An Online Approach for Mining Collective Behaviors from Molecular Dynamics Simulations

Arvind Ramanathan[1], Pratul K. Agarwal[2], Maria Kurnikova[3], and Christopher J. Langmead[1,4⋆]

[1] Lane Center for Computational Biology, Carnegie Mellon University
[2] Computational Biology Institute, and Computer Science and Mathematics Division, Oak Ridge National Laboratory
[3] Chemistry Department, Carnegie Mellon University
[4] Computer Science Department, School of Computer Science, Carnegie Mellon University

Abstract. Collective behavior involving distally separate regions in a protein is known to widely affect its function. In this paper, we present an online approach to study and characterize collective behavior in proteins as molecular dynamics simulations progress. Our representation of MD simulations as a stream of continuously evolving data allows us to succinctly capture spatial and temporal dependencies that may exist and analyze them efficiently using data mining techniques. By using multi-way analysis we identify (a) parts of the protein that are dynamically coupled, (b) constrained residues/ hinge sites that may potentially affect protein function and (c) time-points during the simulation where significant deviation in collective behavior occurred. We demonstrate the applicability of this method on two different protein simulations for barnase and cyclophilin A. For both these proteins we were able to identify constrained/ flexible regions, showing good agreement with experimental results and prior computational work. Similarly, for the two simulations, we were able to identify time windows where there were significant structural deviations. Of these time-windows, for both proteins, over 70% show collective displacements in two or more functionally relevant regions. Taken together, our results indicate that multi-way analysis techniques can be used to analyze protein dynamics and may be an attractive means to automatically track and monitor molecular dynamics simulations.

1 Introduction

With the proliferation of structural information for over 50,000 proteins, a systematic effort to understand the relationship between a protein's three-dimensional structure, dynamics and function is underway. Molecular dynamics (MD) / Monte-Carlo (MC) simulations have become standard tools to gain insight into fundamental behavior of protein structures [30]. With increasing computational power, and the development of specialized hardware and software for MD simulations such as Desmond [15] simulations now easily scale to tens or even hundreds of nanoseconds regularly. The data from these simulations can easily reach several terabytes. Therefore, efficient methods to store, process and analyze this data are needed. There is also a growing interest for development of tools that monitor and track MD simulations, such that rare events within a protein simulation (e.g. a protein undergoing a conformational change) can be automatically detected [38].

Collective behavior in a protein refers to a group of amino-acid residues that may be spatially separate yet exhibit similar dynamics [13]. The similarity in dynamics refers to whether a group of residues are *constrained*, i.e. exhibiting small variance in distances with respect to other residues in the protein, or *flexible*, i.e., showing large variance in distances. Often residues at the interface of constrained/ flexible regions, known as *hinge-sites*, affect protein dynamics and function [22]. Collective behavior in a protein has been assessed primarily using techniques such as principal component analysis (PCA) [31, 29, 10]. Most techniques use a static structure (single snapshot) to analyze intrinsic dynamics and reason about collective behavior [11]. Other techniques use a collection of snapshots from a MD trajectory and perform PCA post-process [26]. The scientific community does not yet possess an efficient technique that provides information on collective behavior in a protein as the simulation is progressing. There are also no automated ways to track and monitor MD simulations on the basis of collective behavior observed in a protein.

⋆ Corresponding Author: cjl@cs.cmu.edu

This paper describes an *online* approach to characterize the collective behavior within proteins as MD simulations are progressing. In our approach, protein structures (snapshots) from a MD trajectory are modeled as a multi-dimensional array or tensor. This representation allows us to capture both spatial and temporal dependencies simultaneously as the simulation is evolving. Using recent advances in tensor analysis and data-mining we show that one can succinctly capture the dynamical behavior of a protein over the simulation. This dynamical behavior captured can be used to (a) conveniently visualize clusters within a protein that exhibit coupled motions or collective behavior, (b) identify residues that may play a significant role in protein's dynamics and (c) identify time-points during a simulation which exhibit a significant deviation from normal behavior of the protein.

Our contributions in this paper introduce a novel representation of protein simulations as streaming data. An approach to mine streaming data allows us to reason about parts of a protein that are more flexible versus parts of the protein that are less flexible. The characterization of flexible/ constrained regions in the protein match well with experimental and prior computational work. We also identify time-points during a simulation where there have been significant changes in the protein's dynamical behavior. The identification of such time-points can be potentially used to fork-off other simulations that may lead to better sampling of the protein's conformational space. Taken together, our approach shows that it is possible to reason about collective behavior as simulations are progressing and this may be of immense use to scientists wanting to understand complex phenomena in protein structures.

2 Tensor Representation of Protein Structures and MD simulations

Tensors are an extension of matrices beyond two dimensions and provide a convenient way to capture multiple dependencies that may exist in the underlying data. Formally, a tensor $\mathcal{X}$ of M dimensions can be defined as a multi-dimensional array of real values,

$$\mathcal{X} \in \Re^{N_1 \times N_2 \times \ldots \times N_M} \tag{1}$$

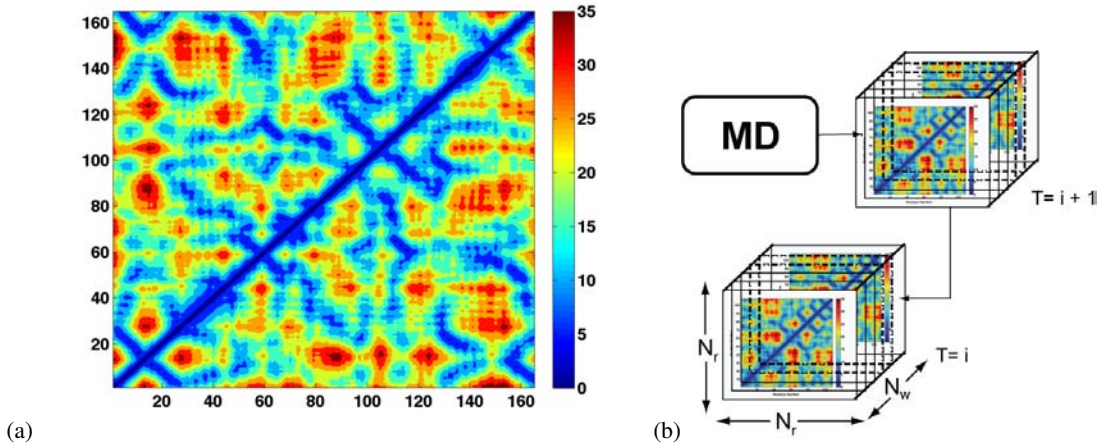
where N_i represents the i^{th} dimension for $(1 \leq i \leq M)$. A discussion of possible operations on a tensor is beyond the scope of this report, however, a review of tensors and related operations are provided in Kolda, et al [32]. In what follows, tensors are represented with calligraphic letters (e.g. $\mathcal{X}$), matrices by bold capital letters(e.g. $\mathbf{X}$), vectors as bold small letters ($\mathbf{x}$) and specific instances as normal text (x).

