UNIVERSITY OF CALIFORNIA, MERCED

Validation of Computational Approaches for Studying Disordered and Unfolded Protein Dynamics Using Polymer Models

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Electrical Engineering and Computer Science

by

Joshua Lee Phillips

Committee in charge:

Professor Shawn Newsam, Chair
Professor Miguel Á. Carreira-Perpiñán
Professor Michael E. Colvin
Professor Ajay Gopinathan

2012

The dissertation of Joshua Lee Phillips is approved:

_____

_____

_____

_____
Chair

University of California, Merced

2012

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my co-advisors Prof. Shawn Newsam and Prof. Michael Colvin for their constant support and advocacy on my behalf and I would like to 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). I also would like to thank Edmond Y. Lau (Lawrence Livermore National Laboratory) for his assistance with molecular dynamics simulations of the FG-Nups.

This work was supported in part by NSF Grant 0960480, NIH Grant GM077520, the U.S. Dept. 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 U. S. Dept. of Energy through the University of California Lawrence Livermore National Laboratory under contract number DE-AC52-07NA27344.

Chapter 3, in part, is a reprint of the of the material as it appears in Analyzing Dynamical Simulations of Intrinsically Disordered Proteins Using Spectral Clustering in the Proceedings of the 2008 IEEE Conference on Bioinformatics and Biomedicine Workshops. Phillips, J. L.; Colvin, M. E,; Lau, E. Y.; Newsam, S., IEEE Computer Society, 2008. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in part, is a reprint of the material as it appears in Validating Clustering of Molecular Dynamics Simulations Using Polymer Models in BMC Bioinformatics 2011. Phillips, J. L.; Colvin, M. E; Newsam, S., BioMed Central, Springer Science and Business Media, 2011. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in part, is currently being prepared for submission for publication of the material. Phillips; J. L.; Colvin M. E.; Newsam, S. The dissertation author was the primary investigator and author of this material.

VITA

2002	Bachelor of Science in Computer Science, Middle Tennessee State University
2004	Master of Science in Computer Science, Vanderbilt University
2012	Doctor of Philosophy in Electrical Engineering and Computer Science, University of California, Merced

PUBLICATIONS

J. L. Phillips, S. Kogekar, and J. A. Adams, "Emergency automated response system (EARS)," in Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society, 2004.

J. L. Phillips and D. C. Noelle, "Reinforcement learning of dimensional attention for categorization," in Proceedings of the 26th Annual Meeting of the Cognitive Science Society, 2004.

J. L. Phillips, "Reinforcement learning of dimensional attention for categorization." M.S. Thesis, Department of Computer Science, Vanderbilt University. Nashville, TN, 2004.

J. L. Phillips and D. C. Noelle, "A biologically inspired working memory framework for robots," in Proceedings of the 27th Annual Meeting of the Cognitive Science Society, 2005.

J. L. Phillips and D. C. Noelle, "Working memory for robots: inspirations from computational neuroscience," in Proceedings of the 5th International Conference on Development and Learning, 2006.

M. Tugcu, X. Wang, J. E. Hunter, J. L. Phillips, D. C. Noelle, and D. M. Wilkes, "A computational neuroscience model of working memory with application to robot perceptual learning," in Proceedings of the 3rd International Conference on Computational Intelligence, 2007.

J. L. Phillips, M. E. Colvin, E. Y. Lau, and S. Newsam, "Analyzing dynamical simulations of intrinsically disordered proteins using spectral clustering," in Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine Workshops, pp. 17-24, 2008.

E. Y. Lau, J. L. Phillips, and M. E. Colvin, "Molecular dynamics simulations of highly charged green fluorescent proteins," Molecular Physics, vol. 107, no. 8, pp. 1233-1241, Jan. 2009.

J. Yamada, J. L. Phillips, S. Patel, G. Goldfien, A. Calestagne-Morelli, H. Huang, R. Reza, J. Acheson, V. V. Krishnan, S. Newsam, A. Gopinathan, E. Y. Lau, M. E. Colvin, V. N. Uversky, and M. F. Rexach, "A bimodal distribution of two distinct categories of intrinsically-disordered structures with separate functions in FG nucleoporins.," Molecular & Cellular Proteomics, Apr. 2010.

J. L. Phillips, M. E. Colvin, and S. Newsam, "Validating clustering of molecular dynamics simulations using polymer models," BMC Bioinformatics, vol. 12, no. 1, p. 445, Jan. 2011.

PUBLISHED ABSTRACTS

J. L. Phillips, E. Y. Lau, V.V. Krishnan, M. Rexach, S. Newsam, and M. E. Colvin, "Characterizing intrinsically disordered FG-nucleoporins using molecular dynamics," in Proceedings of the 22nd Annual Symposium of the Protein Society, 2008. (Poster)

J. L. Phillips, E. Y. Lau, V.V. Krishnan, M. Rexach, S. Newsam, and M. E. Colvin, "Dynamics analysis of unstructured FG-nucleoporins," in Proceedings of the 23rd Annual Symposium of the Protein Society, 2009. (Poster - Winner of the 2009 Best Student Poster Award)

J. L. Phillips, E. Y. Lau, V.V. Krishnan, M. Rexach, S. Newsam, and M. E. Colvin, "Metric scaling for dimensionality reduction of disordered protein dynamics," in Proceedings of the 54th Annual Meeting of the Biophysical Society, 2010. (Poster - Winner of the 2010 Student Research Achievement Award)

J. L. Phillips, E. Y. Lau, M. Rexach, S. Newsam, and M. E. Colvin, "Dimensionality reduction reveals differences between disordered protein dynamics and early-stage protein folding dynamics," in Proceedings of the 24th Annual Symposium of the Protein Society, 2010. (Poster)

J. L. Phillips, E. Y. Lau, M. Rexach, S. Newsam, and M. E. Colvin, "Probing the conformation landscape of the unfolded state: do disordered and unfolded dynamics differ?" in Proceedings of the 55th Annual Meeting of the Biophysical Society, 2011. (Platform Talk)

J. L. Phillips, A. Gopinathan, S. Newsam, and M. E. Colvin, "Dimensionality estimation of disordered protein dynamics," in Proceedings of the 56th Annual Meeting of the Biophysical Society, 2012. (Poster)

FIELDS OF STUDY

Major Field: Electrical Engineering and Computer Science

Studies in Machine Learning
Professors David C. Noelle and Shawn Newsam

Studies in Computational Biology
Professor Michael E. Colvin

Studies in Cognitive Neuroscience
Professor David C. Noelle

ABSTRACT

Validation of Computational Approaches for Studying Disordered and Unfolded Protein Dynamics Using Polymer Models

by

Joshua Lee Phillips

Doctor of Philosophy in Electrical Engineering and Computer Science

University of California, Merced, 2012

Professor Shawn Newsam, Chair

The "protein structure-function" paradigm, which states that proteins adopt nearly rigid 3-dimensional structures that are responsible for their function, is one of the central tenets of molecular biology, yet some proteins and protein domains exist as intrinsically disordered forms. In this dissertation, new approaches to define a metric for the dynamics of disordered proteins are developed which are also readily applicable to the study of non-equilibrium globular protein dynamics. First, standard metrics for comparing protein dynamics are applied to molecular dynamics (MD) simulations of a class of entirely disordered proteins (outside of a small anchoring domain) involved in nucleocytoplasmic transport, the FG-nucleoporins (FG-Nups). After this, clustering and dimensionality reduction techniques are utilized to reveal previously unknown characteristics regarding the convergence properties of disordered protein simulations. Next, the novel application of polymer models is used to assess the efficacy of clustering and dimensionality estimation algorithms applied to MD trajectories. Finally, the results are

used to analyze the differences between FG-Nup dynamics and the dynamics of two fast-folding globular proteins, GB1 and Trp-cage. The results indicate that polymer models are an effective tool for validating computational techniques for studying protein simulations, and that the various proteins can be classified by differences in their underlying dynamics.

Chapter 1

Introduction

Continuing improvements in algorithms and computer speeds promise that an increasing number of biomolecular phenomena can be simulated by molecular dynamics (MD) 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. Progress will therefore 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 (IDPs) – proteins that function without ever folding into a unique structure. There is now growing evidence that some proteins and protein domains exist as "unstructured" or "intrinsically disordered" forms [1]. Indeed, it has been estimated that up to 50% of eukaryotic proteins have at least one region (>50 residues) that is disordered [2] for at least short periods of time. 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.

Traditionally, the structure and function of IDPs is often described in contrast to

"natively folded" proteins (NFPs) which adopt rigid 3-dimensional structures that are responsible for their function. Instead, it is believed that IDPs are best characterized by their dynamics. 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 [3].

Molecular dynamics simulation is a powerful technique for sampling the conformation space of proteins and other biomolecules. All-atom models provide a wealth of structural information at a level of physical detail that is inaccessible to many experimental techniques and can be used to make theoretical predictions for future experimental validation. MD simulation is particularly well-suited for studying the local minima in the free energy landscape (metastable states) and the transitions between these minima (transition states) which characterize how biomolecules perform their requisite functions. These and other dynamical properties can in principle be obtained from the conformational ensembles from MD simulation trajectories; however, calculating them has proven to be a challenge in practice. Nevertheless, there is good reason to believe that we are at the threshold of being able to perform predicatively accurate MD simulations on systems of biologically meaningful sizes (millions of atoms) and timescales (milliseconds-seconds) and therefore new analysis tools are needed to characterize the dynamical properties of biomolecules from MD simulations. *The main focus of this work is to demonstrate the use of machine learning, polymer-based models, and other computational techniques to analyze the data produced from simulations of several forms IDPs and unfolded NFPs.* These methods provide direct, quantitative measures of the dynamics of these proteins extracted from molecular dynamics trajectories and allow IDPs and NFPs to be classified and compared based on their "degree of disorder".

The application of machine learning and data mining techniques to MD trajectories has provided useful tools for studying biomolecular processes, 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. In the chapters that follow, several techniques which possess amenable properties for studying MD trajectories are explored and evaluated. In addition, simple polymer models with well-defined statistical properties will be used to both develop and evaluate the approaches taken. The fusion of these computational methods with polymer theory models provides unique, synergistic insights into both the dynamics of the proteins studied and the validity of the computational methods themselves.

Chapter 2

Quantifying Structural Change in Molecular Dynamics Simulations of Intrinsically Disordered Proteins

In this chapter, several techniques for exploring protein flexibility and structural heterogeneity are described. Many of these approaches are adapted from standard metrics for studying NFPs. Some of these approaches are novel, and others are found in the relevant literature on IDPs and denatured NFPs. The methods can be loosely grouped into two categories, static methods and dynamic methods, which will be described in more detail later in this chapter. These methods are applied to MD simulations of a class of entirely disordered proteins (outside of a small anchoring domain) involved in nucleocytoplasmic transport, the FG-nucleoporins (FG-Nups) *in order to illustrate their relative strengths and limitations for studying non-equilibrium processes.*

2.1 Background

Proteins consist of a set of monomeric units called amino acids (AAs) or residues, that are connected together by amide bonds in order to form a polymer chain. There are

roughly 20 different amino acids, each with unique chemical properties, that can be used to create proteins (the exact number differs slightly across various living organisms). The exact combination and ordering of amino acids that make up a protein is denoted as a protein's primary structure. Since many amino acids have a strong propensity for interacting with other amino acids, the primary structure of a protein might encourage the protein chain to form additional structural motifs beyond the simple polymer structure dictated by the amide bonds.

For NFPs, certain local structural motifs are quite common, such as the α-helix or β-sheet, and are referred to as forms of secondary structure. These two structural motifs are exemplified in Figure 2.1. Additionally, entire secondary structure elements can interact in specific ways due to their constituent amino acids such that they form tertiary structure, or even interact with other proteins to form quaternary structure. The process of forming such structures is called *folding*, and although some proteins require assistance from other proteins to fold, many proteins are known to be able to fold independently. Therefore the information about the folding pathway(s) and final structure is encoded in a proteins's primary structure. The final folded state is of great biological significance because it directly dictates or modulates how the protein interacts with other biomolecules in order to achieve some biologically significant effect. The folded structure is called the *native* state of the protein. It can often be determined experimentally, and serves as the "reference state" for many theoretical and computational methods for studying proteins.

IDPs function under a different structural model, without spontaneously adopting the structural motifs mentioned above, and often interact with other biomolecules in very different ways from their NFP counterparts. If any detailed structural or interactionary patterns among IDPs exist, they are often currently considered beyond the scope of modern experimental techniques. Nevertheless, some patterns are starting to emerge among various classes of IDPs. For example, certain IDPs are known to adopt secondary structure upon binding with partner biomolecules [5]. The prevailing flexibility afforded by intrinsic disorder is thought to allow IDPs to easily search for binding

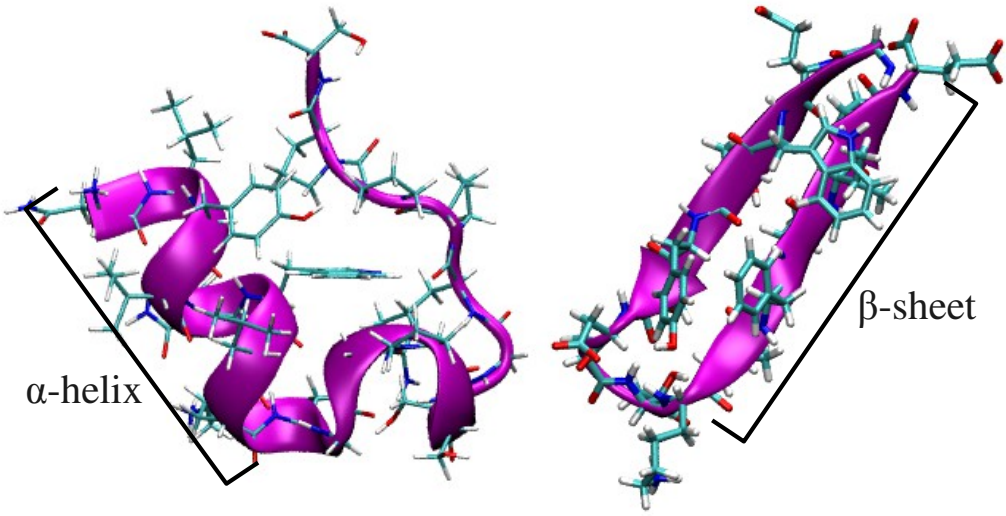

Figure 2.1: Examples of Common Protein Secondary Structures. Left: Folded structure of the Trp-cage mini-protein (RCSB Protein Data Bank ID: 1L2Y) with labeled α-helix secondary structure motif. Right: Folded structure of the GB1 beta-hairpin (RCSB Protein Data Bank ID: 1GB1) with labeled β-sheet secondary structure motif. Images generated using Visual Molecular Dynamics software version 1.9 [4].

partners through a process known as "fly-casting" [6, 7]. More towards the same extreme, some IDPs show gross structural arrangement based on their sequence properties with only short, targeted sequence elements playing the key role in the interaction [8]. Additionally, some research has indicated that some IDPs are characterized by transient secondary structure, which is thought to play the primary role in binding interactions [9]. Across all of these studies, simulation has played a critical role in elucidating these properties of IDPs.

Classical molecular dynamics simulations of biomolecular structures provide a wealth of information on the structure and behavior of biomolecules at the atomic level. The overall approach of molecular dynamics simulation is characterized by the use of classical mechanical laws of motion to model the physical motions of the biomolecules of interest. Complete models of proteins and other biomolecules have been created, and are available in modern simulation software packages which numerically integrate the equations of motions associated with these systems. While the simulation pack-

ages have been heavily optimized and designed to run on distributed parallel processing environments, the spatial/temporal scale of modern simulation is often still small/short compared to the spatial/temporal range needed to capture many interesting biomolecular phenomena. Also, the data produced by simulations that reach useful spatial/time-scales is large and cumbersome to analyze. Therefore, well-validated and efficient methods for analyzing the results of these simulations are of critical importance.

2.2 Methods

2.2.1 Molecular Dynamics Simulations

The model protein domains for this work were 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 [10]. 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.

Classical molecular dynamics simulations were performed on six different protein sequences. Two of these sequences were derived from the 823 amino acid-long yeast wildtype FG-nucleoporin NSP1 (NCBI Assession Number: NP_012494.1, Gene ID: 6322420). This protein is rich in amino acid repeats of the form "FxFG" (F=phenylalanine, G=glycine, x=variable amino acid), and, in general, contains a high number of positively charged K residues (K=lysine). A 105 amino acid-long subsequence of the full-length NSP1 protein (AAs 375-479), referred to as FxFG was used, as well as a mutant obtained from changing all of the F residues in the sequence into A (A=alanine) residues, referred to as AxAG, and a mutant obtained from chaining all of the F residues in the sequence into S (S=serine) resides, referred to as SxSG. Phenylalanine is highly hy-

drophobic and avoids interaction with the surrounding solvent when possible, favoring interactions with other hydrophobic groups in the protein. In FxFG, they are thought to play a key role in driving this protein to be somewhat compact in spite of the relatively large number of positively charged amino acids in the rest of the sequence, which favor interactions with the surrounding solvent. Alanine, and to a greater extent serine, are less hydrophobic, making the mutants more likely to adopt extended conformations. The remaining three sequences were derived from the 1113 amino acid-long yeast wildtype FG-nucleoporin NUP116 (NCBI Assession Number: NP_013762.1, Gene ID: 6323691). This protein is rich in amino acid repeats of the form "GLFG", and, in general, contains a low number of charged amino acids. A 120 amino acid-long subsequence of the full-length NUP116 protein (AAs 346-457 plus a short tag), referred to as GLFG was used, as well as two mutants obtained from changing all of the F residues to A residues (GLAG) or all L (L=leucine) residues to A residues (GAFG). Because GLFG contains only a few charged residues, phenylalanine is thought to play a minor role in keeping these proteins compact, so GLAG should be marginally more extended than GLFG, while GAFG should remain largely unaffected since leucine and alanine have relatively similar chemical properties. The complete sequences are listed below (D=aspartate, E=glutamate, I=isoleucine, M=methionine, N=asparagine, P=proline, Q=glutamine, R=arginine, T=threonine, V=valine, W=tryptophan, and Y=tyrosine).

- FxFG

  SKPAFSFGAK PDENKASATS KPAFSFGAKP EEKKDDNSSK
  PAFSFGAKSN EDKQDGTAKP AFSFGAKPAE KNNNETSKPA
  FSFGAKSDEK KDGDASKPAF SFGAK

- AxAG

  SKPAASAGAK PDENKASATS KPAASAGAKP EEKKDDNSSK
  PAASAGAKSN EDKQDGTAKP AASAGAKPAE KNNNETSKPA
  ASAGAKSDEK KDGDASKPAA SAGAK

- SxSG

```
SKPASSSGAK PDENKASATS KPASSSGAKP EEKKDDNSSK
PASSSGAKSN EDKQDGTAKP ASSSGAKPAE KNNNETSKPA
SSSGAKSDEK KDGDASKPAS SSGAK
```

- GLFG

```
GSRRASVGSG ALFGAKPASG GLFGQSAGSK AFGMNTNPTG
TTGGLFGQTN QQQSGGGLFG QQQNSNAGGL FGQNNQSQNQ
SGLFGQQNSS NAFGQPQQQG GLFGSKPAGG LFGQQQGASY
```

- GLAG

```
GSRRASVGSG ALAGAKPASG GLAGQSAGSK AAGMNTNPTG
TTGGLAGQTN QQQSGGGLAG QQQNSNAGGL AGQNNQSQNQ
SGLAGQQNSS NAAGQPQQQG GLAGSKPAGG LAGQQQGASY
```

- GAFG

```
GSRRASVGSG AAFGAKPASG GAFGQSAGSK AFGMNTNPTG
TTGGAFGQTN QQQSGGGAFG QQQNSNAGGA FGQNNQSQNQ
SGAFGQQNSS NAFGQPQQQG GAFGSKPAGG AFGQQQGASY
```

These disordered proteins span a wide range of sizes as measured by experimental sieving column size-exclusion and solution NMR [8] and are predicted to cover three distinct classes of disordered proteins: GLFG is classified as a *collapsed coil*, FxFG is classified as an intermediate *relaxed coil*, and SxSG is classified as an *extended coil*. The balance between hydrophobic interactions (primarily from the F residues in the FG-Nups examined here) and overall percent of charged content of the proteins is hypothesized to be the driving force for collapsing/extending in these domains [1]. These

classifications also predict that the extended coils should exhibit less frustrated dynamics, with fewer, more shallow minima in the free-energy surface. Likewise, we predict that the collapsed coils should exhibit more frustrated dynamics, with many more, shallow minima.

A total of 40 independent replicate simulations of 5ns classical MD at 300K were performed for each protein using the AMBER software suite [11] and a Generalized Born/Surface Area implicit solvent model, using standard protocols and parameter sets. Fully-extended structures for the simulations were prepared using the AMBER program tleap, with ACE and NME caps on the C and N termini, and subsequently minimized using 10000 steps of steepest descent. Each simulation was then started from the minimized structures using a unique set of random initial velocities. 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. Also, 5 of these replicates from each FG-Nup were extended for an additional 15ns to yield 18000 structures for each of the replicates in order to contrast the results of running many, shorter MD simulations with the results of running fewer, longer MD simulations. An additional 40 independent 2ns simulations for each of the proteins at 350K were also performed using the same protocol to study the effects of high temperature on these IDPs.

2.2.2 Protein Structure Comparison

Computational techniques for studying the conformational heterogeneity of proteins rely heavily on structure comparison metrics. These metrics define the way in which two protein conformations are seen as being similar or dissimilar to one-another. Numerous quantitative techniques exist for computing such distances. While the first of the methods covered below is the most popular in use today, it is not necessarily the best choice in general, and development of better metrics for protein structure comparison remains an active field of research.

Root-Mean-Squared Distance

The most popular distance metric for protein comparisons is the root-mean-squared distance (RMSD). For two protein conformations, $\mathcal{P}$ and $\mathcal{Q}$, RMSD defines the distance between two structures to be the minimum of the root-mean-squared inter-atomic distances over all possible rotations and translations of R^{Q}:

$$\mathrm{RMSD}(\mathcal{P}, \mathcal{Q}) = \min_{\mathrm{R}^{\mathcal{Q}}} \left\{ \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{r}_i^{\mathcal{P}} - \mathbf{r}_i^{\mathcal{Q}}\|^2} \right\} \tag{2.1}$$

where N is the number of corresponding residues in structures $\mathcal{P}$ and $\mathcal{Q}$, R^{Q} is the $N \times 3$ matrix of alpha-carbon (C_α) atom positions for conformation $\mathcal{Q}$, and $\mathbf{r}_i^{\mathcal{P}}$ is the three-dimensional position vector of the ith alpha-carbon atom in conformation $\mathcal{P}$.

Each amino acid contains exactly one C_α atom, which is centrally located between the amide nitrogen (N) and carbonyl carbon (C) atoms. The amide nitrogen from one amino acid is bonded to the carbonyl carbon of another amino acid to form a protein chain, so all of these atoms together are referred to as the protein *backbone*. The set of atoms which give each amino acid its unique chemical properties, known as the *sidechain*, is bonded to the C_α atom as well, but this set differs across amino acids. Hence, the C_α atom is a reasonable atom to choose to represent the overall location of each amino acid. However, a mass-weighted version of RMSD may also be computed if additional atoms (like all backbone or sidechain atoms) are required for more precise calculations. Also, common procedures used to optimize the fit between the two conformations often minimize a different objective function than the one specified above. For example, in practice, the constraint to minimize across all translations is often relaxed by simply mass-centering the two-conformations, and solving for the optimal rotation around the origin, which also results in a symmetric, consistent solution for all pairs of conformations.

RMSD is very effective for discriminating between conformations which are fairly similar to one-another, but often performs somewhat unpredictably for fairly dis-

similar conformations. This problem is caused by the rather naive rigid rotation and translation requirements of the algorithm. For example, if an NFP is undergoing the process of folding, a conformation where half of the protein is completely folded while the other half is completely unfolded may appear just as similar to the final folded structure as a conformation which is tightly collapsed, but lacking relevant secondary structure. In other words, RMSD considers the position of each atom to be of equal importance even though it makes intuitive sense that atoms in the conformation which match closely with the folded structure should not be penalized by the arguably more arbitrary arrangement of the rest of atoms in the conformation. Another example where RMSD is somewhat ineffective considers a protein with two rigid, folded domains separated by a short, flexible linker. The two domains can move rather independently, so that two conformations could be considered more dissimilar than conformations where the domains are still the same relative positions, but one domain is halfway unfolded. Again, it makes intuitive sense that the independent motion of the two domains should not be penalized equally to motions where one of the domains has become unfolded.

Intramolecular Distance Deviation

Another popular metric for comparing protein structures is the intramolecular distance deviation (IDD). For two protein conformations, $\mathcal{P}$ and $\mathcal{Q}$, IDD is defined as:

$$\mathrm{IDD}(\mathcal{P},\mathcal{Q}) = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\|\mathbf{r}_i^{\mathcal{P}} - \mathbf{r}_j^{\mathcal{P}}\| - \|\mathbf{r}_i^{\mathcal{Q}} - \mathbf{r}_j^{\mathcal{Q}}\|\right)^2} \tag{2.2}$$

where N is the number corresponding residues in structures $\mathcal{P}$ and $\mathcal{Q}$, $\mathrm{R}^{\mathcal{Q}}$ is the $N \times 3$ matrix of C_α atom positions for conformation $\mathcal{Q}$, and $\mathbf{r}_i^{\mathcal{P}}$ is the three-dimensional position vector of the ith alpha-carbon atom in conformation $\mathcal{P}$.

IDD has two advantages over RMSD: no optimization procedure is required, and the calculation is rotationally and translationally invariant. Similar to RMSD, the IDD approach also ensures that the result is consistent and symmetric for any pair of

conformations. IDD still faces the same difficulties as RMSD when comparing highly dissimilar structures, and has one additional property that can be problematic in specific cases: it cannot discriminate between mirrored conformations. All amino acids except glycine can exist in one of two possible forms known as optical isomers. This is not commonly a problem when comparing protein structures because, in nature, organisms typically only utilize one of these isometric forms, making any proteins formed from the other isometric form biochemically incompatible. However, for more general polymer models, both forms are often equally likely, and IDD cannot discriminate between them.

Φ-Ψ Angles

While the three dimensional coordinates of protein atoms form an intuitive space for performing conformer comparisons, certain internal coordinates are also commonly employed. One common set of internal coordinates are the backbone Φ-Ψ angles. These angles are determined by computing the dihedral angles formed by the positions of four consecutive atoms along the protein backbone. Figure 2.2 shows a short segment of a protein backbone, with the three constituent dihedral angles (Φ,Ψ,Ω) labeled accordingly. The dihedral angle is obtained by calculating the dot product between the vector normal to the plane formed by the first three atoms and the vector normal to the plane formed by the last three atoms in this quartet. This is a practical and effective way to internally parametrize protein conformations because the bond lengths between pairs of atoms and the bond angles formed by any three consecutive atoms along the backbone are relatively fixed, fluctuating only by very small amounts. Thus, the three-dimensional coordinates of the protein backbone can be effectively reconstructed given just the set of dihedral angles along the backbone of a protein conformation. In addition to this, the peptide bond dihedral angle Ω, formed by the atoms C_α-C-N-C_α, is extremely rigid and fixed at 180 degrees for most residues. (Although it can also take on a value of 0 in rare cases, it is often not modeled in this manner in most molecular dynamics simulations.) Therefore, only the two remaining combinations of atoms may be used to form what are

called the Φ and Ψ angles: $\mathrm{C\text{-}N\text{-}C_\alpha\text{-}C}$ and $\mathrm{N_\alpha\text{-}C_\alpha\text{-}C\text{-}N}$, respectively. This means that a protein conformation can be effectively internally parametrized using the vector of Φ-Ψ angles of length $2 * N - 2$, where N is the number of residues in the protein.

There are two typical methods of employing the Φ-Ψ angles for computing dissimilarity metrics for conformations. The first is the Euclidean distance between the $\mathrm{sin\text{-}cos}$ transform of the Φ-Ψ angles [12]. Since the angles can range from $(-\pi, \pi]$, they must be projected onto the unit circle by taking the sin and cos of each angle. This results in a $4 * N - 4$ dimensional vector of transformed coordinates where simple Euclidean distance between two vectors can be applied to compute dissimilarity. The second metric is the dot product between the inverse Discrete Fourier Transform of the *complex-conjugate* forms of the Φ-Ψ angles. In this approach, a vector of complex values, $\mathbf{c}$ is constructed:

$$\begin{aligned} \mathbf{c}_1 &= 1 + 0i \\ \mathbf{c}_{x+1} &= e^{i\theta_x} \\ \mathbf{c}_{2M+1-x} &= e^{-i\theta_x}, \forall x = 1 \ldots M \end{aligned} \tag{2.3}$$

where $M = 2N - 2$, which is the total number of elements in the vector of all Φ-Ψ angles, and θ is the vector of all Φ-Ψ angles. By construction, taking the normalized inverse DFT of $\mathbf{c}$ results in a $4 * N - 3$ dimensional vector of *real* numbers that lies on the unit hypersphere. The dot product between the transformed vectors can then be used a as metric of dissimilarity between conformations [13].

Both of the above methods for calculating dissimilarity using Φ-Ψ angles have the advantage of using a rotationally and translationally invariant parametrization of conformational structure, just like IDD. However, unlike IDD, these methods are able to discriminate between isometric forms of the polymer chains. In addition to this advantage, the problems with RMSD and IDD mentioned above involving rigid domains are largely overcome. This advantage is limited somewhat because a chain where every other angle along the backbone matches well is the same as a chain where the first half

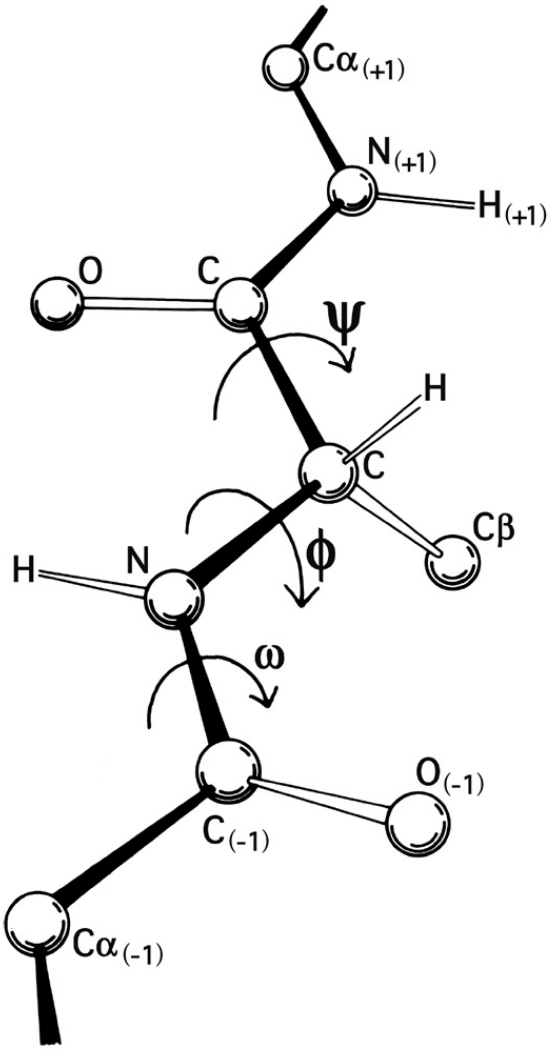

Figure 2.2: Pictorial description of the three rotational degrees of freedom captured by the dihedral angles (Φ,Ψ,Ω) along a protein backbone. Image courtesy of Wikimedia Commons user Dcrjsr (Jane S. Richardson, Professor, Duke University) on the web at http://upload.wikimedia.org/wikipedia/commons/c/c0/Protein_backbone_PhiPsiOmega_drawing.jpg according to the Creative Commons Attribution 3.0 Unported license (http://creativecommons.org/licenses/by/3.0/deed.en).

of the chain matches well, but the second half does not. So, some additional weighting would be needed to lower the dissimilarity for the intuitively preferred latter case. Because the Φ-Ψ angle space has advantages over the IDD approach and mitigates the same problems, IDD isn't utilized in any analyses presented here.

MAMMOTH

Another technique for conformation comparison developed recently by Ortiz et al. is MAMMOTH [14]. Generally speaking, this technique aims to fix the structural domain issue faced by RMSD and IDD by breaking the protein into substructures of seven contiguous residues. These substructures are classified based on the internal coordinates system of $\mathbf{C}_\alpha$-$\mathbf{C}_\alpha$ virtual bond vectors, and then dynamic programming is used to determine a global alignment using the local substructure similarities and an additional penalty term for internal gaps. Finally, the maximum subset of similar substructures within 4 angstroms of each another is found and used to compute a percentage of structural identity index and corresponding z-score assuming a null model based on random structural alignment. While it is not clear that MAMMOTH would necessarily overcome the limitations often found when seeking a single global structural alignment, it calculates information about local structural features, like secondary structures, which are then leveraged to calculate the similarity z-score.

2.2.3 Static Methods for Analyzing Disordered Proteins

Radius of Gyration (R_g) and Shape (S) Parameters

A commonly employed metric for quantifying the size of a single protein conformation is the radius of gyration, R_g. This descriptor quantifies the general size of the conformation, allowing easy comparison between conformations. It can be computed

from the eigenvalues of the gyration tensor, $\mathbf{T}$:

$$\mathbf{T} = \frac{1}{N}\sum_{i=1}^{N}(\bar{\mathbf{r}} - \mathbf{r}_i)(\mathbf{r}_i - \bar{\mathbf{r}})^\top \tag{2.4}$$

where N is the number of C_α atoms in the protein, $\mathbf{r}_i$ is the position vector of the ith C_α atom, and $\bar{\mathbf{r}}$ is the position vector the center of mass of the C_α atoms. Let $\lambda_{1,2,3}$ denote the eigenvalues of $\mathbf{T}$, then R_g can be computed as follows:

$$R_g = \sqrt{\lambda_1 + \lambda_2 + \lambda_3} \tag{2.5}$$

Another metric is the shape parameter S, which provides an indication of the general shape of the conformation and can also be derived from the eigenvalues of the gyration tensor:

$$S = 27\left(\frac{\prod_{i=1}^{3}(\lambda_i - \bar{\lambda})}{(\lambda_1 + \lambda_2 + \lambda_3)^3}\right) \tag{2.6}$$

where $\bar{\lambda}$ is the mean of the three eigenvalues of the gyration tensor. The shape parameter falls within the range $[-0.25, 2]$ where a value of $S = 0$ indicates a perfectly spherical conformation, $S < 0$ indicates an oblate, or flattened, conformation, and $S > 0$ indicates a prolate, or extended, conformation. Therefore, while the radius of gyration may be interpreted as a measure of the size of a sphere that would circumscribe the conformation, the shape parameter describes the shape of a circumscribing ellipsoid [15]. Several example protein conformations and their calculated R_g and S values are shown on the left-hand side of Figure 2.3 and some examples of the three classes of ellipsoids captured by the S parameter are shown on the right-hand side.

Secondary Structure

Secondary structures for the proteins were examined using the program DSSP [16]. This program takes a single conformation, and assigns each residue to a secondary structure class based on its geometry and locality in relation to other residues. While the two

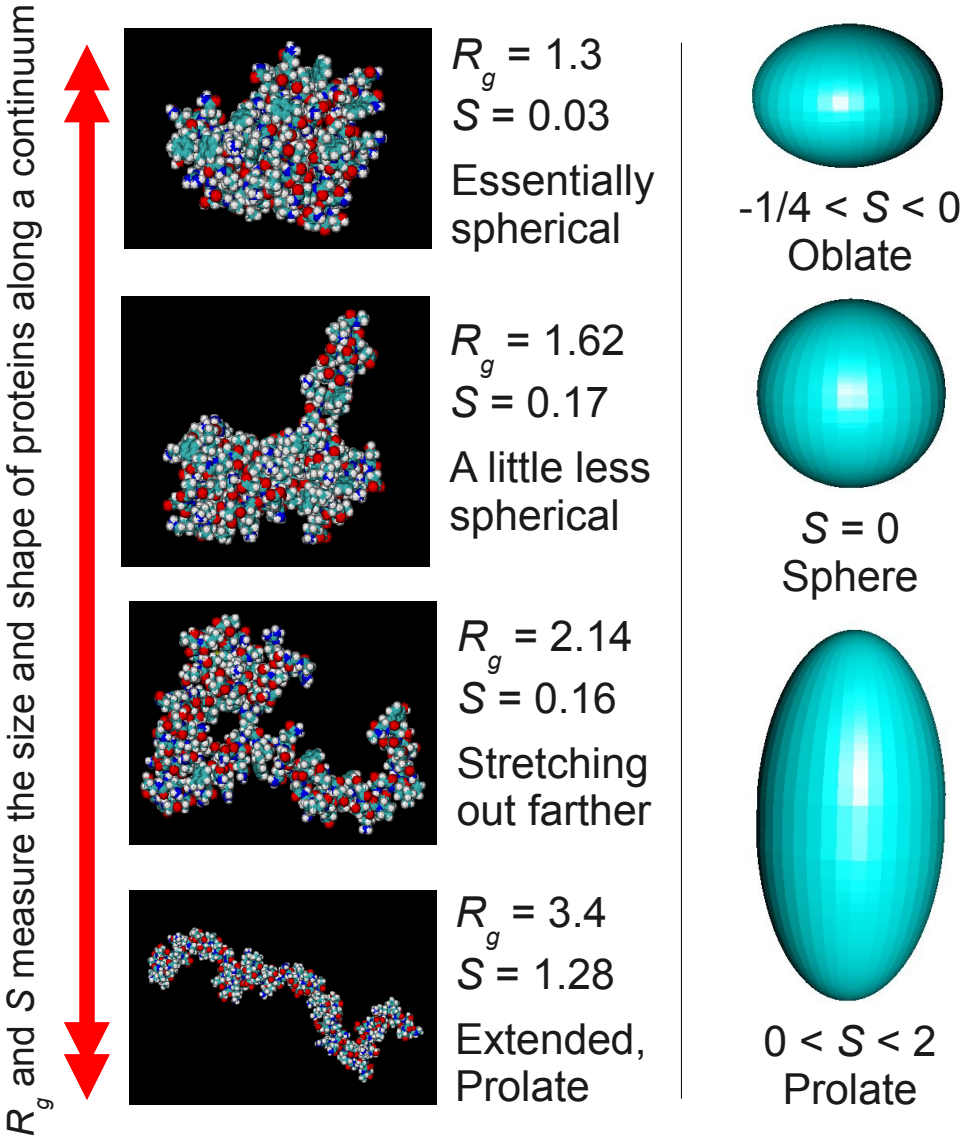

Figure 2.3: Examples of various intrinsically disordered protein structures and their corresponding radii of gyration (R_g) and shape parameters (S) are shown on the left, emphasizing the continuum of compaction and internal arrangement that disordered proteins exhibit. On the right are examples of the three different classes of ellipsoidal structure captured by the shape parameter, S.

primary types of secondary structure are α-helices and β-sheets, DSSP can classify residues into many additional structural motifs. For example, two additional kinds of helical structures that it can classify are the Π-helix and 3-10-helix. There is also the "turn" class assigned to residues that are similar to one of the helical structures, but that don't make the necessary contacts with other residues to be classified as one of the three types of helices. Besides the β-sheet, there is also a similar class called the β-bridge, as well as a "bend" class for β-like structures that don't make the necessary contacts with other residues, analogous to the "turn" class for helices. Finally, the program will classify any residue which doesn't fall into any of the above categories as a "coil". Disordered proteins exhibit only transient secondary structural arrangement, so examining the fraction of the time that these structures are observed during the simulations provides additional details about the internal arrangement of the protein chains.

Interresidue Distance Maps

A commonly employed method for extracting structural information from protein simulations involves measuring the distances between all pairs of residues along the protein chain. Often, these distances are converted into a set of "contacts" where only residues within a small specified distance (ϵ) are said to be in contact with one another. Using this formalism along with an ensemble of protein structures generated from molecular dynamics simulations, this binary contact information can be used to extract the probability with which any pair of residues are in close proximity. By creating an image where each pixel p_{ij} corresponds to a pair of residues along the chain, namely residue i and residue j, and setting the intensity value of this pixel to be proportional to the probability of residues i and j being in contact ($p(R_{ij} <= \epsilon)$), the structural properties of the protein can be quantified in a succinct manner that also lends itself to visualization. This method has been employed extensively for studying the structure of natively folded proteins.

While contact maps are routinely applied to the study of natively folded proteins,

their usefulness in studying disordered proteins is diminished. For a folded protein, certain residues will tend to stay in contact over a persistent period of time, which can be observed visually by the strong intensity values in the protein's contact map. In contrast, IDPs do not show persistent local arrangement. Instead, pairs of residues will often be in contact for only short periods of time, or the protein may adopt many different global arrangements over extended periods of time. In order to quantify and examine the structural variation of IDPs, a novel, generalized version of contact maps is developed here called *interresidue distance maps*.

Distance maps set each pixel, p_{ij}, to a color-mapped value of the average Euclidean distance between residues i and j: $\langle R_{ij} \rangle$. This would result in a distance map image that is mirrored across the diagonal. However, information about the variation in $\langle R_{ij} \rangle$ is also included by placing the standard deviation of R_{ij} in the upper triangle of the distance map, and keeping the average values in the lower triangle. Therefore, distance maps contain not only more detailed information about the distances between residues than standard contact maps, but also information about the distribution around these averages. An example of how to construct an interresidue distance map is shown in Figure 2.4.

R_g–S Histograms

Another recent technique for studying the dynamic structure of IDPs is examining the joint distribution of R_g and S for a conformational ensemble [17]. This can be accomplished by generating a two-dimensional histogram where each bin corresponds to a unique range of R_g and S values. While distance maps provide some insight into the average structural arrangement of IDPs and fluctuations around this average structure, R_g–S histograms lend themselves to interpretation of structural dynamics by resolving the distribution of these conformational structures.

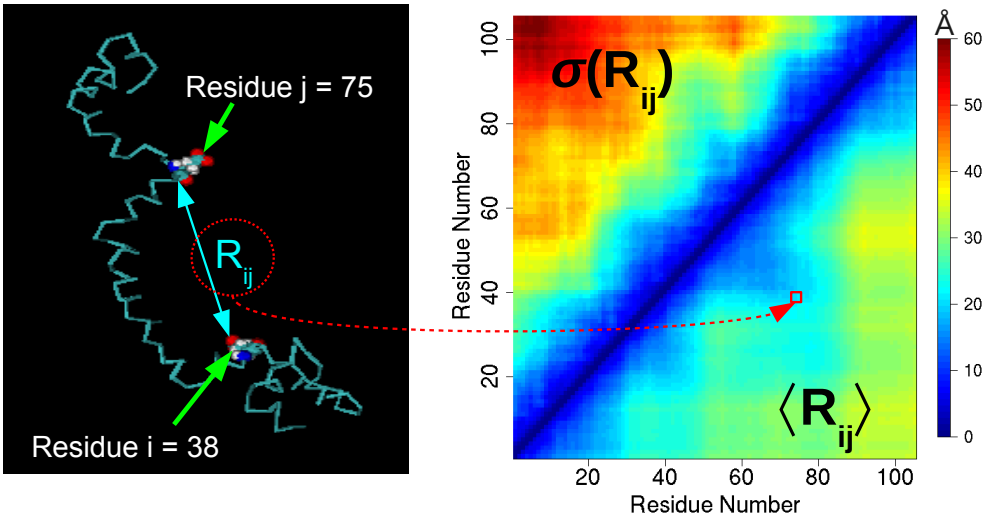

Figure 2.4: Example of distance map construction. The distance between residue i and residue j is denoted as R_{ij}. This values is computed for each structure in an ensemble, the values are averaged ($\langle R_{ij} \rangle$), and the result is reported as a color-coded picture in the corresponding location on the lower diagonal of the distance map. The standard deviation is also calculated, $\sigma(R_{ij})$ and plotted in the corresponding location in the upper diagonal as well.

2.2.4 Dynamic Methods for Analyzing Disordered Proteins

Distance from Previous Structures

A measure of dynamic change in protein conformation is revealed by plotting the structural dissimilarity between the initial and current structures versus time. The results of multiple simulations can be averaged to both visually and quantitatively confirm the rate of change. However, the measure of dissimilarity used may have a significant impact on both the consistency and reliability of the results. Typically, RMSD is employed to assess structural dissimilarity, but this method often performs poorly for very dissimilar structures. Therefore, two additional techniques for assessing structural dissimilarity are also employed here: MAMMOTH [14] and backbone Φ-Ψ angle distances [12]. MAMMOTH is calibrated to include secondary structure information, but still relies heavily on structural alignment similar to RMSD and produces a z-score instead of a geometry-based distance. The backbone angle distance measure is rotationally and

translationally invariant, but is not as commonly used or intuitive as alignment-based techniques.

Instead of measuring the structural dissimilarity between structures in the simulation from the initial structure, one may also desire to look at the structural dissimilarity at a specific time interval in the past. In this case, every structure at time t is compared to the structure at time $t - \Delta t$ where Δt is the time-scale of interest. Therefore, plotting these values as a function of time for several values of Δt can potentially uncover certain conformational changes that might not be observed when simply comparing to the initial structure only.

Backbone Angle Autocorrelation

Autocorrelation functions are a well-established method for estimating the dynamic change in some measurable quantity. For proteins, the calculation of autocorrelation functions using the backbone Φ-Ψ angles has intuitive appeal. The autocorrelation for each angle can be computed independently, but this approach would not take interactions between the various residues into consideration. Therefore, a vector containing information from all Φ-Ψ angles can be constructed using the *conjugate-complex* form of the angles. If there are a total of n angles in the protein of interest, then denote θ to be the n-element long vector of these angles. Then, construct the $2n + 1$-element complex-conjugate vector of these angles, $\mathbf{c}$, as follows:

$$\begin{aligned} \mathbf{c}_1 &= 1 + 0i \\ \mathbf{c}_{x+1} &= e^{i\theta_x} \\ \mathbf{c}_{2n+1-x} &= e^{-i\theta_x}, \forall x = 1 \ldots n \end{aligned} \tag{2.7}$$

where $i = \sqrt{-1}$. By taking the inverse Discrete Fourier Transform of $\mathbf{c}$, the result is, by construction, a unit-length real vector of length $2n + 1$ for each conformation. This vector can be used to calculate the autocorrelation function for all Φ-Ψ angles simultaneously, thus including all internal correlations among the angular dynamics [13].

Structural Decorrelation Time

Another measure of protein dynamics is the structural decorrelation time (τ_{dec}) of Lyman and Zuckerman [18]. This measure is calculated using a simple binning technique to construct structural histograms for different time-spans in a simulation. The general idea behind this approach is that the relative populations of the bins should be quite different at short time intervals, but should be similar at long time intervals. In particular, at very long time scales, a sample should appear independently and identically distributed, and this assumption allows the calculation of an observable average normalized variance of bin occupancy, $\sigma^2_{\rm obs}(t)$, where t is the time interval between subsequently sampled structures from a simulation. The value of $\sigma^2_{\rm obs}(t)$ is high for small t, indicating that the simulation is not making much progress through the conformation space over this amount of time (structures separated at short times are typically still assigned to the same bin or same set of bins). However, as t is increased, $\sigma^2_{\rm obs}(t)$ will eventually decrease to 1, indicating the time at which the simulation is equally likely to end up in any of the structural bins, i.e. after time t two structures appear to be effectively decorrelated. Hence, a plot of $\sigma^2_{\rm obs}(t)$ versus t will allow one to determine the structural decorrelation time, $\tau_{\rm dec}$, for a simulation. Additionally, the number of structures $N = 2$ separated by t can also be made larger in order provide a more robust estimate of τ_{dec} by decreasing the variance in the estimates of $\sigma^2_{\rm obs}(t)$.

In a sense, the decorrelation time can be thought of as a general measure of structural change, not just as a convergence statistic. For systems that have a large number of degrees of freedom to search over, it could take an exceptionally long time to search over the parameter space of the system. IDPs can be thought of as proteins which have a large equilibrium conformation space, and the decorrelation time for these proteins should be quite large for single simulations. NFPs on the other hand, when simulated starting from the folded structure, will not stray far from the folded conformation and quickly sample all of the conformation space that one can reasonably expect from these proteins. So, NFPs should exhibit rather short decorrelation times. How-

ever, there is no method that can determine definitively whether or not a simulation has converged [19], and more complex systems should provide even more of a challenge in this regard. Therefore, interpreting the structural decorrelation time for IDP simulations might prove more challenging than interpreting it for NFP simulations.

2.2.5 Boxplots

In many figures, the boxplot is utilized to represent data distributions [20]. The colored box represents the data range from the first quartile to the third quartile, with the median represented by a black line across the central box region. The notches in the sides of the box roughly approximate a 95% confidence interval, extending around the median by $\pm\frac{1.58\times R_{IQ}}{\sqrt{n}}$, where R_{IQ} is the interquartile range which is defined as the difference between the third and first quartiles and n is the number of data elements. The bottom and top whiskers each extend an additional 1.5 times the distance from the median to the first and third quartiles, but they are truncated to the minimum and maximum data values, respectively, if there are no outliers present. Outliers are plotted as circles above and below the whiskers.

2.3 Results

2.3.1 Static Methods

Radius of Gyration

The radius of gyration for all structures across all of the simulations was calculated, and an average R_g value was then computed for each simulation independently. The distribution of the average R_g for each of the different proteins is shown in Figure 2.5. Overall, the 3ns and 18ns simulations are in good agreement across all of the proteins, although the 18ns simulations did not cover as broad of a range of R_g values as the 3ns simulations. While the distributions for GLAG, GAFG, and GLFG are heavily

overlapping for the 3ns and 18ns simulations, the 2ns simulations provide some additional insight. Taken together, all of the R_g results indicate that the proteins in order from most extended to least extended is: SxSG, AxAG, FxFG, GLAG, GAFG, GLFG.

Shape Parameter

Just as the radius of gyration was computed for all structures, the shape parameter, S, was also computed for all structures. The average value for each simulation was taken and the distributions of these average values are shown in Figure 2.6. Interestingly, the S data are somewhat different than the R_g data for the proteins. In particular, while the SxSG protein was more extended than AxAG according to R_g, it appears that the shapes of the SxSG conformations are more "rounded" than those for AxAG. It could be that certain specific contacts are being formed by AxAG that give it a more prolate arrangement, but additional data are needed to determine if this is the case. Additionally, FxFG displays the widest variation in shape, suggesting that interesting structural rearrangements occur during the 3ns and 18ns simulations, but this is not shown to be as prevalent in the 2ns simulations. Meanwhile, GLAG displays more prolate arrangement than either GAFG or GLFG which are both very spherical for the 18ns simulations. This is consistent with the R_g results above which place GLAG in between its two mutants and FxFG in terms of overall compaction.

Secondary Structure

The secondary structure assignment for all residues of the proteins across all structures in the simulations was computed, and the results are plotted in Figures 2.7–2.12. Overall, there is clearly more prevalence for helical structures in FxFG and its mutants than GLFG and its mutants. In fact, Figures 2.7 and 2.8 show that the strength of this helical propensity correlates with the R_g data in Figure 2.5. Therefore, SxSG has the most α-helix content and GLFG has the least. In addition, there is relatively little β-sheet structure in these simulations, only some bend structures which only loosely

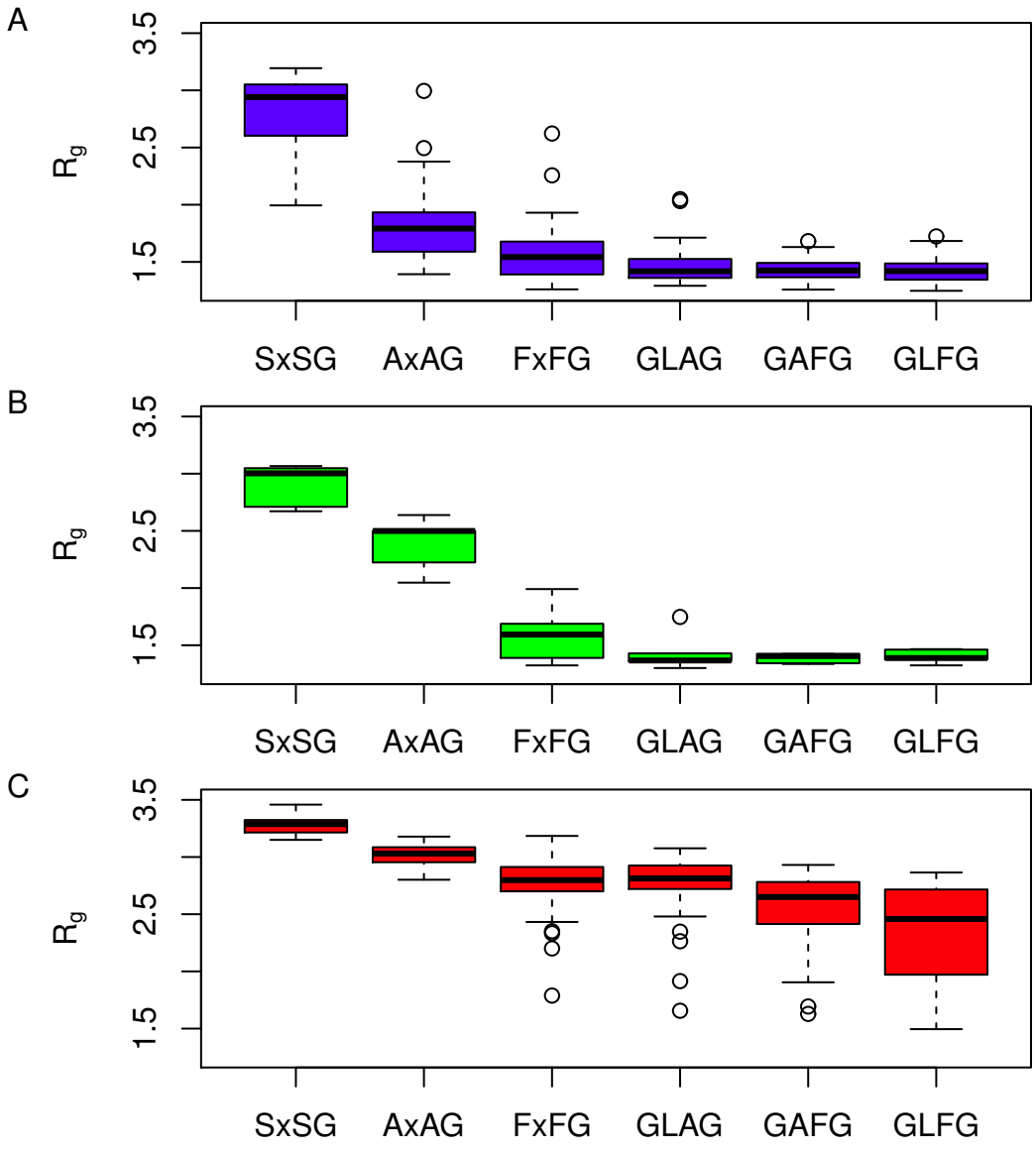

Figure 2.5: Radius of Gyration – (A) 3ns @ 300K (B) 18ns @ 300K (C) 2ns @ 350K

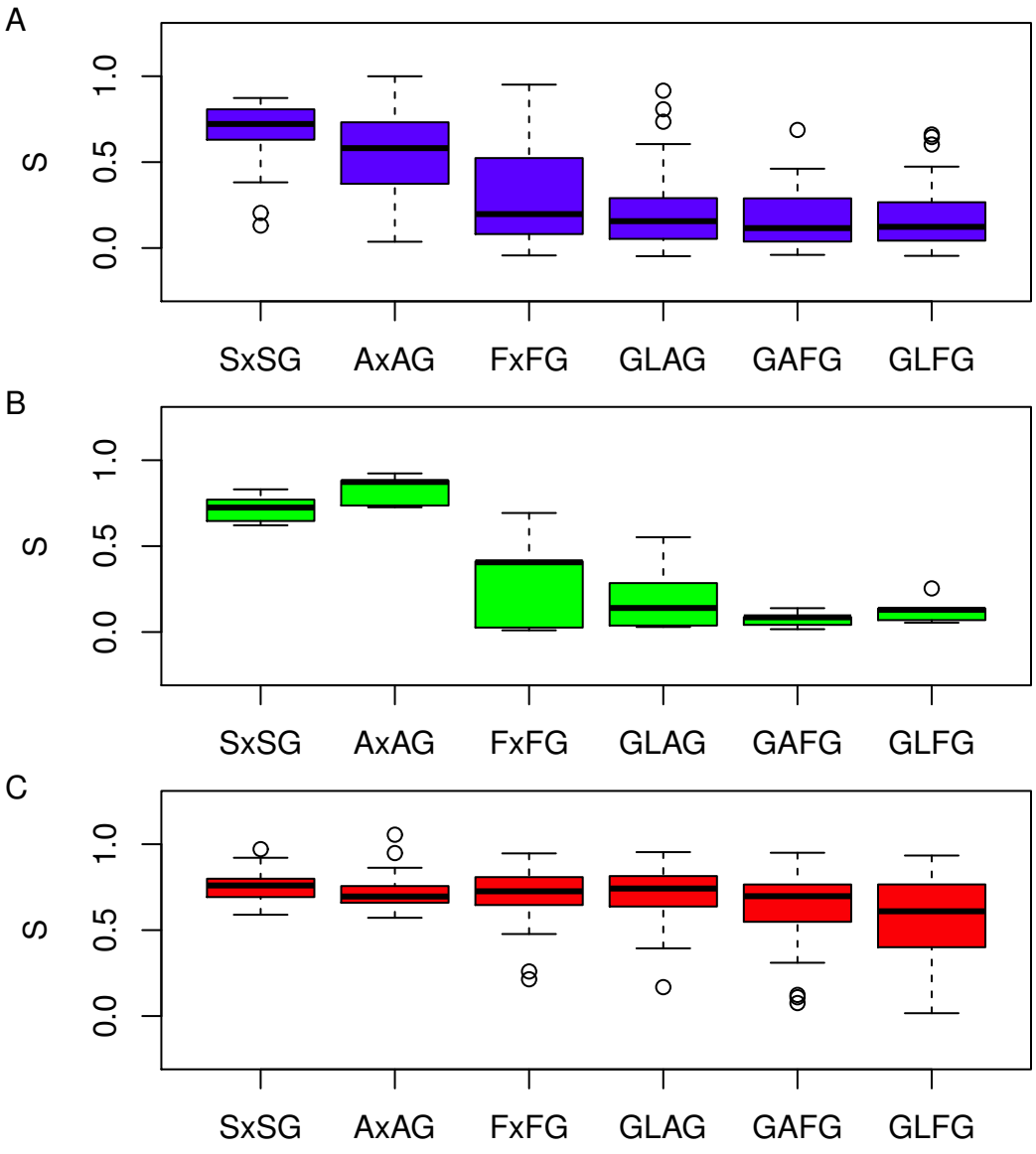

Figure 2.6: Shape Parameter – (A) 3ns @ 300K (B) 18ns @ 300K (C) 2ns @ 350K

resemble β-structures. Again, this correlates with the R_g data, making the protein with the least β-content SxSG, and GLFG having the least. In addition, there are regular patterns of helical propensity along the protein sequences, but the sequences are also spotted with marked decrease in the propensity for any structure at regular intervals. An examination of the sequences of these regions confirms them to be heavy in proline and glycine residues, which are known helix-disrupting amino acids.

Interresidue Distance Maps

Distance maps were generated for each of the FG-Nups by averaging distances over all replicates. The standard deviation of these distances is plotted in the upper triangle, and the mean in the lower triangle of each plot in Figures 2.13–2.15. The standard deviation has been scaled by a constant factor of 3.0 to make the upper triangle more readable. The units for all reported distance values are in angstroms (Å).

