Probing Protein Folding Dynamics Using Multivariate Statistical Techniques

Ahmet Palazoglu*, Yaman Arkun, Burak Erman*****
Attila Gursoy****

**Dept. of Chemical Engineering and Materials Science, University of California, CA 95616 USA (Tel: 530-752-8774; e-mail: anpalazoglu@ucdavis.edu)*
***Dept. of Chemical and Biological Engineering, Koç University, Turkey (e-mail: yarkun@ku.edu.tr)*
****Dept. of Chemical and Biological Engineering, Koç University, Turkey (e-mail: berman@ku.edu.tr)*
***** Dept. of Computer Engineering, Koç University, Turkey, (e-mail: agursoy@ku.edu.tr)}*

Abstract: The study of protein folding and its ramifications in biological contexts is at the heart of computational biology. In this paper, we discuss a number of tools in systems engineering that would provide an analysis framework to help explain the observed dynamic behavior of the protein, ultimately making the connection between protein structure and functionality. A case study of villin headpiece folding using principal components analysis as well as clustering demonstrates the potential of these tools in responding to this challenge.

Keywords: optimal folding trajectories, dynamic simulations, principal components analysis, clustering.

1. INTRODUCTION

The study of proteins is easily justified by the fact that they constitute an essential element of all living beings. Specifically, proteins are responsible for controlling gene expression, allow transmission of signals between cells and organs, transport and store other species and defend the body against microbes, among many other functions. Thus, due to their universal significance, understanding the relationships between their sequence, configurations and the vital functions they play in the body, can help development of new therapies and novel biomaterials. As such, uncovering the mysteries of proteins requires an interdisciplinary approach, enlisting not only biologists and medical professionals but also engineers, mathematicians and computer scientists. Recent studies include bioinformatics approaches that explore data mining (Brito, Dubitzky et al. 2004) and evolutionary algorithms (Pal, Bandyopadhyay et al. 2006), in addition to structure prediction problems (Krogh, Larsson et al. 2001; Floudas 2007) and computational techniques focusing on optimization (Krogh, Larsson et al. 2001; Greenberg, Hart et al. 2004).

All protein molecules are chains of amino acids and referred to as linear heteropolymers due to the unbranched nature of their monomeric units (amino acids) (Creighton 1993). The amino acid building blocks consist of a central α-carbon (C^{α}) atom surrounded by four groups: an amino group ($-NH_2$), a carboxyl group (-COOH), a hydrogen atom and a fourth group (-R) that can be one of twenty specific molecules, and is referred to as the side group. The specific side group gives the amino acid its unique characteristic. The sequence of amino acids (also called residues) as read from the amino (N-terminus) to the carboxyl (C-terminus) is referred to as the primary structure. Helices, β-strands, loops, etc. are the secondary structures. Organization of the secondary structures in space to form a stable 3-D structure leads to the tertiary structure. The lowest free energy tertiary structure is the unique native conformation with which the protein performs its function.

The type of a protein and its folding characteristics are determined by its primary structure, i.e., the sequence of amino acids (Dill, Bromberg et al. 1995). It is also noted that the folding process is often aided by molecular chaperons that help the protein fold correctly as it exits the ribosome by minimizing the influence of other nearby proteins as well as by binding to the protein to prevent misfolding (Shinde and Inouye 2000). This is especially important as incorrectly folded proteins resulting from errors during folding are responsible for illnesses such as Creutzfeldt-Jakob disease, Bovine spongiform encephalopathy, Parkinson's and Alzheimer's diseases. Due its implications in understanding such diseases, the dynamics of folding has received substantial attention in recent years (Karplus and Kuriyan 2005; Colombo and Micheletti 2006).