In a protein, apart from spatial dependencies, an explicit temporal dimension is needed to study the evolution of collective behavior over time. A protein's spatial description can be captured as a *distance map*. A distance map is a matrix where each entry (i, j) is the distance between atom i and j, defined as follows:

$$\mathbf{C}_{ij} = \sqrt{(\mathbf{r}_i - \mathbf{r}_j)^2} \tag{2}$$

where, $\mathbf{C}_{ij}$ is a second order tensor (matrix), $\mathbf{r}_i$ represents the position vector of atom i. For simplicity, we chose to represent the distances between C^α atoms, bringing the total number of entries in the $\mathbf{C}$ matrix to be $N_r \times N_r$, where N_r is the number of residues in the protein. Note that by representing the distances between all C^α atoms, we are able to account for local (immediate neighborhood) as well as global (over the entire protein) dependencies that may exist. An example distance map for the protein cyclophilin A (pdb id: 1AWQ) is shown in Fig. 1(a). There are distinct advantages of such a simple representation: one, distance maps are independent of rotations which may happen during the course of the simulation and two, distance maps also capture information about the entire conformation of a protein.

To capture temporal dependencies, we note that an MD simulation updates the coordinate positions at every time step t. An entire MD simulation can be thought of as a discrete collection of the distance maps $\mathbf{C}(t)$



(a) (b)

Fig. 1. Distance matrix and Tensor Representation of MD simulations. (a) Distance map representation of the enzyme cyclophilin A (pdb: 1AWQ). Distant residues are identified via darker shades of red. Distances are in $\mathring{A}$ (b) shows the streaming representation of MD simulations used in this paper. As new tensors keep arriving at every time interval $T = i + 1$, they are appended to the end of the current stream $T = i$. N_r is the number of residues, N_w represents the size of the window. This can be set by the end user depending on how often the user wants the analysis to run.

defined above. We may also define a discrete window in time, such that a time-slice from the MD simulation constitutes a third order tensor, with dimensions $N_r \times N_r \times N_w$, where N_r is the number of residues in the protein and N_w is the size of the window. A time-slice representation provides us with a mechanism to track collective behavior at short time scales and also at the time scale of the entire simulation.

If we define a discrete sized time window of size w, then the entire simulation can be considered a collection of $T = n/w$ tensors, where n is the total number of snapshots in the simulation. An ordered collection of such tensors is commonly referred to as a *tensor stream*. Formally a tensor stream is a discrete collection of M^{th} order tensors $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T$ where T is an integer that increases with time. The tensor stream representation used for MD simulation is illustrated in Fig. 1(b). Each tensor represents a discrete time window within the simulation. As the MD simulation progresses, more and more tensors become available, and the new tensors are appended to the end of the tensor stream.

3 Tensor Analysis of Protein Dynamics

A simple way to detect patterns is to analyze the overall variance in the underlying data. In two dimensions, PCA identifies patterns by minimizing the least-squared error with respect to the overall variance observed in the data. For tensors, it is possible to use an extension of PCA in multiple dimensions commonly referred to as *tensor analysis*. The objective function in tensor analysis is very similar to that of PCA - we minimize the error with respect to the observed variance in every one of the M dimensions. Formally, given a collection of tensors $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T$, each of dimension $N_1 \times N_2 \times \ldots \times N_M$, tensor analysis will determine orthogonal matrices $\mathbf{U}_i$ for each dimension N_i such that the reconstruction error e is minimized as follows:

$$^5 e = \sum_{t=0}^{T} \|\mathcal{X}_t - \mathcal{X}_t \prod_{i=1}^{M} \times_i (\mathbf{U}_i \mathbf{U}_i^T)\|_F^2 \tag{3}$$

[5] $\|\mathcal{X}\|_F^2$ is the square *Frobenius Norm* and is defined as $\|\mathcal{X}\|_F^2 = \sum_{i_1=1}^{N_1} \ldots \sum_{i_M=1}^{N_M} \mathcal{X}(i_1, \ldots i_M)^2$

The operation $\mathcal{X}_t \prod_{i=1}^{M} \times_i (\mathbf{U}_i \mathbf{U}_i^T)$ [6] yields the approximation of $\mathcal{X}_t$ spanned by the orthogonal matrices $\mathbf{U}_i$. The orthogonal matrices so determined will also reveal any underlying patterns that may be present within the data.

If all of the simulation data is already available, we may use multi-dimensional PCA to obtain insights into what parts of the protein are flexible or what parts of the protein are held relatively rigid. However, our goal is to characterize collective behavior in a dynamic environment such as MD. A recent algorithm called Dynamic Tensor Analysis (DTA) [42] has been successfully used to extract patterns from time-evolving data. This algorithm exploits two main features of the underlying data: (a) it is possible to quickly compute PCA in two dimensions if the variance matrices are available and (b) one can update the variance matrices *incrementally* without having to store any previous history. These two observations are particularly relevant for MD simulations: we compute only variance for pairwise distances, which is easy to compute. MD simulations are ergodic and therefore no historical information needs to be assessed. The algorithm we use to perform online analysis of MD simulations is outlined in Algorithm 1.

Algorithm 1 Online Analysis of Molecular Dynamic Simulations

1: **for** every incoming MD data at time-window T **do**
2: Convert MD data into tensor $\mathcal{C}^{(T)}$ of dimension $d = N_r \times N_r \times N_w$
{/* Perform Tensor Analysis using DTA*/}
3: **for** $i = 1$ to d **do**
4: Matricize $\mathcal{C}^{(T)}$ as $\mathbf{C}_{(i)}^{(T)} \in \Re^{(\prod_{j \neq d} N_i) \times N_i}$
5: Reconstruct variance $\mathbf{V}_{(i)}^{(T-1)} \leftarrow \mathbf{U}_i^{(T-1)} \mathbf{S}_i^{(T-1)} (\mathbf{U}_i^{(T-1)})^T$
6: Update variance matrix $\mathbf{V}_{(i)}^{(T)} \leftarrow \mathbf{V}_{(i)}^{(T-1)} + (\mathbf{C}_{(i)}^{(T)})^T \mathbf{C}_{(i)}^{(T)}$
7: Diagonalize $\mathbf{V}_{(i)}^{(T)} \leftarrow \mathbf{U}_i^{(T)} \mathbf{S}_i^{(T)} (\mathbf{U}_i^{(T)})^T$
8: **end for**
9: Calculate core tensor: $\mathcal{Y} = \mathcal{C}^{(T)} \prod_{i=1}^{d} \times \mathbf{U}_i^{(T)}$
{/* Identify restrained residues */}
10: Compute $\mathbf{f}_j = max(\mathbf{U}_i^{(T)}[j, :])$ to find maximum distance variance
11: **if** $\mathbf{f}_j \leq mean(\mathbf{f}_j) - std(\mathbf{f}_j)$ **then**
12: Identify residue j as restrained
13: **end if**
{/* Identify rigid and flexible regions */}
14: Use k-means algorithm to cluster $\mathbf{U}_r^{(T)}$.
15: Map output of k-means onto the protein structure to identify dynamically coupled regions
{/* Identify time points of interest */}
16: **if** $e_T \geq mean(e_i|_{i=1}^{T}) + \alpha.std(e_i|_{i=1}^{T})$ **then**
17: Identify T as an "event of interest"
18: **end if**
19: **end for**

