

# Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction

Payel Das\*, Mark Moll<sup>†‡</sup>, Hernán Stamati<sup>†</sup>, Lydia E. Kavraki<sup>†§¶||</sup>, and Cecilia Clementi\*<sup>¶||</sup>

Departments of \*Chemistry, <sup>†</sup>Computer Science, and <sup>§</sup>Bioengineering, Rice University, Houston, TX 77005; and <sup>¶</sup>Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030

Communicated by James L. Kinsey, Rice University, Houston, TX, May 1, 2006 (received for review February 6, 2006)

**The definition of reaction coordinates for the characterization of a protein-folding reaction has long been a controversial issue, even for the “simple” case in which one single free-energy barrier separates the folded and unfolded ensemble. We propose a general approach to this problem to obtain a few collective coordinates by using nonlinear dimensionality reduction. We validate the usefulness of this method by characterizing the folding landscape associated with a coarse-grained protein model of src homology 3 as sampled by molecular dynamics simulations. The folding free-energy landscape projected on the few relevant coordinates emerging from the dimensionality reduction can correctly identify the transition-state ensemble of the reaction. The first embedding dimension efficiently captures the evolution of the folding process along the main folding route. These results clearly show that the proposed method can efficiently find a low-dimensional representation of a complex process such as protein folding.**

reaction coordinate | transition state | manifold | embedding | ISOMAP

The folding of a protein to its functional (native) state can be viewed as a chemical reaction, for which the ensemble of unfolded configurations constitutes the reactant and the native state is the product. Generally, the characterization of chemical reactions requires the location of the reactants, products, and transition states on a free-energy surface. Simple models (so-called “reaction profiles” or “reaction coordinate diagrams”) are often used to describe the change in free energy as a function of the progress of the reaction from reactant to product. Clearly, a reaction profile is meaningful if the process of interest can be described in terms of one or a few collective coordinates. For instance, in a dissociation reaction in which a diatomic molecule splits into the constituent atoms, the distance between the two atoms provides a natural choice for the reaction coordinate, and the progress of the reaction can be quantitatively characterized in terms of this coordinate. For more complex reactions, the definition of a set of reaction coordinates is a nontrivial task. Because of the high dimensionality of a protein configurational space, this problem is particularly challenging and is a source of significant debate in protein-folding studies.

Here, we present an approach to the definition of reaction coordinates for the theoretical characterization of a protein-folding free-energy landscape based on the idea of nonlinear dimensionality reduction. Modern dimensionality reduction techniques allow us to define a fast and efficient procedure that uses a significant sample of configurations along the folding to extract the most relevant global coordinates that can effectively describe the process. We prove the efficiency and robustness of this method by applying it to study the folding of the src homology 3 (SH3) domain, as obtained from simulation with a coarse-grained protein model (1).

The possibility of using only a few global coordinates to characterize the mechanism through which a protein “organizes” its constituent atoms into a compact functional structure has important practical implications. It is worth mentioning, for example, that a quantitative comparison between simulation and experiment in

protein folding often relies on the assumption that it is possible to identify the folding transition state and/or intermediate state ensembles from the analysis of the simulated folding (and/or unfolding) trajectories. However, the definition of these ensembles is generally based on the choice of the reaction coordinates (2–4). Alternative definitions of reaction coordinates have been discussed in the literature (2, 5–8), as have different methods for the identification of a set of transition-state structures (3, 4, 9). Most of the discussion revolves around the validity of empirical reaction coordinates that are commonly used in this endeavor. Commonly used empirical reaction coordinates in folding studies are defined to condense in a parameter the information on the degree of similarity with the native structure. Examples of such coordinates include the fraction of native contacts formed,  $Q$  (2, 5, 10–12), the average shortest path length,  $\langle L \rangle$  (13, 14), the radius of gyration,  $R_g$  (12), or the partial contact order  $pCO$  (13, 15). The theoretical justification for the use of these structural reaction coordinates relies on the fact that, generally, proteins are minimally frustrated systems and their folding mechanism can be described as a diffusion process in a funnel-like energy landscape where the potential depth is strongly correlated with the degree of nativeness (16–19). This argument is not sufficient to ensure a perfect *a priori* correspondence between a given ensemble of structures experimentally detected [as, for instance, the transition-state ensemble, experimentally characterized by  $\Phi$ -value analysis (20, 21)] and the corresponding ensembles obtained on a low-dimensional landscape defined through these reaction coordinates.

The parameter  $P_{\text{fold}}$ , defined as the probability of a protein structure to reach the folded state before the unfolded state, has been proposed as an ideal reaction coordinate for protein-folding studies (2, 8, 14, 22, 23). However, the calculation of  $P_{\text{fold}}$  is computationally so expensive that it becomes unfeasible for many systems of interest. The definition of new strategies to estimate the intrinsic dimensionality of a folding reaction and the definition of the reaction coordinates themselves remain paramount issues in folding studies (5). The approach presented here addresses both of these questions.

## What Is the Intrinsic Dimensionality of a Folding Landscape?

A protein conformation is usually described by the Cartesian coordinates of its constituent atoms; a protein structure with  $N$  atoms is therefore completely specified by  $3N$  parameters. However, these parameters are not independent of each other. Clearly, the constraints of maintaining intact the covalent bonds and angles and other steric factors effectively reduce the degrees of freedom

Conflict of interest statement: No conflicts declared.

Abbreviations: SH3, src homology 3; PCA, principal component analysis; SciMAP, scalable ISOMAP.

