustration of transition path theory on a collection of simple examples. J Chem Phys 2006;125:084110.
47. E W, Vanden-Eijnden E. Towards a theory of transition paths. J Stat Phys 2006;123:503–523.
48. Bolhuis P, Dellago C, Chandler D. Reaction coordinates of biomolecular isomerization. Proc Natl Acad Sci USA 2000; 97:5877–5882.
49. Bolhuis P, Chandler D, Dellago C, Geissler P. Transition path sampling: throwing ropes over rough mountain passes, in the dark. Annu Rev Phys Chem 2002;53:291–318.
50. Dellago C, Bolhuis P, Geissler P. Transition path sampling. Adv Chem Phys 2002;123:1–78.
51. Dellago C, Bolhuis PG. Transition path sampling simulations of biological systems. Top Curr Chem 2007;268:291–317.
52. Voter AF, Doll JD. Dynamical corrections to transition state theory for multistate systems: surface self-diffusion in the rare-event regime. J Chem Phys 1985;82:80–92.
53. Hinrichs NS, Pande VS. Calculation of the distribution of eigenvalues and eigenvectors in markovian state models for molecular dynamics. J Chem Phys 2007;126:244101.
54. Singhal N, Snow CD, Pande VS. Using path sampling to build better markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. J Chem Phys 2004; 121:415–425.
55. Elber R. A milestoning study of the kinetics of an allosteric transition: atomically detailed simulations of deoxy scapharca hemoglobin. Biophys J 2007;92:L85–L87.
56. West AMA, Elber R, Shalloway D. Extending molecular dynamics time scales with milestoning: example of complex kinetics in a solvated peptide. J Chem Phys 2007;126:145104.
57. Vanden-Eijnden E, Venturoli M, Ciccotti G, Elber R. On the assumptions underlying milestoning. J Chem Phys 2008;129: 174102.
58. Henkelman G, Uberuaga BP, Jónsson H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. J Chem Phys 2000;113:9901–9904.
59. Sheppard D, Terrell R, Henkelman G. Optimization methods for finding minimum energy paths. J Chem Phys 2008;128:134106.
60. E W, Ren W, Vanden-Eijnden E. String method for the study of rare events. Phys Rev B 2002;66:052301.
61. Maragliano L, Vanden-Eijnden E. On-the-fly string method for minimum free energy paths calculation. Chem Phys Lett 2007; 446:182–190.
62. Miller TF, Vanden-Eijnden E, Chandler D. Solvent coarse-graining and the string method applied to the hydrophobic collapse of a hydrated chain. Proc Natl Acad Sci USA 2007;104:14559–14564.
63. Ma A, Dinner AR. Automatic method for identifying reaction coordinates in complex systems. J Phys Chem B 2005;109:6769–6779.
64. Parida L, Zhou R. Combinatorial pattern discovery approach for the folding trajectory analysis of a β-hairpin. PLoS Comput Biol 2005;1:32–40.
65. 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.
66. Carreira-Perpinan MA. Dimensionality reduction, 1st ed. Chapman & Hall/CRC; 2010, 320 p.
67. Lee JA, Verleysen M. Nonlinear dimensionality reduction. Springer, Information Science and Statistics series; 2007, 310 pp.
68. Benito M, Pena D. Dimensionality reduction with image data. Lect Notes Comput Sci 2004;3177:326–332.
69. Cho E, Kim D, Lee S. Posed face image synthesis using nonlinear manifold learning. Lect Notes Comput Sci 2003;2688:946–954.
70. Kirby M, Sirovich L. Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Trans Pattern Anal Mach Intell 1990;12:103–108.
71. Turk M, Pentland A. Face recognition using eigenfaces. In: Proceedings of the IEEE conference in computer vision and pattern recognition. Maui, HI; 1991. pp 586–591.
72. Weinberger KQ, Saul LK. Unsupervised learning of image manifolds by semidefinite programming. In: Proceedings of the IEEE conference in computer vision and pattern recognition. Washington, DC; 2004. pp 988–995.
73. Teodoro M, Phillips G, Jr, Kavraki L. Understanding protein flexibility through dimensionality reduction. J Comp Biol 2003;10:617–634.