The dynamics of protein folding have been studied extensively in vitro, where the protein is denatured to assume an arbitrary initial configuration and then as the natural conditions are restored, folds into its native configuration. This refolding process has been explored both by molecular dynamics (MD) simulations (Duan and Kollman 1998; Pan and Daggett 2001; Mori, Colombo et al. 2005) and using mostly stop-flow experiments and NMR spectroscopy (Eaton, Thompson et al. 1996; Plaxco and Dobson 1996) and the results provided unique insight towards the folding

dynamics. During the refolding process, the simulations explore the conformational energy landscape accessible to the protein molecule and all-atom MD simulations with explicit solvent can only feasibly achieve time scales shorter than about 1μs for relatively small proteins which leaves out a number of phenomena inaccessible and poorly understood.

In this paper, we show how systems engineering tools can be used to probe the dynamics of protein folding to provide a better understanding of the key mechanisms. The next section discusses protein folding simulations and the type of information gathered as a result. Dynamic folding trajectories that result from such simulations can be interrogated by a number of analysis tools, and we focus on the use of Karhunen-Loeve and clustering to extract spatial and temporal features to help explain the folding dynamics.

2. SIMULATIONS OF PROTEIN FOLDING

Folding of a protein takes place in the form of a competition between the loss of configurational entropy and the decrease of energy due to the formation of inter-residue contacts. Consequently, a free energy barrier separates the unfolded and folded states. The energy surface, as a function of the variables active in folding, such as the $3N$ coordinates of an N atom protein and a multitude of additional dimensions describing the surrounding water molecules, is called the 'energy landscape'. The competition of entropy and energy results in a rugged landscape, and leads to transient trapping of structures that are either partially folded or misfolded. A comprehensive account of protein folding simulations can be found in a recent article (Scheraga, Khalili et al. 2007).

The protein can be modeled at different levels of complexity ranging from all-atom to coarse-grained representations. The all-atom visualization coupled with MD simulations gives the most detailed picture of folding but the computational time is a serious bottleneck. The only full-trajectory molecular dynamics simulation in the presence of explicit water up to this date is that of a 35-residue protein (Duan and Kollman 1998). In typical coarse-grained approximation approaches (Haliloglu and Bahar 1998; Doruker, Jernigan et al. 2002), the protein consists of N beads that represent the amino acids joined into a linear chain by virtual bonds analogous to the chemical bond. A virtual bond joins two consecutive alpha carbons, C^{α}, along the chain. The length of a virtual bond is fixed, a condition referred to as the 'fixed bond length condition'. Each bead has a finite volume. No bead shares its own volume with any other bead. This is called the 'excluded volume condition'. Folding of the protein progresses from a random initial state at $t = 0$ to the final state at $t = t_f$, subject to the fixed bond length and excluded volume conditions at all stages of folding. Folding to the native configuration requires the specification of interactions between pairs of amino acids. This information is based on empirical energy functions, chosen such that the unique native state corresponds to the minimum of total energy (Erman and Dill 2000).

The protein folding problem in its simplest form may be viewed as *a constrained optimization problem:* We are given an initial configuration of N beads connected in the form of a linear chain. The beads want to move towards their specified final destinations by spending minimal energy subject to the (i) connectivity between beads, (ii) fixed bond length and (iii) excluded volume constraints. Each bead obeys Newton's second law of motion throughout the folding trajectory. The forces acting on each bead are received either from the other beads or they are external interaction forces with the environment. Under these conditions, one needs to determine the optimal forces acting on the beads and the resulting optimal trajectory of the beads from their initial configuration to their final native states.

Here, we analyze the optimal pathways followed by the protein during folding. These pathways were generated using the optimal control framework proposed in our earlier work (Guner, Arkun et al. 2006). A coarse grained dynamic model based on Newton's equation of motion is used to make the dynamic optimization manageable.