<sup>¶</sup>Present address: Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292.

<sup>||</sup>To whom correspondence may be addressed. E-mail: kavraki@rice.edu or cecilia@rice.edu.

© 2006 by The National Academy of Sciences of the USA

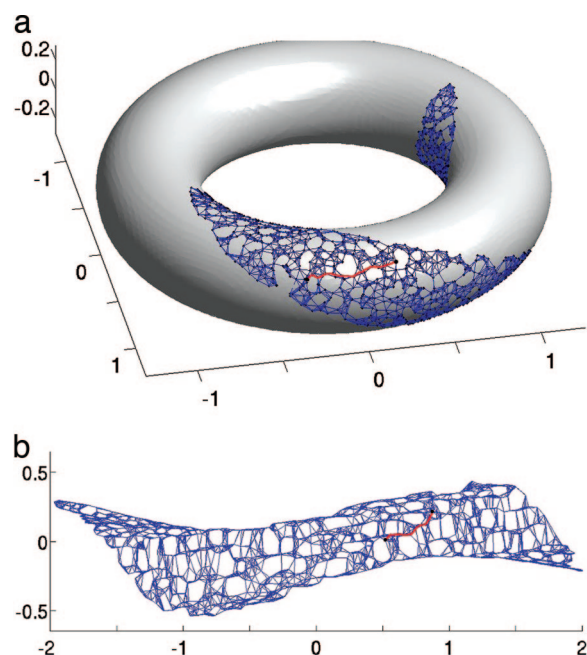
of a protein molecule. In addition, the high cooperativity of the folding process strongly suggests that the motion of different parts of the protein is correlated along the productive folding route(s), further reducing the effective dimensionality of the configurational space. These considerations lead us to assume that most of the relevant conformations visited by a protein throughout the folding process lie on a low-dimensional manifold embedded in the much higher-dimensional space described by the Cartesian coordinates.

In folding/unfolding simulations, molecular dynamics trajectories provide a sampling of configurations populating the embedded manifold that we wish to characterize. Given a sample of protein configurations along the folding process, we address the problem of finding a low-dimensional embedding such that the shape of the underlying manifold is preserved. The final goal is to rigorously define a low-dimensional, free-energy landscape that could be used to quantitatively characterize a folding simulation. The density of states populated on the manifold needs to be preserved as well so that free energy can be estimated directly from the low-dimensional embedding. In practice, the main question underlying the definition of this embedded folding landscape is whether a base set of coordinates exists in which very few show considerable variation and all of the others remain almost constant during the considered reaction; mathematically, this is a problem of dimensionality reduction. Similar problems are common in a number of disparate fields. For instance, dimensionality reduction plays an important role in image analysis and recognition, where the essential information distributed over a large number of pixels needs to be captured by few global parameters that can be quantitatively and meaningfully compared (24–27).

The definition of an embedded, folding, free-energy landscape by dimensionality-reduction techniques can reduce the systematic error associated with the choice of empirical reaction coordinates in the calculations of ensemble averages on particular regions of the landscape (such as, for instance, transition-state ensembles). An important feature of dimensionality reduction is that, usually, the quality of the embedding can be expressed as a function of the number of dimensions chosen, which allows one to estimate *a priori* the error associated with a set of reaction coordinates. Ideally, one could automatically compute an embedding that preserves, say, 99% of the features of the original data. Unlike empirical reaction coordinates, the dimensions of an embedding are completely uncorrelated so that the number of dimensions of an accurate embedding is the same as the number of dimensions of the submanifold. Minima and saddle points of a specified function of the embedding coordinates (such as free energy) can be automatically identified in an embedding, which is important if more than two or three dimensions are needed to capture the features of the original data, in which case it is not possible to visually identify the folded and unfolded minima or transition paths between them.

### Dimensionality Reduction of Folding Simulations

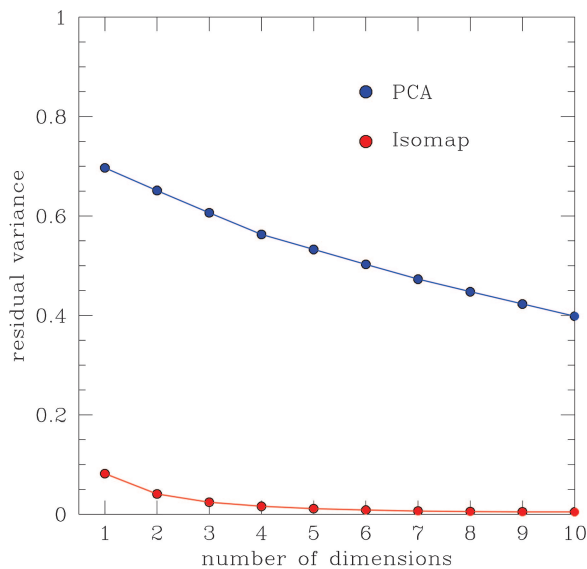
The problem addressed by dimensionality-reduction techniques is to find the best  $d$ -dimensional embedding for  $N$  objects in an  $n$ -dimensional space. Ideally, the embedding is much more compact than the original representation, and dependencies between dimensions are removed. Dimensionality-reduction techniques fall broadly into two categories: linear and nonlinear. Principal component analysis (PCA) (28) is probably the best known (and widely used) linear technique. Essentially, PCA computes a hyperplane that passes through the data points as best as possible in a least-squares sense. The principal components are the tangent vectors that describe this hyperplane. These vectors are ordered by the amount of variance they exhibit on the data. So the first principal component corresponds to the best possible projection onto a line, the first two correspond to the best possible projection onto a plane, and so on. Clearly, if the manifold of interest is inherently nonlinear, the low-dimensional embedding obtained by means of PCA is severely distorted. PCA is commonly used in the