Overall, there are a few additional structural details that can be observed using this analysis. There is a trend to see an off-diagonal depression in the mean distances near one point for GLFG and its mutants in Figures 2.13D, 2.13E, and 2.13F. This indicates a propensity for the protein to fold over in this location. However, FxFG shown in Figure 2.13A, also shows a single off-diagonal depression. This one indicates that the fold over point for this protein is more in the center of FxFG, while the fold over point in GLFG and its mutants was slightly shifted to one side of the protein. The 18ns data confirms these locations, but also indicates an additional off-diagonal depression for GLFG and its mutants. This extra fold over point explains why the 18ns S parameters indicated strongly spherical conformations. This little fold over point must have often been loose in the 3ns simulations, making the structures more prolate. Overall, the standard deviation of the distances is higher for FxFG and its mutants than GLFG and its mutants, indicating larger amplitude conformational changes in FxFG and its mutants. The 2ns data in Figure 2.15 adds little additional information to the results as the off diagonal depressions are only faintly visible. However, these results

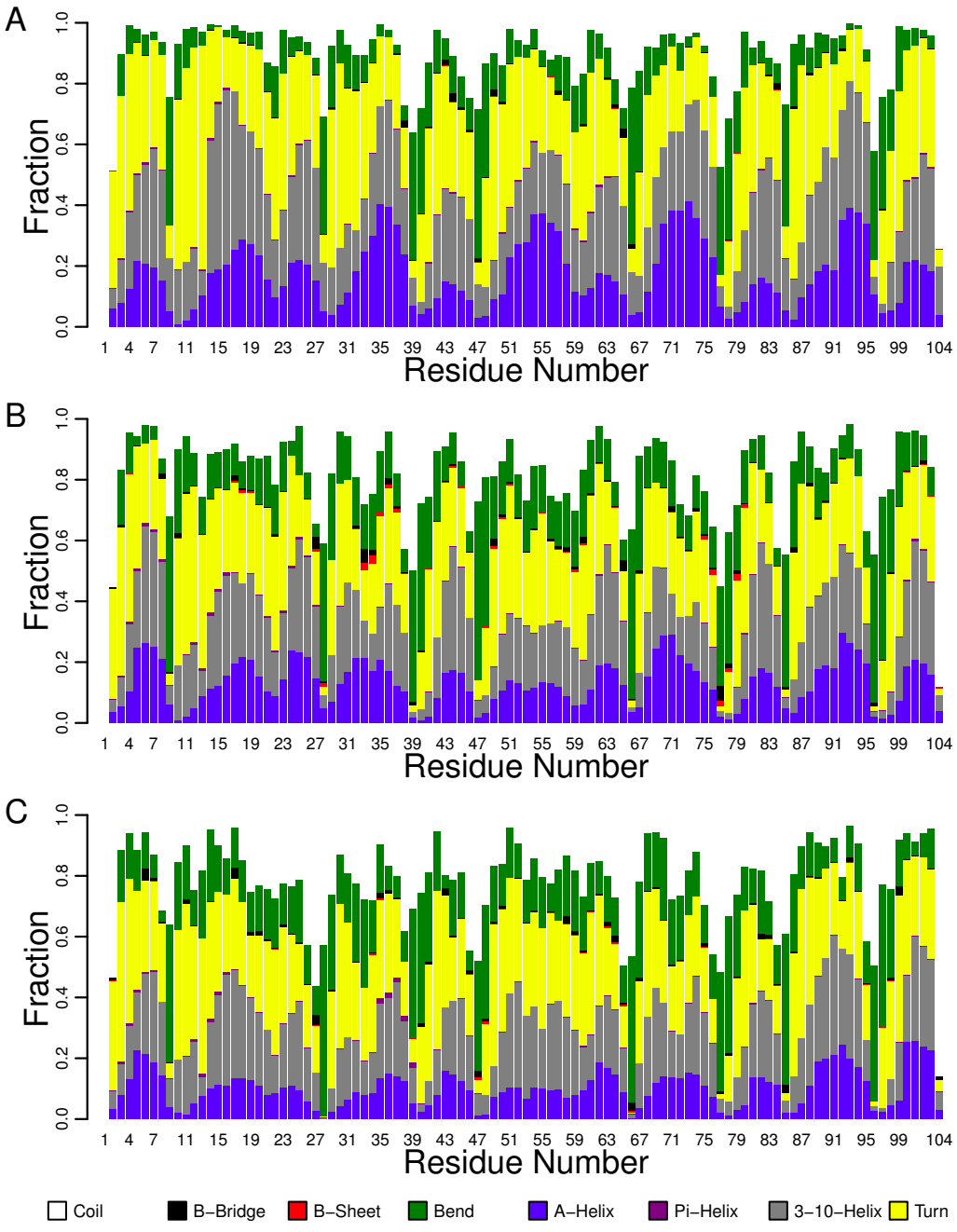

Figure 2.7: Secondary Structure – 3ns @ 300K – (A) SxSG (B) AxAG (C) FxFG

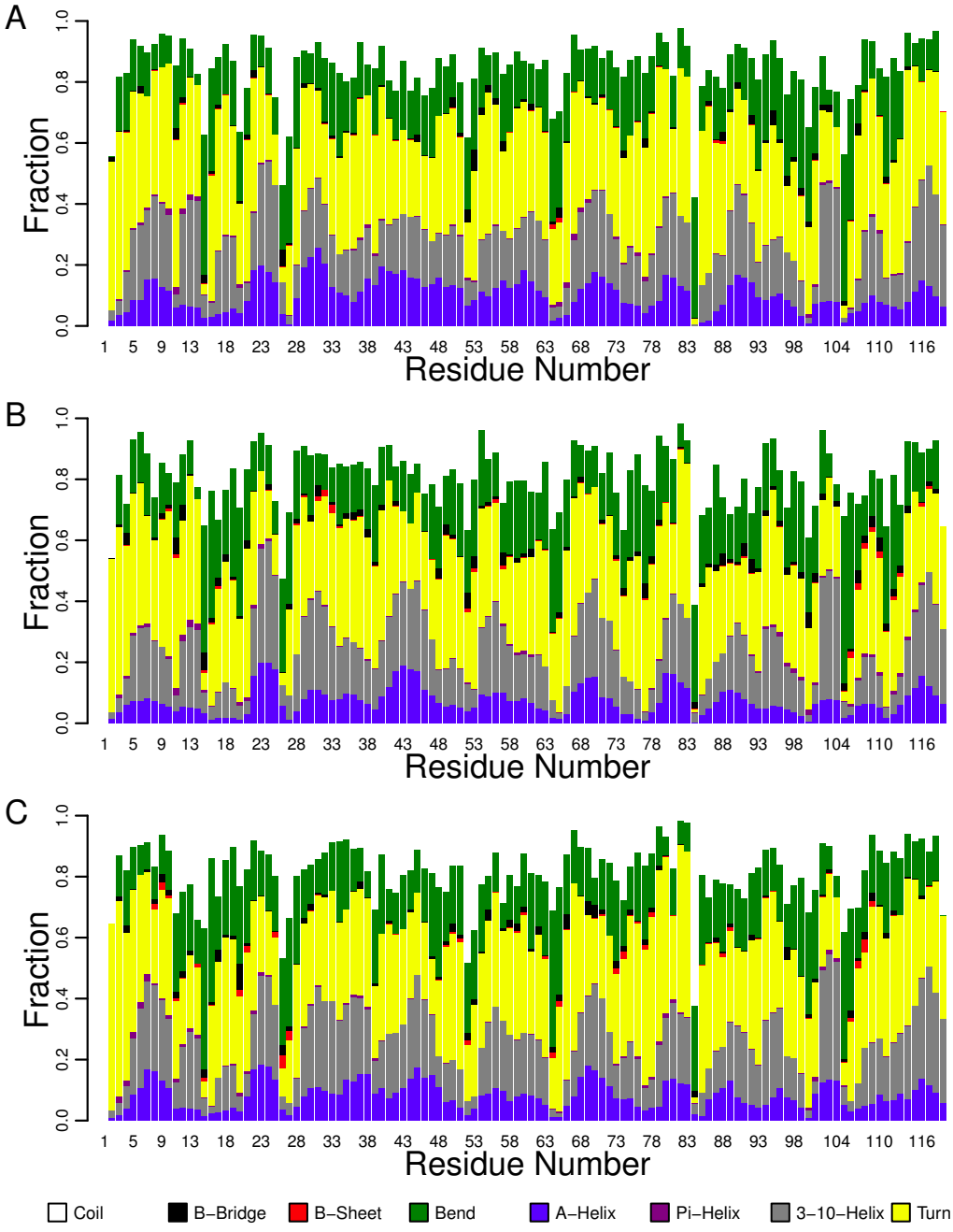

Figure 2.8: Secondary Structure – 3ns @ 300K – (A) GLAG (B) GAFG (C) GLFG

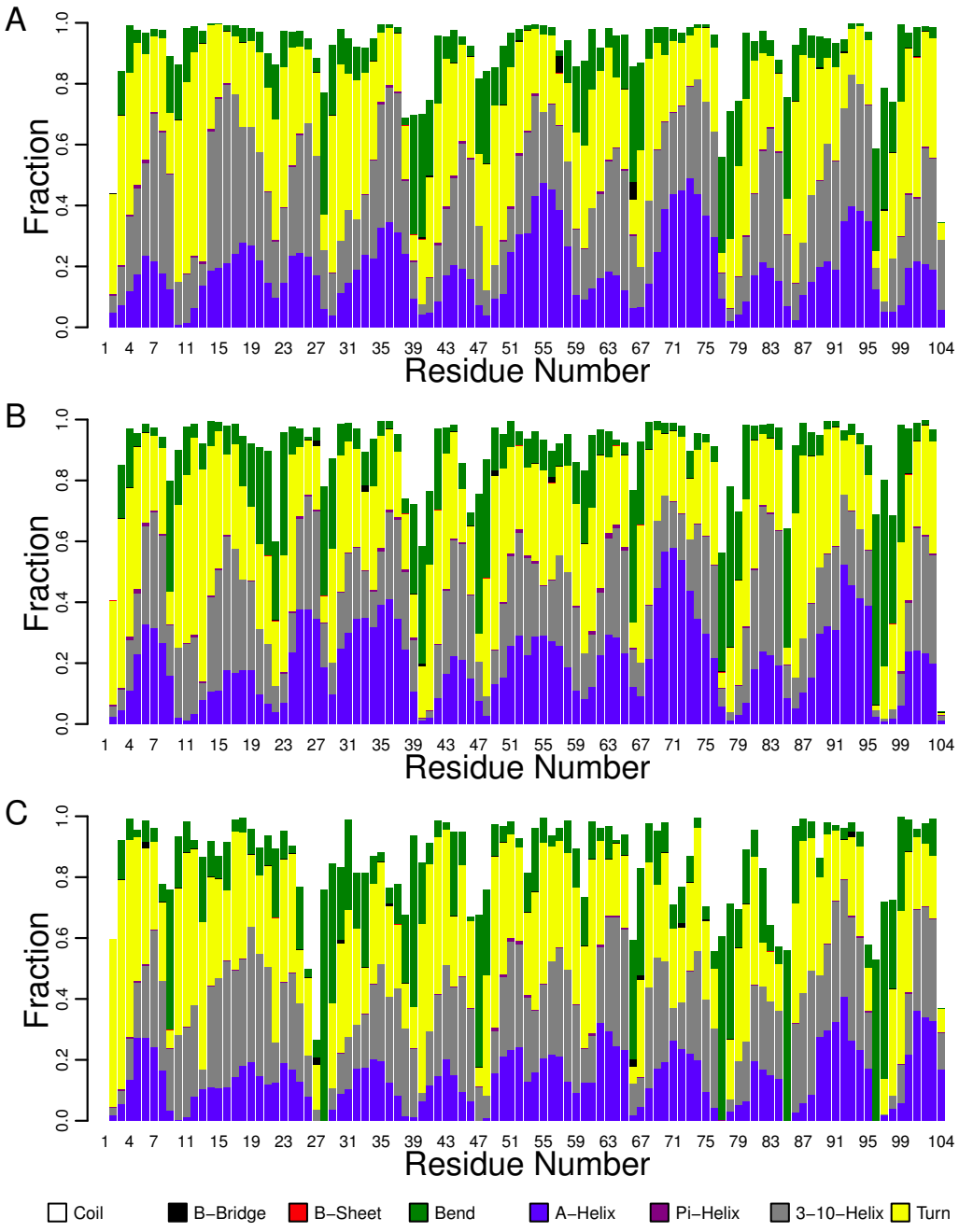

Figure 2.9: Secondary Structure – 18ns @ 300K – (A) SxSG (B) AxAG (C) FxFG

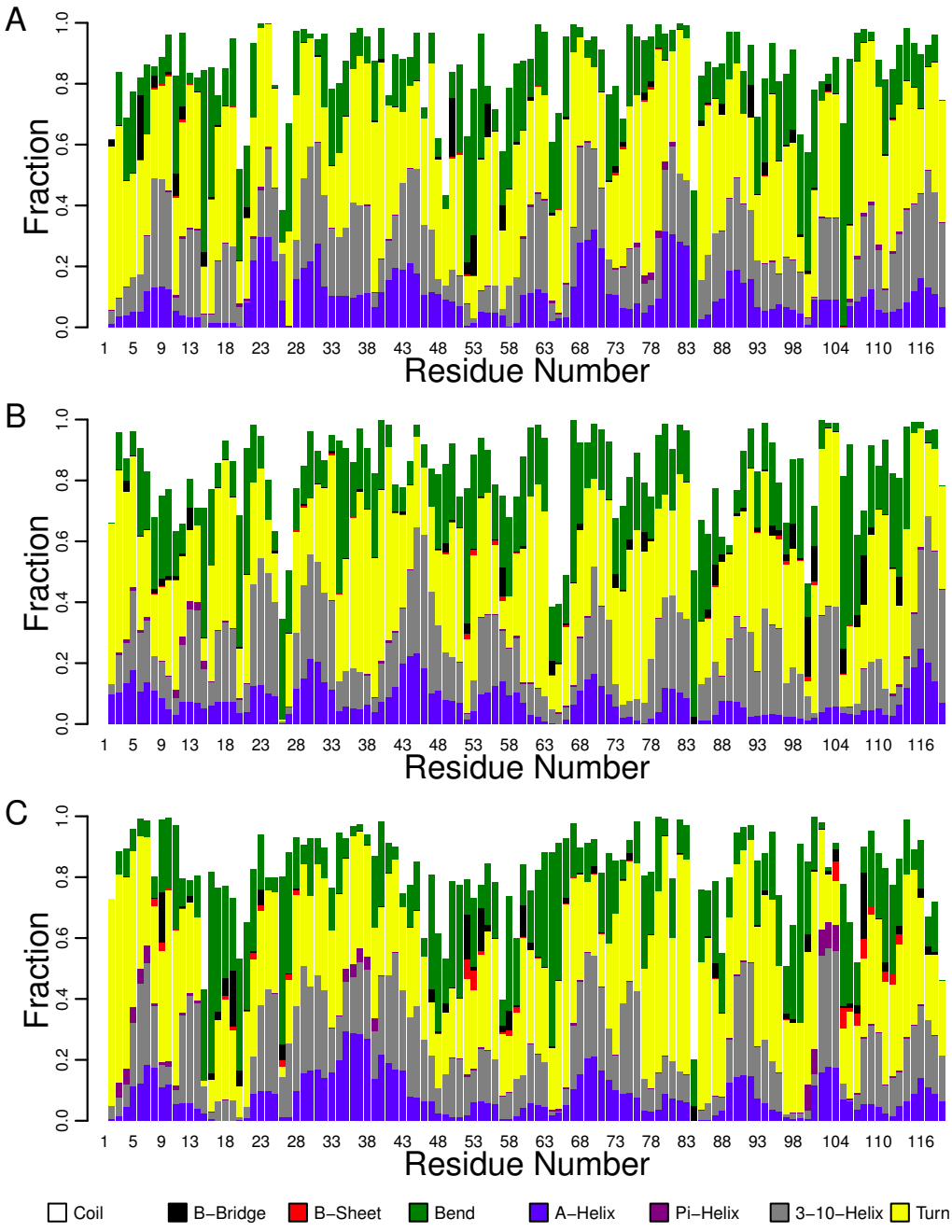

Figure 2.10: Secondary Structure – 18ns @ 300K – (A) GLAG (B) GAFG (C) GLFG

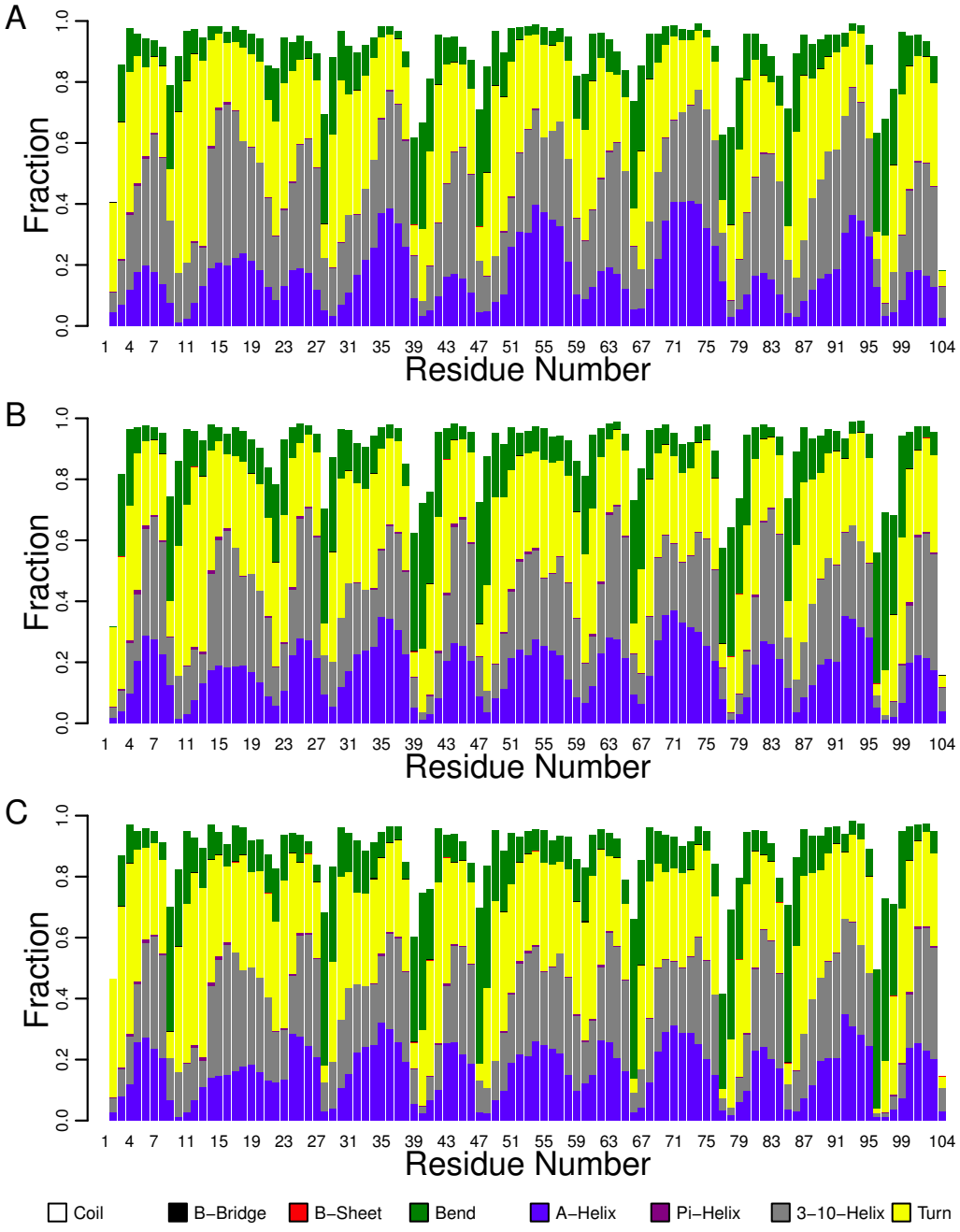

Figure 2.11: Secondary Structure – 2ns @ 350K – (A) SxSG (B) AxAG (C) FxFG

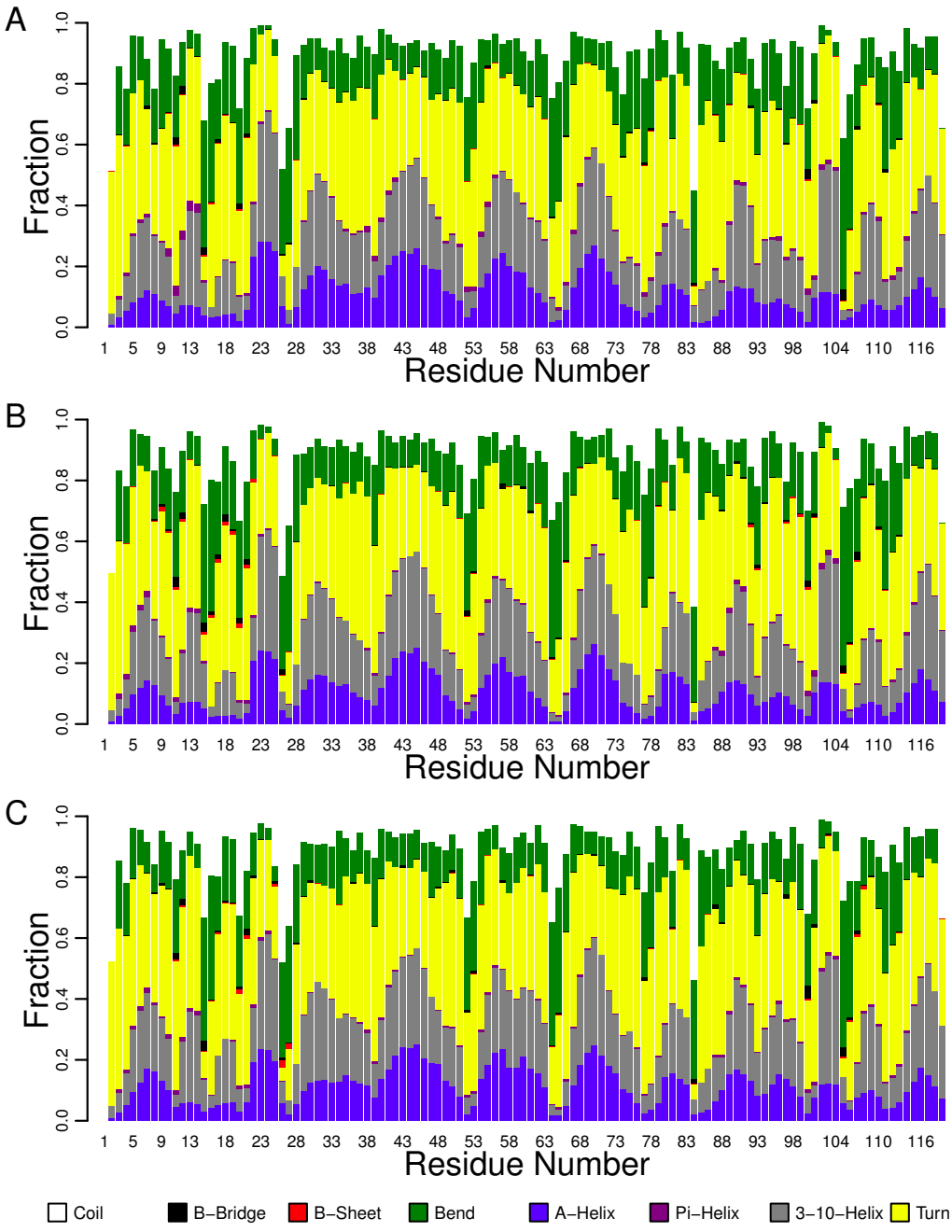

Figure 2.12: Secondary Structure – 2ns @ 350K – (A) GLAG (B) GAFG (C) GLFG

do indicate that the more compact structures exhibited by GLFG and its mutants, as described by the R_g calculations, are due to the propensity of this first fold over point which allows persistent contacts to form even at high temperatures.

R_g–S Histograms

R_g–S histograms for each of the FG-Nups were generated as well. For a single FG-Nup, all of the structures sampled from our simulations were taken and their R_g and S values were calculated. The two sets were then used to generate a single normalized 2D R_g–S histogram which shows the probability of observing each particular combination of R_g and S. Zero-probability regions in the plots are kept white (instead of the defined dark blue) in order to allow the boundary of the structural space to be easily identified.

Interestingly, when comparing the results for the 3ns simulations in Figure 2.16 to the 18ns simulations in Figure 2.17, the distributions for FxFG and its mutants seem to be more broad for the 18ns simulations, while GLFG and its mutants seem more peaked. In essence, the long runs let GLFG and its mutants settle into more compact, spherical structures. However, FxFG and its mutants would explore more extended, prolate conformations. Thus, the time scales of the simulations seem to affect the conformational sampling in different ways depending on the properties of the protein being simulated. In addition, the 2ns results in Figure 2.18 indicate that the conformations explored by the more extended proteins, AxAG, SxSG, and to a lesser extent, FxFG, show a stronger correlation between R_g and S. Thus, the strength of this correlation at higher temperatures is also a good indicator of structural differences between the proteins.

2.3.2 Dynamic Methods

Distance from Previous Structures

The distance from the initial structure as a function of time was computed for all replicate simulations for all of the FG-Nups. After this, the replicate results were aver-

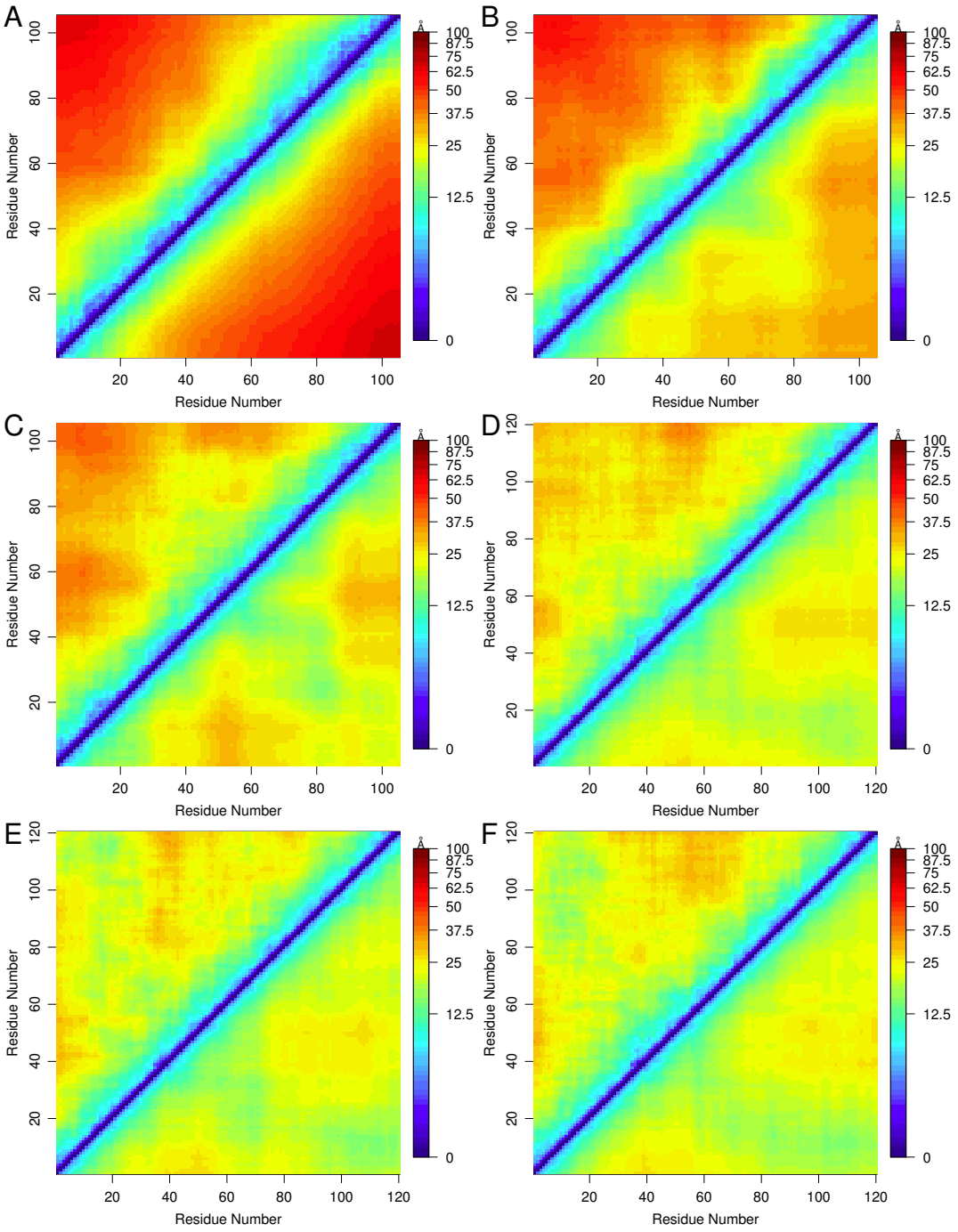

Figure 2.13: Interresidue Distance Maps created by averaging across all 3ns simulations at 300K for (A) SxSG, (B) AxAG, (C) FxFG, (D) GLAG, (E) GAFG, and (F) GLFG. The lower/upper diagonal shows the mean/standard deviation of distances between all pairs residues (standard deviation is scaled by a factor of 3 to enhance detail.)

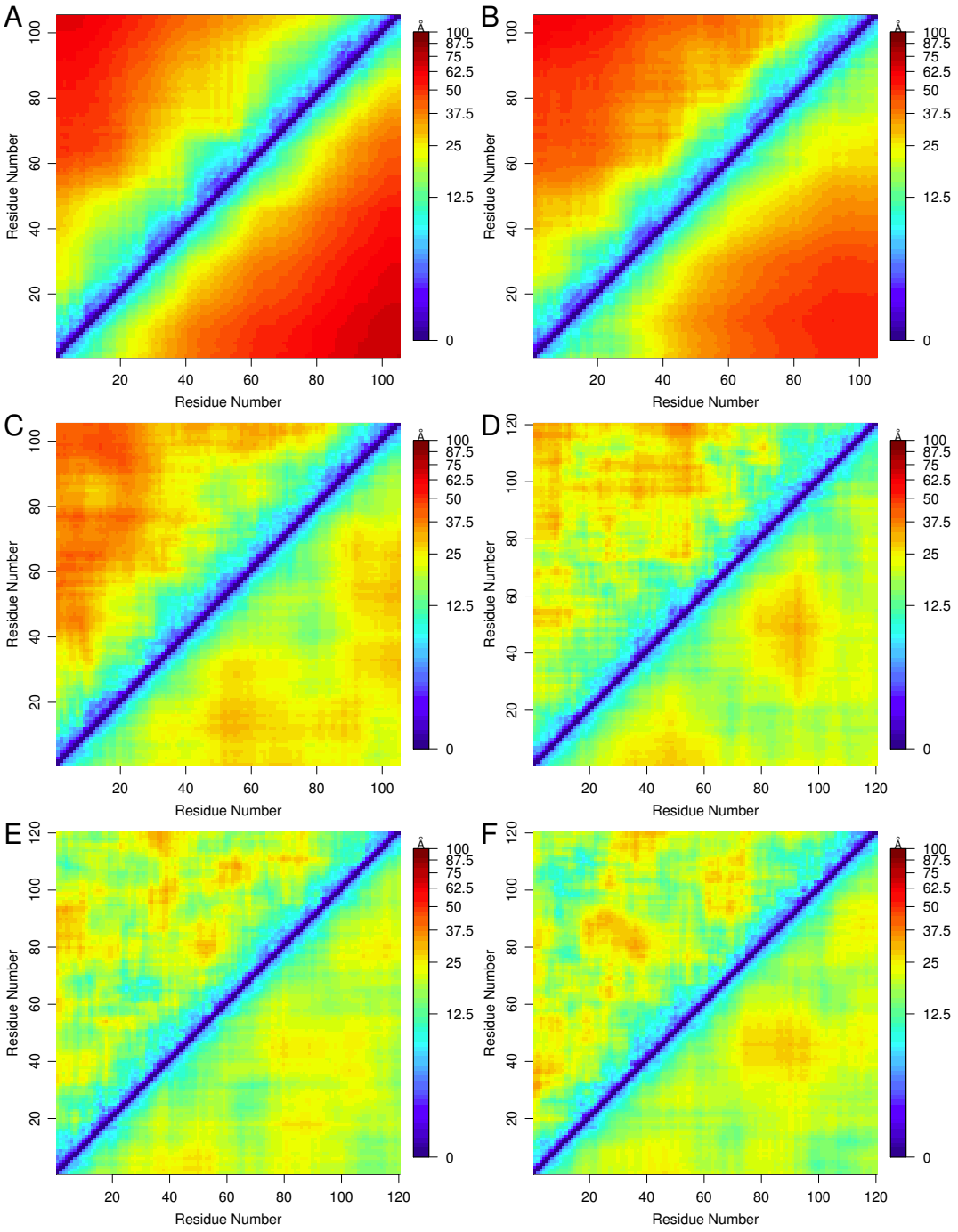

Figure 2.14: Interresidue Distance Maps created by averaging across all 18ns simulations at 300K for (A) SxSG, (B) AxAG, (C) FxFG, (D) GLAG, (E) GAFG, and (F) GLFG. The lower/upper diagonal shows the mean/standard deviation of distances between all pairs residues (standard deviation is scaled by a factor of 3 to enhance detail.)

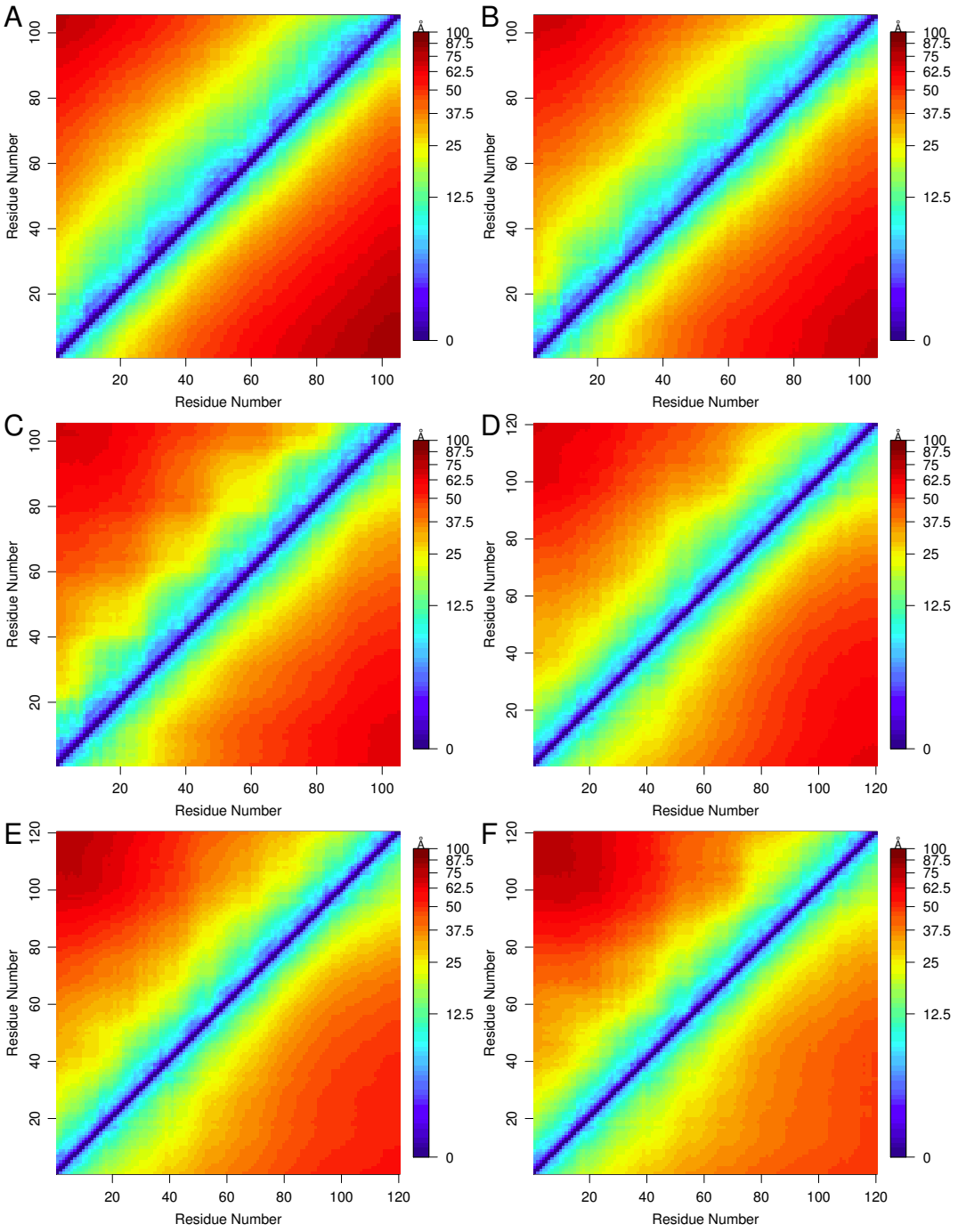

Figure 2.15: Interresidue Distance Maps created by averaging across all 2ns simulations at 350K for (A) SxSG, (B) AxAG, (C) FxFG, (D) GLAG, (E) GAFG, and (F) GLFG. The lower/upper diagonal shows the mean/standard deviation of distances between all pairs residues (standard deviation is scaled by a factor of 3 to enhance detail.)

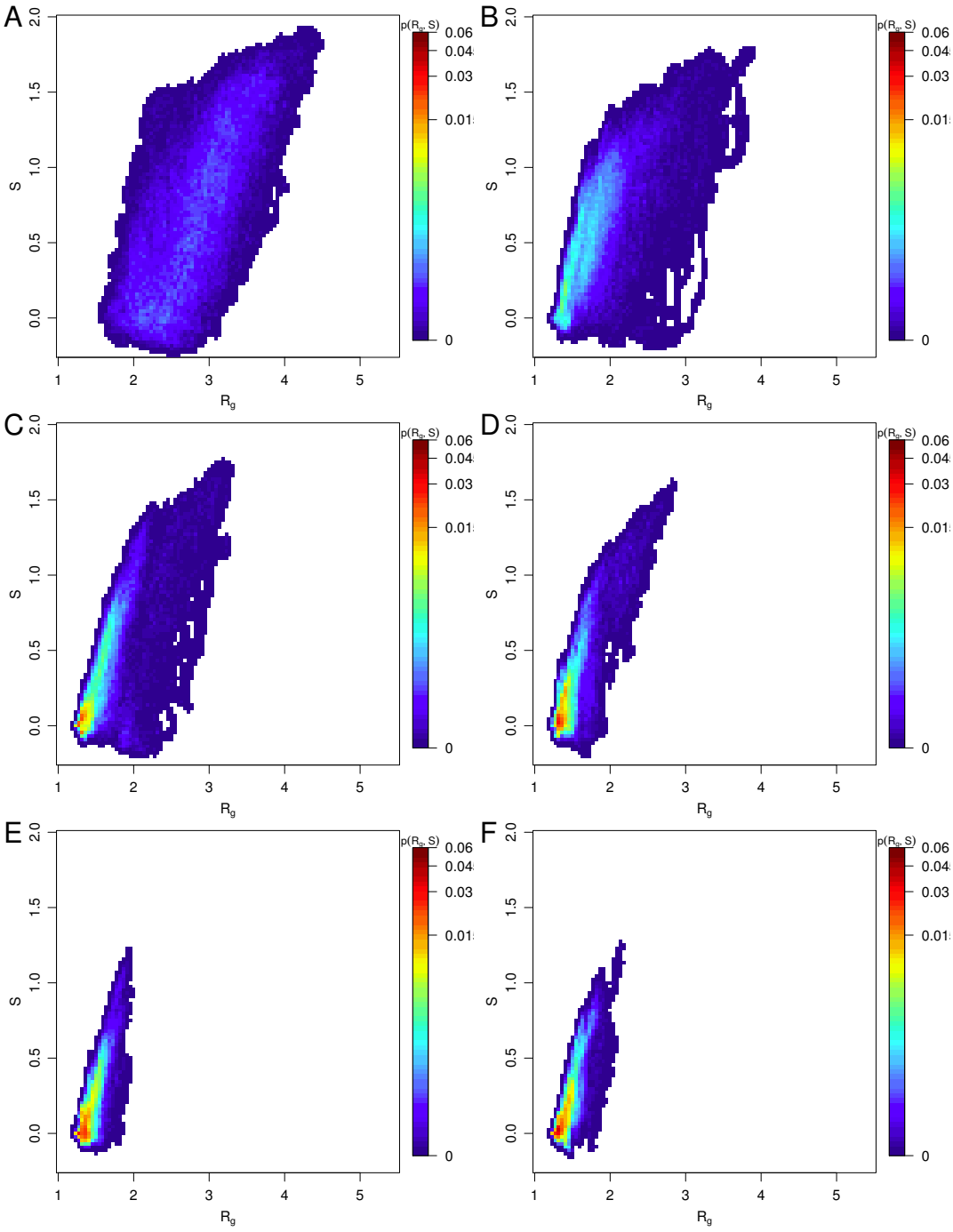

Figure 2.16: R_g–S histograms showing the probability of structures arising with particular combinations of R_g and S compiled from all 3ns simulations at 300K of (A) SxSG, (B) AxAG, (C) FxFG, (D) GLAG, (E) GAFG, and (F) GLFG

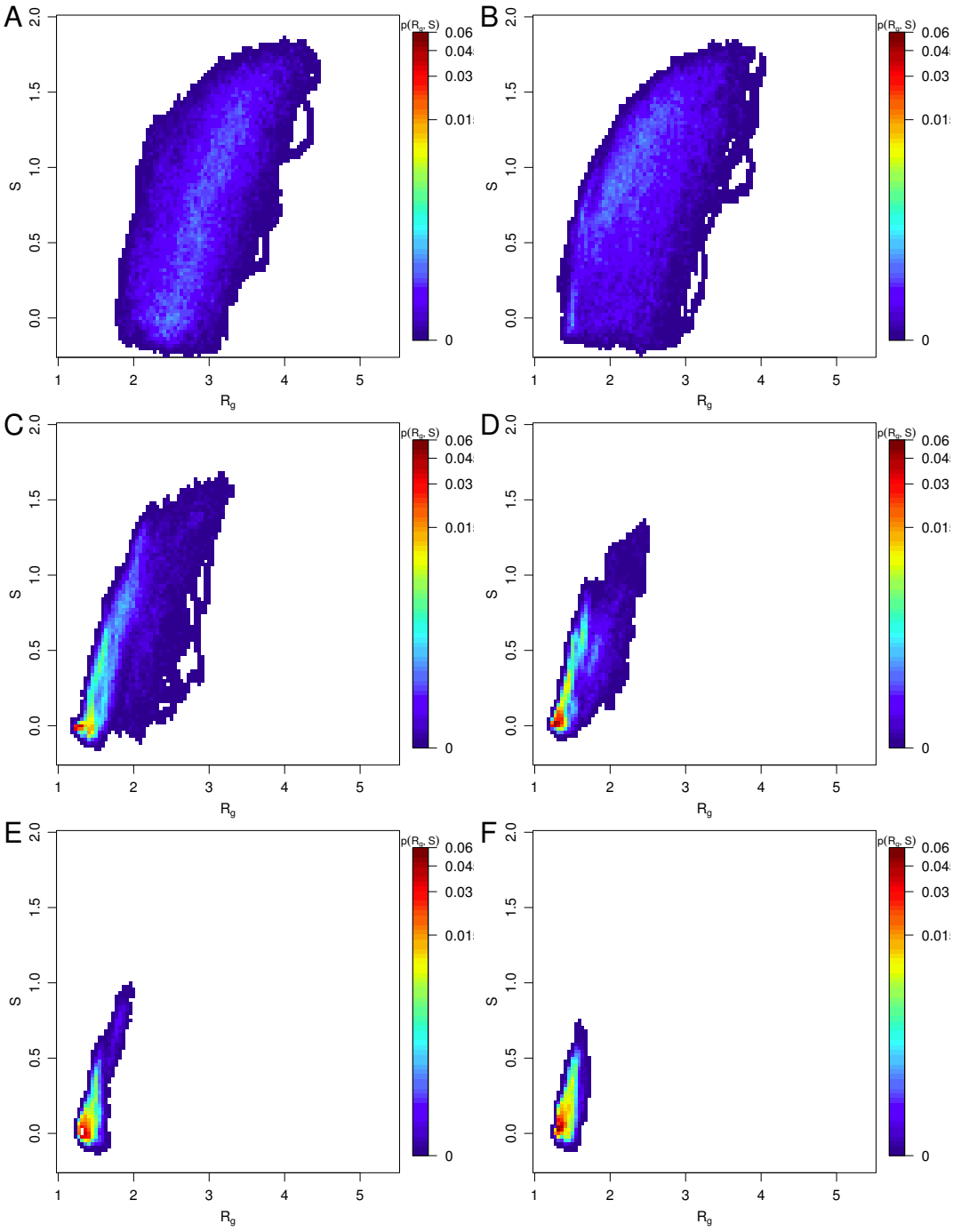

Figure 2.17: R_g–S histograms showing the probability of structures arising with particular combinations of R_g and S compiled from all 18ns simulations at 300K of (A) SxSG, (B) AxAG, (C) FxFG, (D) GLAG, (E) GAFG, and (F) GLFG

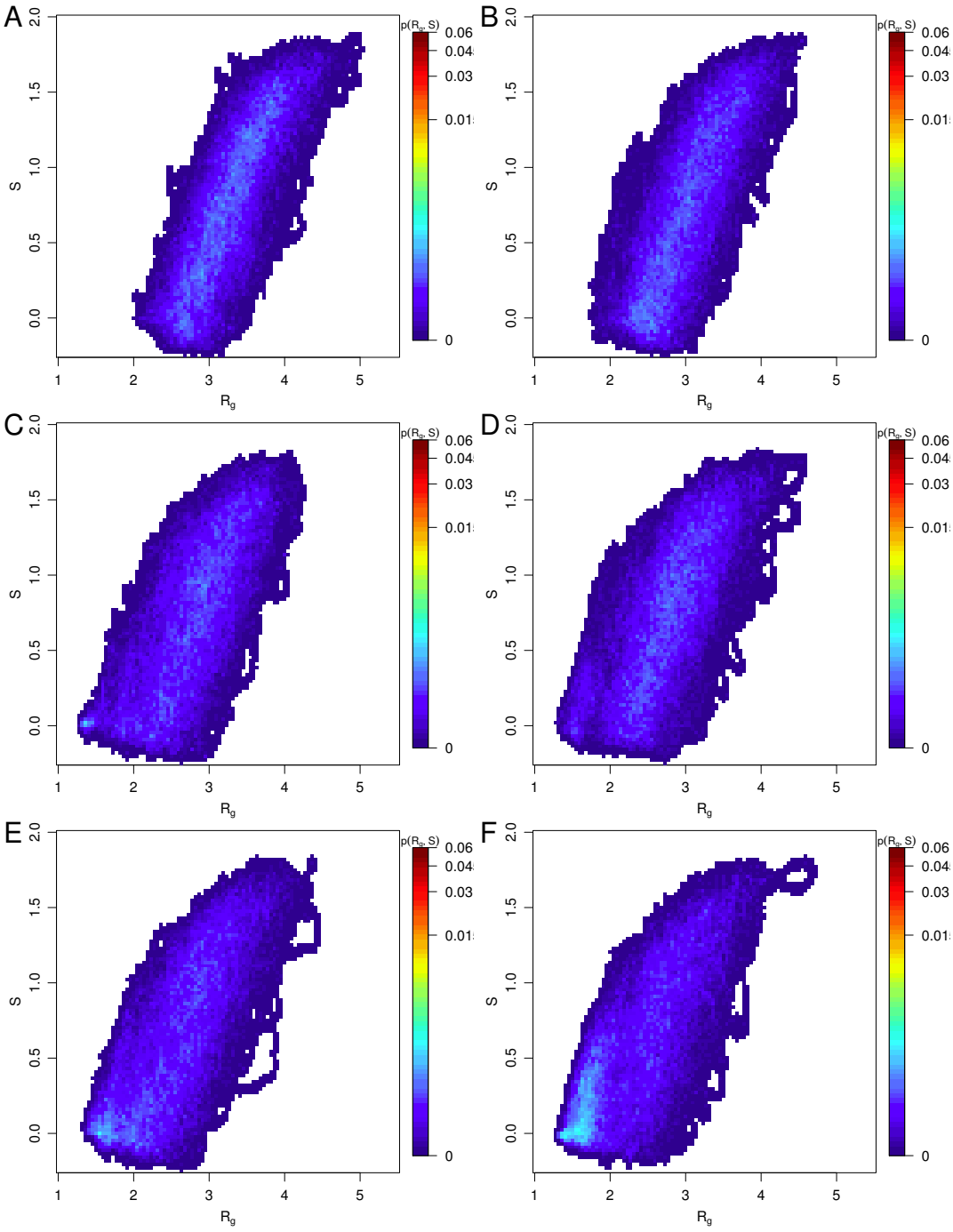

Figure 2.18: R_g–S histograms showing the probability of structures arising with particular combinations of R_g and S compiled from all 2ns simulations at 350K of (A) SxSG, (B) AxAG, (C) FxFG, (D) GLAG, (E) GAFG, and (F) GLFG

aged at each time point to create the plots of RMSD (Figure 2.19), MAMMOTH z-score (Figure 2.19), and the Euclidean Φ–Ψ distance (Figure 2.19) versus time. The first thing to note is that MAMMOTH acts as a similarity measure instead of a dissimilarity measure. So, high z-scores indicate high probability of structural overlap while low z-scores indicate a low probability of structural overlap. For the 3ns and 18ns data, the RMSD and MAMMOTH results are in close agreement. However, the Φ–Ψ distance results for these simulations are not in agreement with the RMSD and MAMMOTH results. Both RMSD and MAMMOTH indicate that the most structural change is occurring in the SxSG simulations and the least in the GLFG simulations, with the other proteins in between following the same order as their order according to R_g. While the 2ns data is less conclusive for the RMSD and MAMMOTH results because the results for all of the proteins are heavily overlapping, the Φ–Ψ distance reports GLFG and its mutants to be more dynamic than FxFG and its mutants. This is surprising, but perhaps there are certain traits of the motion that are not observable using alignment-based methods.

Figures 2.22, 2.23, and 2.24 show the results for RMSD, MAMMOTH, and Φ–Ψ distance, respectively, for distances calculated between t and $t - \Delta t$ where $\Delta t = 100$ps. These plots corroborate the results of the distance from initial structure, although the relatively flat lines generated by this analysis may be better suited for obtaining an average distance for some Δt. Similar plots for $\Delta t = 1$ns are shown in figures 2.22, 2.23, and 2.24, which also all agree with the results from above and for $\Delta t = 100$ps.

Backbone Angle Autocorrelation

The autocorrelation function of the backbone angles using the DFT technique described in the Methods section was calculated for all FG-Nup simulations. The average for each FG-Nup was then obtained by averaging the autocorrelation function results for all simulations of the same length. The plot shown in Figure 2.28A shows the average autocorrelation function for the 3ns simulations using a 200ps sliding window, Figure 2.28B shows the average autocorrelation function for the 18ns simulations using

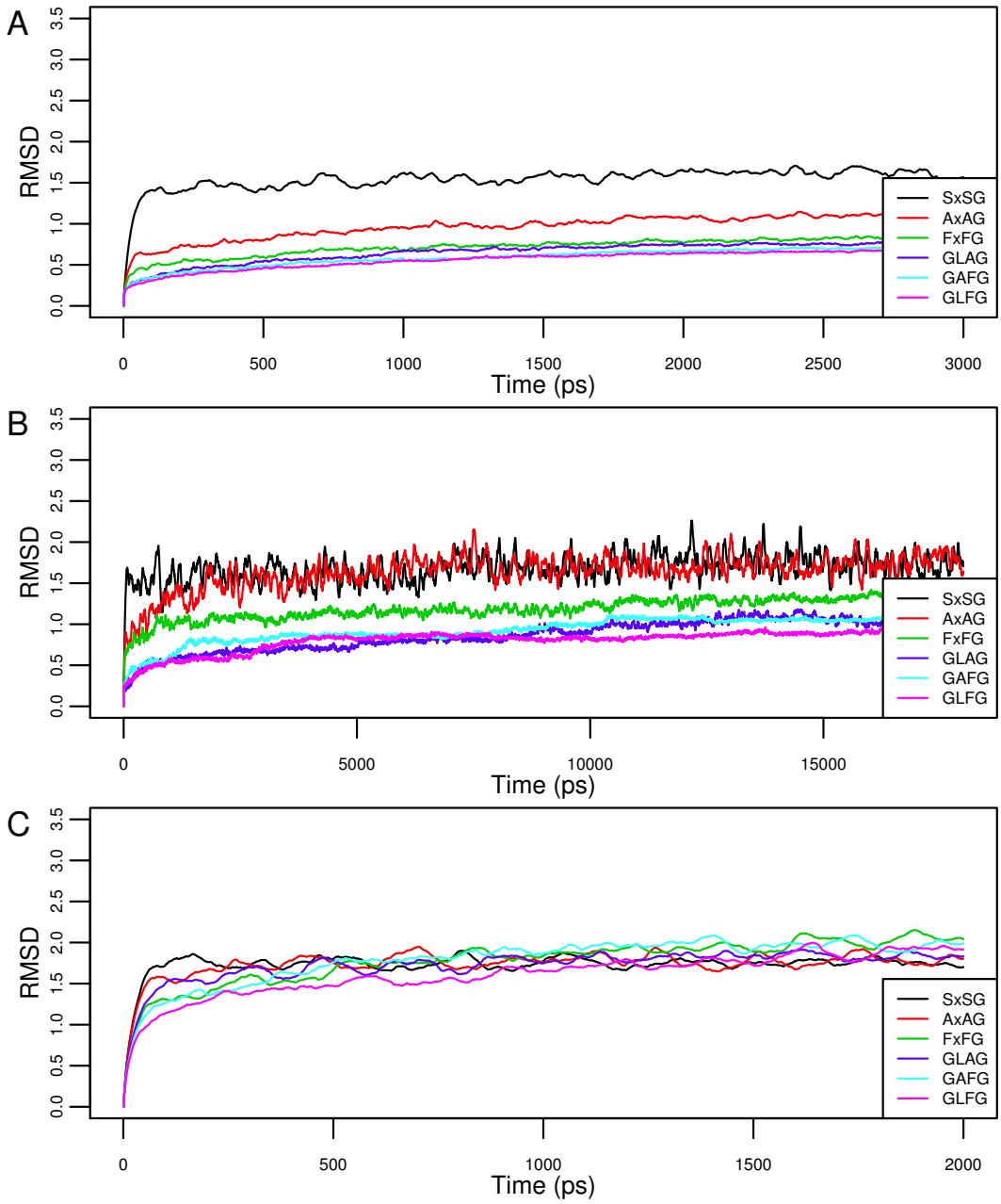

Figure 2.19: RMSD from initial structure as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

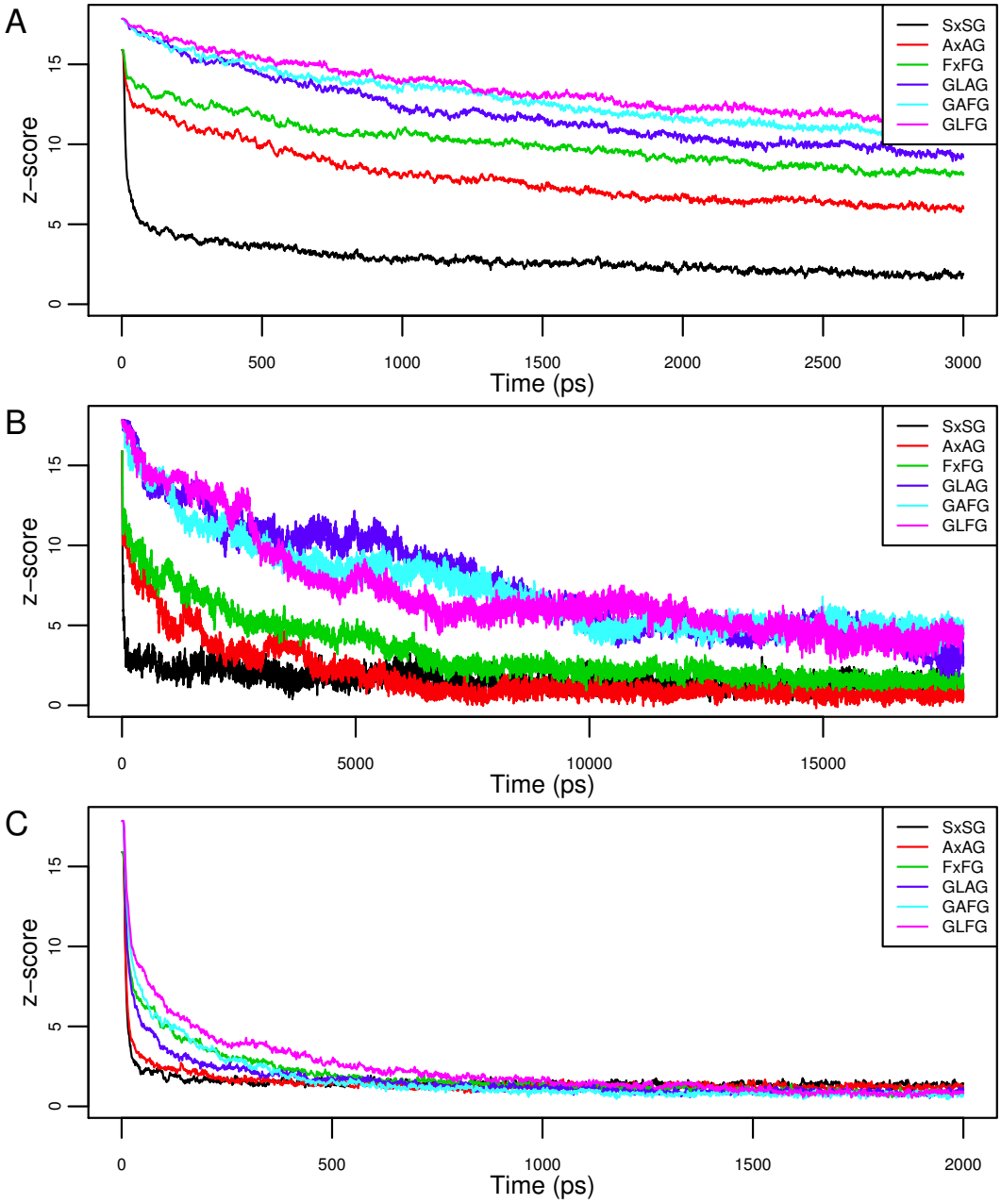

Figure 2.20: MAMMOTH z-score from initial structure as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

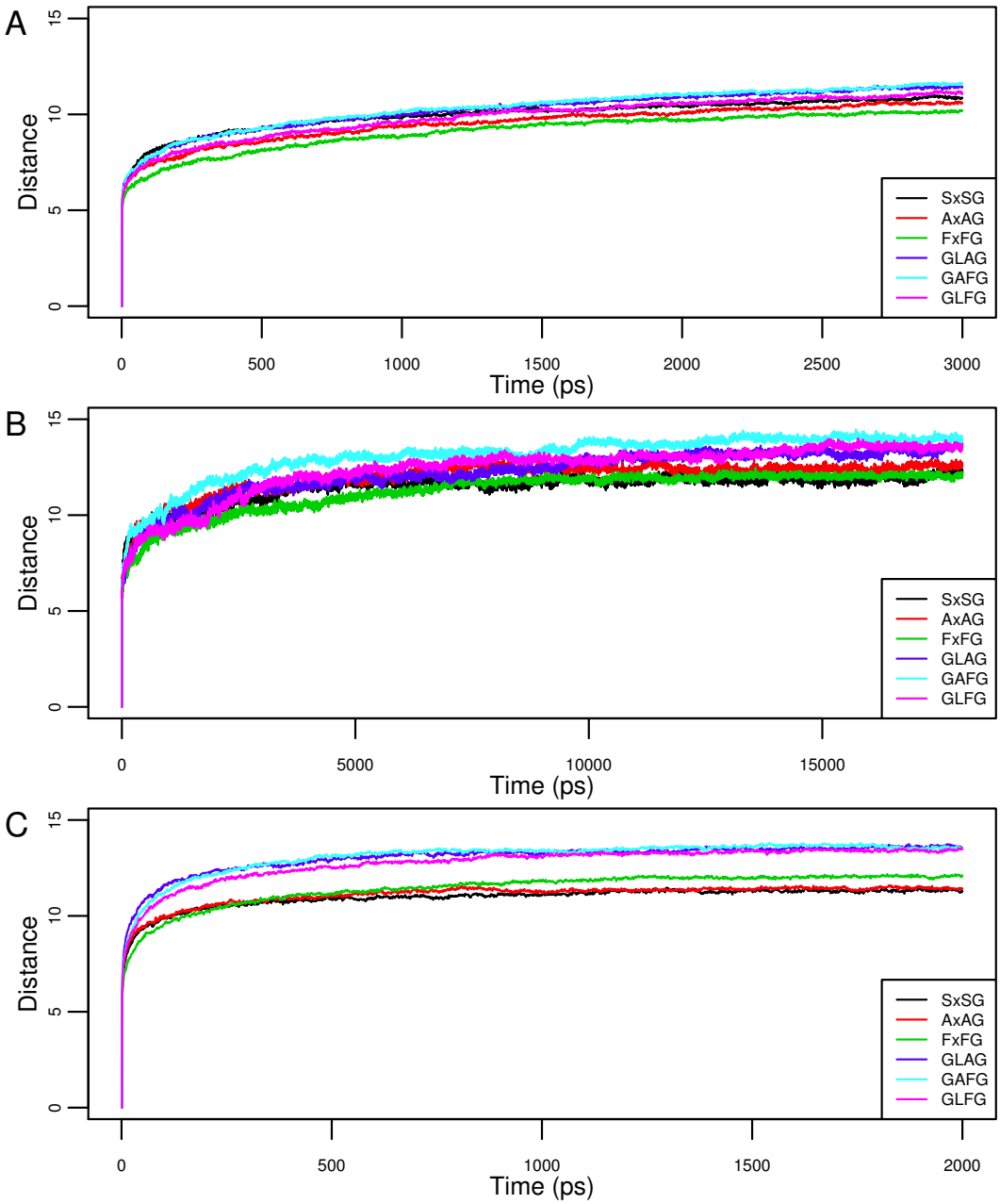

Figure 2.21: Φ–Ψ distance from initial structure as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

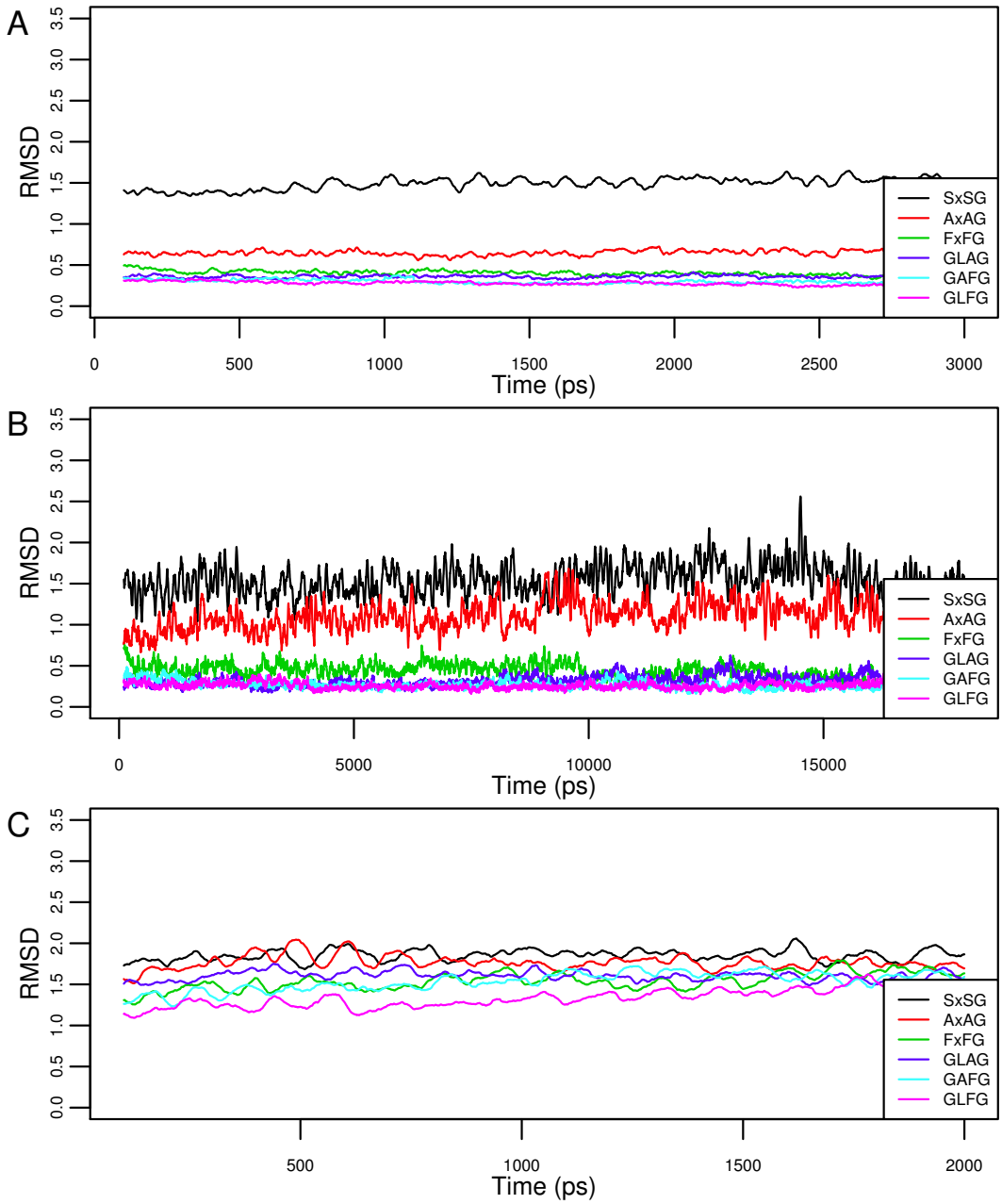

Figure 2.22: RMSD from structure at $\Delta t = 100$ps as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

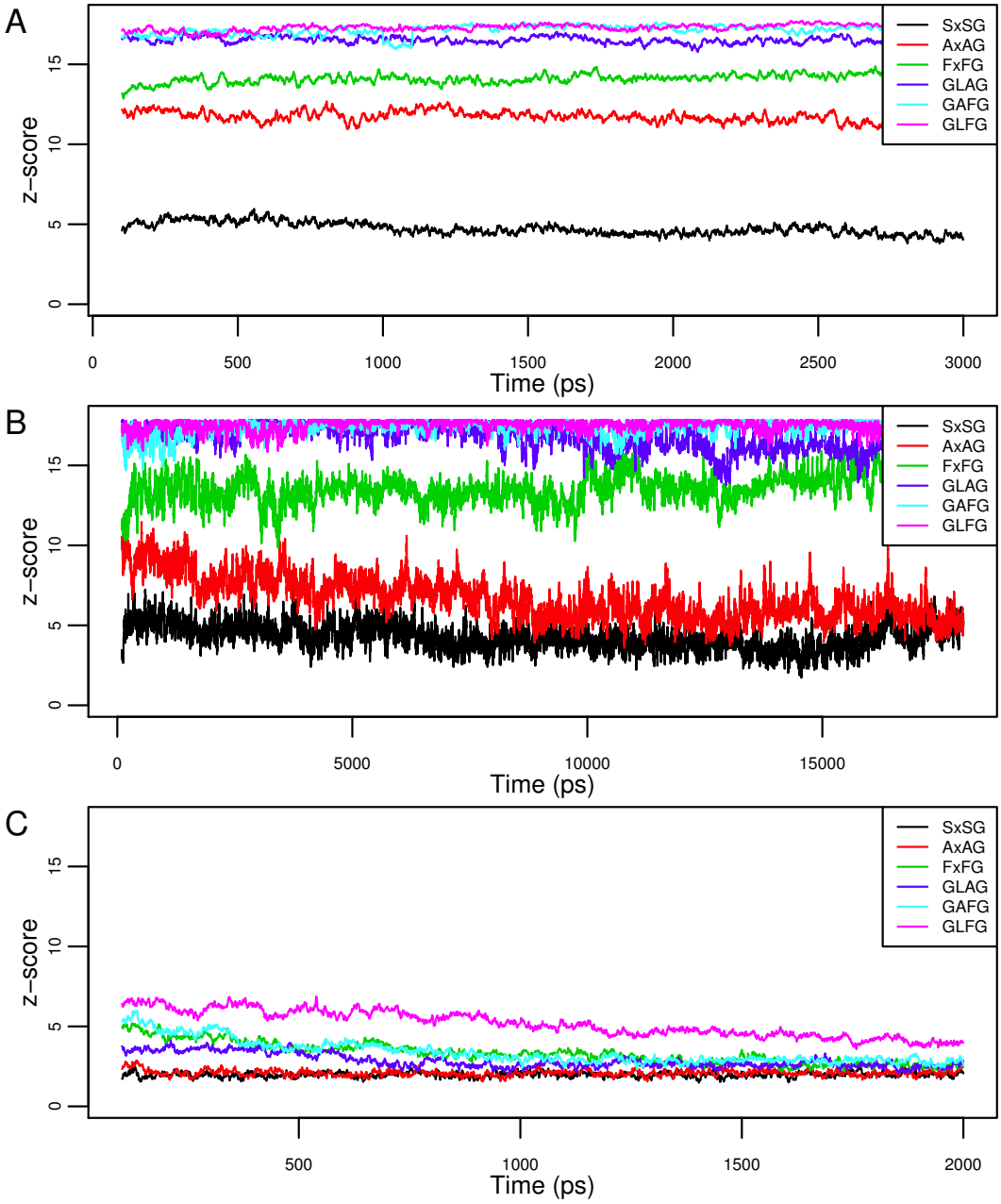

Figure 2.23: MAMMOTH z-score from structure at $\Delta t = 100$ps as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

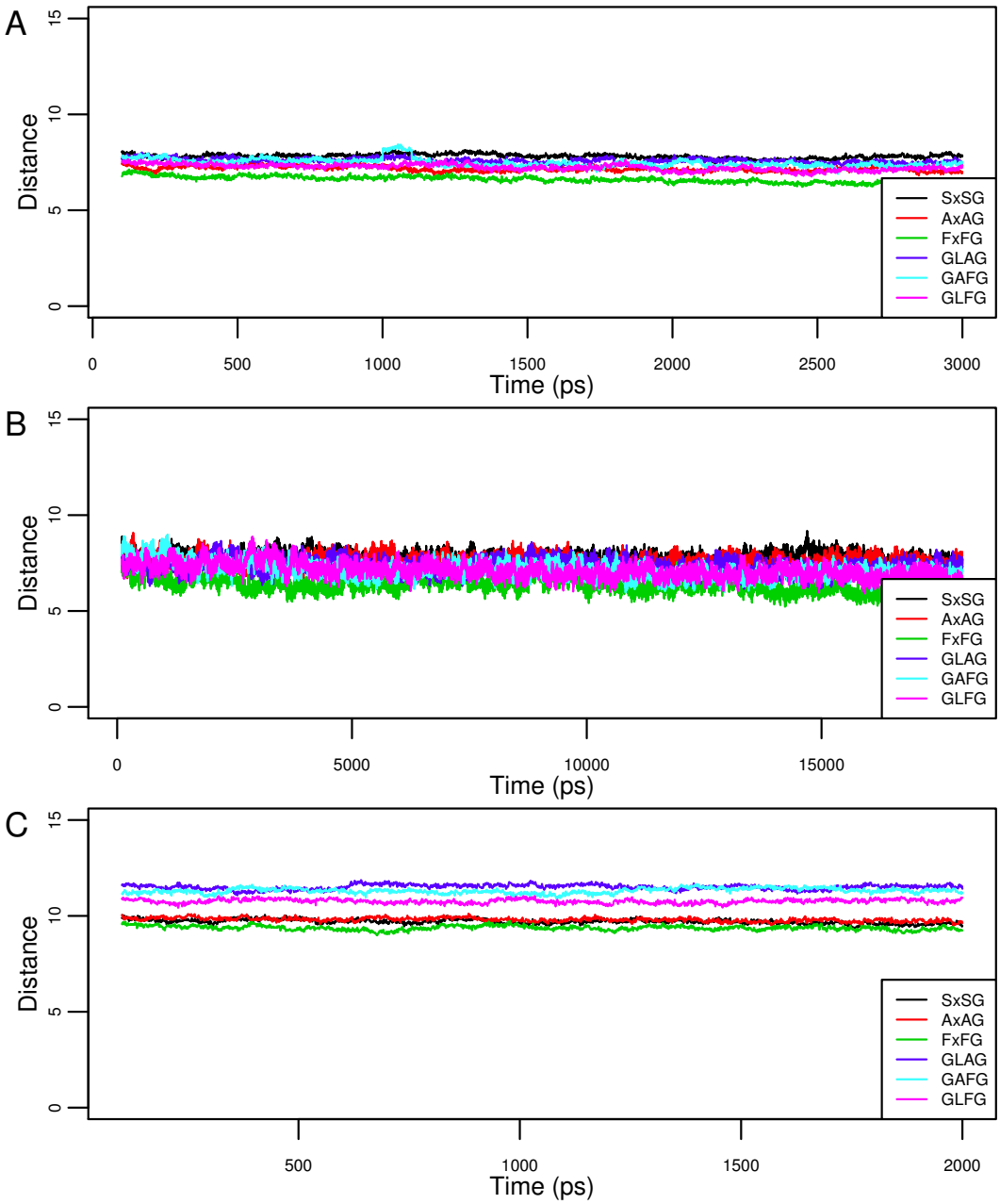

Figure 2.24: Φ–Ψ distance structure at $\Delta t = 100$ps as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

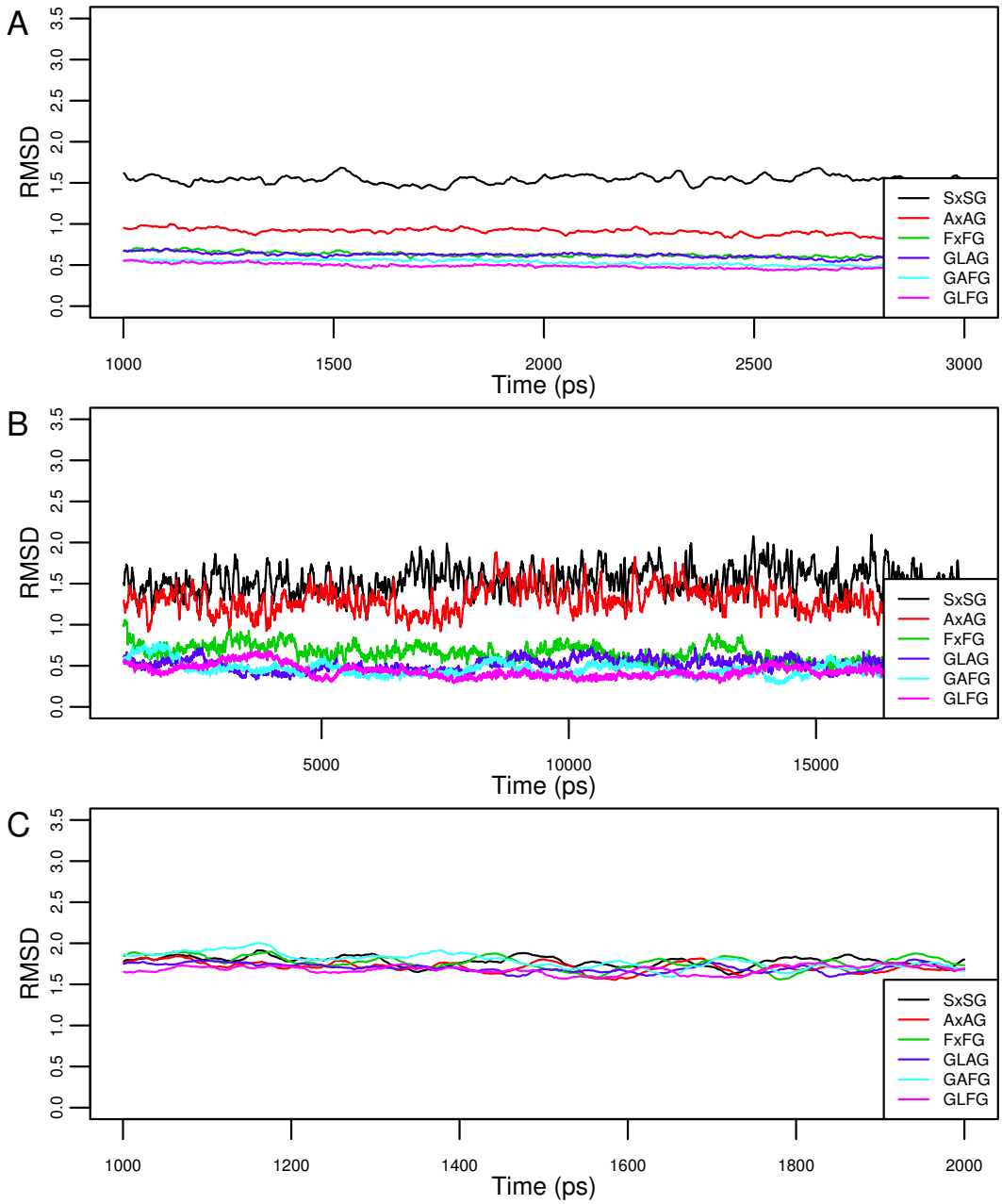

Figure 2.25: RMSD from structure at $\Delta t = 1$ns as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

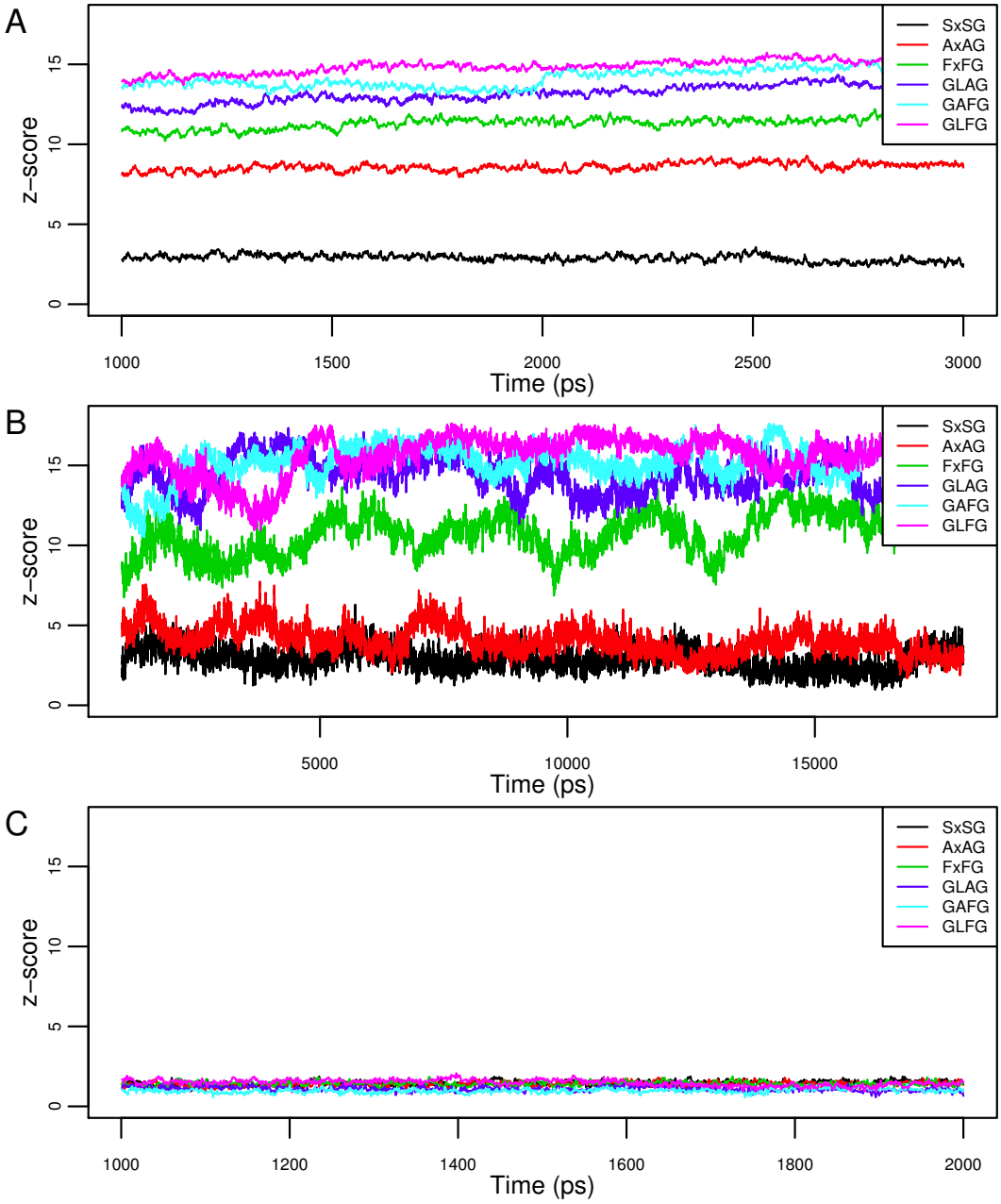

Figure 2.26: MAMMOTH z-score from structure at $\Delta t = 1$ns as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

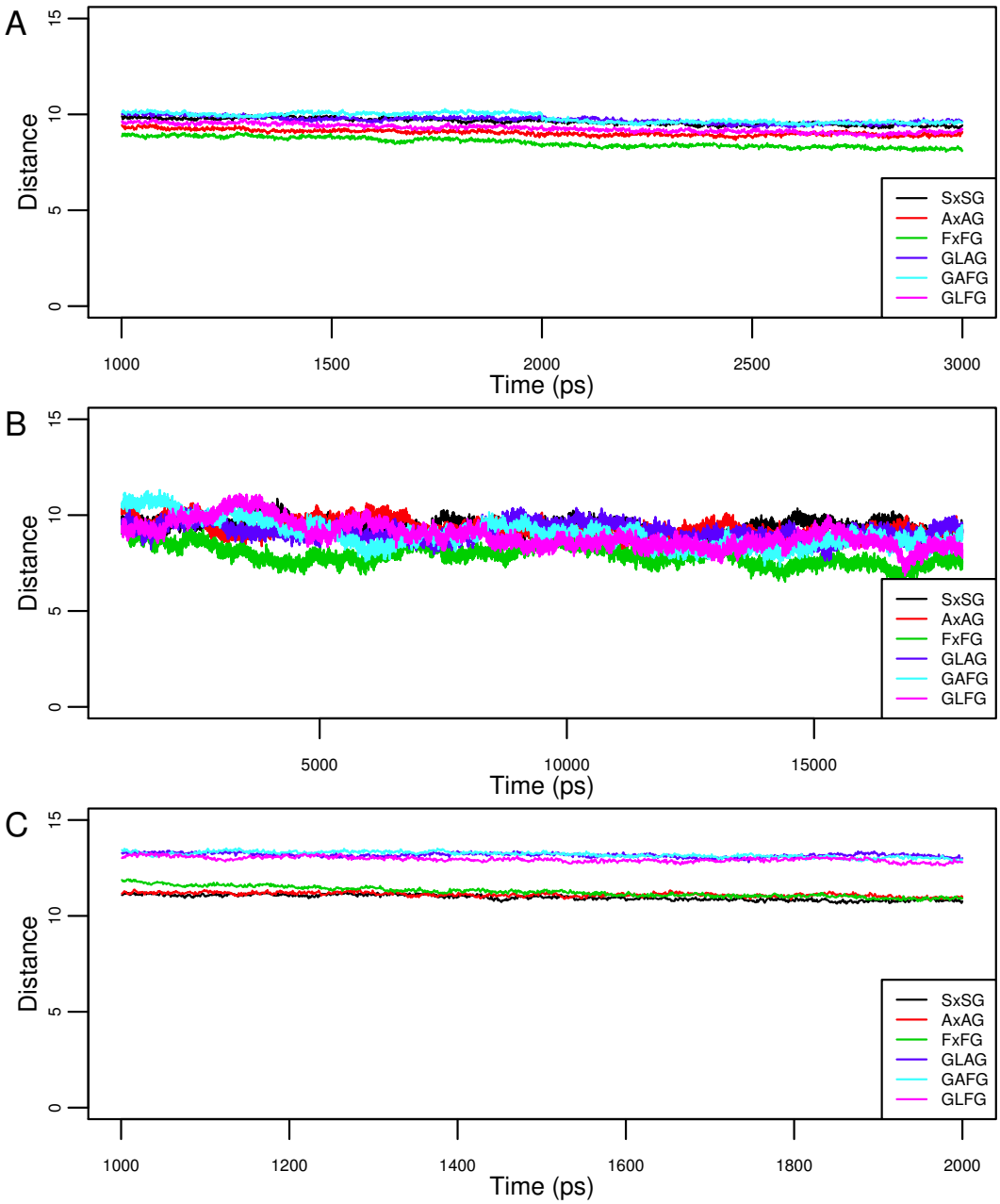

Figure 2.27: Φ–Ψ distance structure at $\Delta t = 1$ns as a function of simulation time averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

a 1ns sliding window, and Figure 2.28C shows the average autocorrelation function for the 2ns simulations using a 100ps sliding window.

