Distinguishing biomolecular pathways and metastable states

Antonio B. Oliveira Jr,[†] Huan Yang,[‡] Paul C. Whitford,[‡] and Vitor B. P.

Leite^{*,†,¶}

[†]*Departamento de Física, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, São Paulo, Brazil.*

[‡]*Department of Physics, Northeastern University, Boston, MA 02115, United States.*

[¶]*Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, United States.*

E-mail: vitor.leite@unesp.br

Abstract

Protein folding occurs in a high dimensional phase space, and the representation of the associated energy landscape is non-trivial. A widely applied approach to studying folding landscapes is to describe the dynamics along a small number of reaction coordinates. However, other strategies involve more elaborate analysis of the complex phase space. There have been many attempts to obtain a more detailed representation of all available conformations for a given system. In this work, we address this problem using a metric based on internal distances between amino acids to describe the differences between any two conformations. Using an effective projection method, we are able to go beyond the usual one-dimensional representation and thus visualize landscapes in two dimensions. We refer to this method as Energy Landscape Visualization Method (ELViM). We have applied this methodology using C_α structure-based model (SBM)

to study the folding of two well-known proteins: SH3 domain and Protein-A. Our visualization method yields a detailed description of the folding process, making possible the identification of transition state regions, and establishing the paths that lead to the native state. For SH3, we have analyzed structural differences in the distribution of folding routes. The competition between the native and mirror structures in protein A is discussed. Finally the method is applied to discuss the conformational changes of the Elongation Factor Thermally unstable (EF-Tu) and its transition states. Distinct features of the ELViM are that it does not require or assume an existence of a reaction coordinate, and it does not require kinetic information regarding the investigated system.

Introduction

Protein folding occurs in a high-dimensional phase space, and it involves a conformational search across a broad range of structurally-distinct states. Within this vast space, the energy landscape is shaped such that a protein can reliably fold and perform its various functions.¹⁻³ Through the course of evolution, proteins have achieved this ability by selecting for energy landscapes that are dominated by the contacts formed in the native conformation. Thus, a long-term goal of the field has been to understand the character of folding dynamics, in the context of “minimally-frustrated”, or “funnel-like” energy landscapes.^{4,5} However, even though with this general character of the landscape, the precise dynamics results from an interplay of energetic, entropic and topological aspects. Thus investigating the balance between these factors has inspired the development of a large number of energetically “smooth” models for studying protein dynamics,⁶ which includes a broad range of native-centric “structure-based” models. While these models are simplified, comparison of simulations^{7,8} and experimental measurements^{9,10} has shown that these models are able to capture many aspect of the folding process, including folding rates, Φ -values, landscape roughness, and topological impacts on folding mechanisms.¹¹⁻¹³

It is important to recognize that energy landscape techniques are not restricted to the study of protein folding. Rather, they may also be used to study functional dynamics, which can involve large-scale conformational changes,¹⁴ motor-like energy transfer¹⁵ and assembly.¹⁶ In addition, this strategy can be used to investigate a variety of other biomolecular systems, including RNAs,¹⁷ DNAs,¹⁸ and complex assemblies, such as the ribosome¹⁹ and chromatin.²⁰ Accordingly, these models represent a means for characterizing the dynamics of any number of large-scale and complex biomolecular processes.

Even though folding and functional dynamics are intrinsically multi-dimensional, it is often possible to accurately describe the kinetic and thermodynamic properties in terms of a few key quantities. For example, structural parameters, such as radius of gyration and fraction of native contacts (Q), are commonly used to measure protein folding dynamics.²¹ While these low-dimensional representations can capture the dominant kinetic properties of a system,^{22,23} and topological features of transition state ensembles,¹¹ applying putative coordinates has the potential to mask the richness of the dynamics. However, techniques are available for systematically identifying suitable reaction coordinates. For example, time-correlation analysis²⁴ can be used to define classes of collective variable and then identify which describe the slowest motions. Transition path analysis may also be used to determine which coordinates capture an underlying free-energy barrier,²⁵ though exhaustive coordinate comparisons can become prohibitively expensive. One limitation with these techniques is that these methods require general definitions of coordinates be provided *a priori*. Accordingly, there is still a need for automated techniques that are capable of rapidly mapping the broad character of an arbitrary landscape.

An alternate strategy for mapping the landscape is to construct connectivity maps that are based on energetics or kinetics. For example, one may define Markov State Models (MSMs), where the system is described in terms of large numbers of relatively long-lived states.²⁶⁻²⁸ Another approach is to identify local minima individually in order to graph the high-dimensional space.^{29,30} For both of these techniques, more accurate representations

are obtained when studying rough landscapes that possess many well-defined free-energy minima. However, when applying minimally-frustrated models, the minima tend to be characterized by broad and shallow minima, making the utility of these approaches unclear.

