DYNAMIC INVARIANTS IN PROTEIN FOLDING PATHWAYS REVEALED BY TENSOR ANALYSIS

Arvind Ramanathan
Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA
Email: aramanat@andrew.cmu.edu

Christopher James Langmead*
Computer Science Department, Carnegie Mellon University, and Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA
**Email: cjl@cs.cmu.edu*

Recent advances in molecular dynamics simulation technologies (e.g., Folding@Home, NAMD, Desmond/Anton) have, for the first time, enabled scientists to perform all-atom simulations over timescales relevant to protein folding. Unfortunately, the concomitant increase in the size of the resulting data sets presents a barrier to understanding the molecular basis of folding. In particular, long simulations make it harder to identify and characterize important microstates, and the collective conformational dynamics that influence and enable the transitions between them. We address these problems by introducing a novel tensor-based method for performing a spatio-temporal analysis of protein folding pathways. We applied our method to folding simulations of the villin head-piece generated by the Pande group using Folding@Home. Using our method, we were able to identify three regions in this protein that exhibit similar collective behaviors across multiple simulations. We were also able to identify cross-over points in these simulations leading to different conformational subspaces. Our results indicate that these three regions may act as folding units, and that the observed collective motions may represent important dynamic invariants in the folding process. Thus, our spatio-temporal analysis method shows promise as a means for obtaining novel insights into protein folding pathways.

1. INTRODUCTION

The physical process by which a nascent polypeptide folds into a functional protein is a long standing question in biology[10]. Failure to fold into a functional protein is linked to cell toxicity as well as several diseases. Given the enormous importance attached to correct folding pathways within a cell, there is considerable interest in understanding all-atom mechanisms of how a protein folds into its functionally relevant conformations.

Protein folding pathways span over 15 orders of magnitude in time, ranging from femto-seconds to even seconds and beyond. Experimental techniques (e.g., FRET), while extremely useful in providing certain insights into the folding process, generally probe a narrow range of time-scales and lack full atomic detail. Fortunately, recent advances in both hardware (e.g., Anton[9]) and software (e.g., Folding@Home[12], NAMD[6], Desmond[1]) have enabled, for the first time, all-atom molecular dynamics (MD) simulations over time scales relevant to folding. Unfortunately, the concomitant increase in size and complexity of the resulting data sets presents considerable barrier to understanding the molecular basis of folding pathways. In particular, long simulations make it harder to characterize important microstates and the collective conformational dynamics that influence and enable transitions between them.

In this paper, we address this challenge by introducing a novel tensor-based method for performing spatio-temporal analysis of protein folding pathways. Previously, our approach has proven successful in characterizing collective conformational dynamics in equilibrium simulations across multiple proteins[7, 8]. Our method is also capable of detecting changes in these collective motions, signaling a change between meta-stable states in the energy landscape. Taken together, our approach provides a unique way to analyze MD simulations while providing biologically relevant insights.

Here, we analyze protein folding pathways of a 35-residue fast-folding variant of villin head piece which is known to fold into three α-helices[2]. A large repository of folding pathways for this protein has been re-

*Corresponding author.

cently made available by Pande and co-workers using the Folding@Home platform. For these folding pathways, we identify meta-stable states and simultaneously characterize the collective conformational dynamics in these states.

2. METHODS

Our approach to analyze protein folding pathways builds a multi-dimensional representation of these trajectories. Tensors are an extension of matrices beyond two dimensions and provide a convenient means to capture multiple dependencies that may exist in the underlying folding pathway. 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 protein's spatial description can be captured as a *distance map*. 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. 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. The end-user may choose an appropriate time-slice depending on the type of simulation; for this paper, we chose a time-slice of 100 snapshots from the simulation.