The results of this analysis differ between all three simulation conditions, similar to the Φ–Ψ distance plots above. However, the differences are more distinct for the autocorrelation functions. For instance, FxFG has the most slowly decaying function in all three conditions relative to its mutants, indicating the smallest amount of structural change. GLFG shows the same trend among its mutants across the simulation conditions. This is consistent with the previous analyses. However, the overall dynamics of GLFG and its mutants seem to be more greatly affected by the increase in temperature in the 2ns simulations to such an extent that the autocorrelation functions surpass even SxSG. Therefore, these results indicate distinct dynamical properties between FxFG and its mutants versus GLFG and its mutants.

Decorrelation Time

The $\sigma^2_{\text{obs}}(t)$ value for several values of t were calculated with sample sizes of $N = 2$, using 10 independent histograms of 20 bins each for every 3ns simulation. The $\sigma^2_{\text{obs}}(t)$ values at corresponding values of t across all 10 histograms and across all 3ns simulations ($40 \times 10 = 400$ histograms in total) were then averaged. The results are shown in Figure 2.29A. The same protocol was followed for all 18ns simulations and 2ns simulations, and those data are shown in Figures 2.29B and 2.29C, respectively. The decorrelation time, τ_{dec}, is the point where $\sigma^2_{\text{obs}}(t)$ decreases to 1. For the 3ns simulations, this appears to be roughly around 450ps for all of the FG-Nups. The 18ns simulation results indicate that $\tau_{dec} \approx 2000$ps, so there is not much consistency between these results. The 2ns simulation results show a $\tau_{dec} \approx 200$ps.

Additional analysis was performed by running the algorithm for both $N = 4$ and $N = 10$ to see if the increased sample sizes would lead to more consistent estimates of τ_{dec}. The results are shown in Figures 2.30 and 2.31, respectively. In both cases τ_{dec} decreased, and was still inconsistent between the 3ns and 18ns results. For $N = 4$,

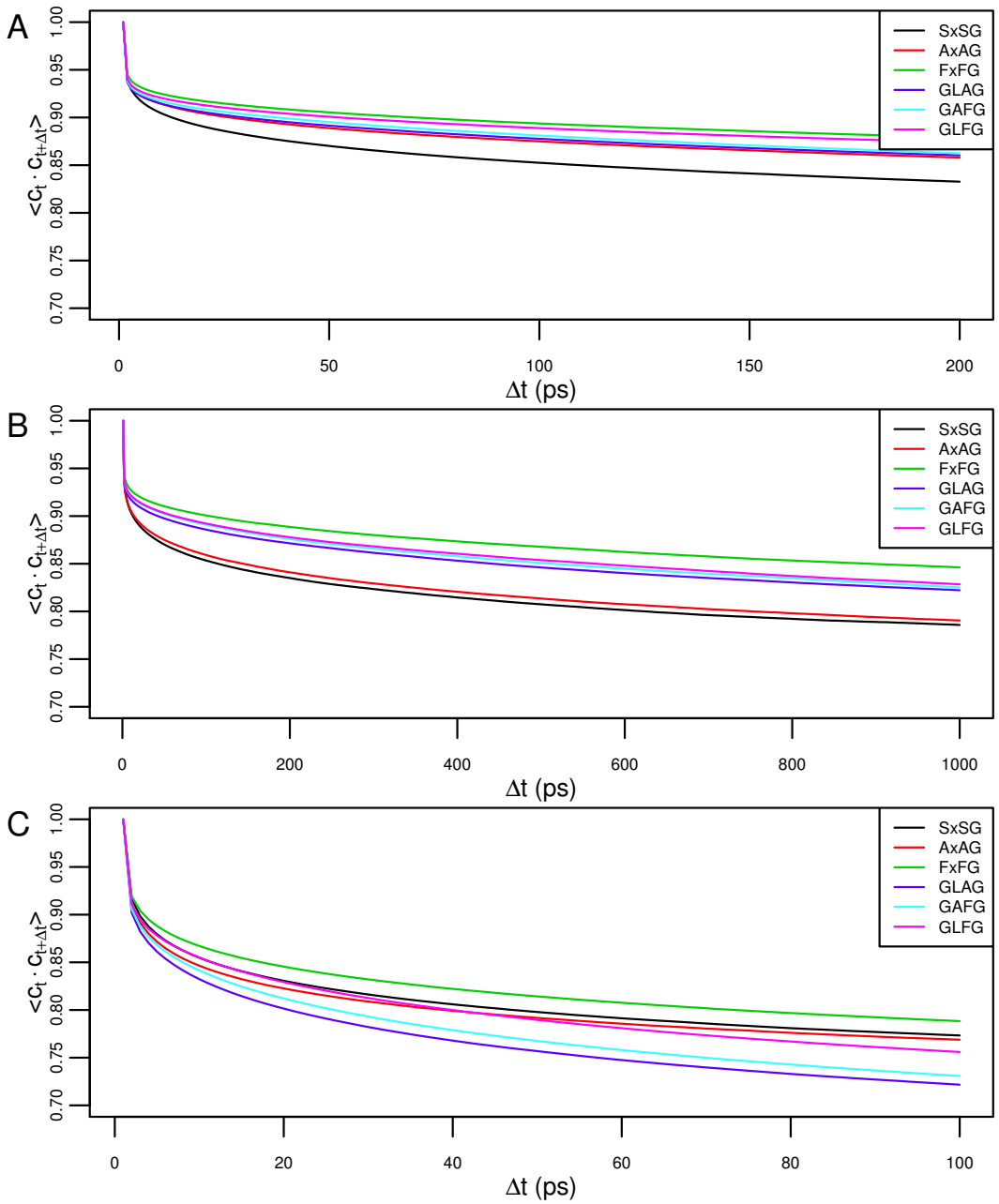

Figure 2.28: Plots of the backbone angle autocorrelation function averaged across all (A) 3ns @ 300K, (B) 18ns @ 300K, and (C) 2ns @ 350K simulations.

$\tau_{dec} \approx 350$ps for the 3ns simulations, $\tau_{dec} \approx 1500$ps for the 18ns simulations, and $\tau_{dec} \approx 150$ps for the 2ns simulations. For $N = 10$, $\tau_{dec} \approx 180$ps for the 3ns simulations, $\tau_{dec} \approx 800$ps for the 18ns simulations, $\tau_{dec} \approx 75$ps for the 2ns. In general, the method predicts that the simulations are well-converged at surprisingly short time scales, but τ_{dec} seems far too sensitive to both simulation length and sample size to be a reliable estimator.

It is interesting to note that while the estimates for τ_{dec} were inconsistent, there was one pattern observed across all of the conditions examined: the rate of convergence for each FG-Nup correlated well with earlier measures of protein dynamics. While this pattern is observed in all of the decorrelation time results, it is perhaps clearest in Figure 2.31A, where $\sigma^2_{\text{obs}}(t)$ converges from slowest to fastest in the following order: GLFG, GAFG, GLAG, FxFG, AxAG, and SxSG. Therefore, the decorrelation time algorithm predicts that GLFG explores its accessible conformation space more slowly than any of the other FG-Nups and SxSG explores its a accessible conformation space more quickly than any of the other FG-Nups. The 2ns simulation results show that this pattern changes slightly at higher temperatures. FxFG and GLAG switch places in the ordering, and SxSG and AxAG are virtually indistinguishable. Both patterns are consistent with the prior R_g, S, distance map, and distance versus time results presented above, as well as earlier experimental and theoretical results [21, 8].

2.4 Discussion

The analysis of a set of six FG-Nups indicates that these proteins span a wide range of structural disorder. The R_g and S data show that the structures span from prolate, extended conformations to collapsed, almost spherical conformations, with various structures in-between. This data is congruent with past experimental studies of FG-Nup compaction [21, 8]. The main structural feature of compaction was the fold over points found using the inter-residue distance map analysis. However, there was also a difference in secondary structure that correlated well with the R_g data. Namely, more

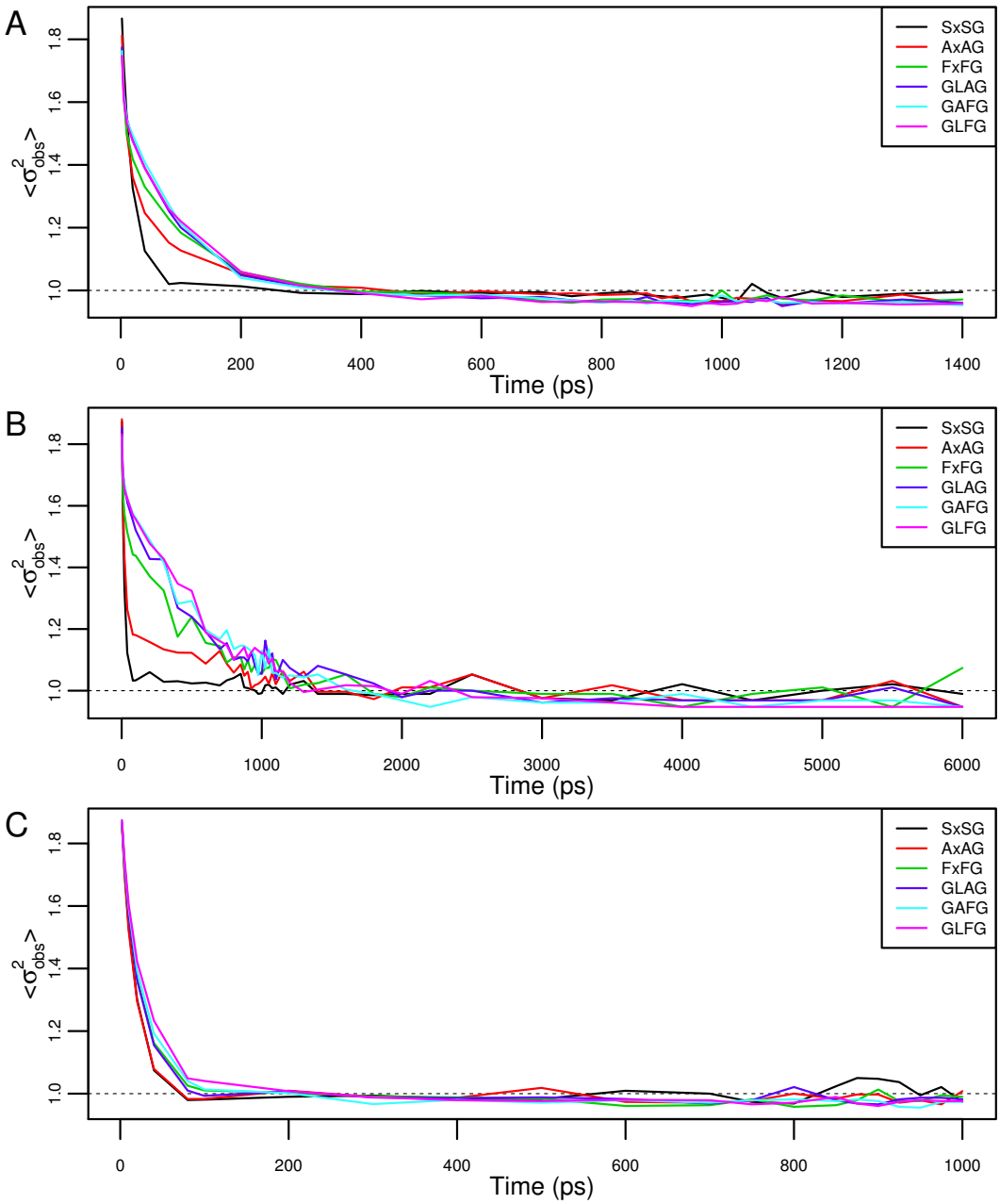

Figure 2.29: Decorrelation Time – $N = 2$ – (A) 3ns @ 300K (B) 18ns @ 300K (C) 2ns @ 350K

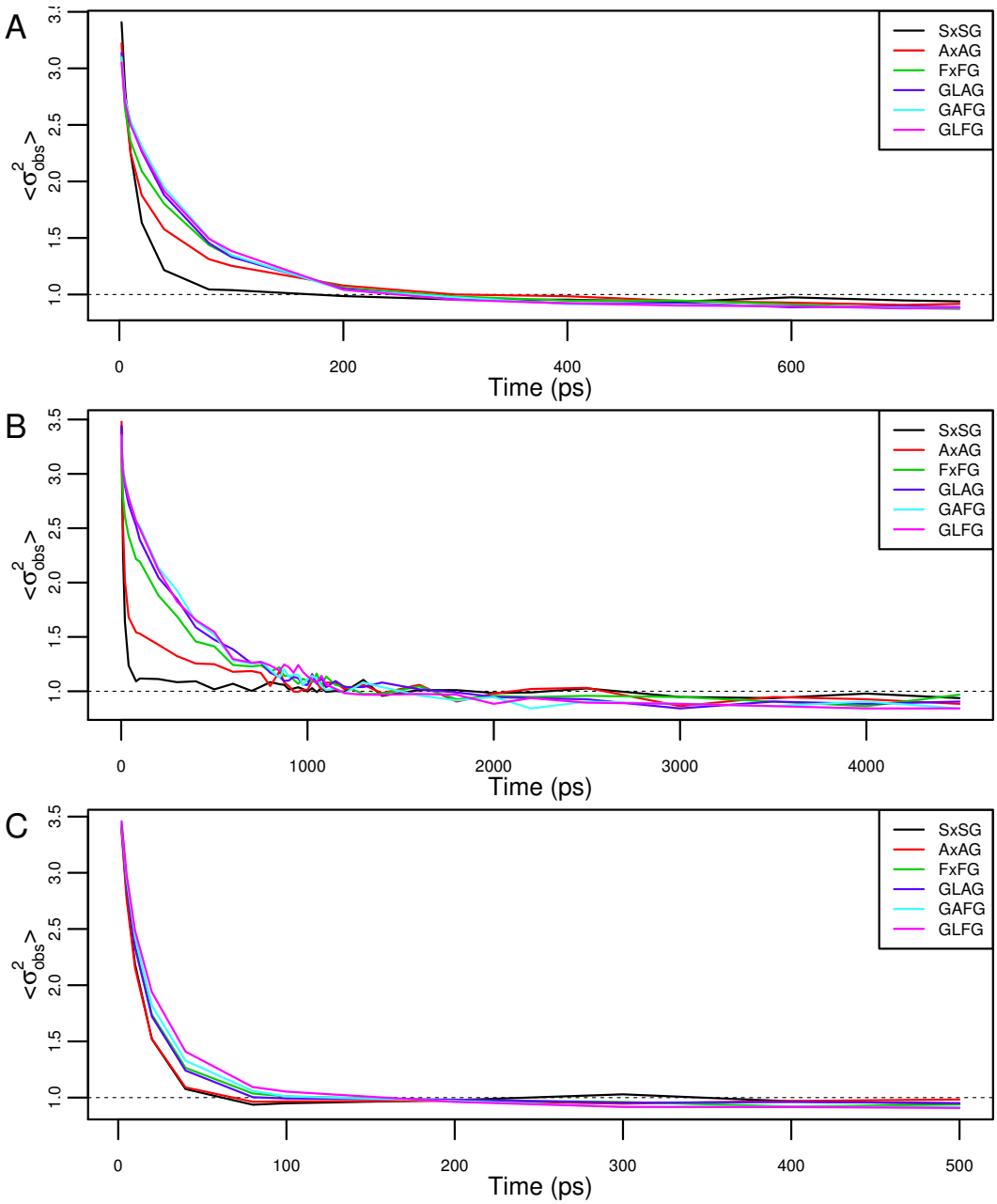

Figure 2.30: Decorrelation Time – $N = 4$ – (A) 3ns @ 300K (B) 18ns @ 300K (C) 2ns @ 350K

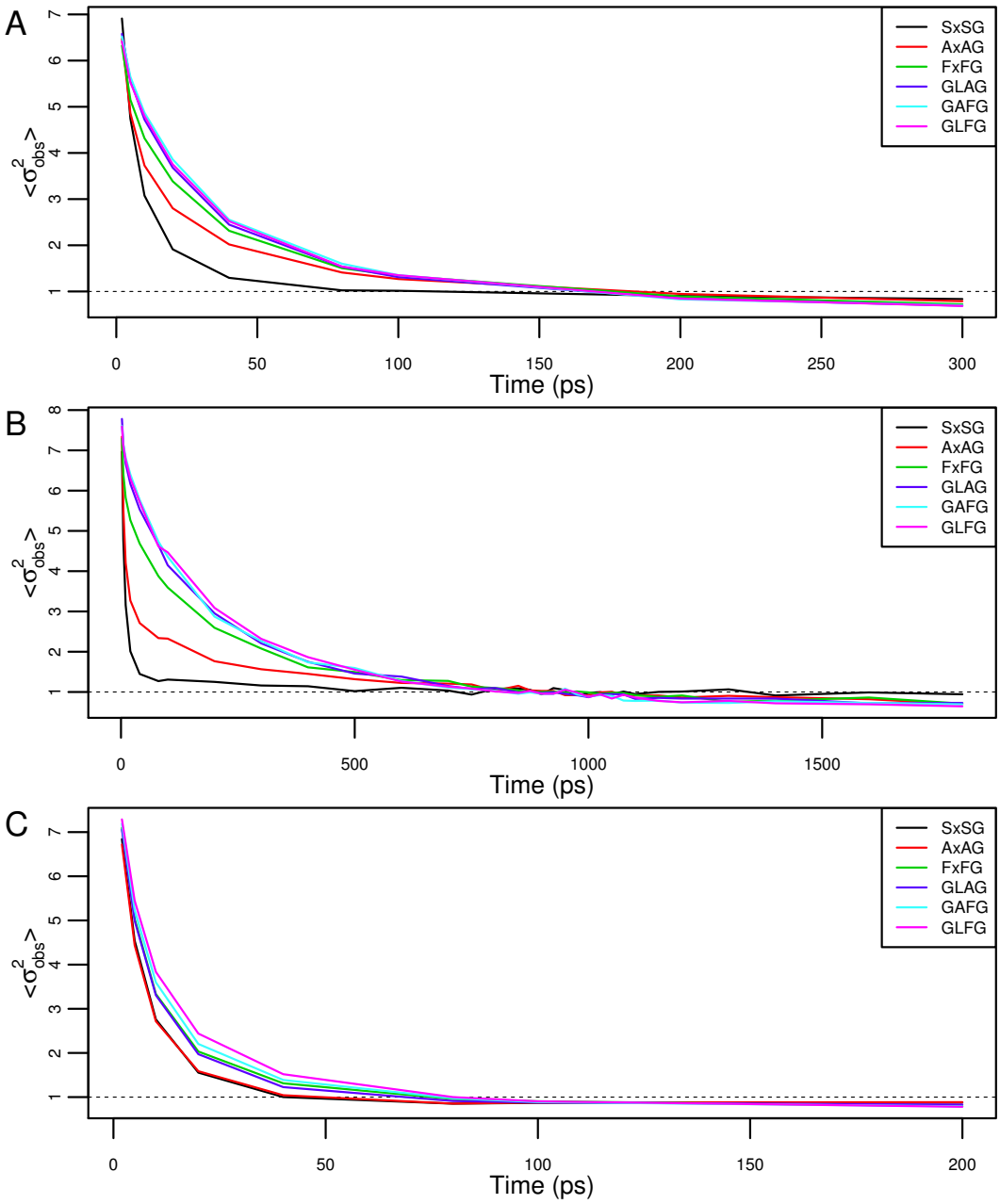

Figure 2.31: Decorrelation Time – $N = 10$ – (A) 3ns @ 300K (B) 18ns @ 300K (C) 2ns @ 350K

extended structures exhibited more α-helical content, and the more collapsed structures, although still mainly helical, showed more propensity for β-content as well. In addition, dynamical features such as the RMSD and MAMMOTH distances versus time corroborated these results. While the Φ–Ψ angle based distance analysis was inconclusive, the autocorrelation functions indicated that the GLFG and FxFG wildtype FG-Nups were less dynamical than corresponding mutants, and also that the GxxG-type FG-Nups were more heavily affected by increases in temperature. Also, while the decorrelation time algorithm failed to reliably determine the convergence properties of the simulations, the results were still consistent with prior data on these proteins. Overall, these results indicate that computational approaches to comparing and contrasting protein dynamics can reliably classify the structural dynamics of both compact and extended IDPs.

Chapter 3

A Clustering Approach for Estimating the Convergence of Protein Simulations

The previous chapter introduced several common techniques for quantifying protein dynamics using both static and dynamic features extracted from simulation trajectories, but only briefly touched on the concept of simulation convergence. In particular, the decorrelation time measure was implemented to aid in determining if and when the conformation space of a protein has been adequately explored by a simulation. However, this technique was unable to determine any significant difference between the convergence rates of the simulations across the proteins studied even though many of the other metrics explored indicated that such differences might exist.

In this chapter, a framework for assessing the convergence properties of protein simulations based on clustering methods and basic information theory is developed [22]. The approach differs from standard applications of clustering MD trajectories in that it does not try to ascertain what conformations are similar within a single replicate simulation of a particular protein. Neither does it attempt to find the differences or similarities in structure between the different proteins. Instead, clustering is used 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 estimation of the

amount of unique structural space being sampled in the simulations, and is particularly relevant to non-equilibrium simulations such as disordered or intrinsically disordered proteins where the conformation space might be largely unexplored even after many other simulation properties have converged. The approach used in this chapter is related to the structural histogram approach of Lyman and Zuckerman [19] to determine when a set of MD simulations is at equilibrium, but differs in that data is clustered from multiple replicates in order to understand the trade-offs between running many, shorter simulations and running fewer, longer simulations. Specifically, *the clustering results are used to determine to what extent replicate MD simulations of a single protein sample independent regions of structural phase space in order to assess the overall convergence of the simulations.*

3.1 Background

Clustering has been widely used to analyze MD simulations of biopolymers, particularly for determining the conformational states of the trajectories. Karpen, et al. made use of a self-organizing neural network to cluster structures based on backbone and side-chain dihedral angles of a small pentapeptide [23]. Best and Hege analyzed simulations of a small tri-ribonucleotide by bi-partitioning the similarity graph defined by the vector of intramolecular distances [24]. Lei et al. used hierarchical clustering based on structural root-mean-squared distance (RMSD) to study folding via replica exchange MD simulation of the villin headpiece subdomain [25]. The same system was studied in a similar manner by Freddolino et al. using MD simulations on the microsecond timescale [26]. While data clustering has been used to perform a variety of analyses on MD simulations of proteins and other polypeptide structures, it has only recently been applied to the study of intrinsically disordered proteins [27]. This list of approaches is by no means exhaustive, and simply serves to illustrate the importance of clustering in simulation analysis as well as the great variation in algorithms utilized across MD studies.

Clustering has also been used to study the convergence properties of simulation trajectories. Lyman and Zuckerman [19] 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. [28] 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.

The study described in this chapter in the first to our knowledge to apply *spectral clustering* to IDP simulations. Spectral clustering consists of three general steps. First, the dissimilarities between all pairs of structures in an ensemble are computed. Root-mean-squared distance (RMSD) is used for computing dissimilarities for all results presented in this study. Second, the matrix of pairwise similarities (obtained directly from the dissimilarities) is normalized and its spectral decomposition is computed to obtain the top k eigenvectors. Third, standard k-means clustering is applied to the (normalized) points described by the top k eigenvectors. The optimal number of clusters is unknown beforehand for most interesting phenomena, so one must examine the results for a range of numbers.

Spectral clustering possesses several attributes that make it particularly well-suited for clustering polymer simulations. First, it shares a formal relationship with Markov-chain models where the dynamics are viewed as a random walk on a structure-transition graph (or matrix) [29] which is also frequently expressed as random diffusion on a free energy surface [30, 31]. Specifically, spectral clustering operates on the Lapla-

cian of the graph of pairwise structural similarities which is analogous to the transition matrix in the Markov-chain model. If the sampling of the simulation is sufficient, this matrix defines a random walk on the free energy surface. Second, once the eigen decomposition step is complete, repartitioning the ensemble into different numbers of clusters, k, is fast, allowing the data to be easily examined at various levels of granularity. Third, since the dissimilarity between all pairs of structures is calculated, disordered systems which lack reference structures can be studied without introducing an unfavorable bias due to the selection of a single reference structure (for the ensemble as a whole or for each cluster), as must be done in most other clustering techniques. Finally, spectral clustering is more informative of the local density of structures than other clustering techniques. A byproduct of the algorithm is a similarity scaling parameter σ. This parameter is computed for each structure and characterizes the local density. Low values of σ indicate that a structure resides in a densely populated region of structural space while high values indicate the region is relatively sparse. When averaged over all structures belonging to a cluster, the similarity scaling parameter can be used to characterize the cluster as corresponding to a metastable or transition state.

3.2 Methods

3.2.1 Spectral Clustering

Spectral clustering is a powerful methodology for partitioning data. Application of this method results in a set of clusters, each of which contains a subset of the data that is considered to show strong intra-cluster similarity and weak inter-cluster similarity according to some metric (ex. Euclidean distance). The name "spectral" refers to the use of eigen decomposition to compute the eigenvectors of the Laplacian matrix obtained from an adjacency matrix (graph) representation of the data. The resulting top few eigenvectors describe a nonlinear projection of the data onto a low dimensional manifold. Applying a standard clustering algorithm to the projected data typically results

in a more intuitive and useful partitioning, compared to applying a standard clustering algorithm in the original data space [32, 33, 34].

Clustering algorithms often have to be adapted to deal with the structure-comparison methods used in MD simulation, such as root-mean-squared distance (RMSD) or MAMMOTH [14], and often these modifications are not trivial [28]. Projected data does not suffer from this drawback since any clustering algorithm which operates on real data vectors can be used.

Research into spectral methods has resulted in a broad number of ways to define the adjacency matrix and its respective Laplacian matrix [34]. A wide range of standard clustering algorithms exist for processing the projected data as well. The methodology outlined in [33] was chosen, which in turn is based on the algorithm in [32], with one modification outlined below. This methodology presents several advantages over other approaches:

- The projection step requires a single free parameter for defining a fully-connected adjacency matrix, which can also be made sparse for most data sets.
- A normalized Laplacian matrix is used so that the resulting projection is a relaxed solution to the normalized cut problem from graph theory.
- The k-means clustering algorithm, a well-understood and commonly used clustering algorithm, is used for processing the projected data.

The method proceeds as follows:

1. Consider P to be the set of n polymer or protein structures that are subject to clustering.
2. Construct the dissimilarity matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ where $x_{ij} = RMSD(P_i, P_j)$.
3. Construct the sorted distance matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ by sorting each row of $\mathbf{X}$ in ascending order.

4. Construct the scaling parameter vector $\sigma \in \mathbb{R}^n$ where $\sigma_i = \frac{1}{q}\sum_{j=2}^{q+1} s_{ij}$ and $q \in \mathbb{Z}, 0 < q < n$.

5. Construct the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ where $a_{ij} = exp(-x_{i,j}^2/2\sigma_i\sigma_j)$ for $i \neq j, \mathbf{A}_{ii} = 0$.

6. Construct the normalized graph Laplacian $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j a_{ij}$.

7. Compute the eigen decomposition of $\mathbf{L} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$

8. Construct the projected data matrix $\mathbf{Y} \in \mathbb{R}^{n \times k}$ by stacking the k eigenvectors associated with the k largest eigenvalues by column and normalize each of the rows to unit length.

9. Apply k-means clustering to the row vectors in $\mathbf{Y}$.

The above approach differs from the approach of Zelnik-Manor and Perona [33] in step 4. While they use $\sigma_i = s_{i(q+1)}$ (the distance between the qth closest structure to structure i and structure i itself), here instead $\sigma_i = \frac{1}{q}\sum_{j=2}^{q+1} s_{ij}$ (the average distance from structure i to the q closest structures to structure i). This modification makes the algorithm more robust to the choice of q which is especially important for exploratory data analysis. A value of $q = 10$ was chosen for all analyses presented here, as this value produced results similar to earlier work where a single σ value for all structures was determined by manual search and found to be rather insensitive for the simulation data presented here [22].

It is also worth noting that step 2 is not limited to any particular pairwise distance function for computing dissimilarity. RMSD is used here because of its ubiquitous application in MD simulation studies. However, any dissimilarity function could be chosen, and may vary depending on the particular application. The systems studied here display large amplitude motions, and RMSD has been criticized in the past for performing poorly when comparing very dissimilar structures. In essence, two structures

that are very different from one another might both appear relatively similar to a third, not necessarily intermediate, structure. This limitation does not prove to be a problem in the context of graph-based clustering methods, such as spectral clustering. The Gaussian kernel in step 5, combined with the locally-scaled parameters from step 4, allows the algorithm to focus on the local, valid structural comparisons and ignore the more distant, less discriminative comparisons. This kernel function is essentially a *soft* version of the *hard* RMSD cutoff used in many other clustering methods, but it can be locally adapted to the data at hand via the scaling parameters, σ. The potential sparsity that can be induced upon the matrices by applying a cutoff for the small Gaussian kernal affinity values also affords the use of fast sparse linear algebra routines, greatly reducing the computational demands of the algorithm.

In step 9, k-means clustering is utilized to perform the final partitioning in the projected data space. The reader is directed to the seminal paper by MacQueen for the details of the algorithm [35]. The k-means clustering algorithm requires specification of several parameters:

- The number of clusters, k.
- The number of times to run the algorithm with different initial positions for the k cluster centroids.
- The maximum number of iterations for the algorithm.

The last two of these must be chosen so that there is a reasonable expectation that the optimal solution is obtained. A random selection of k points from the row vectors of $\mathbf{Y}$ is used to initialize the algorithm. This is done ten times and the result with the smallest sum of the inter-cluster centroid-point distances is considered: $\sum_i^k \sum_{\mathbf{y}_j \in C_i} \|\mathbf{y}_j - \mu_i\|^2$, where C_i is the set of points partitioned into the ith cluster and μ_i is the mean, or centroid, of the points in C_i. The choice of ten restarts is a conservative number of iterations given that the dimensionality of the projected space is equal to k, which, in this work, is always at least two orders of magnitude smaller that

the number of points. However, it is impossible to prove that the algorithm has indeed found the optimal partitioning, which is a recognized short-coming of many clustering approaches. The algorithm is run for a maximum of thirty iterations or until the partitioning does not change between the last and most recent iterations. This final parameter is a practical way to avoid the rare occurrence of infinite oscillations, but the algorithm always terminated prior to thirty iterations for all analyses presented here.

3.2.2 Direct Application of K-means Clustering

Previous work has focused on applying clustering techniques directly to MD trajectories without first applying spectral decomposition. K-means is one of the algorithms studied by Shao et al. [28] so in addition to the spectral clustering approach outlined above, k-means was applied directly to the 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 in order to save overall computational time by only needed to run the algorithm once and also in order to limit the number of free parameters that the user must specify when using their software; however, to have confidence in the results, one 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 the relationships between the atomic coordinate such that simply averaging dissimilar structures results in gross violations to these constraints. 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 the coordinate positions.

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 fast in practice (although not as fast as a direct application of k-means) 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, simply Euclidean distance is used.

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.

3.2.3 Molecular Dynamics Simulations

For all of the work in this chapter, the simulation data set from Section 2.2.1 is utilized. This data set consists of a set of intrinsically disordered protein fragments simulated at short and long time-scales that have been shown to exhibit different dynamical properties. Please see Section 2.2.1 for the details on the proteins simulated, protocols employed, and general motivation for this approach.

3.3 Results

3.3.1 Spectral Clustering of FG-Nups

In order to assess the diversity of structures explored by each FG-Nup, each of the five FG-Nups was clustered separately. To make the trajectory data tractable for clustering, every tenth frame was sampled from each replicate and the forty subsampled trajectories were concatenated 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. Spectral clustering was then applied to this composite set of structures. The number of clusters was specified to be $k = 40$ (one cluster per replicate) for each protein with the hope of understanding how much the replicates overlap in conformation space, or if they are disjoint. Likewise, every tenth frame was sampled from each of the 18ns replicates and concatenated these five trajectories into one set of 9000 structures for each FG-Nup. These trajectories were then clustered with $k = 5$.

Graphs showing cluster membership for the five proteins are shown in Figures 3.1 and 3.3.[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.

[1] The algorithms do not naturally order the clusters so cleanly as is displayed in these graphs. Rather, each cluster is relabeled 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.

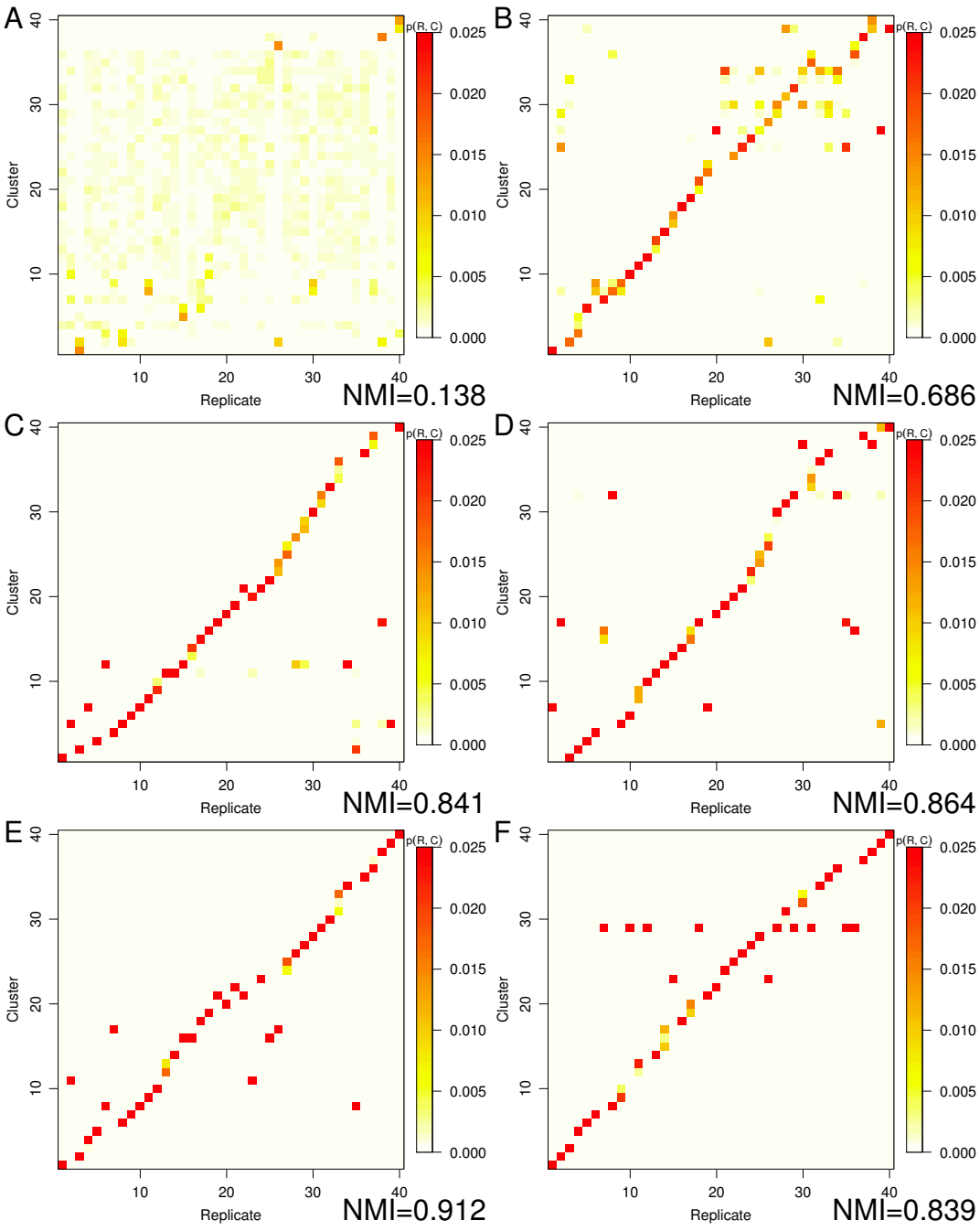

Figure 3.1: Results from applying *spectral clustering* to the *3ns* trajectory data.

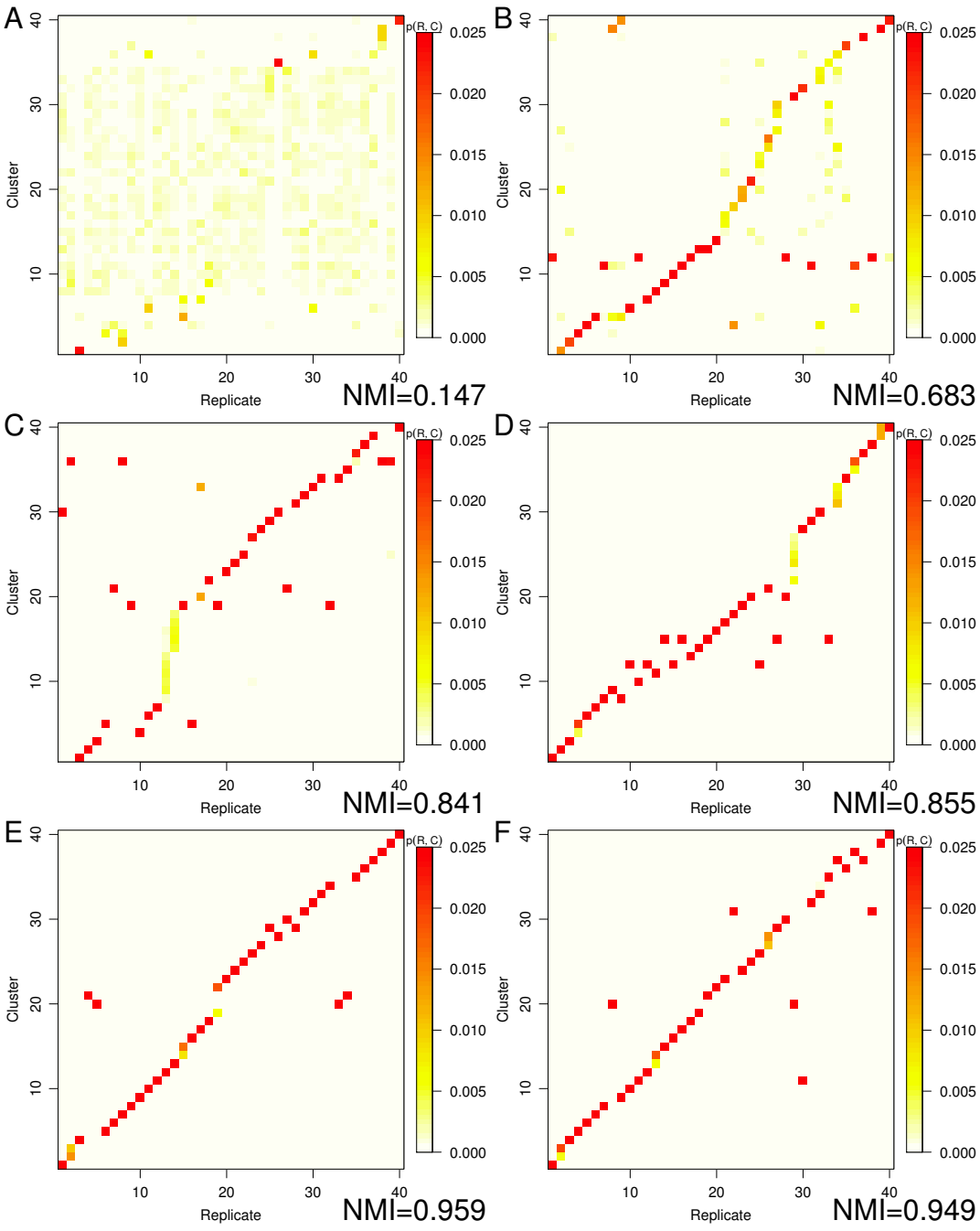

Figure 3.2: Results from applying *k-means clustering* to the *3ns* trajectory data.

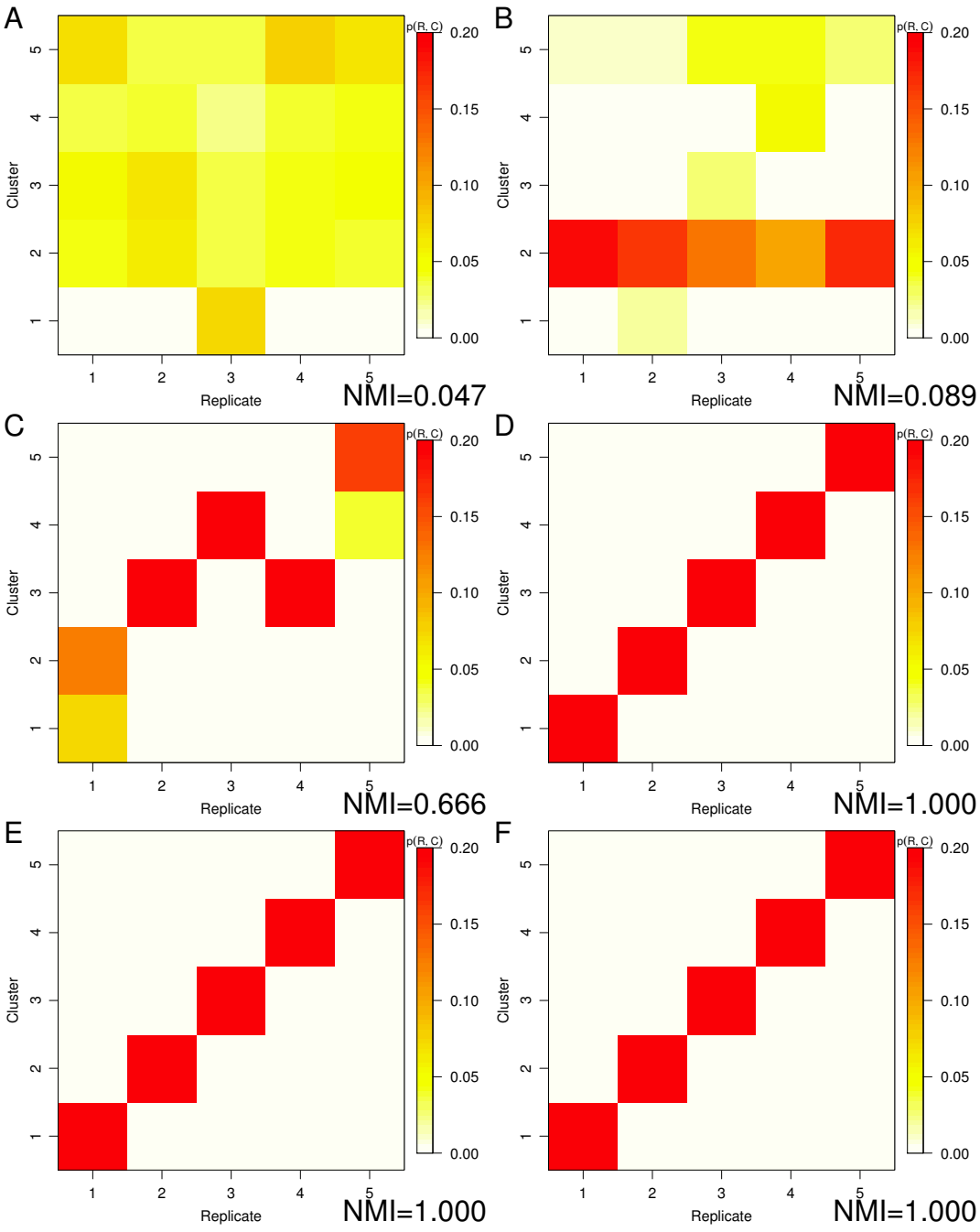

Figure 3.3: Results from applying *spectral clustering* to the *18ns* trajectory data.

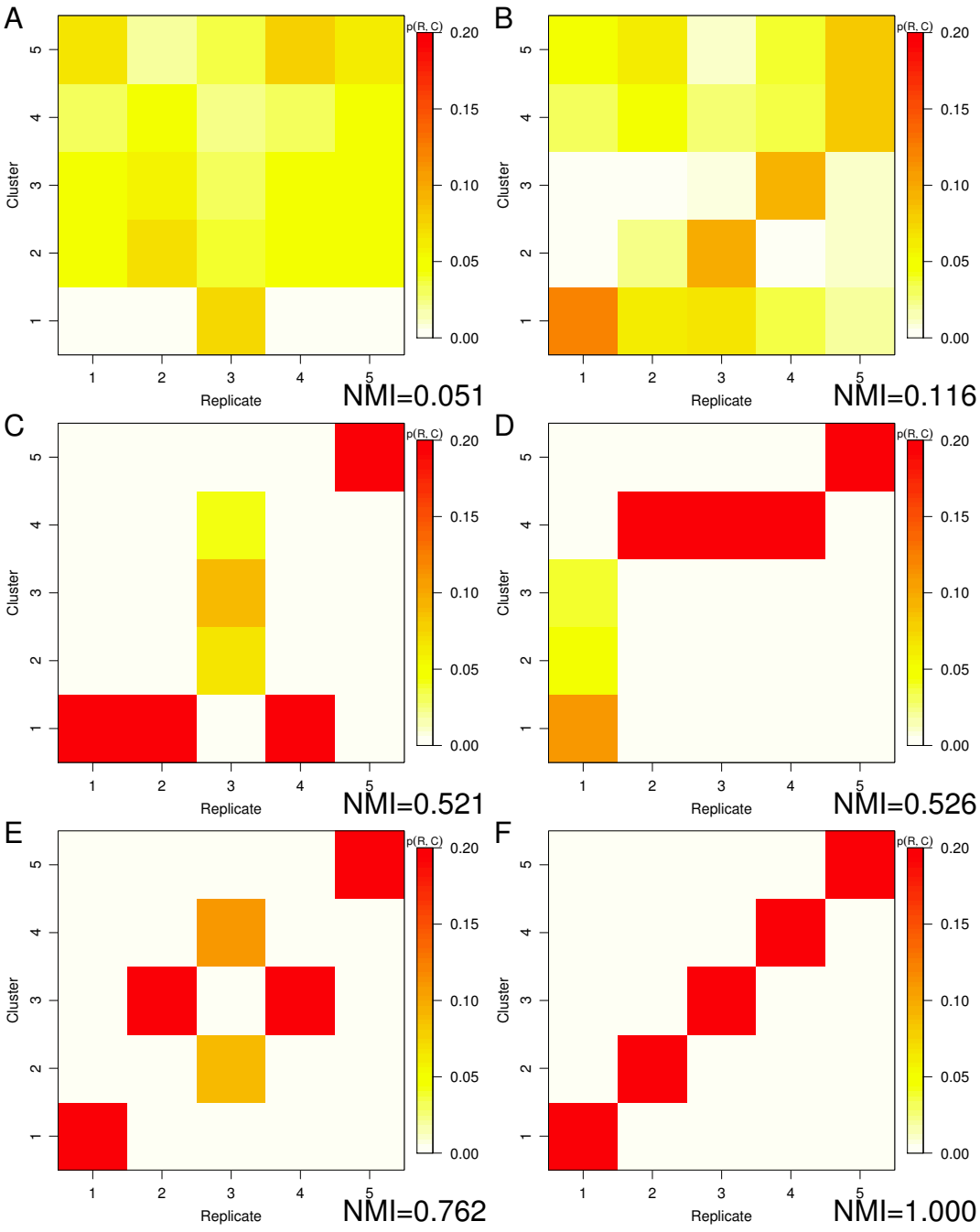

Figure 3.4: Results from applying *k-means clustering* to the *18ns* trajectory data.

Table 3.1: FG-Nup clustering results, including 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	Spectral 3ns	Spectral 18ns	K-means 3ns	K-means 18ns
SxSG	0.138	0.047	0.147	0.051
AxAG	0.686	0.089	0.683	0.116
FxFG	0.841	0.666	0.841	0.521
GALG	0.864	1.000	0.855	0.526
GAFG	0.912	1.000	0.959	0.762
GLFG	0.839	1.000	0.949	1.000

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, one might expect the GLFG trajectories to begin to overlap in structural similarity as well. This hypothesis was tested by applying the same clustering approach to the five 18ns simulations for each FG-Nup (five replicates of each protein). It is clear from Figure 3.3 that simply extending the length of the simulations does not result in structural overlap. This reinforces the 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).

3.3.2 K-means Clustering of FG-Nups

Using the same protocol described above, k-means was directly applied to the MD trajectories using the clustering software developed by Shao et al. [28]. Similar trends can be found in the clustering results obtained from this program as those found in the spectral clustering results. Figures 3.2 and 3.4 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, the mutual information [36] between the replicates and clusters was calculated. Mutual information is a measure of independence of two random variables – the replicate and cluster labels in this 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. This value is then normalized 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 3.1, 3.2, 3.3, and 3.4 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 3.1. 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 3.1 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 discrepancy 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 and quantitative distinguish between the two functional protein classes studied here than standard k-means due to the spectral methods overall higher NMI values for the 18ns simulations and overall lower NMI values for the 3ns simulations.

3.3.3 Comparison of Clustering and Standard Metrics

The previous chapter 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 [21]. 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, *these 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 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 por-

tion 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.

In contrast, the FxFG structures that appear to be the most structurally diverse FG-Nups based on 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 show a great diversity in conformational shape and size (R_g and S) sample fewer distinct regions of structural space than the GLFG-like FG-Nups. In other words, there are no high energy barriers separating one conformational state from another, and the situation seems reversed for GLFG-like FG-Nups.

3.4 Conclusions

While standard metrics of protein size and structure yield some information about the structural variation among the FG-Nups 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 the clustering results it is clear that GLFG and FxFG sample the simulation conformation 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, thereby slowing the structural evolution of these IDPs.

The 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 there begins to be conformational overlap across replicates, this does not guarantee that the space is sampled effectively. Yet, a lack of overlap necessarily means that there is a 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. The clustering tools described here clearly corroborate the observation that these proteins cover two distinct dynamical and functional classes of IDPs.

Chapter 4

A Dimensionality Reduction Approach to Comparing Intrinsic Protein Disorder

In Chapter 3, a clustering framework for assessing the convergence of a set of independent replicate simulations was presented, and its application to a set of IDP simulations suggested that they some were not well-converged, which was not the same result obtained from calculating the decorrelation time of the simulations. While some of the other structural metrics investigated in Chapter 2 back up the results of the clustering framework, methods which provide a visual representation of this result add additional confidence in the results.

In this chapter, the technique of dimensionality reduction is used to visualize the conformation space explored by protein simulations. While several techniques exist for performing dimensionality reduction, some of the properties of unconverged simulations are shown to make the application of certain approaches difficult in practice. Given this situation, one technique is chosen for the task which provides an adequate solution under these constraints, as well as several variations which have yet to be applied to the study of protein dynamics. None of the variations are shown to provide any significant

improvement, but this fact indicates that the standard application of the method is sufficient for visualizing the conformation space explored by the unconverged simulations of much higher dimension than are practically realizable regardless of the particular dimensionality reduction technique being used. In conclusion, *the convergence properties of the IDP simulations can easily be confirmed using two-dimensional projections of the conformation space.*

4.1 Background

4.1.1 Dynamics of Globular Proteins

Several techniques for examining the structural dynamics of simulated proteins have been developed, and many have obtained wide applicability to the study of globular proteins. Perhaps the most widely used methods are those based on harmonic analysis or normal mode analysis. These methods can be used to extract the low-frequency (large-amplitude) motions of a protein, very often the biologically relevant functional properties. These methods are often portrayed as describing the fundamental dynamics of proteins. However, anharmonic motions are common in protein dynamics so these methods often fail to recover these kinds of motions. See [37] for a review of these methods and their application to biomolecular dynamics. Lyman et al. [38] propose an interesting extension to the elastic network model (a form of harmonic analysis also described in [37]) that utilizes data from an MD simulation trajectory to tune various model parameters. Relevant motions are more precisely determined using this approach compared to standard elastic networks.

Since protein motion is often anharmonic, other methods often provide a better picture of the underlying dynamics. One of the earliest and clearest examples is the work by Amadei et al. [39] which describes the application of principal component analysis (PCA) via the $3N \times 3N$ covariance matrix of atom positions for a 900 picosecond solvated simulation of lysozyme. Their analysis reveals the presence of biologically rel-

evant, nonlinear, large-amplitude motions along the first few principal components with motions along the remaining principal components being more constrained and essentially isotropic. This approach was later extended to utilize the space of Φ-Ψ dihedral angles [12] or the vector of interatomic distances [40] instead of the raw atom positions with slightly better results. Feher and Schmidt also used multidimensional scaling approaches, but focused primarily on using the results for effective clustering instead of visualization [41]. More recently, Benson and Daggett [42], use a novel method of applying PCA developed by Teodoro et al. [43] to study the motions of individual atoms across many protein simulations, encompassing a large variety of different folds. They discover concerted directions of motion for residues in α-helices and β-sheets which move primarily perpendicular to the principal axes of these secondary structure elements, while the entire secondary structure elements still show concerted motion in the parallel direction. Even more interesting, they also find that many loop regions, which are often associated with protein flexibility, are sometimes quite a bit more rigid than even α-helices and β-sheets. The use of non-linear approaches, such as Isomap [44, 45] and Laplacian Eigenmaps [46, 47], has also received recent attention in the literature. However, it has also been observed that protein systems larger than just a few residues will often produce conformation landscapes which are typically poorly described by low-dimensional representations using any of the methods mentioned above [48].

4.1.2 Challenges of Disordered Protein Dynamics

While both harmonic analysis, PCA, and non-linear methods provide insights into the motions of globular proteins or short peptides, it is not clear how to apply these methods to large disordered systems. One important reason is that they require a canonical structure which acts as the central reference point for determining motion during the analysis. All motions are deviations around this reference structure which results in certain motions appearing more or less harmonic than they actually are for a disordered ensemble. While a canonical structure could be chosen or created from a

conformational ensemble to use as the reference, it is not clear whether these approaches are well-suited to the task.

Stamati et al. recently presented a method for studying the structural ensembles of short peptides [44]. By applying Isomap directly to the RMSD graph of pairwise structures, the need to determine or in any way define a canonical structure is avoided. The results of Isomap are then used reduce the dimensionality of the data by projecting the structures onto the 2 most relevant non-linear dimensions for visualizing the results.

The approach taken in this work is similar to that of Stamati et al. in two respects: (1) it works with a structural ensemble that isn't tied to any specific canonical structure and (2) it operates on a reduced dimensionality embedding of the structural ensemble of interest. Yet, the method also differs from the approach of Stamati et al. in several respects: (1) metric scaling is used instead of Isomap for constructing a low-dimensional embedding, and (2) the visualization of the conformation landscape is used to validate the convergence of the simulations. This approach is taken because the dimensionality of the conformation spaces studied here are much higher than those used by Stamati et al.. In brief, the gaps or holes in the conformation landscape from unconverged simulations render the results of Isomap essentially equivalent to the simpler metric scaling approach. Also, while the use of other nonlinear techniques which have been developed to deal with disjoint manifolds (e.g. Stochastic Neighbor Embedding [49]) might provide a small amount of additional information on simulation convergence, the investigation of several correction methods for non-metric distances for metric scaling (discussed below) are of current interest since the structure comparison method employed here is non-metric.

4.1.3 Recent Developments in Metric Scaling

Metric scaling provides an elegant, closed-form solution to the problem of embedding pairwise distance data into a Euclidean space. The resulting embedding can used to extract geometrical properties (intrinsic dimensionality, volume, density, etc.)

of the data. Also, the embedded data can be used in conjunction with other machine learning or statistical methods that cannot operate on pairwise distance data since metric scaling casts the data into a more typical vector space representation. The embedding produced by metric scaling is similar to PCA in that the first dimension of the embedding is the one that spans the direction of largest variance in the data, and subsequent dimensions are orthogonal directions stacked in order of decreasing variance. Metric scaling is typically more expensive to compute than PCA (except in the case where the number of data points is less than the dimensionality of the data), but it only requires pairwise distance information instead of vector coordinates. This makes it particularly well-suited for studying systems like disordered proteins, which lack a canonical reference state from which to determine vector coordinates.

While metric scaling is an attractive technique for studying protein dynamics, certain constraints on its operation need to be addressed. Let X be a set of objects and $d(x_i, x_j)$ be the distance between objects x_i and x_j. The complete set of pairwise distances for all objects in X is considered *metric* if:

$$d(x_i, x_j) \geq 0 \qquad \forall x \in X \tag{4.1}$$

$$d(x_i, x_j) = 0 \qquad \text{iff } x_i = x_j \tag{4.2}$$

$$d(x_i, x_j) = d(x_j, x_i) \qquad \forall x \in X \tag{4.3}$$

$$d(x_i, x_j) \leq d(x_i, x_k) + d(x_j, x_k) \quad \forall x \in X \tag{4.4}$$

Given a set of metric pairwise distances, metric scaling can construct a Euclidean space that perfectly preserves these distance relationships. The precision and efficiency of metric scaling make it an attractive method for embedding pairwise data. However, most pairwise distance data from real-world problems do not obey the above constraints due to noisy measurements, missing data, or deficiencies in the distance measure, d. A set of distances for which any element violates one or more of the four constraints above is considered *non-metric*. (One can likewise label the distance measure, d, as metric or

non-metric.) While the nonlinear nature of proteins simulation data presents the most serious drawback to the application of metric scaling, the need for *metric* data is a also a practical drawback to applying metric scaling to protein simulation data.

While the formal underpinnings of metric scaling preclude the use of non-metric input data, previous work has shown metric scaling to nonetheless be a robust embedding method for non-metric distances as well by performing "corrections" to allow metric scaling to compute a Euclidean embedding [50]. Two correction methods have been developed that effectively cast non-metric pairwise distances into a Euclidean space [51, 52]. These will be discussed in more detail but both have a common focus: adding a constant value to all pairwise distances (or to all squared pairwise distances). The scale of the resulting embedding is much larger, but this scaling can easily be factored out in later analyses.

Recently, Roth et al. [53] apparently independently discovered the correction method used by Lingoes [51]. However, the method of Roth et al. differs in its practical implementation. Instead of adding a positive constant to the squared pairwise distances, they simply add this constant to the resulting eigenvalues of the double-centered, squared distance matrix. These modified eigenvalues are then used to construct the Euclidean coordinates as in typical metric scaling (the eigenvectors do not need to be recomputed since they are in fact the same for both the corrected and uncorrected matrices). The result is a more efficient correction algorithm since the limiting step, eigen decomposition, only needs to be computed once (the other correction methods require the eigen decomposition to be recomputed once more on the corrected distance matrix). However, the analyses indicate that the technique of Roth et al. slowly breaks down as higher dimensional embeddings are required, making Lingoes's or Cailliez's [52] correction a more sound option. A formal explanation of this technique in the context of metric scaling will be provided in the next section.

4.2 Methods

4.2.1 Metric Scaling

Metric scaling is a method, summarized by Gower [54], for recovering a set of coordinates in Euclidean space given only the distances between all pairs of coordinates. Consider the coordinate matrix $\mathbf{X}$ of size $N \times M$ where each row vector, $\mathbf{x}$, in $\mathbf{X}$ is a coordinate in $\Re^M$ and N is the number of coordinates. Construct the distance matrix $\mathbf{D}_1$ of size $N \times N$ where each element d_{ij} is the Euclidean distance between points $\mathbf{x_i}$ and $\mathbf{x_j}$:

$$d_{ij} = \sqrt{\sum_{m=1}^{\mathbf{M}} |x_{im} - x_{jm}|^2} \tag{4.5}$$

Likewise, let $\mathbf{D}_2$ be the matrix of *squared* Euclidean distances, where each element is d_{ij}^2. Apply a translation and scaling to $\mathbf{D}_2$ known as double-centering:

$$\hat{\mathbf{D}} = -\frac{1}{2}\mathbf{J}\mathbf{D}_2\mathbf{J} \tag{4.6}$$

where

$$\mathbf{J} = \mathbf{I} - \frac{1}{N}\mathbf{1} \tag{4.7}$$

and $\mathbf{1}$ is the $N \times N$ matrix of ones and $\mathbf{I}$ is the $N \times N$ identity matrix. Then, compute the eigen decomposition of $\hat{\mathbf{D}}$:

$$\hat{\mathbf{D}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}' \tag{4.8}$$

where $\mathbf{Q}$ is the $N \times N$ matrix of eigenvectors and $\mathbf{\Lambda}$ is the $N \times N$ diagonal matrix of eigenvalues. The coordinates of the original points, $\mathbf{X}$, can be recovered by the following:

$$\mathbf{Y} = \mathbf{Q}\mathbf{\Lambda}^{1/2} \tag{4.9}$$

where $\mathbf{Y}$ is an $N \times N$ matrix of Euclidean coordinates. The first M columns of $\mathbf{Y}$ correspond to $\mathbf{X}$, and the remaining columns are all zero. Note that performing dimen-

sionality reduction on the embedding is done by choosing only the first L columns from $\mathbf{Y}$ where $L < M$, similar to PCA.

