Simplifying the representation of complex free-energy landscapes using sketch-map

Michele Ceriotti[a], Gareth A. Tribello[b,1], and Michele Parrinello[b]

[a]Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QZ, United Kingdom; and [b]Computational Science, Department of Chemistry and Applied Biosciences, Eidgenössiche Technische Hochschule Zurich, Università della Svizzera Italiana Campus, Via Giuseppe Buffi 13 C-6900 Lugano, Switzerland

Contributed by Michele Parrinello, June 3, 2011 (sent for review April 28, 2011)

A new scheme, sketch-map, for obtaining a low-dimensional representation of the region of phase space explored during an enhanced dynamics simulation is proposed. We show evidence, from an examination of the distribution of pairwise distances between frames, that some features of the free-energy surface are inherently high-dimensional. This makes dimensionality reduction problematic because the data does not satisfy the assumptions made in conventional manifold learning algorithms We therefore propose that when dimensionality reduction is performed on trajectory data one should think of the resultant embedding as a quickly sketched set of directions rather than a road map. In other words, the embedding tells one about the connectivity between states but does not provide the vectors that correspond to the slow degrees of freedom. This realization informs the development of sketch-map, which endeavors to reproduce the proximity information from the high-dimensionality description in a space of lower dimensionality even when a faithful embedding is not possible.

nonlinear dimensionality reduction | proteins | molecular dynamics

The dynamics of many of the molecules that appear in biology, materials science, and chemistry are highly complex. These molecules can undergo transitions involving large numbers of atoms between an enormous number of different configurations (1), which makes it difficult to comprehend these motions using only chemical intuition. Nevertheless, within this data there is a lot of correlation, and there is a strong body of evidence that the energetically accessible regions of phase space lie on a structure that has a low dimensionality (2–6). Therefore, low-dimensionality representations of the free-energy surface can give meaningful insight into phenomena and can provide collective variables (CVs) that can be used to accelerate the dynamics and to reconstruct the free-energy landscape. Methods exist for extracting this low-dimensionality structure by postprocessing the results of long unbiased molecular dynamics trajectories in which the entirety of the landscape is explored (3, 6–8). Unfortunately however, for many systems—in particular for atomistic simulations—obtaining information on interesting, long-time-scale motions using unbiased simulations requires heroic amounts of computational time (9). Therefore, for these types of problems one would ideally like to use dimensionality reduction in tandem with accelerated sampling. This has to work both ways—the method must be able to analyze data from accelerated sampling simulations on very rough free-energy surfaces. Furthermore, it should produce a mapping of phase space that can serve as an optimized, bespoke set of CVs for calculations that extract quantitative free energies.

Experiments have shown that the low-free-energy part of phase space has a complex structure with a nonuniform dimensionality (8), that it is nonlinear (2, 4), that it is nonuniformly sampled (8, 10), and that it is possibly fractal (4, 11). It therefore seems likely that three, four, or even more vectors would be required to faithfully describe these complex topologies using the currently available dimensionality-reduction technologies. In fact, even for relatively simple systems, which can be sampled using unbiased dynamics, a very careful analysis is required to obtain a satisfactory three-dimensional description (7). This is problematic when it comes to using these methods to educate accelerated sampling algorithms because these methods work best with very low numbers of CVs—ideally one or two (12). Furthermore, it is of paramount importance that these CVs map all the basins in the free-energy surface to different parts of the *xy* plane as barriers to motion in transverse degrees of freedom can hinder the convergence of the free energy. Hence, in this paper we introduce an algorithm, sketch-map, that endeavors to reconcile these two conflicting aims. In doing this we first present an analysis of an enhanced-sampling trajectory, which explores the energetically accessible configurations for a simple polypeptide. This analysis demonstrates that there is a characteristic length scale at which the most valuable topological information about the free-energy landscape is encoded. Therefore, the design of sketch-map is predicated on the assumption that it is not necessary to produce an isometric embedding of the high-dimensionality manifold. Rather, one must preserve the proximity information and ensure that points closer than this characteristic distance are mapped close together, while simultaneously ensuring that the farther apart points are well separated in the projection.

Background

The only dimensionality-reduction algorithm that has been widely adopted within the simulation community is principal component analysis (PCA) (2–5). In this method one runs a simulation trajectory and calculates the means and variances for a large number of collective coordinates. By diagonalizing the resulting covariance matrix one can obtain the directions in which there are the largest structural fluctuations—the directions that are assumed to span the essential substance of the dynamics. However, the assumption that low-energy regions lie in a linear subspace of the full dimensionality space renders PCA appropriate in local regions but results in a poor characterization of the global, nonlinear features (6).

These deficiencies of PCA have led researchers to investigate other, nonlinear manifold learning algorithms and in particular locally linear embedding (LLE) (13), Isomap (6, 14, 15), and diffusion maps (7, 8, 10, 16). The first of these, LLE, is a nonlinear approach, which seeks to combine a set of locally linear descriptions in the vicinity of each trajectory frame into a single, unified embedding (13). It is common knowledge that algorithms like this one are very sensitive to noise (17). This forces one to question how effective this algorithm can be for molecular trajectories, which are typically very noisy (8). The alternative then are global

Author contributions: M.C., G.A.T., and M.P. designed research; M.C., G.A.T., and M.P. performed research; M.C., G.A.T., and M.P. analyzed data; and M.C., G.A.T., and M.P. wrote the paper.