In the current study, we describe a method for visualizing energy landscapes, called the Energy Landscape Visualization Method (ELViM). While this technique has previously been applied to lattice models of proteins,³¹ the current work extends this technique to study off-lattice models for folding and conformational rearrangements. ELViM seeks to convey a meaningful 2D (and/or 3D) projection of the landscape that is based on an internal metric to assign distances between sampled conformations. In contrast to other methods, ELViM does not require intuitive designation of coordinate classes, nor does it require the sampling of long-lived minima, making it technique for studying smoother landscapes. To illustrate the applicability of this approach, we apply the method to three classes of conformational processes in proteins. First, we describe the folding of a well-studied protein: SH3. Next, we apply the method to study the folding of protein A, which can fold to two competing configurations: “native” and “mirror”. Finally, we apply ELViM to describe large-scale ($\sim 60\text{\AA}$) domain rearrangements in elongation factor Tu (EF-Tu). Together, these applications demonstrate how ELViM enables broadly-applicable analysis that can distinguish the rich characteristics of biomolecular dynamics.

Results

One goal of the present study is to develop and apply a method that is capable of identifying parallel pathways in nominally-smooth landscapes. As described in the introduction, simplified structure-based models can provide physically-meaningful descriptions of dynamics that are often consistent with experimental results. However, the challenge of identifying paths that are associated with broadly-defined ensembles is a difficult task. To approach this, we apply ELVIM, an analysis approach inspired by our earlier studies of lattice models

for proteins. As described below, we find that this technique allows one to infer details of energy landscapes in a variety of contexts, including protein folding, folding to competing structures and large-scale rearrangements in a multi-domain protein.

An essential aspect of this approach is to have an appropriate metric to represent effective distances between pairs of conformations in a sampled data set. In this study, such metric corresponds to the internal distance introduced by the Wolynes group.³² Using this distance measure, a multidimensional reduction is performed to obtain a 2D landscape representation, which provides easily interpretable descriptions of complex systems. In the present study, we restricted the metric to consider only structural information of each conformation, though subsequent extensions could include kinetic weighting. As we show below, even though kinetic information is not considered, the method still allows one to capture free-energy barriers, transition states and parallel paths associated with folding and functional rearrangements.

Figure 1: **Typical folding dynamics.** (a) Sample of a SH3 trajectory, described in terms of commonly-used structural coordinates: RMSD (black) and Q (red). (b) Native (N) and Unfolded (U) states occupy different regions in the 2D phase space, but there is not apparent separation of parallel pathways. Tube representation of native SH3 structure (pdb 1FMK) is shown in the inset.

Figure 2: **Folding landscape of SH3 using ELViM.** (a) Folding trajectory (Fig. 1b) mapped into 2D phase space using ELViM, which allows a more detailed folding landscape to emerge. (b) Contour plot calculated by 2D Histogram using the 2D projection as reference, the color scheme represents the values of the free energy. The three marks (A, B and C) correspond to the preferred pathways.

Detecting Parallel Routes During Protein Folding

When describing folding, the “mechanism” consists of a distribution of parallel transition paths.³³ There are cases in which visual inspection can suggest appropriate effective coordinates for identifying these parallel folding processes. However, it is desirable to establish protocols and methods that can identify parallel routes without prior knowledge of the system. One example of parallel folding paths can be found in the context of the SH3 domain of tyrosine kinase c-Src. The folding mechanism of SH3 has been well characterized, both experimentally and computationally,^{34–37} making it an ideal benchmark for evaluating the robustness of the ELViM methodology. SH3 is a 57-residue protein that forms a five stranded beta barrel (PDB ID: 1FMK,³⁸ Fig. 1b). To study the folding of SH3, we apply molecular dynamics simulations using a C- α structure-based (G $\ddot{o}$ -like) model.¹¹ When using this model, the folding dynamics are two-state in character, and there are sharp transitions between the folded (high Q) and unfolded (low Q) ensembles (Fig. 1a). Similar dynamics is also observed when using the RMSD from the folded conformation (Fig. 1a). Further, when describing the dynamics in terms of both Q and RMSD (Fig. 1b), there does not appear to be an obvious

Figure 3: **Contact maps of SH3 structures in A,B and C transition regions.** The maps are colored by the fraction of formed contact in each transition state region. A representative structure for each region is shown in the lower part of each panel. D) Overall contact formation at the transition state, as defined by Q . Two groups of contacts are highlighted: Early group (in green) represent the contacts preferentially formed (strands β_b , β_c and β_d), while Late group (in red) remain mostly unstructured (strands β_a and β_e)

distinction between parallel folding routes.

When we use ELViM to project SH3 conformations into a 2D map, a rich picture of the underlying landscape is revealed, where the unfolded and folded ensembles are connected through a number of transition regions (Fig. 2a). One may also represent projection as a 2D free energy, where $\Delta\mathcal{F} = -\log(n_i/N)$, n_i is a local density of points, and N is the total number of conformations. This reveals three major paths (A, B and C) with similar transition barrier heights ($\sim 3 k_B T$) separating the unfolded and folded ensembles (Fig. 2b).