A detailed description of DTA is provided in [42]; a brief description of DTA is given here (lines 3 to 9 in Algorithm 1). The algorithm proceeds by minimizing the variance in every dimension i, $(1 \leq i \leq M)$. The tensor $\mathcal{C}$ is matricized[7] in the selected dimension, say d to obtain matrix $\mathbf{C}_{(d)}$. Then, using the previous orthogonal matrices at time $T-1$, $\mathbf{U}_d$, the variance matrix is reconstructed to obtain $\mathbf{V}_d \leftarrow \mathbf{U}_d \mathbf{S}_d \mathbf{U}_d^T$. Using the matrix $\mathbf{C}_{(d)}$, we now update the variance matrix at time T in the d^{th} dimension as: $\mathbf{V}_d \leftarrow \mathbf{V}_d + \mathbf{C}_{(d)} \mathbf{C}_{(d)}^T$. We now diagonalize the updated variance matrix $\mathbf{V}_d$ to obtain the new projection matrices $\mathbf{U}_d$ and $\mathbf{S}_d$. In

[6] $\mathcal{X} \prod_{i=1}^{M} \times_i \mathbf{U}_i$ represents tensor multiplied by a set of matrices, defined as $\mathcal{X} \times_1 \mathbf{U}_1 \ldots \times_M \mathbf{U}_M$.

[7] Matricizing in dimension d is the process of unfolding a M^{th} dimension tensor $\mathcal{C}$ obtained by keeping d fixed and varying all other dimensions. Thus we obtain a vector in $\Re^{N_d}$, represented as $\mathbf{C}_{(d)} \in \Re^{(\prod_{i \neq d} N_i) \times N_d}$

order to capture a succinct representation of the data seen so far, we also construct a core tensor using $\mathcal{Y} = \mathcal{C} \prod_{i=1}^{M} \times \mathbf{U}_d$.

3.1 Outputs of DTA

The orthogonal matrices $\mathbf{U}_i$ describe the underlying correlations within the data. The first two orthogonal matrices are identical (since the data is symmetric). Both these matrices ($\mathbf{U}_r$) describe the inter-residue distance variance in time. The criterion for identifying rigid vs. flexible residues is shown in Algorithm 1 (lines 10 through 13). Lower values along the rows of the orthogonal matrices imply the residue is rigid, whereas, higher values in the orthogonal matrices imply that the residue is flexible. The third orthogonal matrix represents variances in time. This is used to identify time-points along the simulation where there is significant deviation in the protein's collective behavior.

It is also possible to cluster the orthogonal matrices to obtain insights into how different parts of the protein may be dynamically coupled (Algorithm 1; lines 14 and 15). Distance variance is a measure for how far two residues moved during the simulation with respect to each other. The correlations from the underlying variance must provide quantitative view of how much each residue moved with respect to its immediate neighbors as well as the entire protein. To visualize this, we used k-means clustering [25] to identify clusters of rigid/ flexible residues. Residues within the same cluster must exhibit similar fluctuations in distances. The individual cluster centers identify the average distance fluctuations within each cluster; larger distance fluctuations imply flexible regions, whereas smaller distance fluctuations imply rigid regions in the protein.

In order to identify time points where there are significant deviations from the normal dynamical behavior, we use the reconstruction error metric defined in Eqn. 3; (Algorithm 1, lines 16 through 18). This metric shows how different the incoming tensor was compared to the previous ones in the list. If the reconstruction error at time T is above α standard deviations from the mean reconstruction error observed thus far, we define T to be an event of interest. A formal definition of this threshold is given below:

$$e_T \geq mean(e_i|_{i=1}^{T}) + \alpha.std(e_i|_{i=1}^{T}) \tag{4}$$

3.2 Related Work

A number of techniques have been developed to analyze and reason about collective behavior from MD simulations. PCA based techniques (reviewed in [13]) are very popular in the MD community and are used to visualize collective behavior in proteins. Other techniques based on spectral analysis [16] as well as mutual information [34, 33] are also widely used in studying collective motions. However, there are two limitations to these approaches.

The first limitation relates to the fact that PCA can be done only in *two dimensions*; hence it is possible to obtain only insights into collective behavior from the perspective of spatial variance, no insight can be drawn from the temporal aspects of the simulation. This limitation can be overcome by using multi-dimensional PCA or multi-way analysis [39]. Multi-way analysis has been applied to assign nuclear magnetic resonance (NMR) spectra [35, 40]. These techniques have not been used to study protein dynamics to our knowledge, however they are popular within the cheminformatics community [24]. Tensor analysis has also been applied in a variety of problem domains including visual analysis of images from electroencephalogram [1] and systems biology [46]. The identification of rigid/ flexible sub-structures within a protein is a well studied problem, especially from the viewpoint of rigidity theory [44, 28]. Application of these techniques to multiple snapshots of proteins proves to be unreliable [36] and similarly, it is not possible to obtain information about time-points where deviation in collective behavior is observed.

The second limitation arises from the fact that analysis techniques developed for MD data can be applied *post-process*. The online techniques that are available provide only minimal feedback - either in terms of displacement vectors or individual snapshots from the simulation [12] and the end user is responsible to check whether these snapshots are of further interest; or improve performance of simulations [23]. By treating MD simulations as streaming applications, we have been able to apply techniques from data-mining to provide automated feedback to the end-user about how the behavior of the protein has changed with respect to time.

4 Implementation and Results

We were provided access to a set of MD simulations from two independent groups. This was done in order to ensure good variability in the available data and also to test the applicability of the method on different types of simulations. The first set of simulations constituted an equilibrium simulation for a 108 residue protein - barnase [36], with a total length of 10 nanoseconds (2,500 snapshots). The second set of simulations consisted of a free-energy profile for an enzymatic reaction catalyzed by the enzyme cyclophilin A (172 residues) [5], with a total of 18,000 snapshots. Both proteins have been studied extensively using both experimental and theoretical/ computational techniques and are ideal to test the utility of the technique outlined above. We first processed the MD trajectories to obtain the distance maps for every individual snapshot.

Third order tensors were then constructed with every 10 snapshots ($N_w = 10$) aggregated into one tensor, thus providing a total of 250 tensors for barnase. For cyclophilin A, we chose evenly spaced snapshots to yield 1800 snapshots to yield 180 tensors. The window size can be provided by the user; in several situations it is desirable to set larger window sizes. It is also dependent on the end-user how often the user would like to run the analysis. The tensors were processed using the tensor toolkit libraries [7, 8] and DTA algorithm [42] in MATLAB.

4.1 Analysis of Collective Behavior in Proteins

A plot of the rigid versus flexible residues for the protein barnase is shown in Fig. 2(a). An examination of the plot reveals that certain residues show lower distance variance within the protein. A low distance variance implies that the residue has a lesser degree of freedom in terms of moving around compared to the other residues and hence, is rigid. We considered those residues below one standard deviation interval from the plot. A total of 15 out of 108 residues were found to be constrained. The maximally constrained residues namely Tyr24, Ala32, Gly52, Gly53, Arg87 and Thr103 are known to be important for barnase's folding with Ala30 being a nucleation site [19]. Similarly, Gly52 and Gly53 form a hinge-site of the protein, implicated in the stabilization of the protein's motions [37]. It is also interesting to note that the catalytically important residue Glu71 is also constrained [20, 21], however Lys25 is not constrained to the extent of the residues that are implicated in protein folding or the functional hinge-site.