So long as $\mathbf{D}_1$ is metric, the correct Euclidean embedding, $\mathbf{Y}$, will be recovered exactly if a sufficient number of latent dimensions (L) is chosen. Metric distances guarantee that the double-centered matrix, $\hat{\mathbf{D}}$, is positive semi-definite, making the resulting eigenvalues, $\mathbf{\Lambda}$, and eigenvectors, $\mathbf{Q}$, all non-negative, real numbers. If $\mathbf{D}_1$ is non-metric, then $\hat{\mathbf{D}}$ is not positive semi-definite, and $\mathbf{\Lambda}$ will have at least one negative value and/or $\mathbf{Q}$ will have at least one complex value. Therefore, equation 4.9 will not recover an embedding that is in $\Re^N$. Therefore, some pairwise distance data will not have a corresponding Euclidean embedding.

4.2.2 Metric Scaling Correction Methods

There are conditions under which corrections can be made to $\mathbf{D}_1$ (or $\mathbf{D}_2$) in order to construct a Euclidean embedding using metric scaling. If $\mathbf{D}_1$ only violates the constraint in equation 4.4, then $\mathbf{\Lambda}$ will have at least one negative value, and $\mathbf{Q}$ will only contain real values. In this case, the only problem with recovering an embedding in $\Re^M$ using equation 4.9 is the negative eigenvalues. As a naive solution, one could simply discard the negative eigenvalues and corresponding eigenvectors. The resulting embedding is only an approximation of the true embedding, and, while still useful in the context of dimensionality reduction, it is not guaranteed to preserve the pairwise distances in any reasonable way.

Lingoes presents the first technique for addressing the problem of negative eigenvalues [51]. First, partially compute the metric scaling solution using equations 4.6, 4.7, and 4.8. Then compute an adjusted $\mathbf{D}_2$ by adding a constant value $-2\lambda_{\text{min}}$ to all *off-diagonal* elements of $\mathbf{D}_2$ where λ_{min} is the smallest eigenvalue in $\mathbf{\Lambda}$. Then repeat metric scaling on the adjusted squared distance matrix. This method ensures that $\hat{\mathbf{D}}$ will be positive semi-definite, and all eigenvalues will now be non-negative.

Cailliez later presented an analytical technique for calculating an adjustment to

the original distances rather than the *squared* distances [52]. This constant is obtained by solving for the largest eigenvalue (λ_{max}) of the following matrix:

$$\begin{bmatrix} \mathbf{0} & 2\hat{\mathbf{D}} \\ -\mathbf{I} & -\mathbf{J}\mathbf{D}_1\mathbf{J} \end{bmatrix} \tag{4.10}$$

The new squared distance matrix can then be obtained by adding λ_{max} to all off-diagonal elements of $\mathbf{D}_1$ and then squaring element-wise. Then repeat metric scaling on the adjusted squared distance matrix. Again, $\hat{\mathbf{D}}$ will be positive semi-definite, and all eigenvalues will be non-negative.

The method of Roth et al. is similar [53]. After computing equation 4.8, add $-2\lambda_{\text{min}}$ to all diagonal elements of $\mathbf{\Lambda}$. Then compute the embedding $\mathbf{Y}$ using the adjusted $\mathbf{\Lambda}$.[1] The theoretical motivation for this method appears to be identical to those proposed by Lingoes, but it is computationally less intensive.

Computing metric scaling for large sets of structures can be computationally demanding if N is large. However, efficient methods exist for extending metric scaling embeddings with out-of-sample points [55]. This would allow the incorporation of the remaining structures into the results. Such methods were unnecessary for the analyses presented here, but will no doubt be critical to the assessment of large-scale simulations such as those obtained via coarse-graining or enhanced-sampling simulation techniques. Such techniques are beyond the scope of this work, and the reader is referred to [56] and [57] for recent applications of these techniques.

4.2.3 Molecular Dynamics Simulations

Just as in Chapter 3, the simulation data set from Chapter 2 is utilized. This data set consists of a set of intrinsically disordered protein fragments simulated at short and long time-scales that have been shown to exhibit different dynamical properties.

[1] Roth et al. actually throw out two of the eigenvectors in their original formulation, corresponding to the last eigenvalue and the one zero eigenvalue in the standard metric scaling solution. Here, all eigenvectors are retained for generality, though this has little effect on the results.

Please see Section 2.2.1 for the details on the proteins simulated, protocols employed, and general motivation for this approach.

4.3 Results

4.3.1 Dynamics of a Single Trajectory

First, a single 18ns trajectory of each FG-Nup is analyzed using metric scaling. Structures were sampled from the trajectory every 10ps. The RMSD between every pair of structures was computed using only the backbone C_α atoms in Angstroms (Å). RMSD is a non-metric distance measure[2], and the resulting distance matrices were all non-metric and symmetric.

Figure 4.1 shows a plot of the percentage of variance accounted for by each subsequent dimension. (Only the non-corrected method is considered for now. See section 4.3.3 for a more complete discussion of the correction methods.) Reducing the dimensionality of the data set can be achieved by excluding dimensions that account for small fractions of the variance, similar to PCA. We utilize this technique by examining the data reduced to just two dimensions. Of course, this removes a large percentage of the useful distance information from the embedding, but it is shown to still provide more insight into the structural dynamics of the proteins compared to standard methods.

The top of Figure 4.2 shows the change in RMSD from the initial structure over time for GLFG (left) and SxSG (right), a common method of determining if the motions of a protein are reasonably equilibrated. Over time, the plot should level off, indicating that the amount of structural change has become constant. For folded proteins this typically converges to 2-3Å RMSD. Compare these plots to those in the bottom of Figure 4.2 which shows plots of the first two dimensional components from metric scaling. The color shift indicates progression in time.[3] The RMSD plot for SxSG shows that

[2]RMSD violates the constraint in equation 4.4, but none of the remaining three.

[3]All 2D maps incorporate some information from the third dimensional component as well since all coordinates were sorted along this dimension and then plotted in order. Hence, the lines are "stacked"

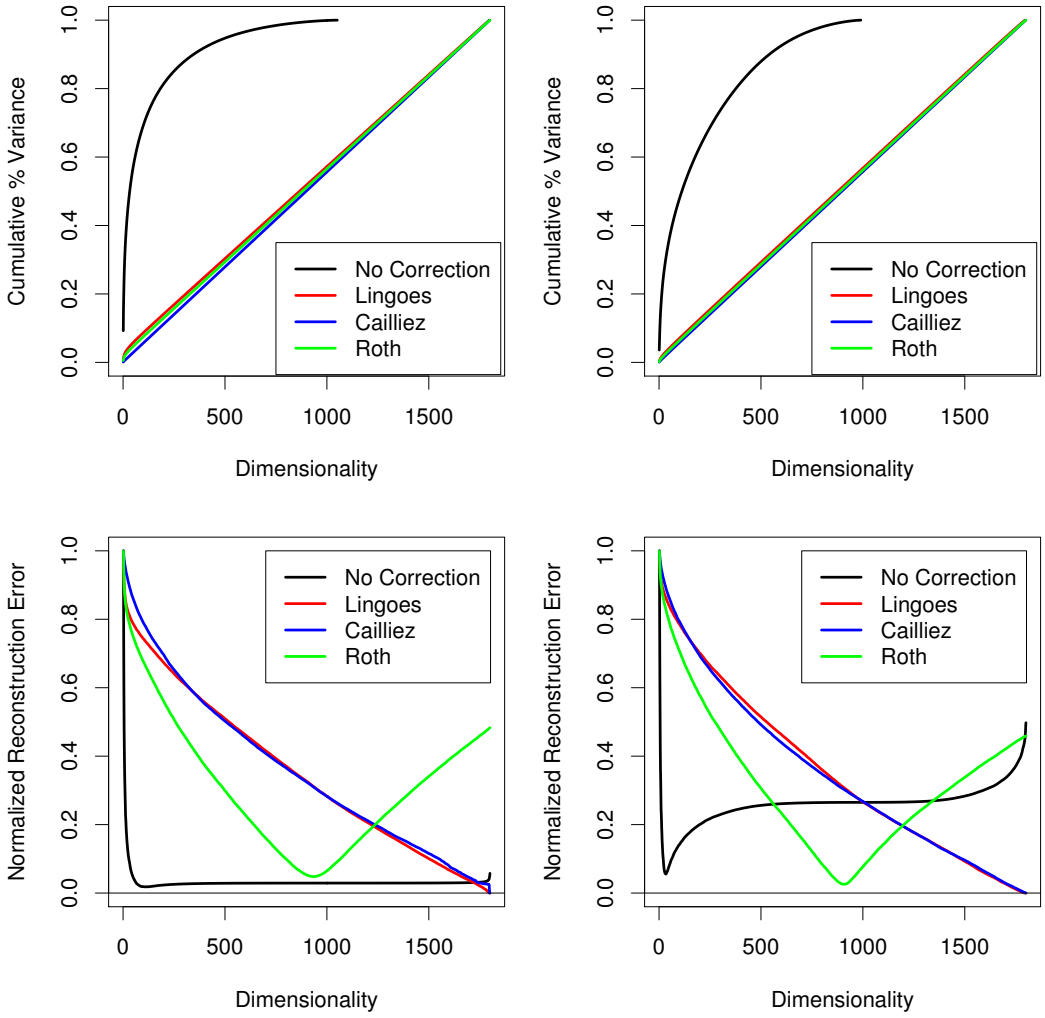

Figure 4.1: Top: Cumulative percentage of variance accounted for by each dimension for GLFG (left) and SxSG (right). Bottom: The reconstruction error attained by utilizing the specified number of dimensions. The noisy nature of molecular motion makes it difficult to determine an optimal number of dimensions, but these results indicate that the main modes are captured as effectively by metric scaling *without* corrections as metric scaling with corrections.

the RMSD quickly converges to a large value, consistent with expected values for a disordered protein, but the one for GLFG is showing transitions between intermediate RMSD values. However, the metric scaling results from these two simulations clearly show that GLFG has not converged.

The metric scaling results also reveal differences between the motions exhibited by GLFG and SxSG. First, the motions of GLFG are much more constrained than those of SxSG. This can be seen by comparing the scale along each dimension. The range of the motion is roughly three times greater for SxSG. Second, the motion of SxSG is fairly unconstrained, and it quickly covers the accessible conformation space. It also regularly revisits areas of conformation space. However, GLFG motions are highly constrained and the simulation has trouble exploring the accessible conformation space, becoming trapped in several conformations. Third, it is clear that there is no single reference structure that can summarize the conformations explored by these proteins. While GLFG might be described in terms of a few structures due to the tendency to remain trapped in a few conformations, it is not clear how many structures would be needed to similarly summarize SxSG. Even if such a set was generated for SxSG, it is unclear how this would be useful for comparing the dynamics of two disordered proteins. Overall, GLFG behaves like a premolten globule while SxSG displays no conformational trapping, exploring the available conformation space like an extended coil.

4.3.2 Comparing Multiple Trajectories

Structures sampled every 10ps from all forty 3ns simulations of SxSG and GLFG were combined for metric scaling. The same was done for the five 18ns simulations. The resulting two-dimensional embeddings are shown in Figure 4.3. It is clear that the SxSG simulations sample the available conformation space more effectively than GLFG. GLFG consistently becomes trapped in some conformations as shown earlier. However, the effect is even more pronounced in the 18ns simulations. The results indicate that

into the third dimension. This additional information makes the temporal effects more clear in the 2D plots.

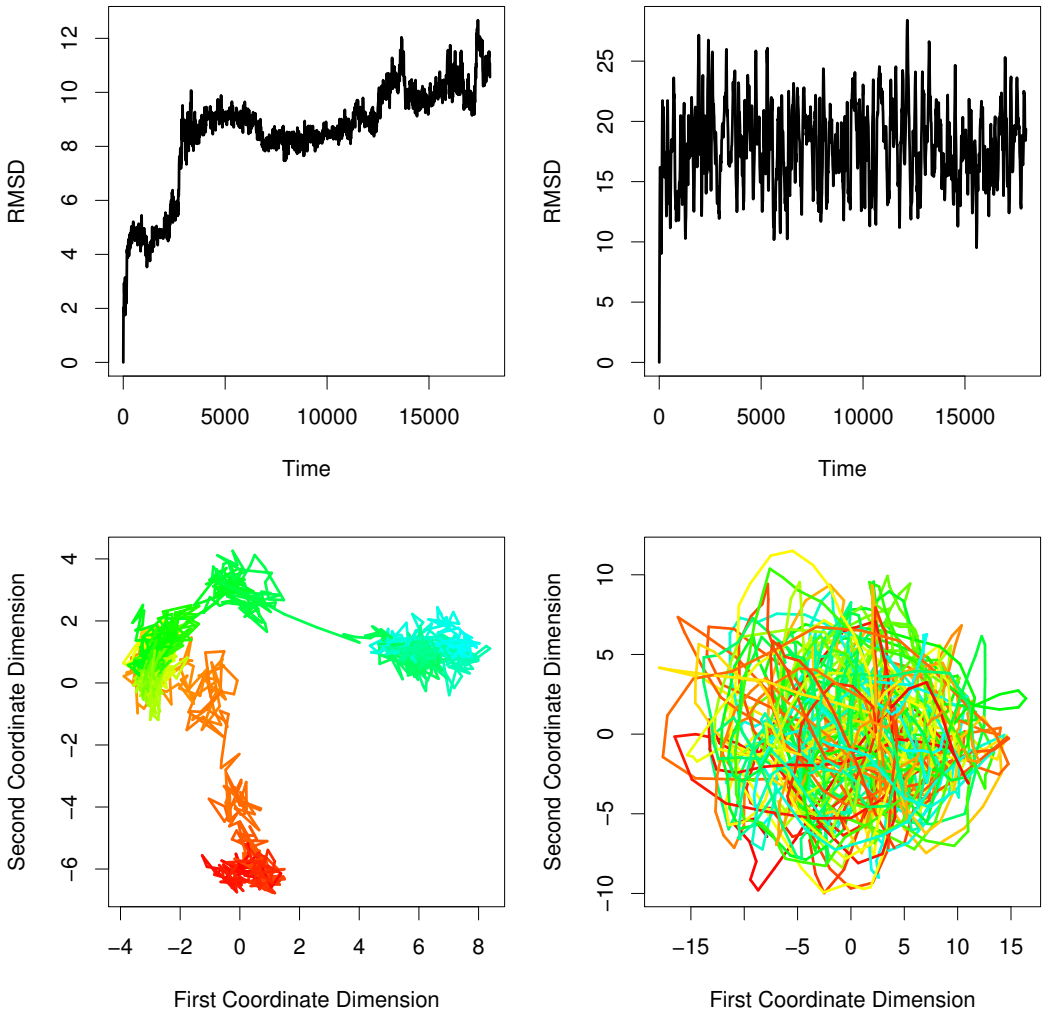

Figure 4.2: Top: RMSD from starting structure versus time for GLFG (left) and SxSG (right), a typical method of determining if conformational sampling has converged. Bottom: Two-dimensional maps provided by metric scaling for GLFG (left) and SxSG (right). The RMSD plot for GLFG is ambiguous as to whether the simulation is converging, but the metric scaling results clearly show that the GLFG simulation is not converging. However, SxSG has clearly converged.

simulation protocols favoring longer simulations are in danger of undersampling the conformational space of certain disordered proteins. In order to effectively sample the conformation space of cohesive disordered proteins (similar to GLFG), many shorter simulations would be far more adequate than longer simulations which would become trapped in rather static conformations. However, extended coils (similar to SxSG) would sample conformation space more effectively in longer simulations.

4.3.3 Comparing Correction Methods

The individual 18ns trajectories examined earlier were also tested with all three correction methods for non-metric distances. The top of Figure 4.1 shows the percentage of variance accounted for by each subsequent dimension while the bottom shows the normalized reconstruction error attributed to using each of the correction methods compared to standard metric scaling. The reconstruction error is the sum-squared difference between the original distance matrix used to obtain the embedding and one that is calculated from the metric scaling embedding itself. Increasing numbers of dimensions can be added and compared in order to determine the "optimal" size embedding, potentially signaled by a minimum in reconstruction error.

Using no corrections, a minimum is observed that corresponds to a reasonably small number of dimensions and reconstruction error. For Lingoes and Cailliez corrections, the minimum is not observed until all dimensions are used. The minimum is actually zero in these cases, which indicates that the corrected embeddings are perfectly Euclidean. However, the need to use so many dimensions is impractical for application here. The two dimensional embeddings were the same as the non-corrected results to a scaling factor (data not shown). Therefore, for examining protein dynamics in just a few dimensions, non-corrected metric scaling works well compared to the corrected versions which show larger reconstruction error for the same number of dimensions. However, future work is needed to understand if higher dimensional embeddings would need the corrections to accurately describe the dynamics.

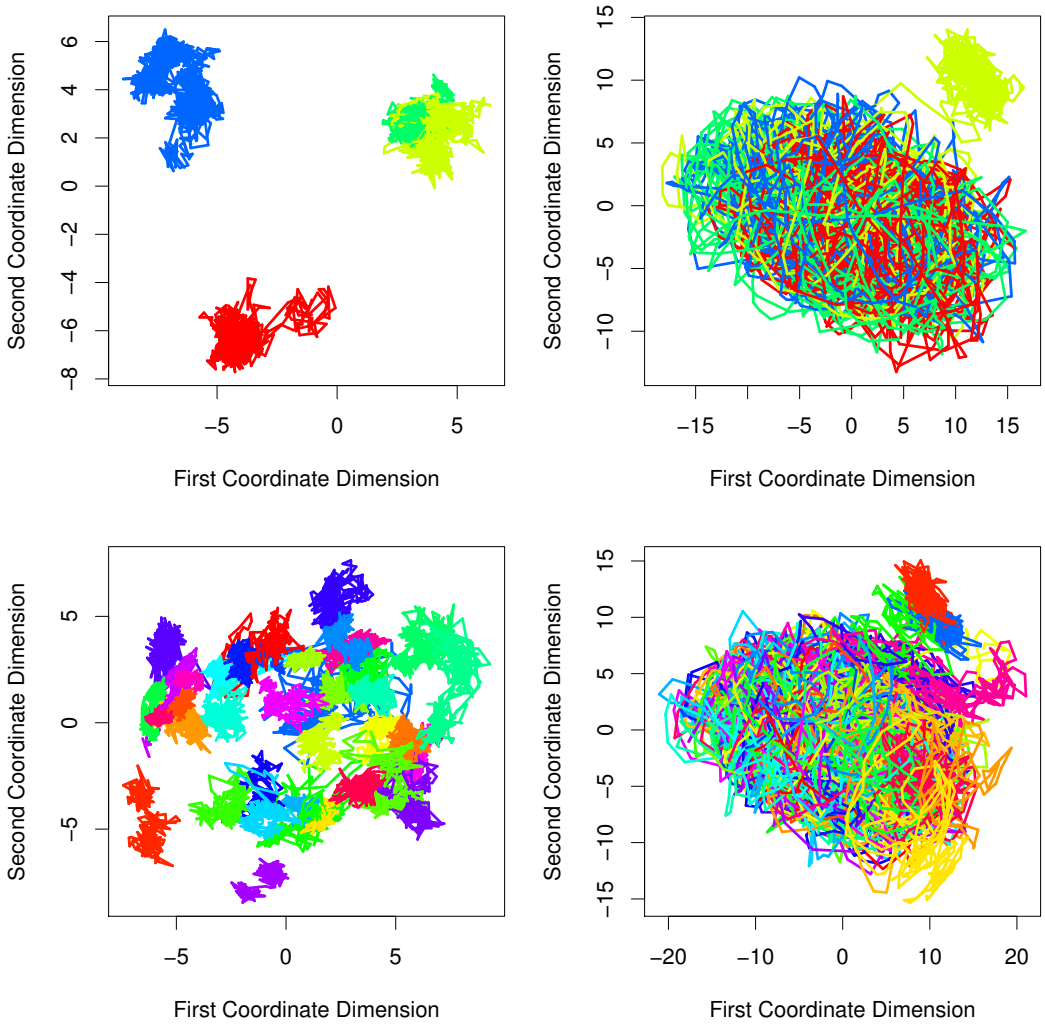

Figure 4.3: Two-dimensional embedding results from metric scaling of structures sampled across all trajectories for GLFG (left) and SxSG (right). Each trajectory was plotted as a uniquely colored line through the embedding. Top: five 18ns simulations Bottom: forty 3ns simulations

Since the method proposed by Roth et al. has the same underlying theory as Lingoes method, the reconstruction error curve should be similar, but this is found to not be true in practice. Notice the deviations from Lingoes method for higher dimensions, a result of the fact that the eigenvectors associated with smaller eigenvalues need to be recomputed for accurate reconstruction to occur. By avoiding this calculation, the method of Roth et al. is much more efficient, but is no longer able to properly reconstruct the corrected distances for higher dimensional embeddings.

4.4 Conclusion

Metric scaling provides fundamental insights into the dynamics of disordered proteins. These dynamics are not revealed using traditional MD simulation analysis methods. The results have shown that metric scaling can distinguish between the dynamics of differing disordered proteins both quantitatively and qualitatively. The method directly computes the size of the structural space sampled by MD simulations and the density of the resulting embeddings can inform protocol development for the simulation of disordered biomolecules.

The results in this chapter confirm earlier results concerning the convergence properties of this set of IDP simulations. In particular, it is clear by visual inspection that GLFG simulations are less converged than the AxAG simulations, which is in concert with the earlier clustering results in Chapter 3. While the visual confirmation of this fact was informative, the clustering protocol provided a more precise quantitative discrimination between the different proteins, and the correction methods for metric scaling don't appear to provide any significant improvements to the resulting embeddings. In fact, the effects seem to be greatest in the least-significant dimensions rather than the most significant ones.

In the future, non-linear dimensionality reduction methods could be utilized [58, 59, 60] for studying IDP dynamics, especially since fast computational approaches for these methods have been recently introduced [61]. In addition, methods for studying

manifolds of mixed density and dimensionality [62, 63] may be more appropriate for MD simulations than those that focus on learning embeddings with a single, fixed number of dimensions. Another major consideration is the effect of noise on these techniques. While low-pass filtering is utilized in later chapters as a fast method for for reducing noise, other techniques for smoothing noise [64] have been developed which may prove more effective for future dimensionality reduction studies of protein dynamics.

Chapter 5

Validation of Clustering Algorithms for Protein Simulations Using Polymer Models

Chapter 3 illustrated how clustering techniques can be used in to calculate useful properties of molecular simulations. MD simulation is particularly well-suited for sampling the meta-stable and transitional conformations which characterize IDP dyanmics and are relevant to performing their requisite functions, and computational data clustering has emerged as a useful, automated technique for determining the meta-stable and transition states from MD simulations. Clustering methodologies applied to the results of MD simulations focus on partitioning structural ensembles into groups of structures which share similar conformational features. It is hoped that when applied to simulations of biomolecules, the clustering results in partitions which correspond to the descriptive – metastable and transition – states of the system. However, clustering the trajectories of real biomolecules typically does not readily provide such a straightforward partitioning due to the high dimensionality of the conformational space, thermal noise, and other factors. Identifying the descriptive states also requires an understanding of the clustering process itself. *A primary goal of this chapter therefore is to provide*

such an understanding through a detailed analysis of data clustering applied to a series of increasingly complex biopolymer models.

In this chapter, a novel series of models is developed using basic polymer theory that have intuitive, clearly-defined dynamics and exhibit the essential properties that one might seek to identify in MD simulations of the real biomolecules. Importantly, these models allow us to determine the properties a clustering algorithm can reliably extract from polymer data, unconfounded by the computational complexities and limitations of all-atom simulation. A series of models is created where each new model increases upon the complexity of the previous so that the dynamics and properties start to approach that of all-atom simulation dynamics. Spectral clustering is applied to the model polymers, and statistics from the various clusters are computed in order to determine which statistical features link clusters to the properties displayed by the polymer models. Finally, the clustering results from the polymer studies are compared to clustering results from all-atom MD simulations of several IDPs. While spectral clustering is again the clustering algorithm of choice for this study, it is not the only algorithm that could be used, and another algorithm might actually be more appropriate and accurate for computing these properties. This protocol allows us to determine if and when the clustering method is no longer able to determine the descriptive states of the systems, as well as the underlying reasons for these limitations.

5.1 Background

While it is clear that clustering has been widely used in the field of MD simulation, relatively little work has been done to determine if the clustering algorithms are actually extracting useful information. For instance, Shao et al. provide one of the few (if not the only) in-depth studies of clustering for MD simulation [28]. They compare various clustering algorithms to determine how well these algorithms can adequately separate structures in ensembles taken from manually concatenated, remarkably distinct MD trajectories. Even though the trajectories cover very different portions of conforma-

tion space, there is no clear winner among the algorithms they chose to study. In fact, all of the algorithms perform well on some problems, but not so well on others. Therefore, it is clear that, while comparing algorithms might yield the "best-case" algorithm for a particular system where the solution is known or anticipated, the ability to determine exactly which properties can be determined using a particular clustering algorithm more generally remains to be investigated.

The recent focus on MD as a tool for exploring nonequilibrium processes has driven the simulations to longer timescales than ever before [65, 66, 67, 27]. The data gathered from such simulations can be extensive so clustering has a key role to play in summarizing the simulation output without losing the key properties and behaviors of interest. Since clustering algorithms are a form of unsupervised learning, where there is no additional evidence or knowledge guiding the algorithm aside from the data itself, and since experimental information may not be available at the spatial and temporal resolution of MD simulation, additional insight and understanding are needed to interpret the clustered data. This study proposes that polymer models which exhibit simplified and/or well-understood structural dynamics can be used to study clustering techniques, and help to bridge the gap between using clustering to confirm established results and using clustering to make theoretical predictions concerning the dynamics of biopolymers.

5.2 Methods

5.2.1 Polymer-based Validation Framework

The following framework and procedure for using polymer models and simulations to guide clustering based approaches to identify the descriptive states of all-atom biomolecular simulations is used. See also Figure 5.1.

1. A polymer model is used to create a structural ensemble with well-characterized properties such as identifiable metastable and transition states.

2. The polymer model ensemble is clustered.

3. Various statistical properties are calculated for the resulting clusters.

4. The statistical properties of clusters known to correspond to metastable and transition states are identified.

5. MD simulations of a chosen biopolymer system are used to generate a structural ensemble.

6. The MD ensemble is clustered.

7. The same statistical properties are calculated for the resulting clusters from the MD ensemble.

8. Correlations between the statistics from steps 3 and 4 are used to characterize the clusters from the MD ensemble as corresponding to metastable or transition states.

Simple polymer models are focused on first with few interesting features, and then incrementally features are added to create a range of polymer simulations. These extended models are designed to possess densely populated metastable states and the sparsely-populated transitions states that lie in between. We repeat the above process for each model so that the analysis of the more complex models always builds upon the analysis of the simpler models.

Two polymer models are developed where the pairwise dissimilarities can be computed analytically and two polymer models where the pairwise dissimilarities can be computed from analytically derived polymer structures. Also, one polymer model is utilized where the pairwise dissimilarities can be computed from polymer structures derived from simulation. Although these models are described in detail in the next section, the following list summarizes all of the models used in this work:

- *Linear Model* - This analytic model is the simplest dynamical model considered here. It does not exhibit any metastable or transition states.

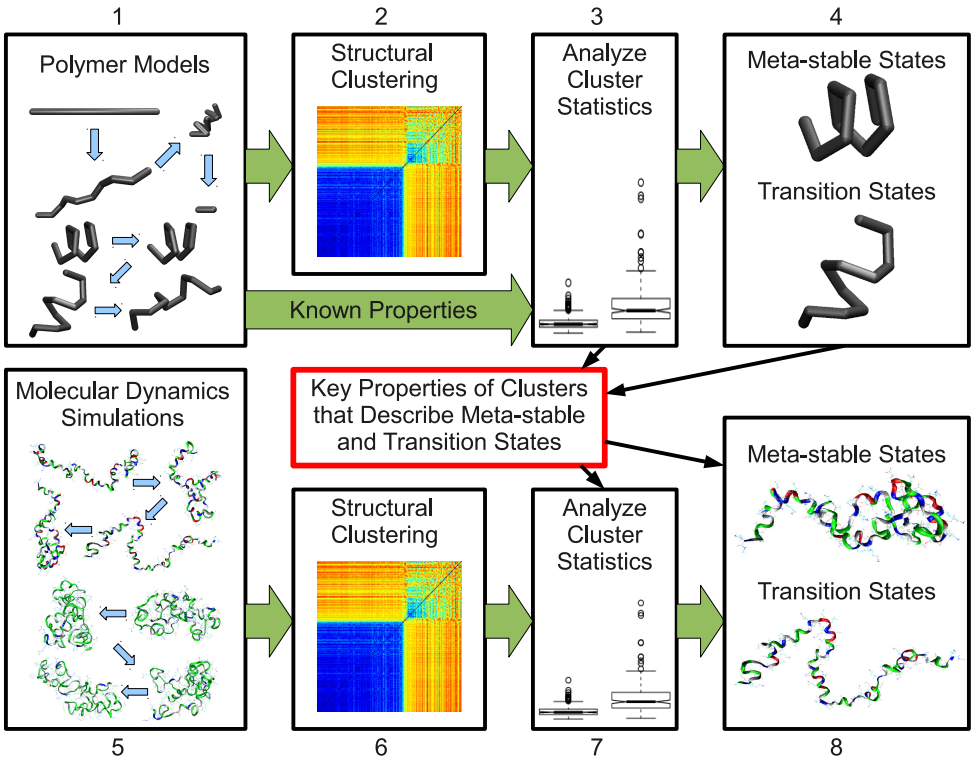

Figure 5.1: Clustering validation protocol. This diagram outlines the process of verifying the results of a clustering algorithm applied to molecular dynamics simulations. The model polymers have certain known properties (metastable states, transition states, etc.) that are known beforehand due to model construction. The statistics and features of clusters that describe these properties can then be used to inform the analysis of molecular dynamics simulations. If the properties displayed by the polymer models are also be present in the molecular dynamics simulations, then the statistics and features of the clusters will also share similarities. Any clustering algorithm could be used, but here the focus is on spectral clustering.

- *Sinusoid Model* - This analytic model builds upon the linear model by the addition of metastable and transition states.
- *Rotation Model* - This model consists of polymer structures generated by changing the polymer link angles in well-behaved manner. It also does not exhibit any metastable or transition states.
- *Cyclical Model* - This model extends the rotation model by revisiting the visited conformational states several times over.
- *Dynamic Model* - This model consists of a helix-favoring polymer that "folds" and then "unfolds" over the course of a simulation.

5.2.2 Polymer Models

Linear Model

The linear model is one of the simplest models of polymer dynamics that can be constructed. It is not necessary to generate the actual structures since the dissimilarity matrix, $\mathbf{X}$, of the system can be determined analytically according to the following equation:

$$\mathbf{X} \in \mathbb{R}^{n \times n} \text{ where } \mathbf{X}_{ij} = \|i - j\| \tag{5.1}$$

with $n = 1000$ for the results presented here.

Sinusoid Model

The sinusoid model exhibits two of the key features of interest to MD simulation studies: metastable and transition states. Like the linear model, the dissimilarity matrix can be constructed analytically, so the generation and comparison of actual polymer

structures is not necessary:

$\mathbf{X} \in \mathbb{R}^{n \times n}$ where

$$\mathbf{X}_{ij} = \left\| \sum_{u=1}^{i-1} \left[\cos\left(\frac{6\pi(u-1)}{n-2} \right) + z \right] - \sum_{v=1}^{j-1} \left[\cos\left(\frac{6\pi(v-1)}{n-2} \right) + z \right] \right\| \quad \text{and } z \in \mathbb{R}, z > 1 \quad (5.2)$$

with $n = 1000$ for the results presented in this paper.

The parameter z is added to the cosine functions in order to ensure that the contributions to their respective sums are always positive values, thus ensuring that $x_{ij} < x_{i(j+1)}$ for all i, j. If we allowed $-1 \leq z \leq 1$ then some temporally adjacent structures would actually be moving toward the origin or stay in the same locations, rather than continuing to evolve away from the origin. It turns out that $z < -1$ also produces reasonable results, where the starting and ending structures reside in metastable states, and two metastable states are created in the middle of the trajectory. However, since we want to observe three metastable states in the middle of the trajectory, z is constrained to be greater than one. Note also that this model asymptotically converges to a linear model as z goes to infinity (or negative infinity), but this property serves no practical purpose in this study. Therefore, $z = 1.01$ (a value slightly larger than 1) for all results presented here.

The corresponding polymer "simulation" shows dynamics indicative of three distinct metastable states and four transition states (one at the beginning, one at the end, and two in-between the three metastable states). This model is similar to the linear model because the simulation is always progressing into new areas of structural space. However, the distance between successive frames is adjusted according to a nonlinear, sinusoidal pattern. This produces the three distinct metastable states by compressing the distances between frames in three regions, while the intervening regions, corresponding

to the transition states, are produced by dilating the distances between successive frames in these regions. These dynamics resemble diffusion on a glassy free-energy surface, which is a feature purportedly common among disordered proteins [17].

Rotation Model

The rotation model is the first polymer model used in this study where 3D polymer structures were constructed for comparison using RMSD. The model defines a polymer structure by a set of consecutive links, each 3.88Å long, analogous to the C_α trace of a protein. There is no steric exclusion, so links may overlap without penalty. The angle between successive links is governed by the polar angle (ϕ) and the azimuthal angle (θ) which range from $[0, 2\pi)$ and $[0, \pi)$, respectively. These two angles are initially set to 0 degrees, resulting in a fully extended chain. The angles are then incremented at each time step by a small amount (2ϵ and ϵ) until the chain completely winds into a tight helical configuration. The number of links used was 10 (11 particles). Here, $\epsilon = \pi/(n-1)$ and $n = 1000$ for the results presented here‘.

This model is similar to the linear model presented earlier because the amount of structural change between successive structures is constant. However, using RMSD to compute dissimilarity between structures results in a nonlinear distortion of the polymer similarity space. Therefore, this model can be utilized to determine if the use of RMSD presents a challenge to clustering the structures in a manner that fully captures the underlying linear model.

Cyclical Model

The cyclical model is an extension of the rotation model in which the ϕ and θ angles are incremented until reaching their maximal values and then subsequently decremented until reaching zero. This process is repeated three times so that the polymer cycles through three phases of collapsing and extending. In this work, we utilize $\epsilon = 6\pi/(n-1)$ and $n = 1000$ for the cyclical polymer model. It is important to

recognize that incrementing the angles ϕ and θ by 2ϵ and ϵ, respectively, beyond their maximal values results in creating a left-handed "helix", while the earlier conformations simulated during collapse (also from incrementing the angles) were all right-handed. The angle decrementing phase is necessary to avoid this problem, resulting in a model where all structures are of the same handedness, similar to biopolymers. This ensures that the structures sampled during the expansion phase of the model are the same as those sampled during the collapse phase.

The cyclical model is similar to the linear model because the angle parameters are adjusted in a linear fashion, but it has several interesting additional properties. First, several *false* metastable states are created. This arises from the use of RMSD for comparing the polymer structures, which again results in a nonlinear distortion of the underlying linear process. Second, the model revisits these false states several times. Therefore, this model is useful for determining how sensitive a clustering algorithm is to the nonlinear distortion of RMSD and how these "false" states differ from the metastable states in the sinusoid model.

Dynamic Model

The dynamic model is the bridge between the analytical models described above and the all-atom MD simulations. The details of the model are fully described in [68]. In the dynamic model, a polymer consists of a string of particles connected by rigid bond constraints, analogous to the links of the previous analytical models. For the purposes here, a link length, l, of 1.3 units is utilized. A soft pairwise potential is applied to eliminate the overlap between the particles and a torsional potential is also applied to the bonds is specified to favor a helical conformation. The periodicity of the helix, $h_p = 5$, to consist of five consecutive links. Therefore, the polar angle $\phi = 2\pi/h_p$ radians, which remains fixed throughout the simulation. The azimuthal angle, θ, is allowed to vary, but has an equilibrium value of $\theta_0 = \arcsin(1.1 r_{cut}/(h_p * l))$ radians, where $r_{cut} = \sqrt[6]{2}$ is the distance cutoff for the neighbor-list. The particles are assigned

initial velocities according to the Maxwell distribution. Newton's equations of motion are integrated using the leap frog method and velocity scaling is used on each time-step in order to keep the average kinetic energy in the system at the desired level.

This model was utilized to perform both a freezing and melting simulation. These two simulations demonstrate two commonly studied phenomena for proteins: folding and unfolding. Initial particle positions are assigned to be either a random coil or folded helix, respectively. The random coil is generated by uniformly sampling the space of torsional angles and the folded helix is generated by setting the torsional angles equal to θ_0. The temperature is slowly annealed every 4000 steps following the first 10000 steps in the simulations according to the following relationship: $T_{current} = \gamma T_{previous}$. For the freezing simulation, $\gamma = 0.925, T_0 = 6$, and for the melting simulation, $\gamma = 1.0811, T_0 = 0.1217$. Each simulation is run for 210000 steps and structures are saved every 400 steps after the initial 10000 steps, for a total of 500 structures per simulation. The polymer consists of 10 links (11 particles), similar to the analytical models above, completing two complete helix turns in the folded state. The integration time step size is set to 0.004, the size of the simulation box is set to 12 units along each side, and the torsional force constant is set to 5. This set of parameter values, and the above annealing schedule allows the freezing simulation to quickly fold the polymer without becoming trapped in local minima in the potential energy surface (kinked helices). The final temperature of the melting simulation is approximately equal to the starting temperature of the freezing simulation, and vice-versa. Therefore, the folding/unfolding events occur at approximately the same number of steps into the simulations. Finally, the two simulations were concatenated to create a single freezing-melting simulation with a total of $n = 1000$ structures.

5.2.3 Spectral Clustering

For all of the work in this chapter, the spectral clustering approach outlined in Section 3.2.1 is utilized. Unlike the prior chapters, the σ_i values are allowed to adapt to

the data according to the method outlined in step 4 of the algorithm in Section 3.2.1.

5.2.4 Molecular Dynamics Simulations

The MD simulations used in this chapter were a subset of those described in Section 2.2.1. In particular, only one of the 18ns simulations for the each of the GLFG, FxFG, and SxSG proteins was used, and each of these was subsampled every 2ps for a total of 9000 structures in each simulation. Please see Section 2.2.1 for the details on the proteins simulated, protocols employed, and other parameters used to generate these trajectories.

5.2.5 Clustering Protocol

Data from all the analytical models, the dynamic model simulation, and the MD simulations were processed using spectral clustering for several values of k (the number of clusters requested): 3, 5, 10, and 15. These values were chosen to examine how a wide range of k can be used to reliably determine the built-in properties of each model. A wide range of values such as these would likely need to be tried for any novel data set since we normally would have no indication of what value of k to choose *a priori*. Features extracted for each of the resulting partitions include: the scaling parameters for each structure (σ_i), the number of structures in each cluster, the distribution of intra-cluster RMSDs, and the distribution of scaling parameters for each cluster.

The polymer models and protein simulations studied here revealed that sampling several values of k was needed to determine the presence of metastable and transition states. In general, some of these states will become discernible at low k, but others will require higher k in order to properly partition these states into separate clusters. However, some other heuristics could be used to constrain the space of k values to explore. For example, the need to gather adequate statistics will somewhat constrain the search along k. If too many (or too few) clusters are requested, then the confidence intervals of the various statistics for each cluster would begin to consistently overlap.

Such heuristics were not employed in the work here since the approximate confidence intervals calculated by the box plots showed sufficient statistical confidence for at least one of the selected values of k for each model. However, it may be possible to utilize such statistics to find a preferred value (or subset) of k, instead of manually examining a range of values as is done here. Exploring the adequacy of this and other heuristics will be the subject of future work.

5.3 Results

5.3.1 Linear Model

First, the performance of spectral clustering on the linear model is examined. The "simulation" corresponding to the linear model possesses dynamics where the structural dissimilarity differs by a constant amount between successive frames, and the polymer is always progressing into new areas of structure space. This can be observed in the linear increase in RMSD from the initial structure as a function of simulation time as shown in Figure 5.2A. This is also observed in the linear increase in RMSD as a function of the difference in time between pairs of structures as shown in Figure 5.2C.

Spectral clustering is shown to behave as expected for the linear model. Figure 5.2 shows the structure assignment and cluster sizes for various values of k (the number of clusters). Each cluster consists of a temporally contiguous set of structures that share no similarity to the structures in the remaining clusters. The cluster sizes at the start and end of the simulation are slightly lower, which occurs because of clustering start- and end-effects.

The clusters at the beginning and end of the simulation are both less structurally diverse as indicated by the narrow intra-cluster pairwise RMSD distributions for these clusters shown in Figure 5.2. Both of these clusters also have a few structures with rather large scaling parameters relative to other clusters and structures as indicated by outliers in the intra-cluster scaling parameter, σ (box plots shown in Figure 5.2). These

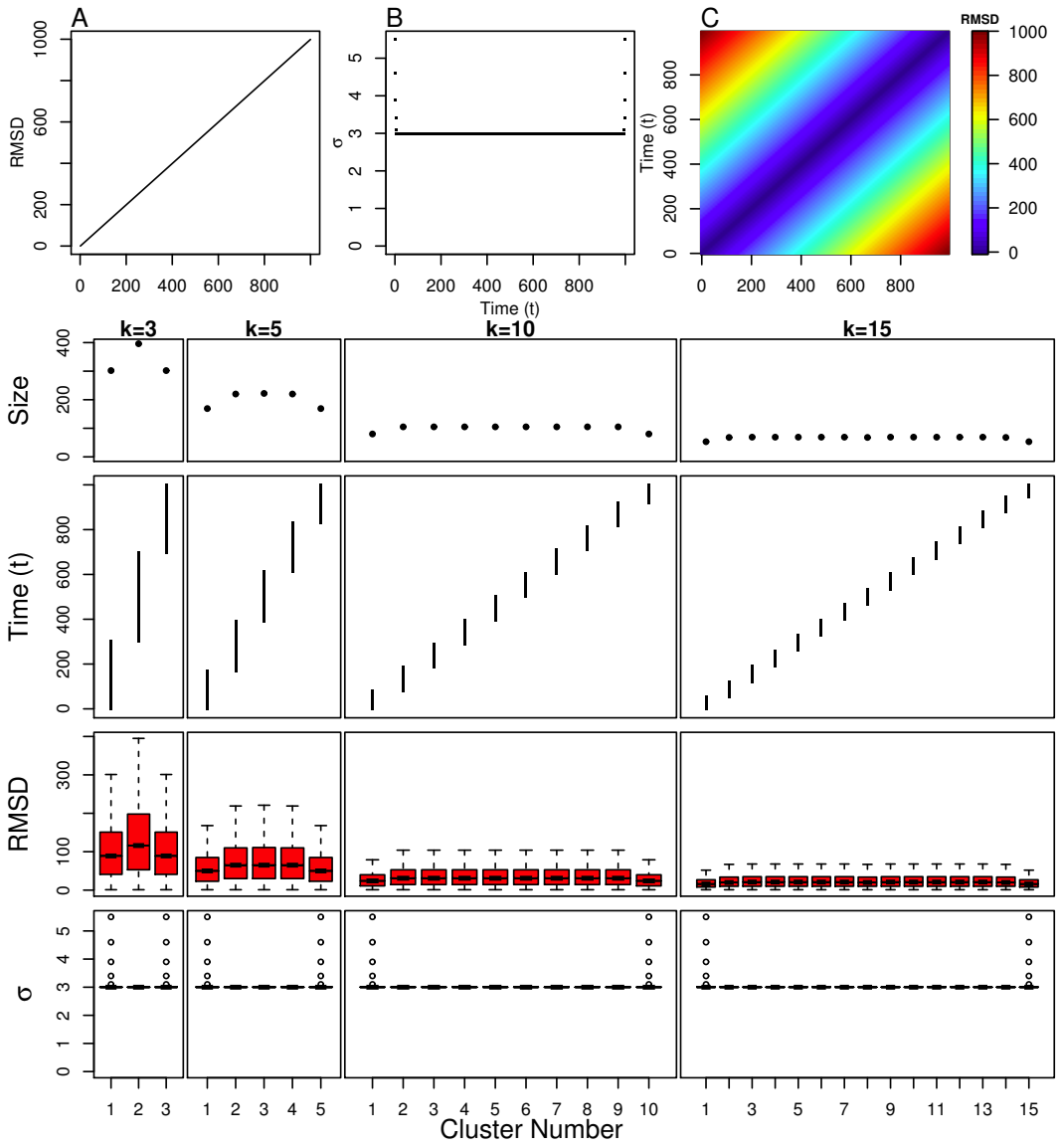

Figure 5.2: Linear model clustering results. Top: (A) Distance from initial structure, (B) local scaling parameters and (C) pairwise distance between all structures. Bottom: Cluster assignment and statistics for several values of k.

results also indicate that the structures at the beginning and end of the simulation have fewer close neighbors than structures in the middle of the simulation, which is confirmed by plotting the scaling parameters as a function of time as shown in Figure 5.2B. The effect is mild, and suggests that spectral clustering does not let the edge-effects of the simulation override the importance of partitioning the structures into clusters that all have a common implicit degree of similarity. The metastable or transition states at the edges of the sampled conformation space are not unduly penalized nor overly favored by spectral clustering.

5.3.2 Sinusoid Model

Next, the performance of spectral clustering on the sinusoid model is examined. The sinusoid model shares one key property with the linear model: the polymer is always progressing into new areas of structure space. However, the distance between successive structures is now varied so that at certain times in the simulation, successive structures are closer together, representing a dense region of highly similar structures akin to a metastable state. At other times in the simulation, successive structures are further apart, representing a sparse region of dissimilar structures akin to a transition state. In particular, this model exhibits three metastable states with two transition states in-between. The beginning and end of the simulation are both at points where the structural distance between successive structures is quite high and are both characteristic of a transition state as well. These properties can be observed in the RMSD plots shown in Figures 5.3A and 5.3C. The centers of the metastable states are found at $t = 133, 500, 833$, which is where the slope of the line describing RMSD from the initial structure versus time is zero, and also where the lowest values of pairwise RMSD are found (Figure 5.3C).

It is clear that the clustering algorithm is able to extract the metastable states from the sinusoid model. Figure 5.3 shows that spectral clustering divides this simulation into clusters of temporally contiguous structures and that these clusters contain similar

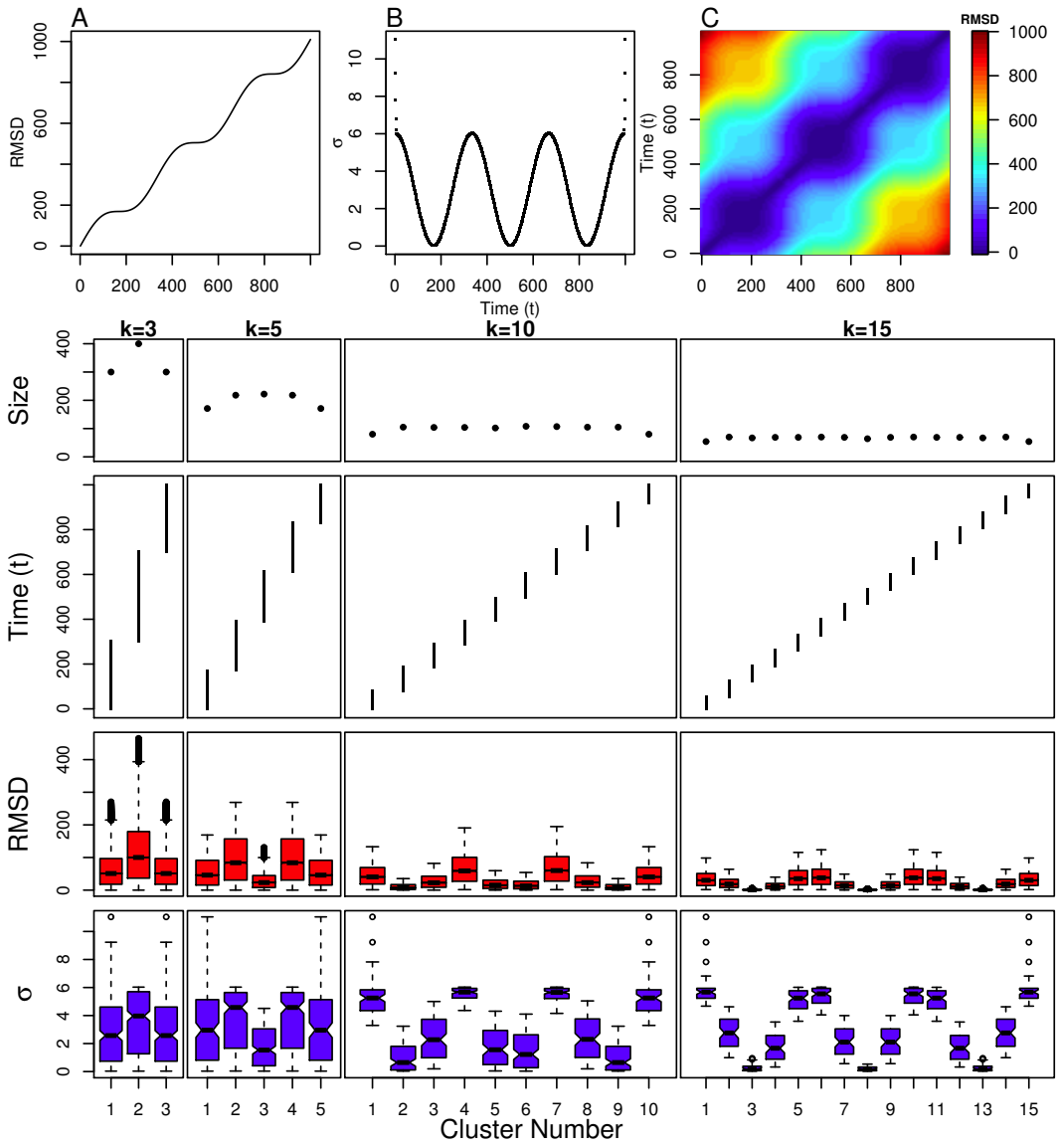

Figure 5.3: Sinusoid model clustering results. Top: (A) Distance from initial structure, (B) local scaling parameters and (C) pairwise distance between all structures. Bottom: Cluster assignment and statistics for several values of k.

numbers of structures. These results are almost identical to those obtained from the linear model. Even the slight end-effects that were observed from the linear model are also replicated, including the sharp increase in the scaling parameters at the beginning and end of the simulation (see Figure 5.3B).

However, Figure 5.3 shows clear differences between the sinusoid model and the linear model in terms of the intra-cluster RMSD and intra-cluster scaling parameters. The intra-cluster RMSDs and scaling parameter values are quite low for the metastable states. In particular, for the $k = 15$ case, clusters 3, 8, and 13 are in the center of the metastable states and the distribution of scaling parameters for these three clusters indicates that these structures are in a densely populated region of structure space. *Therefore, it is stipulated here that a metastable state is described by clusters with low intra-cluster RMSD and low scaling parameter values.* The transition states can also be discerned from these statistics. The $k = 10$ case indicates that the structures in clusters 4 and 7 have large scaling parameter values. *Therefore, it is also stipulated here that large values are indicative of a sparsely populated region of the structure space or a transition state.*

These results are consistent across both the RMSD distributions and scaling parameters, but the results are more evident from the scaling parameters than the RMSD distributions. For example, the distribution of scaling parameters is narrowly distributed around the median for both the metastable and transition state clusters. This is not true for the RMSD distributions, where the transition state clusters have RMSD distributions that are widely distributed around their medians. A large RMSD distribution might indicate that more clusters are needed (higher k) to properly partition the region covered by the corresponding cluster, and such a distribution cannot guarantee that a cluster is not a mixture of transition and metastable states. Therefore, the scaling parameter distribution of a cluster provides better evidence of whether that cluster belongs to a metastable state, transition state, or something in-between. Examples of these in-between clusters are 2, 4, 7, 9, 12, and 14 for the $k = 15$ case.

5.3.3 Rotation Model

The results above correspond to analytic models in which the inter-structure distances are specified directly. Now, a polymer model where RMSD is used to calculate the distances between generated structures is studied. The use of RMSD presents challenges for clustering based analysis. While RMSD is reported to be quite sensitive to small structural differences and, therefore, performs well for distinguishing between structures which are similar, it is less effective for comparing structures with relatively large structural variation. Development of new approaches for structural comparison is an active area of research, and a thorough comparison of these techniques is beyond the scope of the work here. Nonetheless, it is important that clustering-based analysis be as robust as possible to deficiencies in the underlying structural comparison whether it be RMSD or another method.

The rotation model was utilized to determine the effect of the RMSD structural comparison metric on clustering performance. The model consists of a set of consecutive links, each approximately 3.88Å long, analogous to the C_α trace of a protein. Steric exclusion is not considered in this model and the links may overlap with one another without penalty. The angle between successive links is governed by the polar angle (ϕ) and azimuthal angle (θ) which range from $[0, 2\pi)$ and $[0, \pi)$, respectively. These two angles are initially set to 0 degrees, resulting in a fully extended chain. The angles are then incremented on each time step by a small amount (2ϵ and ϵ) until the chain completely winds into a tight helical configuration.

This linear walk through conformation space clearly highlights the nonlinear effects of RMSD. Figures 5.4A and 5.4C show the RMSD from the initial (extended) structure as a function of time and the pairwise RMSD between all structures in the trajectory. Comparisons between (early) extended conformations result in relatively high similarity as compared to (later) collapsed structures which differ by the same distance in time and conformation angle space. Also, the most collapsed, tightly wound structures exhibit a slight bias to consider most intermediately collapsed conformations

to be equally similar even though more extended conformations, separated in time and conformation angle space by the same amounts as the intermediate conformations, are considered to be quite dissimilar.

Example structures are shown in Figure 5.5A for $t = 330$ and $t = 660$ which correspond to unnaturally extended and collapsed configurations respectively. Structures along the helical continuum that correspond to physiologically realizable biopolymers lay approximately between $t = 400$ and $t = 600$. It should be noted that this region is still in danger of improper clustering due to RMSD bias, as indicated by Figure 5.4C, where the pairwise RMSDs to structures from earlier portions of trajectory are still very small. Therefore, the rotation model confirms the observation that RMSD does not possess the ability to discriminate effectively between certain kinds of structures.

Spectral clustering is able to effectively overcome these problems in two specific ways. First, while RMSD is unable to discriminate between extended structures effectively, the metastable states of biopolymers would not typically be composed of extended structures. Second, spectral clustering utilizes the distribution of structures in localized regions to determine cluster membership, as illustrated by the Gaussian kernel employed to transform RMSD into a similarity metric (see step 5 of the algorithm in Section 3.2.1). Even if a biopolymer richly sampled extended conformations, as might be the case for highly disordered systems, only those structures closest in structural similarity would be considered by the algorithm. Therefore, large and mid-range RMSD differences that might bias many clustering algorithms will simply be ignored by spectral clustering, effectively mitigating any problems that result from the RMSD bias.

Upon applying spectral clustering, it is clear that the algorithm is only mildly sensitive to the nonlinear effects of RMSD. Figure 5.4 shows that spectral clustering divides this simulation into clusters of temporally contiguous structures and that these clusters contain similar numbers of structures. This follows the same trend as the linear and sinusoid models, which is encouraging since this model also exhibits a property shared with these models of always progressing into new areas of structure space. The

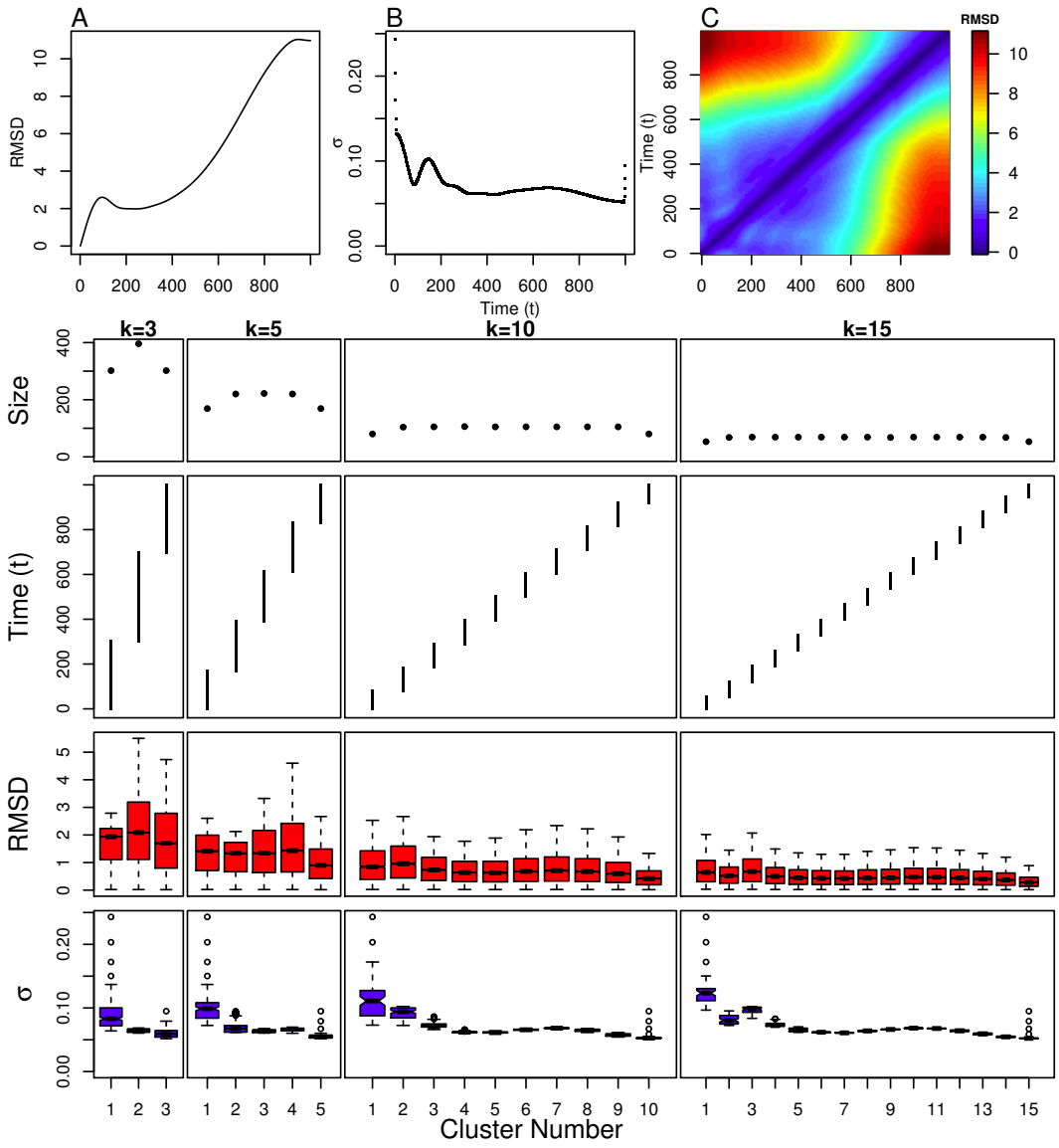

Figure 5.4: Rotation model clustering results.Top: (A) Distance from initial structure, (B) local scaling parameters and (C) pairwise distance between all structures. Bottom: Cluster assignment and statistics for several values of k.

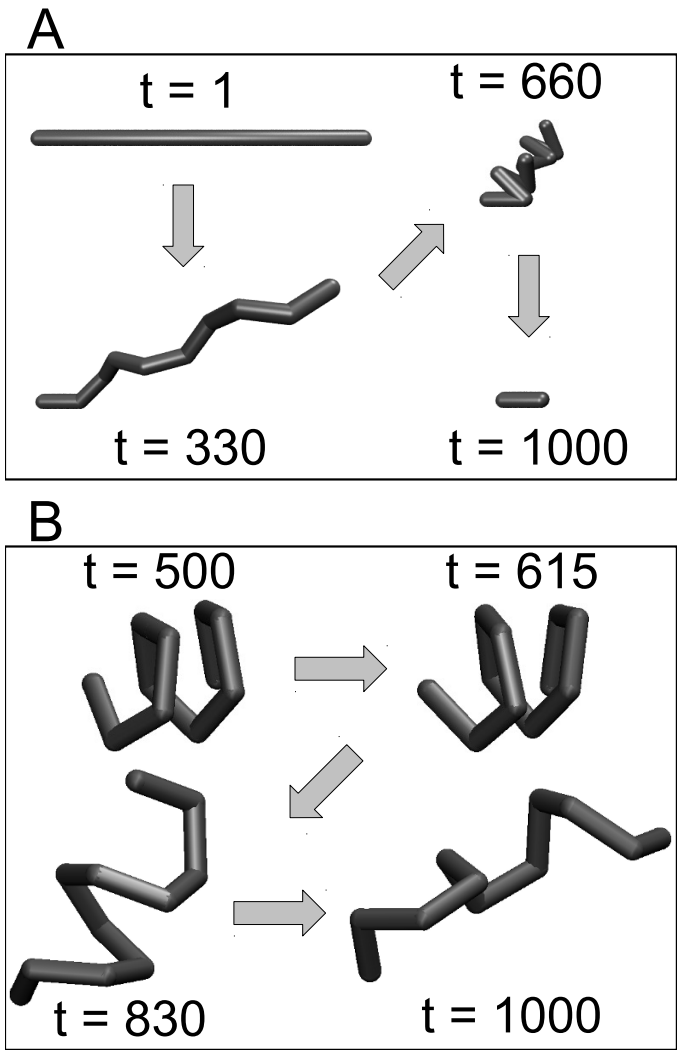

Figure 5.5: Sampled structures from two polymer models. (A) The Rotation Model displays a controlled collapse from a completely extended polymer ($t = 1$) to a tightly wound helix ($t = 1000$). (B) The Dynamic Model is a simulated polymer that starts from a random coil configuration and then "folds" into a helix ($t = 500$) as the temperature of the simulations is lowered. The temperature is then raised for the remainder the of the simulation allowing the polymer to "unfold" back to the random coil ($t = 1000$) conformation.

scaling parameters in Figure 5.4B indicate that the bias is strongest for abnormally extended structures before $t = 300$, where the scaling parameter fluctuates quickly over time.

The intra-cluster RMSD plots for this model, shown in Figure 5.4, verify that the structural diversity in the physiologically relevant region (approximately $t = 400$ to $t = 600$) is still quite large even though it consists of approximately only 200 structures. For instance, for the $k = 3$ case, cluster 2 has the broadest distribution of RMSD values. Increasing k confirms that the diversity of structures is at least on par with the remainder of the simulation, so one can be confident that this region provides a good representation of spectral clustering performance for helical structures.

The intra-cluster scaling parameter distributions, shown in Figure 5.4, make it clear that this region is largely unaffected by any RMSD bias. For the $k = 3$ case, cluster 1 covers the region of extended structures, cluster 2 covers the region of intermediate structures, and cluster 3 covers the region of collapsed structures. Only cluster 1 shows an appreciable bias, which is indicated by the large spread in the intra-cluster scaling parameter distribution. Cluster 3, shows a slight bias as well. However cluster 2 shows almost no bias at all, with a very tight distribution around the median, similar to the results for the linear model. These results are also maintained across the $k = 5$, 10, and 15 cases, where the clusters in the central, physically realizable region show little spread in their intra-cluster scaling parameters distributions. Instead, the bias becomes only mildly evident for the physiologically abnormal structures at both ends of the trajectory.

The potential problems observed from using RMSD on the most extended structures in the trajectory are effectively overcome by spectral clustering. This can be observed from the results for the rotation model, where the structural diversity and total number of structures for clusters in the middle, most relevant portion of the trajectory are on par with the remaining clusters. However, unlike the remaining clusters, the middle clusters did not show any appreciable bias due to the use of RMSD. *Therefore, the conclusion here is that the ability of spectral clustering to utilize localized regions of structure space, and ignore more distant regions and structures, can overcome the*

known problems with using RMSD to compare conformations.

5.3.4 Cyclical Model

The results above were all gathered for models where the simulation is always progressing to new areas of structure space, but biopolymers often do not exhibit this behavior. Instead, most biopolymers, especially highly disordered systems, will revisit certain areas of structure space. The cyclical model is intended to model this process by building on top of the rotation model. This model simply undergoes the same linear change in angle space as the rotation model, but the polymer is reextended following collapse. This process is repeated three times in order to revisit the same region of structure space during the course of the simulation. This revisiting of earlier regions of structure space can be observed in the RMSD from the initial structure as a function of time shown in Figure 5.6A and the pairwise RMSD for all structures shown in Figure 5.6C.

During the initial collapse of the polymer, the clusters are temporally contiguous and contain approximately the same number of structures, as shown in Figure 5.6. This is true for all values of k. These clusters are then re-visited in reverse order during the subsequent phase where the polymer returns to an extended state. The same pattern is observed for the remaining two collapse-extend cycles.

The intra-cluster RMSD distributions shown in Figure 5.6 indicate that the important set of structures identified from the rotation model maintain the same properties as in the cyclical model. The most structurally diverse cluster (the one with the broadest RMSD distribution) is number 2 for the $k = 3$ case. This cluster occurs in the region of the trajectory that corresponds to the physiologically relevant region that the cyclical model shares with the rotation model. Clusters 2 and 3 for the $k = 5$ case are also found in this region, and have the largest structural diversity as well. The effect is less clear for the $k = 5$ and $k = 15$ case, because the clusters covering regions in the fully collapsed state are also highly insensitive to RMSD. Again, this result is consistent with

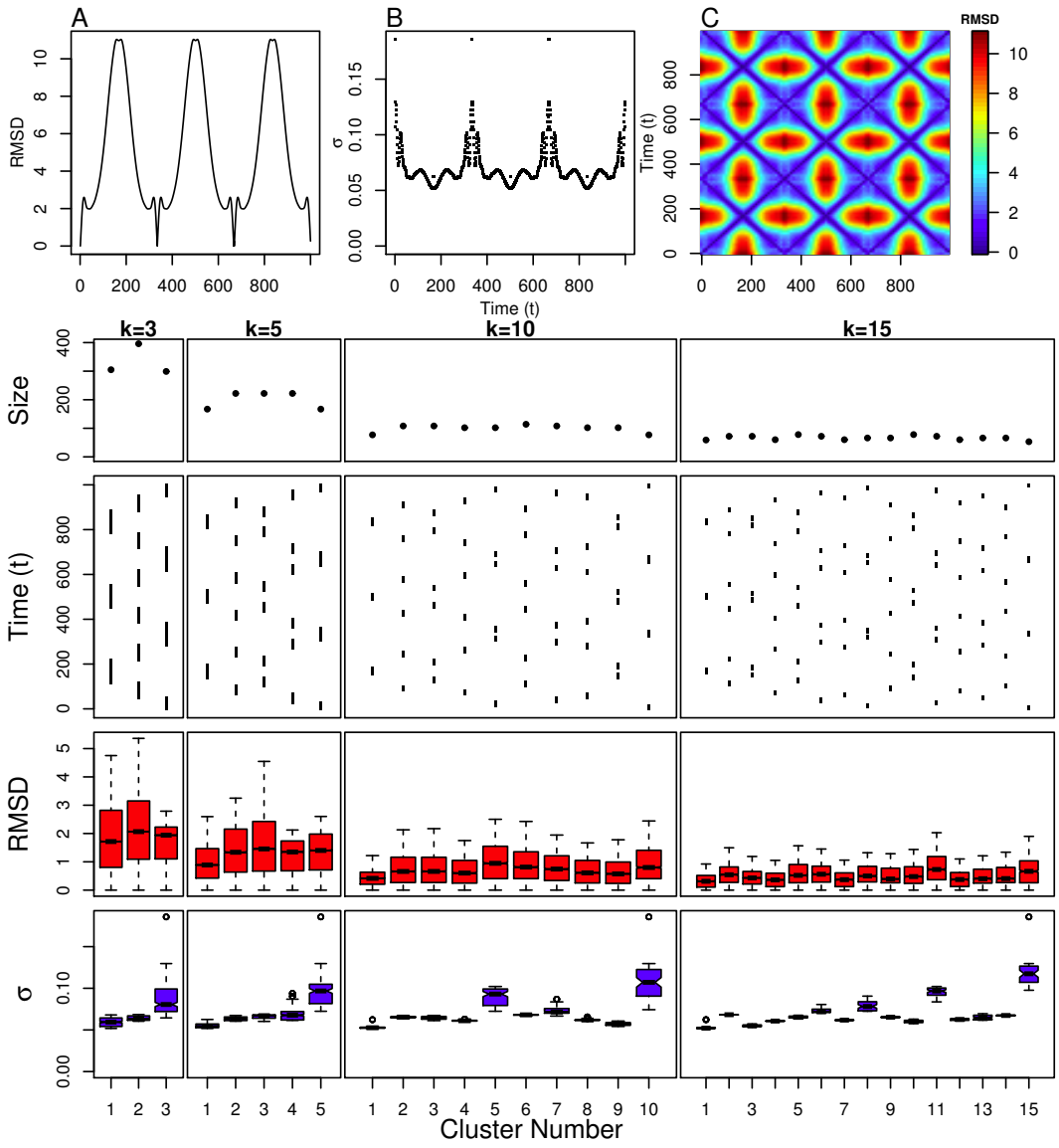

Figure 5.6: Cyclical model clustering results. Top: (A) Distance from initial structure, (B) local scaling parameters and (C) pairwise distance between all structures. Bottom: Cluster assignment and statistics for several values of k.

the rotation model, and can be verified by observing the smoother changes in the structural scaling parameters in both of these regions compared to the extended regions (see Figure 5.6B.)

