Appeared in PNAS 103(26):9885-9890, 2006

Low-Dimensional Free Energy Landscapes of Protein Folding Reactions by Nonlinear Dimensionality Reduction

Payel Das[1], Mark Moll[2*], Hernan Stamati[2], Lydia E. Kavraki[2,3,4,†], Cecilia Clementi[1,4,‡]

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

[*] Current address: Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292
[†] Electronic mail: kavraki@rice.edu, Phone: +1-713-348-5737, Fax: +1-713-348-5930
[‡] Electronic mail: cecilia@rice.edu, Phone: +1-713-348-3485, Fax: +1-713-348-5155

Abstract

The definition of reaction coordinates for the characterization of a protein folding reaction has long been a controversial issue, even for the “simple” case where one single free energy barrier separates the folded and unfolded ensemble.

We propose a new and general approach to this problem based on nonlinear dimensionality reduction. Essentially, the configurational space spanned by a protein during folding can be imagined as a low-dimensional non-linear manifold, embedded in a much higher-dimensional space. Taking advantage of recent advances in nonlinear dimensionality reduction we infer reaction coordinates directly from molecular dynamics simulation data.

We apply this method to characterize the folding landscape associated with a coarse-grained src-SH3 protein model, as sampled by molecular dynamics simulations. The folding free energy landscape projected on the coordinates extracted from the embedding can correctly distinguish the folding transtition state ensemble from the folded and unfolded state ensembles. The first embedding dimension efficiently captures the evolution of the folding process along the main folding route.

These results clearly show that a complex process such as protein folding can be essentially described by a very low-dimensional free energy landscape.

1 Introduction

The folding of a protein to its functional (native) state can be viewed as a chemical reaction, where the ensemble of unfolded configurations constitutes the reactant and the native state is the product.

Generally, the characterization of chemical reactions requires to locate the reactants, products, and transition states on a free energy surface. Simple models (so-called “reaction profiles”, or “reaction coordinate diagrams”) are oftentimes 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 where 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 source of significant debate– in protein folding studies.

We present here a new approach to the definition of reaction coordinates for the theoretical characterization of a protein folding free energy landscape, based upon the idea of non-linear dimensionality reduction. Modern dimensionality reduction techniques allows 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 src-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 oftentimes relies upon 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 upon the choice of the reaction coordinates [2–4]. Alternative definitions of reaction coordinates have been discussed in the literature [2, 5–8], as well as 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 Φ-value analysis [20, 21]) and the corresponding ensembles obtained on a low-dimensional landscape defined through these reaction coordinates.

It has been argued that the parameter P_{fold}, defined as the probability of a protein structure to reach the folded state before the unfolded state, would serve as ideal exact reaction coordinate for protein folding studies [2, 8, 14, 22, 23]. However, the calculation of P_{fold} is computationally so expensive that it becomes unfeasible for most systems of interest. Moreover, it has been shown recently that the parameter P_{fold} does not capture the essential features of a folding landscape if the folding mechanism is intrinsically multidimensional (as for instance is the case when folding occurs via multiple routes), or in the presence of intermediate states [5]. The definition of new strategies to estimate the intrinsic dimensionality of a folding reaction and the definition of the reaction coordinates themselves are paramount issues in folding studies. The approach presented here addresses both these questions.

2 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 $\mathscr{N}$ atoms is thus completely specified by $3\mathscr{N}$ parameters. However, these parameters are not independent on each other. Clearly, the constraints of maintaining intact the covalent bonds and angles and other steric factors effectively reduce the degrees of freedom 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 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 (MD) trajectories provide a sampling of configurations populating the 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 while all 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 the number of dimensions chosen. This 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 sub-manifold. Minima and saddle points of a specified function of the embedding coordinates (such as a free energy) can be automatically identified in an embedding. This is important if more than two or three dimensions are needed to capture the features of the original data, as in that case it is not possible to identify visually the folded and unfolded minima or transition paths between them.

3 Dimensionality reduction of folding simulations: fundamental and technical aspects

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

[§] "Feature" is a rather vague term, as a number of essential properties of the original data that one wish to preserve need to be specified, and the choice may be system-dependent.

representation and dependencies between dimensions are removed. Dimensionality reduction techniques fall broadly into two categories: linear and nonlinear techniques. 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 non-linear the low-dimensional embedding obtained by means of PCA is severely distorted. PCA is commonly used in the analysis of near-equilibrium fluctuations sampled by MD simulations [29–33], as 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 non-linear techniques need to be used[¶]. The fact that empirical reaction coordinates routinely used in protein folding studies can not be reduced to a linear combination of the Cartesian coordinates underscores the inadequacy of linear dimensionality reduction techniques to characterize a folding landscape.

3.1 The Underlying Idea: Isomap Algorithm

Although several non-linear dimensionality reduction techniques have been proposed (especially in the context of image analysis [34], speech recognition [35], visualizing word usages [36], climate data analysis [37, 38]) the development of new methods is still an active area of research. The technique we use here for the characterization of folding landscapes is based upon the recently proposed Isomap algorithm [39]. 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 [39]. Since we start from the assumption that our data lie on a low-dimensional manifold, a good approximation of the "geodesic distance" between a general pair of points, say x and y, on the manifold can be obtained by adding up the short distances connecting neighboring points throughout the shortest sequence of segments connecting x and y. This assumption implies that the geodesic distance for a couple of neighboring points can be approximated by the distance between them in the initial (high dimensional) space. Figure 1 illustrates this idea on a

[¶] One of the technical disadvantage of PCA is that it works on the covariance matrix of mean-centered data, thus requiring the data to be of an euclidean nature, since it uses coordinates to compute the optimal projections. This in turn prompts the need for alignment of the whole set of conformations with some "reference" protein conformation to allow the aligned Cartesian coordinates to be used as the input space. The alignement biases the principal components by centering them around the reference structure. The nonlinear method we consider here preserves the original RMSD distance between conformations in a global sense, and does not rely on a reference structure and Cartesian coordinates.

simple case of embedding. The data points shown in Figure 1(a) lie on a 2-dimensional torus embedded in a 3-dimensional space. The application of the Isomap algorithm to this set of data produces the 2-dimensional manifold shown in Figure 1(b), where the network of neighboring points is also shown.

The approximation for the geodesic distance holds provided that the data represents a good sampling of the embedded manifold (*i.e.*, the sampling needs to be sufficiently dense). In addition the neighborhood size cannot be too small or too large, that is, a robust definition of “neighboring points” is required. These issues are discussed in details in the Supplementary Material, where several tests on the validity of the approximations used are presented.

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

i) First, all nearest neighboring points (according to some distance metric) are computed for each point. We can either choose a fixed number of neighbors or fix a cut-off distance for a neighborhood around a point. The nearest neighbors induce a graph where the nodes correspond to the data points. There exists an edge between two nodes if they are nearest neighbors. The edges are weighted by the distance between them.

ii) The second step consists in computing the shortest paths between every pair of nodes. These shortest paths approximate the geodesic distances within the underlying manifold.

iii) The final step is to apply Multidimensional Scaling[‖] (MDS) to the matrix of pairwise shortest path distances to obtain a low-dimensional embedding.