We now look at clusters of residues that are dynamically coupled. Using k-means clustering on the orthogonal matrices $\mathbf{U}_r$, we identified a total of four clusters. The selection of the best value of k for clustering was based on the mean value of cluster separation; for any value of k, if the mean value of the cluster separation was higher than that of $k-1$, another round of k-means clustering was applied, otherwise, k was chosen to be the optimal number of clusters. The assignment of clusters obtained from k-means was then mapped onto the three-dimensional structure of the protein and visualized using PyMOL [17]. As shown in Fig. 2(b), $\alpha 1$ and β-sheet ($\beta 1 - \beta 5$) form separate clusters. Residues 21-48 formed of $\alpha 2$ and $\alpha 3$ show higher flexibility than the hydrophobic core of the protein and is known to be a functional domain in the protein [45]. The residues 49-52 are clustered into three different groups. These residues form a functional hinge-site of the protein [45, 21]. The loop regions (L1-L4; N- and C-termini) are identified to be very flexible.

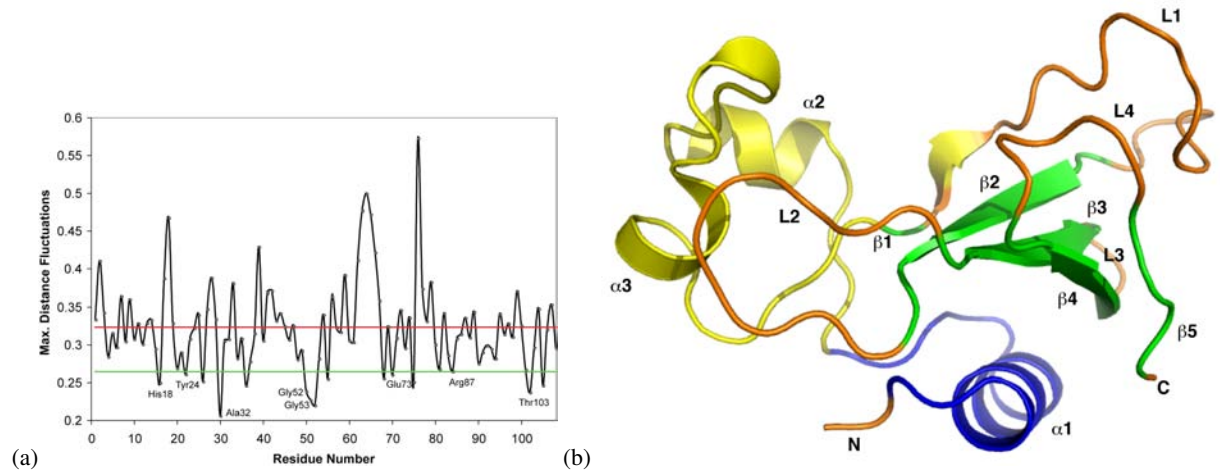


Fig. 2. Barnase constrained residues and flexible regions. (a) shows constrained residues in barnase. The red line indicates the mean distance variance and the green line indicates the first standard deviation interval below the mean value. Residues that are constrained are marked on the plot. Only a few residues are marked for clarity; also note that the first two residues were not present in the simulation. (b) indicates flexible clusters in barnase. Four clusters are identified; $\alpha 1$ and $\beta 1$-$\beta 5$ form two clusters shown in blue and green form the hydrophobic core of the protein, showing low distance variance. $\alpha 2$-$\alpha 3$ (residues 21-48) shown in yellow forms a separate functional domain and loops L1-L4 form the most flexible parts of the protein.

For the enzyme Cyclophilin A (Fig.3(a)), we were able to identify two distinct groups of residues; the first set of residues consisted of the enzyme itself. A second set of residues consisting of higher distance variance than the enzyme turned out to be the substrate bound to the enzyme. Of the enzyme residues, we identified a total of 24 constrained residues. Of these Arg55, Thr68, Gly74, Asn102 and Gly109, Leu122, His126, Val132 and Phe145 are part of a connected network of interactions identified in previous experimental [18] and computational work [5, 2] and are known to be of functional importance.

Based on our approach, we were able to identify six dynamically coupled regions within Cyclophilin A (Fig. 3(b)). Of these, the substrate and the N/C- termini of the protein formed a separate cluster. The β-sheet in the protein clustered into two different regions; the first cluster (shown in blue; $\beta 1$-2, $\beta 8$) represents the hydrophobic core of the protein. The second cluster (shown in cyan; $\beta 3$-7) represents the surface region of the protein in close contact with the substrate. The α-helices form two clusters; one constrained via hydrophobic interactions ($\alpha 1$) and held rigid whereas the second helix ($\alpha 3$) exhibits much more flexibility and coupled with the loop at the active site of the protein (L4, L7; shown in green). Note that the loops that belong to the same cluster show high correlations as noted in previous studies [5].

4.2 Comparing with Experimental/ Theoretical Work

A large amount of structural information for barnase is available within PDB [14]. Of the 55 structures available, we used a total of 6 structures (without mutations) for comparing the root mean square fluctuations (RMSF) to that of the distance fluctuations obtained from DTA. The RMSF values were normalized to compare the different RMSF values, as shown in Fig. 4(a). Note that there is a close correspondence in the location of the hinge site as well as the flexible regions of the protein. Further more, plotting the RMSF (Fig. 4(b) top panel) as well as DTA determined distance fluctuations (Fig. 4(b) bottom panel), we observe a close correspondence in the way distance fluctuations reflect the dynamic behavior of barnase. It is also important to note that many of the structures had the protein's substrate bound to it, hence small differences are visible in terms of the location of flexible regions; note that residues 71-81 are more flexible from our simulations.

