Analyzing Dynamical Simulations of Intrinsically Disordered Proteins Using Spectral Clustering

Joshua L. Phillips
School of Engineering
University of California
Merced, CA 95348
jphillips7@ucmerced.edu

Michael E. Colvin
School of Natural Sciences
University of California
Merced, CA 95348
mcolvin@ucmerced.edu

Edmond Y. Lau
Chemistry, Materials, Earth, and Life Science Directorate
Lawrence Livermore National Laboratory
Livermore, CA 94550
lau12@llnl.gov

Shawn Newsam
School of Engineering
University of California
Merced, CA 95343
snewsam@ucmerced.edu

Abstract

Continuing improvements in algorithms and computer speeds promise that an increasing number of biomolecular phenomena can be simulated by molecular dynamics to produce accurate "trajectories" of their molecular motions on the nanosecond to microsecond time scale. An important target for such simulations will be non-equilibrium biochemical processes, such as protein folding, but existing tools for analyzing molecular dynamics trajectories are not well suited to non-equilibrium processes and progress will require improvements in tools for classifying the range and types of dynamics exhibited by these systems. An extreme example of a non-equilibrium biochemical process is the function of "intrinsically disordered" proteins – proteins that function without ever folding into a unique structure. In this paper, we demonstrate the use of spectral clustering methods to analyze the data produced from simulations of several forms from one class of intrinsically disordered proteins, the phenylalanine-glycine nucleoporins (FG-Nups). We explain why such methods are well-suited for the data produced by our simulations and show that clustering methods provide a direct, quantitative measure of how effectively single simulations independently sample regions of structural phase space. Moreover, our clustering results show distinct dynamical behavior in different forms of the FG-Nups, which may provide insights into their biological function.

1 Introduction

Classical molecular dynamics simulations of biomolecular structures provide a wealth of information on the structure and behavior of biomolecules at the atomic level. In the past, most molecular dynamics (MD) simulations involved the study of biomolecules fluctuating around a reference state (eg. a protein's folded state or a canonical form of DNA); however, MD is increasingly applied to non-equilibrium processes, such as protein folding. The trajectories produced by these simulations produce ensembles of diverse structures, with no unique reference structure.

The application of machine learning and data mining techniques to MD trajectories can provide useful tools for analyzing MD trajectories, but the very high dimensionality of the space of molecular structures (up to three times the number of atoms) means that research is needed to determine the appropriate methods. Past research has focused on applying clustering methods to trajectories produced by simulations of various biomolecules, but, to our knowledge, none have focused on a particular class of proteins known as "unstructured" or "intrinsically disordered" proteins (IDPs). The unique physical properties of this class of proteins motivate a novel application of clustering methods to the study of biomolecular simulations. In particular, by applying spectral clustering techniques, we hope to elucidate certain properties of these proteins that are not accessible using standard low-dimensional metrics.

Our approach differs from standard applications of clustering to MD trajectories in that we are not trying to ascertain what conformations are similar within a single replicate

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

simulation of a particular IDP. Neither are we attempting to find the differences or similarities in structure between the different IDPs. Instead, we use clustering to understand how the diversity of structures within a replicate simulation compares to the diversity between replicates. Understanding this is important because it will allow us to estimate the amount of unique structural space being sampled in our simulations. This is related to the approach taken by Lyman and Zuckerman [8] to determine when a single MD simulation is at equilibrium, but differs in that we cluster data from multiple replicates in order to understand the trade-offs between running many, shorter simulations and running fewer, longer simulations. Specifically, we will use the clustering results to determine to what extent replicate MD simulations of a single IDP sample independent regions of structural phase space.

2 Background

One of the central tenets of molecular biology is the "protein structure-function" paradigm, which states that proteins adopt rigid 3-dimensional structures that are responsible for their function. This paradigm has been the basis for our understanding to date of protein function. There is now growing evidence that some proteins and protein domains exist as "unstructured" or "intrinsically disordered" forms [17]. Indeed, it has been estimated that up to 50% of eukaryotic proteins have at least one region (>50 residues) that is disordered [4]. It is clear that the operating principles will be fundamentally different for unstructured protein regions than for folded protein domains, and there is currently very little knowledge of the biophysics of such regions, with many fundamental questions unanswered. Computational simulations must play a central role in studying intrinsically disordered proteins because there is no experimental technique that can directly sample protein structure on the time scale relevant to conformational changes in such regions and therefore experiment provides only indirect information on the unstructured state [10]. Atomistic MD techniques are well developed for simulating the motions of proteins, but the analysis tools for such simulations usually assume that the protein is fluctuating around a well-defined folded structure. In contrast, MD simulations of IDPs will yield a large ensemble of diverse structures and therefore new analysis tools are required if molecular simulations are to achieve their full promise in elucidating the function of IDPs.

