Spectral Techniques to Explore Point Clouds in Euclidean Space, with Applications to Collective Coordinates in Structural Biology

Frédéric Cazals, Frédéric Chazal, Joachim Giesen

Spectral Techniques to Explore Point Clouds in Euclidean Space, with Applications to Collective Coordinates in Structural Biology

F. Cazals*and F. Chazal †and J. Giesen ‡

January 2009

Abstract

Life sciences, engineering, or telecommunications provide numerous systems whose description requires a large number of variables. Developing insights into such systems, forecasting their evolution, or monitoring them is often based on the inference of correlations between these variables. Given a collection of points describing states of the system, questions such as inferring the effective number of independent parameters of the system (its intrinsic dimensionality) and the way these are coupled are paramount to develop models. In this context, this paper makes two contributions.

First, we review recent work on spectral techniques to organize point clouds in Euclidean space, with emphasis on the main difficulties faced. Second, after a careful presentation of the bio-physical context, we present applications of dimensionality reduction techniques to a core problem in structural biology, namely protein folding.

Both from the computer science and the structural biology perspective, we expect this survey to shed new light on the importance of *non linear computational geometry* in geometric data analysis in general, and for protein folding in particular.

Contents

*INRIA Sophia-Antipolis; Frederic.Cazals@inria.fr
†INRIA Saclay; Frederic.Chazal@inria.fr
‡MPI Saarbrucken; jgiesen@mpi-inf.mpg.de

1 Introduction

1.1 Geometric Data Analysis and Spectral Point Cloud Processing

Modeling the climate, understanding the interplay between proteins, metabolites and nucleic acids making up a regulation network within a cell, or unraveling the connexions between spiking neurons are example problems where a large number of variables interplay in a complex non linear way. Developing insights into such systems, forecasting their evolution, or monitoring them is often based on the inference of correlations between these variables. More precisely, learning such correlations from experiments is paramount to model development, as theory and experimental inference are tightly coupled.

Consider a complex system, and assume we are given a number of observations describing different states of the system. In such a setting, we are interested in the question of inferring the effective number of independent parameters of the system (its intrinsic dimensionality) and the way these are coupled. To meet these challenges, a set of new geometric methods, known as manifold learning, have been developed in the machine learning community mainly over the past decade. These methods are based upon the assumption that the observed data –a point cloud in some n dimensional space, lie on or are close to a submanifold M in $\mathbb{R}^d$.

Naturally, given the variety of situations, one cannot expect a single method to meet all needs. Nevertheless, many of the most popular approaches boil down to *spectral methods*. Note that the term *spectral method* is ambiguous and used differently within different communities, e.g., in numerical methods for partial differential equations it often involves the use of the fast Fourier transform. Here we want to use the term in the sense of data analysis similar as van der Maaten et al. did [25]. That is, for us in a spectral method, a symmetric matrix is derived from the point cloud data and the solution to a given optimization problem can be obtained from the eigenvectors of this matrix. We should mention that the term *spectral method* is also used in mesh processing in the geometric modeling community where the symmetric matrix is obtained from the connectivity of the mesh, see [31] for an overview. The geometric optimization problems that lead to a spectral technique are mostly of a *least squares* nature and include the following classical (and archetypical) problems:

(1) Find the k-dimensional subspace that approximates the point cloud best in a least squares sense.

(2) Find the embedding of the point cloud in k-dimensional space that preserves the distances between the points best possible in a least squares sense.

The first problem is called *principal component analysis (PCA)* as it asks for the principal directions (components) of the data. It essentially is a data quantization technique: every data point gets replaces by its projection onto the best approximating k-dimensional subspace. The loss incurred by the quantization is the variance of the data in the directions orthogonal the best approximating k-dimensional subspace. As long as this variance is small PCA can also be seen as *denoising* the original data. Many machine learning techniques including clustering, classification and semi-supervised learning [18], but also near neighbor indexing and search can benefit from such a denoising.

The second problem is called *multi-dimensional scaling (MDS)*. An important application of MDS is visualization of the point cloud data: the data points get embedded into two- or three-dimensional space, where they can be directly visualized. The main purpose of visualization is to use the human visual system to get insights into the structure of the point cloud data, e.g., the existence of clusters or—for data points labeled with discrete attributes—relations between this attributes. MDS visualization remains to be a popular tool for point cloud data analysis, but of course a lot of information will get lost (and in general cannot be restored by the human visual system) if the *intrinsic dimension* of the data points is larger than three.

Recently the focus in point cloud data analysis shifted: more emphasis is put on detecting non-linear features in the data, although processing the data for visual inspection still is important. What drives this shift in focus is the insight that most features are based on *local* correlations of the data points, but PCA and MDS both have only a global view on the point cloud data. The shift towards local correlations was pioneered by two techniques called *ISOMAP* [24, 12] and *Locally Linear Embedding (LLE)* [22, 23]. It is important to note that focusing on local correlations does not mean that one loses the global picture: for example the global intrinsic dimension of the data can be estimated from local information, whereas

it is often (when the data are embedded non-linearly) not possible to derive this information from a purely global analysis. ISOMAP and LLE and their successors (some of which we will also discuss here) can be used both for the traditional purposes data quantization and data visualization. In general they preserve more information of the data (than PCA and MDS) while achieving a similar quantization error or targeting the same embedding dimension for data visualization, respectively.

1.2 Spectral Methods and Alternatives

Advantages of spectral methods. Consider a point cloud P sampled from a manifold M embedded in $\mathbb{R}^d$. In this survey, we focus on a set of quite famous methods following a common thread, as they ultimately resort to spectral analysis. They all intend to find the best embedding of the dataset P into an Euclidean space $\mathbb{R}^k$ with respect to some quadratic constraint reflecting different geometric properties of the underlying manifold M. The embedding of the data that minimizes the quadratic constraint can then be interpreted as the best k-dimensional embedding of the data with respect to the geometric property we aim to preserve. In most cases, the quadratic minimization problem boils down to a general eigenvalue problem ensuring to find a global minimum. Moreover, the embedding can be found by easy-to-implement polynomial time algorithms.

This provides a substantial advantage over iterative or greedy methods based upon Expectation-Maximization like algorithms that do not provide guarantees of global optimality. In particular, for quite large data sets, the methods we consider still provide results when iterative and greedy methods fail due to complexity issues. Another advantage of "spectral methods" is that the quadratic constraint leads to a measurement of the quality of the embedding [1]. At last, "spectral methods" have been widely used and studied in many applications areas (graph theory, mesh processing [31],...) giving rise to a large amount of efficient theoretical and algorithmic tools that can be used for dimensionality reduction.

Approaches not covered. As our focus is on spectral techniques, a number of dimensionality reduction techniques are not covered in this paper. While the reader might consult [25] for a rather exhaustive catalog, the following comments are in order about the missing classes:

- EM-based methods: a large set of manifold learning algorithms developed in the machine learning community adopt a probabilistic point of view, so as to maximize a likelihood (Self Organizing Maps, Generative Topographic Mapping, Principal curves, etc. See [4] for example.). Some of them, like principal curves [17] or generative topographic mapping [5] for example, aim to fit the data set by a parameterized low dimensional (in general 1 or 2) manifold. They usually assume that the topology of the manifold is known and simple (simple curves, planes, discs) and do not allow to deal with data sampled from more complicated shapes.

- Methods related to the Johnson-Lindenstrauss lemma: the Johnson-Lindenstrauss lemma addresses the dimensionality reduction problem of a general point cloud (not necessarily sampled around a low dimensional manifold) in the perspective of preserving the pairwise distances between the points. An extension to points and flats and algebraic surfaces has been proposed in [1].

- Kernel methods: a number of methods, including some of the methods we shall discuss, can be interpreted in the framework of kernel methods. See [16, 27] for example.

- Methods targeting non manifold shapes: more recently, some geometric inference methods have been developed in the case where the shape underlying the data is not assumed to be a smooth manifold. They lead to promising but preliminary results for dimensionality reduction of general shapes [6, 7].

[1] For example, in [24], the quality of the embedding obtained is assessed resorting to the residual variance $\sigma_k(k,d)$ defined by:

$$\sigma_k(k,d) = 1 - R^2(\hat{D}_k, D_d) \tag{1}$$

with $R(\hat{D}_k, D_d)$ the correlation coefficient taken over all entries of matrices $\hat{D}_k$ and D_d. The closer to zero this variance, the better the approximation.

1.3 An Application in Structural Biology: Protein Folding

As an application of dimensionality reduction techniques in general, and of spectral methods in particular, we shall give a detailed account of one of the core open problems in structural biology, namely protein folding: how does a protein reach its folded-, i.e., its biologically active state, from the unfolded one? As of October 2007, about 1,000 genomes have been fully sequenced or are about to be so, while the Protein Data Bank contains (a mere) 40,000 structures. The question of understanding folding so as to predict the structure of a protein from its sequence is therefore central [2], to foster the understanding of central mechanisms in the cell, but also to perform protein engineering with applications ranging from drug design to bio-technologies.