Previous efforts by Lammert and co-workers³⁶ showed that the so-called ‘early’ (‘late’)

group of contacts are preferably formed before (after) the transition state is reached (Fig. 3d). To describe the dominant folding path, they introduced a reaction coordinate Q_{path} , which accounts for the difference between the number of ‘early’ and ‘late’ contacts that are formed. The main transition paths shown in Fig. 2b (paths A,B, and C) are all in agreement with the initial formation of early contacts (Fig. 3d). Folding through the reverse mode (*i.e.* forming “late” contacts first) is associated with a much larger free-energy barrier, which indicates that such a process is not likely to be kinetically competitive (see Fig. 3 and S2).

While all three transition state regions (A,B, and C) are associated with formation of early group contacts, ELViM reveals distinct differences in these parallel paths. For example, as can be seen by a representative structure of each region, shown in the lower part of each contact map (Fig. 3.a, b, and c). These contact maps can give a weighted transition path distribution, which is more meaningful than the average transition-state ensemble (TSE) representation using Q , which eventually is not even the best reaction coordinate.

ELViM takes into account only geometric distances between conformations. If one considers the potential energy associated with each structure, a third dimension is included in the landscape description. The SH3 3D representation presents indeed a funnel like structure, in which the transition state conformations are placed asymmetrically around the bottom of the funnel. This representation can help to visualize the favorable folding pathways, connecting them to the different transition states, see Fig. S3, and Video S1.

Folding in a Multi-basin Landscape

Mirror-like protein folds can form when there are energetically competitive structures accessible that are compatible with the native contact network.^{39,40} As an example, using molecular dynamics simulations, Noel et al.⁴¹ concluded that the native fold and its mirror image have similar stabilities, making them thermodynamically competitive.

To investigate how the interplay of competitive conformations may be described by 2D ELViM projections, we performed simulations using a dual-basin structured based model

Figure 4: **Mirror image analysis of Protein A.** (a) 2D projection of protein A simulations using a dual-basin structure-based model.⁴¹ The color scheme represents the reaction coordinate $Q_N - Q_M$, where red and blue points are native-like and mirror-like structures, respectively. (b) Free energy contour. The letters mark densely populated regions, which are associated with native (N), native-like intermediate (I) and mirror (M) states. Tube representation of the conformational states N, M and I are shown in (c), (d) and (e), respectively.

(model described in⁴¹), where the native and mirror-like conformations represent free-energy minima.

To determine whether ELViM can distinguish between the competing folding routes, the folding landscape of Protein A (Fig. 4) is colored according to $Q_N - Q_M$, where Q_x ($x =$ Native or Mirror) is defined as the sum of all tertiary contacts present in each contact map (coord defined in Ref.⁴¹). The region in blue represents conformations similar to mirrored structures, while the red region describes native-like structures. Given the parameters of the dual basin model,⁴¹ simulation was conducted at temperature below T_f , in which sampling of the Native and Mirror conformations are better probed, though completely unfolded conformations also occur (see Methods and SI). Fig. 4a shows that native-like conformations occur more frequently than the mirror-like ones. Unfolded conformations are all on the outer rim of the figure (see Fig. S4) A contour plot of $\mathcal{F}$ (Fig. 4b) reveals three clear minima: Native (N), a native-like intermediate (I) and the mirror image (M). The emergence of a native-like intermediate is unexpected, and it was not reported in earlier studies that used conventional low-dimensional reaction coordinates (Q , the $Q_N - Q_M$, and RMSDs).⁴¹ This suggest a new mechanism in which the mirror and native states inter-convert into each other. When one inspects representative structures of the I state, one can envision a process by which the Helix 3 (blue Helix) is displaced relative to a nominally-folded Helix 1/2 bundle (red/white bundle) (Fig. 4). When we take into account the potential energies of each state, the 3D visualization of Protein A energy landscape gives insight also on the relative energy differences between mirror and native configurations (Fig. S5 and Video S2). In this example we have seen how this reaction-coordinate free landscape analysis helped to identify an intermediate state, which may help to design probes to detect them experimentally.

Capturing competing pathways during domain rearrangements

Elongation factor-Tu (EF-Tu) is a three-domain protein that is responsible for delivery of aminoacyl-tRNA molecules to the ribosome.^{42–44} In order to accomplish this task, EF-Tu

binds aa-tRNA and GTP to form the ternary complex. During initial association with the ribosome, the aa-tRNA forms base pairs with the ribosome-bound mRNA codon. EF-Tu then hydrolyzes GTP and undergoes a large-scale domain rearrangement process, resulting in the release of the aa-tRNA molecule. Overall, domain rearrangement involves the displacement of domain I, relative to domains II and III (Fig. 5a/b). One may alternately describe the motion as a relative displacement the II/III superdomain, though structural considerations suggest this process is sterically hindered by the ribosome.⁴⁵ In terms of biological function, the balance between EF-Tu dissociation and conformational dynamics is central to aa-tRNA “proofreading,” where incorrect tRNA molecules may be rejected, even if they successfully pass the initial association step.

