UNIVERSITY OF CALIFORNIA, MERCED

DOCTORAL DISSERTATION

Development and Application of Analysis Tools Optimized For Intrinsically Disordered Proteins

Author:
Timothy G. CONNOLLY

Supervisor:
Professor Michael COLVIN

A dissertation submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in

Quantitative and Systems Biology
School of Natural Sciences

Committee:

Ajay Gopinathan, Chair

Michael Colvin

Christine Isborn

Shawn Newsam

April 27, 2018

UNIVERSITY OF CALIFORNIA

MERCED

Development and Application of Analysis Tools Optimized For Intrinsically Disordered Proteins

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Quantitative and Systems Biology

by

Timothy G. CONNOLLY

May 2018

The Dissertation of Timothy G. Connolly is approved:

Ajay Gopinathan, Chair

Michael Colvin

Christine Isborn

Shawn Newsam

Curriculum Vita

Timothy G. Connolly

Quantitative and Systems Biology
School of Natural Sciences
University of California, Merced

email: tconnolly@ucmerced.edu
url: https://github.com/colvin-lab/g_isd

Fields of Study

Computation Biophysics

Molecular Dynamics Simulations

Intrinsically Disordered Proteins

Cellular Signaling Networks

Machine Learning

Mathematical Programming

Education

2012-2018 Doctor of Philosophy in Quantitative and Systems Biology, University of California, Merced, Merced, CA: Molecular dynamics simulations of intrinsically disordered proteins. The continuum of protein disorder and flexibility. Development of analysis tools optimized for proteins with disordered regions. Applying principles of machine learning to biological systems in order to implement dimensionality reduction and novel clustering methods.

2011 Master of Science in Biomedical Engineering, Wayne State University, Detroit, MI: Simulations of cellular signaling networks based on enzyme kinetics.

2006 Bachelor of Science in Bioengineering, University of Washington, Seattle, WA: Molecular dynamics simulations of the von Willebrand Factor responsible for von Willebrand Disease. Supervisor: Wendy Thomas.

Publications

2016 Xingyuan Shen, Timothy Connolly, Yuhui Huang, Michael Colvin, Changchun Wang, Jennifer Lu. (2016) Adjusting Local Molecular Environment for Giant Ambient Thermal Contraction. Macromol. Rapid Commun., 37: 1904-1911. doi:10.1002/marc.201600045

2014 Timothy G. Connolly, David Ando, Robert L. Wang, Ajay Gopinathan, Shawn D. Newsam, Michael E. Colvin. (2014) Identifying Local Regions of Order and Disorder in FG-Nucleoporins and Partially Disordered Proteins Using Molecular Dynamics Simulations. Biophysical Journal, Volume 106, Issue 2, 608a.

2014 Robert L. Wang, Timothy G. Connolly, Joshua L. Phillips, Amanda V. Miguel, Ajay Gopinathan, Shawn D. Newsam, Michael E. Colvin. (2014) Comparison of Metrics of Inter-Structure Distance When Applied to Molecular Dynamics Simulations of Intrinsically Disordered Proteins. Biophysical Journal, Volume 106, Issue 2, 610a - 611a.

2012 Timothy G. Connolly, Robert L. Wang, Amanda V. Miguel, Joshua L. Phillips, Edmond Y. Lau, Michael E. Colvin. (2012) Measuring Disorder and Dynamical Properties of FG-Nucleoporins. Biophysical Journal, Volume 104, Issue 2, 233a - 234a.

Software

g_isd Performs numerous types of protein analyses based on the concept of inter-structure distance. Estimates explored conformational space and dynamics of exploration. Detects local flexibility and disorder in folded proteins. Detects local stiffness and structure in disordered proteins. Assigns order parameter. Implements classical multidimensional scaling and a dimensionality estimator. Performs protein clustering using hierarchical spectral clustering and K-means spectral clustering.

doCluster.m Protein ensemble visualization and clustering tool for the GNU Octave or MATLAB programming environments. Displays the exploration of protein conformational space in 3 or 6 dimensions. Performs several methods of ensemble clustering.