Aside these general incentives, a number of technical ones advocate this particular problem.
First, the question of folding is closely related to a specific d-dimensional manifold which associates an energy to a conformation (the energy landscape), on which point clouds are sampled thanks to simulations techniques like the prototypical molecular dynamics method. Thus, the underlying mathematical structure is a manifold and not a (stratified) complex of arbitrary topology.
Second, assuming the folded and unfolded conformations correspond to (significant) local minima of the energy landscape, the problem is tantamount to understand *transitions* on this landscape, i.e. paths joining these minima. The difficulty of the problem is rooted in two facts: the high-dimensionality of the landscape ($d = 3n$ or $d = 6n$ as argued below, with n the number of atoms), and its complex topography which reflects the complex interactions (forces) between atoms. These intrinsic difficulties call for dimensionality reduction techniques, so as to exhibit a small number of new variables (typically one or two), called the reaction coordinates, accounting for the transition. These coordinates should match the effective *large amplitude - slow frequency* degrees of freedom of the system, thus providing a simplified view of the process, and easing its interpretation. Thus in essence, one wishes to quantize information located on a non linear manifold, while retaining the essential features.
Third, as opposed to a large number of multi-dimensional data sets, folding features a stimulating interplay between modeling and experiments. The point clouds studied in folding are indeed closely related to a number of experiments in bio-physics, so that one can precisely assess the quality and the interest of dimensionality reduction procedures. Example such experimental methods are dynamic NMR, protein engineering (ϕ-value analysis), laser initiated folding, etc. Describing these procedures is clearly beyond the scope of this survey, and the reader is referred to [57, 48] for starting pointers.

1.4 Notations and Paper Overview

Throughout this paper we will be using the following notations:

P	point cloud
n	number of point is P
d	dimension of the Euclidean space form which the points in P are drawn
k	target dimension

The paper is organized as follows. Section 2 presents the two archetypical spectral methods used to explore point clouds, namely PCA and MDS. The question of localizing neighborhoods is discussed in section 3, while methods meant to accommodate non linear geometries are discussed in section 4. The application of dimensionality reduction techniques to protein folding is discussed in section 5. To conclude, section 6 discusses a number of research challenges.

2 PCA and MDS

In the following we assume that the points in P are centered at the origin, i.e., $\sum_{i=1}^{n} p_i = 0$. Note that this can always achieved by a simple translation: let $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$ and $p'_i = p_i - \bar{p}$, then $\sum_{i=1}^{n} p'_i = 0$.

Principal component analysis (PCA) asks for the k-dimensional subspace of $\mathbb{R}^d$ that approximates the point set P best possible in a least squares sense and projects P onto that subspace, whereas multi-dimensional scaling (MDS) in its basic form aims for the k-dimensional embedding of P that preserves

[2] At least for proteins consisting of a single polypeptidic chain, as the formation of multimers also poses docking questions.

the pairwise inner products of the points in P best possible in a least squares sense. In both cases k can range from 1 to $d-1$.

Though different in their motivation and objective, PCA and MDS are almost identical in a technical sense: both can be formulated in terms of eigenvectors of some positive semi-definite matrix derived from the point set P, which itself can be written as a $(d \times n)$-matrix as follows:

$$P = \begin{pmatrix} p_{11} & \cdots & p_{n1} \\ \vdots & & \vdots \\ p_{1d} & \cdots & p_{nd} \end{pmatrix},$$

where p_{ij} is the j'th component of the point $p_i \in P$. From the matrix P one canonically derives two positive semi-definite matrices,

(1) the *covariance matrix* $C = PP^T$, and

(2) the *Gram matrix* $G = P^T P$.

The covariance matrix is a $(d \times d)$-matrix and can also be written as $C = \sum_{i=1}^{n} p_i p_i^T$, whereas the Gram matrix has dimension $n \times n$ and can also be written as $G = \left(p_i^T p_j\right)$. Both matrices are intimately linked also via their eigenvectors and eigenvalues. We have the following observation.

Observation 1 *The matrices C and G have the same non-zero (positive) eigenvalues (and thus the same rank).*

Proof. Let $v \in \mathbb{R}^d$ be an eigenvector of C with eigenvalue $\lambda > 0$, then $P^T v$ is an eigenvector of G also with eigenvalue λ as can be seen from the following simple calculation:

$$GP^T v = P^T PP^T v = P^T Cv = \lambda P^T v.$$

Similarly, if $u \in \mathbb{R}^n$ is an eigenvector of G with eigenvalue $\mu > 0$, then Pu is an eigenvector of C with eigenvalue μ. □

One important issue with both PCA and MDS is how to choose/determine k (the intrinsic dimensionality of the point cloud data). Sometimes there is a "large" gap in the eigenvalue spectrum of C or G, respectively, and k is then often chosen as the number of eigenvalues above this gap.

2.1 PCA

As mentioned earlier PCA asks for the k-dimensional subspace of $\mathbb{R}^d$ that approximates the point set P best possible in a least squares sense. Let us discuss this for the case $k = d-1$ first. In this case we are looking for a unit vector $v \in \mathbb{R}^d$ such that the sum of the squared lengths of the projections $(v^T p_i)v$ is minimized. Formally this can be written as

$$\begin{array}{ll} \min & v^T PP^T v \\ \text{s.t.} & \|v\|^2 = 1 \end{array}$$

From the Lagrange multiplier theorem one derives the following condition for an optimal solution to this optimization problem: $\lambda v = PP^T v = Cv$. That is, an optimal solution is the subspace orthogonal to an eigenvector of the covariance matrix C and the value of the optimization problem at an optimal solution is $v^T PP^T v = v^T Cv = \lambda \|v\|^2 = \lambda$. Hence we are looking for an eigenvector associated to the smallest eigenvalue of C and the optimal solution is spanned by all eigenvectors of the covariance matrix C except the one corresponding to the smallest eigenvalue.

Let $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ be the eigenvalues of C, $v_1, \ldots, v_d \in \mathbb{R}^d$ a corresponding orthonormal eigenbasis and $P_k = \sum_{i=1}^{k} v_i v_i^T, k = 1, \ldots, d$, the projector on the k'th invariant eigenspace, i.e., the eigenspace spanned by the first k eigenvectors. Iteratively it follows that the best approximating k-dimensional subspace of $\mathbb{R}^d$ in a least square sense is spanned by $v_1, \ldots, v_k$. The k'th order PCA is then given as the following transformation:

$$p_i \mapsto P_k p_i = p_i - (\mathbb{I} - P_k)p_i.$$

In a way $P_k p_i$ is seen as the signal conveyed with the point p_i and $(\mathbb{I} - P_k)p_i$ is seen as noise.

2.2 MDS

Multi-dimensional scaling is aiming for a k-dimensional embedding of P that preserves the pairwise inner products $p_i^T p_j$ as well as possible in a least squares sense[3]. Note that all inner products are stored as entries in the Gram matrix G. Let $\mu_1 \geq \ldots \geq \mu_n \geq 0$ be the eigenvalues of G, let $u_1, \ldots, u_n \in \mathbb{R}^n$ be a corresponding orthonormal eigenbasis and let $Q_k = \sum_{i=1}^{k} u_i u_i^T$, for $k = 1, \ldots, n$, be the projector on the k'th invariant eigenspace. We have the following observation:

Observation 2 *The matrix $Q_k G$ is the best rank k approximation of the Gram matrix G in the sense that*

$$\|Q_k G - G\|_2 = argmin_{Q:\,(n\times n)\text{-}matrix\ of\ rank\ k} \|QG - G\|_2.$$

The matrix $Q_k G$ can also be interpreted as a matrix of inner products. To see this we use (a) the projector property $Q_k^2 = Q_k$, (b) symmetry $Q_k^T = Q_k$, and (c) the commutator property $Q_k G = G Q_k$, and get

$$Q_k G = Q_k^2 G = Q_k G Q_k = Q_k P^T P Q_k = Q_k^T P^T P Q_k = (P Q_k)^T P Q_k,$$

which shows that $Q_k G$ is the matrix of inner products of the columns of $PQ_k = (Q_k P^T)^T$. Here the $(n \times d)$-matrix $Q_k P^T$ is the projection of the rows of P onto the space spanned by $u_1, \ldots, u_n$. The k'th order MDS maps the point p_i to the i'th column $Q_k P^T$, i.e., the i'th column of PQ_k. This column is uniquely specified by its coefficients $\alpha_1^i, \ldots, \alpha_k^i$ in the orthonormal basis $u_1, \ldots, u_k$. Representing the points p_i by $(\alpha_1^i, \ldots, \alpha_k^i)$ gives the thought for least squares optimal k-dimensional embedding of the point set P.

3 Localization

3.1 Neighborhood Criteria

In using PCA and MDS, feature preserving data quantization and visualization can be enhanced by taking only local relations among all the data points into account. Localization the relations means choosing neighborhoods for each data point, i.e., building a (in general directed) neighborhood graph on the data points. The right choice of neighborhood is crucial for the localized version of PCA and MDS to work properly. Commonly used methods to define the neighborhoods are:

(1) κ nearest neighbors: connect every p_i to its κ nearest neighbors (in terms of Euclidean distance) in P.

(2) symmetric κ nearest neighbors: connect p_i to its κ nearest neighbors and all points in this neighborhood to each other.

(3) fixed neighborhood: given $\varepsilon > 0$, connect every p_i to all points in P that have distance less than ε to p_i.

(4) relative neighborhood: given $\rho > 1$, connect every p_i to all neighbors at distance ρ times the distance of p_i to its nearest neighbor.

An important observation is that (1) and (2), i.e., κ nearest neighbors and symmetric κ nearest neighbors, respectively, do not automatically adapt to the intrinsic dimension of the point cloud data. Intuitively, if the intrinsic dimension is large also κ needs to be large in order to cover a meaningful neighborhood for a data point (we expect this neighborhood to grow exponentially in the intrinsic dimension), whereas if the intrinsic dimension is small, for the same value of κ one already covers data points quite far away. Methods (3) and (4), fixed- and relative neighborhood, both automatically adapt to the intrinsic dimension, but cannot—in contrast to the κ nearest neighbor methods—adapt to non-uniform or anisotropic spacing of the data points. In practice a good choice for the value of the parameter ρ of (4) may be easier to find than for the value of ε in (3).