While the ability of EF-Tu to adopt competing conformational states has been known for decades,⁴⁶ the precise physical-chemical relationship between large-scale domain motion and the fidelity of aa-tRNA selection is not fully understood. Recent simulations,^{47,48} as well as experiments,⁴⁹ have suggested that the timing of EF-Tu release from the ribosome may influence the accuracy of aa-tRNA selection. Experiments have shown that domain rearrangement in EF-Tu and dissociation from the ribosome are kinetically competitive processes.⁵⁰ Previous simulations with explicit-solvent models⁵¹ have implicated transient domain separation during this rearrangement. Similarly, simulations with a coarse-grained model predicted that EF-Tu may adopt a disordered intermediate ensemble during domain rearrangement.⁴⁵ Together, these studies suggest that the precise level of proofreading exhibited by the cell will be determined by an elaborate interplay of multiple large-scale processes.

The complicated character of domain rearrangements in EF-Tu makes it an ideal candidate to test the ability of ELViM to identify parallel pathways during biological function. Here, we analyze a recently-reported data set in which over 300 spontaneous domain rearrangement events were observed.⁴⁵

In contrast to previous studies, where intuitive choices of reaction coordinates had been constructed,^{45,51} we find that ELViM can automatically distinguish between the competing

Figure 5: **Distinguishing pathways during large-scale domain rearrangements.** a) Elongation Factor Tu (EF-Tu) is a three-domain (labeled I, II and III) protein that is responsible for delivering tRNA molecules to the ribosome. In order to deliver tRNA, Domain I (red, residues 1-190) must be displaced, relative to the II/III superdomain (blue, residues 200-350), where the system transitions between the PRE (a) and POST (b) orientations. To highlight the change in orientation of Domain I, residues 1, 2 30-33 and 64-66 are shown as yellow spheres. c) 2D projection following the metric defined by eq. 1. With this metric, there is a clear separation of two distinct pathways of interconversion. Both plots are colored based on RMSD from the PRE structure. d) Comparison of average distance of each residue pair between TSE1 and TSE2 regions. Representative structures taken from the TSE1 (e) and TSE2 (f) ensembles reveals a full rotation of Domain I in the TSE1 path. In contrast, TSE2 is associated with transient dissociation of the domains, though the overall orientation of Domain I is similar to the PRE and POST configurations.

pathways of domain rearrangement (Fig. 5c/d). When using ELViM with the internal distances defined by eq. (1) and (2), there is clear separation of the endpoint configurations (PRE and POST), as well as two distinct transition-state ensembles (TSE1 and TSE2). To measure the differences in the identified ensembles, we next considered the average distance of each residue pair: $\langle r_{ij} \rangle_n$, where i and j are the residue indices and n is the ensemble (PRE, TSE1, TSE2, POST). We find that, on average, almost all distances between domain I and the II/III superdomain are larger in TSE1 and TSE2, than in the PRE and POST conformations (not shown). This is consistent with the extended character of the TSEs, where the domains are transiently separated.^{45,51} Relative to TSE2, we find that there is a combination of extended and contracted distances in TSE1 (Fig. 5d). Specifically, we find that the distances between the II/III superdomain and some Domain I residues (1, 2, 30-33 and 64-66) are decreased, while nearly all other distances are increased. This implicates relative rotation of the domains, which is consistent with visual inspection of the ensembles (Figure 5e/f). Overall, there is a roughly 180° rotation of domain I, which is consistent with previous descriptions that implicated extended (TSE1) and disordered (TSE2) ensembles.

In summary, ELViM is able to automatically detect parallel pathways and competing transition state ensembles, without the need for manual intervention. Rather than having to intuitively design coordinates, which can involve visual inspection of large numbers of transition events, the ELViM method provides a means for efficiently mapping structurally-distinct pathways during elaborate conformational rearrangement in multi-domain, or multi-component, assemblies.

Discussion

In this work we presented a method to visualize biological macromolecule energy landscapes. The method ELViM is based on the assumption that energy states are highly correlated with its molecular conformations. If one can have a suitable measurement of distances between

any two configurations, it is possible to have a proper multidimensional representation of its energy landscape that may then be reduced to a more comprehensible 2D or 3D perspective. We have shown that, for the systems analyzed, a suitable metric is based on the inner distances between its atoms.

While we have focused our analysis on molecular dynamics trajectories, ELViM is applicable to any sampling technique, such as Monte Carlo or Discrete Path Sampling.^{52,53} Even though the method has been targeted primarily to proteins, we anticipate this strategy will prove useful for other molecular systems, such as RNA, DNA, or biomolecular assemblies. Further, ELViM is not restricted to any particular level of molecular resolution (C_α or all-atom). Rather, the strength of the method relies on the way the distances between any two conformations are measured, where we have found a suitable metric is q -Wolynes ($q_{i,j}$), eq.(1). This metric was initially developed to compare conformational differences between predicted protein structures. In principle, other metrics could be considered for other systems. One strength of ELViM is that it does not require or assume knowledge of an appropriate reaction coordinate. This is particularly beneficial when studying a new system for which minimal quantitative analysis is available.