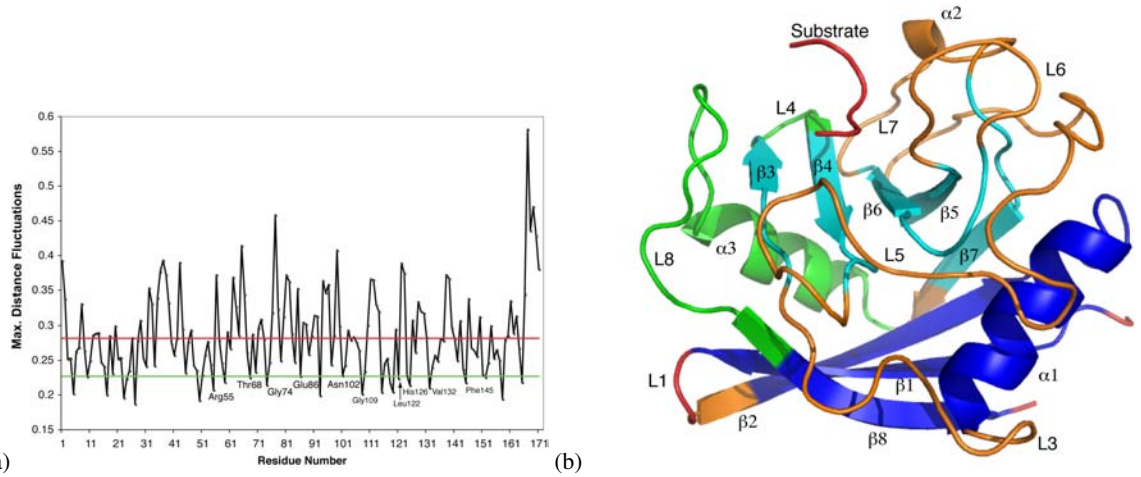


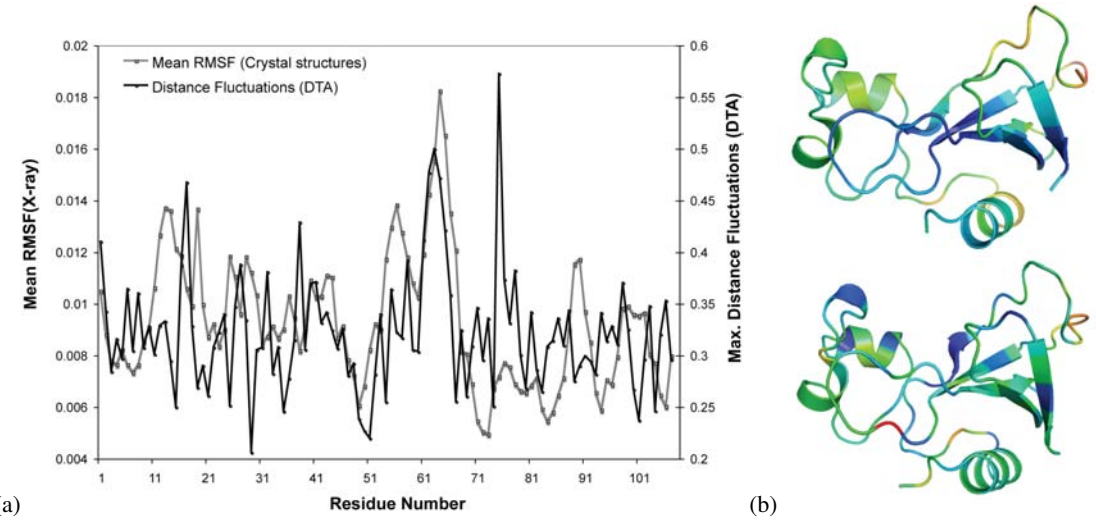
Fig. 3. Cyclophilin A constrained residues and flexible regions. (a) Constrained residues below the first standard deviation interval (green line) form a conserved network of interactions in the enzyme known to affect catalytic function [5]. (b) shows regions of the protein that are dynamically coupled; the hydrophobic part formed by β1-2, 8 and α1 experience low distance variance, β3-7 undergo slightly larger distance variance, L4 and L8 are grouped into one cluster, L3-7 are indeed very flexible whereas loop L1 and the substrate form the most flexible part of the protein.

We were able to compare our results with previously published work based on Gaussian Network Model (GNM) [9], MD simulations and NMR experiments [37, 48]. As illustrated in Fig. 5(a), an overlap of the root mean squared fluctuations (RMSF) determined from GNM and distance fluctuations computed via DTA from the MD simulation, show considerable agreement in the location of hinge site (Gly52/ 53). The identification of flexible loop regions also show good agreement. The residues 21-48 exhibited different dynamical properties compared to the rest of the protein and hence this region was clustered into a separate region via DTA. Previous experimental work [37, 48] also indicates that these residues are indeed dynamically distinct unit, forming a separate domain.

GNM uses the graph laplacian to predict positional fluctuations of residues based entirely on their topology. However, DTA's approach allows it to track distance variations observed during MD simulations and provides insights into constrained/ flexible regions. Note that while we have used only C^{α} distances as input to DTA, the method is quite general and can be used to track even hydrogen bond donor-acceptor distances/ hydrophobic interactions.

We were also able to compare for rigid/ mobile regions within barnase using the program Floppy Incursions and Rigid Sub-structure Topography (FIRST) [28, 27, 36]. An overlap of the flexibility scores determined via the Tool for Identifying Mobility in Macromolecular Ensembles (TIMME) as part of the FIRST suite of programs is shown in Fig. 5(b). Note that we used the same MD trajectory that we used for DTA as input to the program. A casual inspection of the plot indicates that there is good agreement between flexible regions of the protein. We also compared the rigid clusters formed by FIRST for barnase from previous work [28, 27, 36] and found that the largest rigid cluster identified by FIRST consisting of α1 helix and β1-β5 indeed exhibited overall smaller distance fluctuations in DTA. We also note that comparing our results with that of FIRST for cyclophilin A [47], we found good agreement in terms of the location of flexible versus constrained regions in the enzyme.

Note that FIRST can only show what parts of a protein tend to be constrained and cannot distinguish domains on the basis of dynamics exhibited by a protein. FIRST uses a local constraint counting approach to estimate the flexibility/ rigidity of a particular region within a protein and thus, cannot make predictions about



(a) (b)

Fig. 4. Baranse Experimental versus DTA comparisons. (a) DTA distance fluctuations are plotted (black solid line) against the average root mean square fluctuations (RMSF) determined from 6 different crystal structures of barnase. (b) shows regions of barnase that are flexible determined from X-ray crystallography (average B-factors of 6 structures). Note the close correspondence in the location of constrained regions. All the constrained residues identified from Fig. 2(a) are also constrained from the X-ray determined RMSF. Particularly, at the hinge site (48-51) as well as the flexible loop regions, there is good agreement. There is some disagreement in terms of the flexibility observed from the crystal structures in regions 71-81; this is mainly because the crystal structures had the substrate bound.

global constraints that may influence distally separated regions in the protein. On the other hand, DTA includes the notion of distance dependency between every C^{α} atom and hence global behavior from the simulation can be analyzed. Since the determination of constrained/ flexible regions is determined by an ensemble of structures, the approach is not sensitive to changes in placement of hydrogen or other atoms which is often a problem with approaches such as FIRST [36].

It is also relevant to point out the timescales of performing DTA and MD simulations. While our DTA was performed on a dual processor Intel machine running at 2.4 GHz, the MD simulations were performed on two different supercomputers. With code optimization, one could run about 1 ns of simulation on a dual core processor, it would take approximately 5 days to obtain 5 ns simulation for barnase. DTA takes less than 1 s to process a window from the simulation, and thus, we believe that there would be no significant overhead of using DTA to analyze the data.