More neighborhood graphs are discussed by Yang [30] who also provides experimental results.

[3]Observe that completely preserving the inner products allows us to recover P up to a rotation, i.e., completely preserving the pairwise inner products also preserves the pairwise distances $\|p_i - p_j\|$.

3.2 Dimension Detection Using PCA

We have seen above that knowing the local dimension at a data point can guide the right choice of parameter κ when computing the κ nearest neighbors neighborhood. On the other hand, using that given $p \in M$, there exists a small neighborhood of p in which M is close to its tangent space at p, it is appealing to use localized versions of PCA to infer the local intrinsic dimension of M at p from the point cloud data P. With a good neighborhood $N(p) \subset P$ of $p \in P$ one can estimate the intrinsic dimension at p by a localized version of PCA. The localized version uses the local covariance matrix C_p of the points

$$p_i' = (p_i - p) - \frac{1}{n} \sum_{p_i \in N(p)} (p_i - p) \text{ for } p_i \in N(p).$$

Intuitively, if the local dimension at p is k, then we expect a gap in the eigenvalue spectrum of C_p in the sense that k'th largest eigenvalue is much larger than the $(k+1)$'st eigenvalue and the k largest eigenvalues are roughly of the same magnitude. That is, we expect a threshold θ such that

$$\frac{\lambda_j}{\lambda_1} \geq \theta \text{ for } j \leq k \text{ and } \frac{\lambda_j}{\lambda_1} \leq \theta \text{ for } j > k.$$

Indeed, Cheng, Wang and Wu [8] were able to prove the existence of such a threshold θ under the assumption that the data are sampled from a smooth manifold and obey a sampling condition. The sampling condition rules out locally non-uniform or anisotropic spacing of the sample points. Under this assumption fixed- and relative neighborhoods should work. Cheng et. al use the relative neighborhood for their proof. Though their threshold parameter θ depends on parameters of the sampling condition they report good results in practice using a threshold of $\theta = 1/4$.

It is important to remark that when the sampling conditions are not fullfilled or when the size of the neighborhoods are not well-choosen, the previous method usually leads to unclear and confusing estimations. In particular the dimension estimation may depend on a "scale" (in the previous case the size of the neighborhoods) at which the data is considered: assume that P samples a planar spiral with gaussian noise in the normal direction to the curve. At a "microscopic" scale, P just looks like a finite set of points and its dimension is 0. At a scale of the size of the standard deviation of the noise, P seems to locally sample the ambient space and the localized PCA method will probably estimate M to be 2-dimensional. At a higher, but not too big, scale the localized PCA will provide the right estimation and at large scales, it will again provide a 2-dimensional estimation. Various notions of dimension (q-dimension, capacity dimension, correlation dimension, etc...) have been introduced to define the intrinsic dimension of general shapes (including non smooth shapes and fractal sets). They give rise to algorithmically simple methods that simultaneously provide dimension estimations at different explicit scales allowing the user to select the one which is most relevant for his purpose. An introduction to this subject may be found in [20].

4 Turning Non-Linear

The linear and global aspects of PCA and MDS make them inefficient when the underlying manifold M is *highly non linear*. Designing non-linear dimensionality reduction methods that lead to good results for non linear smooth manifolds is an active research area that gave rise to a big amount of literature during the last decade. In this section, we quickly present a set of quite famous dimension reduction methods that take advantage of the localization techniques presented in the previous section and that have interesting geometric interpretations. They also have the advantage of leading to easy to implement polynomial time algorithms that prove more efficient with larger data sets than the ones usually involved in iterative or greedy methods (like e.g. the ones involving EM or EM-like algorithms). We also discuss the guarantees provided by these methods.

Recall that in the following the considered data sets $P \subset \mathbb{R}^d$ are assumed to be sampled on/around a possibly unknown smooth manifold M of dimension k. The common thread of the few methods presented below is that they all aim to find a projection $\hat{P} \subset \mathbb{R}^k$ of the data set minimizing a quadratic functional $\phi(\hat{P})$ that intends to preserve (local) neighborhood information between the sample points.

4.1 Maximum Variance Unfolding (MVU)

PCA and MDS perform poorly when data points are not close to an affine subspace, i.e., they are both based on an inherent linearity assumption. Especially, both methods fail when the data points are close to a "curled up" linear space—the most famous example is the so called Swiss roll data set, points sampled densely from a curled up planar rectangle in $\mathbb{R}^3$. The idea behind *maximum variance unfolding (MVU)*, introduced by Weinberger and Saul [26, 29, 28], is to unfold the data, i.e., to transform the data set to a locally isometric data set, that is closer to an affine subspace. The unfolding aims at maximizing the distance between non-neighboring points (after some choice of neighborhood) while preserving the distances between neighboring points.

Technically MVU proceeds as follows: let $D = (d_{ij} = \|p\_p_j\|^2)$ be the symmetric $(n \times n)$-matrix of pairwise distances. Choose a suited neighborhood for each point in P (Weinberger and Saul choose the symmetric κ-nearest neighbors) and let the indicator variable n_{ij} be 1 if either p_i is in the neighborhood of p_j or p_j is in the neighborhood of p_i, and 0 otherwise. From D an *unfolding*, a positive semi-definite $(n \times n)$-matrix $K = (k_{ij})$ (interpreted as the Gram matrix of the unfolded point set) is computed through the following semi-definite program (SDP)

Maximize the trace of K subject to

(1) K is positive semi-definite

(2)
$$\sum_{i,j=1}^{n} k_{ij} = 0$$

(3)
$$k_{ii} - 2k_{ij} + k_{jj} = d_{ij} \text{ for all } (i,j) \text{ with } n_{ij} = 1$$

From K a lower dimensional embedding can be computed as described for MDS.

4.2 Locally Linear Embedding (LLE)

LLE is a method introduced in [22, 23] that intends to take into account the local linearity of the underlying manifold M to perform the reduction of dimension. In a first step, LLE discards pairwise distances between widely separated points by building a neighborhood graph G (see Section 3). The goal of this first step is to connect only close points of P so that the neighbors of each vertex p_i in G are contained in a small neighborhood of p_i which is close to the tangent space of the underlying manifold M at p_i. To take this local linearity into account, LLE computes for each vertex p_i of the graph its best approximation as a linear combination of its neighbors. More precisely, one computes a sparse matrix of weights $W_{i,j}$ that minimize the quadratic error

$$\varepsilon(W) = \sum_{i=1}^{n} \|p_i - \sum_{j \in N(p_i)} W_{i,j} p_j\|^2$$

where $N(p_i)$ is the set of the vertices that are connected to p_i in G. This is a simple least square problem. Solving it with the additional constraint

$$\forall i, \sum_{j \in Ngb(p_i)} W_{i,j} = 1$$

makes the weights invariant to rescaling, rotations and translations of the data (the weights thus characterize intrinsic properties of the data). The weights matrix is then used to perform the dimensionality reduction: given $k < d$, the points p_i are mapped to the points $\hat{p}_i \in \mathbb{R}^k$ that minimize the quadratic function

$$\Phi(\hat{p}_i) = \sum_i \|\hat{p}_i - \sum_j W_{i,j} \hat{p}_j\|^2$$

This quadratic minimization problem classically reduces to solving a sparse $n \times n$ eigenvalue problem. As for MDS, the LLE algorithm projects the data in a low dimensional space, no matter what the mapping is. To provide satisfactory result, the data have to be sufficiently dense to insure that the neighbors of a given point provide a good approximation of the tangent space of M. Moreover, even if the data are dense enough, the choice of the neighbors may also be awkward: choosing a too small or too large neighborhood may lead to very bad estimates of the tangent space.

4.3 ISOMAP

ISOMAP is a version of MDS introduced in [24, 12], where the matrix of inner products or Euclidean distances, respectively, is replaced by the matrix of the geodesic distances between data points on M. In a first step, ISOMAP builds a neighborhood graph such that the distances between points of P in the graph are close to the geodesic distances on M. Once the geodesic distance matrix has been built, ISOMAP proceeds like classical MDS to project P in $\mathbb{R}^k$.
One of the advantage of ISOMAP is that it provides convergence guarantees. First, it can be proven that if the data are sufficiently densely sampled on M, the distance on the neighbor graph is close to the one on M [11, 21, 15]. Nevertheless, in practice robust estimation of geodesic distances on a manifold is an awkward problem that requires rather restrictive assumptions on the sampling. Second, since the MDS step in the ISOMAP algorithm intends to preserve the geodesic distances between points, it provides a correct embedding if M is isometric to a convex open set of $\mathbb{R}^k$. The convexity constraint comes from the following remark: if M is an open subset of $\mathbb{R}^k$ which is not convex, then there exist a pair of points that cannot be joined by a straight line contained in M. As a consequence, their geodesic distance cannot be equal to the Euclidean distance. It appears that ISOMAP is not well-suited to deal with data on manifolds M that do not fulfill this hypothesis. Nevertheless some variants (conformal ISOMAP [12]) have been proposed to overcome this issue. Note also that ISOMAP is a non local method since all geodesic distances between pairs of points are taken into account. As a consequence ISOMAP involves a non-sparse eigenvalue problem which is a main drawback of this method. To partly overcome this difficulty some variant of the algorithm using landmarks have been proposed in [12].