The basic Isomap algorithm as described above suffers from two performance bottlenecks and can not be directly applied to the study of folding reactions. The major bottleneck is the computation of the shortest paths for all pairs. This operation requires $O(kN^2 \log N)$ time, where k is the neighborhood size (see Supplementary Material §C.4) and N the number of points. Another, relatively smaller bottleneck is the eigenvalue calculation that is part of the MDS algorithm. Computing all eigenvalues takes $O(N^3)$ time, but computing just the first m eigenvalues, $m \ll N$, can be done much more efficiently using iterative Arnoldi methods [40, 41]. These bottlenecks render computationally impossible the application of the Isomap algorithm to study protein simulations, where the number of conformations sampled is generally $N \gg 100,000$. We use here the basic idea of Isomap as a starting point to define a procedure amenable to deal with very large sets of protein conformations and computationally efficient. The procedure we propose

[‖] Multidimensional Scaling computes an embedding such that distances in the embedding correspond to *dissimilarities* between the original data points. The dissimilarity function used can be Euclidean distance (in which case PCA and MDS are equivalent), but in general does not need to be a metric. All that is required of a dissimilarity function $d(i,j)$ is that $d(i,j) = d(j,i)$, it returns 0 if $i = j$, and is positive otherwise. The “geodesic distance” as defined in the text can be used as dissimilarity function. Given a matrix of pairwise dissimilarities D, MDS computes the largest eigenvalues and corresponding eigenvectors of $B = -\frac{1}{2}HD^2H$, where H is the “centering matrix” (i.e., the sum of each row and column of B is 0). The embedding coordinates of point i are $(\sqrt{\lambda_1}\vec{v}_1, \ldots, \sqrt{\lambda_d}\vec{v}_d)$, where λ_k and $\vec{v}_k$ are the k^{th} largest eigenvalue and corresponding eigenvector of B.

is outlined in next section.

Before we proceed to present our method and the results obtained in the application to the analysis of folding simulations, it is worth mentioning another nonlinear dimensionality reduction technique that has received much attention in recent years: Locally Linear Embedding (LLE) [42]. This algorithm computes an embedding in three stages. Like Isomap, it first computes the nearest neighbors of each point. It then computes the best possible linear construction of each point in terms of its neighbors. Given the reconstruction weights associated with each point, the final step is then to compute embedding coordinates such that the difference between the embedding coordinates and the reconstruction from the neighbors' embedding coordinates is minimized. This problem can be reduced to finding the smallest eigenvalues and corresponding eigenvectors of a sparse matrix. Isomap and LLE algorithms may seem similar, but they are fundamentally different. While Isomap preserves *global* properties (geodesic distances between points), LLE preserves *local* properties by considering the differences between a point and its neighbors. LLE uses only sparse matrices, but it requires the computation of the *smallest* eigenvalues. That unfortunately becomes much more difficult than computing the *largest* eigenvalues (as in the Isomap algorithm) as the matrix size increases. Basically, if the matrix is ill-conditioned (which tends to be the case if the intrinsic dimensionality is much smaller than the dimensionality of the original space), the computation is limited by numerical precision problems. The Isomap algorithm does not suffer from this problem. For all these reasons our procedure is based upon the Isomap rather than the LLE algorithm.

4 Application of Nonlinear Dimensionality Reduction to Large Folding Simulation

As discussed in the previous section, the basic Isomap algorithm can not be straightforwardly applied to the analysis of folding/unfolding trajectories. However, the algorithm becomes suited to this purpose when a number of non-trivial modifications are introduced.

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 landmarks reduces the computational time by a factor n_L/N. A slightly modified version of MDS then computes from the $n_L \times N$ distance matrix an embedding for all N points. 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 section C.2 in the Supplementary Material).

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 (that 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.

We scale up the procedure by implementing the following three steps:

i) Filtering and reinsertion of redundant configurations:

The objective of the dimensionality reduction is to preserve the shape of the conformational manifold and the density of samples on that manifold as best as possible. We expect that low free-energy areas on a folding landscape will have a very high sampling density. The high number of conformations sampled in these areas are redundant to infer the topology of the manifold under consideration. The recovery of the embedded manifold becomes computationally efficient when most of the many similar conformations visited along folding simulations are filtered out (see Supplementary Material C.4 for detail on the filtering procedure). However, the density of states around a particular structure is an essential piece of information for free energy calculations. Discarding configurations in high density regions does not significantly affect the recovery of purely geometric information but strongly distorts calculations of thermodynamic quantities on the embedded manifold. This problem is circumvented by reinserting all the discarded conformations onto the embedded manifold after the embedding is found, in order to preserve the density of state as a function of the newly found coordinates. The reinsertion of points on the embedded manifold can be done by using a local fitting much in the spirit of the LLE procedure[42] described above. In practice, the embedding coordinates of the filtered-out conformations are computed as follows. For a conformation c to be reinserted we compute the k nearest neighbors –in the original space– among the conformations used to compute the embedding. The next step is to express c as a weighted sum of its neighbors as best as possible (see [42] for detail). Finally, we compute the new embedding coordinates of c by applying the same weights to the corresponding embedding coordinates of the neighbors.

The insertion of conformations into a 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 the configurations initially filtered out could be reinserted without experiencing such a problem.

ii) Reducing the size of each conformation: The Isomap algorithm requires to determine the nearest neighbors of each conformation. In the work presented here distance is measured using root-mean-square-deviation (RMSD). This metric is relatively expensive to compute compared to, say, Euclidean distance[**]. Fortunately, we can use a lower bound on the RMSD between two conformations that can be computed much faster. The bound is obtained by averaging the positions of groups of atoms, and computing the RMSD of these averaged conformations. Further discussion on this point, a validation of the approximation involved, and a rigorous proof of this statement is provided in the Supplementary Material. It is worth mentioning that this issue may be particularly relevant for the application of the dimensionality reduction procedure to all-atom protein simulations.

iii) Computing the embedding in parallel on a cluster of machines: We have developed a parallel implementation of the algorithm. Our implementation uses MPI for inter-process communication and can be run on a large cluster of machines.

The first two steps above involve approximations. The last step is exact. It entails transforming the landmark Isomap algorithm into an output-equivalent parallel algorithm. The validation of these steps is discussed in the Supplementary Material C.4, where more details on the motivation and implementation of the procedure are also presented.

5 Results: Folding landscape of SH3 as a low-dimensional embedded manifold

We tested the non-linear dimensionality reduction procedure outlined above by applying it to characterize the protein folding landscape obtained from simulation of a coarse-grained model of src-SH3 domain. The basic ideas and computational details of the model are briefly described in the Supplementary Material A, and detailed in a recent publication [1], where a comparison of the results with experimental data is also presented.

The purpose of the application presented here is not to further validate this coarse-grained protein

[**] Computing RMSD involves: re-center all the conformations to the origin, computing a 3×3 covariance matrix from the two conformations under consideration, computing an optimal alignment from this covariance matrix, applying the alignment to one of the conformations, and, finally, computing the Euclidean distance between the aligned conformations.

model, rather to show how non-linear 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. Algorithm details and the values of the parameters used are provided in the Supplementary Material C.4.

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, d, of dimensions considered for the embedded manifold. Following the definition used in [39] 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 the entries of matrices $\hat{D}_d$ and D_n. The matrix $\hat{D}_d$ contains all the pairwise distances as obtained on the d-dimensional embedding, while the matrix D_n stores the corresponding geodesic distances as computed in the original (n-dimensional) space. In the case of 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 n of embedding dimensions considered 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.

Figure 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, $\sigma_r(1,n) \simeq 0.08$, $\sigma_r(2,n) \simeq 0.04$, and $\sigma_r(3,n) \simeq 0.02$. 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 to the corresponding residual variance obtained when PCA is used on the same data (red points in Figure 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, 47].

Free energy surfaces can be computed as a function of the embedding coordinates. Figure 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_{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_{fold} by means of a set of 100 ancillary simulations starting from it. For each small interval $x_1 \pm dx_1$ an average value of P_{fold} is computed over all conformations with a corresponding x_1 within that interval, while the variance is reported as error bar. The inset of Figure 3 shows that the range of values on the first embedding coordinate $x_1 \simeq -4$ corresponding to the of the free energy barrier has an associated value of $P_{fold} \simeq 0.5$. The red circle in the inset identifies the P_{fold} values corresponding to the top of the free energy barrier (that is, around $x_1 \simeq -4$). Remarkably, the transition state identified by means of the 1-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_{fold} analysis: the top of the free energy barrier corresponds to $P_{fold} \simeq 0.5$. The theoretical folding probability [48] $P_t(x_1) = \frac{\int_{x_1}^{x_U} \exp(F(y)/k_B T)dy}{\int_{x_N}^{x_U} \exp(F(y)/k_B T)dy}$ associated to the one-dimensional free energy $F(x_1)$ is also shown in the inset of Figure 3. The folding probability P_t is in agreement with the calculated P_{fold} values on most of the interval considered (particularly, at the transition state). Deviation between the average value of the calculated P_{fold} and the theoretical folding probability $P_t(x_1)$ are observed around the folded state ($x_1 \simeq -6$) and can be explained in terms of the variation of free energy along the second embedding dimension in this region (see Figure 4).