The authors declare no conflict of interest.

See Commentary on page 12969.

[1]To whom correspondence should be addressed. E-mail: gareth.tribello@phys.chem.ethz.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1108486108/-/DCSupplemental.

approaches, which seek to reproduce all the pairwise distances between the D-dimensional frames by distributing their embeddings in a lower, d-dimensional space. The grandfather of these methods is multidimensional scaling (MDS) (18), which can be solved as an eigenvector problem or by minimization of a stress function. When Euclidean distances are used the eigenvector solution is equivalent to PCA, so approaches involving stress function minimization are often preferred because they are more flexible. By using a different metric to calculate distances, one can use MDS to fit nonlinear manifolds (19, 20). For instance, assuming the manifold is isometric with a linear space, one can use the geodesic distance (the distance along the manifold). This idea is the basis of the Isomap algorithm in which geodesic distances are obtained by calculating the length of the shortest path through a fully connected graph that is created by joining the points that are closest together (19). Calculating geodesics in this way assumes that the high-dimensionality points lie in a convex subset of R^D—i.e., it assumes that the low-dimensional manifold is uniformly sampled and there are no "holes." Donoho and Grimes (21, 22), in the context of image articulation, have demonstrated that for relatively simple cases this approximation is not valid and that in these cases Isomap fails to find the correct parameter space up to a linear mapping.

Currently the most promising approach for trajectory dimensionality reduction is diffusion maps (23–25), which can be formulated in a way that makes it resilient to noisy and nonuniformly distributed data (8). In this approach one defines a weighted graph on the simulation data and then uses the first few eigenvalues of the Laplacian of the manifold as the embedding coordinates. This approach is exciting because for the systems examined the vectors spanning the low-dimensionality manifold are those in which large barriers to motion make diffusion slow (8, 16). That said, the method has thus far only been applied to relatively simple systems and not to systems that require one to use accelerated sampling to explore phase space.

To demonstrate our method we use in this paper the folding of polyalanine-12, modeled with a distance-dependent dielectric ($\epsilon_{ij} = r_{ij}$ in Angstroms) that mimics some of the solvent effects. This system has been extensively studied (26) and has been shown to have a complex, funnel-shaped energy landscape with an alpha-helical global minimum that does not form during long MD simulations started from a random configuration (27). To accelerate the dynamics we therefore use the recently developed reconnaissance metadynamics method (see *Materials and Methods*) because with this method one can use a large number of CVs to characterize configurations and still obtain a qualitatively correct mapping of the free-energy surface (27). Furthermore, unlike in other papers on dimensionality reduction, we take advantage of the fact that changes in bond lengths, bond angles, and rigid peptide bond dihedrals along with the rotations of methyl groups are uninteresting. We therefore use only the 24 backbone dihedral angles (Fig. 1*A*) to characterize the various configurations visited during the trajectories.

The Free-Energy Landscape of a Polypeptide

Before introducing our dimensionality-reduction algorithm it is perhaps useful to step back for a moment and to examine some qualitative features of the protein's free-energy landscapes in detail. Therefore in Fig. 1*B* we project the set of configurations obtained from our reconnaissance metadynamics simulations onto two dihedrals. We find that, even for a trajectory in which relatively high energy states are sampled and regardless of the pair of dihedrals selected, the resulting distribution of angles is very similar to the Ramachandran plot. Hence, angles are not uniformly distributed across the available space and there are instead regions of high and low probability. This behavior was also seen by Sims et al. (28) when they examined the distribution of torsional angles for short peptide chains in higher dimensional spaces.

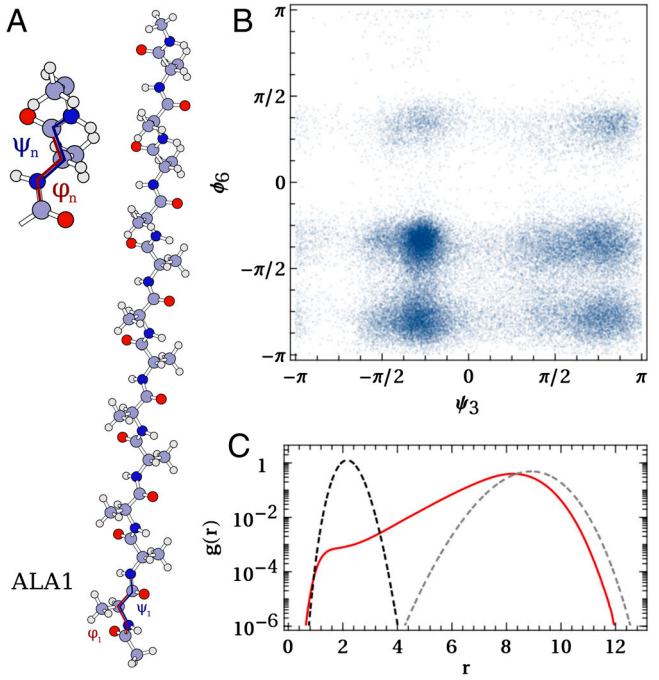

Fig. 1. Information on the distribution of torsional angles found in our reconnaissance metadynamics simulations. *A* shows the ala12 system examined and the backbone dihedral angles that were used as CVs. *B* shows a 2D projection of the distribution of angles found during reconnaissance. Here we show the distribution as a function of ψ in the third residue and of ϕ in the sixth residue, although the distribution of any pair of angles shows the same qualitative features. *C* shows (in red) a histogram for the distribution of distances between pairs of frames. Also shown in this figure is the distribution expected for a 24-dimensional, isotropic Gaussian with a standard deviation equal to 0.5 (black) and the distribution of distances expected for a set of points distributed uniformly across the 24-dimensional space (gray).