**Fig. 1.** A simple case of embedding. (a) The data points (blue dots) are shown “live” on the two-dimensional surface of the torus, although they are embedded on a three-dimensional space. The application of the ISOMAP algorithm (42) to this set of data defines two independent coordinates on which all points are mapped. (b) The resulting two-dimensional embedded space. These two embedding coordinates cannot be reduced to a linear combination of the original coordinates. The network of neighboring points (used to compute the geodesic distances) is shown both in the original (a) and embedded (b) space.

analysis of near-equilibrium fluctuations sampled by molecular dynamics simulations (29–36), because one can usually assume that the manifold of interest can be reasonably approximated by its tangent hyperplane around an equilibrium point. However, the extent of conformational changes involved in a folding process prohibits any *a priori* linearization of the manifold, and nonlinear techniques need to be used. The fact that empirical reaction coordinates routinely used in protein folding studies cannot be reduced to a linear combination of the Cartesian coordinates underscores the inadequacy of linear dimensionality reduction techniques to characterize a folding landscape.

**The ISOMAP Algorithm for Nonlinear Dimensionality Reduction.** Although several nonlinear dimensionality reduction techniques have been proposed [especially in the context of image analysis (37), speech recognition (38), and climate data analysis (40, 41)], the development of new methods is still an active area of research. The technique we used here for the characterization of folding landscapes is based on the recently proposed ISOMAP algorithm (42). The basic idea of ISOMAP is to define a low-dimensional embedding that preserves as best as possible geodesic distances between all pairs of data points in the sample under consideration (42). The geodesic distance between a pair of points that lie on a surface (or, more generally, a manifold) is defined as the length of the shortest path between them, when the path is confined to lie on the surface. In the context of the ISOMAP algorithm, we assume that a low-dimensional embedded manifold exists, but we do not have an explicit mathematical formulation for the manifold; rather, we have a discrete set of points sampled from this manifold. Therefore, we can only approximate the geodesic distance on the manifold between a general pair of points, say  $x$  and  $y$ . The idea of ISOMAP is to estimate the shortest possible path from  $x$  to  $y$  by “hopping” through neighboring points and adding up the short distances between them. Fig. 1 illustrates this idea on a simple case of an



**Fig. 2.** Residual variance as a function of dimensions considered in the embedding, as obtained when the PCA (blue dots) or our nonlinear dimensionality reduction (red dots) [based on the idea of the ISOMAP algorithm (42)] is used to characterize the space sampled in extensive folding/unfolding simulations of a SH3 protein model.

embedding. The data points shown in Fig. 1*a* lie on a two-dimensional torus embedded in a three-dimensional space. The application of the ISOMAP algorithm to this set of data produces the two-dimensional manifold shown in Fig. 1*b*, where the network of neighboring points is also shown. A pair of points has been chosen to show the approximation of the geodesic distance between them. The red curve connecting this pair of points represents the shortest “hopping” distance within the network of neighbors, both in the original three-dimensional space (Fig. 1*a*) and on the resulting two-dimensional embedded surface (Fig. 1*b*).

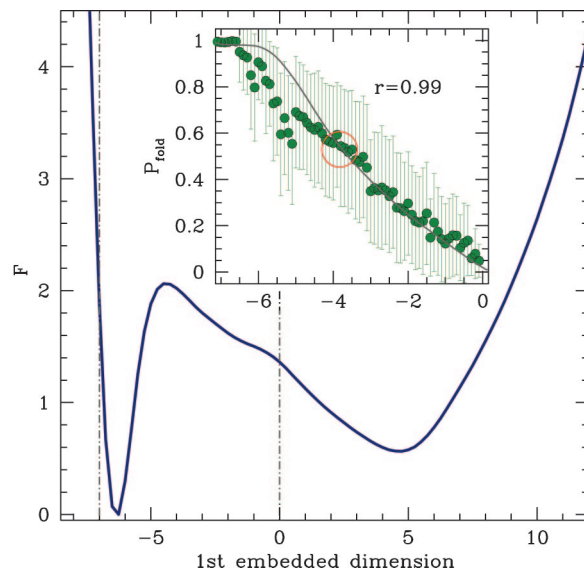
This approximation for the geodesic distance holds provided that the data represents a good sampling of the embedded manifold (that is, the sampling of the manifold is sufficiently dense). If a poor data sampling is used, then connecting a point to its nearest neighbors may result in taking “shortcuts” across the surface, artificially shortening distances.

In addition, this approximation for the geodesic distance relies on the definition of neighboring points. Each point should be connected to a number of neighboring points that is not too small (to ensure the connectivity of the manifold) or too large (to minimize the occurrence of shortcuts), that is, a robust definition of “neighboring points” is required. Neighboring points can be defined around each point either by choosing a fixed number of neighbors (that is, the  $k$  closest points according to a distance measure, as for instance the least rms deviation), or by assigning a cut-off distance and considering all of the points residing within it.

In the work presented here, for each point we have defined its nearest neighbors as the  $k$  points with the smallest rms deviation to it, with  $k = 9$ . The robustness of this definition is discussed in detail in the supporting information, which is published on the PNAS web site, where several tests on the validity of the approximations used are presented.