It is worth noting that for the protein model considered here the P_{fold} analysis required $> 12,000$ CPU hours (Intel Xeon 2.2 GHz) and was performed for a small subset of configurations[††] while the embedding procedure was completed in < 500 CPU hours (less than 24 CPU hours running on 20 processors) and provides information on the whole configurational space.

Not surprisingly, the transition state ensemble from the 1-dimensional embedded manifold of the SH3 model is also in good agreement with what obtained using the parameter Q as 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, 47], SH3 among them. However, this may not be the case in general, particularly for more complex folding reactions where more than one reaction coordinates is needed.

Additional information on the folding process is obtained when the first two embedding dimensions are considered as reaction coordinates in the free energy calculation. Figures 4(a)-(c) present the 2-dimensional embedded free energy landscape as a function of the first two embedding dimensions. Figure 4(a) shows a contour plot of the free energy. Again, as expected for a two-state folding protein, two distinct free energy minima appear: a more localized one corresponding to the folded state, and one with a larger

[††] The P_{fold} parameter was computed for about 8000 protein configurations. The total number of configurations used in the definition of the embedded free energy landscape is $1,818,000$.

basin corresponding to the unfolded state. The free energy gradient field is superimposed to the free energy contour plot in Figure 4(b). The transition state ensemble on this 2-dimensional landscape can be defined by considering the "Continental Divide", that is, 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 Figure 4(b).

It is clear from the figure 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, as a non-negligible amount of structures lie outside the minimum free energy path (about 15% of structures lie within the the light orange free energy level on figure 4).

It is important to clarify that the existence of a main folding route doesn't 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 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 folding process simply means that 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.

Figure 4(c) presents the results from the P_{fold} analysis superimposed on the 2-dimensional embedded landscape. The comparison of Figures 4 (b) and (c) reveals that the region with $P_{fold} \simeq 0.5$ matches the separatrix region identified by the diverging gradient fluxes. The variance of P_{fold} measured in each 2-dimensional interval $(x_1 \pm dx_1, x_2 \pm dx_2)$ is $\delta P_{fold} \simeq 0.12$, significantly lower than the variance $\delta P_{fold} \simeq 0.2$ observed in the 1-dimensional case (see Figure 3). The larger uncertainty obtained when only one embedding dimension is considered accounts for the fluctuations observed along the second embedding coordinate.

Finally, Figure 5 shows the free energy landscape obtained when the first three embedding dimensions are considered as reaction coordinates. The third dimension spans a small range, less than 1/4 of the range spanned by the first dimension. Moreover, deviations from the 2-dimensional landscape involve only high free energy regions, that are populated with low probability. Similarly to the 2-dimensional landscape, the lowest free energy regions clearly identify the main folding route, evolving along the first embedding coordinate. Alternative routes have a lower probability (*i.e.*, higher free energy).

6 Conclusions

We propose a general procedure to obtain a low-dimensional free energy landscape associated with a simulated protein folding reaction. By using non-linear dimensionality reduction methods (based on Isomap [39]) an embedded folding manifold is extracted from a large set ($\simeq 2,000,000$) of protein conformations sampled throughout extensive folding/unfolding simulations of a coarse-grained model of src-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 this procedure to the folding of a coarse-grained protein model of 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_{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_{fold} analysis confirms that protein configurations in the transition state region as identified on the embedded free energy landscape have $P_{fold} \simeq 0.5$. Moreover, fluctuations around this average value of P_{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. Applications of this procedure to the characterization of the folding mechanism of larger systems and more complex reactions are in progress.

Acknowledgments

This work was supported by grants from NSF (CC Career CHE-0349303, LK ITR-0205671, LK and CC CCF-0523908 and CNS-0454333), ATP (003604-0010-2003), the Robert A. Welch Foundation (CC Norman Hackermann Young Investigator award, and C-1570), and the Sloan Foundation (LK). The Rice Terascale Cluster used for the calculations is supported by NSF under Grant EIA-0216467, Intel, and Hewlett Packard.

We acknowledge Alessandro Mossa for contributions to initial stages of this project. We thank Peter Wolynes, Attila Stzabo, Yannis Kevrekidis, Tolya Kolomeisky and Michele Vendruscolo for their insighful comments, and members of Kavraki's and Clementi's groups for stimulating discussions.

References

[1] P. Das, S. Matysiak, and C. Clementi. Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc. Natl Acad. Sci. USA*, 102:10141–10146, 2005.

[2] R. Du, V.S. Pande, A.Yu. Grosberg, T. Tanaka, and E.I. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108:334–350, 1998.

[3] G. Hummer. From transition paths to transition states and rate coefficients. *J. Chem. Phys.*, 120: 516–523, 2004.

[4] R.B. Best and G. Hummer. Reaction coordinates and rates from transition paths. *Proc. Natl Acad. Sci. USA*, 102:6732–6737, 2005.

[5] S.S. Cho, Y. Levy, and P.G. Wolynes. P versus q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl Acad. Sci. USA*, 103:586–591, 2006.

[6] A. Baumketner, J-E. Shea, and Y. Hiwatari. Improved theoretical description of protein folding kinetics from rotations in the phase space of relevant order parameters. *J. Chem. Phys.*, 121:1114–1120, 2004.

[7] A. Ma and A.R. Dinner. An automatic method for identifying reaction coordinates in complex systems. *J Phys B*, 109:6769–6779, 2005.

[8] Y.M. Rhee and V.S. Pande. One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution. *J Phys B*, 109:6780–6786, 2005.

[9] D.K. Klimov and D. Thirumalai. Progressing from folding trajectories to transition state ensemble in proteins. *Chem Phys*, 307:251–258, 2004.

[10] C. Clementi, P.A. Jennings, and J.N. Onuchic. Prediction of folding mechanism for circular-permuted proteins. *J. Mol. Biol.*, 311:879–890, 2001.

[11] C. Clementi and S.S. Plotkin. The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Sci*, 13:1750–1766, 2004.

[12] V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich. A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Fold. Des.*, 3:183–194, 1998.

[13] L.L. Chavez, J.N. Onuchic, and C. Clementi. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J Am Chem Soc*, 126:8426–8432, 2004.

[14] N.V. Dokholyan, L. Li, F. Ding, and E.I. Shakhnovich. Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, 99:8637–8641, 2002.

[15] P. Das, C.J. Wilson, G. Fossati, P. Wittung-Stafshede, K.S. Matthews, and C. Clementi. Characterization of the folding landscape of monomeric lactose repressor: Quantitative comparison of theory and experiment. *Proc. Natl Acad. Sci. USA*, 102:14569–14574, 2005.

[16] J.D. Bryngelson and P.G. Wolynes. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J.Phys. Chem.*, 93:6902–6915, 1989.

[17] H. Nymeyer, A.E. García, and J.N. Onuchic. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl Acad. Sci. USA*, 95:5921–5928, 1998.

[18] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem*, 48:545–600, 1997.

[19] J-E. Shea and C.L. Brooks III. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu Rev Phys Chem*, 52:499–535, 2001.

[20] A.R. Ferst and S. Sato. ϕ-value analysis and the nature of protein-folding transition states. *Proc. Natl Acad. Sci. USA*, 101:7976–7981, 2004.

[21] A.R. Fersht, R.J. Leatherbarrow, and T.N.C. Wells. Quantitative analysis of structure-activity relationships in engineered proteins by linear free-energy relationships. *Nature*, 322:284–286, 1986.