RanchaR An efficient, parallelized random polymer chain generator for the R environment. Generates ensembles of polypeptides with entropic chain behavior using a self-avoiding random walk. Applies filters to generate ensembles with specified sizes.

g_dijkstra Computes correlated protein motion based on mutual information theory. Applies Dijkstra analysis to identify communication pathways in proteins.

g_shape A port of the pre-existing g_shape analysis tool for compatibility with Gromacs versions 5.1+ and 2016.X.

Teaching

2015-2017 Calculus. Lecture. Undergraduate.

2016 Introduction To Scientific Data Analysis. Computer lab. Undergraduate.

2015 Biostatistics. Computer lab and lecture.

2014 Molecular Dynamics. Computer lab. Graduate and undergraduate level course.

UNIVERSITY OF CALIFORNIA, MERCED

Abstract

Quantitative and Systems Biology

School of Natural Sciences

Doctor of Philosophy

Development and Application of Analysis Tools Optimized For Intrinsically Disordered Proteins

by Timothy G. CONNOLLY

The study of intrinsically disordered proteins has rapidly advanced since the identification of the role they play in neurodegenerative diseases. Molecular dynamics simulations of disordered proteins have become common, but analysis tools optimized for their study have lagged behind. Both fully and partially disordered proteins present similar challenges: a vast fold space and difficultly in distinguishing meaningful protein motion. We have implemented an analysis tool based on inter-structure distance. This tool, *g_isd*, quantifies the differences between protein conformations. Our analysis is able to identify local regions that are flexible or disordered in otherwise folded proteins by employing a universal parameter that we developed to describe disorder. This order parameter has been scaled to be comparable between all proteins regardless or size or sequence length. We present one of the only clustering algorithms truly optimized to study protein dynamics and are making it available as open source software. This hierarchical spectral clustering applies empirically-derived data to estimate meaningful protein motion allows unsupervised molecular dynamics clustering in reduced dimensional space. We apply our approach to the disordered loop region of a cystine knot protein. Analysis describes the dynamics of this loop containing a targeted binding sequence for the cancer-associated integrin $\alpha v \beta 6$ protein. A sequence of steps to dock the cystine knot protein to its target as a large ligand is characterized. Finally, we analyze the disorder of a synthetic polymer with the useful property of thermal contraction. Molecular dynamics studies with a customized force field explain that a small difference in a single bond leads to significant disorder. The efficiency of thermal contraction can be modulated by varying levels of disorder in the material.

Acknowledgements

This work would not have been possible without my collaborators. Dr. Xingyuan Shen, Dr. David Ando, Dr. Fabian Filipp, and Robert Wang all provided direct assistance on portions of the project described here. Thanks to Dr. Joshua Phillips who initially got me started on this project with some sample source code and materials to base my work on. Dr. Ajay Gopinathan, Dr. Shawn Newsam, Dr. Fabian Filipp, Dr. Christine Isborn, and Dr. Jennifer Lu have advised either myself or members on my list of collaborators over the years. They initiated many of the projects described here. Dr. Michael Colvin was the unstuckness provider. Running into roadblocks in programming, software, or mathematics generally resulted in a visit to get unstuck before I was able to resume progress. Thanks to all. Research was funded in part by NSF grant #0960480.

More specific acknowledgements are included in individual chapters.

Contents

List of Figures

List of Tables

List of Abbreviations

ISD	interstructure distance
ISDM	interstructure distance measure
RMSD	root-mean-square deviation
MD	molecular dynamics
IDP	intrinsically disordered protein
PDP	partially disordered proteins
SARW	self-avoiding random walk
pdb	protein database
VMD	Visual Molecular Dynamics
NVT	constant N, volume, and temperature
NPT	constant N, pressure, and temperature
RDF	radial distribution function

ISDMs:	
ang	backbone angles
dih	backbone dihedrals
angdih	combined ang and dih
phipsi	ϕ and ψ angles
rmsd	RMSD-based ISDM (lower case)
grot	grid-search rotated RMSD
rrot	randomly-rotated RMSD
distance	distance RMSD
rg	radius of gyration
e2e	end-to-end distance
esa	elastic shape analysis
pcor	position correlation
acor	backbone angle correlation