Our results suggests a variety of potential applications for this method. For instance, the detailed information of transition state paths, given by the contact analysis and typical structures at TS, can be used to engineer mutations to favor one TS with respect to another.³⁷ We also envision it will be useful for studying more complex proteins, such as those that contain topological knots,^{54,55} or proteins where backtracking mechanisms may occur.^{56,57} In closing, we should note that the metric used in this study is based only on structural features, and the mapping of the landscape was not based on kinetic properties. However, there are many avenues by which to further extend this technique, such as including kinetic information to weight the effective distances between conformations, which may enable a more broad thermodynamic and kinetic view of biomolecular dynamics.

Methods

The visualization methodology is based in four steps: (I) Generate an ensemble of structures, which in general are obtained through simulations; (II) Calculate the dissimilarity matrix by applying a metric throughout the simulated trajectory; (III) Clustering data, when there is a large number of structures, a data processing procedure is carried out to cluster very similar structures into a single conformation, which represents each cluster; (IV) Perform a multidimensional projection, which transforms the dissimilarity matrix into a 2D projection. Each step is discussed in detailed below.

I. Generating a data ensemble

As a first step, one obtains an ensemble of structures. This step is not part of the ELViM itself, in the sense that many techniques may be used to generate the data ensemble. In protein folding investigations, typically one uses a C_α , $C_\alpha + C_\beta$, or all-atoms representation. However, ELViM can, in principle, be used in any molecular system; for example it can be used to study energy landscapes of RNA, DNA, macromolecular assemblies and chromatin.

In this work, we use a structure-based model that represents each protein residue by a single bead placed at the position of its C_α atom.¹¹ This model defines an experimentally derived structure as the global potential energy minimum.⁵⁸ All simulations were performed using the Gromacs (v4.5.4) software package,⁵⁹ with force field files generated by the SMOG web server,^{60,61} and the native contact map made by CSU algorithm.⁶²

We performed two sets of simulations, the first one was made to find the folding temperature (T_f). We carried out short simulations (5.0×10^8 time steps) in a range of temperature, with T_f being calculated using eSBMTools.⁶³ With T_f defined, we carried out long simulations (5.0×10^{10} time steps) in this temperature. This second set of simulations is performed in order to have the conformational space as thoroughly visited as possible.

Two proteins were chosen to test the method in the context of folding: src SH3 domain

(pdb 1FMK), Protein A (pdb 1BDD); these proteins have been chosen because their folding is well known and they represent basic structural motifs of proteins.^{11,36,41} A third protein, the GTP-binding domain of EF-Tu (elongation factor thermo unstable) was also used, since it undergoes a dramatic and large-scale rearrangement during the transition between GTP-bound and GDP-bound conformations transition (EFTu-GDP PDB ID: 1DG1, and EFTu-GTP PDB ID: 5YUM).

II. Dissimilarity matrix

A critical feature of ELViM is the definition of a metric that coherently and robustly describes the data. For off-lattice systems, such a metric takes into account the protein structure and establishes a distance based in internal conformational differences. This dissimilarity measurement is inspired on replica symmetry breaking parameter of spin glasses, that has been used to constrain clusters of protein states within a given degree of similarity.³² In the present method, this metric is extended to describe the distance between any two conformations in the entire configuration phase space. The measure of comparison between conformations k and l is given by:

$$q^{k,l} = \frac{1}{N} \sum_{(i,j)} \exp \frac{-(r_{i,j}^k - r_{i,j}^l)^2}{\sigma_{i,j}^2}, \quad (1)$$

where $r_{i,j}^{k(l)}$ is the distance between the atom i and j of the conformation $k(l)$. N is the normalization over all non-neighboring pairs of residues and $\sigma_{i,j} = \sigma_0 |i - j|^\epsilon$, being $\sigma_0 = 1$ and $\epsilon = 0.15$. The dissimilarity between any two structures is described as

$$\delta^{k,l} = 1 - q^{k,l}, \quad (2)$$

where $\delta^{k,l}$ ranges from 0 (identical) to 1 (very different). The dissimilarity measure is calculated between all the structures of a given trajectory, and it yields a $n \times n$ triangular matrix, where n is the total number of conformations.

III. Preliminary data clustering

This step is necessary if one has a large number of structures. Currently available multidimensional projection procedures are limited to about 50,000 data points, and a well probed phase space may present a much larger number of configurations. In these cases, it is required that similar structures are clustered into single representative configurations prior to performing multidimensional reduction. While this reduces the computational requirements, reducing the number of points does not lead to a loss of information, the number of states associated with each cluster is recorded, and all states of the initial data set are considered in the analysis.

This procedure is consistent with the other steps of the method, since the same distance criteria between conformations is used throughout the entire method, eq.(2). Conformations close to each other by less than a cut-off distance δ_{cut} are clustered into a single conformation. Starting with a very small δ_{cut} , we group the data so that conformations close to each other ($\delta^{k,l} < \delta_{cut}$) belong to the same cluster. Through iterations δ_{cut} is gradually increased. A condition is set such that if two conformations have a small distance ($\delta^{k,l} \leq 0.1$) and the difference between distances against all other structures are smaller than ± 0.05 , this pair of structures is clustered. The clustering is performed until it reaches a stop parameter, which can be a maximum cut-off distance δ_{cut}^{MAX} or a defined number of structures that will be used in the next step. In this way, we preserve all the phase space sampled in the simulation, and at least one representative structure of each region is used to generate the landscape visualization. At the end of this step we have a set of distances between each cluster δ_c , which consist of a $N \times N$ triangular matrix, where N is the total number of clusters.