4.3 Identifying events of interest in MD simulations

In order to identify time-points where there was significant deviation from the dynamical behavior observed so far, we plotted the reconstruction error(Eqn. 3) against the tensor windows. For barnase, as shown in Fig. 6(a), most tensors (236 or 94%) fall within the mean and second standard deviation interval. Beyond the second standard deviation interval, we find about 14 tensor windows. Of these small number of tensors, we found that about 70% of structures within the tensor windows show an average RMS deviation of about 1 Å compared to the immediate predecessor window; the rest of the structures show RMS deviation of ≥ 0.6 Å. We selected some of these structures for detailed analysis. In the tensor window 83 (highlighted by a red circle), the greatest distance deviation observed was the displacement of the flexible loops in residues 21-48 with respect to the loops L1 and L2 towards each other as seen in Fig. 6(b) (top panel). In another window (215), loop L1 had moved with respect to $\alpha 1$ helix compared to the predecessor window (214), illustrated in

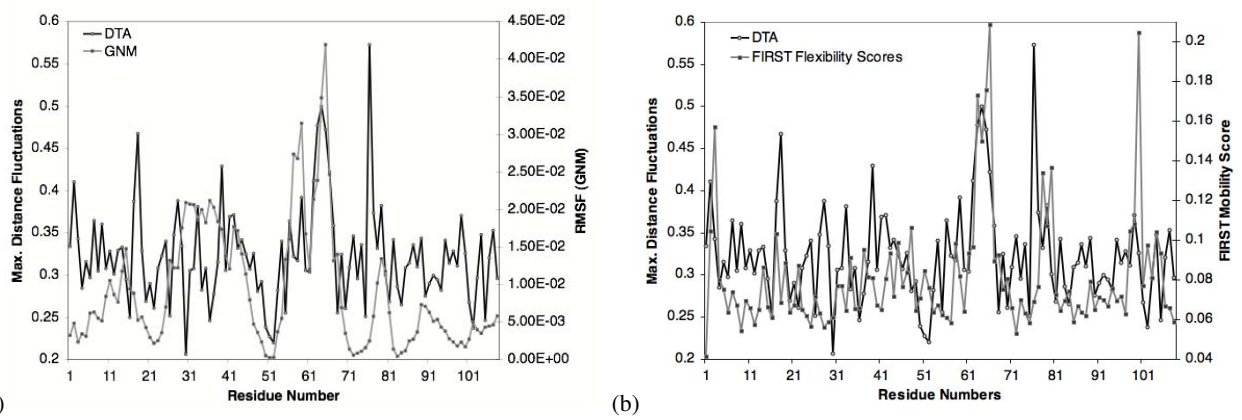


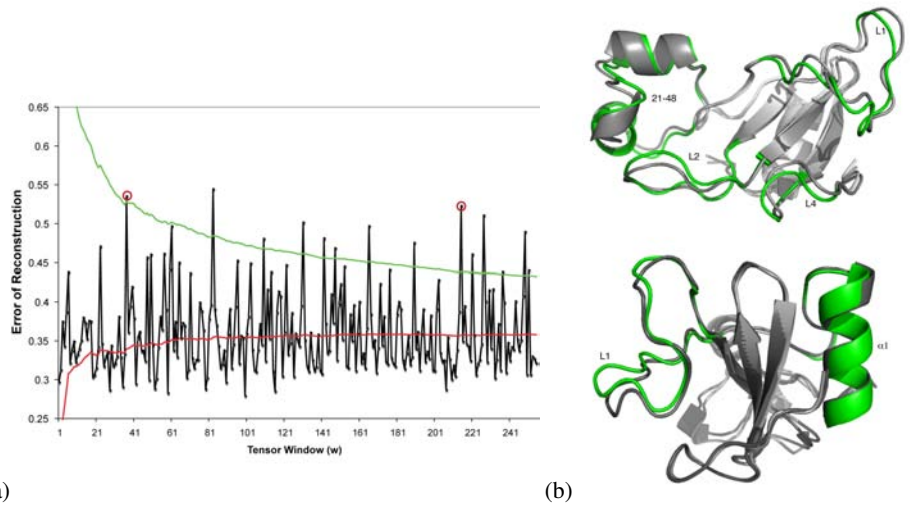
Fig. 5. Comparison of DTA to GNM and FIRST. (a) The predicted root mean squared fluctuations (RMSF) determined via GNM is plotted along with the distance fluctuations determined from DTA. Note the correspondence in flexible regions versus constrained regions. Closely agreeing regions observed from DTA and GNM are between 41-71, 76-84. Other regions show low positional fluctuations mostly depending on how they are connected to their neighbors. This is true especially with the secondary structure regions of the protein. (b) Constrained residues identified via FIRST and DTA. The flexible regions show close correspondence in both the methods, however, FIRST based flexibility scores do not correspond well with the hinge site of the protein. FIRST based rigid clusters identify the region from 48-51 to be highly flexible.

Fig. 6(b) (bottom). The collective displacement of these loops with respect to residues 21-48 is of importance since they represent inter-domain movements [37].

A similar analysis was done for cyclophilin A (Fig. 7(a)). We identified a total of 17 out of 180 tensors showing deviations beyond the second standard deviation interval. The average structural deviation in each of these windows was about 0.8 Å, compared to its immediate predecessor window. For example, a comparison of the structures from $w = 10$ with that of $w = 9$ (Fig. 7(b); top panel), indicated motions of the flexible loop L1 with respect to that of α3, L4 and L8, followed by motions across the L5-L6 regions of the protein. When we compared the two windows $w = 40$ with $w = 39$ (Fig. 7(b); bottom panel), we found that the substrate molecule showed large deviations with respect to both its interacting partners, namely β3 and L6-L8. The other windows reflected similar motions in these regions of the enzyme. The motions associated with these regions of the enzyme are known to be correlated with the catalytic process [5, 2] and therefore seem to indicate functional importance.

RMS deviations are a good measure of how a protein's structure evolves during the course of a simulation. However, it is important to note that RMS values do not provide any information on the collective behavior that influences spatially separate regions of a protein. To show that our approach is indeed sensitive to such changes, we provide two specific examples from each of our simulations. In barnase, the time point between $w = 30$ and $w = 40$ (Fig. 6(b) top panel), show an RMS deviation of 1.033 Å, whereas the window $w = 214$ and $w = 215$ (Fig. 6(b) bottom panel) shows an RMS deviation of just 0.612 Å. Although this RMSD is quite small, there is a large movement with respect to the loop structures of the protein.

In cyclophilin A, the time point between $w = 9$ and $w = 10$ shows an RMS deviation of 0.617 Å (Fig. 7(a) top panel) and the window $w = 30$ and window $w = 40$ shows an RMS deviation of 0.650 Å (Fig. 7(b) bottom panel). It is important to note that even though the overall RMSD for cyclophilin A in the two windows is relatively small, the substrate peptide highlighted in Fig. 7(b) (bottom panel), shows a distinct conformational rearrangement which could not have been detected by using RMS deviations alone. Since DTA incorporates the deviations in the distances to analyze simulations, we can keep track of changes that may relate two distally located residues.