Chapter 1

Measures of Inter-structure Distance

1.1 Background

Observations of the structural similarities between proteins have been widely utilized in bioinformatics to reveal evolutionary relationships and categorize protein fold motifs [1–3]. While there are numerous accurate and useful measures of inter-structure distance (ISD) between like protein structures [4], inferring relationships between distantly related proteins has required the development of tools to quantify the similarities and differences between highly divergent structures [5, 6]. Natively folded proteins exist in a highly constrained dynamical subspace with only a few essential degrees of freedom [7]. On the other hand, intrinsically disordered proteins (IDPs) exist in a naturally unfolded and unconstrained state consisting of an ensemble of transient, dissimilar conformations [8–12]. These ensembles are in many ways analogous to sets of distantly related proteins, and therefore some of the common methods to calculate ISD may not be appropriate.

Knowledge of IDPs has matured in recent years with advances in proteomics towards predictions of disorder [13] and the establishment of databases for disordered proteins [14, 15] and native protein flexibility [16]. Molecular dynamics (MD) simulations of IDPs are capable of estimating subsets of these ensembles [17, 18], but most current tools to analyze MD trajectories are optimized to study folded proteins sampling relatively small conformational spaces. In recent years, several algorithms have been developed with the purpose of analyzing MD simulations of disordered proteins which use aligned root-mean-square deviation (RMSD) as the underlying method to measure the difference between structures [19–22]. The original motivation for the work presented here was the observation that the commonly used method of measuring ISD, RMSD, appears to saturate to a maximal distance within tens to hundreds of nanoseconds in simulations of highly disordered proteins (**Figure 1.1**). Once the method of comparison is saturated to a maximal value, RMSD is unable to differentiate between structures by via ISD. For MD simulations of IDPs for even moderate durations, RMSD computes that the majority of paired clusters are nearly equally distant from one another. RMSD is also highly dependent on the size of the structures being compared while MD simulations of IDPs show a wide variance in the radius of gyration (R_g) [23]. For IDPs with nearly entropic chain behavior, the structural ensemble will contain both extended and compact globular conformations

The dynamics of IDPs are frequently described as similar to random coils [24]; however, observations that IDPs are often more compact than random coil behavior would predict [25, 26] suggests the behavior of IDPs is more complex: possibly modulated by charge content and solvent quality [27, 28]. Both computational and experimental approaches have revealed short-lived, metastable conformations when studying specific systems of IDPs [29, 30] which indicates that there will be utility in the future for clustering methods optimized for the unique energetic and conformational landscape of IDPs [21]. As it becomes clear that the fold space of proteins is continuous rather than discrete [31] and that protein disorder is a continuum rather than a transition [32], tools optimized for the study of IDPs can be applied to certain folded proteins as well.

We have created and validated several tools based on the libraries and interface of the MD simulation software Gromacs [33–38]. In addition, we have completed a library which implements 16 measures of inter-structure distance (ISDM). The ISDMs that we have investigated include modifications to the standard RMSD, various algorithms using internal coordinates such as backbone angles and dihedrals, correlation coefficients based on internal and aligned external coordinates, and two additional measures of structure comparisons based on recent publications: MAMMOTH [39] and elastic shape analysis [40]. By implementing a wide variety of approaches to calculating ISD, our analysis is robust to many types of biological systems. Internal proteins coordinates using backbone angles are less size-dependent; however, they are generally only applicable to complete stretches of protein polymers. On the other hand, binding sites tend to be spatially associated without being directly bonded. Therefore, external coordinate based systems such as RMSD can be applied where internal coordinate based systems would fail. Some ISDMs may have more sensitivity to large changes in structure (IDPs) while others may have more sensitivity to small changes in structure (folded proteins). The ISDMs are tested based on: (1) their ability to differentiate more disordered proteins from less disordered one, (2) how quickly they saturate by measuring a maximal distance between proteins, and (3) their ability to compare different protein systems based on size-independence.

The ISDM options were tested on a set of three simplified model IDPs with a known spectrum of disorder. These model IDPs are homopolymer systems that include a highly disordered and flexible polypeptide (polyglycine), a polymer with significant transient structure and backbone rigidity (polyglutamine), and a third polymer which displays a mixture of the two behaviors (polyalanine).