While data clustering has been used to perform a variety of analyses on MD simulations of proteins and other polypeptide structures, to the best of our knowledge it has not been used to analyze simulations of IDPs. Clustering has been applied to better understand conformational states and the transitions between these states in trajectories that are expected to stabilize, such as the folding process of structured proteins. Karpen et al. [6] use clustering to analyze simulations of a pentapeptide structure by enforcing a cutoff radius on cluster size in a metric space related to the backbone and side-chain dihedral angles. Best and Hege [1] analyze simulations of a small tri-ribonucleotide by recursively bipartitioning a similarity graph in which the similarity between any two conformations is based on the difference between their complete atom-atom distance matrices. Both of these works identify and enumerate the distinct conformational states of the trajectories.

Clustering has also been used to study the convergence properties of simulation trajectories. Lyman and Zuckerman [8] cluster simulations of met-enkephalin, a pentapeptide neurotransmitter, by enforcing a cutoff radius on cluster size in a space determined by the root mean-square distance (RMSD) between conformations. The resulting quantization of the conformational space is used to compute structural histograms. Analysis of convergence is performed by comparing the histograms corresponding to different temporal windows of the simulation. Finally, Shao et al. [15] perform an extensive comparison of different clustering techniques to MD simulations of various DNA systems. Eleven different clustering algorithms are considered all of which use RMSD to compute the similarity between conformations. Their objective is to better understand the different clustering algorithms rather than to gain insight into the simulations. They conclude that there is no one perfect "one size fits all" algorithm but that the results depend on the choice of atoms for the RMSD calculation and knowledge of the number of clusters, among other things.

3 Materials and Methods

3.1 Molecular Dynamics Simulations

We are evaluating trajectory clustering strategies using the phenylalanine-glycine nucleoporins (FG-Nups)–intrinsically disordered proteins that fill the core of the Nuclear Pore Complex (NPC). The NPC facilitates selective transport of 5-40 nanometer diameter molecular "cargo" between the cytoplasm and nucleus only if the cargo carries a specific transport signal [14]. The FG-Nups are believed to form an impermeable gel-like mesh that fills the NPC core and undergoes an as-yet-unknown change when it binds to a cargo possessing a transport signal. The FG-Nups are characterized by different 4-amino acid motifs that are repeated throughout the proteins. In this study, we evaluate our clustering techniques using wild type FG-Nups with GLFG and FxFG (x=variable amino acid) motifs as well as several mutants with altered motifs, for a total of five distinct FG-Nups as shown in table 1. For each FG-Nup we performed 40 independent replicates of 5ns classical MD at 300K using the AMBER software suite [2] and a Generalized Born/Surface

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

Area implicit solvent model, using standard protocols and parameter sets. In each MD simulation, structures were saved every 1 picosecond for the final 3ns of simulation, to yield 3000 structures from each of the 40 replicate simulations. We also extended 5 of these replicates from each FG-Nup for an additional 15ns to yield 18000 structures for each of the replicates to contrast the results of running many, shorter MD simulations with the results of running fewer, longer MD simulations.

In previous work, we analyzed the structures sampled from the MD simulations using standard metrics of protein size, shape and structural difference [7]. A standard metric of protein size is the radius of gyration (R_g), which can be calculated using the RMSD of each atom from the overall center of mass of the protein. It is common to calculate protein R_g using only the carbon atoms in the protein backbone, which is independent of the very rapid structural fluctuations due to amino acid side chain motions. To characterize the shape of the protein configurations we have used a parameter (S) that is calculated from the three moments of inertia and is a measure of the overall shape of the protein ($S < 0$ oblate; $S = 0$ spherical; $S > 0$ prolate) [3]. Our previous work has shown that such computed metrics of size and shape are validated by experimental measurements and can be used to distinguish some functional sub-classes of FG-nups.