A simple way to detect patterns is to analyze the overall variance in the underlying data. For tensors, it is possible to use an extension of PCA in multiple dimensions commonly referred to as *tensor analysis*[11]. 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 error of reconstruction (EoR) e is minimized as follows:

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

We note that for MD simulations and protein folding pathways, we need not store any historical information about how the simulations have progressed. Hence, it is possible to analyze the data as and when it becomes available using an online algorithm called *dynamic tensor analysis*. The algorithm, at every step constructs the variance matrix and does a principal component analysis, which provides insights into collective motions in the MD simulations seen until the current time window.

The eigenvectors from the resulting analyses describe inter-residue distance fluctuations; lower values indicate residues are constrained while large values indicate higher flexibility in the protein. The eigenvectors can also be clustered to identify regions of a protein that show coupled motions. The EoR defined in Eq. 2 can identify time-points during which collective motions during the simulations have significantly changed. Snapshots between two time-points with significant deviations in EoR identify meta-stable states along a folding pathway, which may show collective dynamical in common.

3. RESULTS

The data consists of 8 different runs, each with a unique starting structure. Each run is in turn comprised of 100 different clones, while having the same starting structure, differ in their initial velocities. The simulations represent a total of over 354 μs, with an average length of 863 ns[2]. This unprecedented scale of data available provides an ideal platform to understand the nature of collective conformational dynamics and their effects on the folding process. Here, we have analyzed a total of 10% of the total data, corresponding to a total of 35 μs. Even within this data, we observe considerable diversity in the structural and dynamical characteristics of the folding pathways.

3.1. Dynamic Invariants along Folding Pathways

The inter-residue distance fluctuations (of C^{α} atoms) from the eight folding pathways reveals several similarities in the dynamical nature of several residues (Fig. 1). For example, runs 4 and 7 exhibit strong similarities in terms of the average dynamical fluctuations, exhibiting a correlation coefficient of 0.95. Significantly, the starting structures for runs 4 and 7 are very different (see Fig. 2).

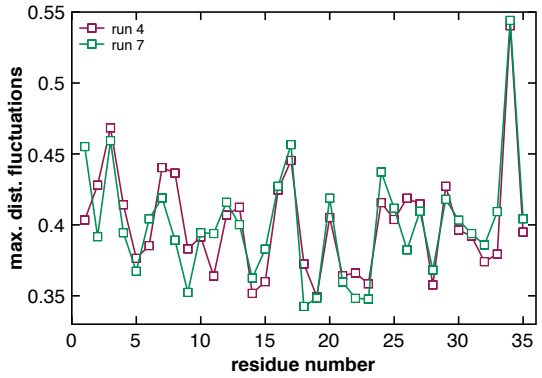

Fig. 1. Similarity between fast folding pathways. The maximum inter-residue distance fluctuations is plotted for the fast folding pathways. Run 4 is shown in purple, while run 7 is shown in green.

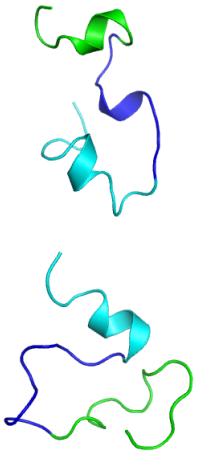

Fig. 2. Collective dynamics in fast folding pathways. Three dynamically coupled clusters of residues are identified via tensor analysis for run 4 (top) and run 7 (bottom). The cluster boundaries corresponding to the folded protein are: α_1: 2-10, α_2: 13-20 and α_3: 21-33. Note, the clusters are shown on the starting structures, but correspond to the three α helices in the folded structure.

This similarity in collective conformational dynamics becomes clearer when we examine the dynamically coupled regions along these two folding pathways. In a majority of the clones from runs 4 and 7, the residues cluster into three regions. These clusters are destined to form the three α helices in the native state ensemble (Fig. 2). An obvious question, therefore, is to ask whether such collective behaviors constitute "*dynamic invariants*" along the folding pathways.

To answer this question, we examined those runs that eventually led to the native folded structure (runs 0, 2, 3, 5 and 6), albeit more slowly than runs 4 and 7. We observed that the slow pathways did, in fact, exhibit collective behaviors similar to those of the fast pathways. Specifically, the collective behaviors of three regions (residues 19-22, 27-30, and 32-25; Fig. 3 highlighted in blue rectangles), which flank the hydrophobic core of the folded protein have similar dynamics in both slow and fast pathways.