High-dimensionality spaces can often display very nonintuitive properties, which challenge our understanding of distance and proximity (29). We therefore cannot possibly expect to understand what structures are present simply by visualizing 2D projections. One quantity that can give us some feel as to whether or not it is feasible to represent the data in the lower dimensionality state is the histogram of pairwise distances, which is shown in Fig. 1*C*. Remarkably, the long range part of this distribution resembles that obtained from a uniform distribution of points in the full, 24-dimensional space.* In fact, only when r is less than eight is there a marked deviation from the uniform distribution—a slower decay toward zero. For values of r of about one this decay resembles that of a Gaussian distribution in the full, 24-dimensional space in agreement with what one would expect for the fluctuations within a harmonic basin. We therefore postulate that the most interesting distances are those between about two and eight because only here does the histogram resemble neither the Gaussian or uniform distribution.

Fig. 1*C* suggests that fitting protein free energy surfaces using dimensionality-reduction methods based on pure distance matching is impossible. The plain fact is that certain features of the distribution of distances are characteristic of points distributed in the full dimensionality space. This histogram can thus not be reproduced by projecting points in a space of lower dimensionality. In addition, it would appear that the free-energy surface has

*The distribution of distances between uniformly distributed points in a periodic space is related to the surface area of a diced hypersphere (30). In contrast to the distribution in a nonperiodic space there is a maximum in this function after which the function decays to zero at $r = \pi\sqrt{D}$.

a complex topology. This appears in our analysis because we use torsional angles that are inherently non-Euclidean to characterize configurations. However, there is evidence from the literature that protein potential energy surfaces have fractal dimensionalities (4, 11) or an otherwise intrinsically non-Euclidean topology.

The theory of energy landscapes suggests that energetically accessible configurations take up only a tiny fraction of phase space because these configurations are clustered together in basins, in which fluctuations take place in a high-dimensionality space, that are themselves connected by a spider's web of transition pathways (1). This picture is far more consistent with the information coming from our analysis of Fig. 1*C* and the structure of the Ramachandran plot than any picture in which all the low-energy regions of phase space lie on a low-dimensional, Euclidean manifold. Therefore, to test whether this is a realistic model for the energy landscape of ala12 we generated a set of points from a model potential that exhibits these features by importance sampling at a sufficiently large temperature for both basins and low-lying transition states to be sampled. The resulting collection of points thus resembles what could have been obtained from enhanced-sampling calculations and can be compared with the histogram of distances obtained for ala12 (Fig. 2). Similarly to what was observed for the protein (Fig. 1*C*) the distribution of pairwise distances only deviates from the histogram for a uniform distribution in the full-dimensionality, periodic space at short r, and at the shortest r the decay resembles that observed for the distribution of distances in a multivariate Gaussian in the full-dimensionality space. In fact the main qualitative difference for the two systems is that the deviation here is less pronounced, which is simply a consequence of the lower dimensionality of this potential. The similarities thus give us confidence in our conceptual picture for the shape of the protein free energy landscape in the high-dimensionality space.