74. Garcia AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett 1992;68:2696–2699.
75. Zhang Z, Wriggers W. Local feature analysis: a statistical theory for reproducible essential dynamics of large macromolecules. Proteins: Struct Funct Biol 2006;64:391–403.

76. Wu D, Su W, Carpuat M. A Kernel PCA method for superior word sense disambiguation. In: Proceedings of the 42nd annual meeting of the association for computational linguistics. Morristown, NJ: Association for Computational Linguistics; 2004. p 637.
77. Jolliffe I. Principal components analysis. New York: Springer-Verlag; 1986.
78. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal component analysis and long time protein dynamics. J Phys Chem 1996; 100:2567–2572.
79. Scholkopf B, Smola A, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 1998;10:1299–1319.
80. Tenenbaum J, de Silva V, Langford J. A global geometric framework for nonlinear dimensionality reduction. Science 2000; 290:2319–2323.
81. Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290:2323–2326.
82. Passerone D, Ceccarelli M, Parrinello M. A concerted variational strategy for investigating rare events. J Chem Phys 2003;118:2025–2032.
83. Bolhuis PG, Dellago C, Chandler D. Reaction coordinates of biomolecular isomerization. Proc Natl Acad Sci USA 2000;97:5877–5882.
84. Drozdov AN, Grossfield A, Pappu RV. The role of solvent in determining conformational preferences of alanine dipeptide in water. J Am Chem Soc 2004;126:2574–2581.
85. Cox T, Cox M. Multidimensional scaling, 2nd ed. Chapman & Hall; 2000.
86. de Silva V, Tenenbaum J. Global versus local methods in nonlinear dimensionality reduction. In: Becker S, Thrun S, Obermayer K, editors, Advances in neural information processing systems 15. Cambridge, MA: MIT Press; 2002. pp 705–712.
87. de Berg M, van Krefeld M, Overmars M, Schwarzkopf O. Computational geometry: algorithms and applications, 2nd ed. Springer-Verlag, Berlin, Heidelberg, New York; 2000. 379 p.
88. Plaku E, Bekris KE, Kavraki LE. OOPS for motion planning: an online open-source programming system. In: IEEE international conference on robotics and automation. Rome, Italy; 2007. pp 3711–3716.
89. Plaku E, Stamati H, Clementi C, Kavraki LE. Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction. Proteins: Struct Funct Bioinform 2007;67:897–907.
90. Plaku E, Kavraki LE. Quantitative analysis of nearest-neighbors search in high-dimensional sampling-based motion planning. Intl Workshop on the Algorithmic Foundations of Robotics. New York, NY 2006. Springer Tracts in Advanced Robotics, 2008;47:3–18.
91. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem 2004;25:1157–1174.
92. Schiffer CA, Caldwell JW, Stroud RM, Kollman PA. Inclusion of solvation free energy with molecular mechanics energy: alanyl dipeptide as a test case. Protein Sci 1992;1:396–400.
93. Hummer G, Kevrekidis IG. Coarse molecular dynamics of a peptide fragment: free energy, kinetics, and long-time dynamics computations. J Chem Phys 2003;118:10762–10773.
94. Ferrenberg A, Swendsen R. Optimized Monte Carlo data analysis. Phys Rev Lett 1989;63:1185–1198.
95. Ferrenberg A, Swendsen R. New Monte Carlo technique for studying phase transitions. Phys Rev Lett 1988;61:2635–2638.
96. Roux B. The calculation of the potential of mean force using computer simulations. Comput Phys Commun 1995;91:275–282.
97. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem 1992; 13:1011–1021.
98. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. Multidimensional free-energy calculations using the weighted histogram analysis method. J Comput Chem 1995;16:1339–1350.
99. Ciccotti G, Lelievre T, Vanden-Eijnden E. Projection of diffusions on submanifolds: application to mean force computation. Commun Pure Appl Math 2008;61:371–408.
100. Guo ZY, Thirumalai D, Honeycutt JD. Folding kinetics of proteins—a model study. J Chem Phys 1992;97:525–535.
101. Mazzoni LN, Casetti L. Geometry of the energy landscape and folding transition in a simple model of a protein. Phys Rev E 2008;77:051917.