4.4 Laplacian Eigenmaps

This method introduced in [3, 2] follows the following general scheme: first a weighted graph G with weights $W_{i,j}$ is built from the data. Here the weights measure closeness between the points: intuitively the bigger $W_{i,j}$ is, the closer p_i and p_j are. A classical choice for the weights is given by the Gaussian kernel $W_{i,j} = \exp(-\frac{\|p_i - p_j\|^2}{4\sigma})$, where σ is a user-defined parameter [4]. Second the graph G is embedded into $\mathbb{R}^k$ in such a way that the close connected points stay as close as possible. More precisely the points p_i are mapped to the points $\hat{p}_i \in \mathbb{R}^k$ that minimize

$$\phi(\hat{P}) = \sum_{i,j} \|\hat{p}_i - \hat{p}_j\|^2 W_{i,j}.$$

There is an interesting and fundamental analogy between this discrete minimization problem on the graph G and a continuous minimization problem on M. Indeed, it can be seen that minimizing ϕ on the functions defined on the vertices of G corresponds (in a discretized version) to minimizing $\int_M \|\nabla f\|^2$ on the space of functions f defined on M with L^2 norm $\|f\|_{L^2}^2 = \int_M \|f\|^2 = 1$. From the Stokes formula, this integral is equal to $\int_M \mathcal{L}(f)f$, where $\mathcal{L}$ is the Laplace-Beltrami operator on M and its minimum is realized for eigenfunctions of $\mathcal{L}$. Similarly the minimization problem on G boils down to a general eigenvector problem involving the Laplacian of the graph. Indeed the Laplace operator on G is the matrix $L = D - W$, where D is the diagonal matrix $D_{i,i} = \sum_j W_{i,j}$. It can be seen as an operator acting on the functions f defined on the vertices of G by subtracting from $f(p_i)$ the weighted mean value of f on the neighbors of p_i. By a classical computation, one can see that $\phi(\hat{P}) = tr(\hat{P}^T L \hat{P})$, where $\hat{P}$ is the $n \times k$ matrix with i-th row given by the coordinates of $\hat{p}_i$. It follows that, given $k > 0$, the minimum of ϕ is deduced from the computation of the $k+1$ smallest eigenvalues of the equation $Ly = \lambda Dy$ (the

[4] To obtain a sparse matrix W the values of $W_{i,j}$ that are smaller than some fixed small threshold are usually set to 0.

smallest one corresponding to the eigenvalue 0 has to be removed). The analogy between the discrete and continuous setting extends to the choice of the weights of G: choosing $W_{i,j} = \exp(-\frac{\|p_i - p_j\|^2}{4\sigma})$, where σ is a user-defined parameter, allows to interpret the weights as a discretization of the heat kernel on M [3]. From the side of the guarantees, the Laplacian eigenmaps only involve intrinsic properties of G so they are robust to isometric perturbations of the data. Moreover, the relationship with the Laplacian operator on M provides a framework leading to convergence results of L to the Laplace operator on M [2].

4.5 Hessian Eigenmaps (HLLE)

ISOMAP provides guarantees when the unknown manifold M is isometric to a convex open subset of $\mathbb{R}^k$. Although the hypothesis of being isometric to an open subset of $\mathbb{R}^k$ seems to be rather reasonable in several practical applications, the convexity hypothesis appears to be often too restrictive. HLLE is a method introduced in [14] intending to overcome this convexity constraint. The motivation of HLLE comes from a rather elementary result stating that if M is isometric to a connected open subset of $\mathbb{R}^k$ then the null-space of the operator defined on the space of $\mathcal{C}^2$-functions on M by

$$\mathcal{H} : f \to \int_M \|Hessf(m)\|^2 dm$$

where $Hessf$ is the Hessian of f, is a $(k+1)$-dimensional space spanned by the constant functions and the "isometric coordinates" of M. More precisely, if there exists an open set U in $\mathbb{R}^k$ and an isometric embedding $\phi : M \to U$ then it can be proven that the constant functions and the functions $\phi_1, \cdots \phi_k$, where ϕ_i is the i-th coordinate of the map ϕ, are contained in the null-space of $\mathcal{H}$. Moreover, the constant functions span one dimension of this null-space and the k functions ϕ_i span the k other dimensions. It is thus appealing to estimate this null space in order to recover these isometric coordinates to map M isometrically on an open subset of $\mathbb{R}^k$. To do this the algorithm follows the same scheme as LLE and the estimation of the null-space of $\mathcal{H}$ reduces to an eigenvalue computation of a sparse $n \times n$ matrix. As a consequence HLLE allows to process dimensionality reduction for a larger class of manifolds M than ISOMAP. The quality of the reduction is obviously closely related to the quality of the approximation of the kernel of the operator $\mathcal{H}$. Nevertheless, it is important to notice that the algorithm involves the estimation of second order differential quantities for the computation of the Hessian while LLE requires only first order ones to approximate the tangent space of M. To be done efficiently this usually needs a very dense sampling of M. At last, note that HLLE is the same as Laplacian Eigenmaps where the Laplacian operator has been replaced by $\mathcal{H}$.

4.6 Diffusion Maps

Diffusion maps [9] provide a method for dimensionality reduction based upon Markov random walks on a weighted graph G reflecting the local geometry of P . The graph G is built in a similar way as for Laplacian Eigenmaps: the larger is the weight of an edge, the "closer" are its endpoints. In particular G can be built using the discretization of the heat kernel on M (see section 4.4). From the weights matrix W one constructs a Markov transition matrix Π by normalizing the rows of W

$$\Pi_{i,j} = \frac{W_{i,j}}{d(p_i)} \quad \text{where} \quad d(p_i) = \sum_k W_{i,k} \quad \text{is the degree of the vertex} \quad p_i$$

$\Pi_{i,j}$ can be interpreted as the probability of transition from p_i to p_j in one time step. The term $\Pi_t(i,j)$ of the successive powers Π^t of Π represent the probability $\Pi_t(p_i, p_j)$ of going from p_i to p_j in t steps. The matrix Π can be seen as an operator acting on the probability distributions supported on the vertices of G. It admits an invariant distribution ϕ_0 defined by $\phi_0(p_i) = \frac{d(p_i)}{\sum_j d(p_j)}$. The idea of diffusion maps is thus to define a metric between the points of P which is such that at a given $t > 0$ two points p_i and p_j are close if the conditional distributions of probability $\Pi_t(p_i, .)$ and $\Pi_t(p_j, .)$ are close. The choice of a weighted L^2 metric between the conditional distributions allows to define a *diffusion metric* between the

points of P

$$D_t^2(p_i, p_j) = \sum_k \frac{(\Pi_t(p_i, p_k) - \Pi_t(p_j, p_k))^2}{\phi_0(p_k)}$$

which is closely related to the spectral properties of the random walk on G given by Π. Intuitively, two points p_i and p_j are close if there are many paths connecting them in G as illustrated on Fig. 1. Note that the parameter t representing the "duration" of the diffusion process may be interpreted as a scale parameter in the analysis. Given k and $t > 0$, the *diffusion map* provides a parameterization and a projection of the data set which performs a dimensionality reduction that minimizes the distortion between the Euclidean distance in $\mathbb{R}^k$ and the diffusion distance D_t. The diffusion map is obtained from the eigenvectors of the transition matrix Π and the eigenvalues to the power t of the transition matrix. The diffusion maps framework reveals deep connections with other areas (such as spectral clustering, spectral analysis on manifolds,...) that open many questions and make it an active research area. For a more detailed presentation of diffusion maps and its further developments the reader is referred to [9, 10, 19].

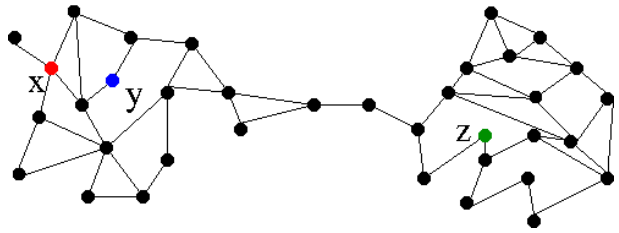

Figure 1: An example of a graph G (the weights are given by the heat kernel approximation, see text) with points that are close or far to each other with respect to the diffusion metric: the points x and y are close to each other while the points x and z are far away because G is "pinched" between the two parts containing x and z. So there are few paths connecting x to z.

5 Applications in Structural Biology: the Folding Problem

In this section, we first recall the intrinsic difficulty of folding proteins on a computer –section 5.1, and bridge the gap between folding and dimensionality reduction –section 5.2. We then proceed with a detailed account of the bio-physical context by discussing the question of cooperative motions within a protein –section 5.3, and make the connexion to Morse theory and singularity theory along the way. Finally, we review techniques to derive meaningful so-called reaction coordinates –section 5.4.

5.1 Folding: from Experiments to Modeling

Anfinsen was awarded the 1972 Nobel prize in chemistry *for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation* [5]. Since then, Anfinsen's dogma states that for (small globular) proteins, the sequence of amino-acids contains the information that allows the protein to fold i.e. to adopts its (essentially unique) native structure, or phrased differently, the 3d structure that accounts for its function. At room temperature, the folding of a protein typically requires from millisecond to seconds, while the time-scale of the finest (Newtonian) physical phenomena involved is the femtosecond.

When compared to femtoseconds, folding times are rather slow, which points towards a process more complex than a mere descent towards a minimum of energy. On the other hand, such folding times are definitely too fast to be compatible with a uniform exploration of an exponential number conformations [6]. This observation is known as Levinthal's paradox [63], and scales the difficulty of folding from a computational perspective.