The intra-cluster scaling parameter values in Figure 5.6 confirm these results as well. Cluster 3 for the $k = 3$ case has the broadest distribution of scaling parameter values and covers the structurally extended regions of the trajectory. Clusters 4 and 5 do likewise for the $k = 5$ case, as do clusters 5, 7, and 10 for the $k = 10$ case, and clusters 8, 11, and 15 for the $k = 15$ case. Clusters 6 and 13 for the $k = 15$ case are also slightly broadened, and are located in regions temporally and structurally adjacent to the extended regions. Cluster 2 for the $k = 3$ case and clusters 2 and 3 for the $k = 5$ case, all have narrow scaling parameter distributions and cover regions corresponding to the intermediate helical structures. For the $k = 10$ and $k = 15$ cases, clusters not covering the extended regions (listed above) have relatively narrow scaling parameter distributions, indicative of relatively little RMSD bias even for extremely collapsed regions.

5.3.5 Dynamic Model

The above models are completely deterministic. Therefore, a dynamic model is now considered which also repeatedly transitions between fully extended and fully collapsed configurations but whose dynamics are stochastic like MD simulations of biopolymers. A simple potential-energy function is utilized which favors a particular orientation of the ϕ and θ angles, combined with a soft-core pairwise repulsive interaction so that the lowest-energy conformation is a helical structure. A temperature bath is applied to the system and we anneal the temperature over time to produce a simulation that initially models an extended coil at high temperature which then "folds" into the final helical conformation at low temperature. As long as the temperature is annealed slowly, the system reliably folds into the native helix conformation. The annealing schedule is then reversed to "unfold" the polymer, allowing it to return to the extended

state. Figure 5.5B shows example structures at different points in time from this second phase of the annealing process.

The simulation exhibits one clearly defined metastable state: the helical conformation that (by construction) is present in the middle of the trajectory ($t \approx 500$). Figure 5.7A confirms that this state is reached as the RMSD from the native state approaches zero at $t \approx 500$. However, it is also the case that the initial, folding transition is much more gradual than the unfolding transition. While there is a slightly abrupt structural transition at $t \approx 180$, the remaining portion of this folding transition smoothly approaches the folded states. This transition occurs because the forces exerted by the potential function begin to overcome the effect of the temperature bath at this point in the simulation, but the soft-core interactions still allow the helix to be quite flexible and dynamic, similar to a weak spring. The unfolding transition does not display this folding intermediate, but instead abruptly shifts from a collapsed coil to an extended coil at $t \approx 800$. The pairwise RMSD for the simulation shown in Figure 5.7C, and the structure scaling parameters shown in Figure 5.7B also confirm this pattern.

Spectral clustering clearly identifies the metastable, folded state of the polymer, and identifies the folding intermediate state as structurally distinct from the folded and unfolded states. The clustering assignment in Figure 5.7 indicates that, as k is increased, the structures associated with the intermediate state segregate into separate clusters. At $k = 3$, cluster 3 covers the extended state, cluster 2 covers the folded state, and cluster 1 covers intermediate structures for both folding and unfolding. However, at $k = 5$, cluster 1 populates the region of the folding intermediate but is not well-populated by structures from the unfolding portion of the simulation. By increasing k to 15, clusters 2, 3, and 4 are almost exclusively populated by the folding intermediate. Clusters assigned to the folded state become slightly more populated (with more total structures) than the intermediate states with increasing k, as shown in Figure 5.7. For $k = 3$ the cluster assigned to the folded state, cluster 2, was the least populated state. However, the population of the folded state cluster, 3 for $k = 5$, was above the intermediate state cluster (2 and 4) populations. The same trend is observed for the $k = 10$ and $k = 15$

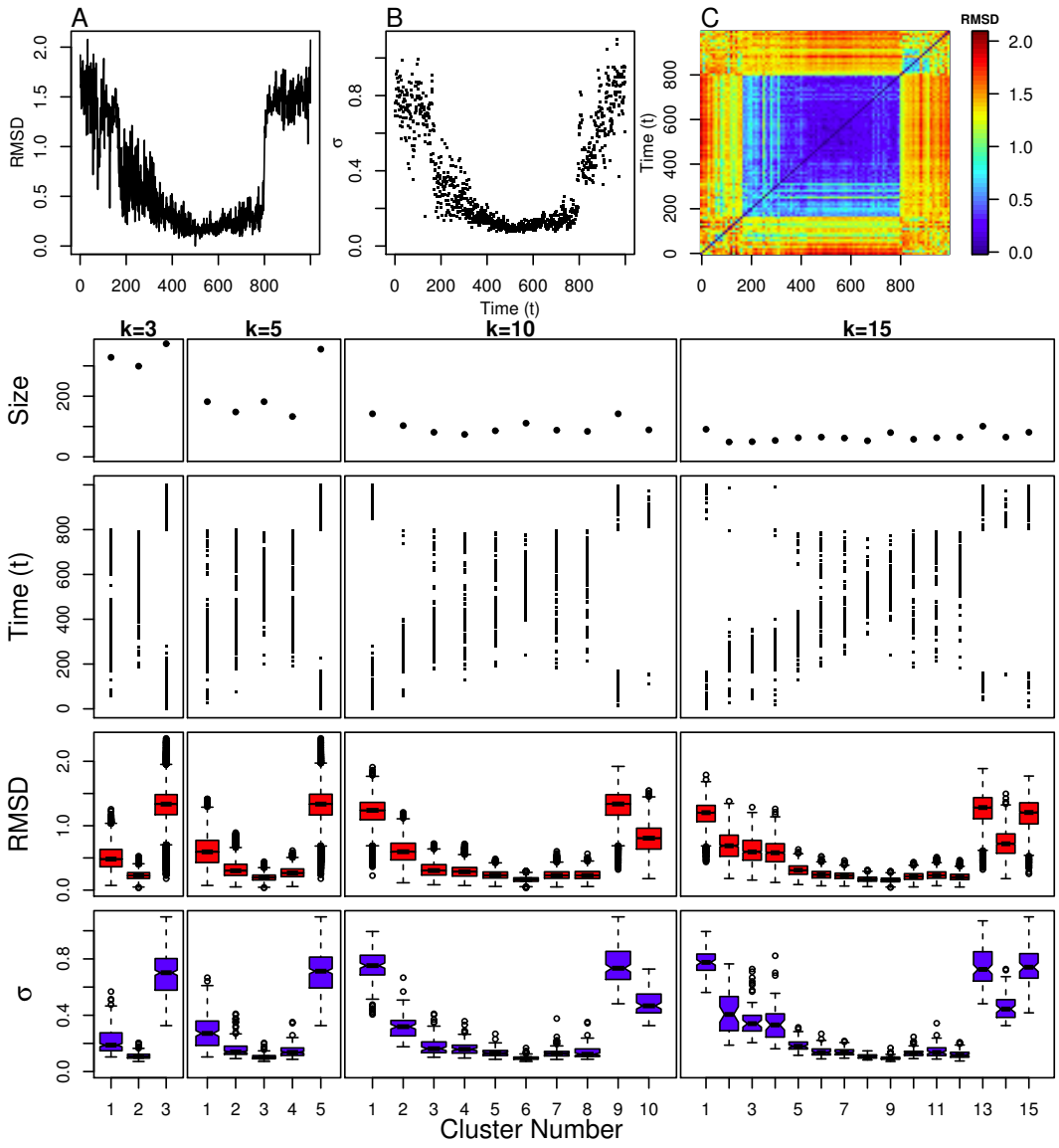

Figure 5.7: Dynamic model clustering results. Top: (A) Distance from folded structure, (B) local scaling parameters and (C) pairwise distance between all structures. Bottom: Cluster assignment and statistics for several values of k.

cases. More importantly, the intra-cluster scaling parameter distributions in Figure 5.7 indicate that the most structurally homogeneous clusters contain structures in or close to the folded state because the distributions for these clusters are much more narrow than clusters corresponding to extended states. So, these distributions indicate that the number of structures assigned to a cluster is not indicative of whether a cluster corresponds to a metastable state since, for the $k = 15$ case, cluster 9 is just as heavily populated as cluster 15. The same results can be observed in the intra-cluster RMSD distributions.

The transition states are more difficult to observe in this model, but there are indicators of the transition ensembles for the $k = 15$ case in clusters 3 and 14, which both have more narrow distributions than one would expect in the temporal regimes that they cover. Cluster 3 heavily covers the folding intermediate state right at the $t \approx 180$ transition, and cluster 14 covers the extended state just after the abrupt transition at $t \approx 800$. Since this transition is so abrupt, there isn't sufficient sampling to capture the transition within its own cluster. However, a sharp jump in the median scaling parameter values between temporally adjacent clusters, such as between clusters 8 and 9 for $k = 10$ and between clusters 12 and 13 for $k = 15$, is a clear indicator of a significant structural transition. These results are in agreement with the sinusoid model as well since such sharp jumps in the median scaling parameters for temporally adjacent clusters are observed there too, even though the sampling was sufficient to create unique clusters for the transition states in that model as well as the metastable states. Therefore, we can see evidence of the transition states, though these states are not easily identified without combining the results of the cluster assignments and scaling parameters in Figure 5.7.

5.3.6 GLFG Simulation

The clustering protocol is now applied to an 18ns simulation of GLFG, a collapsed-coil FG-nucleoporin. The cluster assignments and scaling parameter distributions for $k = 10$ and $k = 15$ are investigated first, since these were the most informative cases for the polymer models. The smaller values of $k = 3$ and $k = 5$ were also investigated

and were consistent with results for $k = 10$ and $k = 15$, but were not as informative as the results for these larger values of k (a property that was also observed for the polymer models).

GLFG undergoes significant structural changes over the duration of the simulation. The differences between the structures can be difficult to describe based on observations of snapshots of the system, shown in Figure 5.8, since the structures all look equally dissimilar to one-another. However, the RMSD from the initial structure as a function of time shown in Figure 5.9A indicates significant structural divergence. The pairwise RMSD in Figure 5.9C additionally reveals that several metastable regions are present, but the dynamics in some regions ($t \approx 6000$-13000) are quite complex, with the simulation potentially revisiting previously explored regions of conformation space. Scaling parameter values shown in Figure 5.9B, indicate that the local structural density is quite sensitive to these metastable/transition regions.

The cluster assignments in Figure 5.9 indicate that this protein continues to move into new structural regions over time, similar to many of the polymer models. An interesting structural transition occurs at around 6ns, observed in the pairwise RMSD plot where there is a sharp increase in RMSD from structures explored previously in time. This is the only part of the simulation that deviates from this continual structural evolution. In particular, for the $k = 10$ case the simulation begins to explore cluster 7 at around 6ns, a little before settling into cluster 6 for a few nanoseconds. This cluster is revisited again at around 10ns, eventually making the transition to cluster 8 at around 12ns. The same pattern can be observed in the $k = 15$ data, where clusters 9 and 10 more clearly indicate the intermediate transition state between these two metastable states. Another distinct structural transition occurs at around 15ns as well. This final 3ns of the simulation is consistently partitioned into a single cluster for all examined values of k.

The intra-cluster scaling parameter distributions in Figure 5.9 validate these claims where clusters 1, 3, 6, 8, and 10 for $k = 10$ have the lowest median values compared to their temporal neighbors, indicating that these are metastable states. The

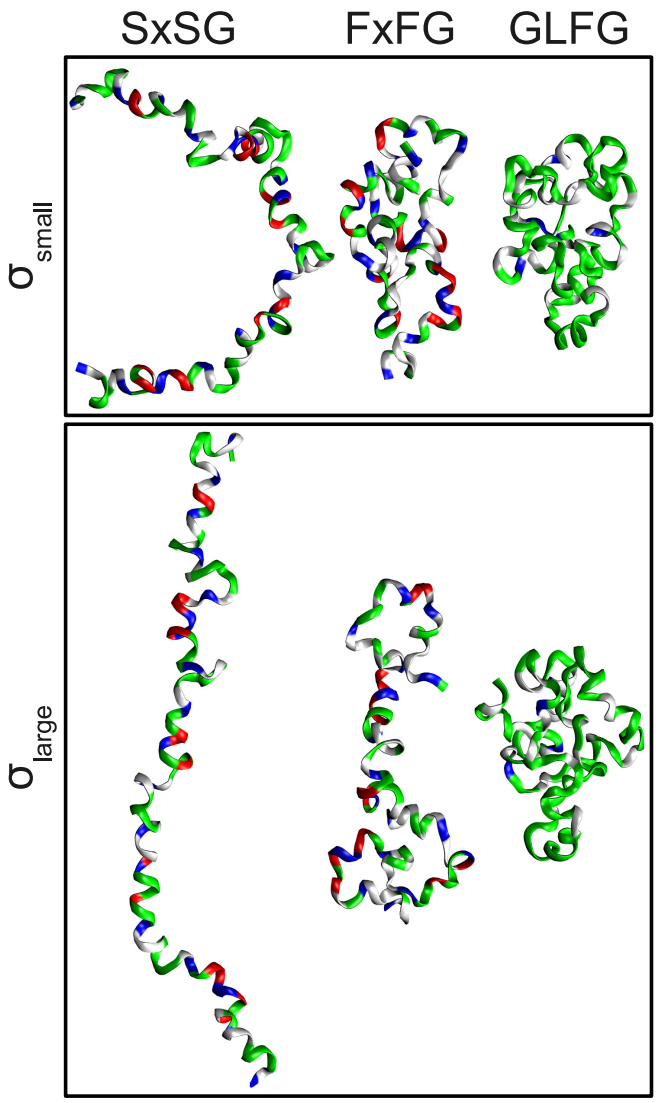

Figure 5.8: Representative structures from the FG-Nup simulations. Structures shown are representatives from clusters with the smallest (σ_{small}) and largest (σ_{large}) median scaling parameters obtained from applying spectral clustering with $k = 15$. Representatives for each cluster were obtained by selecting the structure which had the smallest sum of intracluster distances. These structures clearly show that smaller scaling parameters are associated with collapsed metastable states for all three FG-Nups studied, while large scaling parameters are associated with extended transition states.

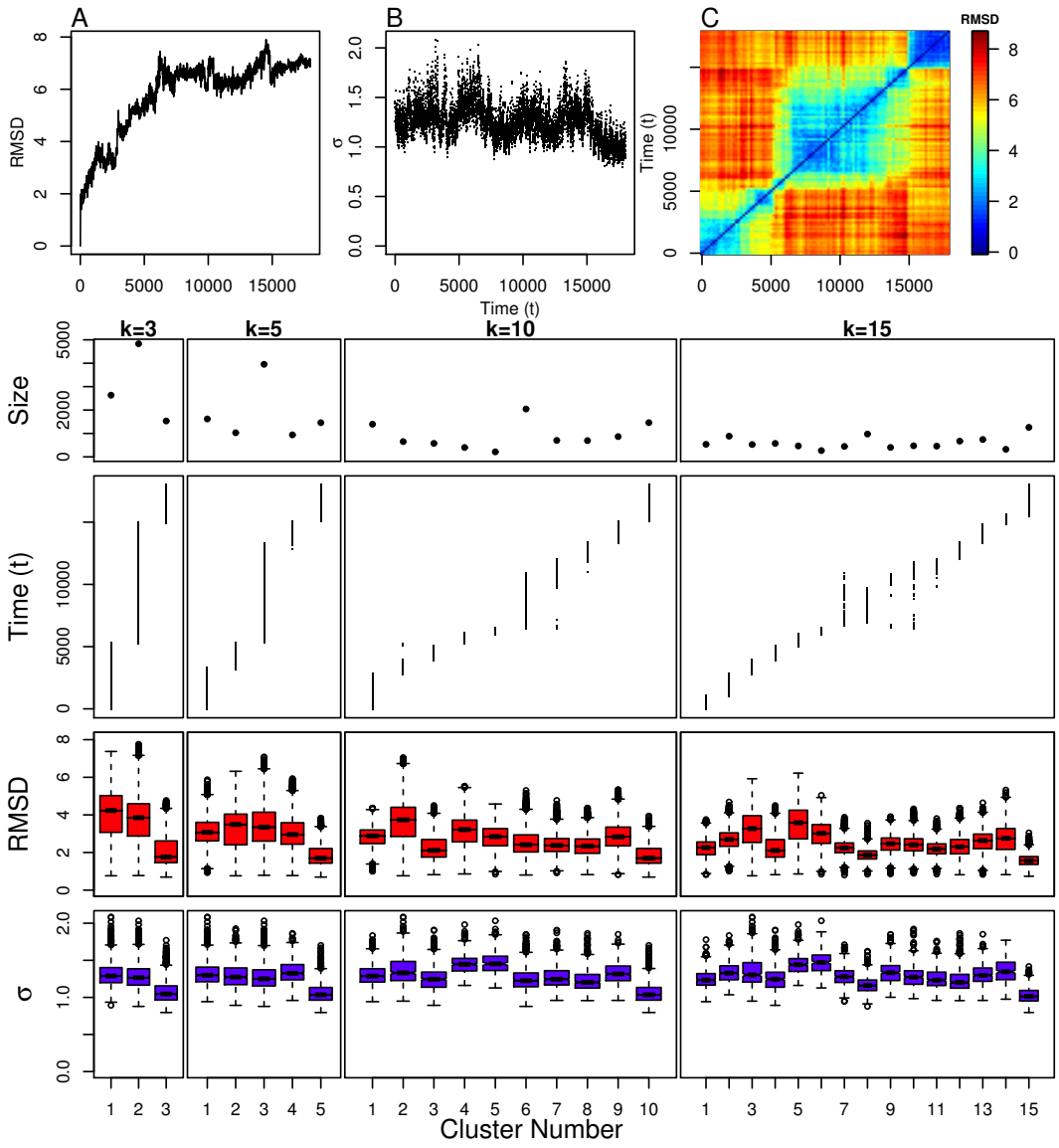

Figure 5.9: GLFG clustering results. Top: (A) Distance from initial structure, (B) local scaling parameters and (C) pairwise distance between all structures. Bottom: Cluster assignment and statistics for several values of k.

same property is observed for the clusters subtending the final 3ns of the simulation across all values of k, indicating that these clusters correspond to a metastable state as well. This is the same pattern observed in the dynamic model where transition and metastable states can be determined by comparing the scaling parameter distributions for clusters that are adjacent in time. The revisited transition state observed in the clustering assignment is explicitly assigned its own clusters (9, 10, and 11) in the $k = 15$ case, and the higher scaling parameters for these three clusters make it clear that this is indeed a transition state. The radius of gyration distributions in Figure 5.10A indicate that two of these clusters (9 and 10) are more extended than the surrounding clusters (8 and 12). However, it is also clear that cluster 11 contains very collapsed structures and is relatively short-lived. Therefore, cluster 11 probably represents a set of collapsed conformations which are energetically unfavorable compared to clusters 8 and 12 which are both more heavily populated.

Overall, the scaling parameters for each cluster are distributed around their medians in a similar manner across all clusters, which is similar to the Linear and Cyclical models, and indicate that the metastable states are representative of shallow minima on the free-energy surface. The values of the scaling parameters are relatively small, indicating that both metastable and transition states are populated with collapsed-coil configurations. The representative structures from the clusters with the highest and lowest median scaling parameters ($k = 15$) shown in Figure 5.8, confirm this result. However, one cluster (11) is composed of highly collapsed structures in terms of radius of gyration (Figure 5.10A) even though it is part of a transition state ensemble based on observations of small shifts in the median scaling parameters of neighboring clusters. Even though these shifts are small, some reasonable statistical confidence in these results is present because the confidence intervals (shown by the notches in the boxplots) between these neighboring clusters are not overlapping.

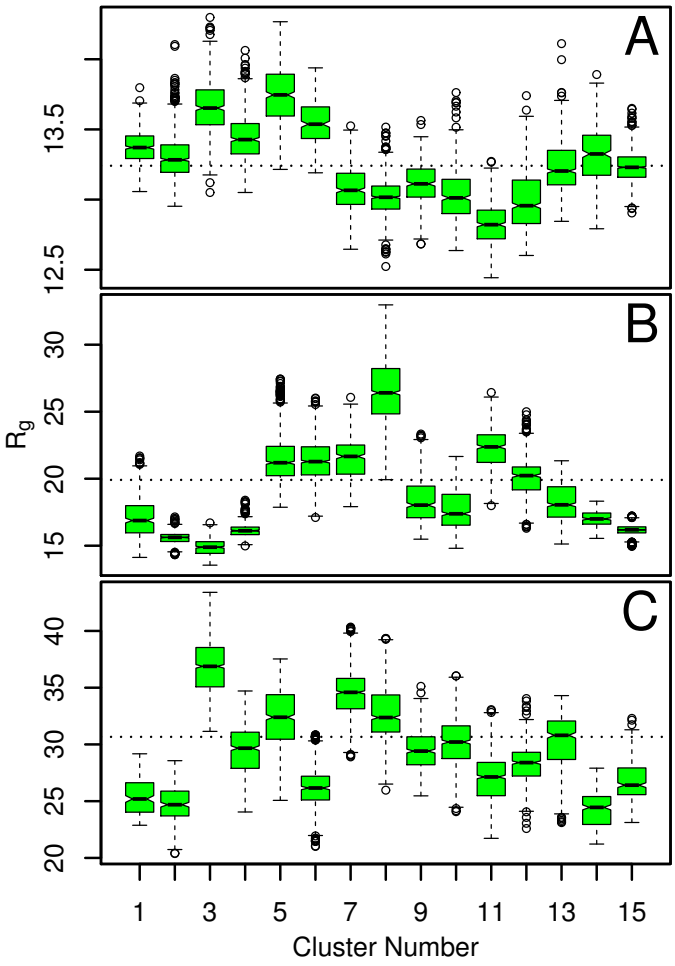

Figure 5.10: Radius of gyration statistics. Distributions of the radius of gyration (R_g) for the FG-Nups (A) GLFG, (B) FxFG and (C) SxSG obtained from MD simulations. Results are shown for each individual cluster obtained from spectral clustering with $k = 15$. A dotted line indicates the mean R_g for the entire simulation.

5.3.7 FxFG Simulation

The FxFG simulation, which undergoes even more significant structural changes than GLFG, was analyzed using the same protocol as the GLFG simulation. The differences between structures are easy to characterize based on observations of snapshots of the system, shown in Figure 5.8. This is primarily because FxFG samples extended conformations that form patterns of long-range contacts that are often discernible from the snapshots. The RMSD from the initial structure as a function of time shown in Figure 5.11A indicates significant structural divergence, even greater than what was observed for GLFG. The pairwise RMSD in Figure 5.11C additionally reveals that several metastable regions are present, but that the dynamics are even more complex than GLFG, with the simulation clearly revisiting previously explored regions of conformation space. The scaling parameters values shown in Figure 5.11B indicate that the local density is different within these metastable/transition regions.

FxFG appears to mostly move into new structural regions over the course of the simulation similar to GLFG, but also seems to revisit previous conformational states more often. The cluster assignments in Figure 5.11 indicate that this is true since for $k = 10$ cluster 5 is heavily revisited during the simulation. Clusters 3, 4, and 7 also possess this property but to a lesser degree. The results for $k = 15$ make this even more clear, with clusters 7, 8, and 11 occupying the same regions in time as the revisited clusters from the $k = 10$ case. However, the cluster assignments alone do not indicate which clusters are potential metastable or transition states.

Again, one needs to consider the differences in the intra-cluster scaling parameter distributions between temporally adjacent clusters in order to characterize clusters as corresponding to metastable or transition states. These distributions are shown in Figure 5.11. The most likely candidates for metastable states for the $k = 10$ case are clusters 2, 7, and 10 due to their low medians. Clusters 2 and 10 both have narrow distributions, clearly indicative of metastable states. However, cluster 7 is not quite as clear because the distribution is broad, opening the possibility that temporally adjacent

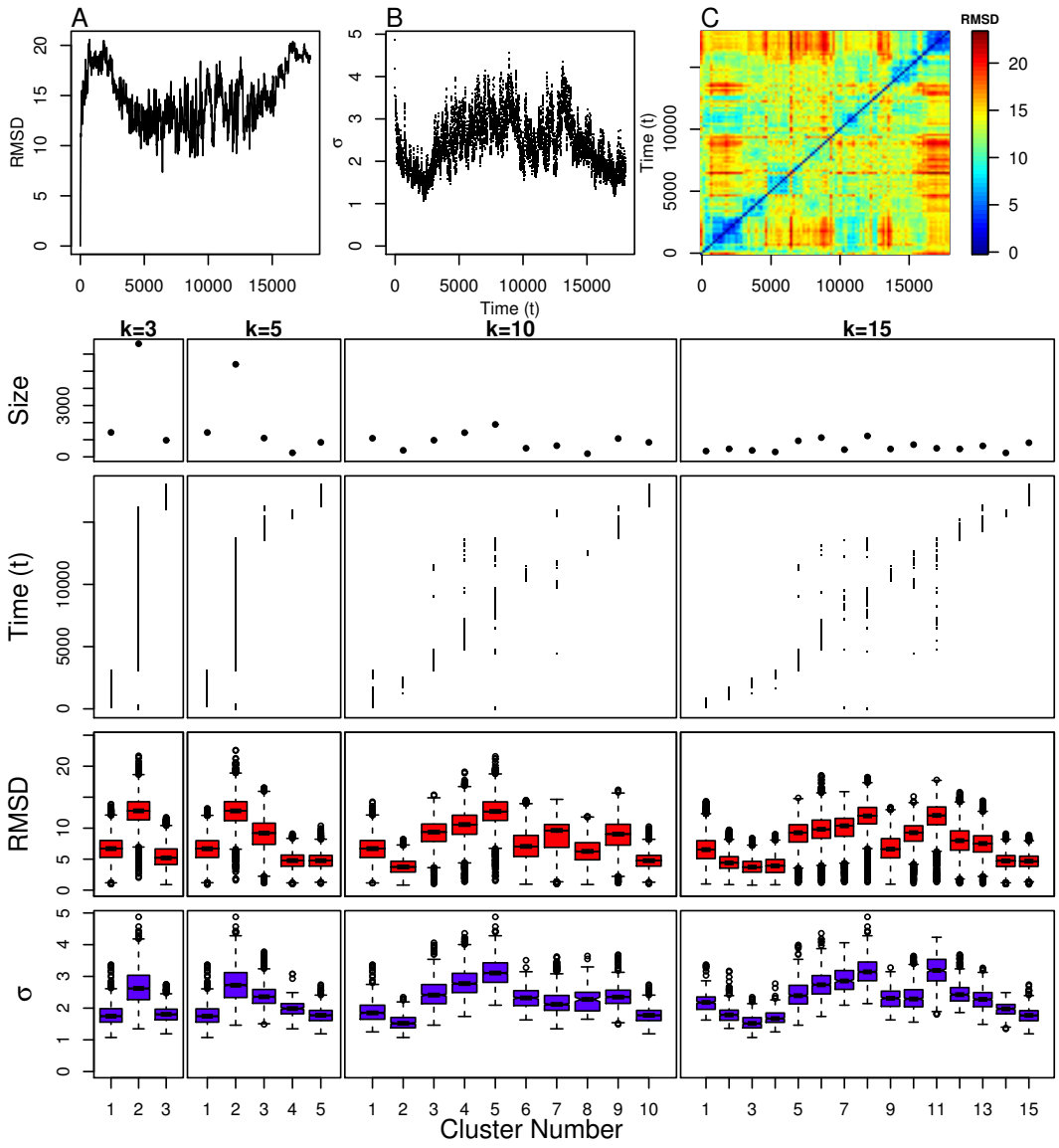

Figure 5.11: FXFG clustering results. Top: (A) Distance from initial structure, (B) local scaling parameters and (C) pairwise distance between all structures. Bottom: Cluster assignment and statistics for several values of k.

clusters 6, 8, and possibly even 9 could also describe this metastable state. The results for $k = 15$ resolve this ambiguity by splitting this region into two different clusters, 10 and 11. The sharp increase in the median, and the broad distribution for cluster 11 indicate that this region corresponds to a transition state, and that cluster 10 is a preliminary move towards this transition. Instead, cluster 9 with its low median, and narrow distribution, displays all of the properties of a metastable state in this regime. These results are congruent with the analysis for the dynamic model where adequate sampling combined with results for various values of k is needed in order to begin extracting transition states that occupy their own distinct clusters. The structures in Figure 5.8 indicate that more extended conformations are often associated with larger scaling parameters, and a more thorough comparison with the cluster radius of gyration ($\mathrm{R_g}$) distributions in Figure 5.10B indicates that this is definitely the case for this protein.

5.3.8 SxSG Simulation

Finally, the simulation of SxSG is examined, which is even more flexible than the wild type FxFG. This is clearly seen in Figure 5.12 where the RMSD value from the initial structure quickly diverges and levels off. This indicates that this simulation is devoid of metastable states. The pairwise RMSD values in Figure 5.12C indicate that there is not only a wide variation in the structural ensemble, but that it is difficult to identify when particular structural regions are revisited. The scaling parameters shown in Figure 5.12B vary consistently over time in an almost cyclical manner. This could indicate rapid transitions into and out of metastable states, but we need to look at the clustering assignments to know this for certain.

The clustering assignments for SxSG are shown in Figure 5.12. The $k = 10$ case indicates that the simulation is devoid of any metastable states since almost any chosen 1ns time window from the simulation spans all 10 clusters. The $k = 15$ case slightly diverges from this result in that clusters 1, 2, 14 and 15 are more sparsely populated. However, when comparing the scaling parameter distributions for these clusters in Fig-

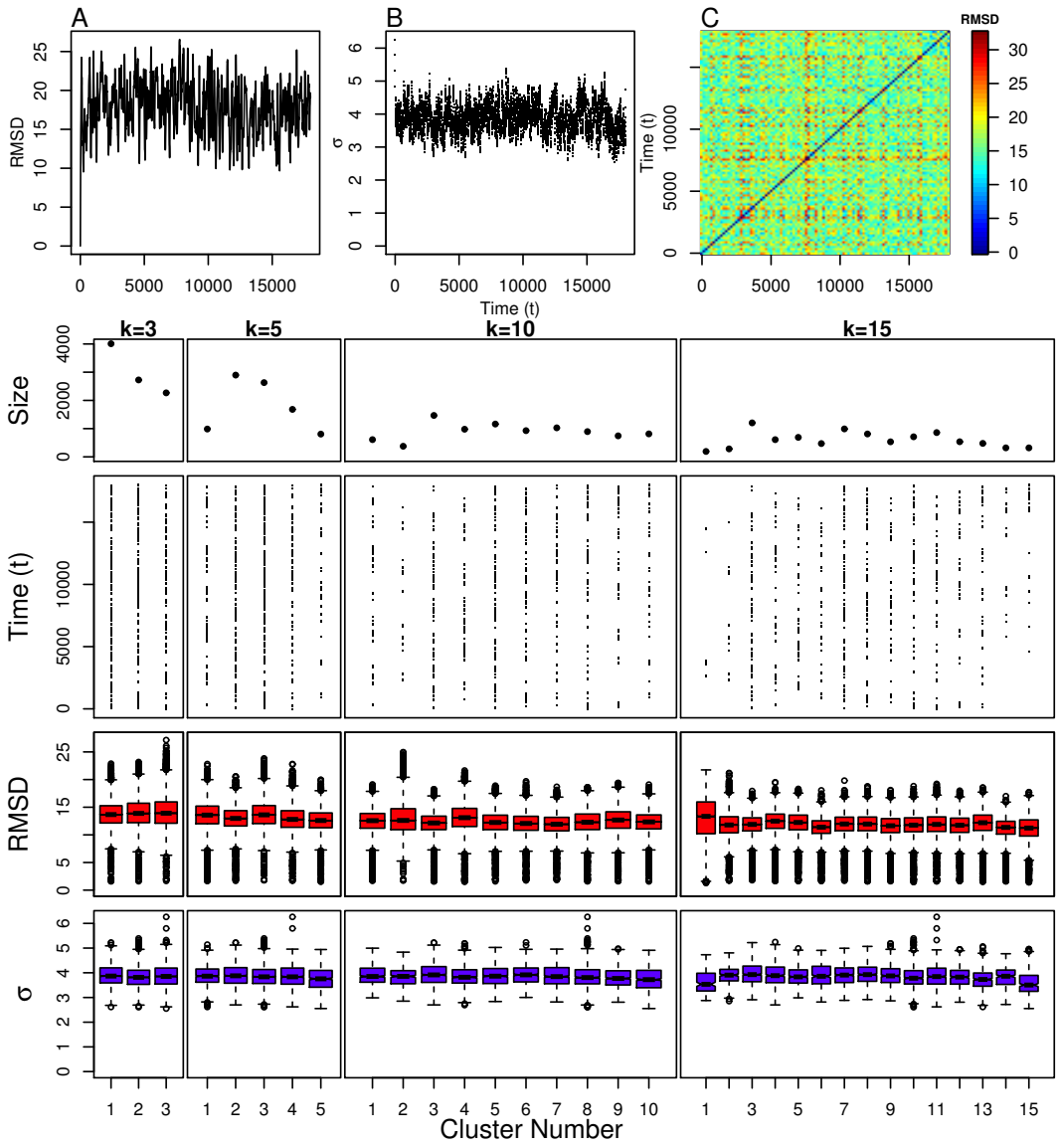

Figure 5.12: SXSG clustering results. Top: (A) Distance from initial structure, (B) local scaling parameters and (C) pairwise distance between all structures. Bottom: Cluster assignment and statistics for several values of k.

ure 5.12, it is clear that these sparsely populated clusters differ only slightly from the remaining clusters. The radius of gyration distributions in Figure 5.10C indicate that these clusters consist of the most compact conformations visited by the simulation. This result is also not due the end-effects like those observed in the linear model because none of these clusters is heavily populated by structures at the beginning or end of the simulation. The large overall values of the scaling parameters indicate that all clusters are consistent with extended coil conformations.

5.4 Conclusions

A framework is developed here for validating the performance and utility of clustering algorithms for studying molecular biopolymer simulations. The key contribution of this framework is the development and use of several analytic and dynamic polymer models which exhibit well-behaved dynamics including: metastable states, transition states, helical structures, and stochastic dynamics. These models provide an informative framework for testing the ability of spectral clustering, a promising clustering algorithm that has received much attention recently in the machine learning community, to partition the polymer model structural ensembles into clusters whose statistical properties reveal the underlying metastable and transition state ensembles. The models have also been used to address potential problems that arise due to RMSD bias and shown there is little adverse effect for spectral clustering. In all of the polymer models, spectral clustering found clusters that corresponded to metastable states, most clearly recognized by comparing the distributions of intra-cluster similarity scaling parameters, σ, between temporally adjacent clusters. Transition states were sometimes not assigned to clusters due to the sparse sampling of these states in the ensembles.

These methods are also utilized to determine the metastable and transition states for simulations of several FG-Nups, and found that the statistical properties of the resulting clusters allowed similar comparisons and predictions to be made for these systems as well. The metastable states could often be predicted quite easily, while the transi-

tion states were again somewhat difficult to determine due to under-sampling. While experimental data for these proteins at the level of detail needed for direct comparison is not available, the results for the three proteins studied here are in agreement with past experimental and computational studies on these proteins [8, 21]. In particular, GLFG is a collapsed coil that slowly explores the free-energy landscape by climbing relatively small barriers between shallow metastable states. FxFG is an extended coil that often revisits previously explored collapsed metastable states and utilizes extended conformations to transition between these states. SxSG is an extended coil that never explores collapsed conformations.

Clustering has been widely used to partition structural ensembles obtained from MD simulations, but few studies have been performed to rigorously determine the utility of various clustering methods for studying MD simulations. The framework provides a novel approach to address this concern that is computationally efficient and highly predictive of success or failure for individual algorithms. While most of the polymer models in this study focused on unfolded and helical conformations, novel polymer models might be developed in the future for assessing simulations involving loop and sheet conformations as well. The framework could also be used to compare different clustering algorithms to better understand their relative strengths and weaknesses. Finally, it is hoped that these results can be brought to bear on simulations of previously unstudied biopolymer systems where predictions can be made concerning metastable and transition states that can be subsequently verified using experimental techniques.

Chapter 6

Validation of Dimensionality Estimators for Protein Simulations Using Polymer Models

Chapter 4 pointed out that dimensionality reduction techniques offer a unique way to visualize the conformation landscape, and can even be used to visually confirm the convergence of a set of independent replicates which corroborates the results of the clustering methodology discussed in Chapter 3. However, the true dimensionality of the space traversed by disordered proteins is often of much higher dimensionality than the two or three dimensions that can be visualized using these techniques. Hence, a more useful measure of the complexity of the protein's motion would be the dimensionality of the conformation space. To this end, various dimensionality estimation techniques have been developed which attempt to address this very problem, but their utility for protein simulation studies has not been validated. The previous chapter introduced the idea of using polymer models for validating clustering methodologies for MD simulations. This method can readily be extended to validate other tools for studying MD simulations which operate on protein structural dynamics

In this chapter, a polymer-based framework similar to the one proposed in the

previous chapter is developed to validate the use of dimensionality estimation techniques for the study of unfolded protein dynamics. First, several techniques in the machine learning and mathematical literature for estimating intrinsic dimensionality are reviewed. One of these methods is chosen that has seems to perform well. However, some minor changes to the algorithm are required to study some of the simulation properties worth capturing. Next, several polymer models are developed that possess clearly defined dynamics and dimensionality by construction, and are used to validate the dimensionality estimation algorithm. In addition, some techniques for dealing with inherent difficulties in the data, noise and relative motions, are developed. The dimensionality estimation algorithm is applied to the polymer models, and the effectiveness of the algorithm is examined in light of the underlying properties of the polymer models. The results of the polymer studies help guide dimensionality analysis of extensive simulations of two IDPs and two unfolded NFPs in order to test the hypothesis that IDP motion should remain of relatively high dimensionality compared to unfolded NFP motion, which should show a decline in dimensionality over time. The proteins and simulation protocol are discussed, as well as the protocol for processing the resulting the trajectories. The dimensionality estimation algorithm and a few other standard metrics from Chapter 2 are applied to the trajectories and the results from the polymer data are used to calibrate the results for the trajectories generated here. The properties of the IDP simulations and NFPs are contrasted based on the results, and the original hypothesis regarding IDP and NFP dynamical differences is revisited.

6.1 Background

Dimensionality reduction has been applied to molecular simulations in many past studies. Given that protein simulations are inherently non-linear systems of high dimensionality, this could be considered somewhat surprising. Indeed, the 2D maps of the conformation landscape produced in most studies exhibit difficulties in mapping very distinct conformations onto different portions of the landscape, making such approaches

limited to extremely small peptides [69].

On the other hand, dimensionality estimation has the potential to be a great aid to dimensionality reduction methods. If the dimensionality of the conformation landscape under study was known, then a projection onto the appropriate number of component dimensions could be constructed which, while not visualizeable, would produce a mapping which would be less likely to encounter the difficulties mentioned above. Also, the dimensionality estimates themselves might be able to say something important about the conformation landscapes in question, allowing different proteins to be classified based on their intrinsic dimensionality.

6.1.1 Dimensionality Estimation

The field of dimensionality estimation has roots in the study of fractals: systems whose inherent properties give rise to the interesting theoretical property where the system lays "in-between" two integral dimension, i.e. *fractional* or *fractal dimensionality*. The Koch snowflake is a classic example of a mathematical curve which has fractal dimensionality. The curve is generated by starting with an equilateral triangle and then applying the following three steps to each segment in a recursive fashion:

1. Divide the segment into three equal-length parts.

2. Draw an equilateral triangle that has the middle part as the base, pointing outwards.

3. Remove the line at the base of the new triangle (middle segment from step 1).

As this process is repeated, it is clear to see that this process will result in a closed loop of line segments that is infinite in length. After each iteration, the number of segments increases by four-fold, each one-third the length of the length of the segments from the previous iteration. Therefore, the total length increases by four-thirds, even though the curve appears to the eye to still be composed of a closed loop because the segments become too small to differentiate without zooming in on some portion of the curve. No

matter how close one zooms in to view the curve, it appears to continuously repeat the same self-similar "snowflake" pattern. Despite having infinite length, the interior area of the curve is finite (eight-fifths the area of the original triangle). The scaling of the length of the curve therefore doesn't follow the same properties of a simple one-dimensional curve, actually giving the curve a fractal dimensionality of $4/\log 3 \approx 1.26$.

Many methods have been developed for calculating the dimensionality of various processes from data samples generated by some system of interest. While the primary systems of interest for these techniques may have been fractals since analytical solutions were difficult or impossible to obtain, interest in the estimation of the dimensionality of any process (whether of fractal or integral dimensionality) via sampling has grown as well [70]. Perhaps the most commonly employed methods are actually methods designed to reduce the dimensionality of data. However, the global solution approach of methods like PCA, multidimensional scaling, and non-linear dimensionality reduction methods often do not work well in practice unless the system of interest exhibits a constant integral dimensionality without noise. Additional techniques based on nearest-neighbors have been developed which utilize various properties of the data to estimate dimensionality. Costa and Hero utilize the scaling properties of graph entropy versus dimensionality [71] and Kegl utilizes an approximate solution to the theoretically accurate *packing dimension* [72]. Some techniques have been developed which, with some small modifications, can be applied to local areas of the system manifold, and noise can be handled somewhat more independently than in the global projection methods mentioned above. For example, the correlation dimension [73] has been extended to use an empirical technique that can be applied locally by leveraging the scaling properties of samples drawn from a uniform hypercube [74]. Levina and Bickel utilize maximum likelihood based on the scaling of nearest-neighbor distances [75]. This list of dimensionality estimation methods is by no means exhaustive, but serves to illustrate the importance of dimensionality estimation and the wide variation in the algorithms and techniques employed.

6.1.2 Protein Dimensionality

The interest in the dimensionality of protein dynamics stems from the intuitive idea that NFPs, when folded, will continually retain the same conformation. In essence, folded protein dynamics exhibit a dimensionality of zero because they subsist in one singular point in the high-dimensional conformation space. If the protein was heated, and parts of the tertiary or secondary structure of the protein become perturbed, then the resulting motions would cause the protein to explore new portions of the conformation space. Importantly, the dimensionality of the motion in this case would increase as the protein fully explores newly accessible regions of the conformation space. The more unfolded the protein becomes, then the higher the dimensionality of the motion. Of course, other transient, non-native structures might form, and, while it might take a while for the protein to explore this new space fully, it is highly improbable that these structural ensembles would be of lower dimensionality than the folded state. In contrast, IDPs would never exhibit zero-dimensional (folded) dynamics since no native structure exists. IDPs which form some regular contacts between residues or unique, transient structures would then naturally be of lower dimensionality to their completely disordered counterparts. Even these collapsed IDPs should exhibit higher dimensionality motion than an NFP of equivalent length in terms of the number of residues. Therefore, dimensionality of protein dynamics could potentially be used to distinguish between different kinds of IDPs or between IDPs and NFPs.

While it is clear that dimensionality estimation could be great interest as a metric for comparing the dynamics of different classes of proteins, there is still a need to determine if dimensionality techniques are actually capable of extracting the dimensionality of the underlying dynamics. In particular, given that there are many different algorithms for estimating the dimensionality of protein dynamics, a framework for assessing and comparing the utility of various dimensionality algorithms is needed. To this end, a novel series of polymer models are developed here which display various dynamic characteristics that vary with dimensionality. Importantly, the dimensionality of

the polymers is known, and can explain the behavior and performance of an algorithm in ways that can be mapped directly to the results obtained from applying the estimation algorithm to MD simulation data. This approach is unique because the polymers effectively bridge the gap between the small systems used to initially calibrate the estimators during their formulation and the large, complex, and inherently nonlinear systems they target, such as MD simulations.

The algorithm of choice for this study will be the maximum likelihood estimator (MLE) of Levina and Bickel [75] with some extensions described in detail in the next section to address some concerns that arise when working with MD data. In particular, as proteins form transient contacts and explore new conformational states, the dimensionality of the motion may fluctuate over time. Therefore, the dimensionality of the conformation manifold in the local region surrounding each structure is calculated instead of averaging the estimators across the entire manifold. In addition, since noisy data is known to be particularly difficult for dimensionality estimators and MD simulations are inherently noisy, trajectory smoothing techniques are also utilized to see if the noise can be mitigated independent of alterations to the estimation algorithm.

6.2 Methods

6.2.1 Maximum Likelihood Estimator of Dimensionality

The maximum likelihood estimator is a powerful methodology for estimating the dimensionality of a dataset [75]. Application of this method results in an estimate of the *intrinsic* (or minimum) number of discrete variables needed to effectively generate or model the data in question, regardless of the analytical or numeric form of the underlying model. The intrinsic dimension can be significantly smaller than the number of dimensions of the space in which the data has been transformed or embedded into, known as the *ambient* space, and the corresponding number of dimensions known as the *ambient* dimensionality. Proper estimation of the intrinsic dimension of a dataset can

allow for more efficient data compression, guide the application dimensionality reduction methods, and provide insight into the functional characteristics of the process that generated the data.

While a number of other techniques utilize geometric techniques similar to those employed by the MLE, the use of a likelihood function makes the estimates quite accurate relative to previous approaches and computationally quite efficient. The calculation of the nearest-neighbors for each data point is first required, and is the computational bottleneck for this approach where an estimate of the dimensionality, $\hat{m}_{R_i}(x)$, for each data point, x, is calculated as follows:

$$\hat{m}_{R_i}(x) = \left[\frac{1}{N(R_i, x)} \sum_{j=1}^{N(R_i,x)} \log \frac{R_i}{T_j(x)} \right]^{-1} \tag{6.1}$$

where $N(R_i, x)$ is the function which calculates the number of points around point x that lay within the surrounding sphere of radius R_i, and $T_j(x)$ is the distance from point x to its jth nearest neighboring point. Another formulation that works well in practice is the following:

$$\hat{m}_{k_i}(x) = \left[\frac{1}{k_i - 1} \sum_{j=1}^{k_i - 1} \log \frac{T_{k_i}(x)}{T_j(x)} \right]^{-1} \tag{6.2}$$

where $\hat{m}_{k_i}(x)$ is the dimensionality estimate for the region surrounding point x, with the k_ith nearest neighboring point used to determine the radius of the sphere, $T_{k_i}(x)$, used in this formulation. In either formulation, the estimates must be recalculated for many different values of the parameters which determine the sizes of the surrounding spheres (R_i or k_i) since there is currently no best-practice method for determining the optimal value, similar to the cutoff radius or k parameters of clustering algorithms. Nevertheless, a final estimate for each point can be determined by averaging over a range of small to moderate values of these parameters:

$$\hat{m}_R(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{R_i}(x) \tag{6.3}$$

or

$$\hat{m}_k(x) = \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{k_i}(x) \tag{6.4}$$

where n is the number of parameter values chosen.

A solution similar to that suggested by MacKay and Ghahramani is adopted here, who suggest to average the inverses of these estimates to obtain better solutions for small R and k [76]:

$$\hat{m}_R(x)^{-1} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N(R_i,x)}\log\frac{R_i}{T_j(x)}}{\sum_{i=1}^{n}N(R_i,x)} \tag{6.5}$$

or

$$\hat{m}_k(x)^{-1} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i-1}\log\frac{T_{k_i}(x)}{T_j(x)}}{\sum_{i=1}^{n}(k_i-1)} \tag{6.6}$$

This technique differs slightly from that of MacKay and Ghahramani because the estimate for each point, x, is obtained by first aggregating over all k_i or R_i, instead of first aggregating over all points. Therefore, the approach taken here presents a local estimator of dimensionality instead of a global one. The harmonic mean of the point estimates can be taken to obtain a global estimate if so desired. This approach is suggested by Haro et al.[62, 63] for manifolds of mixed density and dimensionality, albeit derived independently here. A smoothing approach is taken in the sections that follow, where the harmonic mean of estimates within a sliding time window is used in order to obtain reliable statistics. Additionally, when computing the nearest neighbors between all points, it is often prohibitive to store the entire set of distances, making the radius-based estimator $\hat{m}_R(x)$ difficult to calculate. Instead, the set of k_{max} closest distances is kept in practice, making $\hat{m}_{k_i}(x)$ the more natural choice. For all analyses presented here, $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$, so $k_{max} = 1024$.

6.2.2 Polymer Models

The following polymers are used to test the effectiveness of the dimensionality estimator for working with polymer dynamics. The performance of the estimator algorithm can be used to determine which properties of the polymer chains can be accurately predicted from the dimensionality estimates. If similar properties are observed for dimensionality estimates obtained from MD simulations, then there is at least some confidence that the same dynamical features are present in the simulations that were present in the polymer models.

Semirigid Helix

While several polymer models exist which are suitable for the framework [77], a simple model similar to the freely-jointed chain [78] provides the necessary properties with only one free parameter. This model, called the semirigid helix, consists of a set of l virtual bond segments which are all $a = 3.8$Å in length, which is the typical distance between subsequent $\mathbf{C}_\alpha$ atoms along a protein chain. At the junction between two contiguous links, two angles (θ and ϕ) describe the orientation of the second link relative to the first. The angle θ describes the inclination of a link relative to the prior link, while the angle ϕ describes the azimuthal rotation of the link relative to the prior link.

In this model, the set of links formed by the polymer assume a rigid helix, analogous to the folded protein α-helix. This is accomplished by setting all ϕ angles along the chain to be random values chosen from a Gaussian distribution with $\mu_\phi = 0.83$ and standard deviation $\sigma_\phi = 0$, and all θ angles chosen from a Gaussian distribution with mean $\mu_\theta = 1.54$ and standard deviation $\sigma_\theta = 0$. By increasing the standard deviation, ensembles can be generated which model the fluctuation of the polymer around the average, folded conformation. This model can be used to gauge the effect of noise on the estimator in a systematic and physically meaningful way. An example polymer ensemble for this model is shown in Figure 6.1A.

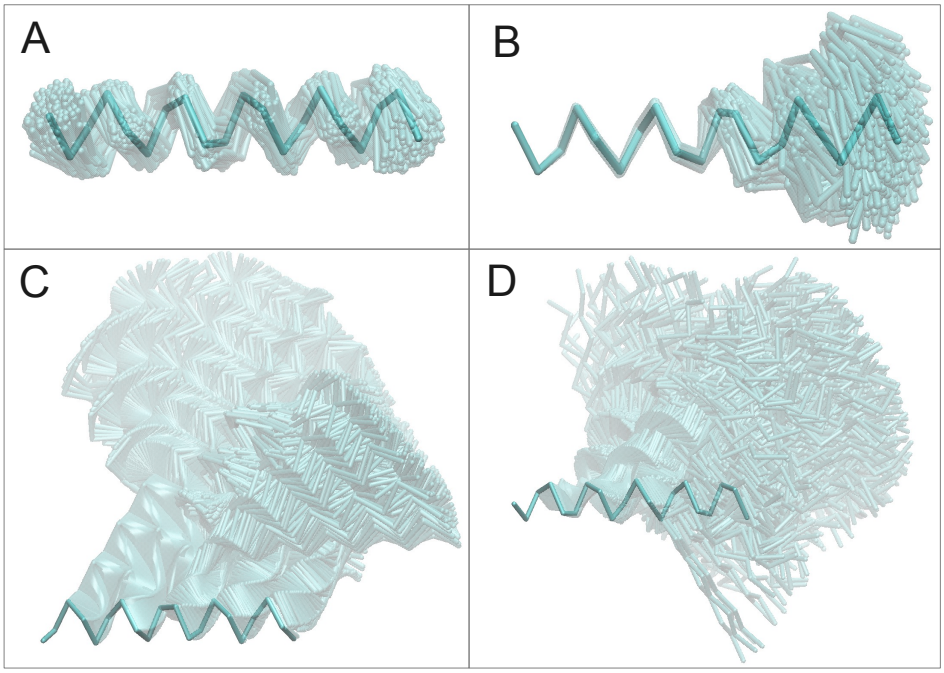

Figure 6.1: Structural ensembles from three polymer models. Two thousand semi-transparent structures from three different polymer models are shown overlaying the solid, fully-folded helix structure for (A) the semirigid helix model with $\sigma_\theta = 0.1$, (B) the half-folded helix model with $\sigma_\theta = 0.01$ and $\sigma_{\theta_{unfolded}} = 0.1$, (C) the correlated helix model with two correlated, folding motions, and (D) the correlated helix model with three correlated, folding motions.

Half-folded Helix

While the semirigid helix models the effects of noise across the entire folded conformation, it is clear that proteins do not fold/unfold in such a global manner. Instead, one portion of the helix may fold/unfold while the other portion remains folded, and the half-folded helix model is designed to investigate this effect on the estimation of dimensionality. This model is formed by first dividing the polymer into two contiguous segments, and applying the standard deviation, $\sigma_{\theta_{unfolded}}$, to the first segment, and then fixing the second segment with $\theta_{folded} = 1.54$. Both segments have a fixed angle $\phi_{polymer} = 0.83$, but zero-mean noise with a standard deviation of 0.01 is then injected into the system at all times, so the structure is not completely rigid. Therefore, by setting $\sigma_{\theta_{unfolded}} = 0.0$, this model becomes equivalent to the semirigid helix model with $\sigma_\theta = \sigma_\phi = 0.01$. However, as $\sigma_{\theta_{unfolded}}$ is increased, the second segment retains a helix-like conformation with minimal noise while the first segment becomes progressively more disordered. Thus, the effects on the dimensionality estimator due to a different number of significant dimensions (or mixed levels of noise) can be investigated using this model. An example polymer ensemble for this model is shown in Figure 6.1B.

Correlated Helix

While the previous two models investigate several variables that impact estimator performance, they both consist of a mean helix-like structure which represents the folded state. However, it is commonly thought that proteins often exhibit coordinated motions, where several parts of the chain would be moving in response to the motions along other parts of the chain. This would naturally have an effect on the underlying dimensionality estimation, especially if such motions were of large amplitude relative to general noise. Methods such as normal mode analysis are effective for finding these coordinated motions for folded proteins [37], and therefore extending this idea to IDPs is desirable. The correlated helix model is designed to effectively validate the performance of dimensionality estimators for the purpose of estimating the total number of

such motions.

The general idea behind the correlated helix model is to generate a polymer trajectory which exhibits coordinated folding and unfolding, i.e. from a helical conformation to a nearly rigid rod, and then in reverse. This is accomplished starting with all angles set so that the polymer is in a helical conformation ($\phi = 0.83$ and $\theta = 1.54$). The theta angles are then decremented by a small amount, $\epsilon_\theta + \mathcal{N}(0, 0.01)$, and a new conformation is generated. This process in repeated until the θ angles are less than 0.087 radians (nearly a straight rod). The angles are then incremented by $\epsilon_\theta + \mathcal{N}(0, 0.01)$ on each step to bring the polymer back to a helical conformation (until $\theta > 1.54$). While this would result in a simple one-dimensional model, the polymer can be broken into distinct segments, s_i, each with a unique increment $\epsilon_\theta^{s_i}$. Even if $\epsilon_\theta^{s_i}$ is set to the same value for all segments, the small amount of Gaussian noise added to the increments at each step ,$\mathcal{N}(0, \sigma_{\epsilon_\theta})$, would make the dimensionality of the system equal to the number of independent segments chosen to exhibit coordinated motion, i.e. the number of independent ϵ_θ^s values used. For simplicity, the total number of independent ϵ_θ^s values used is referred to as the number of correlated dimensions (d_{cor}) in the ensemble. In addition, a small amount of Gaussian noise with zero mean and standard deviation, $\sigma_{\theta,\phi}$, is added to all angles, independent of the coordinated walk in angle space. Thus, the effect of noise can be investigated independent of the coordinated motion of the helix. This model therefore can examine the utility of dimensionality estimators for determining the number of effective coordinated dimensions of a polymer system and what effect noise has on estimates for the exact same system. Two example polymer ensembles for this model is shown in Figure 6.1C and 6.1D.

6.2.3 Noise Screening Using Discrete Fourier Transforms

Noise is an inevitable hurdle to overcome when applying dimensionality estimators to MD simulations. As the definition of the MLE dimension suggests, the data must be examined at different scales in order to ascertain the intrinsic dimension of the

manifold on which it exists. This manifold is of low dimensionality compared to the dimensionality of the ambient space, but noise will typically blur the manifold into the dimensions of the ambient space. That is, at small scales, the points may lay slightly off the manifold in any direction, making the manifold of higher dimensionality. If one looks at slightly larger scales than the magnitude of the noise, but small enough to still not encounter a geodesically distant portion of the manifold, then the intrinsic dimensionality might be recovered. Beyond this scale, multiple areas of the manifold may be taken into consideration, leaving the estimator to yet again assign a dimensionality equivalent to the ambient space dimensionality.

While the estimators cannot help but pay attention to the noisy aspects of the data, some methods for effectively removing the noise or smoothing the data are potentially very helpful. While many techniques exist for dealing with general smoothing of data, a natural method from the domain of signal processing for dealing with polymer data is to utilize the angular representation of the polymer chain combined with Discrete Fourier Transforms (DFTs) to screen out the noise. This can be accomplished by building a vector of complex values by transforming each one of the angles along the chain as $e^{i\theta}$ or $e^{i\phi}$, for each θ or ϕ angle, respectively and where $i = \sqrt{-1}$. By taking the DFT of this vector, all signals below a certain signal threshold or all signals above a certain frequency can be filtered out by setting the corresponding vector elements to zero, and then taking the inverse DFT and extracting the arguments (angles) from the result. Both the frequency cutoff and amplitude cutoff approaches have the potential to effectively screen the noise from the angular space, and produce a smoothed version of the polymer data which should be more amenable to dimensionality estimation. However, these approaches remain to be tested for efficacy.

6.2.4 Molecular Dynamics Simulations

Molecular dynamics simulations were performed on two NFPs and two IDPs which cover the major structural classes of both NFPs and IDPs. The first NFP, GB1

(RCSB Protein Data Bank ID: 1GB1), is a 16 amino acid long fast-folding protein that spontaneously adopts a β-hairpin conformation. The second NFP, Trp-cage (RCSB Protein Data Bank ID: 1L2Y), is a 20 amino acid long fast-folding protein that spontaneously adopts a mainly α-helical conformation. Therefore, These two proteins cover both general structural classes of NFPs. The folded conformations and complete sequences for both GB1 and Trp-cage are shown in Figure 6.2. The first IDP, Nsp1, is a 25 amino acid subseqence of the full length wildtype FG-nucleoporin NSP1 (NCBI Assession Number: NP_012494.1, Gene ID: 6322420), and the second, Nup116, is a 25 amino acid long subsequence of the full length wildtype FG-nucleoporin NUP116 (NCBI Assession Number: NP_013762.1, Gene ID: 6323691). These IDPs have been shown to adopt relaxed coil and collapsed coil structures, respectively, via both theory and experiment [21, 8]. A representative structure for each IDP is shown at the bottom of Figure 6.2, along with the specific amino acid sequences. These structures were chosen since they best match the average R_g and S parameters across all simulations outlined below. See Section 2.2.3 for details on the computation of R_g and S parameters for protein structures.

While the simulation of NFPs has been the focus of MD studies for many decades, the simulation of IDPs presents unique challenges. The two main issues are related to sampling and model selection. The previous chapters have already studied the aspect of sampling for IDPs in some detail, so the main challenge addressed in this chapter is that of model selection. In particular, the parameters governing the simulation of the proteins such as bond strengths, bond lengths, bond angles, atomic charges, van der Walls interactions, etc. have been heavily optimized to effectively simulate folded protein given the abundance of experimental data available for verification of NFP dynamics. However, IDPs have not been given as much attention in this regard, so the typical models employed for NFP protein simulations are often inadequate at accurately capturing the properties of IDPs or even unfolded NFPs [79, 80, 81, 82]. Some models have been developed in recent years to remedy this problem. While the development and testing of these models is beyond the scope of this study, three models have been chosen for the

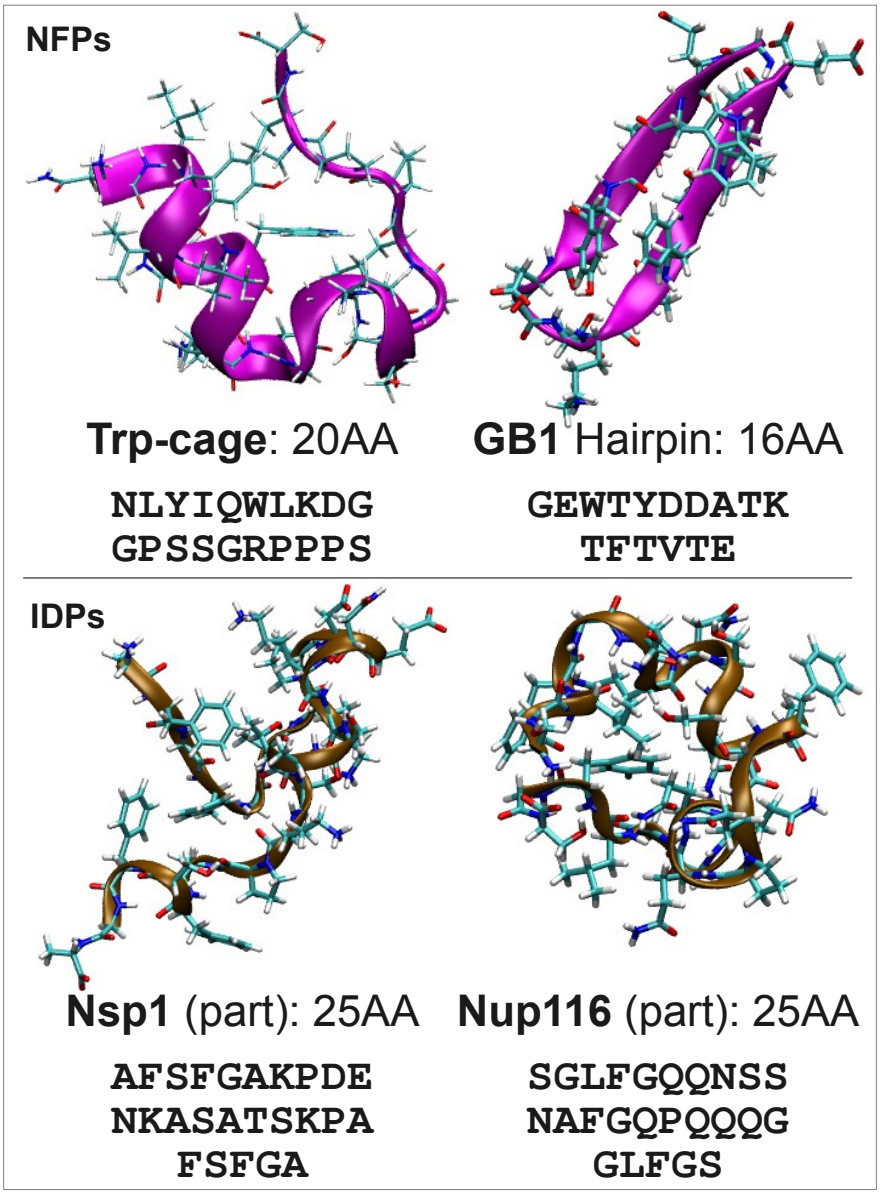

Figure 6.2: Protein structures and sequences examined. Top: The folded conformations of two natively folded proteins with corresponding amino acid sequences. Bottom: Representative structures from MD simulations of two intrinsically disordered proteins with corresponding amino acid sequences. Structures were selected based on best match to the average R_g and S parameters across all IDP simulations.

simulations performed here that show promise in addressing these concerns. Therefore, each protein was simulated using the following three models:

- Amber ff99SB-ILDN [82] – This model is the state-of-the-art in protein folding, and the corrections made compared to traditional models should address some known issues with simulated the unfolded state of proteins. It is based on the traditional Amber ff99SB model [83], but it was not been developed with IDPs in mind.

- Amber ff99SB-PSN [84] – This model makes small changes to the traditional Amber ff99SB model in order to better match experimental results for small disordered peptides. It has been shown to reliably simulate NFPs as well, even though this model was specifically designed for modeling IDPs.

- Amber ff03w[85] – This model makes some changes to the traditional Amber ff99SB model to better model the unfolded state of NFPs. Again, while these changes are thought to be favorable to the study of IDPs as well, this has not been extensively verified.

Ten independent replicate simulations of classical molecular dynamics were performed on the four proteins listed above using all three forcefield models listed above. The simulations were performed using the GROMACS version 4.5.5 software package [86], using the OBC/GBSA implicit solvent model [87]. Charged zwitter ionic termini of the proteins were used even for the disordered proteins to facilitate comparison between NFPs and IDPs, and the because they are needed to adequately stabilize the folded conformations of the NFPs [85]. Hydrogen bonds were constrained using the LINCS algorithm, and the velocity rescale thermostat was used for temperature control, both with default parameter settings. Each simulation started from a completely extended conformation that was minimized for 10000 steps of steepest descent, and then started with a unique set of random initial velocities. The temperature at the beginning of the simulation was set to 300K, but was then linearly increased to 600K for the first

25ns of the simulation, then held at 600K for 50ns before linearly decreasing the temperature for another 25ns back down to 300K. The simulation then continued on for an additional 250ns at a constant 300K. Structures were saved every 2ps, but were subsampled at every 10ps for all analyses presented here for tractability to form a total of 35001 structures per simulation. This is an aggregate total of 42 microseconds of simulation for this study.

6.3 Results

6.3.1 Semirigid Helix

Semirigid helix ensembles were generated for three different polymer lengths (l): 16, 20, and 25 particles. Since the actual number of links in these polymers is one less than the number of particles, there are $2l - 5$ degrees of freedom (DoF) in each of these the polymer chains: 27, 35, and 45, respectively. These lengths were chosen because they model the lengths of some of the proteins in the MD simulations (e.g. GB1, which is 16 residues long). Ensembles of $N = 2000$ and $N = 5000$ structures each were generated for five different values of $\sigma_{\theta,\phi} = [0.0, 0.01, 0.1, 1.0, 3.0, 10.0]$. Each ensemble models the folded structure of a protein with increasing degrees of noise, with a maximum $\sigma_{\theta,\phi} = 10.0$ that is equivalent to a completely random chain. A noise level $\sigma_{\theta,\phi} = 0.1$ is most similar to that observed in MD simulations of folded proteins of corresponding length at 300K (RMSD $\approx$ 0.1nm). In addition, each of these ensembles was subjected to noise smoothing via DFTs at various cutoff fractions $(0.0, 0.01, 0.05, 0.1, 0.5)$ in terms of frequency or amplitude of the component angular signals. While the ensembles are not simulations, they still model a sparsely sampled trajectory, and the results play a large role in deciphering the efficacy of the DFT screening methods.

The dimensionality estimation results for the semirigid helix using $N = 5000$ structures of length $l = 20$ are shown in Figure 6.3. The effect of various levels of noise

on the estimator (Figure 6.3A) indicate that even the smallest noise levels will result in large overestimation of the intrinsic dimensionality of the system. However, noise is inherently high-dimensional, filling the entire ambient space of available degrees of freedom. A system of this length has $2l - 5 = 35$ DoF, so the maximum noise of $\sigma_{\theta,\phi}$ should allow an accurate prediction of the system size. However, only at small values of k can the estimator make an accurate prediction at this level of noise. At larger values of k, the estimates decrease. Since all degrees of freedom are being used, it is clear that the falloff in the estimates is due to a lack of adequate sampling. The results in Table A.5 for the $N = 2000$ case confirm this analysis, as that sample suffers even more at large k. However, Figure 6.3A also shows that the smaller noise levels are more hindered in that regard. In other words, small magnitudes in the fluctuation of the polymer will lead to lower estimations of the dimensionality than very extended ones. Therefore, it is expected that estimates will be a fraction lower than the maximum number of DoF for fully folded systems.

Figure 6.3B shows the DFT frequency smoothing results for the same semirigid helix at various levels of smoothing. As the fraction of the removed signals increases, it is clear that differences are only seen at very large fraction values where the dimensionality estimates become lower. Also, this effect is more dramatic for small k than middle or large values of k. On the whole, it is clear that the smoothing was not very effective at removing the noise. It does create a more consistent estimate across all values of k, but this is not desired since the estimates for small k were clearly more accurate as shown in Figure 6.3.