3. INTERROGATION OF SIMULATION DATA

While simulations provide a wealth of data on the nature of protein motion, extraction of useful information that would shed light on the dominant folding/unfolding mechanisms, evolution of interactions among key residues such as those that determine the hydrophobic core, as well as understanding of the structural conformations and their relationship with biological function is non-trivial. The computational burden and complexity are significant barriers; thus, methods that help reduce dimensionality and provide analytical capabilities in a low-dimensional subspace are largely used. Here, we discuss two techniques. Karhunen-Loeve expansion (KLE) or Principal Components Analysis (PCA) can extract key coordinates (modes) that govern the global motion of the protein. Clustering helps classify large-scale correlated motions that can explain the presence of meta-stable states in which certain protein configurations exist and evolve.

3.1 Principal Components Analysis

The application of PCA to the study of macromolecular motion dates back many years where MD simulations were studied to identify fluctuation modes (Garcia 1992) and to extend simulation time scales (Amadei, Linssen et al. 1993). In the latter work, the conformational space is subdivided into an 'essential' subspace (Van Aalten, De Groot et al. 1997) which contains only a few degrees of freedom, exhibiting unharmonic motion and a 'residual' subspace where the fluctuations are Gaussian. Recent studies explore the energy landscape and the conformational states (Alakent, Doruker et al. 2005; Mu, Nguyen et al. 2005), identify modes contributing to protein fluctuations in MD simulation of apo-adenylate kinase (Lou and Cukier 2006), and extract key mechanistic features from simulations of chemotrypsin inhibitor 2 (Palazoglu, Gursoy et al. 2004). One can also refer to a comprehensive review article for further details (Stein, Loccisano et al. 2006).

The data matrix can be constructed in various forms depending on the information desired. For example, we can construct a matrix of spatial positions as they evolve in time.

The simulations yield the position vectors of N residues for M time intervals and, after subtracting the temporal mean, this results in a $M \times K$ array, where $K = 3N$. One can also use the fluctuation matrix which captures the jump dynamics governing protein folding. The fluctuation matrix becomes $M \times K$ with $K = N$ and has been previously studied (Palazoglu, Gursoy et al. 2004). Another possibility is to form a matrix in which temporal evolution of the magnitude of the distance between the contact pairs is captured. This matrix would have M time intervals and K would correspond to the total number of short- and long-range contact pairs.

The expansion has K modes (eigenvector directions) and each eigenvalue measures the mean energy of its corresponding mode. Among the class of all linear expansions, KLE is optimal in the sense that, on a subspace of lower dimension $L < K$, it retains the most energy possible. One can retain only the first few L modes that extract the important trends and filter the details deemed insignificant by the user.

3.2 Clustering

Cluster analysis (Everitt, Landau et al. 2001) is a class of statistical methods that seeks to partition a set of N observations (objects) into distinct groups. Each observation corresponds to a particular sampling interval (distinct period in time) for which corresponding measurements are available on the same set of parameters. One of the early applications of clustering to MD simulation data is by Karpen et al. (Karpen, Tobias et al. 1993) where feature vectors (dihedral angle time series) are clustered for a 2.2 ns trajectory of the small peptide YPGDV to identify conformational states during unfolding. When applied to protein models, clustering can classify ensembles of structural models based on their backbone structure, using C^{α} distances as the dissimilarity measure (Domingues, Rahnenfuhrer et al. 2004). The molecular motion of proteins can also be classified using clustering to identify functionally relevant structures (Pan, Dickson et al. 2005) and to gain insight towards the shape of the energy landscape (Plaku, Stamati et al. 2007).