There are also significant dynamic differences between the various runs. For example, the dynamics of residues 6-16 exhibit considerable diversity among the runs (see Fig. 3, orange rectangle). Moreover, dynamic differences are observed between the fast and slow folding pathways. Specifically, dynamic couplings exist between residues destined to form the three α-helices are observed in the slow folding pathways, that are not seen in the fast folding pathways (Fig. 1). Conversely, fast folding pathways exhibit a distinct collective pattern that defines the three helices to be independent folding units early on in the simulations.

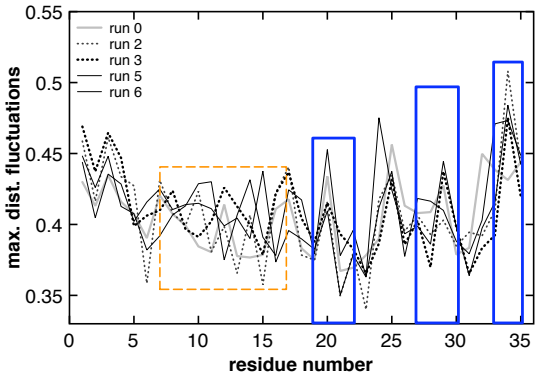

Fig. 3. Comparing slow folding pathways. Three stretches of residues highlighted by blue rectangles show similar collective distance fluctuations over the course of the folding. In contrast, residues 6-16 exhibit diverse collective distance fluctuations in all the simulations.

We also compared the dynamics of those runs that did not lead to folded structures to those that did (see Fig. 4. When we compared run 1, for example, to those that did fold, we observed that the distance fluctuations between residues that formed α_2 and α_3 in the folded protein were much larger compared to the other pathways. This feature is not observed in any other run, raising the possibility that the inability of the two regions to come together (and satisfy the dynamic invariant) leads to a misfolded state.

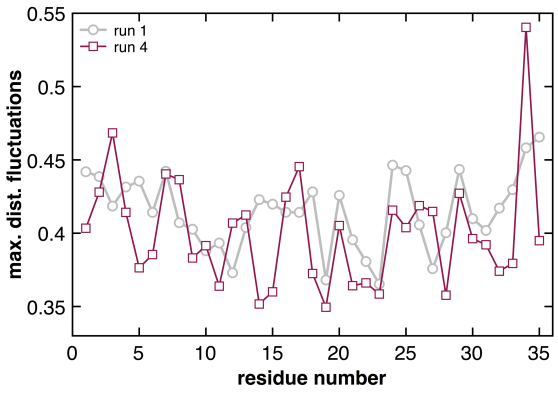

Fig. 4. Comparing a non-folding pathway to fast folding pathway. Non-folding run 1 shows distance fluctuations in residues from α_1 and α_2 that are substantially larger than those from run 4.

3.2. Identifying of meta-stable states along Folding Pathways

The error of reconstruction (EoR) metric from Eq. 2 can be used to identify the presence of meta-stable states in the folding pathways. For this paper, we present an analysis of the EoR metric to identify meta-stable states along the fast folding pathways (4 and 7), since they are of particular interest. In Figs. 5 and 6, we show two clones selected from runs 4 and 7 for which the EoR metric is plotted.

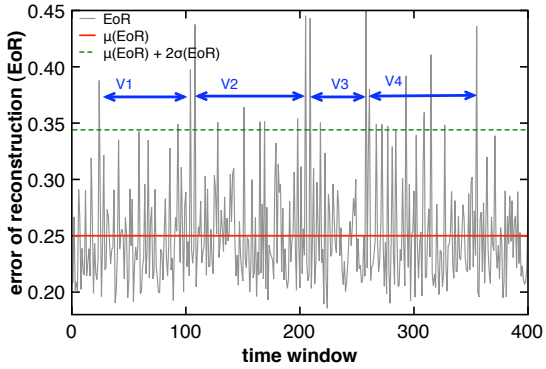

Fig. 5. Meta-stable states in run 4. The EoR plot vs. time reveals the complex dynamics of the folding pathway. Four meta-stable states can be identified (V1 through V4). Note each of the meta-stable states spans a certain time-scale indicated by the blue arrows.