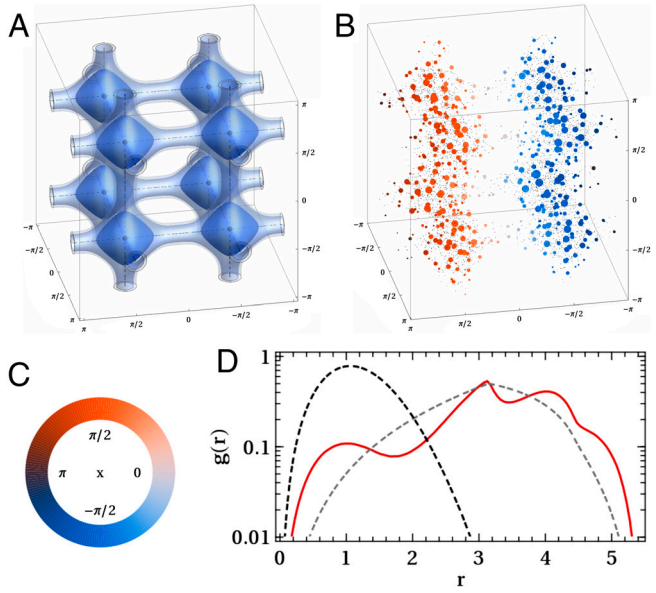

Fig. 2. Information on a model potential ($V(\theta,\phi,\psi) = \exp[3(3 - \sin^4(\theta) - \sin^4(\phi) - \sin^4(\psi))] - 1$), which exhibits many of the features that we believe characterize complex free-energy landscapes. In *A* the isosurfaces that enclose 50, 80 and 90% of the probability density for a particle diffusing about this potential at a temperature of $k_B T = e^3 - 1$ are shown. In *B* the distribution of points extracted from this potential through importance sampling are shown and the 500 landmark points selected using a farthest point sampling strategy are highlighted. In this panel the size of the landmarks is related to their weights and their colors depict the value of one of the angles. A key for the coloring is shown in *C*, and for the remainder of this paper, wherever points are colored according to the value of an angle, we ask the reader to refer to this scale. Finally, in *D* we show a histogram of the distances between pairs of generated points (red). This is again compared with the distribution expected for a 3D, isotropic Gaussian (black) and the distribution for a set of points distributed uniformly across the 3D space (gray).

Dimensionality Reduction Algorithm

One simple way to introduce nonlinearity in manifold learning algorithms is to perform distance matching but with the distances transformed (31) or weighted (32) so as to enhance the importance of certain connections—often the short distances (33). The analysis of the previous sections suggests that, if we could make the algorithm focus on reproducing distances from the interesting part of the histogram (the part where the distribution does not correspond to a high-dimensionality uniform or Gaussian distribution), this would be a useful approach for trajectory data. Furthermore, we can justify this approach based on our picture for the structure of the free-energy landscape by noting that by doing this we are focusing on reproducing the relations and connections between nearby basins and are discarding all the high-dimensionality, unfittable data on the internal structure of basins and the relative positions of distant basins. Our method, sketch-map, then is essentially multidimensional scaling, in which the distances in both the high- and low-dimensional spaces are transformed by a sigmoid function, which maps monotonically $\mathbb{R}^+$ to [0,1). Hence, one produces the mapping by minimizing (for details see *Materials and Methods*) the following stress function:

$$\chi^2 = \left(\sum_{j\neq i} w_i w_j\right)^{-1} \sum_{j\neq i} w_i w_j [F(R_{ij}) - f(r_{ij})]^2 \qquad [1]$$

where w_i is the weight of point i and $R_{ij} = |X_i - X_j|_{(D)}$ and $r_{ij} = |x_i - x_j|_{(d)}$ are the distances between points i and j in the high- and low-dimensionality spaces, respectively.[†] F and f are then both general sigmoid functions of the form:

$$s_{\sigma,a,b}(r) = 1 - (1 + (2^{a/b} - 1)(r/\sigma)^a)^{-b/a} \qquad [2]$$

where $s_{\sigma,a,b}(\sigma) = 1/2$ and the exponents a and b determine the rate at which the function approaches 0 and 1, respectively. The same value of σ is used in both F and f as using different values simply corresponds to a scaling of coordinates. However, we distinguish between the values of a and b in the two functions by using a_D and b_D for F and a_d and b_d for f.

When selecting parameters for the high-dimensionality space sigmoid function F, one is essentially selecting the length scales over which the connectivity data in the high-dimensionality space is interesting. The analysis presented in the previous section would therefore suggest that we should tune σ, a_D, and b_D so that for small values of R_{ij}, where the histogram resembles that of a full-dimensional multivariate Gaussian, $F(R_{ij}) \approx 0.0$, while for large values of R_{ij}, where the histogram of distances resembles that of a set of points uniformly distributed in the full-dimensional space, $F(R_{ij}) \approx 1.0$. This ensures that, once minimized, points that are close together in the D-dimensional space are mapped close together in the d-dimensional space and vice versa. Furthermore, because the error in the reproduction the distance R_{ij} contributes an amount to χ^2 that is proportional to $F'(R)$, a function that is peaked in the vicinity of σ, only a cursory attempt is made to reproduce the precise distribution of near and far neighbors around any given point. Meanwhile, the major focus during optimization is the reproduction of distances close to the value of the method's critical parameter, σ, which selects the interesting length scale for the problem. The values of a_D and b_D are far less important and, much like when similar functions are used to calculate continuous versions of coordination numbers,

[†] $|\cdot|$ does not have to be a Euclidean distance—here, for instance, we apply the minimum image convention to take account of the periodicity of the space.

the performance of the method only depends weakly on their values.

When the same parameters are used in the two sigmoid functions of Eq. **1** sketch-map, like MDS, will reproduce all pairwise distances if the configurations lie in a linear subspace of dimension d. However, given that we know the points are not distributed in this way, this choice is not appropriate and is in fact detrimental because, as shown in Fig. 1*C*, at short-range uninteresting fluctuations occur in the full D-dimensional space. Hence, distance matching involves the impossible task of mapping a manifold, which has parts where the radial density grows as r^{D-1}, into a space where radial density can grow only as r^{d-1}. In sketch-map we therefore use different a and b parameters for the two sigmoid functions to bypass this intractable problem. We note that for any distribution where the radial density around points grows as r^{D-1} the corresponding histogram of distances, transformed by $s_{\sigma,a,b}(r)$, is approximately equal to $s^{D/a-1}$ for small s. Therefore, for small s, the histograms of (differently) transformed distances for two distributions with radial densities that grow as r^{D-1} and r^{d-1} will be similar if $a_d/d \approx a_D/D$.