[22] D. K. Klimov and D. Thirumalai. Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins: Struct. Funct. Genet.*, 43:465–475, 2001.

[23] F. Ding, W. Guo, N. V. Dokholyan, E. I. Shakhnovich, and J-E Shea. Reconstruction of the src-sh3 protein domain transition state ensemble using multiscale molecular dynamics simulations. *J. Mol. Biol.*, 350:1035–1050, 2005.

[24] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12:103–108, 1990.

[25] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[26] M. Benito and D. Pena. Dimensionality reduction with image data. *Lecture Notes in Computer Science*, 3177:326–332, 2004.

[27] E. Cho, D. Kim, and S.Y. Lee. Posed face image synthesis using nonlinear manifold learning. *Lecture Notes in Computer Science*, 2688:946–954, 2003.

[28] I.T. Jolliffe. *Principal Components Analysis*. Springer-Verlag, New York, 1986.

[29] S. Hayward, A. Kitao, and N. Go. Harmonic and anharmonic aspects in the dynamics of bpti - a normal-mode analysis and principal component analysis. *Protein Sci*, 3:936–943, 1994.

[30] S. Hayward, A. Kitao, and H.J.C. Berendsen. Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozymeessential domain motions in barnase revealed by md simulations. *Proteins: Struct. Funct. Genet.*, 27:425–437, 1997.

[31] S.B. Nolde, A.S. Arseniev, V.Y. Orekhov, and M. Billeter. Essential domain motions in barnase revealed by md simulations. *Proteins: Struct. Funct. Genet.*, 46:250–258, 2002.

[32] Y. Levy and A. Caflisch. Flexibility of monomeric and dimeric hiv-1 protease. *J Phys B*, 107:3068–3079, 2003.

[33] Phillips G.N.Jr. Teodoro, M. and L.E. Kavraki. Understanding protein flexibility through dimensionality reduction. *J. Comp. Biol.*, 10:617–634, 2003.

[34] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *Proceedings of the IEEE Conference in Computer Vision and Pattern Recognition*, pages 988–995, 2004.

[35] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comp.*, 15:1373–1396, 2003.

[36] M.E Brand. Continuous nonlinear dimensionality reduction by kernel eigenmaps. *In Proceedings of the Eighteenth International Join Conference on Artificial Intelligence*, pages 547–552, 2003.

[37] A. H. Monahan. Nonlinear principal component analysis by neural networks: Theory and application to the lorenz system. *J Climate*, 13:821–835, 2000.

[38] A. Z. Gamez, C. S. Zhou, A. Timmermann, and J. Kurths. Nonlinear dimensionality reduction in climate data. *Nonlinear Processes in Geophysics*, 11:393–398, 2004.

[39] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(22):2319–2323, December 2000.

[40] W.E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, 9:17–29, 1951.

[41] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, second edition, 2003.

[42] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(22):2323–2326, December 2000.

[43] V. de Silva and J.B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 705–712. MIT Press, Cambridge, MA, 2002.

[44] V. de Silva and J.B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, Mathematics Department, 2004.

[45] A.M. Ferrenberg and R.H. Swendsen. Optimized monte carlo data analysis. *Phys. Rev. Lett.*, 63: 1185–1198, 1989.

[46] A.M. Ferrenberg and R.H. Swendsen. New monte carlo technique for studying phase transitions. *Phys. Rev. Lett.*, 61:2635–2638, 1988.

[47] C. Clementi, H. Nymeyer, and J.N. Onuchic. 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.*, 298:937–953, 2000.

[48] Y.M. Rhee and V.S. Pande. On the role of chemical detail in simulating protein folding kinetics. *Chemical Physics*, in press, 2006.

[49] V. P. Grantcharova and D. Baker. Folding dynamics of the src sh3 domain. *Biochemistry*, 36:15685–15692, 1997.

[50] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatam, D.M. Ferguson, and D.M. Singh. Amber,v.4.1. 1984.

[51] B. Roux. The calculation of the potential of mean force using computer simulations. *Comp Phys Comm*, 91:275–282, 1995.

[52] I. Lotan and F. Schwarzer. Approximation of protein structure for fast similarity measures. *J.Comp.Biol.*, 11(2–3):299–317, 2004.

[53] S. Brin. Near neighbor search in large metric spaces. In *Proc. 21st Conf. on Very Large Databases*, pages 574–584, 1995.

[54] T.H. Cormen, C.E. Leiserson, R.R. Rivest, and C. Stein. *Introduction to Algorithms*. MCG, second edition, 1990.

[55] K.J. Maschhoff and D.C. Sorensen. P_ARPACK: An efficient portable large scale eigenvalue package for distributed memory parallel architectures. In *Proceedings of the Third International Workshop on Applied Parallel Computing, Industrial Computation and Optimization*, pages 478–486. Springer-Verlag, 1996.

[56] R.B. Lehoucq, D.C. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, 1998.

FIGURE CAPTIONS

FIGURE 1: A simple case of embedding: The data points (blue dots) shown in (a) "live" on the 2-dimensional surface of the torus, although they are embedded on a 3-dimensional space. The application of the Isomap algorithm [39] to this set of data defines two independent coordinates on which all points are mapped. The resulting 2-dimensional embedded space is shown in (b). These two embedding coordinates can not 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.

FIGURE 2: Residual variance as a function of dimensions considered in the embedding, as obtained when the Principle Component Analysis (blue dots) or our non-linear dimensionality reduction (red points) –based on the idea of the Isomap algorithm [39]– are used to characterize the space sampled in extensive folding/unfolding simulations of a src-SH3 protein model. The high residual variance associated to PCA is not surprising, as linear dimensionality reduction methods can not be used to analyze intrinsically nonlinear spaces such as the configurational space explored by a protein during folding. In contrast, the low residual variance resulting from our nonlinear dimensionality reduction indicates that this method can successfully define a few embedded coordinates that capture the essential dynamics of the protein model. The residual variance drops very close to zero for more than 3 dimensions, indicating that the first three embedded dimension suffice to describe the process.

FIGURE 3: 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 $x_1 \simeq -4$. The average value of P_{fold} associated to each small interval $x_1 \pm dx_1 \in (-7, 0)$ is plotted in the inset, as a function of x_1. The error bar corresponds to the variance of P_{fold} for a given value of x_1. The continuous gray line is the theoretical folding probability $P_t(x_1)$ associated to the one-dimensional free energy curve $F(x_1)$ (see text for detail). The red circle identify the P_{fold} and P_t values corresponding to the top of the free energy barrier (that is, around $x_1 \simeq -4$): The transition state ensemble identified by the one-dimensional free energy profile corresponds to $P_{fold} \simeq P_t \simeq 0.5$. Deviation between the average P_{fold} and the theoretical folding probability $P_t(x_1)$ around the folded state can be explained in terms of the fluctuations along the second embedding dimension (see Figure 4).

FIGURE 4: 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 the dimensionality reduction procedure. The free energy is shown as a contour plot in (a). Each contour represents an increase of free energy of $1 k_B T$ and colored according to the color map shown at the top of the figure (colors from red to blue

indicate progressively decreasing free energy). The free energy gradient field is superimposed to the free energy contour plot in (b). The thick gray line approximately locates the separatrix between the the folded and unfolded state basins, where gradient fluxes leading to opposite minima meet. The results from the P_{fold} analysis are superimposed on the 2-dimensional embedded landscape in (c). The average value of P_{fold} at a given (x_1, x_2) position on the landscape is color-coded according to the color-map shown at the top of the figure. Colors ranging from red to blue indicates values of P_{fold} increasing from 0 to 1. The comparison of (b) and (c) reveals that the region with $P_{fold} \simeq 0.5$ is fully consistent with the separatrix region.

FIGURE 5: Free energy landscape obtained when the first three embedding dimensions are considered as reaction coordinates. Each isosurface marks an increase of free energy of $1k_BT$. A smaller range of free energy than what used in Figure 4 is shown here, to simplify the image. The third embedded dimension spans a much smaller range than the first two.