Agglomerative hierarchical clustering is used to identify sampling intervals exhibiting similar 'behavior' based on a chosen metric. It accepts as input a symmetric matrix $\boldsymbol{D}$ whose elements D_{ij} indicate the relative dissimilarity between sampling intervals i and j. Matrix $\boldsymbol{D}$ can derive from various parameters in a given simulation, such as the dihedral angles, internal coordinates and potential energies, and must be properly defined for the cluster solution to be physically meaningful. The hierarchical clustering starts with all objects residing in their own cluster, and by using various linkage rules, proceeds by merging the closest objects, and subsequently, the closest clusters, finally terminating when all objects are collected under a single cluster. The output of the hierarchical clustering algorithm is graphical in nature (a dendrogram), and facilitates the visualization of recurring phenomena manifested in the data. Another popular method, k-means clustering (Everitt, Landau et al. 2001), creates clusters based on the maximization of between-cluster variance and minimization of the within-cluster variance, and often gets trapped in local extrema, requiring multiple initializations. The number of clusters needs to be specified a priori for the k-means algorithm and it starts by randomly populating these clusters and proceeding by the optimization step to reform the clusters. These shortcomings were overcome in a recently proposed aggregated k-means clustering strategy (Beaver and Palazoglu 2006) where an ensemble of cluster solutions, generated by performing many randomly initialized runs of the algorithm, can be aggregated to form a single, hierarchical solution. A recent study discusses the performance of different clustering algorithms applied to MD trajectories (Shao, Tanner et al. 2007).

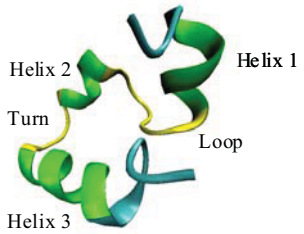


Fig. 1. The structure of villin headpiece.

4. CASE STUDY OF CHICKEN VILLIN HEADPIECE

We consider a 36-residue protein, (PDB code 1Vii.pdb), chicken villin headpiece that is the smallest protein that can fold autonomously. It has been shown through a landmark all-atom explicit water simulation of villin headpiece (Duan and Kollman 1998) that there is a sudden initial hydrophobic collapse followed by longer structural adjustment phase. Other simulation studies also agree with the folding events revealed by Duan and Kollman, e.g., the implicit-water simulation by Shen and Freed (Shen and Freed 2002) and MD scheme integrated with Monte Carlo search by Mori et al. (Mori, Colombo et al. 2005).

Chicken villin headpiece (Figure 1) has 3 short helices, Helix 1, 2 and 3 which contain the residues 4-8, 15-18, and 23-32, respectively. They are held together by a loop between residues 9-14, and a turn between residues 19-22.

Fig. 2. Snapshots from the folding process starting from an arbitrary initial configuration, $t = 0$, followed by $t = 30$, $t = 60$, and $t = 90$, and $t = 150$.

4.1 Folding Trajectories

The optimal folding trajectories were calculated starting from several random initial configurations (Guner, Arkun et al. 2006). Each simulation is performed for 301 time steps. Results include the optimal values for both the position of each bead and the force applied to each bead as a function of time. Figure 2 shows a representative result where the initial denatured configuration is significantly stretched out and the protein starts to establish the helices first. Once the helices form, the loop and the turn secondary structures begin to get established. Finally, the native 3-D structure is reached after refinement of the overall configuration.

The root-mean-square-distance (RMSD) is the distance between the native structure $S_1 = (s_{11} \quad s_{12} \quad \ldots \quad s_{1N})$, and a folding structure $S_2 = (s_{21} \quad s_{22} \quad \ldots \quad s_{2N})$, where s_{ij} is the position of the j^{th} bead in structure i:

$$RMSD = \sqrt{\frac{2}{N(N-1)} \sum_{i}^{N} \sum_{j>i}^{N} \left(\| s_{1i} - s_{1j} \| - \| s_{2i} - s_{2j} \| \right)^2}$$

Figure 3 shows the RMSD variation between the native structure and simulated structures with respect to time for the whole chain. On the average, initial configurations fold around time step 100, with an average final RMSD of 3Å.