The minimization of Eq. **1** scales quadratically with the number of data points so when fitting a trajectory using sketch-map the first step is to select a small number of landmark frames (34), which, as detailed elsewhere, can be done either by selecting points at random or by using a farthest point sampling strategy (FPS) (35, 36). One can then assign weights to the landmarks based either on an estimate of the free energy, if available, or by computing the number of trajectory frames within each landmark's Voronoi polyhedron to ensure that reproduction of the structure in the low-energy parts of the landscape is weighted more in the fitting. Finally, once the minimization is completed, one can calculate the projection, x, of any high-dimensionality point X by minimizing:

$$\chi^2(x) = \left(\sum_{i=1}^{N} w_i\right)^{-1} \sum_{i=1}^{N} w_i [F(|X - X_i|_{(D)}) - f(|x - x_i|_{(d)})]^2 \quad [3]$$

where X_i is one of the landmark points and x_i is its low-dimensional projection. A global minimum for this quantity can be obtained by calculating the value of $\chi^2(x)$ on a grid and then using the lowest-lying point as a start point for a conjugate gradient minimization. The code for performing sketch-map is available online at sketchmap.berlios.de.

Results

Dimensionality Reduction Example. Before fitting the reconnaissance metadynamics data we first fit the data from the model potential shown in Fig. 2. Five hundred landmarks points were selected using a FPS strategy from the 5,000 frames generated by importance sampling. Their weights were then set equal to the number of frames within each landmark's Voronoi polyhedron. In the sketch-map result shown in Fig. 3 all eight basins are well separated and the majority of the connections are reproduced. This is an impressive result as this distribution is periodic in three dimensions and is thus not isometric with a linear, two-dimensional space. Nevertheless, unlike the other manifold learning algorithms we tested (see *SI Text*), sketch-map is able to circumvent this issue by breaking four of the connections between basins. The resulting embedding thus unrolls the box and gives the net shown in Fig. 3 rather than simply squashing the box onto the plane. This clear picture for the shape of phase space that emerges from our sketch-map projection is very appealing from the point of view of our eventual aim of using this method in tandem with biased MD.

Polyalanine-12. For the considerably more complex ala12 landscape we selected 1,000 landmark points from our reconnaissance metadynamics trajectories and again set their weights equal to the number of the remaining frames within each landmark's Voronoi polyhedron. Sketch-map parameters (given in Fig. 4) were then selected based on the shape of the histogram shown in Fig. 1*C*. After fitting we projected the nonlandmarks points, using (Eq. **3**). The final result is shown in Fig. 4, where points are colored in accordance with the number of residues that were identified as being part of an alpha helix or beta sheet by the STRIDE algorithm (37). Fig. 4 shows that embedded points are clustered in basins much like what is observed in full-dimensional description. Furthermore, there is a clear-cut separation between the regions of the plane that correspond to helix-like and sheet-like secondary structures. In the areas around each of these quintessential

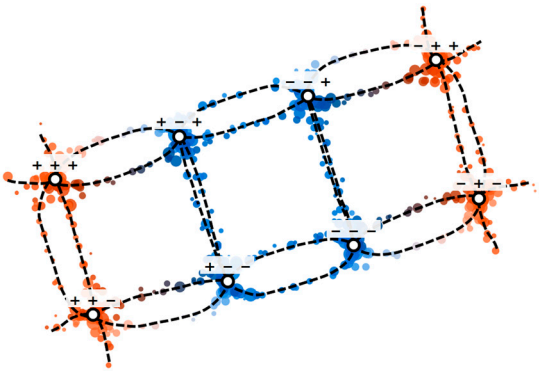

Fig. 3. A 2D, sketch-map projection of the landmark points selected from the dataset depicted in Fig. 2. This model has eight minima in the free-energy surface, which appear at $(\pm\pi/2, \pm\pi/2, \pm\pi/2)$. Projections of these points are indicated on this figure using labeled circles, while the various transition pathways are shown as dashed lines. Parameters for the sigmoid functions were chosen based on the histogram of distances (Fig. 2*D*) as $\sigma = 2$, $a_D = 3$, $b_D = 9$, $a_d = 2$, and $b_d = 2$. Projected points are colored, using the key shown in Fig. 2*C*, in accordance with the value of one of the three underlying variables.