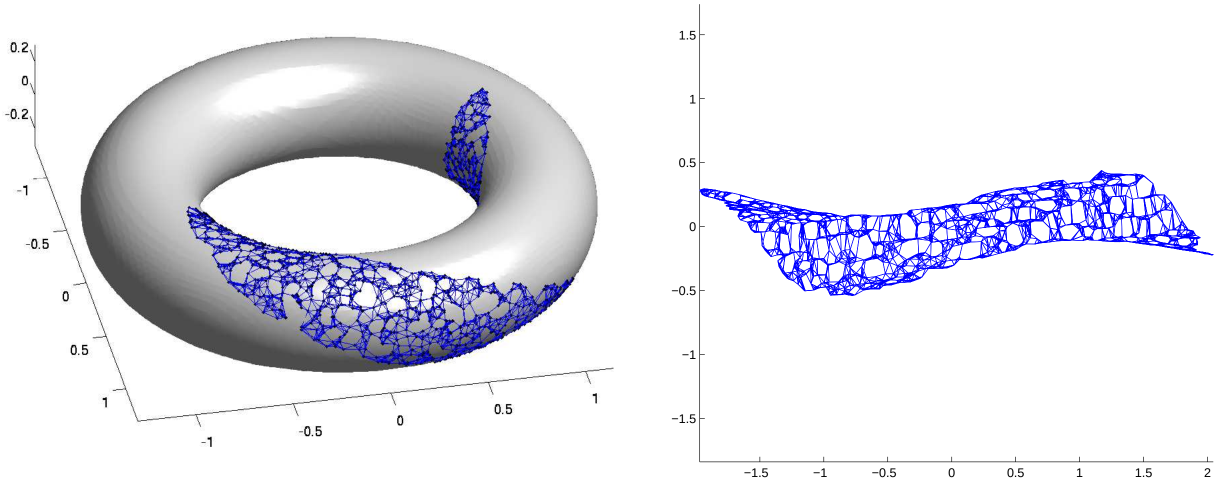


Figure 1:

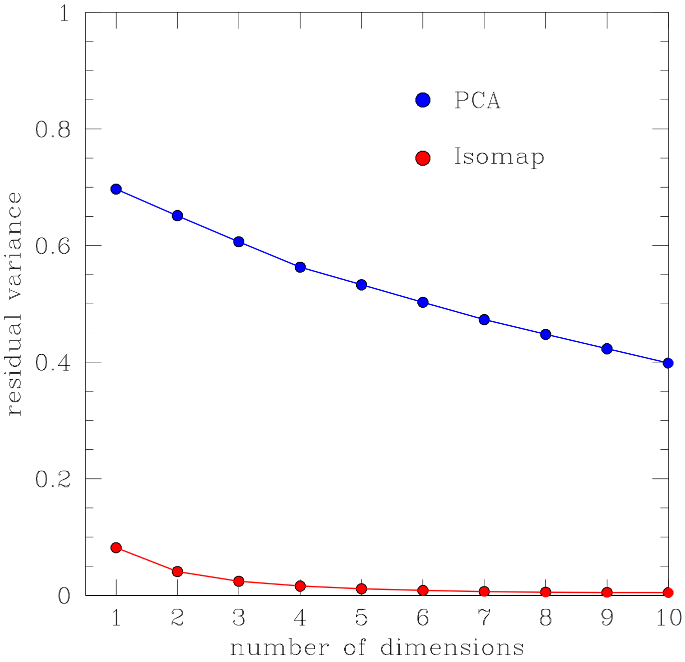

1
0.8
0.6
0.4
0.2
0
residual variance
1
2
3
4
5
6
7
8
9
10
number of dimensions
PCA
Isomap

Figure 2:

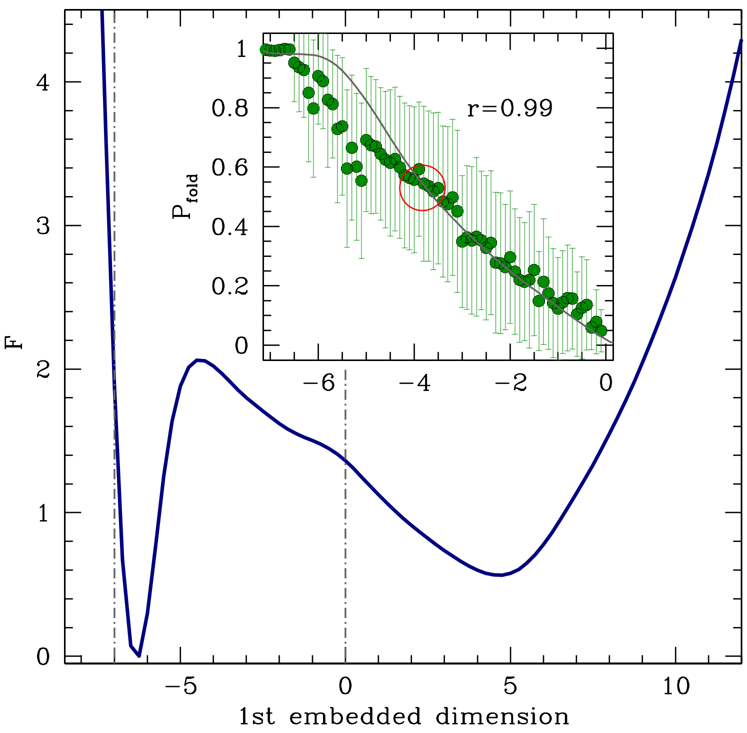
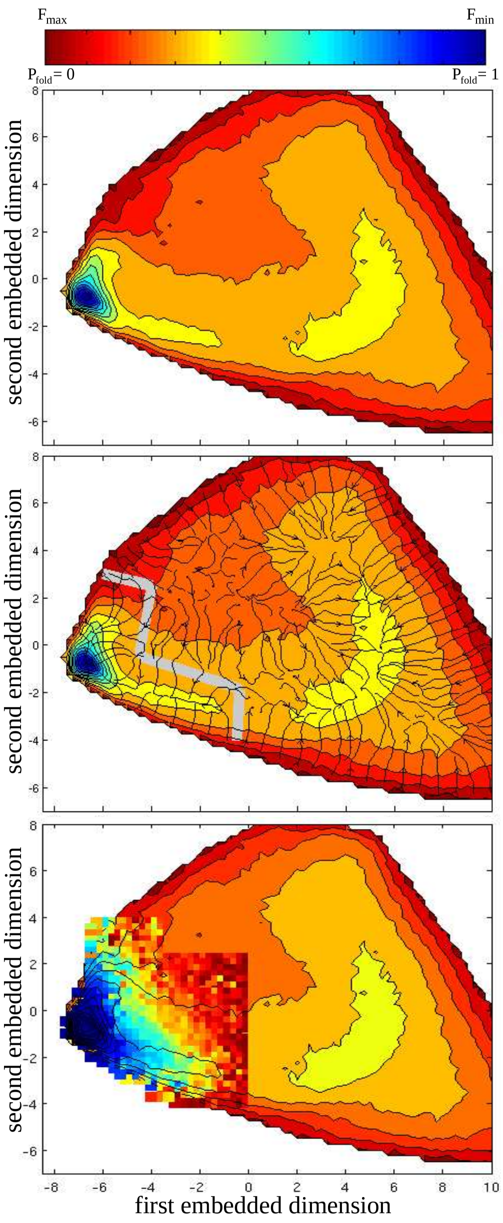

F
0
1
2
3
4
−5
0
5
10
1st embedded dimension
P_fold
0
0.2
0.4
0.6
0.8
1
−6
−4
−2
0
r=0.99

Figure 3:



F_{max}
F_{min}
$P_{fold}= 0$
$P_{fold}= 1$
second embedded dimension
second embedded dimension
second embedded dimension
first embedded dimension
8
6
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
6
8
10

Figure 4:

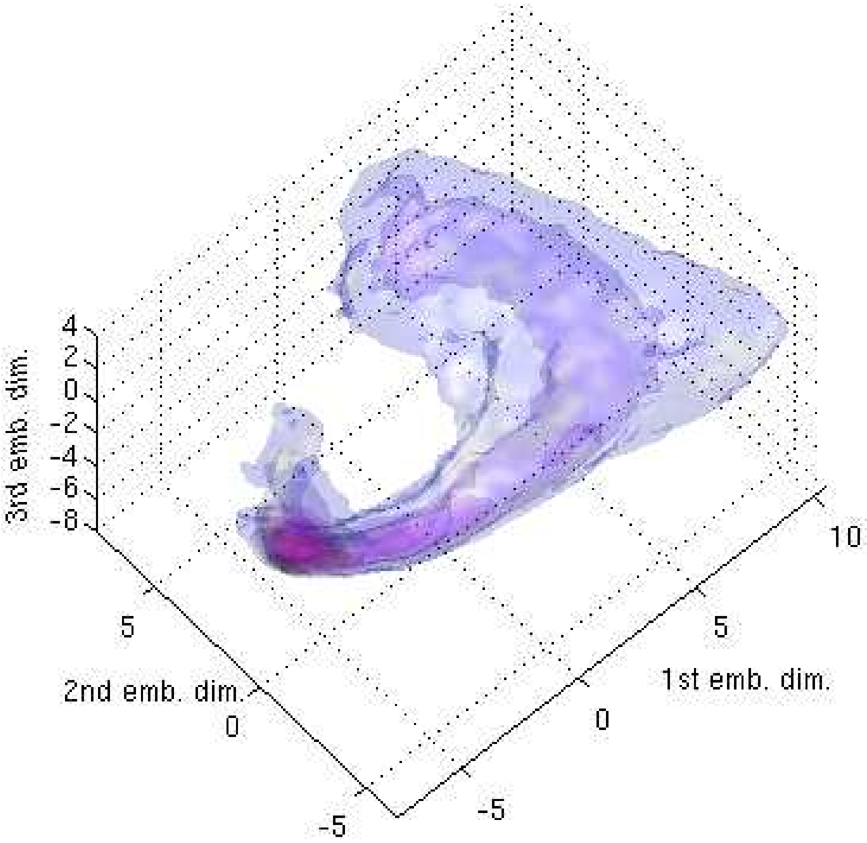


Figure 5:

SUPPLEMENTARY MATERIAL

A Molecular Dynamics Simulations

The dimensionality reduction method described in the paper is applied to analyze the folding landscape of a coarse-grained model of src-SH3, a 57 residue protein domain that is known to fold in a two-state fashion, and has been extensively studied both computationally [1, 11, 47] and experimentally [49]. The coarse-grained model uses an off-lattice simplified representation of the protein, where each amino acid is described by a single bead on a polymer chain located on the C_α position. The potential energy function is comprised of a local and non-local term. The local potential term includes bond, angle and torsional energy terms and is designed to have its absolute minimum in the native state, as in previous models (see *e.g.*, [15, 47]). The non-local potential term describes the interaction between a residue pair separated by at least three residues and is expressed using a pairwise standard 10-12 potential function. The energy parameters of the non-local potential term are optimized by an iterative procedure where the stability of the native state is maximized with respect to the compact non-native structures with low energy[1]. Details on the energy function and the design procedure for the energy parameters can be found in ref. [1].

The folding/unfolding simulations are performed at constant temperature using the MD module SANDER of the simulation package AMBER [50], properly adapted to deal with the minimalist protein representation.

B Calculation of Thermodynamic quantities

The weighted histogram analysis method (WHAM)[45, 46, 51] is used to combine simulation data at different temperatures to estimate the density of states, which is then used to compute thermodynamic quantities over a continuous range of temperatures. In particular, the WHAM procedure is applied to compute the heat capacity as a function of temperature and the free energy profile as a function of one, two, or more embedding reaction coordinates (as resulting from the dimensionality reduction procedure). The folding temperature T_f is estimated as the temperature corresponding to the peak in the heat capacity curve.

C Computational Setup

As stated in the paper, our implementation is based on the landmark Isomap algorithm. Applying the “vanilla version” of the Isomap algorithm would be completely impractical for any trajectory of interest. Both the space and time requirements would be prohibitively large. In our simulations we have applied our algorithm to several MD trajectories, each with up to $606,000$ conformations. The total memory needed to store the pairwise distances for all pairs of conformations in such a trajectory is $606000^2 \times 8/1024^4 = 2.67$ terabyte (assuming that distances are stored in 8 bytes, the typical size of a double precision floating point number on a 32-bit machine). Even distributed over 100 nodes, this is more memory than typically found on a cluster node.

We provide here additional details on the definition and implementation of the nonlinear dimensionality reduction algorithm. Particularly, we discuss the approximations made and the testing performed to ensure that these approximations hold for the application presented in the paper.

C.1 Reducing the size of each protein conformation

As mentioned in section 4, a lower bound for the RMSD distance can be obtained by averaging the positions of groups of atoms, and computing the RMSD of these averaged conformations. A rigorous proof of this statement is provided below (see section D). If we compute the lower bound by averaging every p atoms, then the computation is approximately p times faster than the exact RMSD. This bound can be used in two ways. First, it can be used in nearest neighbor queries to determine if two conformations are “close”. If the lower bound on the RMSD is large, then the true RMSD is definitely large. Otherwise we compute the true RMSD to check whether the conformations are really close together or not. Obviously, the tighter the bound, the larger the performance improvement will be. If we end up computing both the approximate and exact RMSD in practice, there is actually a performance *loss*. The overall performance gain for maintaining a nearest neighbor data structure depends on how many atoms are averaged together as well as the distribution of the conformations. In our experience for the application presented in this paper, the bound is in fact very tight if we average over small groups of consecutive atoms. Figure C.1 shows the correlation between the real RMSD and the RMSD lower bound obtained by averaging p atoms together ($p = 1,\dots,5$), for $10,000$ pairs of conformations selected uniformly at random from a MD trajectory of the SH3 model. These results suggest a second use of the RMSD lower bound: we could use just the bound itself as an approximation of the RMSD. This step would introduce an approximation, but allows to gain significant computational and space savings. This approximation was not used to obtain the results presented in this paper. However, we envision that it may be particularly useful in applications to larger protein systems and/or all-atom protein configurations.

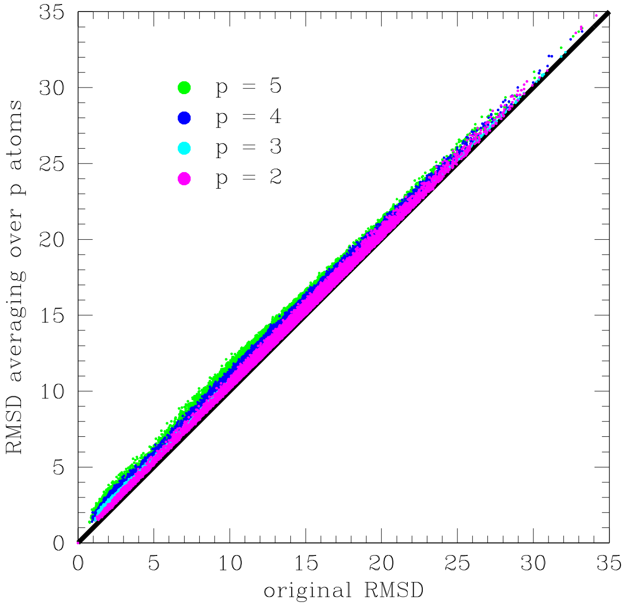


Figure C.1: In order to evaluate the tightness of the RMSD lower bound obtained by averaging beads together, we compare the true RMSD and the approximate RMSD obtained by averaging p beads ($p = 2, \ldots, 5$) for $10,000$ randomly chosen pairs of conformations. Different colors correspond to different values of p. The case $p = 1$ corresponds to the identity line (drawn in black). The small approximation introduced by averaging consecutive atoms allows a faster, less memory intensive run.