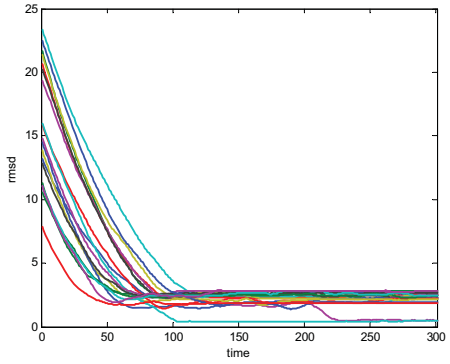
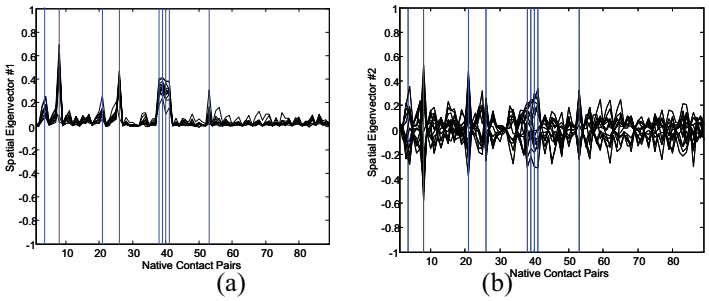


Fig. 3. The evolution of RMSD over all simulations.

4.2 KL Analysis

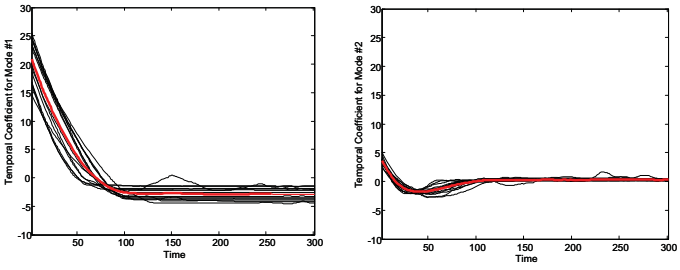
The long-range contact pairs are defined as those residues that are 5 and more residues apart. For chicken villin, there are a total of 89 native contact pairs and 8 are considered as long-range contacts: 2-34, 7-14, 7-34, 10-33, 10-34, 11-33, 11-34, 19-26. Here, we consider the temporal evolution of the magnitude of the contact pairs, r_{ij}'s. The matrix is 301×89 with 301 rows for the time steps and 89 columns for each pair of native contacts. For this analysis, we focused on 14 simulations. We found that, in general, 2 modes capture 99% of the variance in the simulation data. Figure 4 shows the first and second spatial eigenvectors indicating which native contact pairs contribute to these directions the most. The major contributions to the first come from the long-range contacts, as indicated by the vertical lines. The other contact pairs have generally minor contributions to this direction, perhaps with the exception of the pair 2-7 (position 4 on the plot). It is reported (Frank, Vardar et al. 2002) that three phenylalanine residues F47, F51, and F58 (residues 7, 11 and 18) make up the bulk of the hydrophobic core along with the hydrophobic residues L42, V50, and L69 (residues 2, 10 and 29). Thus, it is noteworthy that the first mode is significantly influenced by the interaction between residues 2 and 7 as the hydrophobic collapse occurs. Another observation is that all pairs load positively in this direction, indicating that all move in the same direction, effectively in the direction of reducing the distances among contact pairs. In fact this coordinated collapse is observed in Fig. 1. Another important observation is that this loading behavior is independent of the initial configuration, underscoring the fundamental nature of the collapse. In the second mode, the influence of the long-range contacts is attenuated (especially for 10-33, 10-34, 11-33 and 11-34) and short-range contacts become more important, almost across the board for all such contacts. The native contact pair 2-7 still retains its influence. The contact pairs load in both positive and negative directions, a key difference from the first mode, indicating a more complex motion. It is also important to note that this loading depends on the initial configuration, implying that the formation of secondary structures can follow different paths in time.



(a) (b)

Fig. 4. First (a) and second (b) spatial eigenvectors, vertical lines indicate the position of long-range native contact pairs, and vertical line at position 4 points to the native contact pair 2-7.