In practice, the ISOMAP algorithm consists of the following three steps:

1. First, the nearest neighboring points are computed for each point. The nearest neighbor network (or graph) is represented by the list of neighbors for each point, with their associated distances.



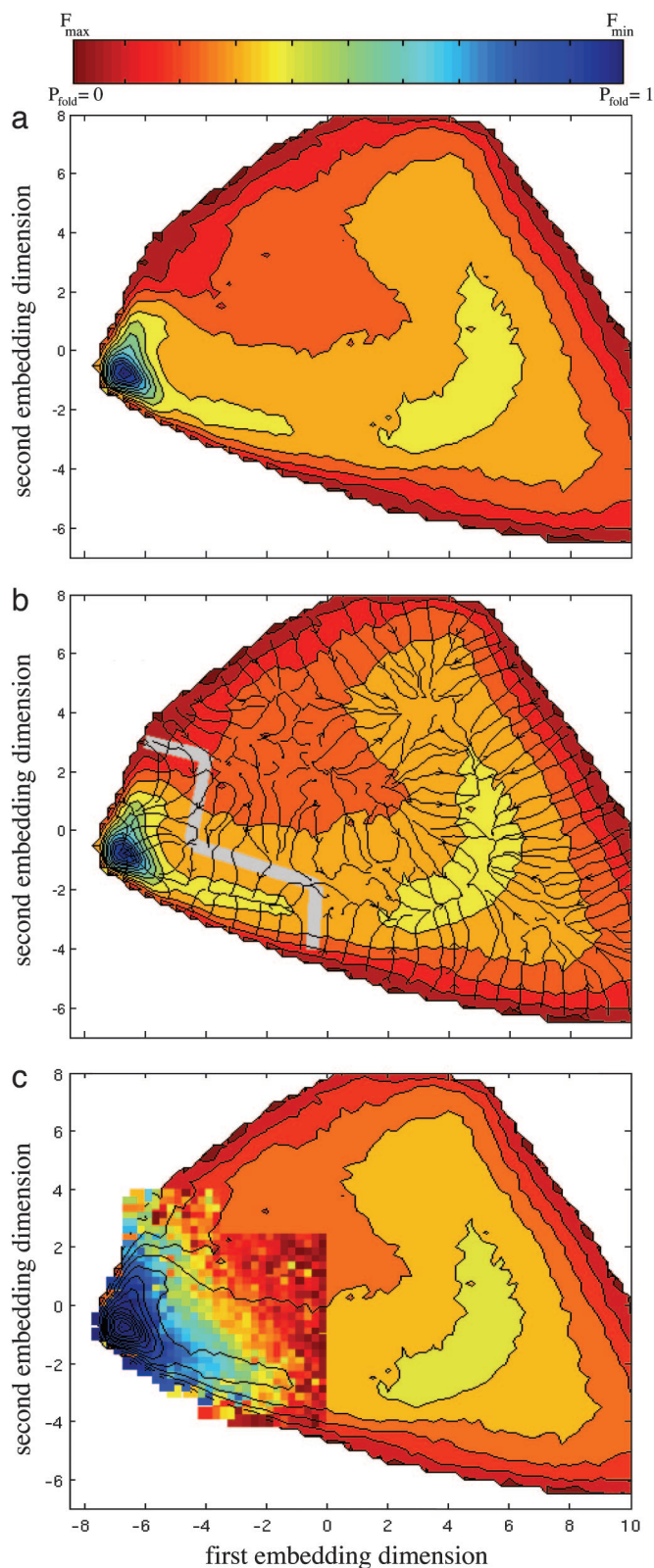
**Fig. 3.** A one-dimensional free-energy profile  $F(x_1)$  as a function of the first embedded dimension,  $x_1$ , as extracted from the dimensionality reduction procedure. One single barrier is detected around the value  $x_1 \approx -4$ . (Inset) The average value of  $P_{\text{fold}}$  associated to each small interval  $x_1 \pm \Delta x_1 \in (-7, 0)$  is plotted as a function of  $x_1$ . The error bar corresponds to the variance of  $P_{\text{fold}}$  for a given value of  $x_1$ . The continuous gray line is the theoretical folding probability  $P_t(x_1)$  associated with the one-dimensional free-energy curve  $F(x_1)$  (see text for detail). The red circles identify the  $P_{\text{fold}}$  and  $P_t$  values corresponding to the top of the free-energy barrier (that is, around  $x_1 \approx -4$ ).

2. The second step involves computing the shortest paths between every pair of points. As explained above, the distances along these shortest paths within the neighbor network are added up to approximate the geodesic distance between the points.
3. Finally, a low-dimensional representation of the points (i.e., protein configurations for the case considered here) is found such that the interpoint Euclidean distances best preserve their geodesic distances.

**Scalable ISOMAP (ScIMAP) and Its Application to Large Folding Simulations.** Although the idea is very straightforward, the basic ISOMAP algorithm as described above suffers from performance bottlenecks when used on large data sets and cannot be directly applied to the study of folding reactions. The major bottleneck when working with molecular data resides in the computation of all neighboring points, for every point, especially if the distance measure is computationally expensive. There are also problems in the computation of the shortest paths for all pairs and in the computation of the final coordinates for the embedding, because these require working with huge matrices (for which, for example, eigenvalues are needed). These bottlenecks render computationally impractical the application of the ISOMAP algorithm to study protein simulations, where the number of conformations sampled is generally  $N \gg 100,000$ , and the protein size yields a very expensive distance measure. We use here the basic idea of ISOMAP as a starting point to define a procedure that is computationally scalable to very large data sets. We introduce a number of nontrivial modifications to suit the basic ISOMAP algorithm to analyze large data sets, such as protein conformations generated in folding/unfolding trajectories. We call this modified procedure ScIMAP.