Experimental results of polyglycine show both a preference for highly flexible, extended conformations [41] and self-aggregation into disordered amyloid-like fibrils [42].

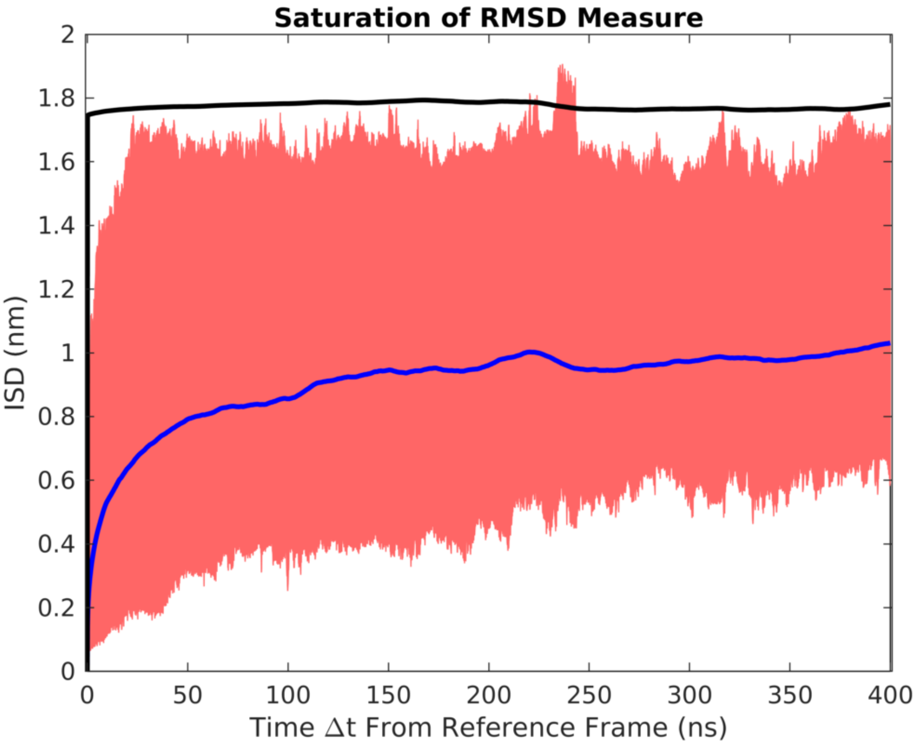


FIGURE 1.1: A 1.0 μs MD simulation of a 50 amino acid fragment of FG-nup nsp1 was sampled every 100 ps. Decorrelation of mean RMSD (blue line) is set against the range from the minimum to maximum values of aligned RMSD (area shown in red). The range of aligned RMSD rapidly approaches a measure of unaligned RMSD (black line) which is a representation of the maximum RMSD possible for a molecule of the given size. The decorrelations are averaged across all pairs of frames separated by the time shown, Δt. Within 30 ns, the aligned RMSD of the maximally distant structures is saturated since the ceiling of the RMSD range is essentially constant and nearly equal to unaligned structures. In addition, the decorrelation of the mean ISD reveals most of the sensitivity to average distance is lost within 50 ns of simulation time.

MD simulations of polyglycine show more compact but still highly flexible and disordered behavior [43].

MD simulations of polyalanine reveal significant proportions of α-helical structure in agreement with experimental results [44, 45]. However, simulations of longer polyalanine chains were shown to have a phase transition to a more complex helix-turn-helix ensemble for greater than 40-45 amino acids [44]. Polyglutamine is a disordered but intrinsically stiff polypeptide [46] which forms collapsed spherical globules in water [47].

MD simulations of polyglutamine confirm that the protein is disordered but tends to form significant proportions of β-sheet secondary structure [48, 49]. In addition, MD simulations of the polypeptide with the addition of small concentrations of NaCl show significant α-helical structure [50]. These three systems were chosen for MD simulations because we believe them to follow a spectrum from most disordered (polyglycine) to least disordered (polyglutamine) without the additional complication of variations of disorder in local regions seen in more complex sequences.