The remaining results for the rest of the semirigid helices are shown in Tables 6.3-6.2. These results corroborate the results observed in Figure 6.3. In particular, the overestimation of the dimensionality due to noise, as well as the underestimation of the total number of degrees of freedom at smaller noise levels repeats in all of the data. The results for the DFT amplitude smoothed estimates are shown in Tables 6.1 and 6.2. This method of smoothing has a smaller effect on the results than the frequency smoothing method, shown in Tables 6.1 and 6.2, in terms of the raw drop in the dimen-

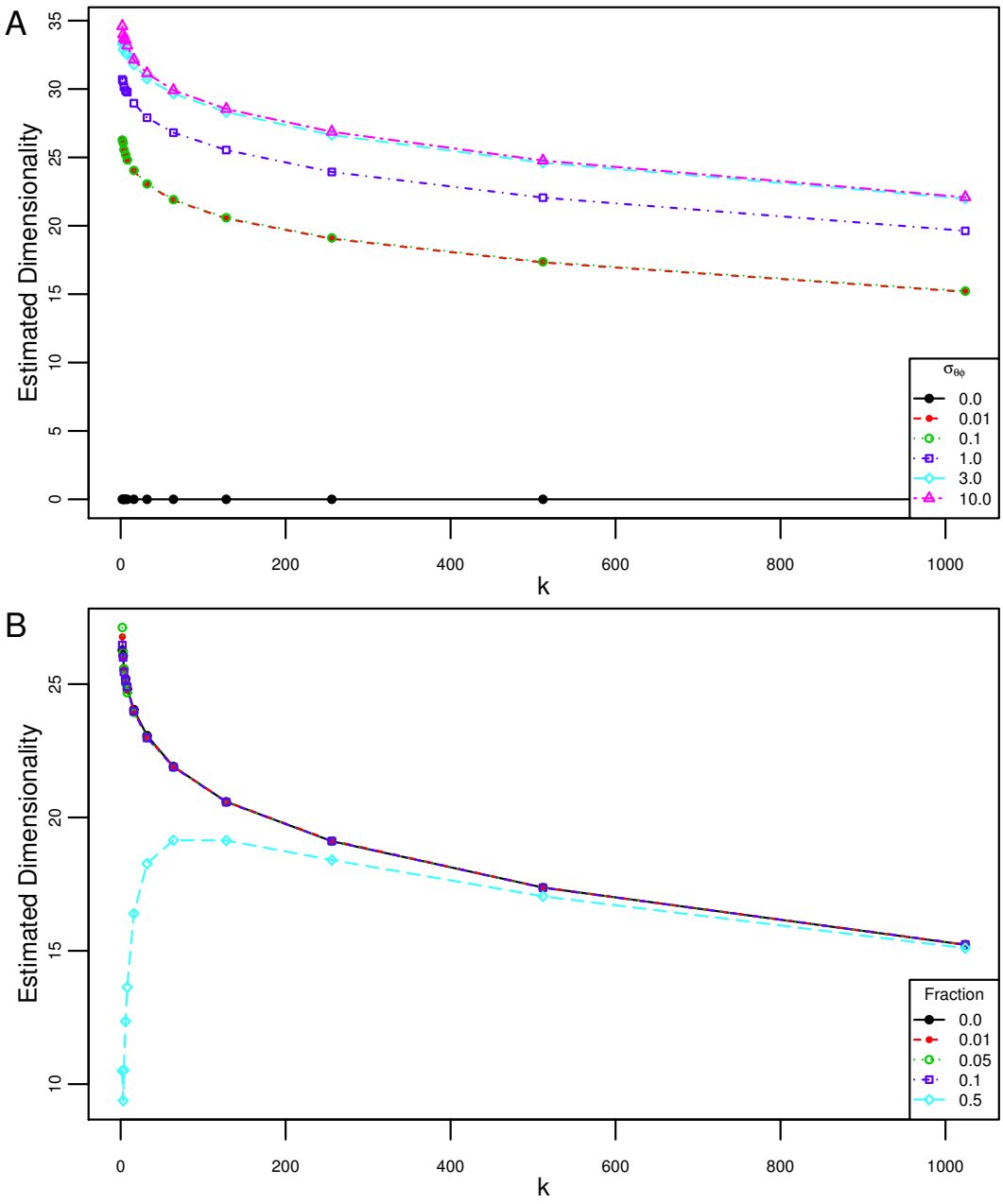

Figure 6.3: Semirigid helix model results. (A) Estimated dimensionality for a semirigid helical polymer of $N = 5000$ structures of length 20 with various amounts of noise ($\sigma_{\theta,\phi}$) injected into the "folded" ensemble. (B) Estimated dimensionality for the same polymer under a noise level of $\sigma_{\theta,\phi} = 0.10$, but with DFT frequency cutoff smoothing applied to smooth out different fractions of the high-frequency motions in order to attempt to eliminate the effect of the noise.

Table 6.1: Dimensionality estimates for the *semirigid helix* model with $N = 2000$ structures across all polymer lengths, all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. The number of degrees of freedom in each ensemble is provided for reference.

		Frequency Smoothing Fraction				
DoF=27		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	17.364	17.375	17.446	17.359	12.584
	0.10	17.396	17.406	17.474	17.385	12.583
	1.00	21.064	21.083	21.063	20.999	13.429
	3.00	23.125	23.094	22.985	22.913	14.574
	10.00	23.379	23.386	23.213	23.047	14.549
DoF=35		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	20.554	20.557	20.654	20.586	13.017
	0.10	20.597	20.595	20.704	20.626	13.020
	1.00	25.407	25.404	25.170	24.912	14.429
	3.00	27.880	27.874	27.861	27.698	15.795
	10.00	28.349	28.338	28.386	28.232	15.760
DoF=45		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	24.142	24.224	24.239	24.044	13.574
	0.10	24.208	24.272	24.276	24.101	13.579
	1.00	30.300	30.301	29.941	30.032	15.418
	3.00	33.255	33.271	33.108	32.746	16.869
	10.00	33.543	33.506	33.396	33.490	17.005

sionality estimates. While frequency smoothing sharply affected small values of k, the effect is more broad for amplitude smoothing, affecting all k values relatively equally, and only noticeably for larger noise values. Therefore, using a frequency cutoff would be considered more useful if one expects noise to dominate the estimates at small k. A breakdown of these results for different scales of k can be found in the corresponding tables in Section A.1.

Table 6.2: Dimensionality estimates for the *semirigid helix* model with $N = 5000$ structures across all polymer lengths, all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. The number of degrees of freedom in each ensemble is provided for reference.

		Frequency Smoothing Fraction				
DoF=27		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	18.885	18.912	18.897	18.811	14.478
	0.10	18.927	18.958	18.932	18.851	14.472
	1.00	22.403	22.421	22.367	22.150	15.232
	3.00	24.846	24.874	24.750	24.634	16.437
	10.00	24.753	24.772	24.684	24.524	16.420
DoF=35		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	22.402	22.420	22.417	22.383	14.991
	0.10	22.437	22.465	22.472	22.424	14.991
	1.00	27.159	27.131	27.004	26.845	16.365
	3.00	29.865	29.864	29.743	29.503	17.684
	10.00	30.374	30.377	30.316	30.113	17.797
DoF=45		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	26.207	26.258	26.353	26.292	15.585
	0.10	26.280	26.310	26.401	26.348	15.585
	1.00	32.423	32.391	32.220	32.109	17.528
	3.00	35.903	35.846	35.682	35.439	19.031
	10.00	36.219	36.147	36.096	36.150	19.093

Table 6.3: Dimensionality estimates for the *semirigid helix* model with $N = 2000$ structures across all polymer lengths, all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. The number of degrees of freedom in each ensemble is provided for reference.

		Amplitude Smoothing Fraction				
DoF=27		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	17.364	17.364	17.350	17.404	17.135
	0.10	17.396	17.395	17.381	17.410	17.341
	1.00	21.064	21.064	21.085	21.115	15.984
	3.00	23.125	23.126	23.110	23.064	17.214
	10.00	23.379	23.380	23.364	23.370	16.876
DoF=35		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	20.554	20.554	20.557	20.578	20.448
	0.10	20.597	20.597	20.603	20.610	20.549
	1.00	25.407	25.407	25.423	25.429	19.985
	3.00	27.880	27.878	27.877	27.922	20.694
	10.00	28.349	28.349	28.340	28.266	20.485
DoF=45		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	24.142	24.141	24.146	24.103	24.031
	0.10	24.208	24.208	24.218	24.162	24.200
	1.00	30.300	30.301	30.305	30.248	22.730
	3.00	33.255	33.255	33.286	33.231	24.111
	10.00	33.543	33.544	33.531	33.566	25.028

Table 6.4: Dimensionality estimates for the *semirigid helix* model with $N = 5000$ structures across all polymer lengths, all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. The number of degrees of freedom in each ensemble is provided for reference.

		Amplitude Smoothing Fraction				
DoF=27		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	18.885	18.885	18.890	18.924	18.614
	0.10	18.927	18.927	18.935	18.956	18.717
	1.00	22.403	22.403	22.400	22.436	17.307
	3.00	24.846	24.846	24.839	24.782	16.464
	10.00	24.753	24.753	24.740	24.752	16.721
DoF=35		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	22.402	22.402	22.408	22.436	22.485
	0.10	22.437	22.436	22.442	22.467	22.455
	1.00	27.159	27.159	27.168	27.151	21.539
	3.00	29.865	29.865	29.865	29.879	21.875
	10.00	30.374	30.375	30.374	30.372	20.788
DoF=45		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	26.207	26.207	26.195	26.112	26.064
	0.10	26.280	26.279	26.267	26.200	26.183
	1.00	32.423	32.423	32.389	32.409	25.983
	3.00	35.903	35.903	35.897	35.851	25.763
	10.00	36.219	36.219	36.241	36.208	25.504

6.3.2 Half-folded Helix

Half-folded helix ensembles were generated for three different polymer lengths (l): 16, 20, and 25 particles. The total number of degrees of freedom in each set was 27, 35, and 45, respectively. Ensembles of $N = 2000$ and $N = 5000$ structures each were generated for five different values of $\sigma_{\theta_{unfolded}} = [0.0, 0.01, 0.1, 1.0, 3.0, 10.0]$. Each ensemble models the folded structure of a protein on one side, i.e. a helix with relatively strong structural integrity, versus the other side which is either folded or unfolded as a function of $\sigma_{\theta_{unfolded}}$. Therefore these ensembles serve as a model for the partially folded state of an NFP, or of proteins which are normally partially folded under native conditions. The ensembles were also subjected to noise smoothing via DFTs at various cutoff fractions $(0.0, 0.01, 0.05, 0.1, 0.5)$ in terms of frequency or amplitude of the component angular signals. Like the semirigid model, this model produces ensembles which are similar to a sparsely sampled, equilibrated MD trajectory, i.e. there are no temporal correlations between successive structures.

The dimensionality estimation results for the half-folded helix using 5000 structures of length 20 are shown in Figure 6.4. The effect of various levels of noise on the estimator (Figure 6.4A) indicate that the introduction of larger amounts of noise into only one half of the structure actually *lowers* the estimates for the polymer. This result is seemingly in direct contrast with the earlier results from the semirigid helix model, which showed an increase in estimated dimensionality with increasing noise. The estimates for the half-folded helix are still, in general, overestimating the intrinsic dimensionality of the system (which is still zero by definition). However, it is clear that a half-folded polymer system (in the presence of small amounts of noise) will have lower estimated dimensionality than either its completely unfolded or completely folded cousins. In addition, this effect is most dramatic at intermediate levels of $\sigma_{\theta_{unfolded}}$ with a minimum in this study at $\sigma_{\theta_{unfolded}} = 0.1$. This is interesting because it models the smoothness of the transition from a folded system with high estimated dimensionality (due to the presence of minor, anticipated noise and $\sigma_{\theta_{unfolded}} = 0.0$), to a system with

low estimated dimensionality as it starts to unfold (intermediate $\sigma_{\theta_{unfolded}}$), and finally back to a system with high estimated dimensionality (high $\sigma_{\theta_{unfolded}}$). Therefore, the results from both the half-folded helix model and the semirigid helix model demonstrate that the estimator is capable of distinguishing between interesting shifts in structure even if it is underestimating or overestimating the dimensionality. However, one should note that while these results hold for polymer ensembles and simulations, and therefore potentially protein simulations, more general data sets will most likely not conform to these trends.

Several predictions can be made from these data. Although the intrinsic dimensionality of a folded system is zero, the estimator will predict extremely high estimates due to the noise (especially for small k). However, the estimates are predicted to drop during the intermediate transitions in the folding process, just like the half-folded helix in the intermediate noise state. Also, since disordered proteins are not truly random coils, their estimated dimensionality is predicted to be somewhat depressed, and should fall in the intermediate regime investigated here with the half-helix model at most times. Therefore, it might be possible to distinguish folding proteins from disordered proteins based on the estimates, even in the presence of noise. Nonetheless, the ability to screen the noise from the models would be of benefit in order to accurately distinguish between a folded protein and completely random coil, even if proteins cannot exhibit such extreme dynamics due to steric exclusion and bond angle constraints.

Figure 6.4B shows the DFT frequency smoothing results for the same half-folded helix model at various levels of smoothing. As the fraction of the removed signal increases, the same effect is observed with this model as with the semirigid helix model above: the smoothing affects mainly the high estimates at small k. Smoothing the data consistently allows the estimator to make lower estimates, and doesn't seem to effect the rank-order of the estimates. However, it also cannot smooth it enough to obtain the desired estimate of zero for these systems which are all essentially oscillating around a helical structure.

The remaining results for the rest of the half-folded helices are shown in Ta-

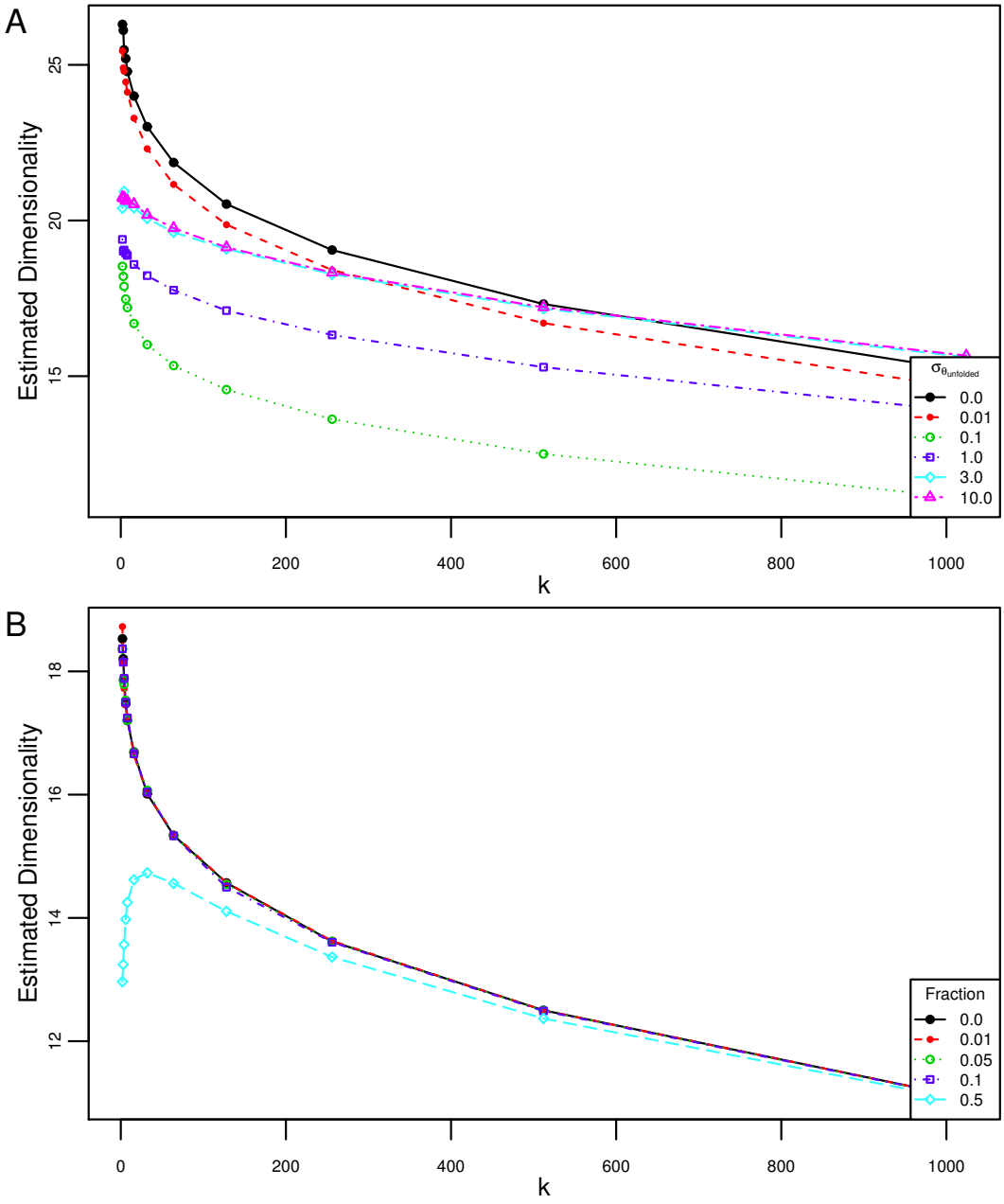

Figure 6.4: Half-folded helix model results. (A) Estimated dimensionality for a half-folded helical polymer of length 20 with various amounts of noise ($\sigma_{\theta_{unfolded}}$) injected into the half-folded ensemble. (B) Estimated dimensionality for the same polymer under a noise level of $\sigma_{\theta,\phi} = 0.10$, but with DFT frequency cutoff smoothing applied to smooth out different fractions of the high-frequency motions in order to attempt to eliminate the effect of the noise.

bles 6.7-6.6. These results corroborate the results observed in Figure 6.4. In particular, the drop in estimated dimensionality due to intermediate levels of $\sigma_{\theta_{unfolded}}$ and subsequent increase at higher values of the overestimation of $\sigma_{\theta_{unfolded}}$. The results for the DFT amplitude smoothed estimates are shown in Tables 6.5 and 6.6. This method of smoothing has a smaller effect on the results than the frequency smoothing method, shown in Tables 6.5 and 6.6 in terms of the raw drop in the dimensionality estimates. While frequency smoothing sharply affected small values of k, the effect is more broad for amplitude smoothing, affecting all k values relatively equally, and only noticeable for larger noise values. Therefore, the results are similar to those obtained from the semirigid helix model, and using the frequency cutoff might be the more useful choice as opposed to an amplitude cutoff. A breakdown of these results for different scales of k can be found in the corresponding tables in Section A.2.

6.3.3 Correlated Helix

Correlated helix ensembles were generated for two different polymer lengths (l): 20 and 25 particles each for differing numbers of correlated dimensions: 2, 3, and 5. The total number of degrees of freedom in each set was 35 and 45 respectively, but the motions of the polymer are constrained onto a space equivalent to the number of correlated dimensions. Ensembles of $N = 2000$ and $N = 5000$ structures each were generated for five different values of $\sigma_{\theta,\phi} = [0.0, 0.01, 0.1, 1.0, 3.0, 10.0]$. Each ensemble models the coordinated folding/unfolding of a set of helices equal to the number of correlated dimensions for the ensemble. Since each structure is successively generated based on the prior structure in this model, it is effectively a trajectory. However, the noise level injected into the model greatly affects the extent to which the ensemble can be recognized as a trajectory. At $\sigma_{\theta,\phi} = 0.0$, the model indeed has this property. But at $\sigma_{\theta,\phi} = 10.0$, the noise is so great that that the model would be equivalent to the semirigid chain model with the same $\sigma_{\theta,\phi} = 0.0$. This model can therefore be used to investigate the effect of correlated motions on the estimator similar to how the effect of interme-

Table 6.5: Dimensionality estimates for the *half-folded helix* model with $N = 2000$ structures across all polymer lengths, all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. The number of degrees of freedom in each ensemble is provided for reference.

		Frequency Smoothing Fraction				
DoF=27		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	17.364	17.375	17.446	17.359	12.584
	0.01	17.057	16.975	16.863	17.039	12.404
	0.10	12.505	12.539	12.630	12.575	10.965
	1.00	13.976	13.942	13.863	13.909	10.838
	3.00	15.415	15.382	15.327	15.325	12.098
	10.00	15.415	15.382	15.327	15.325	12.098
DoF=35		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	20.554	20.557	20.654	20.586	13.017
	0.01	19.837	19.860	19.735	19.898	12.954
	0.10	14.668	14.566	14.629	14.491	11.916
	1.00	16.863	16.849	16.763	16.559	12.394
	3.00	18.556	18.525	18.425	18.361	13.191
	10.00	18.470	18.471	18.375	18.451	13.068
DoF=45		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	24.142	24.224	24.239	24.044	13.574
	0.01	23.193	23.081	23.084	23.126	13.414
	0.10	16.323	16.313	16.279	16.145	12.307
	1.00	19.337	19.252	19.252	18.999	12.988
	3.00	21.578	21.472	21.311	21.105	14.178
	10.00	21.693	21.645	21.536	21.429	14.082

Table 6.6: Dimensionality estimates for the *half-folded helix* model with $N = 5000$ structures across all polymer lengths, all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. The number of degrees of freedom in each ensemble is provided for reference.

		Frequency Smoothing Fraction				
DoF=27		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	18.885	18.912	18.897	18.811	14.478
	0.01	18.535	18.567	18.420	18.409	14.306
	0.10	13.407	13.414	13.401	13.374	12.193
	1.00	14.746	14.715	14.681	14.562	12.094
	3.00	16.057	16.047	16.024	15.923	13.132
	10.00	16.057	16.047	16.024	15.923	13.132
DoF=35		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	22.402	22.420	22.417	22.383	14.991
	0.01	21.670	21.714	21.599	21.628	14.920
	0.10	15.759	15.758	15.714	15.738	13.566
	1.00	17.704	17.715	17.591	17.609	13.863
	3.00	19.462	19.468	19.461	19.408	14.715
	10.00	19.521	19.531	19.423	19.328	14.681
DoF=45		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	26.207	26.258	26.353	26.292	15.585
	0.01	25.331	25.255	25.277	25.269	15.408
	0.10	17.794	17.707	17.779	17.821	14.157
	1.00	20.434	20.345	20.321	20.183	14.676
	3.00	22.686	22.627	22.517	22.413	16.008
	10.00	22.859	22.843	22.809	22.596	15.861

Table 6.7: Dimensionality estimates for the *half-folded helix* model with $N = 2000$ structures across all polymer lengths, all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. The number of degrees of freedom in each ensemble is provided for reference.

		Amplitude Smoothing Fraction				
DoF=27		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	17.364	17.364	17.350	17.404	17.135
	0.01	17.057	17.057	17.070	17.054	13.358
	0.10	12.505	12.502	12.303	12.274	12.135
	1.00	13.976	13.972	13.977	13.959	12.103
	3.00	15.415	15.413	15.425	15.445	10.147
	10.00	15.415	15.413	15.425	15.445	10.147
DoF=35		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	20.554	20.554	20.557	20.578	20.448
	0.01	19.837	19.837	19.815	19.946	16.082
	0.10	14.668	14.667	14.465	14.405	14.447
	1.00	16.863	16.859	16.847	16.838	14.200
	3.00	18.556	18.553	18.528	18.554	13.601
	10.00	18.470	18.467	18.445	18.476	13.285
DoF=45		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	24.142	24.141	24.146	24.103	24.031
	0.01	23.193	23.195	23.218	23.187	17.984
	0.10	16.323	16.321	16.110	16.048	15.999
	1.00	19.337	19.331	19.335	19.364	15.893
	3.00	21.578	21.574	21.584	21.500	15.740
	10.00	21.693	21.689	21.682	21.641	15.339

Table 6.8: Dimensionality estimates for the *half-folded helix* model with $N = 5000$ structures across all polymer lengths, all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. The number of degrees of freedom in each ensemble is provided for reference.

		Amplitude Smoothing Fraction				
DoF=27		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	18.885	18.885	18.890	18.924	18.614
	0.01	18.535	18.535	18.531	18.472	14.605
	0.10	13.407	13.405	13.181	13.115	13.027
	1.00	14.746	14.741	14.747	14.772	12.884
	3.00	16.057	16.054	16.046	15.995	10.009
	10.00	16.057	16.054	16.046	15.995	10.009
DoF=35		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	22.402	22.402	22.408	22.436	22.485
	0.01	21.670	21.670	21.670	21.735	16.364
	0.10	15.759	15.758	15.502	15.462	15.255
	1.00	17.704	17.698	17.696	17.722	15.217
	3.00	19.462	19.458	19.463	19.455	13.349
	10.00	19.521	19.517	19.513	19.567	12.864
DoF=45		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	26.207	26.207	26.195	26.112	26.064
	0.01	25.331	25.330	25.324	25.291	19.305
	0.10	17.794	17.793	17.515	17.387	17.457
	1.00	20.434	20.427	20.420	20.353	16.713
	3.00	22.686	22.681	22.683	22.660	15.686
	10.00	22.859	22.854	22.864	22.854	15.008

diate levels of disorder was investigated using the half-folded helix model. In addition, each of these ensembles was subjected to noise smoothing via DFTs at various cutoff fractions $(0.0, 0.01, 0.05, 0.1, 0.5)$ in terms of frequency or amplitude of the component angular signals.

The dimensionality estimation results for the correlated helix using $N = 5000$ structures of length 20, exhibiting 5 correlated dimensions is shown in Figure 6.5. The effect of various levels of noise on the estimator (Figure 6.5A) indicate that the correlations in the motion of the ensemble allow the estimator to make a fairly accurate assessment of the dimensionality of the system for small levels of noise. It still underestimates the intrinsic dimensionality of 5, but that has been the case consistently for both noise and signal using these models, most likely due to undersampling. Underestimation is also potentially due to the fact that the manifold is not closed, and estimates near the boundaries may therefore be artificially low. This is even true out to the noise level observed to best mimic that of a folded system: $\sigma_{\theta,\phi} = 0.1$. However, the effect at small k is less strong, so the estimator is less accurate in this case. The major cause of this result is the larger amplitude motion that the correlated dimensions are capturing relative to the noise. However, once the noise level becomes sufficiently large, the correlated signal is lost, and the estimates jump to the same value as the equivalent semirigid helix model. Such correlated motion would be exhibited by very extended protein chains, and even short pieces of a folding protein during the folding process. Therefore, in combination with the effects observed from the half-folded helix data, it is clear that intermediate conformations and correlated motion might allow the estimator to adequately produce a signal that ignores the noise to a relative, but significant extent. In this case, undersampling plays a large role in making these estimates inaccurate even still. However, it is also clear by comparing the results across correlated helix ensembles with varying numbers of correlated dimensions (see Tables 6.9,6.10, 6.13 and 6.14) that the rank-order of the dimensionality estimates is retained. So, while an absolute estimate of the dimensionality might be unavailable, it is clear that relative dimensionality is detected.

Figure 6.5B shows the DFT frequency smoothing results for the same correlated

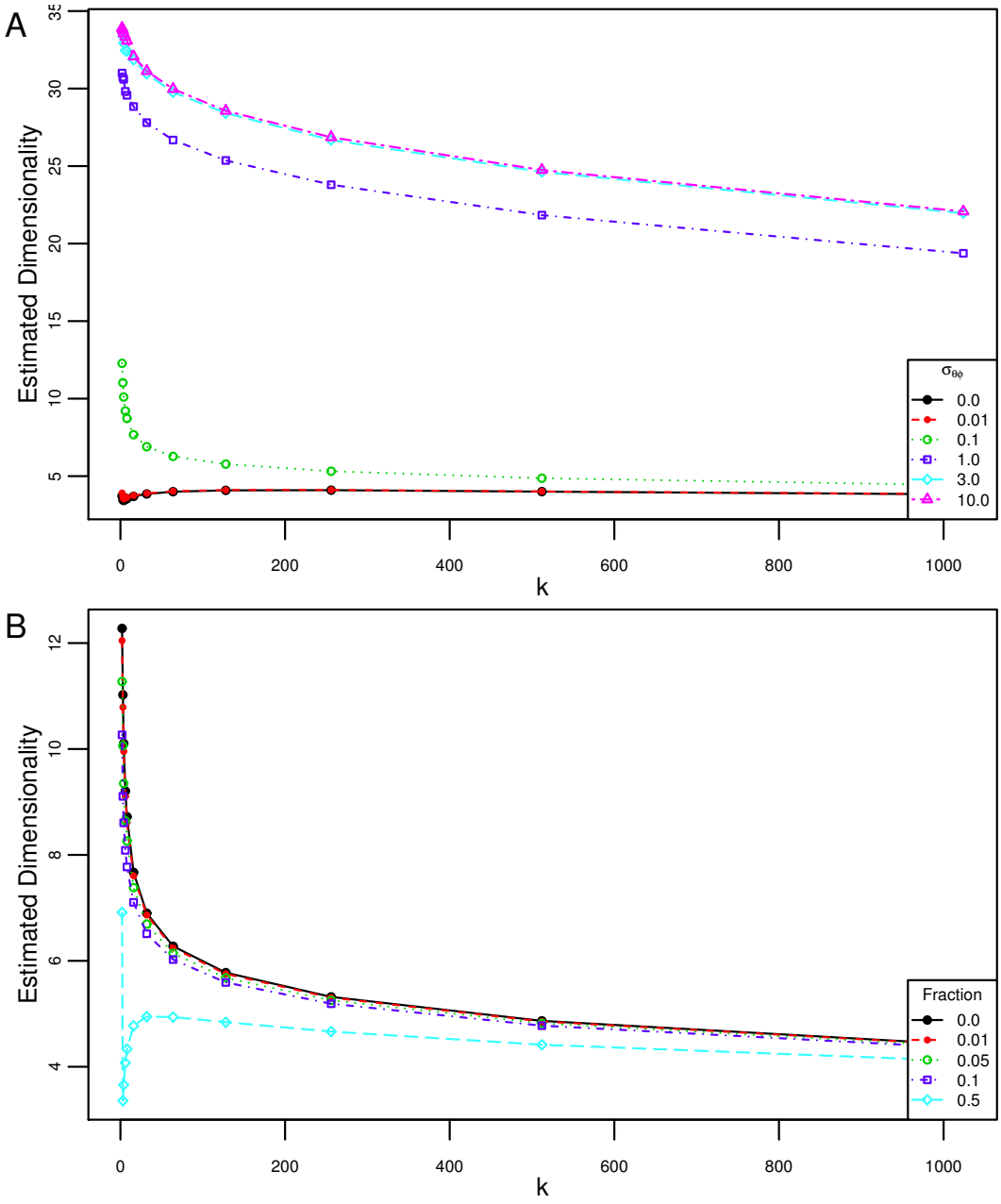

Figure 6.5: Correlated helix model results. (A) Estimated dimensionality for a correlated helical polymer of length 20 and 5 correlated dimensions of motion with various amounts of noise ($\sigma_{\theta,\phi}$) injected into the ensemble. (B) Estimated dimensionality for the same polymer under a noise level of $\sigma_{\theta,\phi} = 0.10$, but with DFT frequency cutoff smoothing applied to smooth out different fractions of the high-frequency motions in order to attempt to eliminate the effect of the noise.

helix model at various levels of smoothing. The pattern is the same as that observed for the semirigid helix model and half-folded helix model. Essentially little effect is garnered by setting the fraction low, but a value of 0.5 has a significant effect at small k, allowing the estimator to more stably reproduce the same (under)estimate at small scales, in addition to large scales.

The remaining results for the rest of the correlated helices are shown in Tables 6.11-6.14. These results corroborate the results observed in Figure 6.5. In particular, there is always a closer estimate of the intrinsic dimensionality at intermediate levels of $\sigma_{\theta,\phi}$. However, there is also a consistent underestimation of the intrinsic dimensionality in the no-noise condition $\sigma_{\theta,\phi} = 0$ due to undersampling. The relative order of the estimated dimensionality across different numbers of correlated dimensions also persists throughout these results The results for the DFT amplitude smoothed estimates are shown in Tables 6.11, 6.12, 6.15, and 6.16. This method of smoothing has a smaller effect on the results than the frequency smoothing method, shown in Tables 6.9, 6.10, 6.13 and 6.14, in terms of the raw drop in the dimensionality estimates. While frequency smoothing sharply affected small values of k, the effect is more broad for amplitude smoothing, affecting all k values relatively equally, and only noticeable for larger noise values. Therefore, the results are similar to those obtained from the semirigid helix model and half-folded helix model. A breakdown of these results for different scales of k can be found in the corresponding tables in Section A.3.

In addition to the summary statistics for the correlated helix data shown in Figure 6.5, the results of the pointwise dimensionality estimate analysis are shown in Figure 6.6. A running harmonic mean with a window size of 500 is shown in Figure 6.6 for different numbers of k. The results indicate that the pointwise method of estimating the dimensionality works as well for estimating the dimensionality as agglomerating the results across all structures. However, the small oscillations in the estimator around the mean value indicate that the estimator is actually biased when it comes into contact with the edge of the manifold. The effect is more clearly seen from the results of applying DFT smoothing to the trajectory (Figure 6.7), where the deviations are quite large

Table 6.9: Dimensionality estimates for the *correlated helix* model with $N = 2000$ structures of length 20 using several correlated folding/unfolding events across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. The number of correlated dimensions in each ensemble is provided for reference.

		Frequency Smoothing Fraction				
Dims=2		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.857	1.852	1.881	1.895	1.922
	0.01	2.519	2.520	2.512	2.493	2.293
	0.10	10.056	10.029	9.895	9.797	6.063
	1.00	24.785	24.734	24.491	24.032	12.084
	3.00	28.156	28.118	28.070	28.051	15.748
	10.00	28.470	28.616	28.262	28.201	15.794
Dims=3		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.737	2.746	2.749	2.751	2.791
	0.01	2.908	2.907	2.903	2.898	2.868
	0.10	7.966	7.895	7.705	7.446	4.666
	1.00	24.821	24.800	24.510	24.199	12.159
	3.00	27.675	27.697	27.757	27.638	15.801
	10.00	28.423	28.243	28.533	27.929	15.746
Dims=5		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.335	3.333	3.328	3.314	3.328
	0.01	3.382	3.380	3.372	3.355	3.340
	0.10	6.396	6.324	6.072	5.771	3.970
	1.00	25.369	25.294	25.067	24.607	12.085
	3.00	28.142	28.066	27.916	27.850	15.744
	10.00	28.226	28.338	28.190	27.862	15.737

even with smoothing. Proteins also do not often explore certain portions of their angular configuration space, so these results predict that the underestimation of the intrinsic dimensionality would persist independent of other factors impacting the estimates.

6.3.4 Molecular Dynamics Simulations

Now that the behavior of the estimator has been verified using the polymer models, the analysis of the protein simulations remain. Several points concerning the behavior of the polymer models should be reiterated for interpreting the results. First, while the original hypothesis that an NFP will eventually fold and undergo zero-dimensional

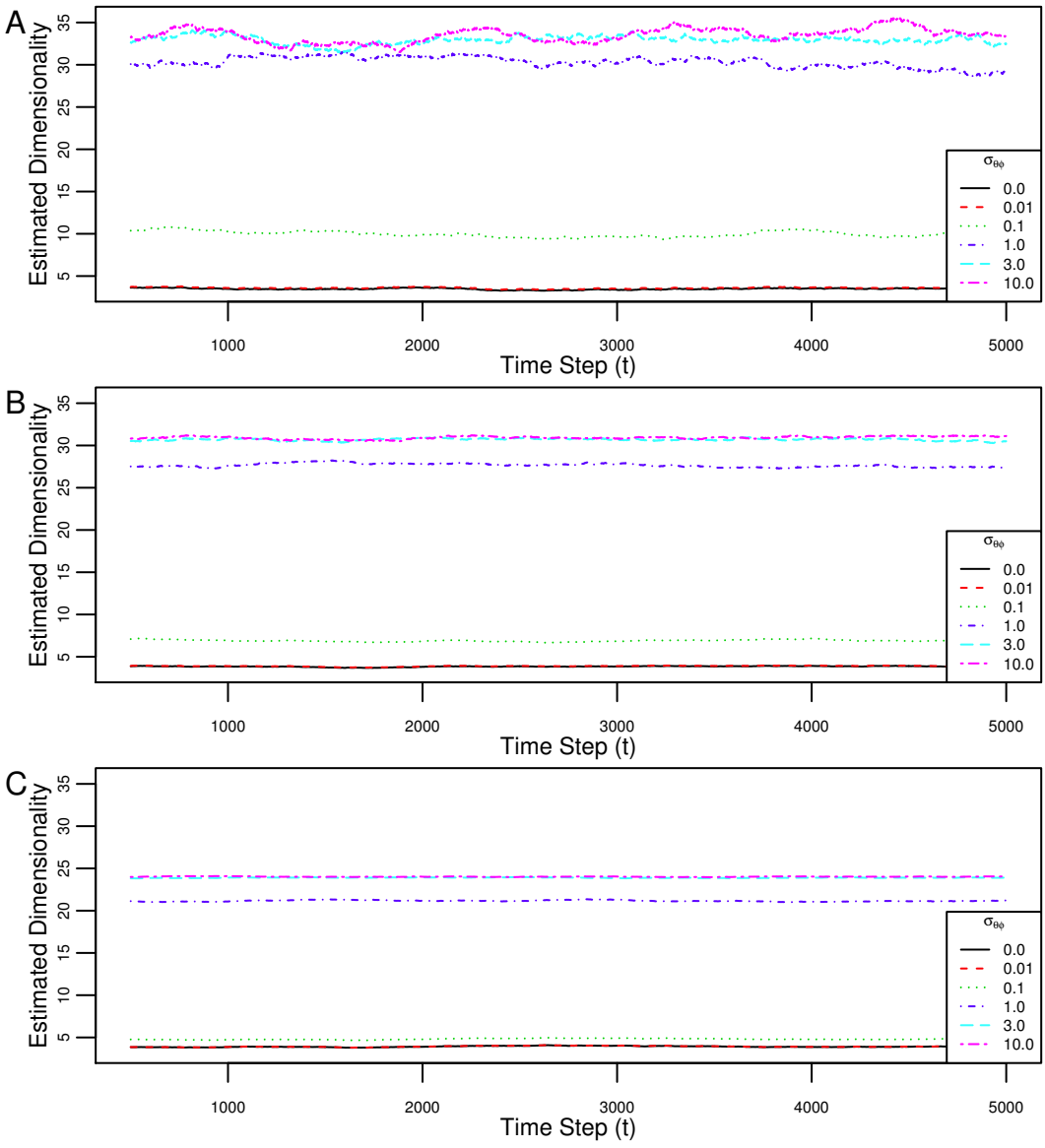

Figure 6.6: Correlated helix model pointwise results. Running harmonic mean of pointwise dimensionality estimates for a correlated helical polymer of length 20 and 5 correlated dimensions of motion with various amounts of noise ($\sigma_{\theta,\phi}$) injected into the ensemble. Estimates were obtained from averaging values for (A) small $k = (2, 3, 4, 6)$, (B) medium $k = (8, 16, 32, 64)$, and (C) large $k = (128, 256, 512, 1024)$.

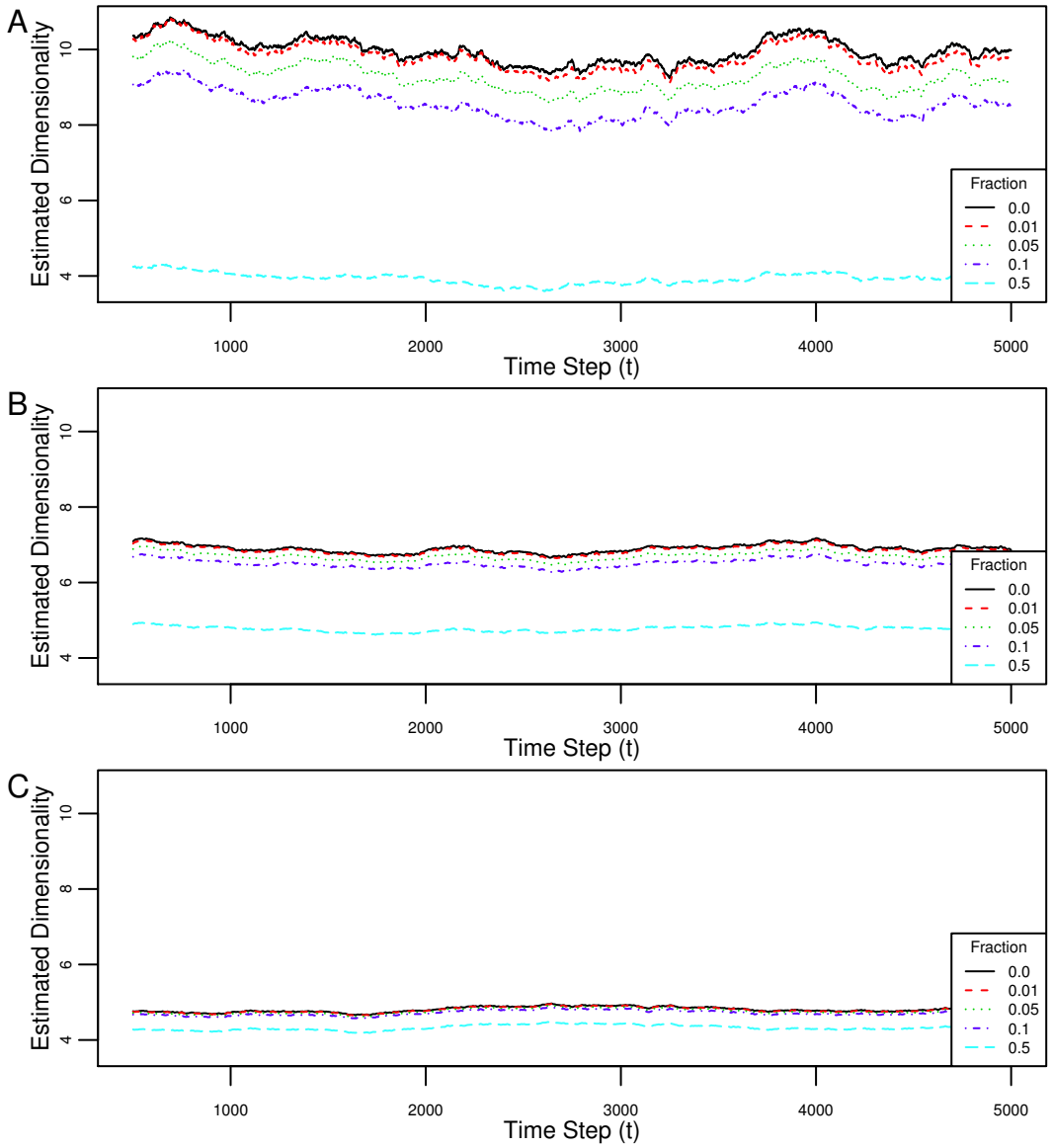

Figure 6.7: Correlated helix model *smoothed* pointwise results. Running harmonic mean of pointwise dimensionality estimates for a correlated helical polymer of length 20 and 5 correlated dimensions of motion under a noise level of $\sigma_{\theta,\phi} = 0.10$, but with DFT frequency cutoff smoothing applied to smooth out different fractions of the high-frequency motions in order to attempt to eliminate the effect of the noise for (A) small $k = (2, 3, 4, 6)$, (B) medium $k = (8, 16, 32, 64)$, and (C) large $k = (128, 256, 512, 1024)$.

Table 6.10: Dimensionality estimates for the *correlated helix* model with $N = 5000$ structures of length 20 using several correlated folding/unfolding events across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. The number of correlated dimensions in each ensemble is provided for reference.

		Frequency Smoothing Fraction				
Dims=2		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.882	1.886	1.910	1.935	1.929
	0.01	3.279	3.271	3.227	3.188	2.711
	0.10	12.353	12.359	12.243	12.129	8.049
	1.00	26.697	26.641	26.415	26.075	13.972
	3.00	29.949	29.951	29.879	29.806	17.718
	10.00	30.354	30.319	30.179	30.076	17.822
Dims=3		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.795	2.797	2.803	2.799	2.828
	0.01	3.053	3.049	3.045	3.031	2.955
	0.10	9.809	9.758	9.572	9.332	5.852
	1.00	26.785	26.738	26.531	26.219	13.824
	3.00	29.973	30.044	30.079	29.905	17.691
	10.00	30.313	30.291	30.303	30.091	17.746
Dims=5		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.769	3.768	3.764	3.755	3.758
	0.01	3.833	3.831	3.821	3.806	3.774
	0.10	7.713	7.629	7.329	6.949	4.586
	1.00	27.117	27.051	26.683	26.307	13.755
	3.00	29.939	29.912	29.875	29.785	17.719
	10.00	30.253	30.295	30.171	29.950	17.741

Table 6.11: Dimensionality estimates for the *correlated helix* model with $N = 2000$ structures of length 20 using several correlated folding/unfolding events across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. The number of correlated dimensions in each ensemble is provided for reference.

		Amplitude Smoothing Fraction				
Dims=2		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.857	1.910	1.901	1.914	1.846
	0.01	2.519	2.268	2.350	2.089	1.852
	0.10	10.056	9.386	4.056	3.912	2.024
	1.00	24.785	24.780	24.621	23.645	3.677
	3.00	28.156	28.157	28.154	28.134	20.352
	10.00	28.470	28.470	28.473	28.397	21.116
Dims=3		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.737	2.798	2.822	2.836	2.828
	0.01	2.908	2.867	2.872	2.999	2.828
	0.10	7.966	7.059	3.605	3.659	3.445
	1.00	24.821	24.823	24.766	23.906	3.928
	3.00	27.675	27.675	27.676	27.585	19.842
	10.00	28.423	28.422	28.425	28.418	21.111
Dims=5		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.335	3.281	3.859	3.863	3.377
	0.01	3.382	3.301	3.869	3.884	3.376
	0.10	6.396	5.889	3.994	4.124	3.442
	1.00	25.369	25.369	25.279	24.591	3.726
	3.00	28.142	28.142	28.108	28.173	19.672
	10.00	28.226	28.226	28.214	28.140	20.971

Table 6.12: Dimensionality estimates for the *correlated helix* model with $N = 5000$ structures of length 20 using several correlated folding/unfolding events across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. The number of correlated dimensions in each ensemble is provided for reference.

		Amplitude Smoothing Fraction				
Dims=2		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.882	1.980	1.938	1.961	1.976
	0.01	3.279	2.662	2.674	2.826	1.983
	0.10	12.353	11.322	5.059	4.917	3.058
	1.00	26.697	26.696	26.538	25.353	4.111
	3.00	29.949	29.949	29.942	29.954	21.520
	10.00	30.354	30.353	30.342	30.329	21.222
Dims=3		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.795	2.830	2.857	2.861	2.906
	0.01	3.053	2.946	3.004	3.217	2.906
	0.10	9.809	7.969	4.153	4.131	3.218
	1.00	26.785	26.778	26.675	24.521	3.801
	3.00	29.973	29.974	29.965	29.933	21.136
	10.00	30.313	30.314	30.315	30.348	20.640
Dims=5		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.769	3.749	4.103	3.980	3.809
	0.01	3.833	3.768	4.122	4.010	3.845
	0.10	7.713	6.287	4.358	4.237	3.905
	1.00	27.117	27.116	27.012	25.556	4.316
	3.00	29.939	29.939	29.924	29.940	22.082
	10.00	30.253	30.253	30.247	30.273	21.320

Table 6.13: Dimensionality estimates for the *correlated helix* model with $N = 2000$ structures of length 25 using several correlated folding/unfolding events across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. The number of correlated dimensions in each ensemble is provided for reference.

		Frequency Smoothing Fraction				
Dims=2		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.857	1.852	1.881	1.895	1.923
	0.01	2.521	2.525	2.524	2.513	2.303
	0.10	11.032	11.014	10.907	10.694	6.231
	1.00	28.861	28.926	28.624	28.465	12.737
	3.00	33.405	33.406	33.501	33.392	16.910
	10.00	33.824	33.935	33.698	33.590	17.028
Dims=3		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.741	2.750	2.752	2.754	2.795
	0.01	2.909	2.908	2.901	2.896	2.868
	0.10	8.361	8.331	8.136	7.808	4.677
	1.00	29.660	29.568	29.229	28.694	12.780
	3.00	33.723	33.724	33.207	32.931	16.824
	10.00	33.479	33.447	33.290	32.942	16.983
Dims=5		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.342	3.341	3.336	3.322	3.335
	0.01	3.387	3.385	3.376	3.360	3.344
	0.10	6.342	6.268	5.977	5.671	3.940
	1.00	29.863	29.801	29.514	29.050	12.815
	3.00	33.445	33.478	33.637	33.455	16.896
	10.00	33.322	33.359	33.368	33.323	17.000

Table 6.14: Dimensionality estimates for the *correlated helix* model with $N = 5000$ structures of length 25 using several correlated folding/unfolding events across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. The number of correlated dimensions in each ensemble is provided for reference.

		Frequency Smoothing Fraction				
Dims=2		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.884	1.886	1.910	1.935	1.930
	0.01	3.304	3.277	3.237	3.210	2.708
	0.10	13.780	13.774	13.659	13.377	8.281
	1.00	31.542	31.580	31.180	31.022	14.823
	3.00	36.336	36.248	36.175	36.174	19.041
	10.00	36.250	36.265	36.088	35.934	19.120
Dims=3		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.796	2.798	2.802	2.799	2.829
	0.01	3.051	3.048	3.042	3.027	2.954
	0.10	10.440	10.385	10.154	9.870	5.914
	1.00	31.579	31.559	31.305	30.938	14.614
	3.00	36.247	36.252	36.056	35.936	18.983
	10.00	36.080	36.068	36.134	35.791	19.088
Dims=5		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.763	3.761	3.755	3.746	3.746
	0.01	3.822	3.820	3.808	3.795	3.761
	0.10	7.739	7.645	7.285	6.859	4.538
	1.00	32.124	32.109	31.861	31.367	14.602
	3.00	35.936	35.908	35.955	35.911	18.978
	10.00	35.815	35.839	35.685	35.706	19.128

Table 6.15: Dimensionality estimates for the *correlated helix* model with $N = 2000$ structures of length 25 using several correlated folding/unfolding events across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. The number of correlated dimensions in each ensemble is provided for reference.

		Amplitude Smoothing Fraction				
Dims=2		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.857	1.907	1.900	1.913	1.843
	0.01	2.521	2.290	2.361	2.126	1.847
	0.10	11.032	10.316	4.159	4.109	1.977
	1.00	28.861	28.857	28.737	27.459	3.559
	3.00	33.405	33.404	33.417	33.380	24.797
	10.00	33.824	33.824	33.843	33.830	24.391
Dims=3		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.741	2.801	2.826	2.840	2.830
	0.01	2.909	2.876	2.888	3.064	3.125
	0.10	8.361	7.280	3.668	3.769	3.274
	1.00	29.660	29.658	29.741	27.922	3.847
	3.00	33.723	33.723	33.728	33.666	25.132
	10.00	33.479	33.478	33.483	33.457	24.323
Dims=5		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.342	3.289	3.915	3.919	3.412
	0.01	3.387	3.309	3.925	3.972	3.412
	0.10	6.342	5.813	4.036	4.202	3.540
	1.00	29.863	29.860	29.978	29.112	3.833
	3.00	33.445	33.445	33.465	33.517	25.010
	10.00	33.322	33.322	33.338	33.354	24.479

Table 6.16: Dimensionality estimates for the *correlated helix* model with $N = 5000$ structures of length 25 using several correlated folding/unfolding events across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. The number of correlated dimensions in each ensemble is provided for reference.

		Amplitude Smoothing Fraction				
Dims=2		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.884	1.980	1.939	1.961	1.976
	0.01	3.304	2.678	2.762	2.941	1.985
	0.10	13.780	12.589	5.434	5.122	2.938
	1.00	31.542	31.544	31.589	29.571	3.902
	3.00	36.336	36.335	36.362	36.335	25.268
	10.00	36.250	36.250	36.261	36.205	25.937
Dims=3		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.796	2.830	2.859	2.861	2.910
	0.01	3.051	2.943	3.014	3.125	2.910
	0.10	10.440	8.312	4.231	4.207	3.219
	1.00	31.579	31.577	31.438	29.062	3.804
	3.00	36.247	36.249	36.267	36.209	25.637
	10.00	36.080	36.081	36.066	36.089	26.478
Dims=5		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.763	3.738	4.144	4.008	3.846
	0.01	3.822	3.757	4.156	4.042	3.846
	0.10	7.739	6.261	4.397	4.280	4.020
	1.00	32.124	32.121	31.958	30.691	4.353
	3.00	35.936	35.935	35.939	35.851	26.042
	10.00	35.815	35.814	35.823	35.764	26.626

dynamics is still the key idea, it has been shown that the the estimator is unable to accurately assess polymer systems of like variety. Instead, given even small amounts of noise, the estimator will ascribe the highest dimensionality that it can (minus some accuracy due to undersampling and open manifold end effects). However, if some portions of the polymers have already folded or correlated motions are being exhibited, the estimator will provide a lower value. The estimator was able to accurately determine relative dimensionality under all conditions, so that these systems were discernible from one another even under these conditions.

Aggregate Estimates

The aggregate dimensionality for each protein simulation (and corresponding forcefield) was determined for the first 100ns annealing phase of the simulations, and for the 250ns production phase of the simulations. The Φ and Ψ angle space of the protein conformations from the simulations was computed, and the nearest neighbors distances for all structures in a single simulation were calculated using Euclidean distance between the $\sin - \cos$ transform of the angle vectors. Pointwise dimensionality estimates were obtained for three sets of k nearest neighbors to assess the results at different scales: small $k = [2, 3, 4, 6]$, medium $k = [2, 3, 4, 6, 8, 16, 32, 64]$, and large $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$. The harmonic mean of the estimates was calculated across for each of the 10 replicate simulations and boxplots of the resulting estimates are shown in Figures 6.8-6.10. In addition, the normalized dimensionality estimates are shown in Figures 6.11-6.13. These are obtained by normalizing the dimensionality estimates by the total number of degrees of freedom in each protein (equivalent to the number of Φ and Ψ angles in each): GB1-30, Trp-cage-38, Nsp1-48, and Nup116-48. This allows for a more balanced comparison between whether the different proteins are of higher/lower dimensionality compared when amongst each other, since the dimensionality estimates tend to remain in relative rank order according to all of the polymer results.

For the unnormalized results reported in Figures 6.8-6.10, it is clear that the annealing portions of the trajectories show different dimensionality than the production portions. This is not surprising since the temperature at the annealing stage is high enough to disrupt any collapsed structures that would normally form at the production temperature of 300K. In particular, the IDPs, Nsp1 and Nup116, are both higher predicted dimensionality than either of the NFPs, GB1 or Trp-cage. However, the lengths of these sets of proteins differ, and a quick comparison to the normalized dimensionality estimates makes it clear that the longer IDP proteins are actually undergoing dynamics that fill fractionally fewer of their available degrees of freedom than the NFPs. Overall, the unnormalized estimates are about 50% lower than their target values. Since the higher temperature annealing stage is very likely to be pressing these systems to utilize all of their degrees of freedom similar to a the high-noise case of the semirigid polymer, this is more than likely due to limitations in sampling and underestimation at the manifold boundaries. These results are robust across all forcefields utilized here. Taking a closer look at the estimates for the production portion of the simulations shows that the estimates now drop considerably, but the trend of the IDPs utilizing fewer degrees of freedom than the NFPs remains. According to the polymer studies, a lower dimensionality would be predicted if there are more partially formed structures, or correlated motions in the IDPs, or essentially less frustrated dynamics. The NFPs would therefore be exhibiting very high dimensionality motion due to remaining in tightly-packed, frustrated, or possibly even folded structures.

In order to assess the structural properties of the simulations, the average radii of gyration, R_g, and shape parameters, S, were calculated according to the methods described Section 2.2.3 [15]. Figure 6.14 shows the distribution of these measurements for the production simulations. The R_g data suggest that both GB1 and Trp-cage are more compact than either Nsp1 or Nup116 in all forcefields, but some of this difference can be attributed to the slightly longer chain lengths for Nsp1 and Nup116. However, the S data indicate that the GB1 structures are very extended, while the Trp-cage protein is very spherical. The IDPs are both fairly spherical as well. So, these data alone cannot

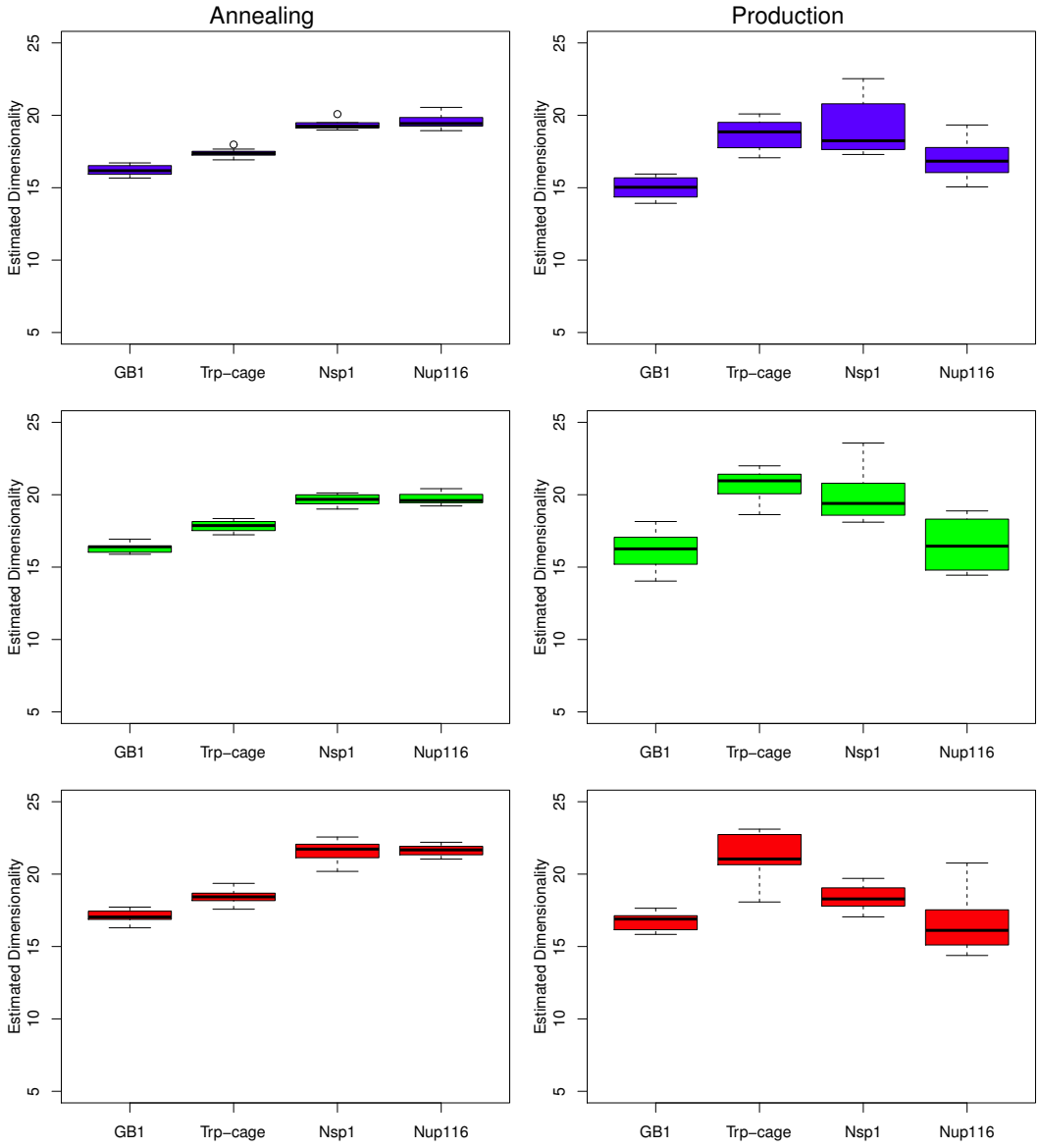

Figure 6.8: Distributions of dimensionality estimation results for (left) 100ns annealing (300K-600K-300K) and (right) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: (top,blue) ff99SB-ILDN, (middle,green) ff99SB-PSN, and (bottom,red) ff03w. Estimates were obtained by taking the harmonic mean across small values of $k = [2, 3, 4, 6]$ for each replicate simulation.

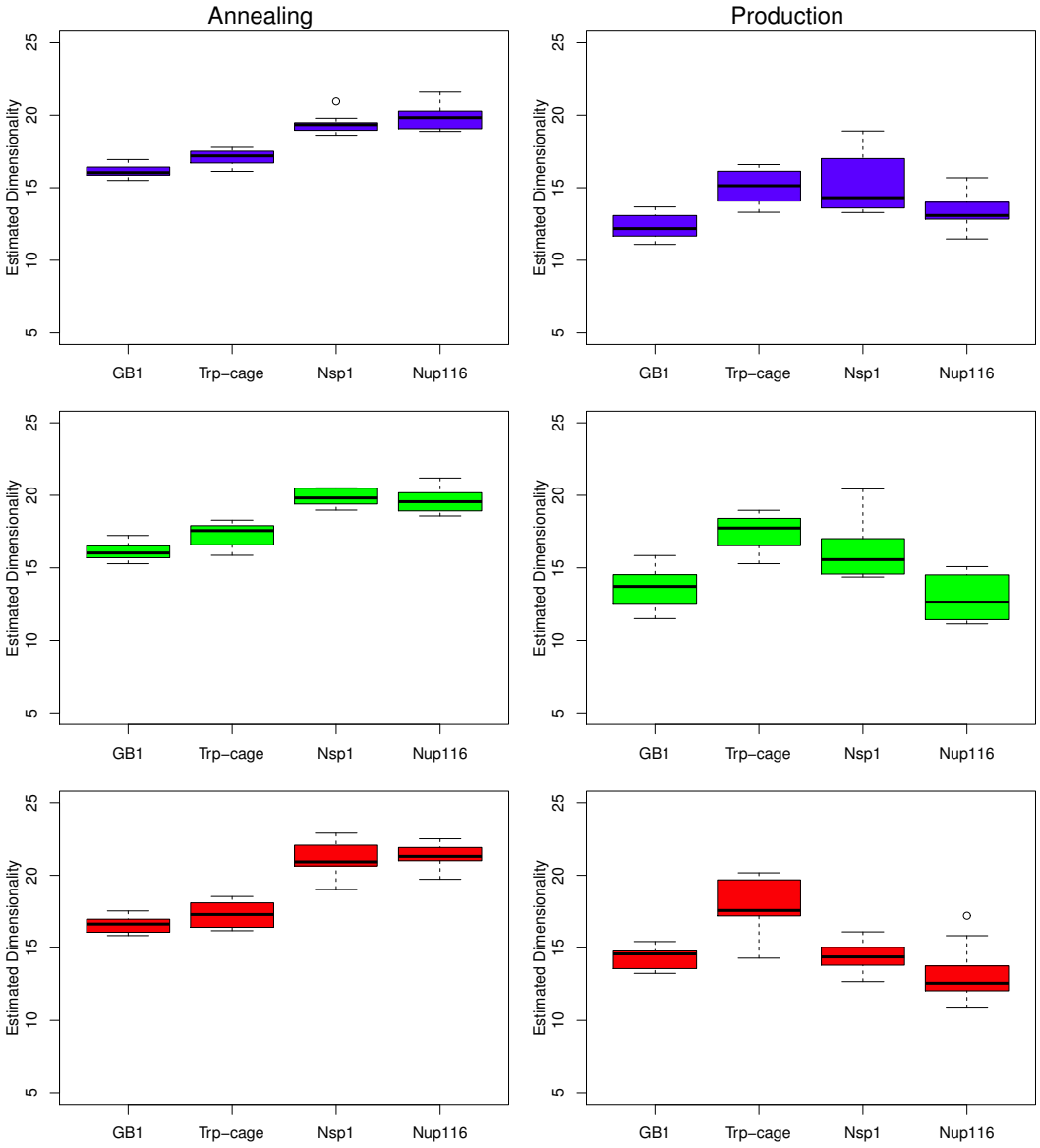

Figure 6.9: Distributions of dimensionality estimation results for (left) 100ns annealing (300K-600K-300K) and (right) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: (top,blue) ff99SB-ILDN, (middle,green) ff99SB-PSN, and (bottom,red) ff03w. Estimates were obtained by taking the harmonic mean across medium values of $k = [2, 3, 4, 6, 8, 16, 32, 64]$ for each replicate simulation.

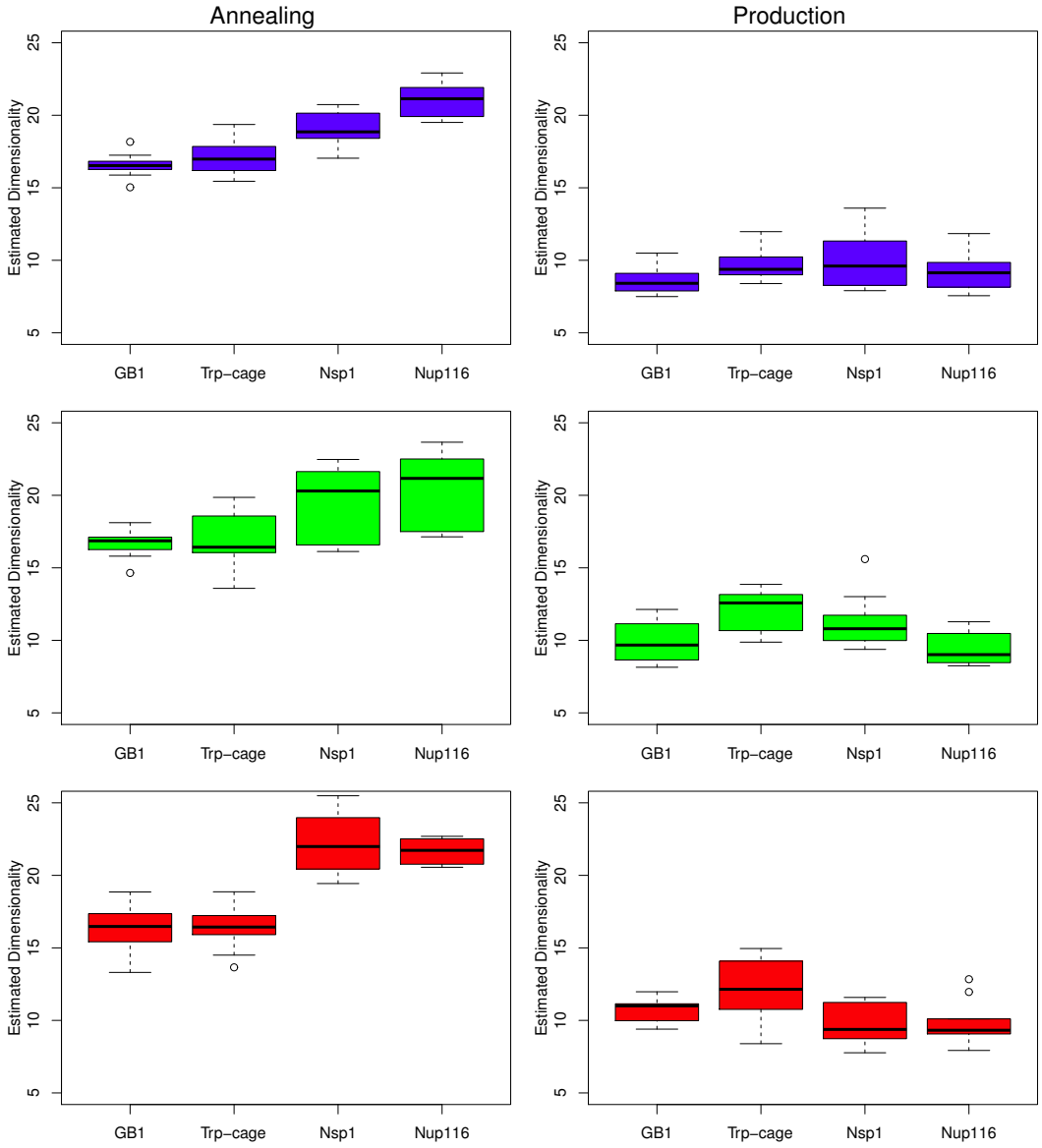

Figure 6.10: Distributions of dimensionality estimation results for (left) 100ns annealing (300K-600K-300K) and (right) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: (top,blue) ff99SB-ILDN, (middle,green) ff99SB-PSN, and (bottom,red) ff03w. Estimates were obtained by taking the harmonic mean across large values of $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$ for each replicate simulation.