As shown in Fig. 5, the temporal coefficient of the first spatial eigenvector decays exponentially, indicating that the manner with which energy is minimized is common for all simulations regardless of the initial configuration. This also supports the all-positive loading directions of the contact pairs as shown in Figure 4a. On the other hand, the temporal coefficient of the second spatial eigenvector shows second-order behavior and is attenuated significantly, underscoring the lesser influence of the second mode. This mode explains the fast dynamics associated with the short-range contacts as secondary structures (helices in this case) are made quickly and then readjusted to conform to the overall formation of the protein structure. Folding dynamics exhibit two time-scales, which is consistent with the two-step folding mechanism of the hydrophobic collapse model (Baldwin 2002).



(a) (b)

Fig. 5. First (a) and second (b) temporal coefficients, bold red line indicates the averages.

4.3 Cluster Analysis

To demonstrate the potential of clustering, we focus on native contact distances analyzed before. For a given simulation, the expectation is to see if the dynamic signature of contact distances can be used to label each contact pair as belonging to a class, distinguished by its characteristic temporal evolution or contact distance. Thus, the feature vector is the

time series of contact distance magnitudes, r_{ij}'s. The simulation data from the 14 runs were stacked into a matrix of dimension 4214×89 and the data were normalized to [0, 1]. The data matrix is then transposed for clustering the 89 contact pairs. Thus, the scaling is in reality performed on the *rows* of the clustered data matrix, as opposed to for the columns as is more typical. A different scaling is used in this analysis because the variables have different mean levels although they are measured in the same units.

Using aggregated *k*-means clustering with average linkage, the dendrogram in Figure 6 is obtained. It shows two coarse and six relatively distinct fine clusters with a cophenetic coefficient (Beaver and Palazoglu 2006) of 0.96, which indicates that the dendrogram is a good representation of the relationships among the objects. The aggregated distances show that the cluster members have short merging distances while the main clusters merge at relatively large distances. This shows that the within-cluster variance is low while the between-cluster variance is high.

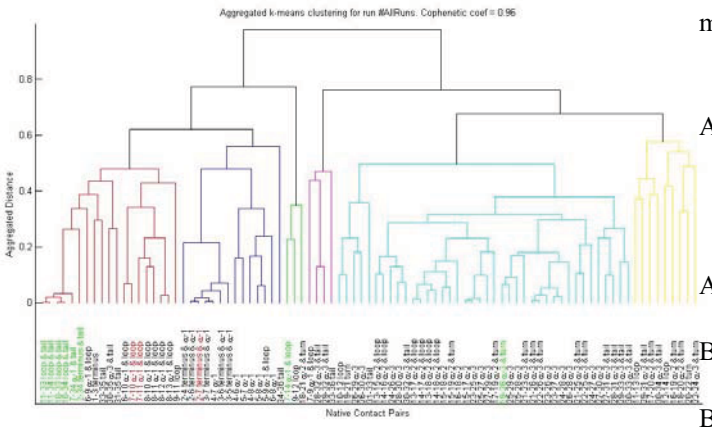


Fig. 6. The dendrogram for clustering the distances of all native contact pairs for 14 simulations. Contacts labelled as green are the long-range native contacts whereas the contacts labelled as red are the ones involving residues in the hydrophobic core.

In Figure 6, it is noted that a large number of the residues that form the long-range native contacts and the hydrophobic core fall in the same cluster. Indeed, first cluster (labeled as red) contains six of the eight long-range contact pairs along with two short-range contact pairs that contain hydrophobic core residues. This cluster captures the concerted motions of the loop and the tail, as well as the loop and helix-1 contacts. The second cluster (dark blue) is largely residues involved in forming the helix-1. It is noted that the long-range native contact pair 7-14 (PHE7-THR14) appears in a rather isolated third cluster (green) and its motion shows greatest similarity to contact pairs 9-12 and 18-21. The long-range interaction, i.e., 7-14, is the only long-range interaction close to the N-terminal, thus having dynamic characteristics different than the other long-range contacts is expected. The small fourth cluster (magenta) contains residues towards the end of the protein chain and also notably differs from the first three clusters. The fifth cluster (light blue) contains all the remaining short-range contacts associated with helix 2 and helix 3, including the long-range contact 19-26. This long-range contact pair shares characteristics common with the other helix 3 contact pairs, thus dynamically acts in concert with a large number of short-range contact pairs specific to the secondary structure with which it is associated. The final cluster (yellow) brings together short-range contacts primarily involving turn and tail secondary structures.