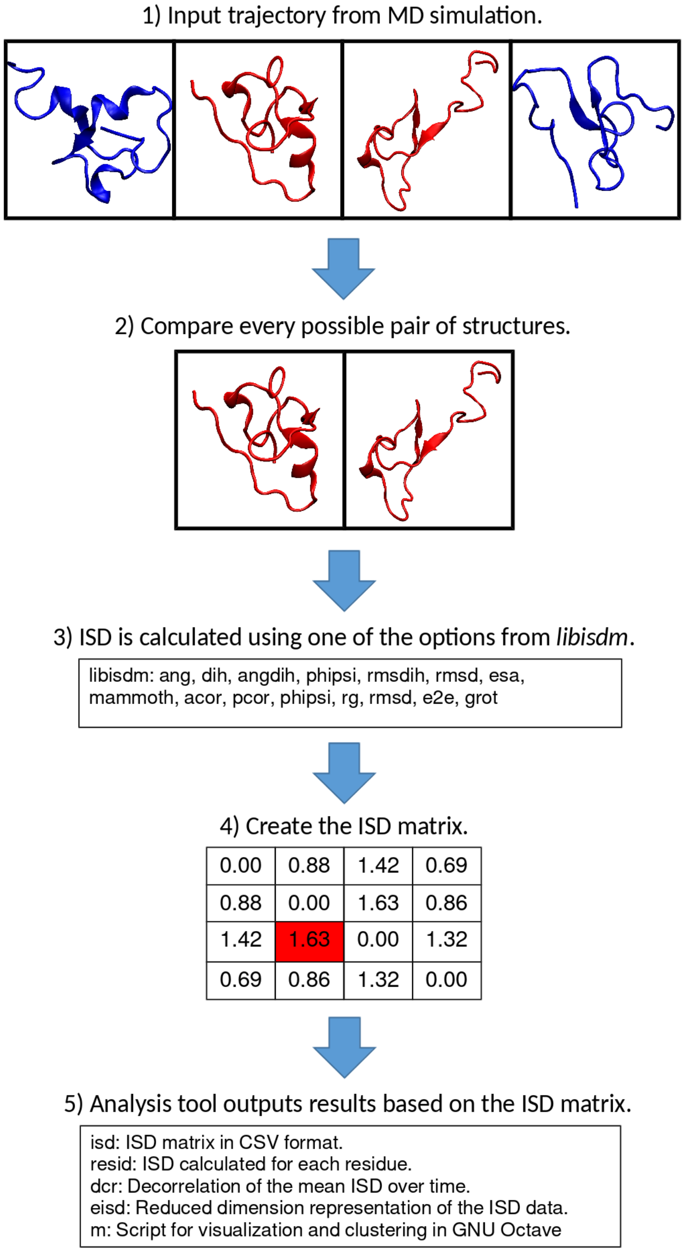
In addition to simplified polymers, we have chosen to test segments of nucleoporin proteins rich in phenylalanine-glycine repeats (FG-nups) as our model system of realistic IDPs. FG-nups are natively unfolded [51] yet have a known function in gating protein diffusion across the nuclear pore complex [52]. FG-nups have complex amino acid sequences with regional variance in charge content and may display consistent compaction under native conditions [53].

1.2 Methods

1.2.1 Implementation of ISDMs

To implement the library of ISDMs described, an analysis tool was developed based on Gromacs functions and libraries [33] called *g_isd*. Analysis requires trajectory and topology files as input in Gromacs-compatible formats. The user must also choose a single option from the implemented methods to compute ISD between structures.

Figure 1.2 explains the basic algorithm used by the *g_isd* analysis tool. Briefly, *g_isd* loads the trajectory and topology information then compares every possible pair of structures using the ISDM option chosen by the user. Each pair of structures results in an ISD calculation that fills in one position of the ISD matrix. The entire ISD matrix represents all possible pairwise comparisons. There are options for *g_isd* to operate on either the entirety or only a portion of the calculated ISD matrix. As an example, a pair of structures is highlighted in red, and the rmsd option is used to calculate their inter-structure distance as 1.63 nm which fills in one position of the ISD matrix. Note that the ISD matrix



1) Input trajectory from MD simulation.

2) Compare every possible pair of structures.