A standard metric for measuring structural change when simulating folded proteins is the RMSD between the C_α backbones of two protein structures. Computing the RMSD between two protein structures involves an alignment of the two protein structures achieved by rotating and translating the two structures to achieve minimal RMSD. Alternate methods of comparing protein structures have been developed that are believed to better measure the fold similarity between two structures (e.g. MAMMOTH [13]) but these methods have not been calibrated for IDPs.

3.2 Spectral Clustering

While many different clustering methodologies exist, *spectral clustering* methods seem particularly well-suited to the classification of protein structures. These methods have emerged in several disciplines as a general means of clustering data based solely on the distances between data points in a high-dimensional space. By constructing the full affinity matrix for the points in this space, we effectively cast the clustering problem into a *spectral graph partitioning* problem, where the second generalized eigenvector of a graph's Laplacian can be used to make an approximately optimal cut through the graph [16]. More recently, this approach has been generalized to compute a k way partitioning for a graph using k way normalized cuts and applying standardized clustering methods to the transformed graph [11]. Therefore, by computing the k way partitioning of the affinity matrix, we can effectively cluster the data points into k clusters.

Consider X which consists of n feature vectors of length l, $X = (x_1, x_2, \ldots, x_n)$. We can construct an affinity matrix $A \in \Re^{n \times n}$ which is defined as $A_{ij} = exp(-d(x_i, x_j)^2/2\sigma^2)$ for $i \neq j$, $A_{ii} = 0$, σ is a scaling parameter, and $d(x_i, x_j)$ for our work is the RMSD distance between structures i and j. Defining D to be a diagonal matrix where $D_{ii} = \sum_j A_{ij}$, we can construct the normalized affinity matrix $L = D^{-1/2}AD^{-1/2}$. By taking the k largest eigenvectors of L, stacked in columns, and normalizing each row to unit length, we produce an $n \times k$ matrix Y. We can then view the rows in Y as points that can be clustered using any general clustering technique that tries to minimize the sum-squared deviation within clusters. In our case, we use k-means clustering with several random initial restarts. We say that the structure x_i is in cluster j if and only if row i of Y was assigned to cluster j. The best scaling parameter, σ, can be found by simply searching for the value which minimizes the sum of intra-cluster variances when k-means is applied [11].

K-means clustering is one of the simplest, most widely understood clustering algorithms [9]. Traditionally, the algorithm is provided with a set X of l-dimensional feature vectors, where each vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ corresponds to a point in $\Re^l$, and also an integer value, $k > 1$, corresponding to the desired number of clusters in which to partition X.

The algorithm then proceeds as follows: 1. Initially select k points in the space spanned by X to be the cluster centroids $(\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_k})$. 2. Assign each vector in X to the closest centroid $\mathbf{c}$. 3. Set each $\mathbf{c}$ to be the average of all vectors in X assigned to itself. 4. Repeat steps 2 and 3 until every feature vector in X is assigned to the same cluster centroid $\mathbf{c}$ as in the previous iteration (or very few assignments change.)

This algorithm does not prescribe a particular distance measure, so we can use any reasonable measure for the data. Also, the need to form an average or "canonical" centroid for each cluster is problematic for certain feature spaces (such as the feature space of 3D structures). This is not a problem for spectral approaches because these calculations are not performed on the original data set X, but on the dimensionally reduced set of points described by the matrix Y above. When the k-means algorithm is applied to Y, distance measurements and the calculation of centroids is done in the space spanned by the matrix Y, and not the original domain space spanned by X. Therefore, we can use the Euclidean distance metric and averaging can be used for computing cluster centroids. We only use RMSD to construct the affinity matrix A at the beginning of the clustering process.

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

3.3 Direct Application of K-means Clustering

Previous work has focused on applying clustering techniques directly to MD trajectories without first applying spectral decomposition [15]. K-means is one of the algorithms studied by Shao et al. so in addition to the spectral clustering approach outlined above, we applied k-means directly to our trajectories for comparison. While k-means clustering often performs well in practice, there are certain details that must be carefully considered. First, the algorithm is known to be quite sensitive to the initial placement of the k centroids. Shao et al. utilize a deterministic heuristic for initialization of the algorithm; however, to have confidence in the results, we would need to run the algorithm multiple times from different random initial conditions and use the solution with the minimum sum of intra-cluster variances.