We must re-iterate that a coarse clustering decision indicates two main clusters, containing the subclusters (1, 2, 3) and (4, 5, 6), respectively. Such a grouping would suggest two classes of residues where the first mainly contains the long-range-contacts and the second, the short-range contacts. Yet, the fact that specific residues (such as long-range contacts) do not appear all in a single cluster and usually appear together with other residues is supported by (Larson, Ruczinski et al. 2002) who claim that both poorly and highly conserved residues are equally likely to participate in the protein folding nucleus. They also note, however, that there is an observable bias in the mean sequence conservation of the residues in the folding nuclei. This is especially consistent with the membership of the large cluster on the left.

REFERENCES

Alakent, B., P. Doruker, et al. (2005). "Application of time series analysis on molecular dynamics simulations of proteins: A study of different conformational spaces by principal component analysis " J. Chem. Phys. **121**(10): 4759-4769.

Amadei, A., A. B. Linssen, et al. (1993). "Essential dynamics of proteins." Proteins **17**: 412-425.

Baldwin, R. L. (2002). "Making a network of hydrophobic clusters." Science **295**(5560): 1657-8.

Beaver, S. and A. Palazoglu (2006). "A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area." Atmospheric Environment **40**(4): 713-725.

Brito, R., W. Dubitzky, et al. (2004). "Protein folding and unfolding simulations: A new challenge for data mining." Omics - J Integrative Biology **8**(2): 153-166.

Colombo, G. and C. Micheletti (2006). "Protein folding simulations: combining coarse-grained models and all-atom simulations." Theor. Chem. Acc. **116**: 75-86.

Creighton, T. E. (1993). Proteins: Structures and Molecular Properties. New York, W.H. Freeman.

Dill, K. A., S. Bromberg, et al. (1995). "Principles of protein folding - a perspective from simple exact models." Protein Science **4**: 561-602.

Domingues, F. S., J. Rahnenfuhrer, et al. (2004). "Automated clustering of ensembles of alternative models in protein structure data." Protein Engineering **17**: 537-543.

Doruker, P., R. L. Jernigan, et al. (2002). "Dynamics of large proteins through hierarchical levels of coarse-grained structures." J. Comp. Chem. **23**: 119-127.

Duan, Y. and P. A. Kollman (1998). "Pathways to a protein folding intermediate observed in a 1-Microsecond simulation in aqueous solution." Science **282**(5389): 740-744.

Eaton, W. A., P. A. Thompson, et al. (1996). "Fast events in protein folding." Structure **4**: 1133-1139.

Erman, B. and K. A. Dill (2000). "Gaussian theory of protein folding." J. Chem. Phys. **112**: 1050-1056.

Everitt, B., S. Landau, et al. (2001). Cluster Analysis. New York, NY, Oxford.

Floudas, C. A. (2007). "Computational methods in protein structure prediction." Biotech. and Bioeng. **97**(2): 207-213.

Frank, B. S., D. Vardar, et al. (2002). "The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain." Protein Science **11**(3): 680-687.

Garcia, A. E. (1992). "Large-amplitude nonlinear motions in proteins." Phys. Rev. Lett. **68**: 2696-2699.

Greenberg, H., W. Hart, et al. (2004). "Opportunities for combinatorial optimization in computational biology." Informs J Computing **16**(3): 211-231.