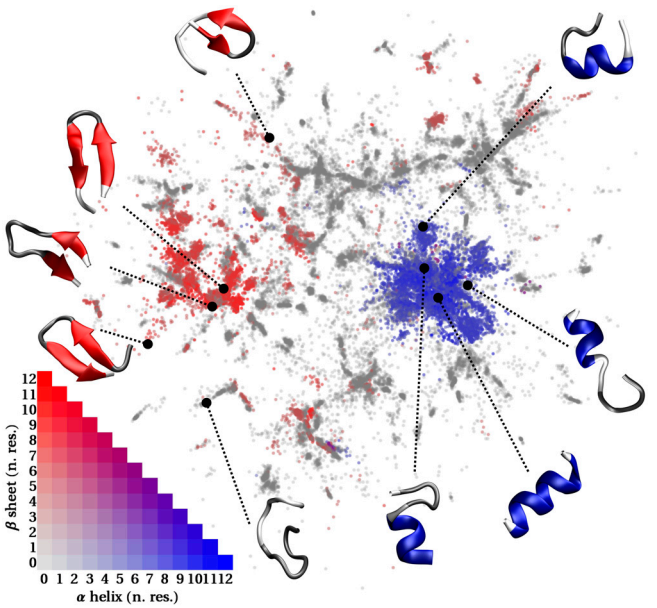

Fig. 4. Results for a projection of the frames obtained from the reconnaissance metadynamics simulations of ala12. The parameters of the sigmoid functions were chosen to be $\sigma = 6$, $a_D = b_D = 12$, $a_d = 1$, and $b_d = 2$. Points not included in the landmark set were embedded with the out-of-sample extension described in the text. The 2D projection is shown, with the embedded points colored in accordance with the number of residues that, according to STRIDE, are part of an alpha helix or beta sheet. A key for the color scheme is also shown, together with snapshots of a few selected configurations.

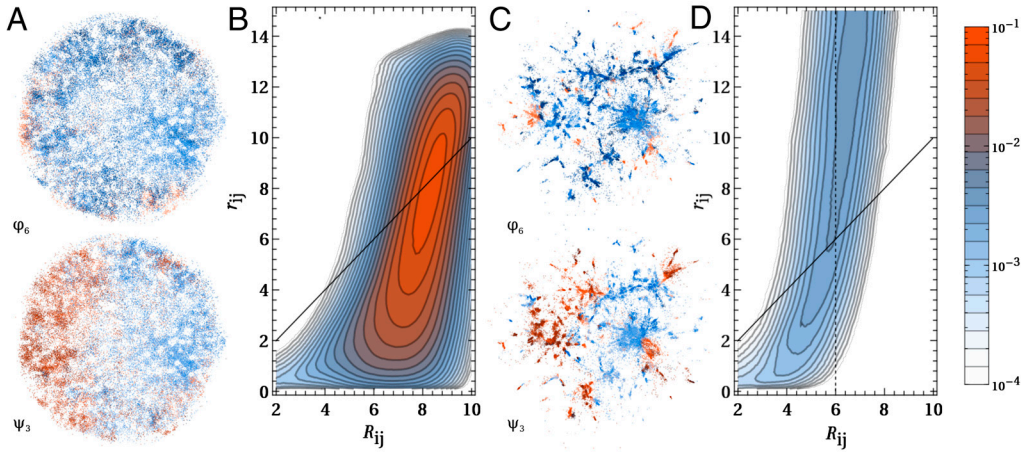

Fig. 5. *A* and *B* show diagnostic information on the embedding of the ala12 trajectory frames obtained by pure distance matching, while *C* and *D* refer to the results of sketch-map. Similar figures for other dimensionality reduction algorithms are provided in *SI Text*. In *A* an *C* the low-dimensional embedding of the frames is shown, colored according to two of the backbone dihedrals. *B* and *D* depict the joint probability distribution of the distances between two frames in high dimension (R_{ij}) and the distance between the corresponding low-dimensional projections, (r_{ij}).

protein configurations STRIDE identifies the structures as being mostly composed of coils and turns.

Fig. 5 gives more detailed diagnostic information and also compares the results of sketch-map with those from pure distance matching.[‡] The panels that show the embedded points colored in accordance with one of the backbone dihedrals demonstrate that sketch-map is better at clustering together points with similar values for a particular dihedral. More revealing, though, is the analysis of the joint probability distribution of low and high-dimensional pairwise distances between frames. Obviously, if the embedding is exact all density should be concentrated along the diagonal. However, as discussed above, this goal cannot be achieved, because of the intrinsically high-dimensional nature of the distribution of configurations at both short and large distances. Fig. 5 shows that, for distance matching, there is a sizable density in the region of the histogram corresponding to projection of points close together when they are in actuality fact far apart. This is disastrous in terms of using these coordinates to provide a coarse-grained description of configuration space as it means that structures that are very different from each other cannot be distinguished. In contrast, the histogram for the sketch-map result (Fig. 5*D*) result demonstrates that this algorithm only projects points that lie closer than σ in the high-dimensionality space close together.

Conclusions

For proteins and other chemical systems the manifold on which the energetically accessible region of phase space lies has a small volume but a very complex structure. It consists of small, high-dimensionality basins that are connected by a spider's web of transition pathways and its structure can be thought of in terms of a hierarchy of different length scales. On the smallest of these scales harmonic fluctuations in the full-dimensionality space take place. Changes in secondary and tertiary structure, meanwhile, take place over longer scales. Evidence presented here and elsewhere has demonstrated that one can recognize these different length scales by examining the distribution of distances between trajectory frames and that estimates of the dimensionality of the manifold depend on the length scale at which one examines the problem. Therefore, we contend that, when creating a low-dimensionality projection of a trajectory, one should first examine the distribution of distances and thereby identify the interesting length scale. Then, when projecting the data, an algorithm like sketch-map can be used so that the fitting effort is directed toward reproducing distances in the range that has been identified as interesting. Using these ideas we were able to produce a 2D mapping from a description of a set of protein configurations based on backbone dihedral angles. This mapping is able to reproduce the qualitative features of the free-energy landscape and clearly separates configurations with different protein secondary structures. Furthermore, in *SI Text* we show that sketch-map produces a qualitatively similar mapping when the set of distances between the C_α atoms is used to describe configurations. It may well be that for larger systems analysis of the distribution of distances will provide evidence of multiple interesting length scales in the problem. For these cases a hierarchical version of the sketch-map approach, which makes use of multiple sigmoid function with different σ parameters could be very useful.