IV. Multidimensional reduction

The last step in the ELViM method is the reduction of the original multidimensional space, to a low-dimensional (2-3) representation. Multidimensional projection techniques address this question by mapping the data into p -dimensional spaces, with $p = \{1, 2, 3\}$, preserving

the maximum information about distance or dissimilarity between the data.⁶⁴

A multidimensional projection technique can be defined as follows: Let X be a set of N objects in $\mathbb{R}^m$, with $\delta : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ a criterion of proximity between two objects in $\mathbb{R}^m$, and Y a set of objects $\mathbb{R}^p$ with $p = 1, 2, 3$ and $d : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ a criterion of proximity in $\mathbb{R}^p$. A multidimensional projection technique can be described as function f , where $f : X \rightarrow Y$ which aims to make $|\delta(x_i, x_j) - d(x_i, x_j)|$ as close as possible to zero, $\forall x_i, x_j \in X$. In our case, this consists of a minimization procedure in which all distances between clusters in the real multidimensional phase space $\delta^{k,l}$ correspond to distances $d^{k,l}$ in a 2D phase space. These methods are generally known as Multi-Dimensional Scaling methods (MDS).^{65,66}

It is important to note that ELViM approach is very different from some dimension reduction methods such as PCA,^{67,68} in which one seeks component vectors to describe a data. Typically, the variables associated with these vectors are dihedral angles (Φ and Ψ) and specific distances between monomers of atoms.

MDS procedures usually vary in the way the pairwise distances among a set of N conformations are computed. The classic MDS approach uses a Euclidian distances in the high dimensional space,⁶⁹ and Metric MDS methods generalize optimization procedures using different cost functions.⁶⁶ Non-linear methods can be broadly classified into two groups: those that provide a mapping (either from the high-dimensional space to the low-dimensional embedding or vice versa), and those that just give a visualization (this is case for the ELViM method). The basic concept of this method is based on proximity data from a dissimilarity matrix (distance matrix). Given a chosen p dimension (usually 2 or 3), the algorithm places each object of the matrix in a position of the p -space dimension, then a minimization is made in order to preserve the distances contained in the original matrix. There are a variety of ways that this procedure is performed, in general, a pre-step is performed in the original matrix, followed by a classic MDS, for example, Isomap⁷⁰ uses the Floyd-Warshall algorithm to estimate the geodesic distance matrix, followed by classic MDS to project it into a low dimension; Locally-Linear Embedding (LLE)⁷¹ is very efficient when using sparse matrix

because it performs a close neighbor check before performing dimensional reduction.

The minimization procedure is analogous to the traveling salesman problem,^{72,73} in which given a set of distances, one tries to minimize the path to go through all of the cities, which is the cost function. In ELViM, the cost function E is associated with the position of the points in the 2D representation $s^{k,l}$, and it is given by

$$E = \sum_{(k,l)} |\delta^{k,l} - s^{k,l}|, \quad (3)$$

where the sum is carried out over all k and l pairs of conformations.

Though this step is also an essential part of the ELViM, it can be performed using a variety of minimization multidimensional projection techniques. The simplest method of them would be following a Monte Carlo/annealing, but it would be inefficient and time consuming. In the present work, we use the Force-Scheme technique,⁶⁴ which is a method based on a standard multidimensional projection technique. It is computationally fast and allows to work with sufficiently large number of conformational states (up to 50,000 points). It offers a balance between precision and computational performance. This iterative minimization technique establishes an approach in which the objects (conformations) are initially positioned, either in a random or heuristic way. Iterations are then performed to introduce perturbations in the positions of each object until the system reaches equilibrium; further details about this technique can be found in our recent work³¹ and discussion on system size restrictions can be found in Tejada et. al.⁶⁴

Acknowledgement

ABOJr was supported by CAPES - Brazil. VBPL was supported by the CNPq and FAPESP Grant 2018/18668-1, 2016/19766-1 and 2014/50739-5. PCW was supported by a NSF CAREER Award (grant No. MCB-1350312). This work was partially supported by the Center for Theoretical Biological Physics sponsored by the NSF Grant PHY-1427654.