C.2 Reducing the number of conformations

As stated in the text, to recover just the shape of the embedded manifold we do not need to preserve the density of samples on that manifold. We filter out a large number of conformations in higher density areas before computing the low-dimensional embedding. Once we have an embedding we can reinsert the conformations that were left out into the embedding, as described in the text (see section 4).

For the application presented in the paper we have used *a priori* information on the folded conformation to filter out data from the folded state ensemble, that we know is a minimum of free energy, thus populated with high probability. However, it is possible to generalize the approach in a way that does not require any *a priori* information on the data, nor the existence of a well defined native state. Such an approach consists on filtering the conformations of a MD trajectory by an unbiased computational "tool", as for instance clustering or indexing data structures (such as trees) that group conformations by metric similarity (using RMSD as the metric, for instance). Conformations lying in a densely populated area (according to the tool used) can be discarded for later reinsertion. We are currently working on the development and testing of a set of these computational tools.

Figure C.2 shows the function $f(Q)$ used to filter out configuration for the src-SH3 simulation data used in the paper. The parameter Q quantifies the similarity to the native structure (pdb structure 1FMK.pdb, residues 84-140). For each configuration, with a corresponding Q value, the probability to be filtered out is given by $f(Q)$. This fraction is monotonically increasing with Q, with a sharp increase for large values of Q, in such a way that only configurations with $q > 0.7$ can be discarded. For $Q = 1$ we reject 80% of the conformations, as all the fully folded configurations are highly similar to each other. Around 20% of the conformations are removed in total.

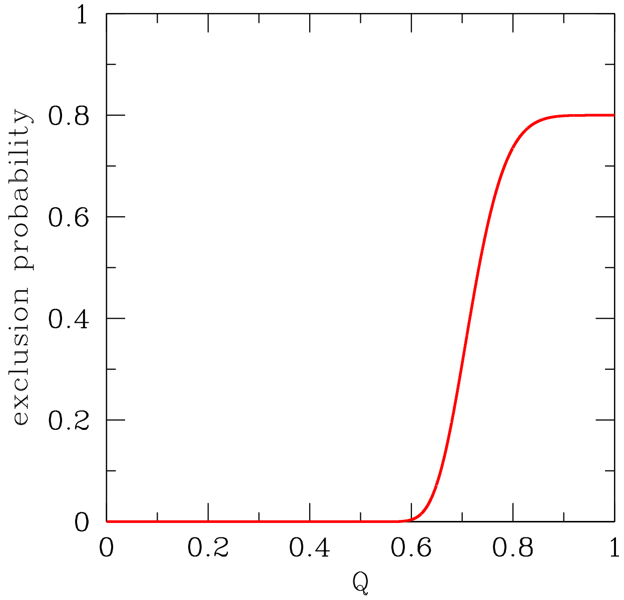


Figure C.2: The function $f(Q)$ used to filter out redundant conformations is plotted as a function of Q, the fraction of native contacts formed in a given structure. The value $f(Q) = 0$ corresponds to not filtering out any conformations. This choice of $f(Q)$ is motivated by the fact that most conformations with high Q-value are very similar to the native structure (and to each other), and contribute very little information to the large-scale manifold shape. A different, more general strategy can be used if no *a priori* information on the data set is available.

C.3 Choice of distance measure

In this paper we use RMSD as a distance measure between protein conformations. The procedure presented in the paper can accommodate different distance measures for the definition of geodesic. In principle, if the data represent a dense sampling of the manifold, neighboring points are all very close to each other, and the results are not strongly affected by the choice of the distance measure.

Another common distance measure for conformations is the difference in intra-molecular distances (sometimes referred to as dRMS). This metric has as the advantage to be translation and orientation invariant; that is, no alignment is necessary. However, for this metric we need to compare a number of distances that is quadratic in the number of atoms.

Like for RMSD, this metric can be approximated by averaging atoms together. The averaging can also be combined with performing PCA on the set of intramolecular distances to determine which residue pairs contribute the most to the overall dRMS, in the same spirit of the approach proposed in [52]. The use of a reduced set of residue pairs in the computation of dRMS can greatly reduce the computational time associated to this distance measure.

Preliminary tests with this alternative definition of distance measure (based on dRMS) did not produce any significant difference in the results, nor did show significant computational gain for the src-SH3 model protein used in the paper. For this reason in the work presented here we have chosen RMSD as a distance metric, as its use is more common in the analysis of MD simulations.

C.4 Definition of Neighboring Network

We have tested the dependence of the results on the neighborhood size, as quantified by the parameter k. For each conformation in our data set, we define a neighborhood as the k closest conformations to the selected one.

Some care needs to be taken in the choice of k. If k is too small, the connectivity of the network may decrease significantly, and the approximation of geodesics as the sum of distances connecting neighboring points throughout the network may be affected. On the other side, increasing k to a very big number would artificially "flatten out" the recovered manifold. In the limit $k = N - 1$ the geodesic distance between any two configurations reduces to the corresponding RMSD, and any nonlinearity of the dimensionality reduction is lost. However, for a dense sampling of an intrinsically low-dimensional manifold, we expect the procedure to be robust and results to be relatively insensitive to the exact value of k within a certain range of values.

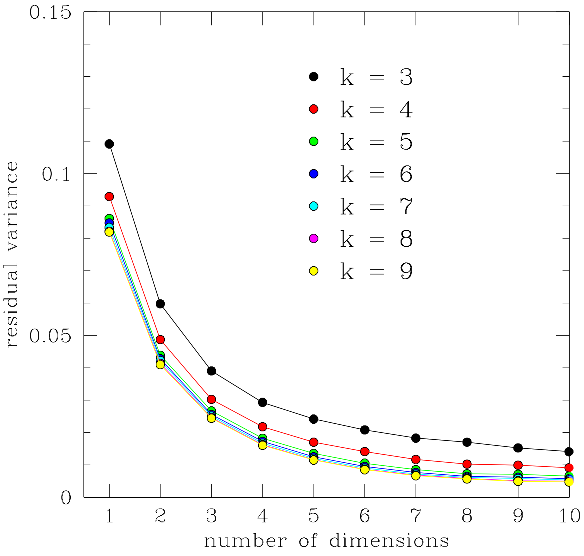


Figure C.3: Residual variance as a function of the number of dimensions, as obtained for different choices of the neighborhood size, k, for the src-SH3 model protein we have used in the paper. The resulting embedding appears completely insensitive to the neighborhood size within the range $5 \leq k \leq 9$. The magnitude of the residual variance should be compared to the residual variance obtained from PCA, shown in Figure 2 of the paper (note the different scale on the y-axis in Figure 2).

Figure C.3 shows the residual variance as a function of the number of dimensions, as obtained for different choices of k for the src-SH3 model protein we have used in the paper. The curves corresponding to $5 \leq k \leq 9$ all collapse into a single curve, suggesting that the embedded manifold as obtained from our procedure is relatively insensitive to the neighborhood size within this range of k. The results presented in the paper are obtained with $k = 9$.

C.5 Number of Landmarks

As stated in the paper, the use of landmark points has been recently proposed [43, 44] to significantly improve the efficiency of the Isomap algorithm when dealing with extremely large data sets.

We designate n_L data-points (*i.e.*, protein configurations in our case) to be landmarks and we compute the shortest path from each landmark to every other point. Ideally, if the manifold being recovered were truly d-dimensional, $d+1$ independent landmarks (i.e. not lying on a $d-1$-dimensional surface) should yield a unique, correct placement of all the points on the manifold. In practice, however, since the data are noisy and the true dimensionality of the manifold is unknown, a large number of landmarks must be used. If the landmarks are chosen randomly, the number of landmarks needs to be sufficiently larger than d to guarantee stability. In order to have a robust procedure, we use an increasing number of landmarks and monitor the results. If the number of landmarks is sufficiently large, no performance improvement is observed upon further increase. Figure C.4 shows the residual variance as a function of the number of dimensions, as obtained when an increasing number of landmarks $n_L \in (3,000-6,000)$ is used. Results obtained for $n_L \leq 4,500$ have an associated error significantly larger than results for $n_L \geq 4,650$. Further increasing the number of landmarks in the interval $(4,650-6,000)$ does not improve the quality of the embedding. Based on this test, we have used a number of landmarks $n_L = 5,000$ for the study presented in the paper.