The meta-stable states in each of these folding pathways were identified by examining the structural transitions at each of the peaks in Figs. 5 and 6. Based on our analysis, we were able to identify four regions (V1 through V4) in run 4 and six regions (V1 through V6) in run 7.

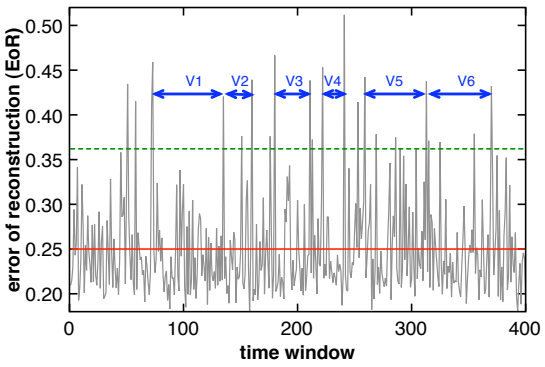

Fig. 6. Meta-stable states in run 7. The EoR plot vs. time reveals the complex dynamics of the folding pathway. Six meta-stable states can be identified (V1 through V6). Note each of the meta-stable states spans a certain time-scale indicated by the blue arrows.

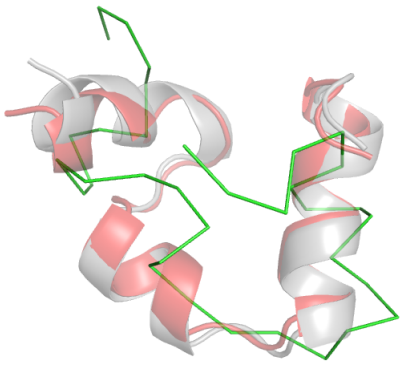

Fig. 7. Structural characteristics of meta-stable states in run 4. The initial structure for run 4 is shown in green ribbon representation. Meta-stable state 1 (V1) is identified by the presence of a partial α_1. Meta-stable state 2 (V2) is comprised of α_1 and α_3 coming closer to each other. Note, here we have shown the average backbone structure of the two states.

Closer examination of the structures in V1 in run 4 revealed the presence of a partial α_1 and a partial α_2. This feature remains invariant until the folding protein makes a transition to a second meta-stable state (V2), where we observe a partial movement of a well-formed α_1 towards α_3 (see Fig. 7). Further, the meta-stable states in run 4 are more stable; the average timescale of V1 through V4 is about 1.6 μs (computed from the width between EoR

spikes). In run 7, we observe that V1, V5 and V6 have an average timescale of 1.1 μs, where as V2, V3 and V4 have a lifetime of about 0.4 μs. This diverse sampling of the folding landscape by run 7 may explain the experimentally observed double exponential kinetics[2].

4. CONCLUSIONS

We have presented a novel approach to analyze collective conformational dynamics along folding pathway for a small autonomously folding protein. Our method has the benefit of being run *online* and in *parallel*. It can be used to suitably analyze and monitor large-scale folding simulations for (a) characterizing collective behavior, (b) identifying meta-stable states and (c) comparing multiple folding simulations for identifying common dynamical patterns.

Previous studies had revealed the heterogeneity in the kinetics of villin folding. Our analysis complements the analysis of kinetics by revealing the complex nature of collective dynamics involved in protein folding. Even though the initial conditions for several folding pathways were very different, we observed that there is significant similarity in collective dynamics exhibited by fast folding pathways (runs 4 and 7). Further, folding pathways that did not show the collective dynamics exhibited by the fast folding pathways either took a long time to converge to the folded ensemble (runs 0, 2, 3, 5 and 6) or did not sample the native state (run 1). This suggests the possibility that certain dynamical features might accelerate the folding process. In case of the villin head piece, we identified some dynamic invariants involving residues flanking the hydrophobic core. These same residues are known to stabilize the interactions necessary to sample the native state ensemble.