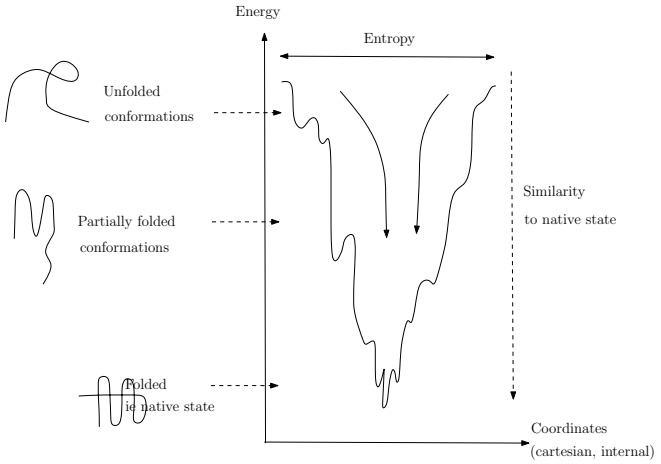

Figure 2: Folding funnel.The variability of conformations encodes the entropy of the system, while its energy level encodes the proximity to nativeness. Adapted from [40].

[5] See `http://nobelprize.org/nobel_prizes/chemistry/laureates/1972/index.html`

[6] Recall that the side-chains of the amino-acids take conformations within finite sets –the so-called rotamers [59, 46], whence a priori an exponential number of conformations.

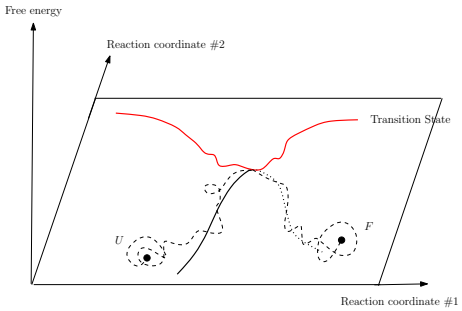

Figure 3: Crossing the Transition State on a rugged energy landscape: the system moves from one watershed (state U) to a neighboring watershed (state F) by crossing the energy landscape pass. Adapted from [42].

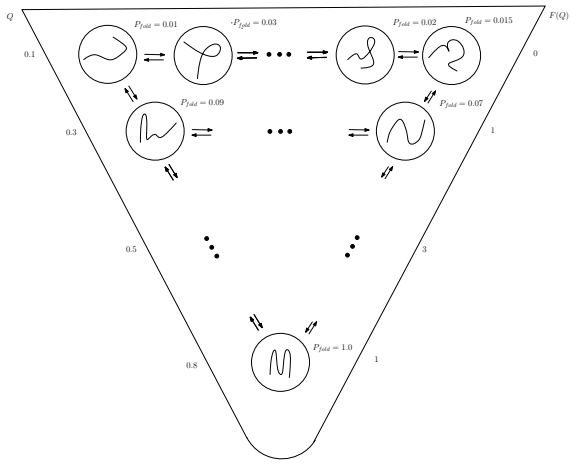

Figure 4: Folding process down a folding funnel: fraction of native contacts Q increases, free energy $F(Q)$ crosses a barrier, p_{fold} increases. Adapted From [42].

5.2 Energy Landscapes and Dimensionality Reduction

5.2.1 Potential and Free Energy Landscape

Consider a system consisting of a protein and the surrounding solvent, for a total of n atoms. Each atom is described by 3 parameters for the position and 3 for its velocity (momentum). In the following, depending on the context, we shall be interested in a parameter space of dimension $d = 3n$ (positions) or $d = 6n$ (positions+velocities), the latter being called the *phase* space. As the system is invariant upon rigid motions, one could work with $d - 6$ degrees of freedom, but we skip this subtlety in the following. From this d-dimensional parameter space, one defines the *the energy landscape* [70], i.e. the d dimensional manifold obtained by associating to each conformation of the system an energy (potential energy or free energy [7]). Since the water molecules are critical to model appropriately the electrostatic interactions, n typically lies in the range 10^4 to 10^5 for a system consisting of a protein and its aqueous environment.

5.2.2 Enthalpy-Entropy Compensation, Energy Funnel, Ruggedness and Frustration

Energy funnels. Folding may be seen as the process driving a heterogeneous ensemble of conformations populating the unfolded state to a homogeneous ensemble of conformations populating the folded or native state. To intuitively capture one major subtlety of folding, it is instructive to examine the variation of the enthalpy H and entropy S of the system *protein+solvent*. While the protein folds, more native contacts

[7] As will be shown with Eq. (2), a free energy landscape is obtained from the potential energy landscape by projecting onto selected coordinates.

between atoms get formed, whence an enthalpy decrease. On the other hand, two phenomena account for an entropic drop down: first, the conformational variability of the protein decreases; second, the structure of the solvent around the protein changes. This latter re-organization, known as the hydrophobic effect, corresponds to the fact that water molecules line-up along the hydrophobic wall formed by the molecular surface of the protein. Overall, the variations of the enthalpy and entropy almost cancel out, resulting in a small variation of the free energy $G = H - TS$ of the system. This phenomenon is known as the *enthalpy-entropy compensation* [48], and can be illustrated using energy landscapes, as seen from Fig. 2. On this figure, the vertical axis features the free energy, and the horizontal one the entropy: while the folding process progresses, the free energy (slightly) decreases and the landscape becomes narrower—the entropy decreases. Such a landscape is generally called a *folding funnel [40].*

While the previous discussion provides a thermodynamic overview of the folding process, Levinthal's paradox deals with a kinetic problem—*how come the folding process is so fast?* Travelling down the folding funnel [8] provides an intuitive explanation: the protein is driven towards the minimum of energy corresponding to the native state by a steep gradient along the energy surface. This intuitive simplified view, however, must be amended in several directions.

Ruggedness and frustration. Two important concepts which help to describe landscapes are *ruggedness* or *roughness* and *frustration*. Ruggedness refers to the presence of local minima, which in a folding process may correspond to partially folded states. Frustration refers to the presence of several equally deep minima separated by significant barriers, which may prevent the system from reaching the deepest one. The fact that most proteins seem to have a single native state [9] seems to advocate a minimal frustration principle. Yet, even for non frustrated landscapes, several levels of *ruggedness* may exist. In particular, on the easy side of the spectrum, one finds proteins folding with a two-states kinetics, i.e. without any intermediates [58].

Ruggedness / frustration may actually come from two sources, namely from the interaction energy between atoms of the protein, and/or from the conformational entropy [41]. The enthalpic frustration comes from local minima of the interaction potential energy. For the entropic frustration, observe that the folding process is accompanied by a loss of conformational entropy (of the protein). If this loss is heterogeneous and larger than the energetic heterogeneity, the corresponding free energetic landscape becomes frustrated.

5.2.3 Cooperativity and Correlated Motions

Another concept related to minimally frustrated folding funnels is that of *cooperative* motions between atoms. Cooperativity stipulates that when one atom is moving, atoms nearby must move in a coherent fashion. This is rather intuitive for condensed states where local forces (repulsion forces as atoms cannot inter-penetrate, hydrogen bonding) are prominent. At a more global scale, cooperation is likely to also be important, e.g. due to electrostatic interactions. From a technical point of view, simple illustrations of correlated motions are provided by normal modes studies [10], as well as correlations between positional fluctuations [11].

Having mentioned correlated motions of atoms, the fact that dimensionality reduction techniques play a key role in modeling folding (and more generally the behavior of macro-molecular systems) is expected. First, the d degrees of freedom are certainly not equivalent, as different time-scales are clearly involved: from small amplitude - high frequency vibrations apart from chemical bonds, to large amplitude - slow frequency deformations of the protein. Second, the constraints inherent to the large amplitude motions

[8] The fact that the kinetic pathway follows the thermodynamic one is non trivial, and in general unwarranted, see [48, Chapter 19].

[9] As opposed to many polymers which exist under a number of energetically equivalent inter-convertible states.

[10] Assuming the system is at a minimum of its potential energy V, the dominant term in the Taylor series expansion of V is the quadratic one. Diagonalizing the corresponding quadratic form yields the so-called normal modes, whose associated eigenvectors are collective coordinates. See for example [69].

[11] Given a molecular dynamics simulation, one may investigate the correlations between the atomic fluctuations —with respect to a reference conformation. Both PCA and MDS have been used for this problem: in [51, 32], the average covariance matrix of the positional fluctuations is resorted to, while [56] computes the average Gram matrix. See also [62] for a characterization of pairwise atomic correlations based of Pearson's coefficients and relatives.

are such that one expects the *effective* parameters to lie on some lower-dimensional manifold representing the *effective* energy landscape, that is the one accounting for transitions.

5.3 Bio-physics: Pre-requisites

5.3.1 Molecular Dynamics Simulations

The simulation data we shall be concerned with are essentially molecular dynamics (MD) data. (The reader is referred to [50] for alternate simulation methods, such as Monte Carlo simulations or Langevin dynamics.) A MD simulation is a deterministic process which evolves a system according to the Newtonian laws of motion. Central in the process is the force field associated to the system, or equivalently the potential energy stemming from the interactions between atoms. A typical potential energy involves bonded terms (energies associated to covalent bonds), and non bonded terms (Van der Waals interactions and electrostatic interactions). From the potential energy V associated to two atoms, one derives an associated force. Given these forces, together with the positions and momenta of the atoms, one determines the configuration of the system at time $t + \Delta t$. Practically, Δt is of the order of the femtosecond, so that in retaining one conformation every 10, long simulations (beyond the nanosecond) result in a number of conformations $> 100,000$.

5.3.2 Models, Potential Energy Landscapes and their Ruggedness

As exploring exhaustively the energy landscape of large atomic models is not possible, a number of coarse models mimicking the properties of all atoms models have been developed. We may cite the united residues model [47]; the BLN model [36] which features three types of beads only (hydrophilic, hydrophobic, neutral); the 20 colors beads model [44], which accommodates anisotropic interactions between residues so as to maximize packing of side chains.