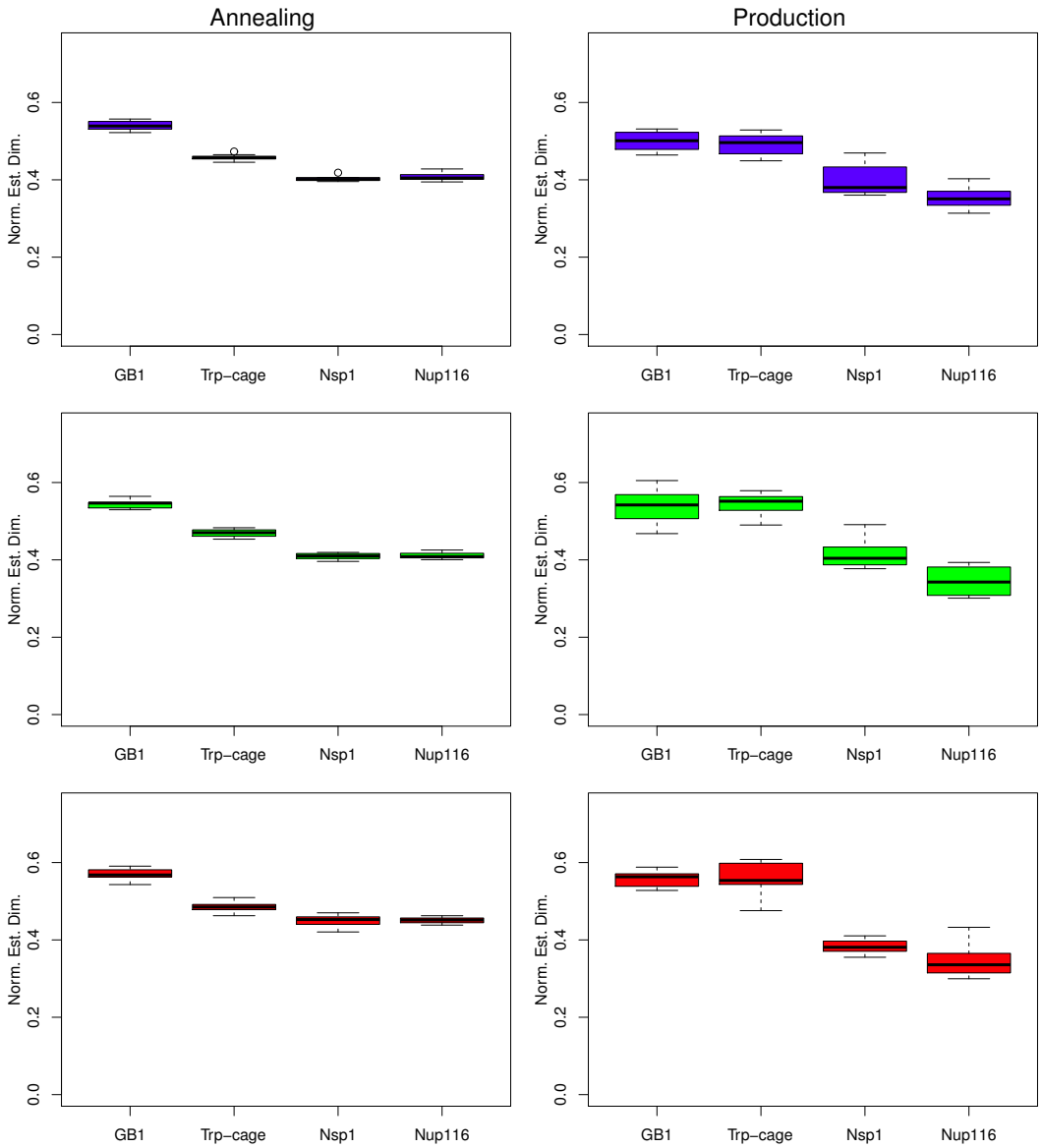

Figure 6.11: Distributions of normalized dimensionality estimation results for (left) 100ns annealing (300K-600K-300K) and (right) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: (top,blue) ff99SB-ILDN, (middle,green) ff99SB-PSN, and (bottom,red) ff03w. Estimates were obtained by taking the harmonic mean across small values of $k = [2, 3, 4, 6]$ for each replicate simulation and then dividing by the total number of degrees of freedom in the respective protein.

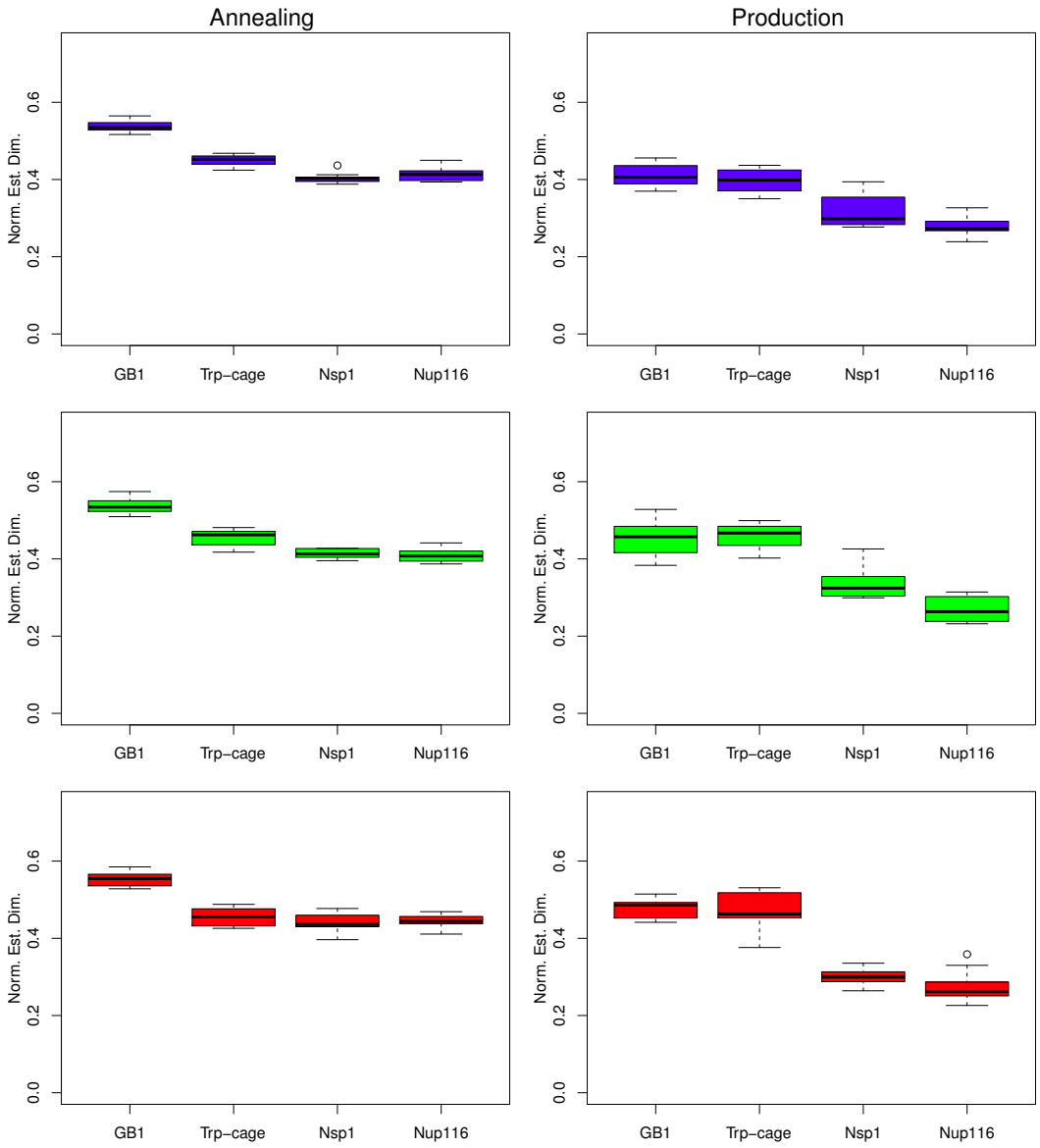

Figure 6.12: Distributions of normalized dimensionality estimation results for (left) 100ns annealing (300K-600K-300K) and (right) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: (top,blue) ff99SB-ILDN, (middle,green) ff99SB-PSN, and (bottom,red) ff03w. Estimates were obtained by taking the harmonic mean across medium values of $k = [2, 3, 4, 6, 8, 16, 32, 64]$ for each replicate simulation and then dividing by the total number of degrees of freedom in the respective protein.

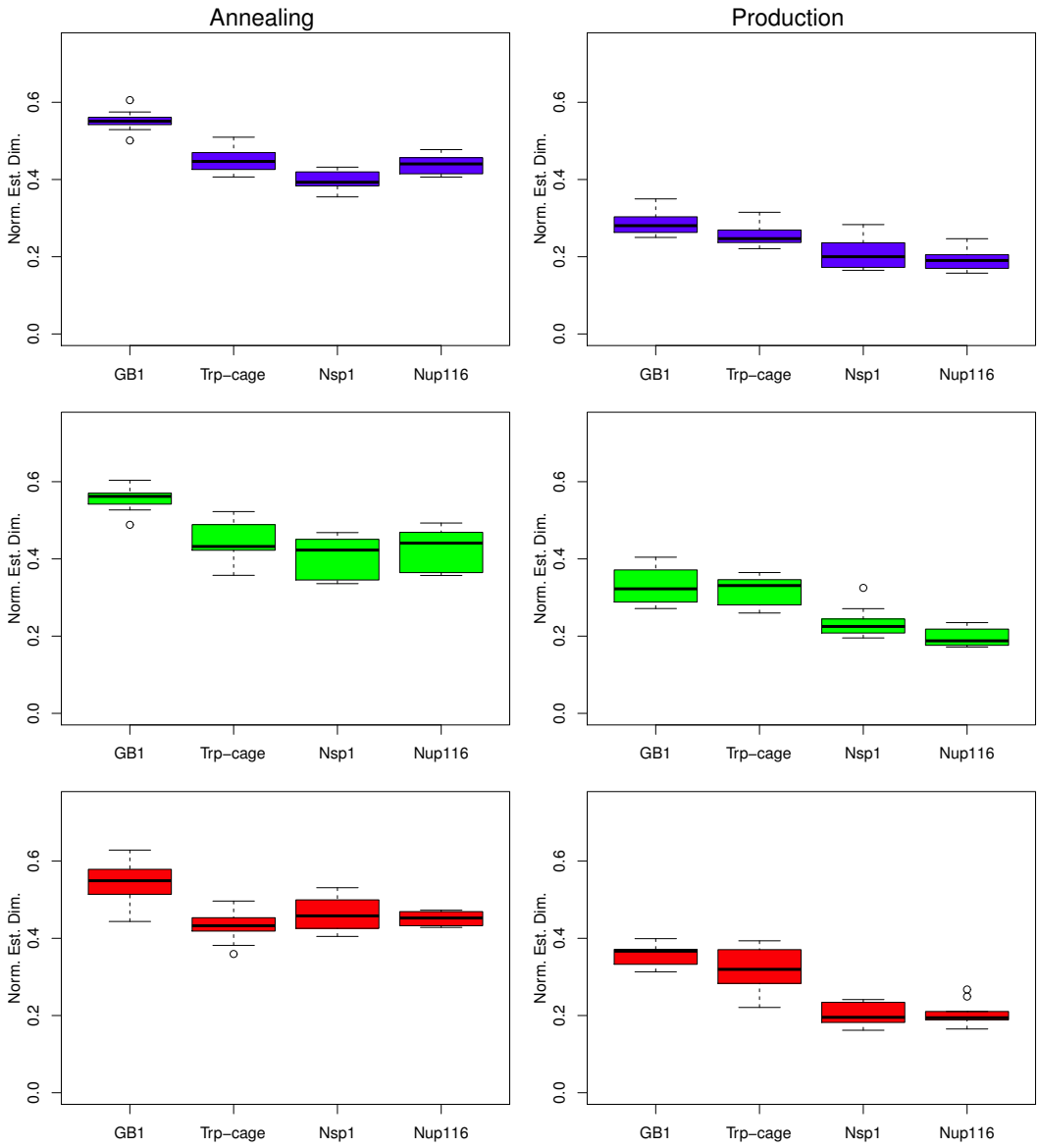

Figure 6.13: Distributions of normalized dimensionality estimation results for (left) 100ns annealing (300K-600K-300K) and (right) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: (top,blue) ff99SB-ILDN, (middle,green) ff99SB-PSN, and (bottom,red) ff03w. Estimates were obtained by taking the harmonic mean across large values of $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$ for each replicate simulation and then dividing by the total number of degrees of freedom in the respective protein.

show how both GB1 and Trp-cage exhibit higher-dimensionality dynamics than the IDPs given their very structural differences.

To help lend more information to this analysis, the secondary structure assignment [16] for all of the structures was computed. The results of this analysis are shown in Figures 6.15-6.18 for all four proteins. Nsp1 exhibits a central, mainly helical region with two flanking helical regions attached at two distinct points, residues 8 and 19, where the protein is very disordered. Nup116 has two helical regions with a disordered center at residue 12. However, GB1 has adopted a long helical conformation, for ff99SB-PSN and ff03w, and less so for ff99SB-ILDN. This would be consistent with the S parameter data showing a more prolate structural arrangement versus the other proteins. Trp-cage has a strong helical region on one end, but also a long tail on the other end that lacks a particular structural class assignment according to DSSP. This is reminiscent of the folded structure of Trp-cage, so this tail may actually be packed against the helix as in the native structure. This corroborates the S data for this protein which indicated a tight, spherical conformation. While the conformation that GB1 adopts is not close to the native state, it is a well-defined structure according to DSSP. Hence both Trp-cage and GB1 appear to be mainly sampling very trapped or folded conformations, which would agree with the dimensionality estimates. On the other hand, the specific disordered regions and end domains of the FG-Nups are consistent with the idea that the structures are intermediate in nature, and lower dimensional, in agreement with the dimensionality estimates as well.

Individual Replicate Estimates

In order to further investigate if the dimensionality estimates are in-fact predicting more tightly-packed, folded structures for the NFPs, the pointwise dimensionality estimates for representative simulations for each forcefield for all four proteins were plotted versus time. In addition, the RMSD from the folded structure versus time, or from the average structure in the case of the IDPs, was also plotted versus time. The re-

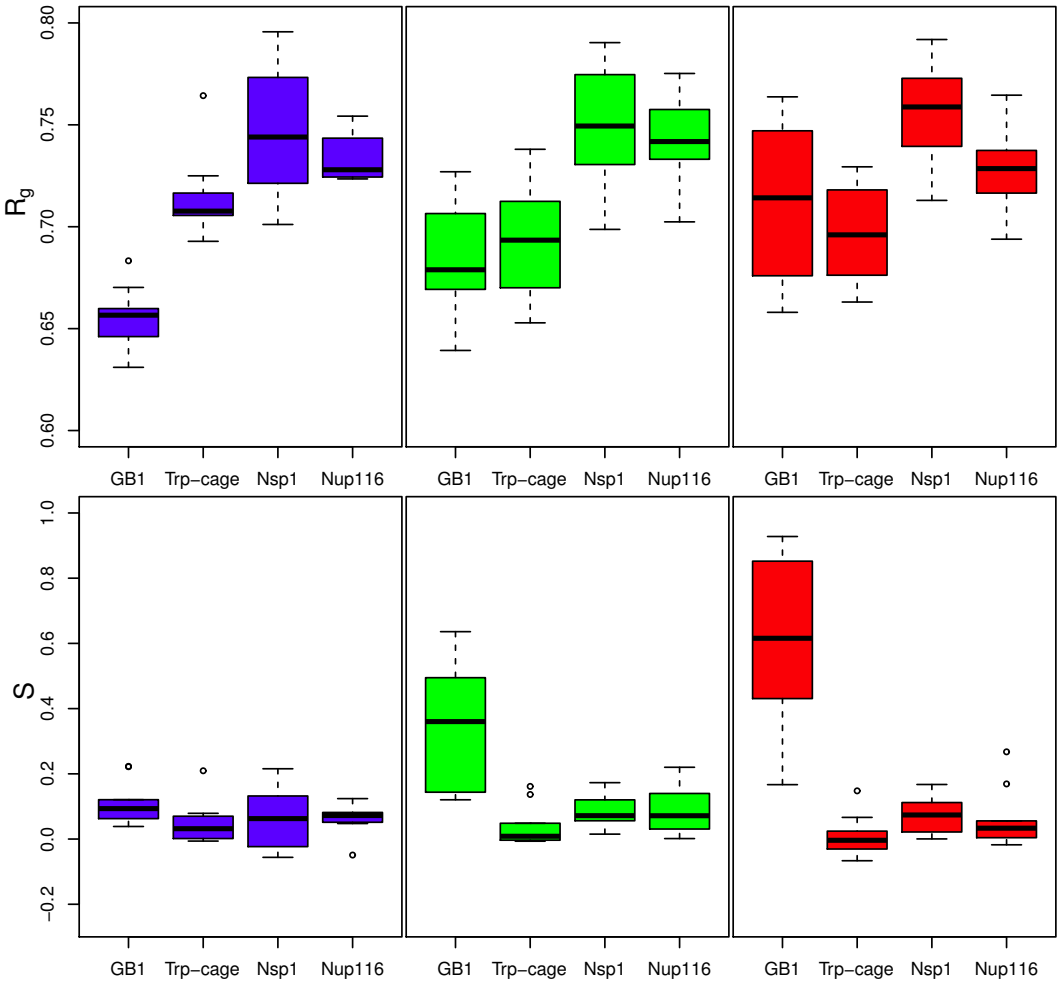

Figure 6.14: Distribution of the average R_g and S parameters across the ten replicate production MD simulations (250ns @ 300K) of GB1, Trp-cage, Nsp1, and Nup116 using the Amber (blue) ff99-ILDN, (green) ff99SB-PSN, and (red) ff03w forcefields.

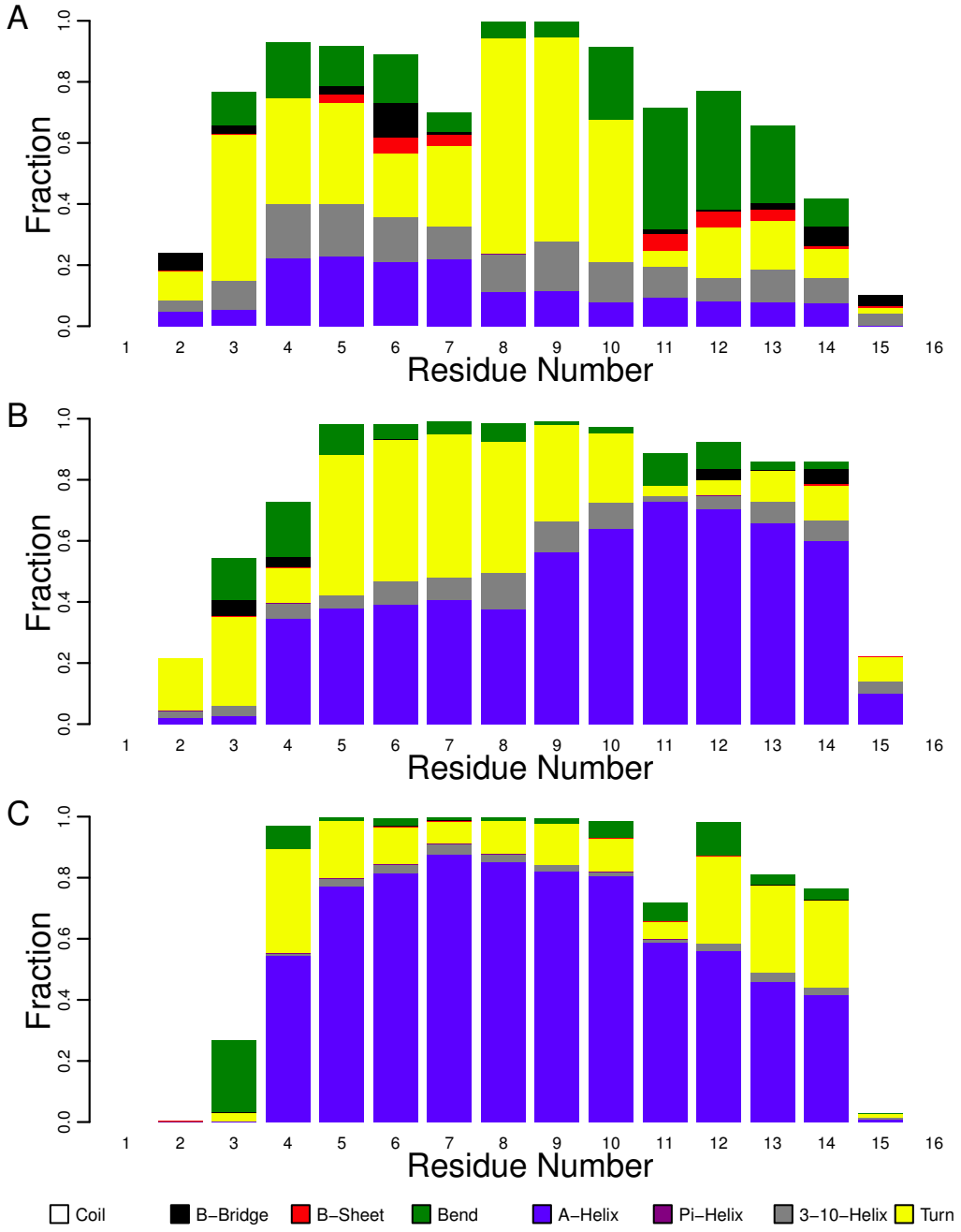

Figure 6.15: GB1 Secondary Structure – 250ns @ 300K – Amber (A) ff99SB-ILDN (B) ff99SB-PSN (C) ff03w

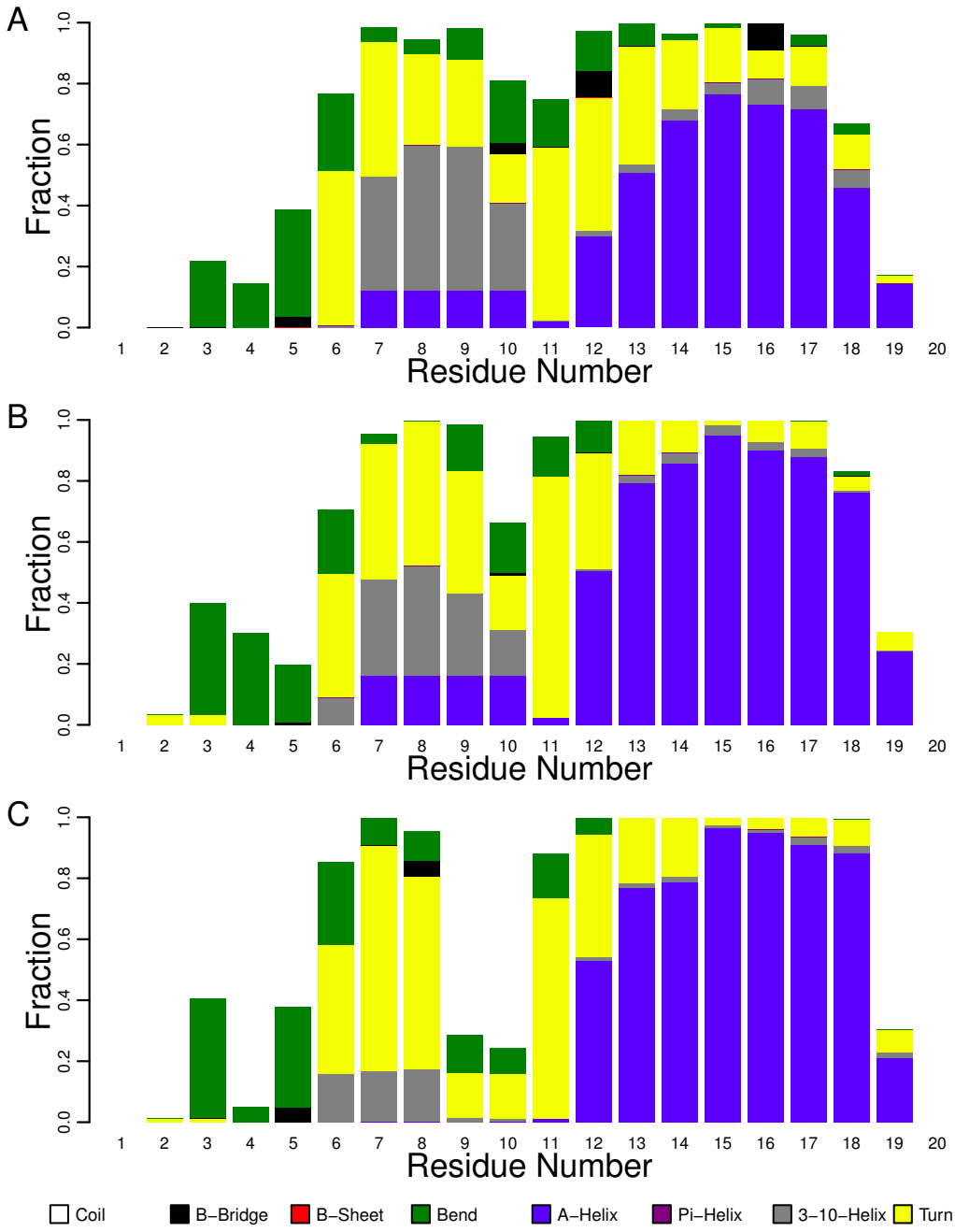

Figure 6.16: Trp-cage Secondary Structure – 250ns @ 300K – Amber (A) ff99SB-ILDN (B) ff99SB-PSN (C) ff03w

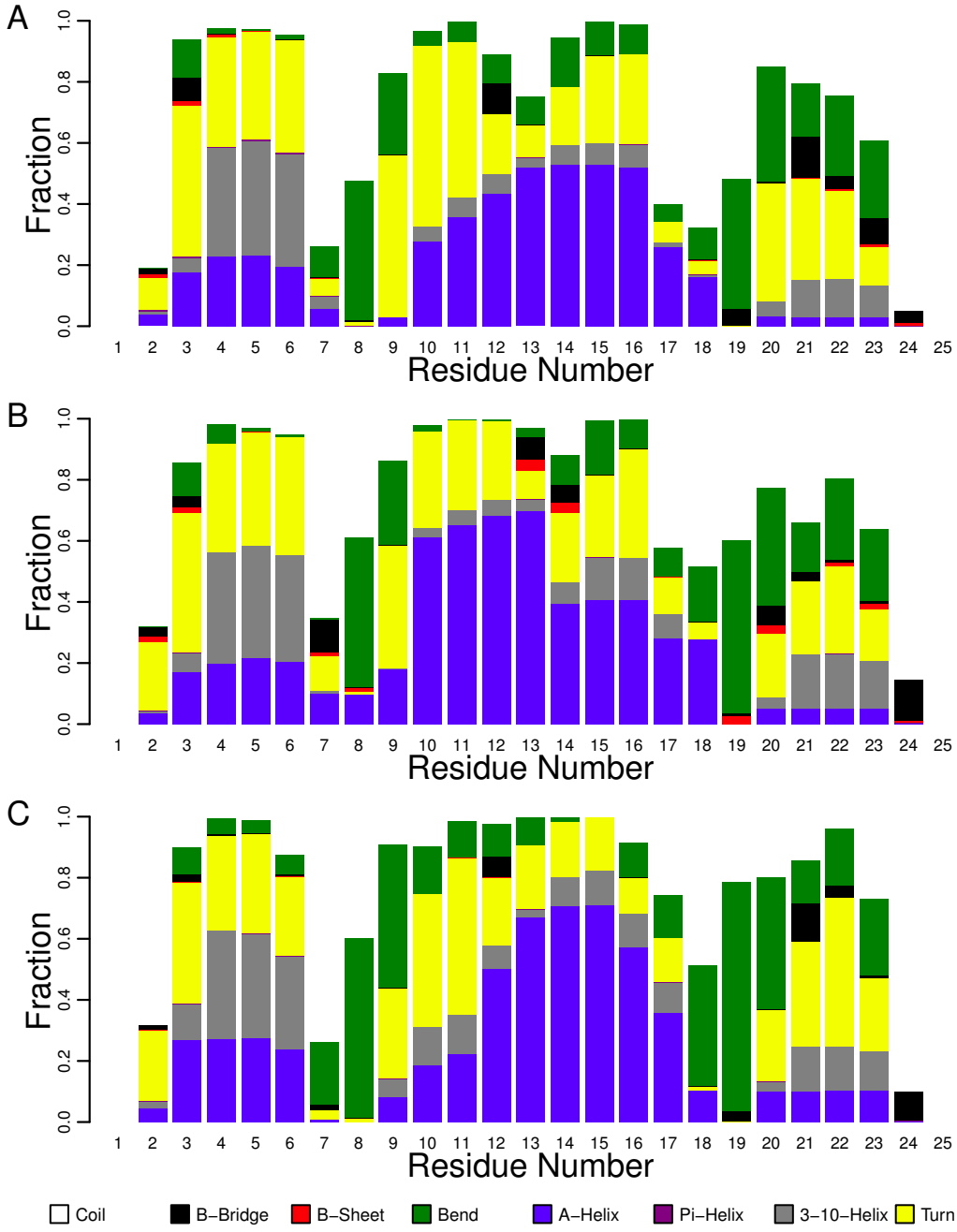

Figure 6.17: Nsp1 Secondary Structure – 250ns @ 300K – Amber (A) ff99SB-ILDN (B) ff99SB-PSN (C) ff03w

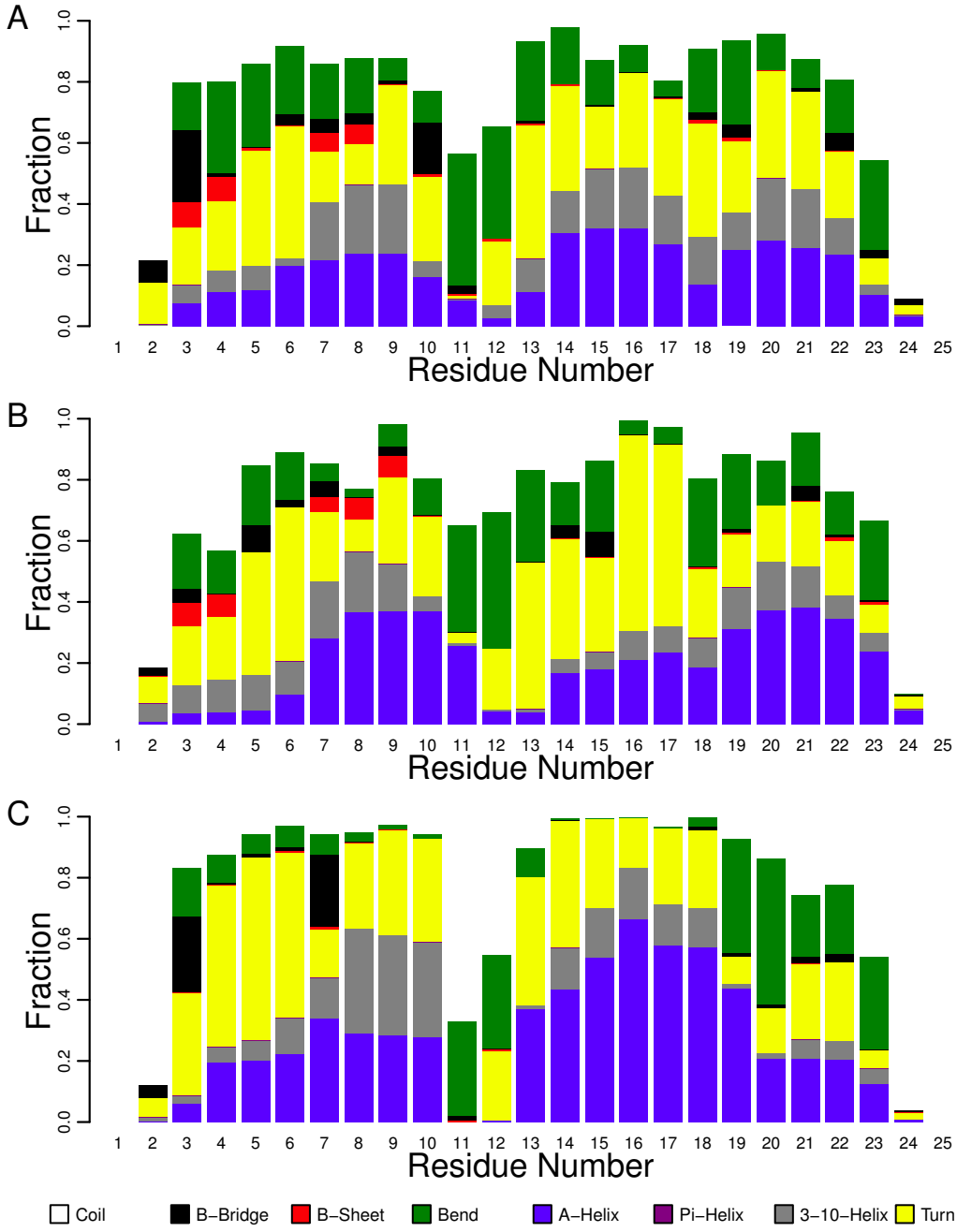

Figure 6.18: Nup116 Secondary Structure – 250ns @ 300K – Amber (A) ff99SB-ILDN (B) ff99SB-PSN (C) ff03w

sults were smoothed using a 0.5ns window for clarity as both RMSD and dimensionality estimates are extremely noisy from frame-to-frame. Side-by-side comparison of these plots facilitates the discovery of portions of the trajectory that display corresponding shifts in RMSD and the dimensionality estimates. The replicates chosen for these plots represented the best fit to the reference structures used. Hence for GB1 and Trp-cage, they were the replicates that best matched the folded state. For Nsp1 and Nup116, the replicates best matching the average conformation were chosen.

The results for GB1 are shown in Figure 6.19. The most striking feature of these plots is the essentially constant estimates of dimensionality across the production portions of the runs even though there are clearly large, fast structural transitions that occur according to the plot of RMSD versus time. This indeed appears to be a very frustrated system that is constantly attempting to fold, but is restricted to suboptimal conformational states. As a result, the thermal fluctuations of the system act as noise, which drive the dimensionality estimates inordinately high.

The results for Trp-cage are shown in Figures 6.20. According to the RMSD versus time plot in part B of these figures, all three simulations come very close to folding the protein. The simulation for ff03w gets especially close, and exhibits the highest estimated dimensionality as a result. The remaining two simulations do not quite adopt the folded structure, and the resulting estimates are a little lower, indicating additional disorder compared to the ff03w simulation.

The results for Nsp1 are shown in Figure 6.21. According to the RMSD plot, this protein undergoes regular, rather minor structural transitions without becoming frustrated at any particular location. Likewise, the dimensionality estimates follow the trend of the RMSD, but remain rather small compared to the two NFPs. The results for Nup116 shown in Figure 6.22 are extremely similar to the results for Nsp1.

Additionally, the effect of DFT smoothing of the Φ-Ψ angles using a fractional frequency cutoff of 0.5 are shown in part D of the individual simulation results. Overall, the largest effects appear minimal for the large k values in these plots, but are more pronounced at smaller scales (smaller values of k). Tables 6.17-6.22 provide a detailed

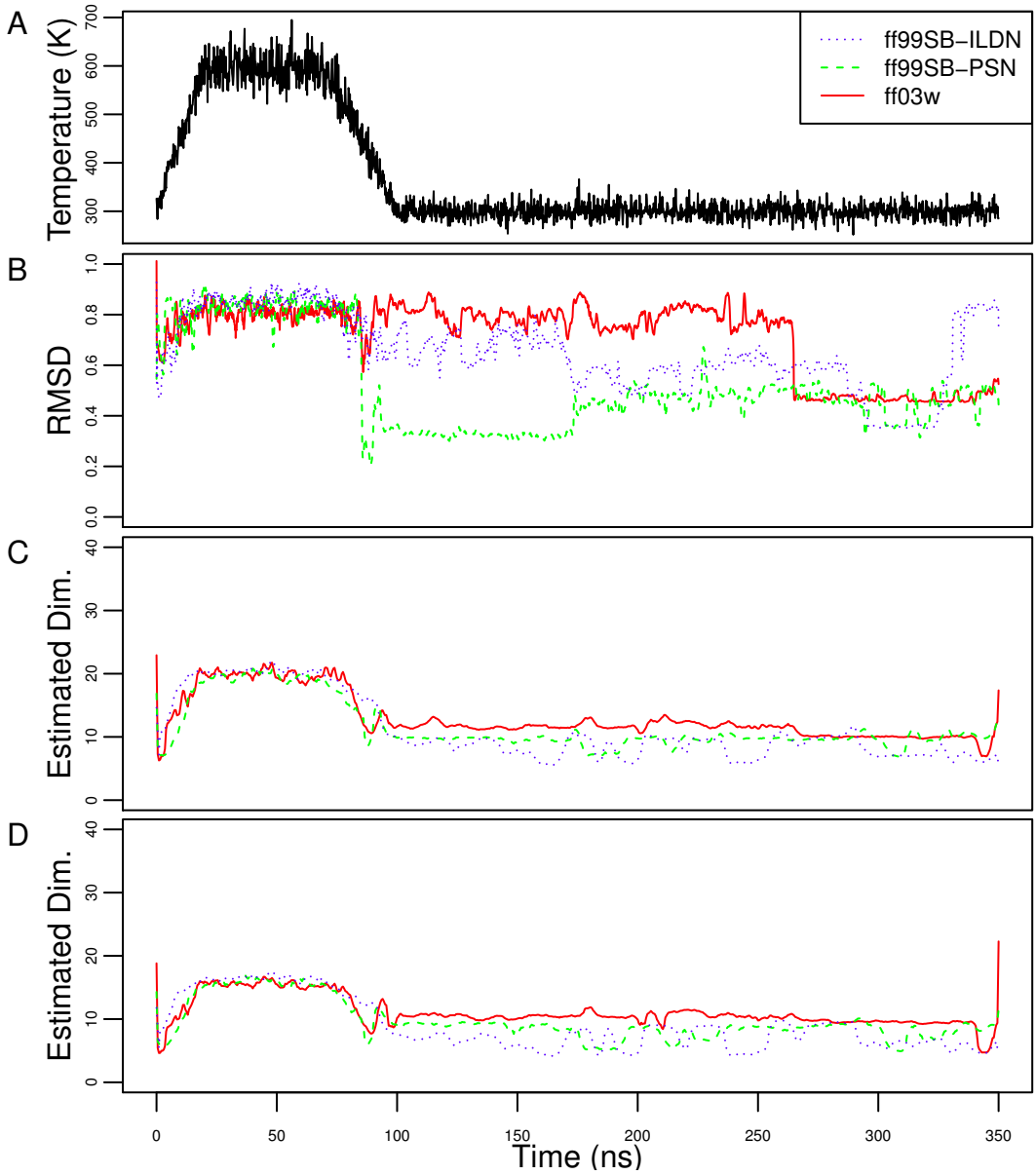

Figure 6.19: Dimensionality estimation results for representative GB1 simulations using each of the three forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. (A) Plot of temperature versus time, (B) RMSD from the folded structure versus time, (C) pointwise dimensionality estimates and (D) pointwise dimensionality estimates after applying DFT smoothing using a fractional frequency cutoff of 0.5. Point estimates were obtained by taking the harmonic mean across all values of $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$.

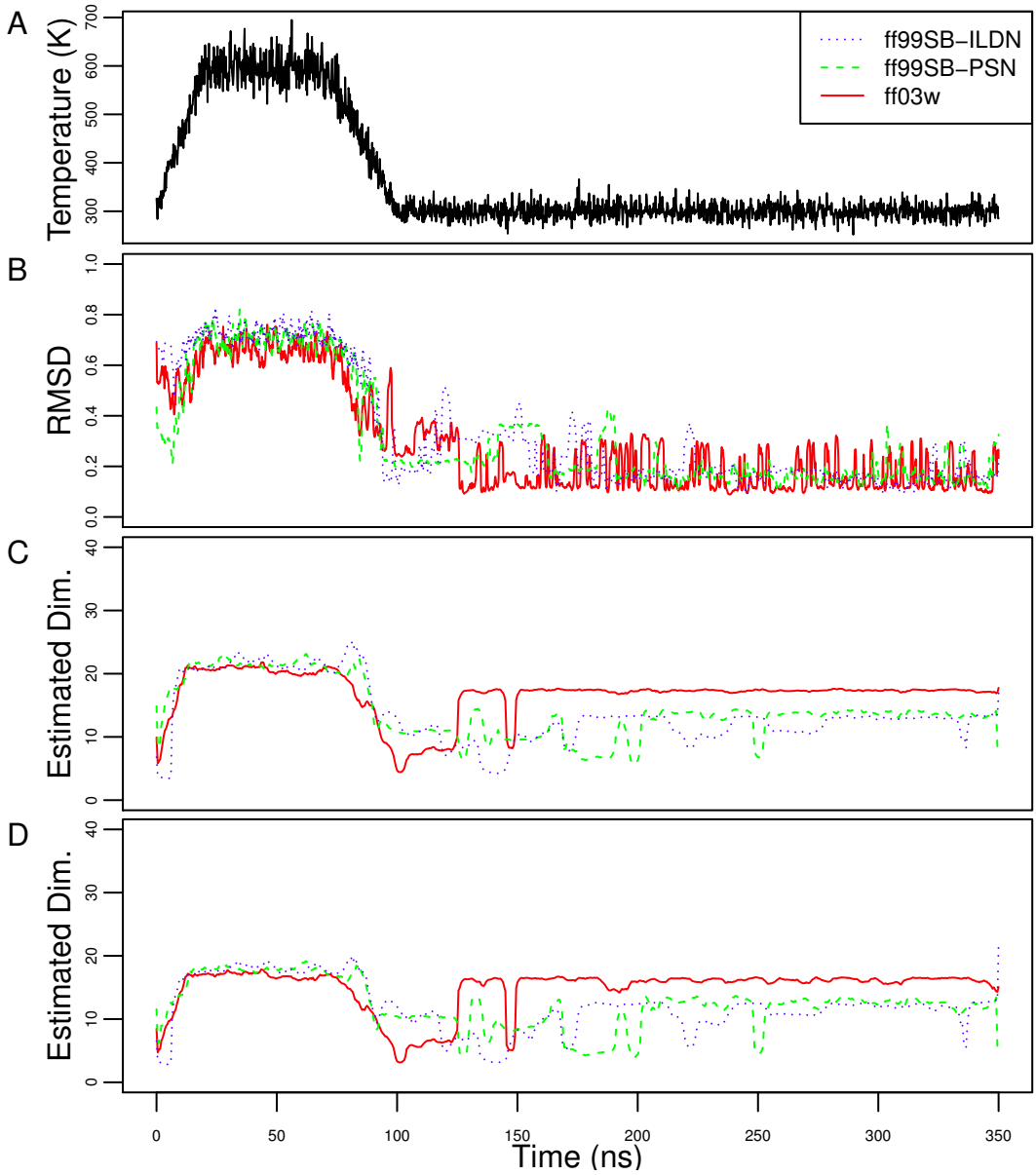

Figure 6.20: Dimensionality estimation results for representative Trp-cage simulations using each of the three forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. (A) Plot of temperature versus time, (B) RMSD from the folded structure versus time, (C) pointwise dimensionality estimates and (D) pointwise dimensionality estimates after applying DFT smoothing using a fractional frequency cutoff of 0.5. Point estimates were obtained by taking the harmonic mean across all values of $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$.

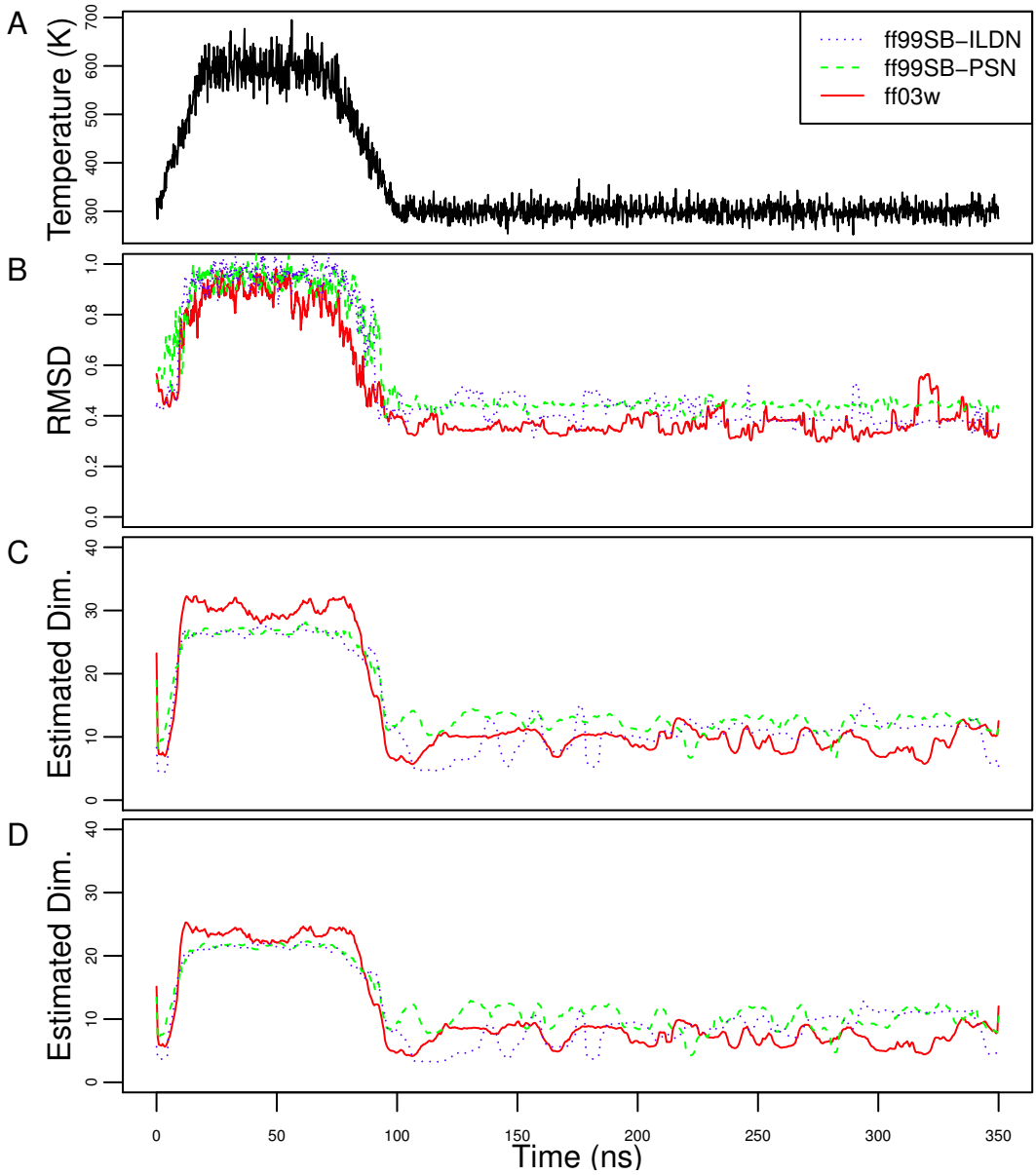

Figure 6.21: Dimensionality estimation results for representative Nsp1 simulations using each of the three forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. (A) Plot of temperature versus time, (B) RMSD from the average structure versus time, (C) pointwise dimensionality estimates and (D) pointwise dimensionality estimates after applying DFT smoothing using a fractional frequency cutoff of 0.5. Point estimates were obtained by taking the harmonic mean across all values of $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$.

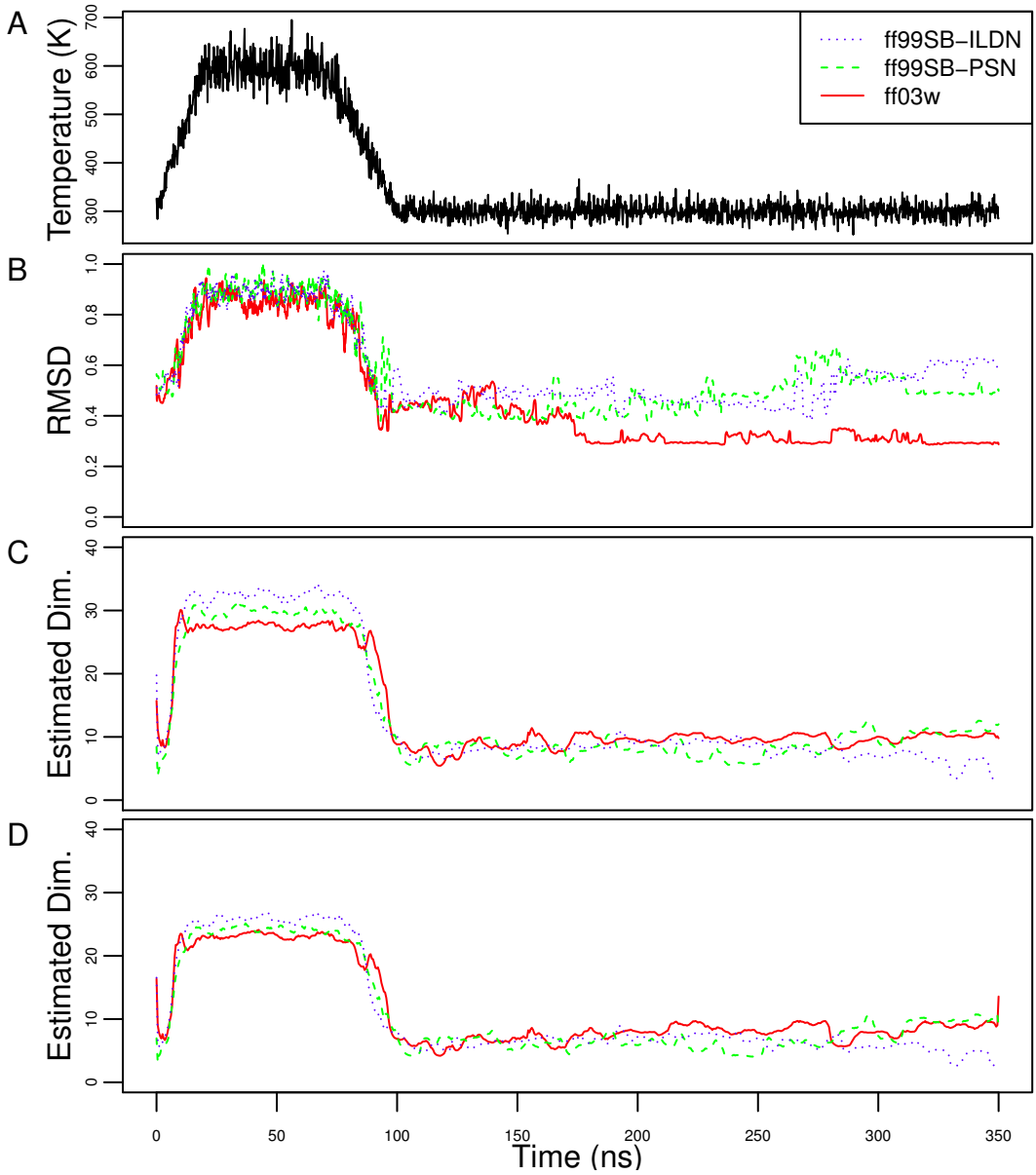

Figure 6.22: Dimensionality estimation results for representative Nup116 simulations using each of the three forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. (A) Plot of temperature versus time, (B) RMSD from the average structure versus time, (C) pointwise dimensionality estimates and (D) pointwise dimensionality estimates after applying DFT smoothing using a fractional frequency cutoff of 0.5. Point estimates were obtained by taking the harmonic mean across all values of $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$.

breakdown of the smoothing results at small, medium, and large k. Possibly the most telling result from this analysis is the fact that the IDPs were much more sensitive to the smoothing effects at small k. In particular, using the frequency fraction cutoff, the dimensionality estimates, which were higher for the IDPs than the NFPs in the unsmoothed results, were lower for the IDPs than the NFPs under heavily smoothed conditions (large fractional cutoffs). Since only the correlated helix model exhibited this degree sensitivity to the smoothing techniques, it might be that the IDPs are exhibiting more *correlated*, large amplitude motions than the NFPs. Although this prediction is based on the polymer results, future work is needed to determine if this result is robust across a wider range of intrinsically disordered proteins.

6.4 Discussion

This chapter introduced a polymer framework for examining the utility of dimensionality estimation algorithms for studying MD simulations of proteins. The key contribution of this framework is the development and use of several polymer models which exhibit well-defined dynamics of known dimensionality. The models include noise, partially and fully folded structures, and correlated motions. The polymers provide an informative framework for testing the ability of the maximum likelihood estimator of dimensionality, a dimensionality estimation algorithm which has received considerable attention in the machine learning and physics communities due to its simplicity and effectiveness, to estimate the dimensionality of several polymer model ensembles where the dimensionality is known *a priori*. The effects of small sample sizes and noisy sampling are treated directly, and the shortcomings of the method in these cases are acknowledged. While precise estimates of the dimensionality were not attained, the systematic bias of the estimator made it effective in maintaining the rank-ordering of the dimensionality estimates. Therefore, it can still be useful for determining differences between protein classes.

The dimensionality estimator was used to compare the dynamics of natively

Table 6.17: Dimensionality estimation results across all levels of DFT smoothing using a fractional *frequency* cutoff for the (top) 100ns annealing (300K-600K-300K) and (bottom) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. Estimates were obtained by taking the harmonic mean across small values of $k = [2, 3, 4, 6]$.

		Frequency Smoothing Fraction				
Annealing		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	15.930	15.812	15.234	14.350	5.244
	ff99SB-PSN	16.648	16.530	15.913	14.989	5.452
	ff03w	17.440	17.363	16.887	16.059	5.727
Trp-cage	**ff99SB-ILDN**	17.315	17.202	16.616	15.618	5.331
	ff99SB-PSN	18.097	17.986	17.465	16.440	5.693
	ff03w	17.950	17.831	17.402	16.606	6.040
Nsp1	**ff99SB-ILDN**	19.134	18.923	17.927	16.385	4.899
	ff99SB-PSN	19.189	18.956	17.876	16.264	4.853
	ff03w	21.588	21.388	20.371	18.843	5.336
Nup116	**ff99SB-ILDN**	19.644	19.436	18.434	16.897	4.980
	ff99SB-PSN	19.934	19.728	18.752	17.201	5.028
	ff03w	21.252	21.063	20.084	18.693	5.351
Production		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	15.492	15.512	15.467	15.330	8.321
	ff99SB-PSN	15.082	15.053	15.036	14.973	7.844
	ff03w	16.668	16.682	16.615	16.521	8.577
Trp-cage	**ff99SB-ILDN**	18.997	18.991	18.881	18.732	8.199
	ff99SB-PSN	19.319	19.292	19.248	19.047	8.374
	ff03w	21.642	21.631	21.529	21.505	9.685
Nsp1	**ff99SB-ILDN**	19.594	19.574	19.451	19.256	8.127
	ff99SB-PSN	19.953	19.914	19.870	19.659	7.498
	ff03w	17.375	17.362	17.210	16.876	6.123
Nup116	**ff99SB-ILDN**	17.129	17.115	16.910	16.553	6.351
	ff99SB-PSN	14.649	14.612	14.394	13.978	5.459
	ff03w	17.270	17.239	17.066	16.770	6.696

folded and intrinsically disordered proteins. While it was hypothesized that the folded state of proteins was of zero dimensionality, the practical limitations of the algorithm

Table 6.18: Dimensionality estimation results across all levels of DFT smoothing using a fractional *frequency* cutoff for the (top) 100ns annealing (300K-600K-300K) and (bottom) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. Estimates were obtained by taking the harmonic mean across small values of $k = [2, 3, 4, 6, 8, 16, 32, 64]$.

		Frequency Smoothing Fraction				
Annealing		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	15.947	15.868	15.505	14.993	9.363
	ff99SB-PSN	16.779	16.696	16.327	15.800	9.846
	ff03w	17.086	17.020	16.676	16.169	10.089
Trp-cage	**ff99SB-ILDN**	17.306	17.228	16.875	16.351	10.148
	ff99SB-PSN	17.873	17.804	17.474	16.975	10.634
	ff03w	16.298	16.233	15.948	15.531	10.073
Nsp1	**ff99SB-ILDN**	19.351	19.254	18.750	18.035	10.471
	ff99SB-PSN	19.452	19.328	18.787	18.007	10.351
	ff03w	21.306	21.198	20.663	19.882	11.193
Nup116	**ff99SB-ILDN**	19.594	19.493	19.022	18.335	10.675
	ff99SB-PSN	20.272	20.166	19.664	18.939	10.910
	ff03w	21.175	21.070	20.586	19.919	11.470
Production		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	12.543	12.542	12.495	12.417	9.943
	ff99SB-PSN	12.409	12.395	12.344	12.252	9.577
	ff03w	14.388	14.385	14.336	14.259	11.081
Trp-cage	**ff99SB-ILDN**	15.450	15.443	15.370	15.278	11.386
	ff99SB-PSN	15.882	15.863	15.792	15.660	11.563
	ff03w	18.528	18.517	18.457	18.361	13.772
Nsp1	**ff99SB-ILDN**	15.809	15.788	15.702	15.549	11.359
	ff99SB-PSN	16.119	16.100	15.998	15.838	11.146
	ff03w	13.216	13.185	13.041	12.815	8.722
Nup116	**ff99SB-ILDN**	13.474	13.445	13.277	13.048	8.970
	ff99SB-PSN	11.285	11.250	11.077	10.836	7.389
	ff03w	13.638	13.608	13.463	13.258	9.365

Table 6.19: Dimensionality estimation results across all levels of DFT smoothing using a fractional *frequency* cutoff for the (top) 100ns annealing (300K-600K-300K) and (bottom) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. Estimates were obtained by taking the harmonic mean across small values of $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$.

		Frequency Smoothing Fraction				
Annealing		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	17.164	17.106	16.845	16.518	13.422
	ff99SB-PSN	17.017	16.956	16.705	16.391	13.381
	ff03w	17.819	17.762	17.501	17.158	13.721
Trp-cage	**ff99SB-ILDN**	17.635	17.583	17.360	17.065	14.218
	ff99SB-PSN	18.861	18.809	18.595	18.302	15.323
	ff03w	16.094	16.043	15.836	15.553	12.788
Nsp1	**ff99SB-ILDN**	18.208	18.145	17.888	17.559	14.403
	ff99SB-PSN	22.348	22.267	21.931	21.486	17.374
	ff03w	21.624	21.553	21.248	20.839	16.796
Nup116	**ff99SB-ILDN**	21.127	21.069	20.823	20.493	17.094
	ff99SB-PSN	22.477	22.416	22.158	21.822	18.244
	ff03w	21.546	21.485	21.224	20.887	17.302
Production		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	8.561	8.542	8.462	8.358	7.222
	ff99SB-PSN	8.639	8.619	8.532	8.420	7.165
	ff03w	11.105	11.089	11.015	10.918	9.516
Trp-cage	**ff99SB-ILDN**	9.634	9.616	9.528	9.406	7.994
	ff99SB-PSN	10.431	10.407	10.297	10.153	8.513
	ff03w	12.934	12.912	12.814	12.679	10.954
Nsp1	**ff99SB-ILDN**	10.286	10.257	10.141	9.988	8.320
	ff99SB-PSN	11.260	11.228	11.106	10.937	9.079
	ff03w	8.382	8.349	8.211	8.032	6.367
Nup116	**ff99SB-ILDN**	9.599	9.570	9.441	9.278	7.601
	ff99SB-PSN	8.537	8.505	8.378	8.211	6.651
	ff03w	9.619	9.586	9.441	9.260	7.517

Table 6.20: Dimensionality estimation results across all levels of DFT smoothing using a fractional *amplitude* cutoff for the (top) 100ns annealing (300K-600K-300K) and (bottom) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. Estimates were obtained by taking the harmonic mean across small values of $k = [2, 3, 4, 6]$.

		Amplitude Smoothing Fraction				
Annealing		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	15.930	3.164	1.779	1.868	1.900
	ff99SB-PSN	16.648	4.344	1.773	1.829	1.901
	ff03w	17.440	2.941	1.820	1.896	1.900
Trp-cage	**ff99SB-ILDN**	17.315	9.347	1.779	1.835	1.894
	ff99SB-PSN	18.097	8.684	1.776	1.834	1.896
	ff03w	17.950	5.338	1.818	1.870	1.900
Nsp1	**ff99SB-ILDN**	19.134	4.660	1.793	1.882	1.899
	ff99SB-PSN	19.189	5.681	1.795	1.842	1.897
	ff03w	21.588	3.759	1.814	1.879	1.897
Nup116	**ff99SB-ILDN**	19.644	5.510	1.809	1.867	1.899
	ff99SB-PSN	19.934	4.729	1.813	1.875	1.897
	ff03w	21.252	9.692	1.803	1.864	1.893
Production		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	15.492	4.918	1.786	1.866	1.901
	ff99SB-PSN	15.082	6.965	1.807	1.835	1.899
	ff03w	16.668	4.035	1.866	1.898	1.901
Trp-cage	**ff99SB-ILDN**	18.997	16.795	1.774	1.841	1.893
	ff99SB-PSN	19.319	15.318	1.773	1.837	1.901
	ff03w	21.642	7.639	1.850	1.884	1.902
Nsp1	**ff99SB-ILDN**	19.594	9.054	1.791	1.878	1.902
	ff99SB-PSN	19.953	10.965	1.808	1.844	1.899
	ff03w	17.375	5.557	1.809	1.877	1.899
Nup116	**ff99SB-ILDN**	17.129	8.121	1.811	1.864	1.899
	ff99SB-PSN	14.649	5.604	1.820	1.872	1.898
	ff03w	17.270	13.816	1.811	1.854	1.896

make it impossible to calculate a dimensionality of zero from inherently noisy systems. Instead, the high-dimensionality of the noise allowed the folded or more frustrated states

Table 6.21: Dimensionality estimation results across all levels of DFT smoothing using a fractional *amplitude* cutoff for the (top) 100ns annealing (300K-600K-300K) and (bottom) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. Estimates were obtained by taking the harmonic mean across medium values of $k = [2, 3, 4, 6, 8, 16, 32, 64]$.

		Amplitude Smoothing Fraction				
Annealing		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	15.947	5.268	1.242	1.183	1.162
	ff99SB-PSN	16.779	6.813	1.326	1.216	1.163
	ff03w	17.086	4.723	1.204	1.162	1.160
Trp-cage	**ff99SB-ILDN**	17.306	11.379	1.301	1.192	1.163
	ff99SB-PSN	17.873	10.989	1.288	1.188	1.162
	ff03w	16.298	7.134	1.222	1.174	1.162
Nsp1	**ff99SB-ILDN**	19.351	8.006	1.212	1.166	1.162
	ff99SB-PSN	19.452	9.000	1.260	1.182	1.162
	ff03w	21.306	6.439	1.193	1.167	1.162
Nup116	**ff99SB-ILDN**	19.594	8.983	1.194	1.173	1.163
	ff99SB-PSN	20.272	7.661	1.218	1.166	1.162
	ff03w	21.175	13.156	1.264	1.171	1.162
Production		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	12.543	7.052	1.303	1.206	1.164
	ff99SB-PSN	12.409	8.400	1.381	1.232	1.165
	ff03w	14.388	6.430	1.176	1.160	1.163
Trp-cage	**ff99SB-ILDN**	15.450	14.152	1.373	1.208	1.162
	ff99SB-PSN	15.882	13.526	1.379	1.199	1.163
	ff03w	18.528	10.616	1.207	1.170	1.165
Nsp1	**ff99SB-ILDN**	15.809	10.552	1.242	1.175	1.165
	ff99SB-PSN	16.119	11.855	1.355	1.210	1.161
	ff03w	13.216	6.552	1.221	1.167	1.163
Nup116	**ff99SB-ILDN**	13.474	8.863	1.240	1.186	1.162
	ff99SB-PSN	11.285	6.389	1.241	1.171	1.163
	ff03w	13.638	11.555	1.386	1.190	1.161

Table 6.22: Dimensionality estimation results across all levels of DFT smoothing using a fractional *amplitude* cutoff for the (top) 100ns annealing (300K-600K-300K) and (bottom) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116 using three different forcefields: ff99SB-ILDN, ff99SB-PSN, and ff03w. Estimates were obtained by taking the harmonic mean across large values of $k = [2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]$.

		Amplitude Smoothing Fraction				
Annealing		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	17.164	9.120	2.160	1.238	1.039
	ff99SB-PSN	17.017	10.253	2.584	1.540	1.039
	ff03w	17.819	8.044	1.757	1.074	1.047
Trp-cage	**ff99SB-ILDN**	17.635	13.336	3.178	1.643	1.043
	ff99SB-PSN	18.861	13.708	3.097	1.559	1.052
	ff03w	16.094	9.197	1.833	1.234	1.043
Nsp1	**ff99SB-ILDN**	18.208	11.374	1.909	1.157	1.046
	ff99SB-PSN	22.348	14.335	2.647	1.495	1.041
	ff03w	21.624	11.074	1.978	1.162	1.053
Nup116	**ff99SB-ILDN**	21.127	14.319	2.077	1.276	1.033
	ff99SB-PSN	22.477	14.250	2.219	1.338	1.040
	ff03w	21.546	16.239	2.660	1.464	1.033
Production		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
GB1	**ff99SB-ILDN**	8.561	5.751	1.956	1.302	1.030
	ff99SB-PSN	8.639	6.216	2.204	1.469	1.039
	ff03w	11.105	7.350	1.533	1.053	1.030
Trp-cage	**ff99SB-ILDN**	9.634	8.295	2.775	1.658	1.074
	ff99SB-PSN	10.431	8.384	2.988	1.673	1.033
	ff03w	12.934	8.926	1.829	1.264	1.027
Nsp1	**ff99SB-ILDN**	10.286	7.109	1.970	1.164	1.026
	ff99SB-PSN	11.260	8.507	2.317	1.566	1.066
	ff03w	8.382	4.720	1.722	1.210	1.040
Nup116	**ff99SB-ILDN**	9.599	6.827	1.991	1.305	1.053
	ff99SB-PSN	8.537	5.788	1.817	1.281	1.042
	ff03w	9.619	7.764	2.252	1.487	1.063

of proteins folding to be predicted to have very high dimensionality relative to less frustrated, and dynamic intrinsically disordered proteins. In particular, even though one of the natively folded proteins used in the study didn't fold, it often assumed structural motifs that were also very rigid and of high estimated dimensionality. Future work is needed to ascertain if this result is consistent across other folding proteins, or the particular properties driving this effect in the simulations.

While dimensionality estimation has the potential to serve as a measure for the classification of protein dynamics, it has not been rigorously tested on MD simulations. The framework developed here provides a novel approach for addressing this concern that is computationally efficient and highly predictive of algorithm efficacy. While the polymer models studied here focused on noise and basic correlated motions, future studies could address folding more specifically by using models where the links have to cooperate to form helices and sheets or even simplified simulations methodologies such as coarse-grained simulations [88] or elastic network models [89]. The framework could also be used to compare different dimensionality estimation algorithms, uncovering their relative strengths and weaknesses. Also, manifold smoothing techniques besides the DFT approach investigated here may prove more useful[64], and their performance could also be compared using the above framework. An additional future aim is to apply these techniques to simulations of previously unstudied systems where predictions can be be made concerning the underlying dynamics using dimensionality estimation that can be verified via experiment.

Chapter 7

Conclusion

The application of machine learning and data mining techniques to molecular dynamics (MD) simulations 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 none have focused on a particular class of proteins known as "unstructured" or "intrinsically disordered" proteins.

Experimental techniques for studying intrinsically disordered proteins have proven useful for determining their general properties (e.g. radius of hydration, aggregation propensity, etc.); however, they have not been able to ascertain structural information with the same precision as is available for globular proteins. Certain disordered proteins adopt rigid structure upon binding, and this has aided progress in determining how the disordered nature of these proteins is advantageous from a functional or evolutionary point of view. However, many disordered proteins function *without* undergoing such transitions. Instead, the biologically relevant structural changes in these proteins occur at a level of detail that is currently beyond the capabilities of modern experimental techniques. Fortunately, simulation provides a reasonable means for assessing disordered protein structure and dynamics in atomic detail and promises to be a key component in

future analyses of disordered protein dynamics.

The unique physical properties of this class of proteins motivate a novel application of polymer-based models combined with statistical, clustering, dimensionality reduction and dimensionality estimation methods for the study of biomolecular simulations. In particular, applying and extending these various techniques has helped to elucidate certain properties of these proteins that are not accessible using standard low-dimensional metrics. Ultimately, the development of robust computational 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 this work is a polymer-based metric of "degree-of-unstructuredness" that could be used to categorize intrinsically disordered proteins in the same way that fold types are currently used to categorize folded proteins (e.g. [90]).

Appendix A

Polymer Dimensionality Estimates

This appendix contains additional data tables for the polymer studies in from Chapter 6. All of the tables presented here utilize the same nomenclature for identifying the k nearest-neighbor values used calculate dimensionality estimates for their respective system: small values of $k = [2, 3, 4, 6]$, medium values of $k = [8, 16, 32, 64]$, and large values of $k = [128, 256, 512, 1024]$. Please see Section 6.2.1 for details on the maximum likelihood estimator of dimensionality used to compute these values, and Section 6.2.2 for details on how these models were constructed.