Supporting Information Available

This material is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.
- (2) Thirumalai, D.; O’Brien, E. P.; Morrison, G.; Hyeon, C. *Annual Review of Biophysics* **2010**, *39*, 159–183, PMID: 20192765.
- (3) Gruebele, M.; Dave, K.; Sukenik, S. *Annual Review of Biophysics* **2016**, *45*, 233–251, PMID: 27391927.
- (4) Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. *Quarterly Reviews of Biophysics* **2014**, *47*, 285363.
- (5) Onuchic, J. N.; Wolynes, P. G. *Curr Opin Struct Biol* **2004**, *14*, 70–75.
- (6) Hills, R. D.; Brooks, C. L. *International Journal of Molecular Sciences* **2009**, *10*, 889–905.
- (7) Nymeyer, H.; García, A. E.; Onuchic, J. N. *Proc Natl Acad Sci U.S.A* **1998**, *95*, 5921–5928.
- (8) Koga, N.; Takada, S. *J Mol Biol* **2001**, *313*, 171–180.
- (9) Nettels, D.; Gopich, I. V.; Hoffmann, A.; Schuler, B. *Proc Natl Acad Sci U.S.A* **2007**, *104*, 2655–2660.
- (10) Chung, H. S.; Louis, J. M.; Eaton, W. A. *Proc Natl Acad Sci U.S.A* **2009**, *106*, 11837–11844.
- (11) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J Mol Biol* **2000**, *298*, 937–953.

- (12) Capraro, D. T.; Gosavi, S.; Roy, M.; Onuchic, J. N.; Jennings, P. A. *PLoS One* **2012**, *7*, e38512.
- (13) Chavez, L. L.; Onuchic, J. N.; Clementi, C. *J Am Chem Soc* **2004**, *126*, 8426–8432.
- (14) Takada, S.; Kanada, R.; Tan, C.; Terakawa, T.; Li, W.; Kenzaki, H. *Accounts of Chemical Research* **2015**, *48*, 3026–3035, PMID: 26575522.
- (15) Kolomeisky, A. B. *Motor Proteins and Molecular Motors*; Taylor & Francis, 2015.
- (16) Hirokawa, N.; Noda, Y.; Tanaka, Y.; Niwa, S. *Nature Reviews Molecular Cell Biology* **2009**, *10*, 682 EP –.
- (17) Chen, A. A.; García, A. E. *Proceedings of the National Academy of Sciences* **2013**, *110*, 16820–16825.
- (18) MacKerell Jr., A. D.; Banavali, N. K. *Journal of Computational Chemistry* **2000**, *21*, 105–120.
- (19) Noel, J. K.; Whitford, P. C. *Nature Communications* **2016**, *7*, 13314 EP –.
- (20) Di Pierro, M.; Zhang, B.; Aiden, E. L.; Wolynes, P. G.; Onuchic, J. N. *Proceedings of the National Academy of Sciences* **2016**, *113*, 12168–12173.
- (21) Head-Gordon, T.; Brown, S. *Current Opinion in Structural Biology* **2003**, *13*, 160 – 167.
- (22) Zhuravlev, P. I.; Papoian, G. A. *Q Rev Biophys* **2010**, *43*, 295–332.
- (23) Potoyan, D. A.; Papoian, G. A. *J Am Chem Soc* **2011**, *133*, 7405–7415.
- (24) Noé, F.; Clementi, C. *Current Opinion in Structural Biology* **2017**, *43*, 141 – 147.
- (25) Best, R. B.; Hummer, G. *Proc Natl Acad Sci U.S.A* **2016**, *113*, 3263–3268.
- (26) Chodera, J. D.; Noé, F. *Curr Opin Struct Biol* **2014**, *25*, 135–44.

- (27) Zimmerman, M. I.; Hart, K. M.; Sibbald, C. A.; Frederick, T. E.; Jimah, J. R.; Knoverek, C. R.; Tolia, N. H.; Bowman, G. R. *ACS Cent Sci* **2017**, *3*, 1311–1321.
- (28) Jacobs, W.; Shakhnovich, E. I. *J Phys Chem B* **2018**, *122*, 11126–11136.
- (29) Wales, D. J. *Phys Rev E* **2017**, *95*, 030105.
- (30) Wales, D. J. *Annu Rev Phys Chem* **2018**, *69*, 401–425.
- (31) Oliveira, A. B., Jr.; Fatore, F. M.; Paulovich, F. V.; Oliveira, O. N., Jr.; Leite, V. B. P. *PLoS One* **2014**, *9*, e100861.
- (32) Hardin, C.; Eastwood, M. P.; Prentiss, M. C.; Luthey-Schulten, Z.; Wolynes, P. G. *PNAS* **2003**, *100*, 1679–1684.
- (33) Chung, H. S.; Eaton, W. A. *Current Opinion in Structural Biology* **2018**, *48*, 30 – 39, Folding and binding in silico, in vitro and in cellula Proteins: An Evolutionary Perspective.
- (34) Tsai, J.; Levitt, M.; Baker, D. *Journal of Molecular Biology* **1999**, *291*, 215 – 225.
- (35) Borreguero, J. M.; Ding, F.; Buldyrev, S. V.; Stanley, H. E.; Dokholyan, N. V. *Biophys J* **2004**, *87*, 521.
- (36) Lammert, H.; Noel, J. K.; Onuchic, J. N. *PLoS Comput Biol* **2012**, *8*, e1002776.
- (37) Guinn, E. J.; Jagannathan, B.; Marqusee, S. *Nature Communications* **2015**, *6*, 6861 EP –.
- (38) Xu, W.; Harrison, S. C.; Eck, M. J. *Nature* **1997**, *385*, 595–602.
- (39) Olszewski, K. A.; Kolinski, A.; Skolnick, J. *Proteins* **1996**, *25*, 286–299.
- (40) Kachlishvili, K.; Maisuradze, G. G.; Martin, O. A.; Liwo, A.; Vila, J. A.; Scheraga, H. A. *PNAS* **2014**, *111*, 8458.