Guner, U., Y. Arkun, et al. (2006). "Optimum folding pathways of proteins. Their determination and properties." J. Chem. Phys. **124**: 134911.

Haliloglu, T. and I. Bahar (1998). "Coarse-grained simulations of conformational dynamics of proteins: application to apomyoglobin." Proteins: Structure, Function, and Genetics **31**: 271-281.

Karpen, M. E., D. J. Tobias, et al. (1993). "Statistical clustering technique for the analysis of long molecular dynamics trajectories: Analysis of 2.2-ns trajectories of YPGDV." Biochemistry **32**: 412-420.

Karplus, M. and J. Kuriyan (2005). "Molecular dynamics and protein function." PNAS **102**(19): 6679-6685.

Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes " J. Mol. Biology **305**(3): 567-580.

Larson, S. M., I. Ruczinski, et al. (2002). "Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation." J Mol Biol **316**(2): 225-33.

Lou, H. and R. I. Cukier (2006). "Molecular dynamics of apo-adenylate kinase: A principal component analysis." J. Phys. Chem. B **110**: 12796-12808.

Mori, G. M. S., G. Colombo, et al. (2005). "Study of the villin headpiece folding dynamics by combining coarse-grained Monte Carlo evolution and all-atom molecular dynamics." Proteins: Structure, Function, and Bioinformatics **58**: 459-471.

Mu, Y., P. H. Nguyen, et al. (2005). "Energy landscape of a small peptide revealed by dihedral angle principal component analysis." Proteins: Structure, Function, and Bioinformatics **58**: 45-52.

Pal, S. K., S. Bandyopadhyay, et al. (2006). "Evolutionary computation in bioinformatics: A review." IEEE Trans. Systems Man and Cybernetics C - Applications and Reviews **36**(5): 601-615.

Palazoglu, A., A. Gursoy, et al. (2004). "Folding dynamics of proteins from denatured to native state: principal component analysis." J. Comp. Biology **11**: 1149-1168.

Pan, P. W., R. J. Dickson, et al. (2005). "Functionally relevant protein motions: Extracting basin specific collective coordinates from molecular dynamics trajectories." J. Chem. Phys. **122**: 034904.

Pan, Y. P. and V. Daggett (2001). "Direct comparison of experimental and calculated folding free energies for hydrophobic deletion mutants of chymotrypsin inhibitor 2: Free energy perturbation calculations using transition and denatured states from molecular dynamics simulations of unfolding." Biochemistry **40**(9): 2723-2731.

Plaku, E., H. Stamati, et al. (2007). "Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction." Proteins: Structure, Function, and Bioinformatics **67**: 897-907.

Plaxco, K. W. and C. M. Dobson (1996). "Time-resolved biophysical methods in the study of protein folding." Current Opinion in Structural Biology **6**: 630-636.

Scheraga, H. A., M. Khalili, et al. (2007). "Protein-folding dynamics: Overview of molecular simulation techniques." Annual Review of Physical Chemistry **58**: 57-83.

Shao, J. Y., S. W. Tanner, et al. (2007). "Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms." Journal of Chemical Theory and Computation **3**(6): 2312-2334.

Shen, M. Y. and K. F. Freed (2002). "All-atom fast protein folding simulations: the villin headpiece." Proteins: Structure, Function, and Genetics **49**: 439-445.

Shinde, U. and M. Inouye (2000). "Intramolecular chaperones: polypeptide extensions that modulate protein folding." Cell and Developmental Biology **11**: 35-44.

Stein, S. A. M., A. E. Loccisano, et al. (2006). "Principal components analysis: A review of its application on molecular dynamics data." Ann. Rep. in Comp. Chemistry **2**: 233-261.

Van Aalten, D. M. F., B. L. De Groot, et al. (1997). "A comparison of techniques for calculating protein essential dynamics." J. Comp. Chemistry **18**(2): 169-181.