Such coarse models deserve a comment about the ruggedness of potential energy landscapes. Ruggedness and frustration are indeed clearly related to the complexity of the force field governing the system, since non local interactions between atoms are likely to yield local minima of the landscape —cf the $G\bar{o}$ models thereafter. On the other hand, non local interactions are likely to help the protein to overcome local energy barriers (to escape the local minima of the rugged landscape) due to solvent collisions, non-native contacts, etc. See for example [65].

Having mentioned energy landscapes and MD simulations, a crucial remark is in order. Following the gradient vector field of the energy on the potential energy surface amounts to a mere energy minimization. But MD simulations are more powerful, since a system evolved by a MD can cross energy barriers thanks to its kinetic energy [12]. Another way to cross barriers is to resort to a Monte Carlo simulation [50].

5.3.3 Morse Theory and Singularity Theory

As outlined above, the properties of a system are described by its energy landscape. To investigate transitions of our macro-molecular system, the topographical features of the landscape i.e. its minima, maxima, and passes are of utmost importance [70]. These features are best described in terms of Morse theory [64] as well as singularity theory [13], which in our setting amounts to studying the gradient vector of the energy function on the manifold.

Following classical terminology, a *critical* point of a differentiable function is a point where the gradient of the function vanishes, and the function is called a *Morse* function if its critical points are isolated and non-degenerate. For a critical point p of such a function, the stable (unstable) manifold $W^s(p)$ ($W^u(p)$) is the union of all integral curves associated to the gradient of the function, and respectively ending (originating) at p. Locally about a critical point of index i (the Hessian has i negative eigenvalues), the (un-)stable manifold is a topological disk of dimension i ($d - i$). The stable and unstable manifolds are also called the *separatrices*, as they partition the manifold into integral curves having the same origin and endpoint. In a more prosaic language, they are also called watersheds, by analogy with water drainage. In particular, under mild non degeneracy assumptions of the energy landscape, a transition between two

[12] If the internal (potential+kinetics) energy remains constant along the MD simulation, the system is Hamiltonian, and a large number of mathematical results apply [66]. We shall get back on this issue in the outlook.

adjacent minima is expected to correspond to the stable manifold of index one saddle joining the minima –a result known as the Murrell-Laidler theorem in bio-physics [70].

If Morse theory provides a powerful framework to describe energy landscapes, the pieces of information provided should be mitigated by the relative energies associated to critical points of various indices. As already noticed at the end of section 5.3.2, the thermal energy of the system indeed allows barrier crossing.

5.3.4 Free Energy Landscapes and Reaction Coordinates

In classical chemistry, a chemical systems moves from one minimum of energy to another following the minimum energy path, which, as just discussed is expected to go through index one saddles and intermediate minima. For complex systems such as a protein in its aqueous solution, things are more involved [45, 38, 42]. The different parameters have different relaxation times: fast parameters are those describing the solvent, as well as the variables accounting for the fast vibrations apart from covalent bonds of the protein; slow ones account for the large amplitude motions of the protein. Because the system equilibrates faster for some coordinates than others, we may partition the parameters as $x = (q, s)$. Denote $V(x)$ the potential energy of the system. By focusing on q and averaging out the other parameters, one defines the free energy landscape, which is the kinetically relevant one, by:

$$W(q) = -kT \ln \int \exp[-\frac{V(q,s)}{kT}]ds. \tag{2}$$

Coordinates q which provide kinetically relevant informations on transitions are called reactions coordinates. Finding such coordinates is challenging, even on simple systems. We illustrate these difficulties with a two dimensional system corresponding to a two states folding protein, whose unfolded and folded states are respectively denoted A and B. If q is the reaction coordinate sought, obvious requirements are (i) q takes different values q_A and q_B for these states, and (ii) q is such that the free energy W has a maximum at some value q^* in-between q_A and q_B. When these conditions are met, q is called an *order parameter*. If q provides in addition informations about the kinetics of the transitions, it is called a *reaction coordinate*. As illustrated on Fig. 5(a,b), these are different notions. In particular, Fig. 5(b) features a parameter q which is a good order parameter but not a reaction coordinate. For example, the dashed trajectory passes through q^* but does not correspond to a transition. In the ideal setting, for a reaction coordinate, the unstable manifold of the index one saddle joining the two minima separates the points which are committed to one state or the other, and thus determines the so-called Transition State Ensemble (TSE).

Practically, dealing with reaction coordinates poses several problems. First, for a system such as a protein and its solvent, one does not know a priori which variables are the slow ones. This issue is further discussed at the end of section 5.4.3. Second, if there is not a unique coordinate which is *slower* than the remaining ones, a multi-dimensional analysis must be carried out. Third, computing a free energy profile from Eq. (2) requires the coordinates over which the integration is performed to be equilibrated.

5.3.5 Folding Probability p_{fold}

To probe the relevance of a parameter as a reaction coordinate, one resorts to the *committor* probability, i.e. the probability of being committed to a given state [45]. More precisely, for any state x in the system, this is the probability of arriving say at B before before arriving at A within some time t_s. If the potential energy depends on positions and momenta, averaging is understood w.r.t. momenta. By studying this probability along a given path, one locates points near the TSE, since such points are equally committed to both states. Denote Dirac's delta function δ, and let $< z >_E$ the average of quantity z over an ensemble E. To probe the interest of an order parameter as a reaction coordinate, one studies the distribution of the committor probability at $q = q^*$, that is

$$P(p_B) =< \delta[p_B(x, t_s) - p_B] >_{q^*},\ p_B \in [0, 1].$$

For a good reaction coordinate, one expects $P(p_B)$ to be sharply peaked at $p_B = 1/2$. This is the case on Fig. 6(a), but not on Fig. 6(b) where $P(p_B)$ is bimodal, meaning that the orthogonal coordinates

are such that commitment to the two states occurs. The reader is referred to [45, 39, 38] for example physical systems featuring various committor's distributions.

The notion of transition state is also closely related to that of transition path [55, 37]. Define a transition path TP as a path in phase space that exits a region about the unfolded state, and reaches a region about the folded state. A collection of transition paths determines a conditional phase space density $p(x \mid TP)$, and one has

$$p(TP \mid x) = \frac{p(x \mid TP)p(TP)}{p_{eq}(x)}, \tag{3}$$

with $p_{eq}(x)$ the equilibrium probability of state x and $p(TP)$ the fraction of time spent on transition paths. Transition states are naturally defined as points maximizing $p(TP \mid x)$. Moreover, denoting $\overline{x}$ a point with same position and reversed momentum, and $p_A(x)$ the probability of reaching state A before state B from x, it can be shown [55] that

$$p(TP \mid x) = p_A(\overline{x})p_B(x) + p_A(x)p_B(\overline{x}). \tag{4}$$

An important property of this equation is that one can project x onto a lower dimensional space –see section 5.4. Denoting $r = r(x)$ such a coordinate, it can be shown [55] that

$$p(TP \mid r) = \frac{p(r \mid TP)p(TP)}{p_{eq}(r)}. \tag{5}$$

Practically, the difficulty with p_{fold} and related quantities are several [42]. First, the concept is bound to simple landscapes corresponding to two states folding processes. Most importantly, estimating p_{fold} requires sampling the TSE, which either requires long simulations –usually out of reach, or some form of importance sampling to favor the rare events corresponding to crossings of the TSE.

5.4 Inferring Reaction Coordinates

In the following, we review some of the most successful techniques to analyze transitions. We focus on the methodological aspects, and refer the reader to the original papers for a discussion of the insights gained, including connexions with experimental facts. As it can be seen from [42] for example, assessing the relevance of a particular coordinate can be rather controversial.

5.4.1 Reaction Coordinates?

In order to prove efficient to investigate folding, funnels such as that of Fig. 2 must be made quantitative, that is, one needs to specify what the axis account for. The variables parameterizing the axis are called *reaction coordinates*, and a quantitative energy landscape is displayed on Fig. 3. We now discuss several ways to design such coordinates.

5.4.2 Contacts Based Analysis

Following the work of G$\overline{o}$ [53], a natural way to tackle Levinthal's paradox consists of introducing a bias in the energy function towards native contacts, i.e. contacts observed in the folded state. More precisely, two residues which are not adjacent along the primary sequence of the protein form a native contact if they are *spatially close* in the protein's native state. Such pairs of residues are associated a favorable interaction energy, while the remaining ones are associated a repulsive, neutral or less attractive interaction energy. Figure 4 illustrates a folding process down a funnel, described using the fraction of native contacts. On one hand, energy landscapes obtained with G$\overline{o}$ models are generally minimally frustrated. On the other hand, as discussed in section 5.2, removing non local contacts may impair the folding process. At any rate and regardless of the energy model used, the fraction of native contacts Q can be used as reaction coordinate. Alternative empirical reaction coordinates, also exploiting the resemblance of a particular conformation with the native state, are being used: the radius of gyration (the root mean square distance of the collection of atoms from their center of mass), the effective loop length and the partial contact order [41]. In particular, the latter two coordinates are used in [41] to

measure the fraction of conformations that are actually accessible amongst the conformations with the same degree of *nativeness* Q. Such measures are directly related to the entropy of the system along the folding route, and thus allow one to assess the entropic ruggedness of the free energy landscape.

The native contacts can be used in a more elaborate fashion. Following [37], denote Q the matrix such that $Q_{ij} = 1$ if the distance between residues i and j is less than some cutoff (e.g. 12Å), and 0 otherwise. Using a weight matrix $W = (w_{ij})$, the contact matrix is projected onto a reaction coordinate defined by $r = \sum_{ij} w_{ij} q_{ij}$. Starting from a random initialization of matrix W, the weights are optimized so as to maximize a Gaussian fit of $p(TP \mid r)$ –see Eq. (5). In doing so, one ensures that all reactive configurations are condensed in a single peak.