First of all, the computational bottlenecks present in ISOMAP can be strongly reduced by using landmark points, as has been proposed in recent literature (43, 44). We designate  $n_L$  data points (i.e., protein configurations) to be landmarks, where  $n_L \ll N$ . Rather than computing all-pairs shortest paths, we just compute the shortest path from each landmark to every other point. The use of





**Fig. 4.** A two-dimensional free-energy profile  $F(x_1, x_2)$  as a function of the first and second embedded dimensions ( $x_1$  and  $x_2$ , respectively) as extracted from ScIMAP. (a) The free energy is shown as a contour plot. Each contour represents an increase of free energy of  $1 k_B T$  and colored according to the color key (colors from red to blue indicate progressively decreasing free energy). (b) The free-energy gradient field is superimposed on the free-energy contour plot. The thick gray line approximately locates the separatrix between the folded and unfolded state basins, where gradient fluxes leading to

landmarks reduces the shortest-path computation time by a factor  $n_L/N$ . The intuition for landmark-based ISOMAP is that if the manifold is low-dimensional, each point can be located by considering its distance to only a small number of landmarks. In theory, if  $n_L \geq d + 1$  and the landmarks are in general position, then there are enough landmarks to locate each point. If the landmarks are chosen randomly, then  $n_L$  needs to be sufficiently larger than  $d$  to guarantee stability (see the supporting information for details).

Although it is more space- and time-efficient than the basic version of the algorithm, landmark-based ISOMAP is still not practical to compute low-dimensional embeddings of large molecular trajectories (typically  $>100,000$  conformations). To obtain a good coverage of the conformational manifold (which is, in turn, essential to ensure the validity of the geodesic approximation and to obtain accurate free energy estimates), it is necessary to compute embeddings of very large trajectories.

To make the problem computationally tractable, we can make the following observation. We expect that low free-energy areas on a folding landscape will have a very high sampling density, resulting in a multitude of conformations. In particular, for a canonical sampling at the temperature close to the folding temperature, a significant fraction of the sampled points is expected to fall near the protein's folded state, typically with little spatial variability. These redundant conformations, although crucial to estimate the density of states (thus to compute all thermodynamic averages), are not necessary to infer the overall shape of the manifold we are trying to recover and can be initially ignored to obtain the embedding geometry. Once the low-dimensional embedding has been found, the discarded conformations can be reinserted into place by computing their low-dimensional coordinates as a combination of the coordinates of their nearest neighbors, which is a less demanding process than solving for the global manifold shape of the original data set. The reinserted points now restore the original density of states of the trajectory around the folded state and allow for the computation of thermodynamic quantities.

The reinsertion of conformations into a precomputed low-dimensional embedding can also be used to further enrich the resolution of the landscape, for instance by adding configurations sampled at different temperatures [that can be combined in free energy calculations (45, 46)]. Moreover, the reinsertion of configurations provides a way to test the robustness of the procedure to extract the low-dimensional embedding. If some of the configurations to be reinserted are in regions where their closest neighbors are in fact far apart, the approximation used is not valid. In the application presented below, all of the configurations initially filtered out could be reinserted without experiencing such a problem.

Finally, after the problem size has been reduced as outlined above, the algorithm can be run in parallel on a cluster of computers. In particular, we adopted parallel, iterative procedures (47, 48) that efficiently compute the top  $m \ll N$  eigenvalues of the distance matrix as required by the embedding method. The details are presented in the supporting information.

### Results: Folding Landscape of SH3 as a Low-Dimensional Embedded Manifold

We tested ScIMAP outlined above by applying it to characterize the protein-folding landscape obtained from simulation of a coarse-grained model of an SH3 domain. The basic ideas of the model are detailed in a recent publication (1), where a comparison of the

opposite minima meet. (c) The results from the  $P_{\text{fold}}$  analysis are superimposed on the two-dimensional embedded landscape. The average value of  $P_{\text{fold}}$  at a given  $(x_1, x_2)$  position on the landscape is color-coded according to the color key. Colors ranging from red to blue indicate values of  $P_{\text{fold}}$  increasing from 0 to 1. The comparison of b and c reveals that the region with  $P_{\text{fold}} \approx 0.5$  is fully consistent with the separatrix region.

results with experimental data also are presented. The purpose of the application presented here is not to further validate this coarse-grained protein model but rather to show how nonlinear dimensionality reduction can be used to estimate the intrinsic dimensionality of the configurational space explored in folding simulations, and to “naturally” define a set of orthogonal reaction coordinates associated to the relevant dimensions.