Along the folding pathways, we were also able to analyze and characterize multiple meta-stable states that show both common structural and dynamical features. For the two fast folding pathways, we observe that even though the collective dynamics is similar, the two pathways exhibit considerable variation in the timescales related to the transitions and existence of meta-stable states. Taken together, our analysis shows promise in revealing novel insights into the dynamical features of the folding landscape of this protein.

5. ONGOING WORK

We are in the process of analyzing multiple protein folding pathways to characterize the overall collective dynamical features for villin folding. Future work on the computational aspects will involve making the analysis platform compatible with multiple simulation software as well as parallelizing the code for exploiting both Folding@Home platform as well as supercomputers. Since the tensor analysis is a variant of principal component analysis, it would also be possible to compress and store only the essential aspects of a simulation, without having to store large trajectory files.

On the scientific front, the use of tensor analysis offers several exciting opportunities. It is well known that solvent fluctuations enslave the protein folding process[3]. Hence, one important aspect of our study would be to analyze the collective behavior of the solvent and its effect on the folding process. It will also be interesting to analyze effects of mutations to the villin head piece[5] and how that affects the collective conformational dynamics in the folding process. Further, our rigorous statistical mechanics based approach to model the protein ensembles[4] could be used here to reason about the enthalpic, entropic and free-energetic contributions that allow the protein to sample the native state ensemble.

We also note that the feature that we chose to analyze here was the collective distance fluctuations across multiple folding pathways. However, tensor analysis is quite flexible and can allow for the analysis of even more interesting features such as electrostatics, forces, and even energy flux. We are in the process of extending the code to handle these features to further our understanding of the protein folding process.

References

1. K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC '06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, page 84, New York, NY, USA, 2006. ACM.
2. D. L. Ensign, P. M. Kasson, and Vijay S. Pande. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics of a fast-folding variant of the villin headpiece. *J. Mol. Biol.*, 374:806–816, 2007.
3. H. Frauenfelder, P. W. Fenimore, G. Chen, and B. H. McMahon. Protein folding is enslaved by solvent motions. *Proc. Natl. Acad. Sci. USA*, 103:15469–15472, 2006.

4. H. Kamisetty, E. P. Xing, and C. J. Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. *J. Comp. Biol.*, 15(7):755–766, September 2008.
5. H. Lei, C. Wu, H. Liu, and Y. Duan. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 104(12):4925–4930, 2007.
6. J. C. Philips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. V. Kale, and K. Schulten. Scalable molecular dynamics with namd. *J. Comp. Chem.*, 26(16):1781–1801, 2005.
7. A. Ramanathan, P. K. Agarwal, M. Kurnikova, and C.J. Langmead. An Online Approach to mine for collective behaviors from molecular dynamics simulations. In *Proc. of the 13th Conference on Computational Molecular Biology (RECOMB)*, pages 138–154, 2009.
8. A. Ramanathan, P. K. Agarwal, and C. J. Langmead. Using tensor analysis to characterize contact-map dynamics in proteins. Technical report, Carnegie Mellon University, 2008.
9. D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. In *ISCA '07: Proceedings of the 34th annual international symposium on Computer architecture*, pages 1–12, New York, NY, USA, 2007. ACM.
10. C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande. How well can simulation predict protein folding kinetics and thermodynamics? *Annual Review of Biophysics and Biomolecular Structure*, 34(1):43–69, 2005.
11. J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: Dynamic tensor analysis. 2006.
12. B. Zagrovic, C.D. Snow, M.R. Shirts, and V.S. Pande. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *Journal of Molecular Biology*, 323(5):927 – 937, 2002.