3) ISD is calculated using one of the options from *libisdm*.

libisdm: ang, dih, angdih, phipsi, rmsdih, rmsd, esa, mammoth, acor, pcor, phipsi, rg, rmsd, e2e, grot

4) Create the ISD matrix.

0.00	0.88	1.42	0.69
0.88	0.00	1.63	0.86
1.42	1.63	0.00	1.32
0.69	0.86	1.32	0.00

5) Analysis tool outputs results based on the ISD matrix.

isd: ISD matrix in CSV format.
resid: ISD calculated for each residue.
dcr: Decorrelation of the mean ISD over time.
eisd: Reduced dimension representation of the ISD data.
m: Script for visualization and clustering in GNU Octave

FIGURE 1.2: Analysis workflow for the *g_isd* analysis tool. The command-line application accepts trajectories from MD simulations in Gromacs-compatible formats. The ISD of every pair of structures is calculated using one of the implemented ISDMs to build an ISD matrix. Several types of analysis utilize the computed ISD. As an example, (1) a pair of structures marked in red is chosen, (2) the pair is compared with (3) the RMSD ISDM for a resulting distance of 1.63 nm. (4) The result is stored in the ISD matrix (red background).

TABLE 1.1: Sequences of simulated nucleoporins.

nsp1	AFSFGAKPDENKASATSKPAFSFGA KPEEKKDDNSSKPAFSFGAKSNEDK
nup116	ASSSGAKPDENKASATSKPASSSGA KPEEKKDDNSSKPASSSGAKSNEDK

is generally symmetric, but some of the implemented ISDMs do not guarantee this. The ISD matrix can be coerced into a symmetric matrix with the -symmetric option.

The analysis tool *g_isd* uses the ISD matrix to carry out a variety of analysis options. The -isd output creates a comma-separated values (CSV) format file containing the entire ISD matrix with a summary of mean and maximum ISD. The ISD matrix in CSV format can be converted into heat map showing groupings of similar conformations over time. Structures retaining similarity over brief periods can be identified as transient metastable states. These metastable conformations appear as squares along the diagonal of the heat map with low ISD scores. The areas along the diagonal between low ISD squares can be identified as transition states. The length along the diagonal corresponds to the period of the transition.

1.2.2 MD Simulations

All-atom MD simulations were run on protein fragments of 50 amino acids in length using the velocity rescaling thermostat [54] and Parinello-Rahman pressure coupling [55]. The yeast nucleoporins nsp1 and nup116 were simulated using three different solvent/forcefield combinations at 300 K. The sequences used in MD simulations are given in **Table 1.1**. As a model system, three simple homopolymer chains of 50 amino acids were also simulated at 300 K: polyglycine, polyalanine, and polyglutamine. All simulations used uncapped sequences without the addition of acetamide or N-methyl groups.

Two groups of MD simulations were used in this analysis. The first group consisted of a larger number of short replicates. The 5 protein chains described were each simulated in 3 different solvent environments for a total of 15 protein systems: (1) the Amber ff99SB-ILDN force field [56–58] with the with the Generalized Born surface area implicit solvent model, (2) the Amber ff99SB-ILDN force field [56–58] with the TIP3P explicit solvent water model [59], and (3) the Amber ff03ws force field [60] with the TIP4P/2005 explicit solvent model [61]. The 15 protein systems were each run in 20 replicates for a total of 300 production MD simulations. Each replicate included a 50 ns equilibration followed by a 100 ns production run for approximately 45 μs of simulation time. The second group of MD simulations used the Amber ff03ws force field [60] with the TIP4P/2005 explicit solvent model [61] for a 1,000 ns MD simulation of each protein. The longer simulations only consist of a single replicate.

Initial starting configuration of the replicates were generated with a custom parallelized R script [62–65] called random chains in R (ranchar) which was optimized to generate large numbers of randomized protein structures. Pulchra version 3.04 [66] converted the initial randomized polymer chains to optimized all-atom protein structures. In order to estimate the necessary simulation box size, the ranchar script was used to generate 100,000 randomized, self-avoiding entropic chains for each protein sequence. The maximum distance between residues was calculated for all generated structures to estimate the periodic boundary conditions large enough to fit more than 95% of the structures. Explicit solvent replicates contained approximately 40,000 to 60,000 atoms depending on the polymer.