The performance of a dimensionality-reduction procedure can be estimated by monitoring the residual variance  $\sigma_r(d, n)$  as a function of the number of dimensions considered for the embedded manifold,  $d$ , and the dimensionality,  $n$ , of the original space. Following the definition used in ref. 42, the residual variance  $\sigma_r(d, n)$  can be computed as  $\sigma_r(d, n) = 1 - R^2(\hat{D}_d, D_n)$ , where  $R(\hat{D}_d, D_n)$  is the correlation coefficient taken over all of the entries of matrices  $\hat{D}_d$  and  $D_n$ . The matrix  $\hat{D}_d$  contains all of the pairwise distances as obtained on the  $d$ -dimensional embedding, whereas the matrix  $D_n$  stores the corresponding geodesic distances as computed in the original ( $n$ -dimensional) space. In the case of the SH3 folding simulations that we are considering here, the original space has dimensionality  $n = 3 \times 57 = 171$ . The function  $\sigma_r(d, n)$  monotonically decreases as the number of embedding dimensions,  $d$ , increases, up to the limit value  $\sigma_r(n, n) = 0$  when  $d = n$ . By definition, the maximum possible value of the residual variance is  $\sigma_r(d, n) = 1$ , if the distances computed on the  $d$ -dimensional embedded manifold are completely uncorrelated with the geodesic distances computed in the original space. If  $\sigma_r(d, n)$  drops close to zero for small values of  $d \ll n$ , then the space of interest can be well approximated by considering only  $d$  embedding dimensions.

Fig. 2 shows that the embedded landscape associated to the folding simulations of the coarse-grained model of SH3 has extremely low residual variance (blue points), even when only one dimension is considered. Namely,  $\sigma_r(1, n) \approx 0.08$ ,  $\sigma_r(2, n) \approx 0.04$ , and  $\sigma_r(3, n) \approx 0.02$ . These values give an estimate of the distortion introduced when one, two, or three embedding dimensions are used as reaction coordinates to describe the folding landscape. The small magnitude of these values is evident when they are compared with the corresponding residual variance obtained when PCA is used on the same data (red points in Fig. 2). These results support the idea that the folding landscape of SH3 can be essentially described by one reaction coordinate, in agreement with results from previous work (1, 49).

Free-energy surfaces can be computed as a function of the embedding coordinates. Fig. 3 shows the free-energy profile obtained when only the first dimension is used as a reaction coordinate for the folding process. These results are obtained for a temperature very close to the folding temperature  $T_f$ . One main barrier separates the free-energy minima corresponding to the unfolded and folded states, as expected in a two-state folding process. On this reaction profile, the transition state can be defined as the ensemble of states with a value of the first embedding coordinate corresponding to the top of the free energy barrier. For a two-state folding process, the parameter  $P_{\text{fold}}$  provides a stringent test for the identification of the transition-state ensemble (2, 8, 10, 14, 22, 23). Each individual configuration around the free-energy barrier (namely, each conformation with a value of the first embedding coordinate  $x_1$  in the range  $-7 < x_1 < 0$ ) is labeled with a value of  $P_{\text{fold}}$  by means of a set of 100 ancillary simulations starting from it. For each small interval  $x_1 \pm \Delta x_1$  an average value of  $P_{\text{fold}}$  is computed over all conformations with a corresponding  $x_1$  within that interval, whereas the variance is reported as an error bar. Fig. 3 *Inset* shows that the range of values on the first embedding coordinate  $x_1 \approx -4$  corresponding to the location of the free-energy barrier has an associated value of  $P_{\text{fold}} \approx 0.5$ . The red circle in Fig. 3 *Inset* identifies the  $P_{\text{fold}}$  values corresponding to the top of the free-energy barrier (that is, around  $x_1 \approx -4$ ). Remarkably, the transition state identified by means of the one-dimensional free energy profile  $F(x_1)$  as a function of the first embedding coordinate,  $x_1$ , is in full agreement with the ensemble obtained by a thorough  $P_{\text{fold}}$  analysis: The top of

the free energy barrier corresponds to  $P_{\text{fold}} \approx 0.5$ . The theoretical folding probability (39)

$$P_t(x_1) = \frac{\int_{x_1}^{x_U} \exp(F(y)/k_B T) dy}{\int_{x_N}^{x_U} \exp(F(y)/k_B T) dy} \quad [1]$$

associated to the one-dimensional free energy  $F(x_1)$  is also shown in Fig. 3 *Inset*. The folding probability  $P_t$  is in agreement with the calculated  $P_{\text{fold}}$  values on most of the interval considered, particularly at the transition state. Deviations between the average value of the calculated  $P_{\text{fold}}$  and the theoretical folding probability  $P_t(x_1)$  are observed around the folded state ( $x_1 \approx -6$ ) and can be explained in terms of the variation of free energy along the second embedding dimension in this region (as it appears from Fig. 4, discussed below).

It is worth noting that for the protein model considered here, the  $P_{\text{fold}}$  analysis required >12,000 central processing unit (CPU) hours on an 2.2-GHz Intel Xeon processor and was performed for a small subset of configurations,\*\* and the embedding procedure was completed in <500 CPU hours (<24 CPU hours running on 20 processors) and provides information on the whole configurational space.

Not surprisingly, the transition-state ensemble from the one-dimensional embedded manifold of the SH3 model also is in good agreement with what can be obtained by using the parameter  $Q$  as an empirical reaction coordinate (data not shown). Previous studies have shown that  $Q$  is a robust reaction coordinate for some two-state folding proteins (5, 10, 49), SH3 being one of them, which may not be the case in general, however, particularly for more complex folding reactions for which more than one reaction coordinate is needed.

Additional information on the folding process is obtained when the first two embedding dimensions are considered reaction coordinates in the free-energy calculation. Fig. 4 presents the two-dimensional embedded free-energy landscape as a function of the first two embedding dimensions.