The most successful approaches for performing dimensionality reduction on trajectory data do not assume that the low-dimensionality manifold, which contains the low-energy configurations, is isometric with a low-dimensional linear space. Instead these methods distort the distances between the high-dimensionality data points so that the essential features in the data can be represented in a low-dimensionality space. Sketch-map works in a similar manner and has this observation at its core. In addition, sketch-map produces an embedding from a very small number of landmark frames and is able to embed further configurations after the projection of this initial training set. This means that one can feasibly imagine combining sketch-map projections with enhanced sampling techniques to calculate the free-energy landscapes for systems in which the interesting events are not observable on the simulation time scale. Consequentially, we are currently working on ensuring this mapping is continuous so that the embedding can be used as CVs for biased MD.

In all of this work we focus on the data output by simulation trajectories, which presents a particular set of problems to manifold learning algorithms. However, the ubiquity of high-dimensionality data in disciplines of science, from chemistry and physics to social sciences and psychology, suggests that there is an abundance of potential applications of sketch-map.

Materials and Methods

Reconnaissance Metadynamics Simulations. All simulations of polyalanine were run using gromacs-4.5.1 (38), the amber96 forcefield (39) and a distance dependent dielectric. A time step of 2 fs was used, all bonds were kept rigid using the LINCS algorithm, and the van der Waals and electrostatic interactions were calculated without any cutoff. The global thermostat of Bussi et al. (40) was used to maintain the system at a temperature of 300 K. Recently we introduced an accelerated sampling method, reconnaissance metadynamics

[‡]Distance matching was performed by linear multidimensional scaling followed by iterative minimization of χ^2 with both of the sigmoid functions set to be the identity.

(27), which can be used with large numbers of collective variables. This method uses a self-learning algorithm to examine the trajectory and to construct an adaptive simulation bias that accelerates the exploration of phase space. We chose to use this method to perform the enhanced sampling calculations in this work and in particular the implementation of it in PLUMED (41). In previous work (27) we have shown that a 50-ns reconnaissance metadynamics simulation started from a random configuration can be used to find the alpha-helical, folded state of polyalanine-12. However, in this work, so as to have an extensive exploration of the region of phase space about the folded state, we took our trajectory data from four reconnaissance metadynamics simulations started from the folded state. In these simulations CVs were stored every 250 steps, whereas cluster analysis was done every 5×10^5 steps. The expansion parameter was set equal to 0.05, only basins with a weight greater than 0.2 were considered, and attempts were made every 1,000 steps to add to these basins hills of height 1.0 kJ mol^{-1} and width 1.5. During all calculations we stored frames every 8 ps for later analysis but discarded the first nanosecond of all simulations so as to ensure that our trajectories were independent. Hence, in this work all analysis is based on a set of 46,182 trajectory frames.

Optimization strategy. Eq. **1** is a nonconvex function and is thus very difficult to optimize. Moreover, the problem becomes stiffer as the sigmoid function becomes steeper at the inflection point. Hence, we have found that a combination of strategies is required to minimize χ^2 effectively. During the early stages of the minimization we introduce the better-behaved, although still nonconvex, merit function for least-squares distance matching: $\chi^2_{id} = (\sum_{i \neq j} w_i w_j)^{-1} \sum_{i \neq j} w_i w_j (R_{ij} - r_{ij})^2$. Iterative minimization of this function can be initialized using the result from classical MDS, which will minimize a χ^2_{id}-like stress in which all weights are equal to one. Once the minimum for the weighted χ^2_{id} is found, we introduce the sigmoid function by performing a series of minimizations of a stress function given by $\alpha\chi^2_{id} + (1 - \alpha)\chi^2$ in which we progressively reduce α from 1 to 0. During our experiments we found that the most effective strategy for iteratively minimizing stress functions like χ^2_{id} and χ^2 is to perform 20–50 steps of conjugate gradient optimization followed by a "pointwise-global" scheme in which the minimum stress position for each landmark point is found by minimizing Eq. **3**, while keeping the positions of the other landmarks fixed. Sweeping through all the landmarks a few times performing this second procedure allows one to quickly find an optimal configuration for all the projected points. This global optimization step becomes costly when the target dimensionality d is increased. However, χ^2 is a relatively smooth function, so an adaptive grid search or a more sophisticated global minimization algorithm could be used to reduce the overhead.

ACKNOWLEDGMENTS. The authors thank Michael Bronstein and Davide Branduardi for useful discussions and also acknowledge funding from the European Union (Grant ERC-2009-AdG-247075), the Royal Society, and the Swiss National Science Foundation.