A.1 Semirigid Helix

This section presents a detailed breakdown of dimensionality estimates at different scales of nearest neighbors for the semirigid helix model described in Section 6.2.2.

A.2 Half-folded Helix

This section presents a detailed breakdown of dimensionality estimates at different scales of nearest neighbors for the half-folded helix model described in Section 6.2.2.

Table A.1: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 2000$ structures of length 16 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 27 total degrees of freedom.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	20.489	20.546	20.773	20.622	10.632
	0.10	20.504	20.559	20.792	20.631	10.606
	1.00	23.905	23.957	24.077	24.293	9.914
	3.00	26.115	26.085	26.046	25.938	9.933
	10.00	26.591	26.591	26.239	25.938	9.874
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	18.517	18.486	18.483	18.380	14.463
	0.10	18.556	18.525	18.505	18.407	14.462
	1.00	22.510	22.553	22.457	22.202	15.688
	3.00	24.594	24.552	24.335	24.295	16.915
	10.00	24.848	24.878	24.762	24.635	16.789
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	13.086	13.092	13.082	13.076	12.656
	0.10	13.129	13.134	13.124	13.117	12.680
	1.00	16.777	16.739	16.655	16.503	14.684
	3.00	18.666	18.645	18.574	18.507	16.876
	10.00	18.699	18.690	18.639	18.568	16.984

Table A.2: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 5000$ structures of length 16 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 27 total degrees of freedom.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	21.546	21.605	21.600	21.461	12.481
	0.10	21.589	21.662	21.636	21.512	12.448
	1.00	24.660	24.717	24.724	24.398	11.391
	3.00	27.212	27.325	27.117	26.962	11.476
	10.00	27.105	27.174	27.052	26.752	11.356
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	19.717	19.723	19.685	19.574	15.968
	0.10	19.756	19.764	19.711	19.606	15.964
	1.00	23.289	23.313	23.247	23.044	17.191
	3.00	25.856	25.843	25.748	25.641	18.373
	10.00	25.650	25.652	25.574	25.462	18.346
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	15.392	15.408	15.407	15.399	14.983
	0.10	15.436	15.448	15.450	15.436	15.003
	1.00	19.259	19.233	19.130	19.009	17.113
	3.00	21.470	21.454	21.385	21.300	19.462
	10.00	21.504	21.491	21.425	21.358	19.559

Table A.3: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 2000$ structures of lengths 16 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 27 total degrees of freedom.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	20.489	20.488	20.432	20.525	20.126
	0.10	20.504	20.502	20.450	20.507	20.666
	1.00	23.905	23.904	23.976	24.050	19.066
	3.00	26.115	26.116	26.063	25.952	18.258
	10.00	26.591	26.592	26.557	26.605	18.465
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	18.517	18.517	18.535	18.596	18.357
	0.10	18.556	18.555	18.570	18.591	18.437
	1.00	22.510	22.511	22.505	22.542	17.015
	3.00	24.594	24.596	24.599	24.579	18.491
	10.00	24.848	24.847	24.839	24.808	17.820
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	13.086	13.086	13.082	13.092	12.923
	0.10	13.129	13.129	13.125	13.132	12.919
	1.00	16.777	16.778	16.776	16.753	11.870
	3.00	18.666	18.666	18.667	18.660	14.892
	10.00	18.699	18.700	18.696	18.696	14.342

Table A.4: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 5000$ structures of length 16 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 27 total degrees of freedom.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	21.546	21.546	21.548	21.654	21.065
	0.10	21.589	21.589	21.611	21.665	21.334
	1.00	24.660	24.661	24.652	24.766	19.999
	3.00	27.212	27.213	27.208	27.089	17.027
	10.00	27.105	27.105	27.076	27.089	17.454
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	19.717	19.718	19.728	19.732	19.463
	0.10	19.756	19.756	19.756	19.775	19.499
	1.00	23.289	23.289	23.288	23.303	18.043
	3.00	25.856	25.855	25.840	25.797	17.021
	10.00	25.650	25.649	25.642	25.672	17.242
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	15.392	15.391	15.393	15.386	15.313
	0.10	15.436	15.436	15.437	15.427	15.318
	1.00	19.259	19.260	19.260	19.237	13.880
	3.00	21.470	21.469	21.468	21.459	15.345
	10.00	21.504	21.504	21.502	21.496	15.469

Table A.5: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	24.388	24.406	24.798	24.539	9.196
	0.10	24.423	24.423	24.825	24.576	9.174
	1.00	29.612	29.660	29.220	28.903	9.130
	3.00	32.015	32.070	32.004	31.764	9.194
	10.00	33.027	32.972	33.268	32.905	9.065
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	22.013	22.006	21.916	21.935	15.131
	0.10	22.058	22.055	21.991	21.966	15.130
	1.00	27.001	26.964	26.778	26.499	17.069
	3.00	29.734	29.693	29.791	29.633	18.409
	10.00	30.007	30.032	29.957	29.938	18.407
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	15.260	15.260	15.247	15.286	14.723
	0.10	15.310	15.308	15.296	15.335	14.756
	1.00	19.607	19.589	19.512	19.334	17.087
	3.00	21.889	21.859	21.787	21.696	19.782
	10.00	22.014	22.009	21.932	21.853	19.808

Table A.6: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	25.773	25.857	25.931	25.727	10.720
	0.10	25.775	25.881	25.976	25.762	10.689
	1.00	30.325	30.316	30.174	30.057	10.402
	3.00	33.025	33.078	32.910	32.515	10.193
	10.00	33.965	33.970	33.964	33.611	10.318
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	23.415	23.376	23.310	23.398	16.857
	0.10	23.461	23.430	23.372	23.438	16.861
	1.00	28.362	28.313	28.173	27.978	18.701
	3.00	31.185	31.141	31.017	30.795	19.939
	10.00	31.592	31.617	31.506	31.337	20.078
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	18.017	18.027	18.011	18.023	17.394
	0.10	18.074	18.083	18.068	18.073	17.421
	1.00	22.790	22.764	22.664	22.500	19.990
	3.00	25.387	25.373	25.302	25.199	22.920
	10.00	25.565	25.546	25.479	25.390	22.995

Table A.7: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	24.388	24.387	24.398	24.495	24.604
	0.10	24.423	24.425	24.437	24.473	24.644
	1.00	29.612	29.613	29.644	29.675	24.060
	3.00	32.015	32.014	32.014	32.140	23.474
	10.00	33.027	33.027	33.004	32.834	23.055
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	22.013	22.014	22.009	21.958	21.632
	0.10	22.058	22.057	22.063	22.039	21.837
	1.00	27.001	27.000	27.015	27.015	21.301
	3.00	29.734	29.733	29.731	29.745	21.714
	10.00	30.007	30.007	30.004	29.942	21.631
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	15.260	15.260	15.262	15.282	15.108
	0.10	15.310	15.310	15.310	15.319	15.165
	1.00	19.607	19.607	19.611	19.597	14.594
	3.00	21.889	21.889	21.885	21.881	16.893
	10.00	22.014	22.014	22.012	22.022	16.767

Table A.8: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	25.773	25.773	25.790	25.847	26.152
	0.10	25.775	25.774	25.783	25.862	26.107
	1.00	30.325	30.324	30.349	30.345	24.707
	3.00	33.025	33.024	33.032	33.078	23.672
	10.00	33.965	33.967	33.955	33.984	22.087
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	23.415	23.415	23.412	23.435	23.315
	0.10	23.461	23.460	23.470	23.461	23.285
	1.00	28.362	28.362	28.363	28.332	22.467
	3.00	31.185	31.185	31.173	31.174	22.673
	10.00	31.592	31.593	31.606	31.585	21.600
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	18.017	18.018	18.023	18.027	17.988
	0.10	18.074	18.074	18.074	18.079	17.973
	1.00	22.790	22.790	22.791	22.777	17.442
	3.00	25.387	25.387	25.389	25.385	19.281
	10.00	25.565	25.565	25.561	25.547	18.677

Table A.9: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	29.522	29.656	29.669	29.123	8.205
	0.10	29.571	29.684	29.662	29.178	8.183
	1.00	36.054	36.099	35.306	35.749	8.419
	3.00	39.027	39.118	38.782	37.901	8.318
	10.00	39.501	39.402	39.244	39.632	8.410
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	25.535	25.613	25.611	25.544	15.777
	0.10	25.616	25.669	25.673	25.601	15.779
	1.00	32.095	32.081	31.946	31.955	18.168
	3.00	35.230	35.202	35.098	35.007	19.547
	10.00	35.474	35.486	35.381	35.385	19.689
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	17.368	17.403	17.436	17.465	16.739
	0.10	17.436	17.464	17.494	17.525	16.775
	1.00	22.752	22.722	22.572	22.393	19.668
	3.00	25.509	25.493	25.444	25.330	22.741
	10.00	25.655	25.630	25.562	25.452	22.917

Table A.10: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	30.451	30.553	30.705	30.467	9.332
	0.10	30.543	30.584	30.739	30.517	9.305
	1.00	36.823	36.784	36.605	36.584	9.492
	3.00	40.487	40.337	40.047	39.668	9.101
	10.00	41.286	41.073	41.103	41.349	9.223
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	27.380	27.419	27.521	27.543	17.524
	0.10	27.439	27.481	27.562	27.598	17.521
	1.00	33.753	33.718	33.556	33.443	19.939
	3.00	37.384	37.383	37.260	37.023	21.318
	10.00	37.477	37.481	37.359	37.399	21.300
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	20.790	20.801	20.834	20.866	19.899
	0.10	20.858	20.866	20.901	20.928	19.929
	1.00	26.693	26.670	26.499	26.300	23.154
	3.00	29.838	29.820	29.739	29.627	26.673
	10.00	29.895	29.886	29.826	29.702	26.755

Table A.11: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	29.522	29.521	29.543	29.527	29.230
	0.10	29.571	29.572	29.609	29.564	29.714
	1.00	36.054	36.055	36.073	35.911	27.570
	3.00	39.027	39.029	39.122	38.938	27.502
	10.00	39.501	39.502	39.461	39.601	28.727
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	25.535	25.535	25.525	25.423	25.552
	0.10	25.616	25.615	25.612	25.500	25.527
	1.00	32.095	32.096	32.090	32.075	24.198
	3.00	35.230	35.228	35.226	35.263	25.468
	10.00	35.474	35.474	35.492	35.458	26.544
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	17.368	17.369	17.371	17.360	17.313
	0.10	17.436	17.436	17.434	17.423	17.359
	1.00	22.752	22.752	22.751	22.760	16.422
	3.00	25.509	25.509	25.511	25.491	19.362
	10.00	25.655	25.654	25.641	25.639	19.812

Table A.12: Dimensionality estimates at different sets of k for the *semirigid helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	30.451	30.450	30.428	30.236	30.211
	0.10	30.543	30.541	30.495	30.336	30.448
	1.00	36.823	36.823	36.750	36.774	30.075
	3.00	40.487	40.487	40.489	40.371	28.320
	10.00	41.286	41.284	41.367	41.306	27.920
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	27.380	27.379	27.365	27.325	27.239
	0.10	27.439	27.438	27.442	27.408	27.321
	1.00	33.753	33.752	33.731	33.767	27.155
	3.00	37.384	37.384	37.362	37.347	26.683
	10.00	37.477	37.479	37.465	37.441	26.481
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	0.000	0.000	0.000	0.000	0.000
	0.01	20.790	20.791	20.791	20.776	20.743
	0.10	20.858	20.858	20.865	20.856	20.781
	1.00	26.693	26.693	26.687	26.685	20.720
	3.00	29.838	29.838	29.840	29.836	22.285
	10.00	29.895	29.895	29.891	29.876	22.112

Table A.13: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 2000$ structures of length 16 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 27 total degrees of freedom, but only 16 are attributed to the unfolded region.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	20.489	20.546	20.773	20.622	10.632
	0.01	20.455	20.274	20.036	20.462	10.723
	0.10	14.554	14.669	14.991	14.777	11.551
	1.00	15.354	15.252	15.198	15.462	9.858
	3.00	16.573	16.458	16.478	16.517	10.776
	10.00	16.573	16.458	16.478	16.517	10.776
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	18.517	18.486	18.483	18.380	14.463
	0.01	18.086	18.028	17.925	18.015	14.266
	0.10	13.289	13.278	13.222	13.294	11.873
	1.00	14.710	14.725	14.607	14.575	12.056
	3.00	16.345	16.374	16.223	16.215	13.267
	10.00	16.345	16.374	16.223	16.215	13.267
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	13.086	13.092	13.082	13.076	12.656
	0.01	12.631	12.624	12.629	12.640	12.223
	0.10	9.670	9.669	9.675	9.655	9.471
	1.00	11.864	11.848	11.784	11.690	10.601
	3.00	13.327	13.315	13.280	13.243	12.251
	10.00	13.327	13.315	13.280	13.243	12.251

A.3 Correlated Helix

This section presents a detailed breakdown of dimensionality estimates at different scales of nearest neighbors for the correlated helix model described in Section 6.2.2.

Table A.14: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 5000$ structures of length 16 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 27 total degrees of freedom, but only 16 are attributed to the unfolded region.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	21.546	21.605	21.600	21.461	12.481
	0.01	21.403	21.511	21.232	21.213	12.602
	0.10	15.083	15.126	15.078	15.007	12.625
	1.00	15.741	15.709	15.734	15.552	11.167
	3.00	16.586	16.569	16.643	16.474	11.770
	10.00	16.586	16.569	16.643	16.474	11.770
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	19.717	19.723	19.685	19.574	15.968
	0.01	19.274	19.263	19.125	19.119	15.796
	0.10	13.913	13.894	13.912	13.911	12.902
	1.00	15.160	15.111	15.052	14.961	12.996
	3.00	16.674	16.679	16.581	16.501	13.943
	10.00	16.674	16.679	16.581	16.501	13.943
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	15.392	15.408	15.407	15.399	14.983
	0.01	14.927	14.927	14.903	14.895	14.520
	0.10	11.226	11.222	11.212	11.205	11.051
	1.00	13.336	13.325	13.258	13.172	12.118
	3.00	14.912	14.893	14.847	14.793	13.684
	10.00	14.912	14.893	14.847	14.793	13.684

Table A.15: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 2000$ structures of length 16 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 27 total degrees of freedom, but only 16 are attributed to the unfolded region.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	20.489	20.488	20.432	20.525	20.126
	0.01	20.455	20.453	20.485	20.521	15.820
	0.10	14.554	14.550	14.240	14.247	14.005
	1.00	15.354	15.348	15.358	15.309	13.962
	3.00	16.573	16.570	16.601	16.635	10.125
	10.00	16.573	16.570	16.601	16.635	10.125
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	18.517	18.517	18.535	18.596	18.357
	0.01	18.086	18.086	18.097	18.038	14.119
	0.10	13.289	13.287	13.086	13.022	12.919
	1.00	14.710	14.706	14.709	14.708	12.842
	3.00	16.345	16.342	16.350	16.369	10.611
	10.00	16.345	16.342	16.350	16.369	10.611
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	13.086	13.086	13.082	13.092	12.923
	0.01	12.631	12.631	12.629	12.602	10.134
	0.10	9.670	9.670	9.584	9.554	9.482
	1.00	11.864	11.862	11.863	11.861	9.506
	3.00	13.327	13.326	13.326	13.332	9.706
	10.00	13.327	13.326	13.326	13.332	9.706

Table A.16: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 5000$ structures of length 16 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 27 total degrees of freedom, but only 16 are attributed to the unfolded region.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	21.546	21.546	21.548	21.654	21.065
	0.01	21.403	21.403	21.412	21.224	16.723
	0.10	15.083	15.078	14.747	14.687	14.509
	1.00	15.741	15.734	15.746	15.819	14.335
	3.00	16.586	16.581	16.556	16.455	9.578
	10.00	16.586	16.581	16.556	16.455	9.578
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	19.717	19.718	19.728	19.732	19.463
	0.01	19.274	19.274	19.259	19.279	15.143
	0.10	13.913	13.911	13.689	13.596	13.554
	1.00	15.160	15.155	15.160	15.160	13.329
	3.00	16.674	16.671	16.674	16.643	10.152
	10.00	16.674	16.671	16.674	16.643	10.152
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	15.392	15.391	15.393	15.386	15.313
	0.01	14.927	14.927	14.923	14.913	11.947
	0.10	11.226	11.225	11.107	11.061	11.018
	1.00	13.336	13.333	13.335	13.336	10.988
	3.00	14.912	14.910	14.908	14.888	10.298
	10.00	14.912	14.910	14.908	14.888	10.298

Table A.17: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom, but only 20 are attributed to the unfolded region.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	24.388	24.406	24.798	24.539	9.196
	0.01	23.814	23.930	23.518	23.889	9.654
	0.10	17.538	17.208	17.364	17.030	11.614
	1.00	19.079	19.021	18.902	18.576	10.821
	3.00	20.601	20.530	20.349	20.284	10.522
	10.00	20.314	20.300	20.092	20.436	10.174
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	22.013	22.006	21.916	21.935	15.131
	0.01	21.087	21.051	21.075	21.151	15.095
	0.10	15.435	15.454	15.496	15.400	13.245
	1.00	17.772	17.802	17.744	17.531	14.055
	3.00	19.568	19.566	19.497	19.418	14.900
	10.00	19.564	19.591	19.557	19.500	14.839
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	15.260	15.260	15.247	15.286	14.723
	0.01	14.611	14.599	14.611	14.653	14.112
	0.10	11.031	11.037	11.027	11.043	10.891
	1.00	13.740	13.724	13.644	13.569	12.305
	3.00	15.498	15.481	15.428	15.382	14.152
	10.00	15.532	15.521	15.477	15.418	14.191

Table A.18: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom, but only 20 are attributed to the unfolded region.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	25.773	25.857	25.931	25.727	10.720
	0.01	24.897	25.017	24.687	24.811	11.199
	0.10	18.023	18.009	17.885	17.976	13.438
	1.00	19.101	19.144	18.963	19.243	12.407
	3.00	20.679	20.698	20.756	20.684	12.129
	10.00	20.705	20.803	20.600	20.439	11.873
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	23.415	23.376	23.310	23.398	16.857
	0.01	22.717	22.726	22.706	22.681	16.784
	0.10	16.312	16.320	16.324	16.321	14.541
	1.00	18.365	18.369	18.248	18.129	15.085
	3.00	20.170	20.183	20.156	20.132	16.000
	10.00	20.273	20.216	20.156	20.099	16.051
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	18.017	18.027	18.011	18.023	17.394
	0.01	17.395	17.400	17.404	17.393	16.778
	0.10	12.942	12.945	12.932	12.917	12.720
	1.00	15.645	15.632	15.563	15.456	14.095
	3.00	17.536	17.523	17.471	17.406	16.015
	10.00	17.585	17.572	17.513	17.445	16.118

Table A.19: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom, but only 20 are attributed to the unfolded region.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	24.388	24.387	24.398	24.495	24.604
	0.01	23.814	23.813	23.758	24.108	19.321
	0.10	17.538	17.538	17.238	17.158	17.365
	1.00	19.079	19.073	19.047	19.033	16.567
	3.00	20.601	20.597	20.532	20.610	14.421
	10.00	20.314	20.310	20.260	20.317	14.188
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	22.013	22.014	22.009	21.958	21.632
	0.01	21.087	21.088	21.075	21.118	17.080
	0.10	15.435	15.433	15.229	15.160	15.111
	1.00	17.772	17.767	17.761	17.756	15.046
	3.00	19.568	19.564	19.555	19.561	14.316
	10.00	19.564	19.561	19.548	19.586	13.887
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	15.260	15.260	15.262	15.282	15.108
	0.01	14.611	14.612	14.612	14.611	11.846
	0.10	11.031	11.030	10.929	10.897	10.865
	1.00	13.740	13.738	13.735	13.725	10.986
	3.00	15.498	15.496	15.496	15.492	12.066
	10.00	15.532	15.530	15.527	15.525	11.781

Table A.20: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom, but only 20 are attributed to the unfolded region.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	25.773	25.773	25.790	25.847	26.152
	0.01	24.897	24.897	24.907	25.114	18.559
	0.10	18.023	18.023	17.676	17.643	17.151
	1.00	19.101	19.093	19.083	19.149	17.110
	3.00	20.679	20.674	20.681	20.663	13.495
	10.00	20.705	20.699	20.687	20.831	12.959
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	23.415	23.415	23.412	23.435	23.315
	0.01	22.717	22.717	22.709	22.722	17.101
	0.10	16.312	16.310	16.039	15.982	15.904
	1.00	18.365	18.359	18.364	18.381	15.772
	3.00	20.170	20.166	20.174	20.180	13.703
	10.00	20.273	20.269	20.274	20.300	13.150
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	18.017	18.018	18.023	18.027	17.988
	0.01	17.395	17.395	17.395	17.368	13.432
	0.10	12.942	12.941	12.791	12.760	12.709
	1.00	15.645	15.642	15.641	15.635	12.769
	3.00	17.536	17.533	17.532	17.522	12.850
	10.00	17.585	17.582	17.579	17.569	12.482

Table A.21: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom, but only 24 are attributed to the unfolded region.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	29.522	29.656	29.669	29.123	8.205
	0.01	28.179	27.939	27.853	27.947	8.361
	0.10	19.178	19.177	19.113	18.810	10.970
	1.00	22.140	21.891	22.052	21.741	10.333
	3.00	24.355	24.099	23.722	23.333	10.313
	10.00	24.596	24.501	24.366	24.298	10.049
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	25.535	25.613	25.611	25.544	15.777
	0.01	24.689	24.613	24.687	24.664	15.732
	0.10	17.475	17.434	17.401	17.312	14.000
	1.00	20.399	20.404	20.323	19.999	15.021
	3.00	22.921	22.873	22.817	22.649	16.264
	10.00	22.972	22.935	22.807	22.621	16.200
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	17.368	17.403	17.436	17.465	16.739
	0.01	16.712	16.691	16.714	16.767	16.149
	0.10	12.317	12.327	12.324	12.314	11.952
	1.00	15.473	15.460	15.382	15.258	13.611
	3.00	17.458	17.445	17.395	17.333	15.956
	10.00	17.511	17.499	17.434	17.368	15.998

Table A.22: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom, but only 24 are attributed to the unfolded region.

		Frequency Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	30.451	30.553	30.705	30.467	9.332
	0.01	29.434	29.240	29.257	29.291	9.588
	0.10	20.326	20.113	20.285	20.397	12.770
	1.00	22.257	22.057	22.139	22.038	11.875
	3.00	24.595	24.449	24.253	24.157	12.140
	10.00	24.865	24.887	24.866	24.447	11.819
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	27.380	27.419	27.521	27.543	17.524
	0.01	26.508	26.478	26.480	26.402	17.417
	0.10	18.502	18.443	18.495	18.512	15.554
	1.00	21.245	21.196	21.140	20.954	16.328
	3.00	23.566	23.553	23.480	23.347	17.705
	10.00	23.711	23.661	23.633	23.495	17.528
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	20.790	20.801	20.834	20.866	19.899
	0.01	20.050	20.046	20.092	20.115	19.221
	0.10	14.554	14.565	14.555	14.553	14.147
	1.00	17.800	17.782	17.685	17.557	15.825
	3.00	19.899	19.880	19.818	19.734	18.180
	10.00	20.001	19.982	19.926	19.844	18.237

Table A.23: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom, but only 24 are attributed to the unfolded region.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	29.522	29.521	29.543	29.527	29.230
	0.01	28.179	28.183	28.209	28.195	21.806
	0.10	19.178	19.176	18.881	18.814	18.751
	1.00	22.140	22.130	22.147	22.221	18.665
	3.00	24.355	24.351	24.370	24.179	17.111
	10.00	24.596	24.590	24.583	24.482	16.336
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	25.535	25.535	25.525	25.423	25.552
	0.01	24.689	24.690	24.738	24.660	19.057
	0.10	17.475	17.471	17.248	17.187	17.084
	1.00	20.399	20.393	20.388	20.410	16.921
	3.00	22.921	22.916	22.926	22.874	16.587
	10.00	22.972	22.968	22.957	22.942	16.269
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	17.368	17.369	17.371	17.360	17.313
	0.01	16.712	16.711	16.707	16.705	13.088
	0.10	12.317	12.316	12.199	12.143	12.161
	1.00	15.473	15.470	15.471	15.461	12.093
	3.00	17.458	17.455	17.457	17.448	13.520
	10.00	17.511	17.509	17.506	17.500	13.411

Table A.24: Dimensionality estimates at different sets of k for the *half-folded helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom, but only 24 are attributed to the unfolded region.

		Amplitude Smoothing Fraction				
Small k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	30.451	30.450	30.428	30.236	30.211
	0.01	29.434	29.432	29.401	29.335	22.357
	0.10	20.326	20.325	19.944	19.734	19.940
	1.00	22.257	22.247	22.214	22.064	18.932
	3.00	24.595	24.589	24.607	24.562	16.195
	10.00	24.865	24.859	24.901	24.881	15.267
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	27.380	27.379	27.365	27.325	27.239
	0.01	26.508	26.508	26.520	26.525	20.089
	0.10	18.502	18.501	18.213	18.098	18.137
	1.00	21.245	21.239	21.251	21.217	17.452
	3.00	23.566	23.560	23.546	23.533	16.212
	10.00	23.711	23.705	23.695	23.692	15.506
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta_{unfolded}}$)	**0.00**	20.790	20.791	20.791	20.776	20.743
	0.01	20.050	20.050	20.051	20.013	15.468
	0.10	14.554	14.553	14.388	14.329	14.293
	1.00	17.800	17.796	17.793	17.778	13.755
	3.00	19.899	19.895	19.896	19.885	14.651
	10.00	20.001	19.997	19.996	19.988	14.250

Table A.25: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom, but 2 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.845	1.842	1.911	1.944	2.012
	0.01	3.449	3.450	3.432	3.394	2.912
	0.10	16.805	16.782	16.565	16.483	8.755
	1.00	29.526	29.484	29.191	28.562	7.227
	3.00	32.539	32.451	32.532	32.683	9.108
	10.00	33.327	33.711	32.979	32.715	9.121
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.914	1.902	1.918	1.927	1.948
	0.01	2.265	2.266	2.262	2.245	2.146
	0.10	10.315	10.269	10.129	9.970	6.946
	1.00	26.246	26.182	25.922	25.467	14.701
	3.00	30.025	30.012	29.856	29.739	18.372
	10.00	30.057	30.126	29.873	29.988	18.427
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.813	1.812	1.812	1.813	1.805
	0.01	1.844	1.843	1.842	1.841	1.822
	0.10	3.047	3.035	2.991	2.939	2.487
	1.00	18.581	18.537	18.359	18.066	14.323
	3.00	21.906	21.890	21.823	21.731	19.765
	10.00	22.026	22.010	21.935	21.898	19.835

Table A.26: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom, but 2 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.879	1.910	1.949	2.002	1.965
	0.01	5.246	5.226	5.107	5.013	3.847
	0.10	18.924	18.995	18.885	18.820	10.886
	1.00	30.517	30.412	30.154	29.732	8.205
	3.00	33.347	33.401	33.338	33.296	10.217
	10.00	34.096	33.955	33.698	33.534	10.360
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.895	1.880	1.910	1.931	1.961
	0.01	2.650	2.644	2.635	2.616	2.388
	0.10	13.559	13.533	13.384	13.215	9.861
	1.00	27.699	27.677	27.445	27.136	16.232
	3.00	31.107	31.089	30.997	30.917	20.017
	10.00	31.430	31.475	31.380	31.315	20.078
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.871	1.870	1.871	1.872	1.862
	0.01	1.942	1.941	1.938	1.935	1.900
	0.10	4.576	4.550	4.459	4.353	3.400
	1.00	21.875	21.833	21.648	21.357	17.481
	3.00	25.393	25.363	25.301	25.205	22.920
	10.00	25.535	25.527	25.460	25.378	23.029

Table A.27: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom, but 2 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.845	1.989	1.979	1.996	1.973
	0.01	3.449	2.848	3.050	2.414	1.989
	0.10	16.805	16.004	6.487	6.128	2.407
	1.00	29.526	29.517	29.225	28.184	4.878
	3.00	32.539	32.541	32.557	32.561	22.640
	10.00	33.327	33.328	33.316	33.042	23.918
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.914	1.943	1.977	1.998	1.906
	0.01	2.265	2.141	2.233	2.099	1.908
	0.10	10.315	9.368	3.762	3.685	2.003
	1.00	26.246	26.243	26.135	25.202	3.309
	3.00	30.025	30.025	30.000	29.944	21.527
	10.00	30.057	30.055	30.081	30.121	22.289
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.813	1.797	1.746	1.746	1.657
	0.01	1.844	1.814	1.767	1.754	1.657
	0.10	3.047	2.785	1.919	1.925	1.664
	1.00	18.581	18.581	18.502	17.548	2.845
	3.00	21.906	21.906	21.906	21.896	16.890
	10.00	22.026	22.026	22.022	22.028	17.142

Table A.28: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom, but 2 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.879	2.083	1.972	2.016	2.021
	0.01	5.246	3.708	3.696	3.983	2.040
	0.10	18.924	17.864	7.799	7.497	3.720
	1.00	30.517	30.510	30.200	28.936	4.796
	3.00	33.347	33.347	33.348	33.374	22.952
	10.00	34.096	34.095	34.067	34.053	22.515
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.895	1.996	2.000	2.015	2.000
	0.01	2.650	2.383	2.442	2.586	2.003
	0.10	13.559	12.238	5.131	4.975	3.163
	1.00	27.699	27.701	27.627	26.572	3.332
	3.00	31.107	31.106	31.092	31.110	22.403
	10.00	31.430	31.428	31.424	31.410	22.134
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.871	1.860	1.841	1.851	1.906
	0.01	1.942	1.895	1.882	1.909	1.906
	0.10	4.576	3.865	2.247	2.279	2.291
	1.00	21.875	21.875	21.788	20.552	4.204
	3.00	25.393	25.393	25.388	25.379	19.206
	10.00	25.535	25.535	25.536	25.525	19.016

Table A.29: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom, but 3 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.850	2.876	2.887	2.900	3.040
	0.01	3.252	3.252	3.244	3.243	3.225
	0.10	13.048	12.921	12.572	12.059	6.294
	1.00	28.718	28.770	28.368	27.985	7.009
	3.00	31.723	31.713	31.917	31.673	9.226
	10.00	33.306	32.922	33.860	32.421	9.054
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.744	2.745	2.746	2.738	2.726
	0.01	2.842	2.838	2.836	2.824	2.768
	0.10	7.381	7.302	7.115	6.890	4.653
	1.00	26.667	26.598	26.298	26.000	14.588
	3.00	29.407	29.499	29.538	29.499	18.401
	10.00	29.974	29.841	29.832	29.573	18.355
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.617	2.616	2.616	2.615	2.605
	0.01	2.630	2.630	2.628	2.627	2.612
	0.10	3.470	3.461	3.428	3.388	3.049
	1.00	19.077	19.031	18.864	18.613	14.881
	3.00	21.894	21.878	21.817	21.740	19.775
	10.00	21.991	21.965	21.907	21.794	19.829

Table A.30: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom, but 3 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.846	2.854	2.870	2.866	2.952
	0.01	3.429	3.420	3.419	3.393	3.245
	0.10	15.438	15.347	15.073	14.699	7.879
	1.00	30.196	30.157	29.834	29.504	7.877
	3.00	33.196	33.446	33.708	33.477	10.181
	10.00	33.765	33.730	33.914	33.569	10.218
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.839	2.839	2.839	2.835	2.846
	0.01	3.008	3.002	2.996	2.982	2.924
	0.10	9.789	9.742	9.518	9.241	6.203
	1.00	27.903	27.850	27.733	27.401	15.907
	3.00	31.300	31.284	31.193	31.013	19.960
	10.00	31.644	31.628	31.550	31.374	19.999
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.700	2.700	2.699	2.697	2.686
	0.01	2.724	2.723	2.721	2.718	2.697
	0.10	4.201	4.185	4.126	4.056	3.474
	1.00	22.255	22.207	22.028	21.751	17.688
	3.00	25.424	25.403	25.337	25.225	22.931
	10.00	25.529	25.514	25.444	25.331	23.020

Table A.31: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom, but 3 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.850	3.069	3.286	3.375	3.537
	0.01	3.252	3.229	3.412	3.768	3.538
	0.10	13.048	11.653	5.144	5.261	4.927
	1.00	28.718	28.728	28.733	28.579	5.996
	3.00	31.723	31.723	31.712	31.494	21.912
	10.00	33.306	33.304	33.330	33.350	23.890
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.744	2.729	2.713	2.753	2.769
	0.01	2.842	2.769	2.733	2.843	2.768
	0.10	7.381	6.293	3.128	3.242	3.109
	1.00	26.667	26.666	26.576	25.355	2.939
	3.00	29.407	29.407	29.427	29.397	21.177
	10.00	29.974	29.972	29.960	29.915	22.240
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.617	2.597	2.469	2.380	2.177
	0.01	2.630	2.603	2.471	2.386	2.177
	0.10	3.470	3.233	2.544	2.473	2.299
	1.00	19.077	19.076	18.989	17.783	2.850
	3.00	21.894	21.894	21.890	21.865	16.438
	10.00	21.991	21.991	21.987	21.990	17.204

Table A.32: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom, but 3 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.846	2.968	3.095	3.156	3.253
	0.01	3.429	3.234	3.440	3.966	3.255
	0.10	15.438	12.829	5.934	5.846	3.890
	1.00	30.196	30.186	30.224	27.900	4.607
	3.00	33.196	33.200	33.193	33.123	22.433
	10.00	33.765	33.768	33.783	33.839	22.001
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.839	2.859	2.932	2.931	2.814
	0.01	3.008	2.932	3.021	3.158	2.815
	0.10	9.789	7.505	3.832	3.885	3.045
	1.00	27.903	27.895	27.708	25.711	3.193
	3.00	31.300	31.299	31.276	31.258	21.999
	10.00	31.644	31.644	31.632	31.680	21.431
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.700	2.663	2.542	2.495	2.649
	0.01	2.724	2.673	2.552	2.526	2.649
	0.10	4.201	3.573	2.691	2.662	2.721
	1.00	22.255	22.254	22.093	19.951	3.604
	3.00	25.424	25.423	25.424	25.418	18.976
	10.00	25.529	25.529	25.530	25.524	18.490

Table A.33: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom, but 5 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.096	3.094	3.093	3.069	3.255
	0.01	3.190	3.189	3.182	3.153	3.274
	0.10	8.914	8.745	8.189	7.517	3.998
	1.00	29.799	29.633	29.397	28.772	6.673
	3.00	32.591	32.441	32.181	32.278	9.171
	10.00	32.648	32.933	32.698	31.917	8.967
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.357	3.354	3.341	3.326	3.202
	0.01	3.395	3.392	3.377	3.358	3.213
	0.10	6.035	5.997	5.827	5.634	4.051
	1.00	26.979	26.958	26.683	26.186	14.333
	3.00	29.911	29.841	29.734	29.516	18.304
	10.00	29.982	30.052	29.928	29.812	18.384
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.553	3.552	3.551	3.548	3.528
	0.01	3.561	3.561	3.559	3.555	3.531
	0.10	4.239	4.231	4.200	4.164	3.861
	1.00	19.329	19.291	19.120	18.863	15.249
	3.00	21.925	21.915	21.833	21.757	19.757
	10.00	22.048	22.029	21.943	21.856	19.859

Table A.34: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 35 total degrees of freedom, but 5 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.534	3.534	3.536	3.525	3.697
	0.01	3.661	3.660	3.650	3.626	3.721
	0.10	10.653	10.473	9.832	9.016	4.501
	1.00	30.536	30.392	29.746	29.266	7.473
	3.00	33.140	33.108	33.124	33.156	10.265
	10.00	33.638	33.763	33.521	33.288	10.256
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.773	3.769	3.757	3.743	3.605
	0.01	3.825	3.820	3.803	3.784	3.621
	0.10	7.391	7.332	7.120	6.853	4.748
	1.00	28.222	28.204	27.924	27.550	15.653
	3.00	31.255	31.219	31.170	30.969	19.982
	10.00	31.561	31.579	31.527	31.201	19.979
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	4.001	4.001	3.999	3.996	3.973
	0.01	4.014	4.013	4.011	4.008	3.979
	0.10	5.095	5.082	5.034	4.978	4.508
	1.00	22.593	22.556	22.378	22.106	18.140
	3.00	25.422	25.409	25.332	25.231	22.911
	10.00	25.561	25.543	25.466	25.363	22.989

Table A.35: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom, but 5 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.096	3.106	5.138	5.266	4.397
	0.01	3.190	3.146	5.166	5.319	4.397
	0.10	8.914	7.967	5.430	5.929	4.563
	1.00	29.799	29.800	29.674	28.968	4.981
	3.00	32.591	32.592	32.489	32.704	21.940
	10.00	32.648	32.648	32.630	32.454	23.321
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.357	3.214	3.055	3.017	2.839
	0.01	3.395	3.230	3.056	3.021	2.839
	0.10	6.035	5.578	3.116	3.081	2.840
	1.00	26.979	26.979	26.887	26.138	2.824
	3.00	29.911	29.910	29.908	29.899	20.749
	10.00	29.982	29.983	29.965	29.919	22.279
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.553	3.524	3.384	3.306	2.894
	0.01	3.561	3.528	3.386	3.313	2.894
	0.10	4.239	4.122	3.438	3.363	2.922
	1.00	19.329	19.329	19.276	18.666	3.372
	3.00	21.925	21.925	21.927	21.915	16.327
	10.00	22.048	22.048	22.046	22.048	17.313

Table A.36: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 20 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 35 total degrees of freedom, but 5 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.534	3.740	5.157	4.949	4.692
	0.01	3.661	3.780	5.206	5.024	4.759
	0.10	10.653	8.097	5.735	5.527	4.820
	1.00	30.536	30.536	30.392	29.017	5.508
	3.00	33.140	33.141	33.112	33.163	24.165
	10.00	33.638	33.637	33.645	33.727	22.717
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.773	3.558	3.389	3.376	3.186
	0.01	3.825	3.572	3.393	3.387	3.203
	0.10	7.391	6.064	3.497	3.491	3.209
	1.00	28.222	28.221	28.145	26.606	2.957
	3.00	31.255	31.256	31.233	31.240	22.735
	10.00	31.561	31.560	31.536	31.551	22.227
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	4.001	3.948	3.763	3.614	3.548
	0.01	4.014	3.953	3.766	3.620	3.574
	0.10	5.095	4.701	3.842	3.694	3.686
	1.00	22.593	22.592	22.497	21.046	4.483
	3.00	25.422	25.422	25.426	25.416	19.346
	10.00	25.561	25.561	25.561	25.541	19.017

Table A.37: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom, but 2 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.844	1.843	1.913	1.945	2.015
	0.01	3.453	3.465	3.468	3.448	2.945
	0.10	18.742	18.771	18.708	18.356	9.053
	1.00	34.645	34.937	34.453	34.713	6.859
	3.00	39.223	39.250	39.804	39.813	8.387
	10.00	39.992	40.315	40.021	39.930	8.416
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.913	1.900	1.918	1.926	1.948
	0.01	2.265	2.265	2.263	2.251	2.144
	0.10	11.302	11.235	11.023	10.785	7.156
	1.00	30.791	30.766	30.595	30.190	15.518
	3.00	35.501	35.491	35.298	35.083	19.582
	10.00	35.819	35.843	35.481	35.346	19.770
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.813	1.812	1.813	1.813	1.805
	0.01	1.844	1.843	1.842	1.841	1.821
	0.10	3.050	3.037	2.992	2.940	2.482
	1.00	21.146	21.074	20.824	20.491	15.834
	3.00	25.491	25.477	25.401	25.279	22.761
	10.00	25.661	25.647	25.592	25.495	22.898

Table A.38: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom, but 2 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.885	1.911	1.950	2.003	1.966
	0.01	5.318	5.242	5.137	5.074	3.839
	0.10	21.625	21.652	21.608	21.068	11.018
	1.00	36.506	36.685	36.029	36.283	7.608
	3.00	41.317	41.183	41.120	41.413	9.187
	10.00	41.158	41.229	40.974	40.722	9.237
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.895	1.879	1.909	1.932	1.961
	0.01	2.654	2.649	2.637	2.620	2.386
	0.10	15.071	15.055	14.851	14.657	10.421
	1.00	32.949	32.941	32.652	32.291	17.185
	3.00	37.817	37.721	37.626	37.470	21.284
	10.00	37.640	37.637	37.446	37.345	21.342
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.871	1.870	1.871	1.871	1.862
	0.01	1.941	1.940	1.937	1.934	1.899
	0.10	4.643	4.614	4.518	4.404	3.405
	1.00	25.171	25.115	24.859	24.493	19.675
	3.00	29.873	29.839	29.779	29.638	26.653
	10.00	29.953	29.930	29.843	29.733	26.779

Table A.39: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom, but 2 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.844	1.980	1.977	1.995	1.971
	0.01	3.453	2.905	3.082	2.491	1.981
	0.10	18.742	18.047	6.819	6.567	2.290
	1.00	34.645	34.635	34.426	33.502	4.661
	3.00	39.223	39.221	39.242	39.219	28.350
	10.00	39.992	39.992	40.037	39.937	28.042
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.913	1.944	1.978	1.999	1.903
	0.01	2.265	2.151	2.235	2.131	1.905
	0.10	11.302	10.117	3.746	3.826	1.981
	1.00	30.791	30.792	30.737	29.310	3.287
	3.00	35.501	35.501	35.519	35.460	26.187
	10.00	35.819	35.819	35.833	35.904	25.685
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.813	1.796	1.745	1.745	1.655
	0.01	1.844	1.813	1.765	1.756	1.655
	0.10	3.050	2.784	1.913	1.935	1.661
	1.00	21.146	21.144	21.049	19.564	2.729
	3.00	25.491	25.491	25.490	25.462	19.853
	10.00	25.661	25.660	25.660	25.649	19.448

Table A.40: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom, but 2 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.885	2.084	1.974	2.015	2.022
	0.01	5.318	3.757	3.903	4.221	2.046
	0.10	21.625	20.422	8.691	7.986	3.629
	1.00	36.506	36.521	36.937	34.930	4.673
	3.00	41.317	41.314	41.414	41.313	27.450
	10.00	41.158	41.157	41.193	41.068	28.519
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.895	1.996	2.001	2.015	2.000
	0.01	2.654	2.384	2.495	2.683	2.004
	0.10	15.071	13.440	5.361	5.107	3.057
	1.00	32.949	32.940	32.804	30.840	3.069
	3.00	37.817	37.817	37.807	37.842	26.246
	10.00	37.640	37.640	37.639	37.621	26.888
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	1.871	1.860	1.841	1.852	1.905
	0.01	1.941	1.894	1.886	1.920	1.905
	0.10	4.643	3.903	2.250	2.274	2.129
	1.00	25.171	25.172	25.027	22.942	3.965
	3.00	29.873	29.873	29.864	29.850	22.110
	10.00	29.953	29.953	29.949	29.928	22.403

Table A.41: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom, but 3 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.854	2.880	2.888	2.896	3.046
	0.01	3.247	3.245	3.234	3.228	3.218
	0.10	14.014	13.992	13.640	12.980	6.310
	1.00	35.705	35.556	35.124	34.379	6.630
	3.00	40.183	40.185	38.870	38.388	8.242
	10.00	39.093	39.056	38.785	38.105	8.437
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.745	2.746	2.746	2.743	2.728
	0.01	2.845	2.842	2.834	2.826	2.769
	0.10	7.586	7.529	7.331	7.046	4.666
	1.00	31.467	31.392	31.030	30.543	15.232
	3.00	35.451	35.456	35.268	35.038	19.493
	10.00	35.693	35.648	35.508	35.260	19.622
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.623	2.623	2.622	2.622	2.612
	0.01	2.637	2.636	2.634	2.633	2.618
	0.10	3.482	3.473	3.438	3.397	3.056
	1.00	21.808	21.755	21.532	21.161	16.477
	3.00	25.535	25.531	25.484	25.366	22.738
	10.00	25.651	25.639	25.578	25.462	22.891

Table A.42: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom, but 3 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.843	2.850	2.866	2.864	2.955
	0.01	3.419	3.413	3.409	3.381	3.239
	0.10	16.854	16.764	16.406	15.938	8.003
	1.00	36.004	36.011	35.803	35.503	7.246
	3.00	41.219	41.190	40.870	40.767	9.067
	10.00	40.630	40.636	41.000	40.259	9.257
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.843	2.841	2.840	2.835	2.844
	0.01	3.008	3.005	2.995	2.982	2.923
	0.10	10.254	10.194	9.917	9.607	6.264
	1.00	33.076	33.061	32.756	32.335	16.732
	3.00	37.638	37.704	37.511	37.371	21.218
	10.00	37.661	37.630	37.544	37.374	21.285
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.701	2.701	2.700	2.699	2.687
	0.01	2.725	2.725	2.723	2.720	2.698
	0.10	4.214	4.198	4.137	4.066	3.474
	1.00	25.656	25.605	25.356	24.976	19.863
	3.00	29.885	29.862	29.787	29.669	26.664
	10.00	29.948	29.938	29.858	29.739	26.722

Table A.43: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom, but 3 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.854	3.071	3.292	3.380	3.538
	0.01	3.247	3.245	3.446	3.926	4.182
	0.10	14.014	12.209	5.291	5.483	4.547
	1.00	35.705	35.700	36.160	34.013	5.881
	3.00	40.183	40.184	40.195	40.036	28.896
	10.00	39.093	39.092	39.096	39.121	27.488
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.745	2.728	2.712	2.756	2.764
	0.01	2.845	2.772	2.740	2.867	2.948
	0.10	7.586	6.393	3.152	3.314	3.046
	1.00	31.467	31.466	31.345	29.636	2.906
	3.00	35.451	35.451	35.459	35.433	26.568
	10.00	35.693	35.691	35.705	35.611	25.855
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.623	2.604	2.475	2.385	2.188
	0.01	2.637	2.611	2.479	2.399	2.245
	0.10	3.482	3.239	2.563	2.510	2.230
	1.00	21.808	21.808	21.717	20.119	2.753
	3.00	25.535	25.535	25.530	25.528	19.932
	10.00	25.651	25.651	25.647	25.639	19.628

Table A.44: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom, but 3 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.843	2.969	3.104	3.159	3.261
	0.01	3.419	3.223	3.464	3.758	3.263
	0.10	16.854	13.638	6.109	6.016	3.912
	1.00	36.004	35.998	35.952	33.716	4.625
	3.00	41.219	41.224	41.254	41.161	28.273
	10.00	40.630	40.632	40.586	40.628	28.776
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.843	2.856	2.934	2.932	2.817
	0.01	3.008	2.931	3.026	3.103	2.817
	0.10	10.254	7.720	3.885	3.937	3.033
	1.00	33.076	33.077	32.895	30.389	3.080
	3.00	37.638	37.639	37.657	37.607	26.504
	10.00	37.661	37.663	37.665	37.702	27.628
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	2.701	2.664	2.541	2.493	2.651
	0.01	2.725	2.674	2.552	2.515	2.651
	0.10	4.214	3.577	2.700	2.668	2.713
	1.00	25.656	25.654	25.466	23.082	3.708
	3.00	29.885	29.885	29.889	29.858	22.135
	10.00	29.948	29.948	29.948	29.936	23.029

Table A.45: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom, but 5 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.062	3.062	3.060	3.038	3.227
	0.01	3.152	3.151	3.140	3.116	3.241
	0.10	8.827	8.657	7.980	7.284	3.914
	1.00	35.635	35.517	35.303	34.831	6.301
	3.00	39.226	39.214	39.710	39.585	8.372
	10.00	38.770	38.905	39.030	38.832	8.394
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.319	3.316	3.304	3.288	3.160
	0.01	3.353	3.350	3.336	3.317	3.169
	0.10	5.862	5.818	5.653	5.465	3.947
	1.00	31.765	31.753	31.310	30.731	15.029
	3.00	35.600	35.720	35.774	35.466	19.538
	10.00	35.546	35.540	35.508	35.663	19.701
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.645	3.644	3.643	3.640	3.619
	0.01	3.654	3.653	3.651	3.648	3.623
	0.10	4.336	4.328	4.299	4.263	3.958
	1.00	22.188	22.134	21.930	21.590	17.116
	3.00	25.509	25.501	25.426	25.313	22.778
	10.00	25.650	25.633	25.565	25.472	22.905

Table A.46: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *frequency* cutoff. This polymer has 45 total degrees of freedom, but 5 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Frequency Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.494	3.494	3.493	3.483	3.645
	0.01	3.613	3.610	3.595	3.577	3.671
	0.10	10.818	10.609	9.810	8.859	4.425
	1.00	36.644	36.724	36.556	35.883	6.879
	3.00	40.479	40.444	40.617	40.748	9.139
	10.00	40.186	40.251	39.972	40.154	9.246
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.742	3.737	3.722	3.707	3.568
	0.01	3.788	3.783	3.765	3.747	3.582
	0.10	7.270	7.209	6.979	6.707	4.643
	1.00	33.528	33.460	33.135	32.693	16.427
	3.00	37.463	37.434	37.480	37.345	21.177
	10.00	37.328	37.355	37.245	37.239	21.364
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	4.054	4.053	4.050	4.048	4.024
	0.01	4.066	4.066	4.063	4.059	4.029
	0.10	5.129	5.116	5.067	5.012	4.546
	1.00	26.200	26.142	25.891	25.526	20.498
	3.00	29.865	29.846	29.769	29.639	26.619
	10.00	29.931	29.912	29.839	29.725	26.774

Table A.47: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 2000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom, but 5 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.062	3.076	5.269	5.411	4.487
	0.01	3.152	3.120	5.300	5.557	4.487
	0.10	8.827	7.815	5.527	6.143	4.791
	1.00	35.635	35.630	36.056	35.185	5.317
	3.00	39.226	39.227	39.299	39.423	28.758
	10.00	38.770	38.770	38.801	38.883	27.931
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.319	3.175	3.003	2.964	2.818
	0.01	3.353	3.185	3.002	2.971	2.818
	0.10	5.862	5.403	3.048	3.019	2.854
	1.00	31.765	31.763	31.737	30.770	2.769
	3.00	35.600	35.600	35.586	35.630	26.521
	10.00	35.546	35.546	35.553	35.538	25.931
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.645	3.616	3.472	3.380	2.932
	0.01	3.654	3.621	3.473	3.389	2.932
	0.10	4.336	4.220	3.534	3.444	2.976
	1.00	22.188	22.188	22.141	21.380	3.414
	3.00	25.509	25.509	25.512	25.498	19.750
	10.00	25.650	25.650	25.659	25.641	19.575

Table A.48: Dimensionality estimates at different sets of k for the *correlated helix* model with $N = 5000$ structures of length 25 across all noise levels and all levels of DFT smoothing using a fractional *amplitude* cutoff. This polymer has 45 total degrees of freedom, but 5 are attributed to large amplitude, correlated folding/unfolding dynamics.

		Amplitude Smoothing Fraction				
		0.00	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.494	3.702	5.288	5.051	4.817
	0.01	3.613	3.741	5.322	5.140	4.817
	0.10	10.818	8.097	5.864	5.664	5.165
	1.00	36.644	36.649	36.525	35.517	5.585
	3.00	40.479	40.480	40.474	40.242	28.534
	10.00	40.186	40.184	40.191	40.085	29.382
Med. k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	3.742	3.514	3.344	3.335	3.170
	0.01	3.788	3.526	3.346	3.344	3.170
	0.10	7.270	5.943	3.447	3.456	3.236
	1.00	33.528	33.518	33.245	31.877	3.027
	3.00	37.463	37.462	37.477	37.459	27.052
	10.00	37.328	37.328	37.350	37.279	27.585
Large k		**0.00**	**0.01**	**0.05**	**0.10**	**0.50**
Noise ($\sigma_{\theta,\phi}$)	**0.00**	4.054	3.998	3.799	3.637	3.549
	0.01	4.066	4.003	3.801	3.641	3.549
	0.10	5.129	4.743	3.882	3.720	3.658
	1.00	26.200	26.196	26.104	24.679	4.446
	3.00	29.865	29.864	29.867	29.850	22.539
	10.00	29.931	29.930	29.929	29.929	22.912

References

[1] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Science*, vol. 11, no. 4, pp. 739–756, Apr. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11910019

[2] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, May 2002. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/bi012159%2B

[3] T. Mittag and J. D. Forman-Kay, "Atomic-level characterization of disordered protein ensembles," *Current Opinion in Structural Biology*, vol. 17, no. 1, pp. 3–14, Feb. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17250999

[4] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–8, 27–8, Feb. 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8744570

[5] V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky, and A. K. Dunker, "Characterization of molecular recognition features, MoRFs, and their binding partners," *Journal of Proteome Research*, vol. 6, no. 6, pp. 2351–2366, Jun. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2570643&tool=pmcentrez&rendertype=abstract

[6] B. A. Shoemaker, J. J. Portman, and P. G. Wolynes, "Speeding molecular recognition by using the folding funnel: the fly-casting mechanism," *Proceedings of the National Academy of Sciences*, vol. 97, no. 16, pp. 8868–8873, Aug. 2000. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16787&tool=pmcentrez&rendertype=abstract

[7] K. Sugase, H. J. Dyson, and P. E. Wright, "Mechanism of coupled folding and binding of an intrinsically disordered protein," *Nature*, vol. 447, no. 7147, pp. 1021–1025, Jun. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17522630

[8] J. Yamada, J. L. Phillips, S. Patel, G. Goldfien, A. Calestagne-Morelli, H. Huang, R. Reza, J. Acheson, V. V. Krishnan, S. Newsam, A. Gopinathan, E. Y. Lau, M. E. Colvin, V. N. Uversky, and M. F. Rexach, "A bimodal distribution of two distinct categories of intrinsically disordered structures with separate functions in FG nucleoporins." *Molecular & Cellular Proteomics*, vol. 9, no. 10, pp. 2205–2224, Oct. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20368288http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2953916&tool=pmcentrez&rendertype=abstract

[9] R. K. Das, S. L. Crick, and R. V. Pappu, "N-terminal segments modulate the α-helical propensities of the intrinsically disordered basic regions of bZIP proteins," *Journal of Molecular Biology*, vol. 416, no. 2, pp. 287–299, Feb. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22226835

[10] M. P. Rout and J. D. Aitchison, "The nuclear pore complex as a transport machine," *The Journal of Biological Chemistry*, vol. 276, no. 20, pp. 16 593–16 596, May 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11283009

[11] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, Dec. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16200636

[12] Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis," *Proteins*, vol. 58, no. 1, pp. 45–52, Jan. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15521057

[13] E. Y. Lau, J. L. Phillips, and M. E. Colvin, "Molecular dynamics simulations of highly charged green fluorescent proteins," *Molecular Physics*, vol. 107, no. 8, pp. 1233–1241, Jan. 2009. [Online]. Available: http://www.informaworld.com/openurl?genre=article&doi=10.1080/00268970902845305&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3

[14] 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*, vol. 11, no. 11, pp. 2606–2621, Nov. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12381844

[15] R. I. Dima and D. Thirumalai, "Asymmetry in the shapes of folded and denatured states of proteins," *The Journal of Physical Chemistry B*, vol. 108, no. 21, pp. 6564–6570, May 2004. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jp037128y

[16] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6667333

[17] A. Vitalis, X. Wang, and R. V. Pappu, "Quantitative characterization of intrinsic disorder in polyglutamine: insights from analysis based on polymer theories," *Biophysical Journal*, vol. 93, no. 6, pp. 1923–1937, Sep. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1959550&tool=pmcentrez&rendertype=abstract

[18] E. Lyman and D. M. Zuckerman, "On the structural convergence of biomolecular simulations by determination of the effective sample size," *The Journal of Physical Chemistry B*, vol. 111, no. 44, pp. 12 876–12 882, Nov. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2538559&tool=pmcentrez&rendertype=abstract

[19] ——, "Ensemble-based convergence analysis of biomolecular trajectories," *Biophysical Journal*, vol. 91, no. 1, pp. 164–172, Jul. 2006. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1479051&tool=pmcentrez&rendertype=abstract

[20] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, *Graphical Methods for Data Analysis*. Boston, MA: Duxbury Press, 1983.

[21] V. V. Krishnan, E. Y. Lau, J. Yamada, D. P. Denning, S. S. Patel, M. E. Colvin, and M. F. Rexach, "Intramolecular cohesion of coils mediated by phenylalanine–glycine motifs in the natively unfolded domain of a nucleoporin," *PLoS Computational Biology*, vol. 4, no. 8, p. e1000145, Jan. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2475668&tool=pmcentrez&rendertype=abstract

[22] J. L. Phillips, M. E. Colvin, E. Y. Lau, and S. Newsam, "Analyzing dynamical simulations of intrinsically disordered proteins using spectral clustering," in *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. Philadelphia, PA: IEEE, Nov. 2008, pp. 17–24. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4686204

[23] M. E. Karpen, D. J. Tobias, and C. L. Brooks, "Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV," *Biochemistry*, vol. 32, no. 2, pp. 412–420, Jan. 1993. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8422350

[24] C. Best and H.-C. Hege, "Visualizing and identifying conformational ensembles in molecular dynamics trajectories," *Computing in Science & Engineering*, vol. 4,

no. 3, pp. 68–75, 2002. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=998642

[25] 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*, vol. 104, no. 12, pp. 4925–4930, Mar. 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1829241&tool=pmcentrez&rendertype=abstract

[26] P. L. Freddolino and K. Schulten, "Common structural transitions in explicit-solvent simulations of villin headpiece folding," *Biophysical Journal*, vol. 97, no. 8, pp. 2338–2347, Oct. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19843466

[27] S. Rauscher and R. Pomès, "Molecular simulations of protein disorder," *Biochemistry and Cell Biology*, vol. 88, no. 2, pp. 269–90, Apr. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20453929

[28] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, "Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms," *Journal of Chemical Theory and Computation*, vol. 3, no. 6, pp. 2312–2334, Nov. 2007. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ct700119m

[29] M. Meila and J. Shi, "A random walks view of spectral segmentation," in *AISTATS*, 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.1501

[30] M. S. Apaydin, D. L. Brutlag, C. Guestrin, D. Hsu, J.-C. Latombe, and C. Varma, "Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 257–281, Jan. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12935328

[31] B. Keller, P. Hunenberger, and W. F. van Gunsteren, "An analysis of the validity of Markov state models for emulating the dynamics of classical molecular systems and ensembles," *Journal of Chemical Theory and Computation*, vol. 7, no. 4, pp. 1032–1044, Mar. 2011. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ct200069c

[32] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2002, pp. 849–856. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=ADF7700D03B148CBA60CF58E57B47480?doi=10.1.1.19.8100&rep=rep1&type=pdf

[33] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*, vol. 2. MIT Press, 2005, pp. 1601–1608. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.7940

[34] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 1–32, Nov. 2007. [Online]. Available: http://arxiv.org/abs/0711.0189

[35] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, no. 233. University of California, 1967, pp. 281–297.

[36] S. Haykin, *Communication Systems*, 3rd ed. Wiley Publishing, 1994.

[37] J. Ma, "Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes," *Structure*, vol. 13, no. 3, pp. 373–380, Mar. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15766538

[38] E. Lyman, J. Pfaendtner, and G. A. Voth, "Systematic multiscale parameterization of heterogeneous elastic network models of proteins," *Biophysical Journal*, vol. 95, no. 9, pp. 4183–4192, Nov. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567941&tool=pmcentrez&rendertype=abstract

[39] A. Amadei, A. B. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," *Proteins*, vol. 17, no. 4, pp. 412–425, Dec. 1993. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8108382

[40] D. F. Lowry, A. C. Hausrath, and G. W. Daughdrill, "A robust approach for analyzing a heterogeneous structural ensemble," *Proteins*, vol. 73, no. 4, pp. 918–928, Dec. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18536020

[41] M. Feher and J. M. Schmidt, "Metric and multidimensional scaling: efficient tools for clustering molecular conformations," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 2, pp. 346–53, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11277721

[42] N. C. Benson and V. Daggett, "Dynameomics: large-scale assessment of native protein flexibility," *Protein Science*, vol. 17, no. 12, pp. 2038–2050, Dec. 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2590920&tool=pmcentrez&rendertype=abstract

[43] M. L. Teodoro, G. N. Phillips, and L. E. Kavraki, "Understanding protein flexibility through dimensionality reduction," *Journal of Computational Biology*, vol. 10, no. 3, pp. 617–634, 2003.

[44] H. Stamati, C. Clementi, and L. E. Kavraki, "Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides," *Proteins*, vol. 78, no. 2, pp. 223–235, Feb. 2010. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2795065&tool=pmcentrez&rendertype=abstract

[45] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11125149

[46] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods," *Proceedings of the National Academy of Sciences*, vol. 102, no. 21, pp. 7432–7437, May 2005. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1140426&tool=pmcentrez&rendertype=abstract

[47] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/089976603321780317

[48] M. Ceriotti, G. a. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," *Proceedings of the National Academy of Sciences*, vol. 108, no. 32, pp. 13 023–13 028, Aug. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3156203&tool=pmcentrez&rendertype=abstract

[49] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15*, 2003, pp. 833–840.

[50] J. C. Gower and P. Legendre, "Metric and Euclidean properties of dissimilarity coefficients," *Journal of Classification*, vol. 3, no. 1, pp. 5–48, Mar. 1986. [Online]. Available: http://www.springerlink.com/index/10.1007/BF01896809

[51] J. C. Lingoes, "Some boundary conditions for a monotone analysis of symmetric matrices," *Psychometrika*, vol. 36, no. 2, pp. 195–203, Jun. 1971. [Online]. Available: http://www.springerlink.com/index/10.1007/BF02291398

[52] F. Cailliez, "The analytical solution of the additive constant problem," *Psychometrika*, vol. 48, no. 2, pp. 305–308, Jun. 1983. [Online]. Available: http://www.springerlink.com/index/10.1007/BF02294026

[53] V. Roth, J. Laub, M. Kawanabe, and J. M. Buhmann, "Optimal cluster preserving embedding of nonmetric proximity data," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol. 25, no. 12, pp. 1540–1551, Dec. 2003. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1251147

[54] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966. [Online]. Available: http://biomet.oxfordjournals.org/cgi/doi/10.1093/biomet/53.3-4.325

[55] M. W. Trosset and C. E. Priebe, "The out-of-sample problem for classical multidimensional scaling," *Computational Statistics & Data Analysis*, vol. 52, no. 10, pp. 4635–4642, Jun. 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167947308001515

[56] L. Miao and K. Schulten, "Transport-related structures and processes of the nuclear pore complex studied through molecular dynamics," *Structure*, vol. 17, no. 3, pp. 449–459, Mar. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19278659

[57] G. Jayachandran, V. Vishal, and V. S. Pande, "Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece," *The Journal of Chemical Physics*, vol. 124, no. 16, p. 164902, Apr. 2006. [Online]. Available: http://link.aip.org/link/JCPSA6/v124/i16/p164902/s1&Agg=doihttp://www.ncbi.nlm.nih.gov/pubmed/16674165

[58] G. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 857–864. [Online]. Available: http://books.nips.cc/papers/files/nips15/AA45.pdf

[59] L. J. P. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[60] M. Á. Carreira-Perpiñán, "The elastic embedding algorithm for dimensionality reduction," in *Proceedings of the 12th International Conference on Machine Learning*. Omnipress, 2010, pp. 167–174. [Online]. Available: http://www.icml2010.org/papers/123.pdf

[61] M. Vladymyrov and M. Á. Carreira-Perpiñán, "Partial-Hessian strategies for fast learning of nonlinear embeddings," in *Proceedings of the 12th International Conference on Machine Learning*. Omnipress, 2012. [Online]. Available: http://icml.cc/2012/papers/199.pdf

[62] G. Haro, G. Randall, and G. Sapiro, "Stratification learning: detecting mixed density and dimensionality in high dimensional point clouds," in *Advances in Neural Information Processing Systems 19*. MIT Press, 2007, pp. 553–560. [Online]. Available: http://books.nips.cc/papers/files/nips19/NIPS2006_0183.pdf

[63] ——, "Translated Poisson mixture model for stratification learning," *International Journal on Computer Vision*, vol. 80, no. 3, pp. 358–374, 2008. [Online]. Available: http://dx.doi.org/10.1007/s11263-008-0144-6

[64] W. Wang and M. Á. Carreira-Perpiñán, "Manifold blurring mean shift algorithms for manifold denoising," in *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1759–1766. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2010.5539845

[65] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, "Atomic-level characterization of the structural dynamics of proteins," *Science*, vol. 330, no. 6002, pp. 341–346, Oct. 2010. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1187409

[66] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten, "Ten-microsecond molecular dynamics simulation of a fast-folding WW domain," *Biophysical Journal*, vol. 94, no. 10, pp. L75–7, May 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18339748

[67] R. Day, D. Paschek, and A. E. Garcia, "Microsecond simulations of the folding/unfolding thermodynamics of the Trp-cage miniprotein," *Proteins*, vol. 78, no. 8, pp. 1889–1899, Jun. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20408169

[68] D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, 2nd ed. New York: Cambridge University Press, 2004. [Online]. Available: http://www.cambridge.org/uk/catalogue/catalogue.asp?isbn=9780521825689

[69] K. Hinsen, "Comment on: "Energy landscape of a small peptide revealed by dihedral angle principal component analysis"," *Proteins*, vol. 64, no. 3, pp. 795–797, Aug. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16456860

[70] F. Camastra, "Data dimensionality estimation methods: a survey," *Pattern Recognition*, vol. 36, no. 12, pp. 2945–2954, Dec. 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0031320303001766

[71] J. Costa and A. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2210–2221, Aug. 2004. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1315941

[72] B. Kegl, "Intrinsic dimension estimation using packing numbers," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003.

[73] P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Physical Review Letters*, vol. 50, no. 5, pp. 346–349, Jan. 1983. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.50.346

[74] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1404–1407, Oct. 2002. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1039212

[75] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005.

[76] D. J. C. MacKay and Z. Ghahramani, "Comments on "Maximum likelihood estimation of intrinsic dimension" by E . Levina and P . Bickel (2004)," pp. 1–5, 2005. [Online]. Available: http://www.inference.phy.cam.ac.uk/mackay/dimension/

[77] J. N. Bright, T. B. Woolf, and J. H. Hoh, "Predicting properties of intrinsically unstructured proteins," *Progress in Biophysics and Molecular Biology*, vol. 76, no. 3, pp. 131–73, Jul. 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11709204

[78] P. J. Flory, *Statistical Mechanics of Chain Molecules*. New York: Wiley Publishing, 1967.

[79] R. B. Best, N.-V. Buchete, and G. Hummer, "Are current molecular dynamics force fields too helical?" *Biophysical Journal*, vol. 95, no. 1, pp. L07–L09, 2008.

[80] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill, "Small molecule hydration free energies in explicit solvent : an extensive test of fixed-charge atomistic simulations," *Journal of Chemical Theory and Computation*, vol. 5, no. 2, pp. 350–358, 2009.

[81] L. Wickstrom, A. Okur, and C. Simmerling, "Evaluating the performance of the ff99SB force field based on NMR scalar coupling data," *Biophysical Journal*, vol. 97, no. 3, pp. 853–856, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.bpj.2009.04.063

[82] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 8, pp. 1950–1958, 2010. [Online]. Available: http://doi.wiley.com/10.1002/prot.22711

[83] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," *Proteins*, vol. 65, no. 3, pp. 712–725, Nov. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16981200

[84] P. S. Nerenberg and T. Head-Gordon, "Optimizing proteinsolvent force fields to reproduce intrinsic conformational preferences of model peptides," *Journal of Chemical Theory and Computation*, vol. 7, no. 4, pp. 1220–1230, Mar. 2011. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ct2000183

[85] R. B. Best and J. Mittal, "Protein simulations with an optimized water model: cooperative helix formation and temperature-Induced unfolded state collapse," *The Journal of Physical Chemistry B*, vol. 114, no. 46, pp. 14 916–14 923, Nov. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21038907

[86] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, Mar. 2008. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ct700301q

[87] A. Onufriev, D. Bashford, and D. A. Case, "Exploring protein native states and large-scale conformational changes with a modified generalized born model," *Proteins*, vol. 55, no. 2, pp. 383–394, May 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15048829

[88] G. Williams and A. J. Toon, "Protein folding pathways and state transitions described by classical equations of motion of an elastic network model," *Protein Science*, vol. 19, no. 12, pp. 2451–2461, Dec. 2010. [Online]. Available: http://doi.wiley.com/10.1002/pro.527http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3009412&tool=pmcentrez&rendertype=abstract

[89] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, "The MARTINI coarse-grained force field: extension to proteins," *Journal of Chemical Theory and Computation*, vol. 4, no. 5, pp. 819–834, May 2008. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ct700324x

[90] 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*, vol. 5, no. 8, pp. 1093–1108, Aug. 1997. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9309224