5.4.3 Dimension Reduction Based Analysis

If one discards the momenta of the points, an important question is to come up with a simplified representation of the $3n$ dimensional energy landscape. Not surprisingly, PCA and MDS have been used for this purpose [13]. A typical illustration is provided by [35], where a PCA analysis of the conformations is first performed. Using the two most informative eigenvectors, an approximation of the landscape termed the *energy envelope* is computed. Fine informations on barriers between watersheds of minima might be lost –the ruggedness observed on a landscape computed from two PCA coordinates is at best questionable, but one expects to retain the overall shape of the watersheds. In [60], a PCA analysis is carried out on the critical points of an energy landscape, rather than on the whole point cloud. This analysis yields new coordinates, which can be plugged into the potential energy function.

One step towards a finer analysis is made in [43], where an adaptation of ISOMAP is used to derive new coordinates. The adaptions w.r.t. the standard ISOMAP algorithm are threefold. First, the computation of the nearest neighbors is done resorting to the least RMSD (lRMSD) [14]. Second, following [12], landmarks are used to alleviate the pairwise geodesic distance calculations. Third, the point cloud is trimmed to get rid of redundancies, which are expected in particular near the minima of potential energy. These conformations are later re-introduced into the low-dimensional embedding, which is important in particular to recover statistical averages. To assess the performance of the dimensionality reduction, a residual variance calculation is performed. For a two states folding protein, the transition state identified from the maximum of the free energy profile $W(x_1)$ associated to the first embedding coordinate x_1 is in full agreement with p_{fold}. (A result also holding for the reaction coordinate Q in this case.)
Motivated by the fact that 95% of the running time is devoted to the calculation of nearest neighbors, a further improvement is proposed in [67]. Assume m landmark conformations have been selected. Following the strategy used by the General Positioning System, each conformation (a point $\mathbb{R}^{3n}$) is represented as a m-dimensional point whose coordinates are the lRMSD distances to the m landmarks. In the corresponding m-dimensional space, the $l > k$ nearest neighbors of a point can be computed using the Euclidean distance, from which the k nearest ones according to the lRMSD are selected.

To finish up this review, one should mention methods which do not provide a simplified embedding of the landscape, but resort instead to a clustering of the nodes in parameter space [54, 52]. Nodes within the same watershed should belong to the same cluster, from which a Configuration Space Network (CSN) can be built. In some cases, quantitative informations (e.g. free energies) can even be retrieved.

Remark 1 *Having discussed dimensionality reduction techniques, one comment is in order. If one does not know a priory which are the slow variables, integrating Eq. (2) is not possible. This accounts for a three-stage strategy which consists of performing a simulation, performing a dimensionality reduction to infer candidate reaction coordinates, and probing them using p_{fold}.*

5.4.4 Morse Theory Related Analysis

Energy landscapes govern the folding process of proteins, but also the behavior of a number of physical systems such as clusters of atoms, ions or simple molecules [40, 70]. For some of these systems which exhibit a small number of stable crystalline geometries and a large number of amorphous forms, exploring

[13] Notice this analysis is different from the investigation of positional fluctuations mentioned in section 5.2.
[14] The Root Mean Square Deviation computed once the two structures have been aligned.

the landscape exhaustively is impossible. Yet, a qualitative analysis can be carried out by focusing on selected critical points. In [61, 33, 36], sequences of triples *minimum - saddle - minimum* are sought, and *super-basins* are built from their concatenation. In a related vein, the relative accessibility of potential energy basins associated to minima is investigated in [34], so as to define the so-called disconnectivity graph (DG). More precisely, two constructions are performed in [34]. The first one, based on the *canonical mapping*, focuses on the relative height of energy barriers, which governs transitions between states, thus encoding the kinetic behavior of the system. The second one, based on the *canonical mapping*, probes the potential energy surface at pre-defined values of the energy, thus encoding global topological properties of the landscape. Mathematically, constructing either DG is tantamount to tracking the topological changes of the set $V^{-1}(]\infty, v])$ when increasing v. As such changes occur at critical values only [64], the graph built when all critical values are available is called the contour tree [15]. In [34], a discrete set of energies are used to probe the topological changes of the level sets, though.

If one focuses on the relative accessibility of basins, one problem arising is that the DG built does not have any privileged embedding —the vertical axis encodes an energy, but the horizontal one does is meaningless. To complement the topological information by a geometric one, the following is carried out in [68]: first, similarly to [60], a PCA of critical points is carried out, from which a two-dimensional embedding of these critical points is derived; next, the DG is rendered as a three-dimensional tree, the z coordinate corresponding to the potential energy. Interestingly, such representations convey the (lack of) frustration of BLN models [36], depending on the interaction energy used.

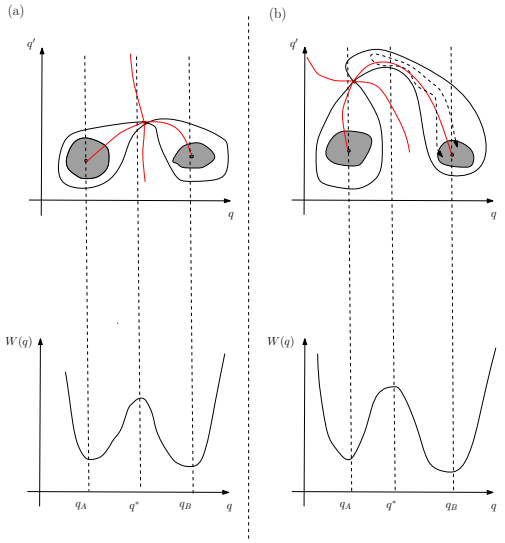

Figure 5: Potential energy landscape, with separatrices of the saddle in red: an order parameter may not be a good reaction coordinate. Adapted from [38].

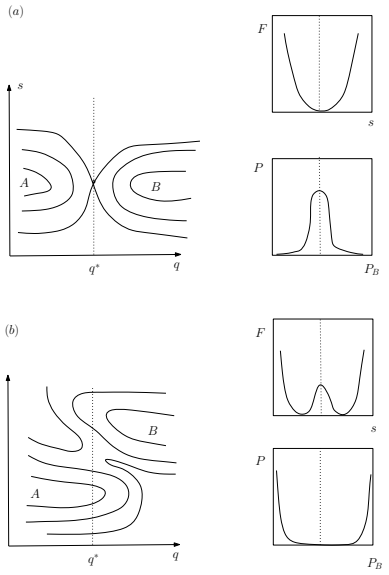

Figure 6: Probing a reaction coordinate by computing the committor probabilities p_B. Adapted from [39].

6 Outlook

Algorithms. Exploring a high-dimensional point cloud with the methods discussed and mentioned raises critical issues which should be kept in mind. First, it is usually assumed that the data points lie on a manifold. But for complex data corresponding e.g. to physical phenomena featuring bifurcations, a stratified complex might actually be the true underlying structure. Even in the manifold case, since the underlying manifold M is unknown, the geometric quantities we aim to preserve have to be estimated from the data set. Coming up with robust estimators poses difficult questions, especially since noisy

[15] Consider the level sets of a Morse function f, and call a connected component of a level set $f^{-1}(h)$ a *contour*. Further contract every contour to a point. The graph encoding the merge/split events between these contours is called the Reeb graph, or the contour tree if the domain is simply connected [49].

data (i.e. not exactly sampled on M) has to be accommodated from a practical standpoint. Worse, the sampling conditions insuring that the geometry can be correctly inferred from the data usually depend on some assumptions on M... which is unknown! These questions have been widely studied in computational geometry, in particular for the three dimensional surface reconstruction problem, but remain largely open in a broader setting.

Closely related to the previous questions are those concerning the convergence and theoretical guarantees. As discussed earlier, dimensionality reductions methods are not well suited for all situations. It is thus important to identify the necessary assumptions on M so as to ensure satisfactory results. We have seen in section 4 that one can answer this question for some of the methods (ISOMAP, HLLE). It is also interesting to have informations on the asymptotic behavior of the considered methods when the samples become denser and denser and converge to M. In this way, Hessian eigenmaps and diffusion maps reveal interesting asymptotic connections with classical operators defined on the underlying manifold M that need to be further explored.

Protein folding. In spite of three decades of intense research, the problem of protein folding is still open. In the context of energy landscapes and dimensionality reduction, a number of further developments are called for.

A variety of (molecular dynamics) simulations are being performed: depending on the system studied (all atoms/coarse, explicit/implicit/no solvent), either the temperature, the pressure or the internal energy of the system are kept constant. For example, if the temperature is held constant using a thermostat —for example the Nose-Hoover, part of the internal energy of the system is dissipated into the thermostat. If the internal energy of the system is conserved, then, the system is Hamiltonian.

For Hamiltonian systems, a large number of mathematical results exist. For example, using the geometrization of Hamiltonian dynamics, a trajectory of the system corresponds to a geodesic of a suitable Riemannian manifold [66]. This point of view is not really used in recent folding studies, which focus on Morse related analysis of potential and free energy surfaces. The study of the relationship between folding properties inferred from energy landscapes on the one hand, and from Hamiltonian dynamics on the other hand deserves further scrutiny.