(a) (b)

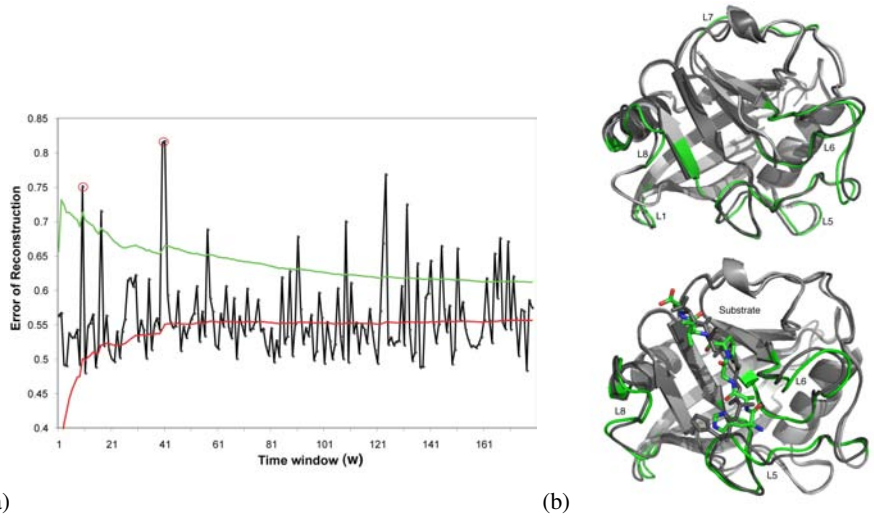
Fig. 6. Reconstruction Error for barnase. (a) shows the reconstruction error plotted as a function of tensor window. The red line indicates the mean reconstruction error as per Eqn. 3 and the second standard deviation above the mean is shown in green. Red circles are used to highlight those tensor snapshots shown in the adjacent panel. (b) Structural changes associated with $w = 40$ (top panel) showing movements in L1 and L2 along with the functional domain 21-48, overall RMSD 1.03 Å; $w = 215$ (bottom) showing movements associated with $\alpha 1$ and L1 with an overall RMSD of 0.612 Å. In both cases, the regions shown in a darker shade of gray represent predecessor windows whereas lighter shade shows the current window.

5 Conclusions

We have shown that it is possible to obtain biologically relevant information about flexible and rigid substructures within a protein as the simulations progress. This was illustrated on two entirely different sets of simulations and in both cases, we were able to identify dynamically coupled regions as well as constrained residues within the protein. We were also able to correlate the residues identified to be constrained to have some functional role in protein dynamics and function as verified via experimental data and prior computational work. We were also able to demonstrate that using the reconstruction error as a metric, the structures identified from different windows of the simulation showed significant deviations in two or more regions of the protein, indicating a change in collective behavior.

There are several problem domains where one may want to apply our approach. First, we note that although we used Cartesian distances as a means to capture spatial dependencies, the method is very flexible and can use any feature that can be represented as a tensor. For example, it may be beneficial to track forces or velocities over different atoms to detect patterns of energy flow in protein structures. We are also extending the method to track multiple features, such as distances between hydrogen bonds/ hydrophobic interactions in the same simulation. By tracking cross correlations that may exist between different streams of data, we may be able to provide detailed information to end-users about correlated changes in non-covalent interactions and how they account for large-scale motions within a protein. This is of fundamental interest to chemists and biologists alike in domains such as protein and drug design where networks of covalent and non-covalent interactions are known to widely affect protein function [4, 3, 48, 41].

The core tensor $\mathcal{X} = \mathcal{C} \prod_{i=1}^{d} \times \mathbf{U}_i$ can be used as an efficient means to approximate the dynamics thus far observed in a simulation to build linear kernels such that one may learn and reason about protein dynamics on a long time scale. This allows us to develop new techniques to perform unsupervised learning from such large-scale data. On the other hand, tensor analysis also provides an ideal handle for data reduction if one were to use a suitable rank approximation on the number of eigenvectors to approximate the space spanned



(a) (b)

Fig. 7. Reconstruction Error for Cyclophilin A. (a) shows the reconstruction error plotted as a function of tensor window. Red line: mean reconstruction error, green line: second standard deviation interval. Red circles are used to highlight those tensor windows shown in the adjacent panel. (b) Structural changes associated with cyclophilin A for two windows, namely $w = 10$ (top; overall RMSD 0.617 Å) and $w = 40$ (bottom; overall RMSD 0.650 Å). The structure from the predecessor window is shown in dark gray and the current window is shown in light gray; regions involved in collective movements highlighted in green. Note the large movement in the substrate molecule, shown as sticks in the bottom panel. This cannot be picked up using traditional metrics such as RMSD since it is only an average measure of structural deviations. However, tracking distance variations, we note a significant difference in the placement of the substrate.

by the current data. This is valuable in case of storing and processing simulation data, which can easily reach terabyte scales, even for small simulations.

The approach outlined here is an unsupervised learning for an MD simulation. The core tensor can be used to summarize the behavior of a protein during the course of a simulation. However, an extension to this method would allow one to check for consistency between two or more simulations based on the same protein. Also, recent work indicates that tensor based methods are being increasingly used in supervised learning [43]. It would be interesting to see if one could use such approaches to learn from MD simulations. This approach will be useful to predict the behavior of a protein over the course of a simulation.

Unlike GNM [9]/ ANM [6], our approach is not limited to studying only biophysical motions involved in binding or molecular recognition. The flexibility of our approach allows us to investigate motions involved in biochemical processes such as catalysis (as was done in this paper) and even protein folding simulations. This will allow one to generalize approaches such as GNM or ANM to time-series data, which may be of potential benefit to analyze complex biological processes.

From the perspective of analyzing simulations online, a direct extension to our method will be to fork simulations from the time-points where significant deviation in dynamics is observed in order to sample a larger conformational space. This in turn is particularly useful to drive simulations involving chemical reactions or folding pathways which are known to be hard to simulate. Further, the identification of *events of interest* within a simulation is particularly useful in folding applications where tracking/ monitoring structural changes is of prime importance. We propose to investigate folding trajectories in the near future to make useful predictions of how one may sample the conformational space and identify structural intermediates that may be important for the protein to fold correctly.

Acknowledgements

We thank Christos Faloutsos and Jimeng Sun from the Computer Science Department at Carnegie Mellon for introducing us to DTA and providing us with the implementation of DTA. This work is supported in part by US Department of Energy (DOE) Career Award and a grant from Microsoft Research to CJL. Pratul K. Agarwal would like to acknowledge the financial support from the Laboratory Directed Research and Development (LDRD) Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. DOE. We thank Tatyana Mamonova from the Kurnikova group for providing us access to MD simulations of barnase. We also thank Hetunandan Kamisetty from the Langmead lab for constructive discussions. We thank the anonymous reviewers for their valuable comments.