1. Wales DJ (2003) *Energy Landscapes* (Cambridge Univ Press, Cambridge, UK).
2. Garcia AE (1992) Large-amplitude nonlinear motions in proteins. *Phys Rev Lett* 68:2696–2699.
3. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. *Proteins* 17:412–425.
4. Hegger R, Altis A, Nguyen PH, Stock G (2007) How complex is the dynamics of peptide folding? *Phys Rev Lett* 98:028102.
5. Zhuravlev PI, Materese CK, Papoian GA (2009) Deconstructing the native state: Energy landscapes, function and dynamics of globular proteins. *J Phys Chem B* 113:8800–8812.
6. Das P, Moll M, Stamati H, Kavraki LE, Clementi C (2006) Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc Natl Acad Sci USA* 103:9885–9890.
7. Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, Kevrekidis IG (2010) Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc Natl Acad Sci USA* 107:13597–13602.
8. Rohrdanz MA, Zheng W, Maggioni M, Clementi C (2011) Determination of reaction coordinates via locally scaled diffusion map. *J Chem Phys* 134:124116.
9. Shaw DE, et al. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* 330:341–346.
10. Zheng W, Rohrdanz MA, Maggioni M, Clementi C (2011) Polymer reversal rate calculated via locally scaled diffusion map. *J Chem Phys* 134:144109.
11. Piana S, Laio A (2008) Advillin folding takes place on a hypersurface of small dimensionality. *Phys Rev Lett* 101:208101.
12. Frenkel D, Smit B (2002) *Understanding Molecular Simulation* (Academic, London).
13. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.
14. Plaku E, Stamati H, Clementi C, Kavraki LE (2007) Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction. *Proteins* 67:897–907.
15. Stamati H, Clementi C, Kavraki LE (2010) Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins* 78:223–235.
16. Singer A, Erban R, Kevrekidis IG, Coifman RR (2009) Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proc Natl Acad Sci USA* 106:16090–16095.
17. Hou C, Wang J, Wu Y, Yi D (2009) Local linear transformation embedding. *Neurocomputing* 72:2368–2378.
18. Cox TF, Cox MAA (1994) *Multidimensional Scaling* (Chapman and Hall, London).
19. Tenenbaum JB, Silva Vd, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–2323.
20. Bronstein AM, Bronstein MM, Kimmel R, Mahmoudi M, Sapiro G (2010) A Gromov–Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *Int J Comput Vis* 89:266–286.
21. Donoho DL, Grimes C (2002) *When Does Isomap Recover the Natural Parameterization of Familes of Articulated Images?* (Stanford University), pp 2002–27.
22. Donoho DL, Grimes C (2003) Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc Natl Acad Sci USA* 100:5591–5596.
23. Coifman RR, et al. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proc Natl Acad Sci USA* 102:7432–7437.
24. Coifman RR, Lafon S (2006) Diffusion maps. *Appl Comput Harmon Anal* 21:5–30.
25. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 15:1373–1396.
26. Mortenson PN, Evans DA, Wales DJ (2002) Energy landscapes of model polyalanines. *J Chem Phys* 117:1363–1376.
27. Tribello GA, Ceriotti M, Parrinello M (2010) A self-learning algorithm for biased molecular dynamics. *Proc Natl Acad Sci USA* 107:17509–17514.
28. Sims GE, Choi IG, Kim SH (2005) Protein conformational space in higher order $\phi - \psi$ maps. *Proc Natl Acad Sci USA* 102:618–621.
29. Bellman R (1961) *Adaptive Control Processes: A Guided Tour* (Princeton Univ Press, Princeton, NJ).
30. Langerholc J (1989) Volumes of diced hyperspheres: resuming the tam-zardecki formula. *Appl Math Comput* 30:1–18.
31. Schölkopf B, Smola A, Muller KR (1999) *Advances in Kernel Methods: Support Vector Learning* (MIT Press, Cambridge, MA), pp 327–352.
32. Rosman G, Bronstein MM, Bronstein AM, Kimmel R (2010) Nonlinear dimensionality reduction by topologically constrained isometric embedding. *Int J Comput Vis* 89:56–58.
33. Sammon JW (1969) A nonlinear mapping for data structure analysis. *IEEE Trans Comput* 18:401–409.
34. Bronstein MM, Bronstein AM, Kimmel R, Yavneh I (2005) A multigrid approach for multi-dimensional scaling. *Proceedings of the Copper Mountain Conference on Multigrid Methods* (Society for Industrial and Applied Mathematics, Philadelphia).
35. Hochbaum DS, Shmoys DB (1985) A best possible heuristic for the *k*-center problem. *Math Oper Res* 10:180–184.
36. de Silva V, Tenenbaum B (2004) Sparse multidimensional scaling using landmark points. *Technical Report, Stanford University.* Available at http://graphics.stanford.edu/courses/cs468-05-winter/Papers/Landmarks/Silva_landmarks5.pdf. Accessed September 10, 2010.
37. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins* 23:566–579.
38. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) Gromacs 4: Algorithms for highly efficient, load-balanced and scalable molecular simulation. *J Chem Theory Comput* 4:435–447.
39. Kollman PA (1996) Advances and continuing challeges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc Chem Res* 29:461–469.
40. Bussi G, Donadio D, Parrinello MJ (2007) Cannonical sampling through velocity rescaling. *J Chem Phys* 126:014101.
41. Bonomi M, et al. (2009) Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Comput Phys Commun* 180:1961–1972.