Practically, one or two reaction coordinates are usually dealt with, a rather stringent limitation. Methods based on manifold learning are appealing in this perspective, since the dimensionality of the embedding can be estimated. But a critical step for these methods is that of the neighborhood selection. On one hand, the samples are generally processed in a uniform way since the same number of neighbors is used for all points. On the other hand, Morse theory tells us that the local density of samples about a critical point is related to its index. Therefore, a segmentation of the point cloud might be in order before resorting to dimensionality reduction techniques. Doing so might allow one to bridge the gap with Morse theory related methods, whose focus has been on the decomposition of the landscape into basins –as opposed to the design of new coordinates accounting for transitions.

Another key problem is that of stability, in the context of rugged / frustrated landscapes. Ideally, multi-scale analysis of landscapes should be developed, so as to asses what is significant and what is not at a given scale. Topological persistence and more generally tools developed in computational topology might be helpful here. Such analysis might also allow one to spot cascades of minor events in the folding process, such cascades triggering major events –cf phase transitions.

Finally, an improved analysis of landscapes would have another dramatic impact, namely on the simulation processes themselves. Should a finer understanding of cooperative motions be available, steered simulations favoring these coordinates should allow one to move faster along a (rugged) landscape.

Hopefully, a finer geometric and topological analysis of non linearities arising on energy landscapes will help in making simulation able to cope with biologically relevant time scales.

Acknowledgments. F. Cazals wishes to acknoledge Benjamin Bouvier, Ricardo Lima, Marco Pettini and Charles Robert for stimulating discussions.

7 Bibliography: Dimensionality Reduction

[1] P.K. Agarwal, S. Har-Peled, and H. Yu. Embeddings of surfaces, curves, and moving points in euclidean space. In *ACM SoCG*, 2007.

[2] M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. In *COLT 2005*.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[5] C. M. Bishop, M. Svensen, and C. K. I. Williams. Gtm: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[6] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in euclidean space. In *Proceedings of the 22nd ACM Symposium on Computational Geometry*, 2006.

[7] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Stability of boundary measures. 2007.

[8] Siu-Wing Cheng, Yajun Wang, and Zhuangzhi Wu. Provable dimension detection using principal component analysis. In *Symposium on Computational Geometry*, pages 208–217, 2005.

[9] R. R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. of Nat. Acad. Sci.*, 102:7426–7431, 2005.

[10] R. R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proc. of Nat. Acad. Sci.*, 102:7432–7437, 2005.

[11] V. de Silva, J.C. Langford, and J.B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. 2000.

[12] V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA, 2003.

[13] M. Demazure. *Bifurcations and Catastrophes: Geometry of Solutions to Nonlinear Problems*. Springer, 1898.

[14] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[15] J. Giesen and U. Wagner. Shape dimansion and intrinsic metric from samples of manifolds with high co-dimension. In *Proc. of the 19th Annual symp. Computational Geometry*, pages 329–337, 2003.

[16] Jihun Ham, Daniel D. Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 47, New York, NY, USA, 2004. ACM.

[17] T. Hastie and W. Stuetzle. Principal curves. *J. Amer. Stat. Assoc.*, 84:502–516, 1989.

[18] Matthias Hein and Markus Maier. Manifold denoising. In *NIPS*, pages 561–568, 2006.

[19] S. Lafon and A.B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning and data set parameterization. *IEEE PAMI*, 28(9):1393–1403, 2006.

[20] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2007.

[21] F. Memoli and G. Sapiro. Distance functions and geodesics on point clouds, 2005.

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

[23] S. T. Roweis and L. K. Saul. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

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

[25] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: a comparative review. 2007.

[26] Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *CVPR (2)*, pages 988–995, 2004.

[27] Kilian Q. Weinberger, Fei Sha, and Lawrence K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 106, New York, NY, USA, 2004. ACM.

[28] K.Q. Weinberger and L.K. Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*, 2006.

[29] K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[30] Li Yang. Building connected neighborhood graphs for isometric data embedding. In *KDD*, pages 722–728, 2005.

[31] Hao Zhang, Oliver van Kaick, and Ramsay Dyer. Spectral mesh processing. *Computer Graphics Forum (to appear)*, 2008.

8 Bibliography: Structural Biology

[32] A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Genetics*, 17(4):412–425, 1993.

[33] K.D. Ball, R.S. Berry, R.Kunz, F-Y. Li, A. Proykova, and D.J. Wales. From topographies to dynamics on multidimensional potential energy surfaces of atomic clusters. *Science*, 271(5251):963 – 966, 1996.

[34] O. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, 106(4):1495–1517, 1997.

[35] O.M. Becker. Principal coordinate maps of molecular potential energy surfaces. *J. of Comp. Chem.*, 19(11):1255–1267, 1998.

[36] R. Stephen Berry, Nuran Elmaci, John P. Rose, and Benjamin Vekhter. Linking topography of its potential surface with the dynamics of folding of a proteinmodel. *Proceedings of the National Academy of Sciences*, 94(18):9520–9524, 1997.

[37] Robert B. Best and Gerhard Hummer. Chemical Theory and Computation Special Feature: Reaction coordinates and rates from transition paths. *Proceedings of the National Academy of Sciences*, 102(19):6732–6737, 2005.

[38] P.G. Bolhuis, D. Chandler, C. Dellago, and P.L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry*, 53:291–318, 2002.

[39] P.G. Bolhuisdagger, C. Dellago, and D. Chandler. Reaction coordinates of biomolecular isomerization. *PNAS*, 97(11):5877–5882, 2000.

[40] C.L. Brooks, J. Onuchic, and D.J. Wales. Statistical thermodynamics: taking a walk on a landscape. *Science*, 293(5530):612 – 613, 2001.

[41] 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(27):8426–8432, 2004.

[42] Samuel S. Cho, Yaakov Levy, and Peter G. Wolynes. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proceedings of the National Academy of Sciences*, 103(3):586–591, 2006.

[43] P. Das, M. Moll, H. Stamati, L. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *PNAS*, 103(26):9885–9890, 2006.

[44] Payel Das, Corey J. Wilson, Giovanni Fossati, Pernilla Wittung-Stafshede, Kathleen S. Matthews, and Cecilia Clementi. Characterization of the folding landscape of monomeric lactose repressor: Quantitative comparison of theory and experiment. *Proceedings of the National Academy of Sciences*, 102(41):14569–14574, 2005.

[45] R. Du, V. Pande, A.Y. Grosberg, T. Tanaka, and E.I. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108(1):334–350, 1998.

[46] R.L. Dunbrack. Rotamer libraries in the 21st century. *Curr Opin Struct Biol*, 12(4):431–440, 2002.

[47] H.A. Scheraga et al. A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. of Computational Chemistry*, 18(7):849–873, 1997.

[48] A. Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. 1999.

[49] A.T. Fomenko and T.L. Kunii. *Topological Modeling for visualization*. Springer, 1997.

[50] D. Frenkel and B. Smit. *Understanding molecular simulation*. Academic Press, 2002.

[51] A.E. Garcia. Large-amplitude nonlinear motions in proteins. *Physical Review Letters*, 68(17):2696–2699, 1992.

[52] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao. Complex network analysis of free-energy landscapes. *Proceedings of the National Academy of Sciences*, 104(6):1817–1822, 2007.

[53] Nobuhiro Go and Hiroshi Taketomi. Respective Roles of Short- and Long-Range Interactions in Protein Folding. *Proceedings of the National Academy of Sciences*, 75(2):559–563, 1978.

[54] Isaac A. Hubner, Eric J. Deeds, and Eugene I. Shakhnovich. Understanding ensemble protein folding at atomic detail. *Proceedings of the National Academy of Sciences*, 103(47):17747–17752, 2006.

[55] G. Hummer. From transition paths to transition states and rate coefficients. *J. Chemical Physics*, 120(2), 2004.

[56] T. Ichiye and M. Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Genetics*, 11(3):205–217, 1991.

[57] C.L Brooks III, M. Gruebele, J. Onuchic, and P. Wolynes. Chemical physics of protein folding. *Proceedings of the National Academy of Sciences*, 95(19):11037–11038, 1998.

[58] S.E. Jackson. How do small single-domain proteins fold? *Fold Des.*, 3(4):R81–91, 1998.

[59] J. Janin, S. Wodak, M. Levitt, and B. Maigret. Conformations of amino acid side chains in proteins. *J. Mol. Biol.*, 125:357–386, 1978.

[60] T. Komatsuzaki, K. Hoshino, Y. Matsunaga, G.J. Rylance, R.L. Johnston, and D. Wales. How many dimensions are required to approximate the potential energy landscape of a model protein? *J. Chem. Phys.*, 122, February 2005.

[61] R.E. Kunz and R.S. Berry. Statistical interpretation of topographies and dynamics of multidimensional potentials. *J. Chem. Phys.*, 103:1904–1912, August 1995.

[62] O.F. Lange and H Grubmller. Generalized correlation for biomolecular dynamics. *Proteins*, 62:1053–1061, 2006.

[63] C. Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, 65:44–45, 1968.

[64] John W. Milnor. *Morse Theory*. Princeton University Press, Princeton, NJ, 1963.

[65] E. Paci, M. Vendruscolo, and M. Karplus. Native and non-native interactions along protein folding and unfolding pathways. *Proteins*, 47(3):379–392, 2002.

[66] M. Pettini. *Geometry and Topology in Hamiltonian Dynamics and Statistical Mechanics*. Springer, 2007.

[67] E. Plaku, H. Stamati, C. Clementi, and L.E. Kavraki. Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction. *Proteins: Structure, Function, and Bioinformatics*, 67(4):897–907, 2007.

[68] G. Rylance, R. Johnston, Y. Matsunaga, C-B Li A. Baba, and T. Komatsuzaki. Topographical complexity of multidimensional energy landscapes. *PNAS*, 103(49):18551–18555, 2006.

[69] M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996.

[70] D.J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.