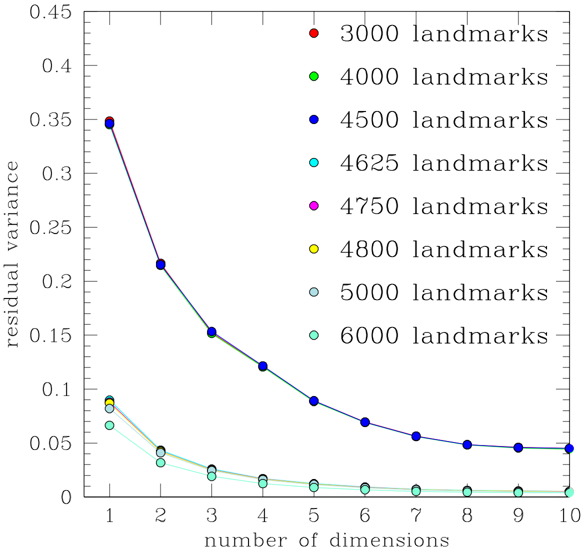


Figure C.4: Residual variance as a function of the number of dimensions, as obtained for different choices of the number of landmarks, n_L, used in the embedding procedure, for the src-SH3 model protein used in the paper. The resulting embedding appears robust against further increase in the number of landmarks for $n_L \geq 4,650$. Based on this test, we have selected a number of landmarks $n_L = 5,000$ for the study presented in the paper.

It is worth noticing that the overall number of configurations considered in the paper (for instance, the total number of configurations used to compute the free energy surfaces shown in Figure 4) is $\simeq 2,000,000$.

C.6 Parallelizing the Isomap Algorithm

Any practical application of the Isomap algorithm to real molecular trajectories will require some form of parallelization both to fit the entire trajectory in memory and to achieve non-trivial speedups in the computation. Broadly, Isomap is divided into three main stages that can be parallelized in different ways; our approach was to do so as follows.

1. *Building the neighborhood graph.* Since we are using RMSD as our distance measure, which is non-Euclidean, we are limited in the choice of data structure to use. In particular, *k*-d trees and their derivatives are not suitable for this metric. Instead, we use a variant of the Geometric Near-Neighbor Access Tree (GNAT) [53]. Each processor can then build a partial neighborhood for its own points (conformations) using a locally built tree as an index, and then ask other processors for neighbors of its points, merging the results into a list of *true k*-nearest neighbors for each point.

2. *Computing Geodesics.* The shortest paths between landmark points and all points can be computed from the neighborhood graph constructed in the first stage by using any graph search algorithm. Our implementation uses Dijkstra's algorithm [54] to do so, and since all processes have a copy of the neighborhood graph this can be done without the need for communication.

3. *Solving the MDS eigenvalue problem.* Each process has a fraction of the rows of the squared distance matrix between landmarks and all conformations. The eigenvector calculations are performed using P_ARPACK [55], a parallel version of the ARPACK library [56] based on iterative Arnoldi methods. From the eigenvectors and eigenvalues, the embedding coordinates can be computed on each processor.

D Proof for the Fast Lower-Bound on RMSD

Let $d(A,B)$ denote the squared RMSD between two conformations A and B. Conformations are represented by $3 \times n$ matrices, where n is the number of atoms. Column i of each matrix contains the position of atom i. Assume we can write $n = p \times q$, where p and q are integers. Suppose we create conformations A_c and B_c consisting of q "super-atoms" whose positions are obtained by averaging every p atoms. We can write

$A_c = AC$ and $B_c = BC$, where C is a $n \times q$ matrix of the following form:

$$C = \begin{pmatrix} 1/p & & & \\ \vdots & & & \\ 1/p & & & \\ & 1/p & & \\ & \vdots & & \\ & 1/p & & \\ & & \ddots & \\ & & & 1/p \\ & & & \vdots \\ & & & 1/p \end{pmatrix}. \tag{D.1}$$

Theorem 1. *The RMSD of the averaged conformations, $d(A_c, B_c)$, is a lower bound for the RMSD of the original conformations, $d(A, B)$.*

Proof. We can write $d(A,B)$ as $\frac{1}{n}\|A - RB\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm[‡‡], and R is the rotation matrix that minimizes this norm for given A and B. Similarly, let R_c be the rotation matrix that minimizes $\|A_c - R_c B_c\|$. It follows that

$$d(A_c, B_c) = \tfrac{1}{q}\|A_c - R_c B_c\|_F^2 \le \tfrac{1}{q}\|A_c - RB_c\|_F^2 = \tfrac{1}{q}\|(A - RB)C\|_F^2 = \tfrac{1}{q}\mathrm{trace}(XCC^T X^T), \tag{D.2}$$

where $X = A - RB$. The matrix CC^T is a block diagonal matrix, where each block has size $p \times p$ and all elements of each block are equal to $1/p^2$. We will show that each element of the diagonal of $\frac{1}{q}XCC^T X^T$ is smaller than the corresponding element of $\frac{1}{n}XX^T$. This is a sufficient condition for $d(A_c, B_c) \le d(A, B)$. Let $Y = CC^T X^T$. We will show that the first diagonal element of $\tilde{Z} = \frac{1}{q}XY$ is smaller than the first diagonal element of $Z = \frac{1}{n}XX^T$; the proof for the other elements is analogous. The first diagonal element of $\tilde{Z}$ can be written as

$$\tilde{z}_{11} = \tfrac{1}{q}\sum_i x_{1i} y_{i1} = \tfrac{1}{q}\big((x_{11} \ldots x_{1p})(y_{11} \ldots y_{p1})^T + (x_{1(p+1)} \ldots x_{1(2p)})(y_{(p+1)1} \ldots y_{(2p)1})^T + \ldots\big) \tag{D.3}$$

[‡‡] The Frobenius norm of a matrix M is defined as the square root of the sum of the squared elements of M: $\|M\|_F = \sqrt{\sum_{i,j} m_{ij}^2}$

The first diagonal element of Z can be written as

$$z_{11} = \frac{1}{n}\sum_i x_{1i}^2 = \frac{1}{q}\big((x_{11}\ldots x_{1p})(x_{11}\ldots x_{1p})^T + (x_{1(p+1)}\ldots x_{1(2p)})(x_{1(p+1)}\ldots x_{1(2p)})^T + \ldots\big) \tag{D.4}$$

To show that $\tilde{z}_{11} \leq z_{11}$, it is sufficient to show that each term on the right-hand side of equation D.3 is less than the corresponding term in equation D.4. We will show that this is true for the first terms; the proof for the remaining terms is analogous. We need to show that

$$\frac{1}{q}(x_{11}\ldots x_{1p})(y_{11}\ldots y_{p1})^T \leq \frac{1}{n}(x_{11}\ldots x_{1p})(x_{11}\ldots x_{1p})^T = \frac{1}{n}\sum_{i=1}^{p} x_{1i}^2 \tag{D.5}$$

By expanding CC^TX^T we see that $y_{11} = \ldots = y_{p1} = \frac{1}{p^2}\sum_{i=1}^{p} x_{1i}$. It follows that

$$\frac{1}{q}(x_{11}\ldots x_{1p})(y_{11}\ldots y_{p1})^T = \frac{1}{np}\big(\sum_{i=1}^{p} x_{1i}\big)^2 \leq \frac{1}{n}\sum_{i=1}^{p} x_{1i}^2. \tag{D.6}$$

The difference between the right side and left side of the inequality is proportional to the variance of $x_{1i},\ldots,x_{1p}$. This gives us an intuitive interpretation of the tightness of the bound. □