Even more problematic is the need to average the feature vectors within a cluster to obtain each cluster centroid for the next iteration of the algorithm. Shao et al. carefully investigate the tradeoffs in methods for calculating an "average" structure but there is no method that can ensure the result will be a physically reasonable protein structure. This is due to the fact that the constraints on bond lengths, atom sizes, torsional angles, etc. in MD trajectories constrain X such that $X \subset \Re^l$. In fact, any clustering approach that relies on an average or canonical structure of a cluster, without constraining such a structure to be a possible conformation of the system, could suffer from severe limitations when dealing with any domain where there are constraints on the values of X.

Spectral clustering effectively overcomes the limitations of the simple k-means approach discussed above. The generation of average or canonical structures for a cluster is avoided because there is no need to calculate a canonical structure for each centroid in the original space of protein conformations. This method is also faster in practice since efficient algorithms exist for computing the first few eigenvectors of a symmetric matrix (often the affinity matrix is sparse as well). Also, the matrix Y is typically much smaller than the original set X, and the points in Y are no longer in the original 3D structural space. So, while RMSD is used to compute the affinity matrix A, it is not used to cluster the points in Y. Instead, we can simply use Euclidean distance.

Hence, while calculating physically meaningful canonical structures seemingly limits the use of k-means for clustering MD trajectories, spectral approaches avoid explicitly calculating average structures and therefore are particularly well-suited for analyzing MD trajectories.

4 Results

4.1 Spectral Clustering

In order to assess the diversity of structures explored by each FG-Nup, we clustered each of the five FG-Nups separately. To make our trajectory data tractable for clustering, we sampled every tenth frame from each replicate and concatenated the forty trajectories into one set of structures. Therefore, the first 300 structures were all from replicate one, the next 300 structures were from replicate two, etc. for a total of 12000 structures. We then applied spectral clustering to this set of structures where each structure corresponds to a single point in X. We specified the number of clusters $k = 40$ (one cluster per replicate) for each protein with the hope of understanding how much the replicates overlap in phase-space, or if they are disjoint. Likewise, we sampled every tenth frame from each of the 18ns replicates and concatenated these five trajectories into one set of 9000 structures for each FG-Nup. We then clustered these trajectories with $k = 5$. The scaling parameter σ was set to 5 in all cases as searching for different values for each FG-Nup did not yield better clustering.

Graphs showing cluster membership for the five proteins are shown in figures 1 and 3.[1] Separate plots are shown for each distinct FG-Nup, and each plot shows the results of clustering all 40 independent replicates. These plots show the joint distribution of replicate to cluster assignment where each element in the image at location (i, j) corresponds to the fraction of the structures in the concatenated trajectory that were both sampled from replicate i and assigned to cluster j. The interesting thing to note here is that the GLFG motif is clustered in such a way that almost every replicate is contained within its own cluster. This shows that the structural diversity within each GLFG replicate is small compared to the diversity between clusters. In contrast, for FxFG and its mutants, the replicates do not cluster into unique clusters. Therefore, it seems likely that the structural space explored by each replicate for FxFG is very diverse compared to the inter-replicate diversity.

While it is not a surprise for the trajectories for each replicate to be very different, as in the GLFG cases, it is surprising that the seemingly more extended and rapidly changing FG-Nups such as FxFG and AxAG do not display this behavior. However, if the simulations were continued for a much more extended period of time, we might expect the GLFG trajectories to begin to overlap in structural similarity as well. We tested this hypothesis by apply-

[1] The algorithms do not naturally order the clusters so cleanly as is displayed in these graphs. Rather, we relabeled each cluster so as to align it to the replicate most commonly associated with itself. This is done by simply relabeling each cluster based on the replicate from which the median structure was taken.