Fig. 4a shows a contour plot of the free energy. Again, as expected for a two-state folding protein, two distinct free-energy minima appear: one that is more localized corresponding to the folded state and one with a larger basin corresponding to the unfolded state. The free-energy gradient field is superimposed to the free-energy contour plot in Fig. 4b. The transition-state ensemble on this two-dimensional landscape can be defined by considering the “Continental Divide,” i.e., the separatrix between the basin corresponding to the folded and unfolded states. In practice, a point on the landscape is considered in the basin of a given minimum if the gradient flux starting from that point leads to the minimum. The transition-state ensemble is then defined as all regions on the landscape where gradient fluxes leading to opposite minima meet. The transition-state region so defined is depicted in Fig. 4b.

It is clear from Fig. 4 that the most populated folding route (defined by the minimum free-energy path on this landscape) closely follows the first embedding dimension. However, deviations from the main folding route are probable, because a nonnegligible amount of structures lie outside the minimum free-energy path ( $\approx 15\%$  of structures lie within the light orange free-energy level on Fig. 4).

It is important to clarify that the existence of a main folding route does not mean that the folding mechanism follows a deterministic pathway where one single protein structure evolves into the next one along the pathway. Each point along this route on the low-dimensional landscape represents a large ensemble of structures

\*\*The  $P_{\text{fold}}$  parameter was computed for  $\approx 8,000$  protein configurations. The total number of configurations used in the definition of the embedded free-energy landscape is 1,818,000.



that are not necessarily similar to each other. The fact that a single parameter (i.e., the first embedding dimension, in this case) captures the evolution of the folding process simply means that it is possible to define a “macroscopic” quantity condensing into a single number the common features of the ensemble of structures populated at a given stage of the folding process. The first embedding coordinate describes the evolution of this parameter from the unfolded to the folded ensembles.

Fig. 4c presents the results from the  $P_{\text{fold}}$  analysis superimposed on the two-dimensional embedded landscape. The comparison of Fig. 4b and c reveals that the region with  $P_{\text{fold}} \approx 0.5$  matches the separatrix region identified by the diverging gradient fluxes. The variance of  $P_{\text{fold}}$  measured in each two-dimensional interval ( $x_1 \pm \Delta x_1, x_2 \pm \Delta x_2$ ) is  $\delta P_{\text{fold}} \approx 0.12$ , which is significantly lower than the variance  $\delta P_{\text{fold}} \approx 0.2$  observed in the one-dimensional case (see Fig. 3). The larger uncertainty obtained when only one embedding dimension is considered accounts for the fluctuations observed along the second embedding coordinate.

## Conclusions

We have proposed a general procedure to obtain a low-dimensional free energy landscape associated with a simulated protein-folding reaction. By using ScIMAP [based on ISOMAP (42)], an embedded folding manifold is extracted from a large set ( $\approx 2,000,000$ ) of protein conformations sampled throughout extensive folding/unfolding simulations of a coarse-grained model of SH3. The first few embedding coordinates provide a set of reaction coordinates independent of each other. The quality of the embedding can be expressed as a function of the number of dimensions considered. This feature provides an estimate of the error introduced when the first few  $d$  embedding dimensions are used as reaction coordinates to describe the simulated folding process. As a consequence, it is possible to estimate the intrinsic dimensionality of a simulated folding process.

The application of ScIMAP to the folding of a coarse-grained protein model of an SH3 domain reveals that its folding landscape

is essentially one-dimensional. The first embedding dimension captures the evolution of the folding process along the main folding route. However, additional features emerge when two or three dimensions are considered. For instance, the two-dimensional free energy landscape as a function of the first two embedding dimensions reveals deviations around the main folding route, populated with a lower probability. The simulated folding reaction considered in this paper is known to be a two-state folding process, where no intermediate states are significantly populated. For such kind of processes, the calculation of the transition probability (or  $P_{\text{fold}}$  parameter) provides a strict *a posteriori* test for the “goodness” of a reaction coordinate on the identification of the transition-state ensemble. Remarkably, a thorough  $P_{\text{fold}}$  analysis confirms that protein configurations in the transition-state region as identified on the embedded free-energy landscape have  $P_{\text{fold}} \approx 0.5$ . Moreover, fluctuations around this average value of  $P_{\text{fold}}$  significantly decrease when the transition-state region is identified on the two dimensional free-energy landscape (defined by means of the first two embedding coordinates), with respect to a one-dimensional, free-energy landscape (where only the first embedding coordinate is used). These results validate the use of the first few embedding dimensions as optimal reaction coordinates to characterize the protein folding reaction, at least for the protein model used here.

We thank Alessandro Mossa for contributions to initial stages of this project; Peter Wolynes, Attila Stzabo, Yannis Kevrekidis, Tolya Kolomeisky, and Michele Vendruscolo for insightful comments; and members of the L.E.K. and C.C. groups for stimulating discussions. This work was supported by National Science Foundation Career Grant CHE-0349303 (to C.C.), National Science Foundation Grants ITR-0205671 (to L.E.K.) and CCF-0523908 and CNS-0454333 (to L.E.K. and C.C.), Advanced Technology Program Grant 003604-0010-2003, Robert A. Welch Foundation Norman Hackermann Young Investigator Award and Grant C-1570 (to C.C.), and the Sloan Foundation (L.E.K.). The Rice Terascale Cluster used for the calculations is supported by National Science Foundation Grant EIA-0216467 and by Intel and Hewlett Packard.