References

1. Evrim Acar, Canan Aykut-Bingol, Haluk Bingol, Rasmus Bro, and Bulent Yener. Multiway analysis of epilepsy tensors. *Bioinformatics*, 23(13):i10–18, 2007.
2. P. K. Agarwal. Cis/trans isomerization in hiv-1 capsid protein catalyzed by cyclophilin a: Insights from computational and theoretical studies. *Proteins: Struct., Funct., Bioinformatics*, 56:449–463, 2004.
3. P. K. Agarwal. Enzymes: An integrated view of structure, dynamics and function. *Microbial Cell Factories*, 5, 2006.
4. P. K. Agarwal, S. R. Billeter, P. T. R. Rajagopalan, S. Hammes-Schiffer, and S. J. Benkovic. Network of coupled promoting motions in enzyme catalysis. *Proc. Natl. Acad. Sci. USA*, 99:2794–2799, 2002.
5. P. K. Agarwal, A. Geist, and A. Gorin. Protein dynamics and enzymatic catalysis: Investigating the peptidyl-prolyl cis-trans isomerization activity of cyclophilin a. *Biochemistry*, 43(33):10605–10618, 2004.
6. A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.
7. Brett W. Bader and Tamara G. Kolda. Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4):635–653, December 2006.
8. Brett W. Bader and Tamara G. Kolda. Efficient MATLAB computations with sparse and factored tensors. *SIAM Journal on Scientific Computing*, 30(1):205–231, December 2007.
9. I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman. Vibrational dynamics of folded proteins. significance of slow and fast modes in relation to function and stability. *Phys. Rev. Lett.*, 80:2733–2736, 1998.
10. I. Bahar and Q. Cui. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. Mathematical and Computational Biology Series. Chapman and Hall/ CRC, New York, 2003.
11. I. Bahar and A. J. Rader. Coarse grained normal mode analysis in structural biology. *Cur. Op. Struct. Biol.*, 15:1–7, 2005.
12. David M. Beazley and Peter S. Lomdahl. Lightweight computational steering of very large scale molecular dynamics simulations. In *Supercomputing '96: Proceedings of the 1996 ACM/IEEE conference on Supercomputing (CDROM)*, page 50, Washington, DC, USA, 1996. IEEE Computer Society.
13. H. J. C Berendsen and S. Hayward. Collective protein dynamics in relation to function. *Current Opinion in Structural Biology*, 10(2):165–169, 2000.
14. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2002.
15. Kevin J. Bowers, Edmond Chow, Huafeng Xu, Ron O. Dror, Michael P. Eastwood, Brent A. Gregersen, John L. Klepeis, Istvan Kolossvary, Mark A. Moraes, Federico D. Sacerdoti, John K. Salmon, Yibing Shan, and David E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. *SC Conference*, 0:43, 2006.
16. J. D Chodera, N. Singhal, V. S. Pander, K. A. Dill, and W. C Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126:155101, 2007.
17. Warren L. DeLano. The pymol molecular graphics system, 2003.
18. E. Z. Eisenmesser, D. A. Bosco, M. Akke, and D. Kern. Enzyme dynamics during catalysis. *Science*, 295(5559):1520–1523, 2002.
19. A. R. Fersht and Valerie Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108(4):573–582, 2002.
20. A. R. Fersht, A. Matouschek, J. Sancho, L. Serrano, and S. Vuilleumier. Pathway of protein folding. *Faraday Discuss.*, 93:183–193, 1992.
21. Alan R. Fersht. Protein folding and stability: the pathway of folding of barnase. *FEBS Letters*, 325(1-2):5–16, 1993.
22. M. Gerstein and W. Krebs. A database of macromolecular motions. *Nucl. Acids Res.*, 26(18):4280–4290, 1998.
23. Weiming Gu, G. Eisenhauer, E. Kraemer, K. Schwan, J. Stasko, J. Vetter, and N. Mallavarupu. Falcon: on-line monitoring and steering of large-scale parallel programs. *Frontiers of Massively Parallel Processing, Symposium on the*, 0:422, 1995.

24. R. Gussio, N. Pattabiraman, G. E. Kellogg, and D. W. Zaharevitz. Use of 3d qsar methodology for data mining the national cancer institute repository of small molecules: Application to hiv-1 reverse transcriptase inhibition. *Methods*, 14:255–263, 1998.
25. J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *App. Stat.*, 28(1):100–108, 1979.
26. S. Hayward and N. Go. Collective variable description of native protein dynamics. *Annual Review of Physical Chemistry*, 46(1):223–250, 1995.
27. B. M. Hespenheide, A. J. Rader, M. F. Thorpe, and L. A. Kuhn. Identifying protein folding cores: observing the evolution of rigid and flexible regions during unfolding. *J. Mol. Graph. and Model.*, 21:195–207, 2002.
28. D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins: Struct., Funct., Genet.*, 44(2):150–65, 2001.
29. Ivan T. Jolliffe. *Principal Component Analysis*. Springer, 2002.
30. M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9:646–652, 2002.
31. Martin Karplus and Joseph N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.
32. Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. Technical report, Sandia National Laboratories, 2007.
33. O. F. Lange and H. Grubmuller. Full correlation analysis of conformational protein dynamics. *Proteins: Struct., Funct. and Bioinformatics*, 70:1294–1312, 2008.
34. T. Lenaerts, J. Ferkinghoff-Borg, F. Stricher, L. Serrano, J. W. H. Schymkowitz, and F. Rousseau. Quantifying information transfer by protein domains: Analysis of the fyn sh2 domain structure. *BMC Struct. Biol.*, 8:43, 2008.
35. D. Malmodin and M. Billeter. Multiway decomposition of nmr spectra with coupled evolution periods. *J. Am. Chem. Soc.*, 127(39):13486–13487, 2005.
36. T. Mamonova, B. Hespenheide, R. Straub, M. F. Thorpe, and M. Kurnikova. Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.*, 2(4):S137–47, 2005.
37. Svetlana B. Nolde, Alexander S. Arseniev, Vladislav Yu, and Martin Billeter. Essential domain motions in barnase revealed by md simulations. *Proteins: Struct., Funct. and Bioinformatics*, 46(3):250–258, 2003.
38. J. Shao, S.W. Tanner, N. Thompson, and T.E. Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation*, 3(6):2312–2334, 2007.
39. Age Smilde, Rasmus Bro, and Paul Geladi. *Multi-way Analysis: Applications in the Chemical Sciences*. J. Wiley and Sons, Ltd., 2004.
40. Doroteya Staykova, Jonas Fredriksson, Wolfgang Bermel, and Martin Billeter. ssignment of protein nmr spectra based on projections, multi-way decomposition and a fast correlation approach. *Journal of Biomolecular NMR*, 2008.
41. G. M. Suel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10:59–69, 2003.
42. Jimeng Sun, Dacheng Tao, and Christos Faloutsos. Beyond streams and graphs: Dynamic tensor analysis. 2006.
43. Dacheng Tao, Xuelong Li, Xindong Wu, Weiming Hu, and J. Maybank, Stephen. Supervised tensor learning. *Knowledge and Information Systems*, 13:42, 2007.
44. W. Whiteley. *Rigidity of Molecular structures: generic and geometric analysis*. Rigidity Theory and Applications. Kluwer Academic/ Plenum, New York, 1999.
45. H. Yanagawa, K. Yoshida, C. Torigoe, J. S. Park, K. Sato, T. Shirai, and M. Go. Protein anatomy: functional roles of barnase module. *J. Biol. Chem.*, 268(8):5861–5865, 1993.
46. Bulent Yener, Evrim Acar, Pheadra Aguis, Kristin Bennett, Scott Vandenberg, and George Plopper. Multiway modeling and analysis in stem cell systems biology. *BMC Systems Biology*, 2(1):63, 2008.
47. M. I. Zavodszky, M. Lei, M. F. Thorpe, A. R. Day, and L. A. Kuhn. Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins: Struct., Funct. and Bioinformatics*, 57(2):243–261, 2004.
48. Anastasia Zhuravleva, Dmitry M. Korzhnev, Svetlana B. Nolde, Lewis E. Kay, Alexander S. Arseniev, Martin Billeter, and Vladislav Yu Orekhov. Propagation of dynamic changes in barnase upon binding of barstar: An nmr and computational study. *Journal of Molecular Biology*, 367(4):1079–1092, 2007.