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

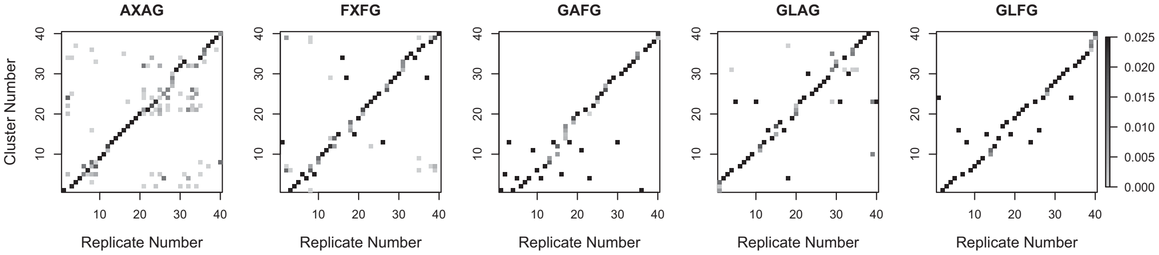

Figure 1. Results from applying *spectral clustering* to the *3ns* trajectory data.

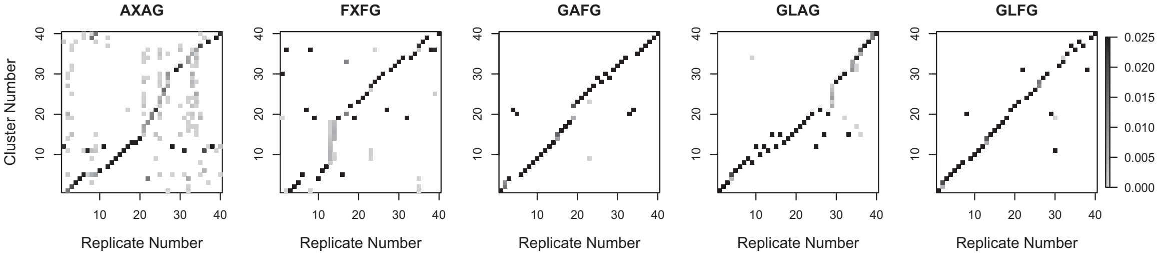

Figure 2. Results from applying *k-means clustering* to the *3ns* trajectory data.

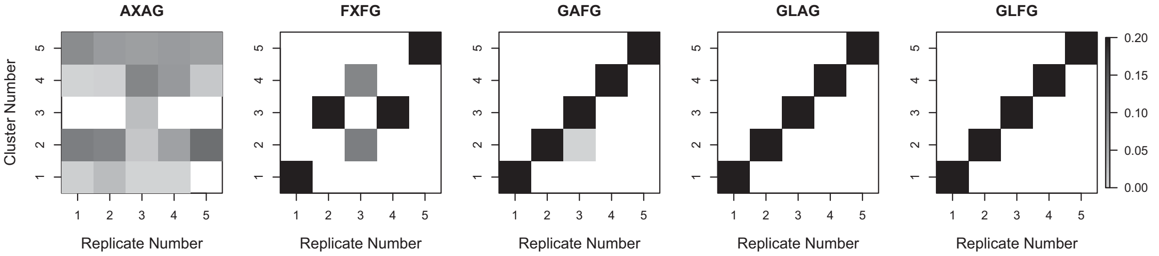

Figure 3. Results from applying *spectral clustering* to the *18ns* trajectory data.

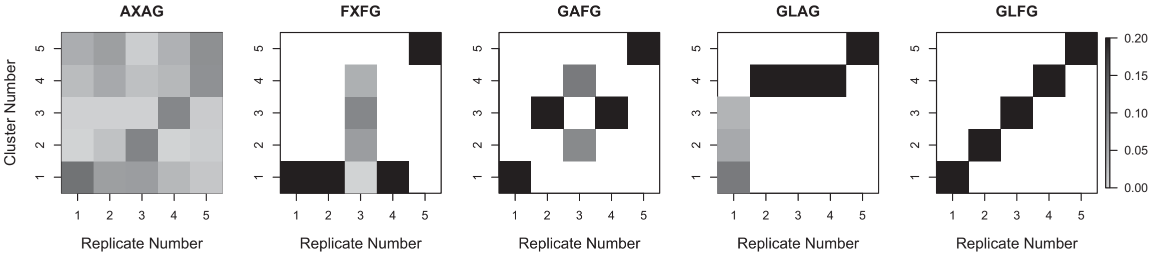

Figure 4. Results from applying *k-means clustering* to the *18ns* trajectory data.