- (41) Noel, J. K.; Schug, A.; Verma, A.; Wenzel, W.; Garcia, A. E.; Onuchic, J. N. *J Phys Chem B* **2012**, *116*, 6880–6888.
- (42) Rodnina, M. V.; Wintermeyer, W. *arbc* **2001**, *70*, 415–435.
- (43) Schmeing, T. M.; Ramakrishnan, V. *Nature* **2009**, *461*, 1234–42.
- (44) Lai, J.; Ghaemi, Z.; Luthey-Schulten, Z. *American Chemical Society* **2017**,
- (45) Yang, H.; Perrier, J.; Whitford, P. C. *Proteins* **2018**, *86*, 1037–1046.
- (46) Stark, H.; Rodnina, M. V.; Rinke-Appel, J.; Brimacombe, R.; Wintermeyer, W.; van Heel, M. *Nature* **1997**, *389*, 403–406.
- (47) Nguyen, K.; Whitford, P. C. *Nat Commun* **2016**, *7*, 10586.
- (48) Yang, H.; Noel, J. K.; Whitford, P. C. *J Phys Chem B* **2017**, *121*, 10593–10601.
- (49) Jeong, K.-W.; Uzun, Ü.; Selmer, M.; Ehrenberg, M. *PNAS* **2016**, *113*, 13744–13749.
- (50) Kavaliauskas, D.; Chen, C.; Liu, W.; Cooperman, B. S.; Goldman, Y. E.; Knudsen, C. R. *Nucleic Acids Res* **2018**, *46*, 8651–8661.
- (51) Lai, J.; Ghaemi, Z.; Luthey-Schulten, Z. *Biochemistry* **2017**, *56*, 5972–5979.
- (52) Wales, D. J. *Mol Phys* **2002**, *100*, 3285–3305.
- (53) Neelamraju, S.; Wales, D. J.; Gosavi, S. *J Chem Inf Model* **2019**,
- (54) Sułkowska, J. I.; Rawdon, E. J.; Millett, K. C.; Onuchic, J. N.; Stasiak, A. *Proc Natl Acad Sci U.S.A* **2012**, *109*, E1715–E1723.
- (55) Jamroz, M.; Niemyska, W.; Rawdon, E. J.; Stasiak, A.; Millett, K. C.; Sułkowski, P.; Sułkowska, J. I. *Nucleic Acids Res* **2015**, *43*, issue.
- (56) Chavez, L. L.; Gosavi, S.; Jennings, P. A.; Onuchic, J. N. *PNAS* **2006**, *103*, 10254–10258.

- (57) Giri Rao, V. V. H.; Gosavi, S. *Proceedings of the National Academy of Sciences* **2018**, *115*, 1998–2003.
- (58) Noel, J. K.; Onuchic, J. N. *SpringerLink* **2012**, 31–54.
- (59) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M.; Smith, J.; M Kasson, P.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics (Oxford, England)* **2013**, *29*.
- (60) Noel, J. K.; Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. *Nucleic Acids Res* **2010**, *38*, Server.
- (61) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. *PLOS Computational Biology* **2016**, *12*, 1–14.
- (62) Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M. *Bioinformatics* **1999**, *15*, 327–332.
- (63) Lutz, B.; Sinner, C.; Heuermann, G.; Verma, A.; Schug, A. *Bioinformatics* **2013**, *29*, 2795–2796.
- (64) Tejada, E.; Minghim, R.; Nonato, L. G. *Information Visualization* **2003**, *2*, 218–231.
- (65) Trevor F. Cox, M. A. A. C. *Multidimensional Scaling*; Taylor & Francis, 2000.
- (66) France, S. L.; Carroll, J. D. *IEEE Trans Syst Man Cybern Part C Appl Rev* **2011**, *41*, 644–661.
- (67) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *The Journal of Chemical Physics* **2007**, *126*, 244111.
- (68) Riccardi, L.; Nguyen, P. H.; Stock, G. *The Journal of Physical Chemistry B* **2009**, *113*, 16660–16668, PMID: 20028141.
- (69) Torgerson, W. S. *Psychometrika* **1952**, *17*, 401–419.

- (70) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. *Science* **2000**, *290*, 2319–2323.
- (71) Roweis, S. T.; Saul, L. K. *Science* **2000**, *290*, 2323–2326.
- (72) Cook, W. J. *In Pursuit of the Traveling Salesman*; Princeton University Press, 2012.
- (73) Applegate, D. L.; Bixby, R. E.; Chvátal, V.; Cook, W. J. *The Traveling Salesman Problem: A Computational Study*; Princeton University Press, 2006; pp 583–593.