- Das, P., Matysiak, S. & Clementi, C. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 10141–10146.
- Du, R., Pande, V., Grosberg, A., Tanaka, T. & Shakhnovich, E. (1998) *J. Chem. Phys.* **108**, 334–350.
- Hummer, G. (2004) *J. Chem. Phys.* **120**, 516–523.
- Best, R. & Hummer, G. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 6732–6737.
- Cho, S., Levy, Y. & Wolynes, P. (2006) *Proc. Natl. Acad. Sci. USA* **103**, 586–591.
- Baumketner, A., Shea, J.-E. & Hiwataru, Y. (2004) *J. Chem. Phys.* **121**, 1114–1120.
- Ma, A. & Dinner, A. (2005) *J. Phys. B* **109**, 6769–6779.
- Rhee, Y. & Pande, V. (2005) *J. Phys. B* **109**, 6780–6786.
- Klimov, D. & Thirumalai, D. (2004) *Chem. Phys.* **307**, 251–258.
- Clementi, C., Jennings, P. & Onuchic, J. (2001) *J. Mol. Biol.* **311**, 879–890.
- Clementi, C. & Plotkin, S. (2004) *Protein Sci.* **13**, 1750–1766.
- Abkevich, V., Gutin, A. & Shakhnovich, E. (1998) *Fold. Des.* **3**, 183–194.
- Chavez, L., Onuchic, J. & Clementi, C. (2004) *J. Am. Chem. Soc.* **126**, 8426–8432.
- Dokholyan, N., Li, L., Ding, F. & Shakhnovich, E. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 8637–8641.
- Das, P., Wilson, C., Fossati, G., Wittung-Stafshede, P., Matthews, K. & Clementi, C. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 14569–14574.
- Bryngelson, J. & Wolynes, P. (1989) *J. Phys. Chem.* **93**, 6902–6915.
- Nymeyer, H., Garcia, A. & Onuchic, J. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5921–5928.
- Onuchic, J., Luthey-Schulten, Z. & Wolynes, P. (1997) *Annu. Rev. Phys. Chem.* **48**, 545–600.
- Shea, J.-E. & Brooks, C., III (2001) *Annu. Rev. Phys. Chem.* **52**, 499–535.
- Ferst, A. & Sato, S. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 7976–7981.
- Fersht, A., Leatherbarrow, R. & Wells, T. (1986) *Nature* **322**, 284–286.
- Klimov, D. K. & Thirumalai, D. (2001) *Proteins: Struct. Funct. Genet.* **43**, 465–475.
- Ding, F., Guo, W., Dokholyan, N. V., Shakhnovich, E. I. & Shea, J.-E. (2005) *J. Mol. Biol.* **350**, 1035–1050.
- Kirby, M. & Sirovich, L. (1990) *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 103–108.
- Turk, M. & Pentland, A. (1991) *J. Cognit. Neurosci.* **3**, 71–86.
- Benito, M. & Pena, D. (2004) *Lect. Notes Comput. Sci.* **3177**, 326–332.
- Cho, E., Kim, D. & Lee, S. (2003) *Lect. Notes Comput. Sci.* **2688**, 946–954.
- Jolliffe, I. (1986) *Principal Components Analysis* (Springer, New York).
- Levitt, M. (1991) *J. Mol. Biol.* **220**, 1–4.
- Garcia, A. (1992) *Phys. Rev. Lett.* **68**, 2696–2699.
- Hayward, S. & Go, N. (1995) *Ann. Rev. Phys. Chem.* **46**, 223–250.
- Hayward, S., Kitao, A. & Go, N. (1994) *Protein Sci.* **3**, 936–943.
- Hayward, S., Kitao, A. & Berendsen, H. (1997) *Proteins: Struct. Funct. Genet.* **27**, 425–437.
- Nolde, S., Arseniev, A., Orekhov, V. & Billeter, M. (2002) *Proteins: Struct. Funct. Genet.* **46**, 250–258.
- Levy, Y. & Caflisch, A. (2003) *J. Phys. B* **107**, 3068–3079.
- Teodoro, M., Phillips, Jr., G. & Kavraki, L. E. (2003) *J. Comp. Biol.* **10**, 617–634.
- Weinberger, K. Q. & Saul, L. K. (2004) *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition* (IEEE, Washington, DC), pp. 988–995.
- Belkin, M. & Niyogi, P. (2003) *Neural Comp.* **15**, 1373–1396.
- Rhee, Y. & Pande, V. (2006) *Chem. Phys.* **323**, 66–77.
- Monahan, A. H. (2000) *J. Climate* **13**, 821–835.
- Gamez, A. Z., Zhou, C. S., Timmermann, A. & Kurths, J. (2004) *Nonlinear Processes Geophys.* **11**, 393–398.
- Tenenbaum, J., de Silva, V. & Langford, J. (2000) *Science* **290**, 2319–2323.
- de Silva, V. & Tenenbaum, J. (2002) in *Advances in Neural Information Processing Systems 15*, eds. Becker, S., Thrun, S. & Obermayer, K. (MIT Press, Cambridge, MA), pp. 705–712.
- de Silva, V. & Tenenbaum, J. (2004) *Sparse Multidimensional Scaling Using Landmark Points* (Stanford Univ., Stanford, CA).
- Ferrenberg, A. & Swendsen, R. (1989) *Phys. Rev. Lett.* **63**, 1185–1198.
- Ferrenberg, A. & Swendsen, R. (1988) *Phys. Rev. Lett.* **61**, 2635–2638.
- Arnoldi, W. (1951) *Q. Applied Mathematics* **9**, 17–29.
- Saad, Y. (2003) *Iterative Methods for Sparse Linear Systems*. (Soc. Ind. Appl. Math., Philadelphia), 2nd Ed.
- Clementi, C., Nymeyer, H. & Onuchic, J. (2000) *J. Mol. Biol.* **298**, 937–953.