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

ing the same clustering approach to the five 18ns simulations for each FG-Nup (five replicates of each protein.) It is clear from figure 3 that simply extending the length of the simulations does not result in structural overlap. This reinforces our previous results. The 18ns GLFG replicates remain clustered in separate clusters, but the 18ns FxFG and AxAG replicates tend to overlap, even more so than in the 3ns simulations (see mutual information results described below).

4.2 K-means Clustering

Using the same protocol described above, we applied k-means directly to the MD trajectories using the clustering software developed by Shao et al. [15]. Similar trends can be found in the clustering results obtained from this program as those found in our spectral clustering results. Figures 2 and 4 show the results of this analysis. While there are general similarities between the spectral and k-means results, there are some notable differences.

In order to quantitatively evaluate how much each clustering algorithm was separating replicates into disjoint clusters, we calculated the mutual information [5] between the replicates and clusters. Mutual information is a measure of independence of two random variables – the replicate and cluster labels in our case – and is computed as:

$$I(A;B) = \sum_{b \in B} \sum_{a \in A} p(a,b)\ \log_2 \left(\frac{p(a,b)}{p_1(a)p_2(b)} \right)$$

where $p(a,b)$ is the joint probability distribution of two discrete random variables A and B, $p_1(a)$ is the marginal probability distribution of A, and $p_2(b)$ is the marginal probability distribution of B. We then normalize this value by dividing by the maximum attainable mutual information ($\log_2(40)$ or $\log_2(5)$ for the 3ns and 18ns results, respectively) so that a value of one indicates perfect mutual information (each replicate placed in a cluster by itself) and zero indicates no mutual information (uniform assignment of replicate structures across all clusters). Taking the replicate-cluster assignment histograms in figures 1, 2, 3, and 4 to represent the joint probability distribution of replicate-cluster assignment, it is possible to compute mutual information from these data. The normalized mutual information for each cluster assignment is shown in table 1. Each FG-Nup examined is identified by a particular 4 amino acid (AA) motif that is repeated often along the protein sequence, and these are listed in column one. The second column describes the length of each fragment (in amino acids) as well as the name of the full-length yeast FG-Nup from which this fragment was taken. Mutants are described in terms of a specific amino acid substitution (eg. phenylalanine to alanine: F⇒A). The remaining columns show the normalized mutual information computed from each of the four clustering experiments in the study.

Comparing the mutual information values in table 1 gives insight into the independence of the replicates and the efficiency with which they are sampling the structural phase space of the FG-Nups. The 3ns replicates show relatively little loss of mutual information indicating that each replicate is providing new sampling of structural space. The 3ns AxAG simulations show the most decline in mutual information, consistent with the highest level of overlap in the replicate clusters. The 18ns simulations show a wider range of diversity in mutual information. In particular, the 18ns spectral clustering results from the AxAG and FxFG simulations show a general loss of mutual information compared to GLFG, GLAG and GAFG. This suggests that, for the former FG-Nups, the longer 18ns replicates are less efficient at sampling structural phase space than the 3ns replicates. However, the k-means results do not show a consistent loss of mutual information between these two groups. Instead, only GLFG shows an increase in mutual information. This discrepency could be arising from the structure averaging process and/or the initialization method used by the k-means clustering software. Regardless, it is clear that spectral methods can more precisely quantify the structural diversity of these proteins than standard k-means.

4.3 Comparison of Clustering and Standard Metrics

Our previous work involving standard metrics of protein size and shape indicated that the mutant varieties seem to express an even broader range of structures than the wild-type, which is consistent with previous hypotheses on the role that the various FG motifs play in structural arrangement [7]. However, one could incorrectly conclude from these data that the structural diversity of FxFG across replicates is much greater than the structural diversity of GLFG across replicates. However, *our clustering approaches yield unique insights into the accessibility of structural regions explored by IDPs that are not readily apparent using standard metrics.*

For example, among the FG-Nups analyzed in our previous work, GLFG was the least structurally diverse and most rigid. These results might be evidence for a lack of diversity in the structural space sampled, but clustering results for GLFG reveal a different picture. Since nearly all of the 40 GLFG replicate trajectories are clustered into separate clusters, most of the GLFG replicates are sampling a distinct and non-overlapping portion of structural space. Therefore, the high dimensional clustering analysis shows that the structural diversity within each GLFG replicate is small compared to the structural diversity between replicates.

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

Table 1. FG-Nup fragment motifs, lengths in amino acids (AA), and normalized mutual information in replicate-cluster assignment. A value of one indicates that each replicate was placed in a cluster by itself and zero indicates a uniform assignment of replicate structures across all clusters.

FG-Nup Motif	Fragment Details	Spectral 3ns	K-means 3ns	Spectral 18ns	K-means 18ns
AxAG	105AA Nsp1p mutant (F⇒A)	0.854	0.780	0.152	0.204
FxFG	105AA Nsp1p	0.901	0.872	0.828	0.590
GAFG	120AA Nup116p mutant (L⇒A)	0.884	0.971	0.999	0.828
GLAG	120AA Nup116p mutant (F⇒A)	0.890	0.882	1.000	0.590
GLFG	120AA Nup116p	0.949	0.962	1.000	1.000

In contrast, the FxFG structures that appear to be the most structurally diverse FG-Nups based on our previous work, do not homogeneously cluster into different replicates. Instead the FxFG clusters show a high degree of structural overlap between the different replicates, which points out a limitation on using low-dimensional aggregate measures of size and shape to categorize protein structure. The replicate simulations of FG-Nups like FxFG which have the most diversity in shape and size (R_g and S), actually sample fewer distinct regions of structural space than the GLFG-like FG-Nups.

5 Conclusions

While standard metrics of protein size and structure yield some information about the structural variation among the FG-nups that we have simulated, the application of clustering to our trajectories provides additional insights into their structural properties. Standard metrics lead us to infer that the FG-Nups characterized by the GLFG motif and its mutants adopt more compact configurations than those containing the FxFG motif and its mutants. However, this tells us little about the dynamic behavior of these FG-Nups. From our clustering results it is clear that GLFG and FxFG sample the simulation phase-space in very different ways. FxFG and its mutants all take on more extended configurations that are highly dynamic and readily cross into and out of structural configurations sampled by other replicates, broadly sampling the space of possible conformations. However, GLFG and its mutants tend to be less dynamic, sinking into local energy minima that are fairly distinct from one replicate to another.

Our results indicate that FG-Nups that are more extended, such as FxFG, tend to broadly sample the space of possible conformations and for these FG-Nups, it doesn't matter whether one runs many, shorter simulations of extended FG-Nups or fewer, longer simulations. In either case, the proteins should quickly sample the conformation space. However, FG-Nups that are more compact, such as GLFG, persist in structural arrangement over an extended period of time. Thus, running fewer, longer simulations will result in sampling only a few small regions of the conformation space. When many more replicates are run, the conformation space of several of the trajectories begins to overlap. Of course, even if we begin to see conformational overlap across replicates, this does not guarantee that we have sampled the space effectively. Yet, a lack of overlap necessarily means that we are in danger of undersampling.

These results provide information on the type and extent of MD simulations required to optimize the sampling of conformational space. Recall that the aim of replicate simulations is to independently sample portions of structural phase space to allow meaningful statistical descriptors of protein properties. The clustering analysis in this study shows that the optimal MD simulation protocol depends on the properties of the IDP being simulated. At one extreme, for GLFG the forty 3ns replicates as well as the five 18ns replicates are mostly clustered separately, indicating that each replicate is sampling a new and independent region of structural phase space. At the other extreme, for AxAG there is some overlap in the clustering of the replicates (and concomitant loss of mutual information) for both 3ns and 18ns, but the loss of mutual information due to this overlap is much more dramatic for the 18ns AxAG replicates, indicating that these longer simulations are not efficiently sampling structural phase space and that more, shorter replicates would be more efficient. Similarly, the 18ns FxFG and GLAG proteins show a large loss of mutual information. These conclusions could not be made without the clustering tools described in this paper.

6 Future Work

As with any physical system, raising the temperature of the FG-Nups increases the number of conformational states the protein has access to and increases the rate of interconversion between these states. As noted in the previous section, the AxAG, FxFG and GLAG proteins behave as if they are at an effectively higher temperature than GLFG, as measured by the greater structural diversity in the former and

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

in the lack of overlap observed in the clustered replicates of GLFG, indicating a slower rate of structural interconversion. To evaluate this relationship between the "effective temperature" and the homogeneity of the replicate clustering, we plan to extend our analysis to previously completed high temperature (350K) MD simulations of the same set of FG-Nups. Standard measures indicate much more dynamical similarity between the different FG-Nups at 350K than at 300K, with the high temperature simulations of GLFG yielding R_g–S distributions similar to FxFG at 300K. Clustering analysis will validate whether this similarity in the R_g–S distributions reflects deeper similarities in the structural dynamics of the different FG-Nups.

A limitation of all-atom MD simulations like those described here, are the relatively short timescales (nanoseconds) compared to protein folding (microseconds-milliseconds). To overcome this limitation we are simulating more simplified "course-grained" models of these proteins that will allow 100-1000x longer simulations. Clustering analysis of such long timescale simulations should allow us to estimate the fraction of available structural space that is being sampled by different FG-Nups.

Ultimately, the development of robust clustering methods for analyzing non-equilibrium MD simulations will open the door to a new language for describing and understanding the increasingly vast amount of MD simulation results made possible by continuing improvements in computer speeds and simulation algorithms. For example, a key long-term goal of our work is a clustering-based metric of "degree-of-unstructuredness" that could be used to categorize IDPs in the same way that fold types are currently used to categorize folded proteins (e.g. [12]).

7 Acknowledgments

The authors acknowledge helpful discussions with Prof. Miguel Carreira-Perpiñán (UC Merced), Prof. Ajay Gopinathan (UC Merced), Prof. Michael Rexach (UC Santa Cruz), and Prof. V.V. Krishnan (CSU Fresno & UC Davis). This work was supported in part by NIH Grants GM077520 and in part by the U.S. Department of Energy, Office of Science, Offices of Advanced Scientific Computing Research, and Biological & Environmental Research through the U.C. Merced Center for Computational Biology. This work was also performed in part under the auspices of the United States Department of Energy through the University of California Lawrence Livermore National Laboratory under contract number DE-AC52-07NA27344.

References

[1] C. Best and H.-C. Hege. Visualizing and identifying conformational ensembles in molecular dynamics trajectories. *Computing in Science and Engineering*, 4(3):68–75, 2002.

[2] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005.

[3] R. I. Dima and D. Thirumalai. Asymmetry in the shapes of folded and denatured states of proteins. *Journal of Physical Chemistry B*, 108:6564–6570, 2004.

[4] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6572–6582, 2002.

[5] S. Haykin. *Communication Systems*. John Wiley and Sons, 1994.

[6] M. E. Karpen, D. J. Tobias, and C. L. Brooks, III. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: Analysis of 2.2-ns trajectories of YPGDV. *Biochemistry*, 32(2):412–420, 1993.

[7] V. Krishnan, E. Y. Lau, J. Yamada, D. P. Denning, S. S. Patel, M. E. Colvin, and M. F. Rexach. Intra-molecular cohesion of coils mediated by phenylalanine-glycine motifs in the natively unfolded domain of a nucleoporin. *PLOS Computational Biology*, 2008. (in press).

[8] E. Lyman and D. M. Zuckerman. Ensemble-based convergence analysis of biomolecular trajectories. *Biophysical Journal*, 91:164–172, 2006.

[9] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[10] T. Mittag and J. D. Forman-Kay. Atomic-level characterization of disordered protein ensembles. *Current Opinion in Structural Biology*, 17(1):3–14, 2007.

[11] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856, Cambridge, MA, 2002. MIT Press.

[12] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH – a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[13] A. R. Ortiz, C. E. M. Strauss, and O. Olmea. MAMMOTH (matching molecular models obtained from theory): An automated method for model comparison. *Protein Science*, (11):739–756, 2002.

[14] M. P. Rout and J. D. Aitchison. The nuclear pore complex as a transport machine. *Journal of Biological Chemistry*, (276):16593–16595, 2001.

[15] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham III. Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J. Chem. Theory Comput.*, 3(6):2312–2334, 2007.

[16] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[17] V. Uversky. Natively unfolded proteins: A point where biology waits for physics. *Protein Science*, 11:739–756, 2002.

978-1-4244-2890-8/08/$25.00 ©2008 IEEE