All replicates were run through a similar set of steps involving energy minimization; MD simulation with position restraints and a short thermalization up to 300 K; several short NVT MD simulations with 1 fs, 2 fs, and 5 fs time steps to slowly increment the time steps; a short NPT simulation with Berendsen pressure coupling [67]; and a 50 ns equilibration simulation using Parinello-Rahman pressure coupling [55]. Explicit solvent production runs used several optimization to accelerate simulation times. Explicit solvent production runs of 100 ns and 1,000 ns used virtual sites, heavy hydrogens, bond constraints, and 5 fs time steps. MD simulations using implicit solvent were run in Gromacs version 4.5.5 with 2 fs time steps; MD simulations using explicit solvent were run in Gromacs version 2016.3 [33–38].

1.2.3 Figure Abbreviations

Figures use abbreviations for the names of ISDMs: backbone angles (ang), backbone dihedral angles (dih), combined backbone angles and dihedrals (angdih), $\phi - \psi$ angles (phipsi), RMSD (rmsd), RMSD of backbone dihedrals (rmsdih), grid search rotations RMSD (grot), randomly rotated RMSD (rrot), backbone angle correlation (acor), position correlation (pcor), radius of gyration (rg), distance root mean square deviation (drms), end-to-end distance (e2e), elastic shape analysis (esa), and MAMMOTH (mmth or mammoth). The ISDMs grot, rg, and e2e were implemented as intentionally poor measures of ISD.

1.2.4 Further Details

Details of the individual ISDM implementations are discussed in the **Implementation Details For ISDMs** section.

1.3 Results And Discussion

1.3.1 Distinguishing Polymers By Quantified Disorder

We would like to be able to use our measures of ISD to quantify the conformational sampling of systems and differentiate disordered proteins from flexible or stable ones. Therefore, the ability to distinguish systems with varying levels of disorder is one of their most important performance indicators.

Figure 1.3 presents the results of 13 ISDMs which were applied to 15 simulated systems of 20 replicates each. The 15 sets consist of 5 polymers each simulated in 3 different solvent models. For each ISDM, the mean ISD was calculated for all 300 trajectories. For the 15 sets of MD simulations, there are a total of 105 possible paired comparisons. We quantified the ability of our implemented ISDMs to distinguish systems based on their disorder. Since the mean ISDs cannot be reliably predicted to follow normal distributions, the Wilcoxon Rank Sum statistical test was applied rather than the more ubiquitous Student's t-test. The Wilcoxon Rank Sum test [68, 69] was applied to compare each of the 105 possible pairs of systems. The test sample sizes were 20 replicates for each system. Using $\alpha = 0.05$ and a Bonferroni correction of $n = 105$, results were compiled if at least one ISDM option returned a p-value of $p < 0.05/105$. **Figure 1.3** includes 102 out of the possible 105 comparisons.

The detection rate for each ISDM is defined as the number of p-values where $p < (0.05/105)$ divided by the 102 compiled results. Note that there are no false positives, as all tests compare different sets of simulations. The ISDMs based on internal coordinates (ang, dih, angdih, phipsi, and rmsdih) generally performed well at distinguishing different systems. The ISDMs that incorporated information about the backbone dihedrals were the optimal choice for this test. As expected, the intentionally poor measures of ISD–rg and e2e–performed poorly. The relatively poor performance of the MAMMOTH algorithm was unexpected; however, MAMMOTH is optimized to compare molecules with different sequences to look for genetic relationships [39]. Time steps of a single IDP will have zero sequence variance but high tertiary structure flexibility. The study of these dynamics is perhaps not a good usage for MAMMOTH. RMSD performed as poorly at this test as the rg ISDM. Note that the rg ISDM only uses information about the differences in R_g of the compared conformations.

The results of acor and pcor ISDMs consistently did not stand out among the other ISDMs as particularly poor or impressive in both this and the following analyses. Results for these options have been removed from the other figures to focus on the most interesting ISDMs.

Due to the rg ISDM performing nearly as well as several other options, we attempt to explore the relationship between quantified ISD and the size of compared structures as

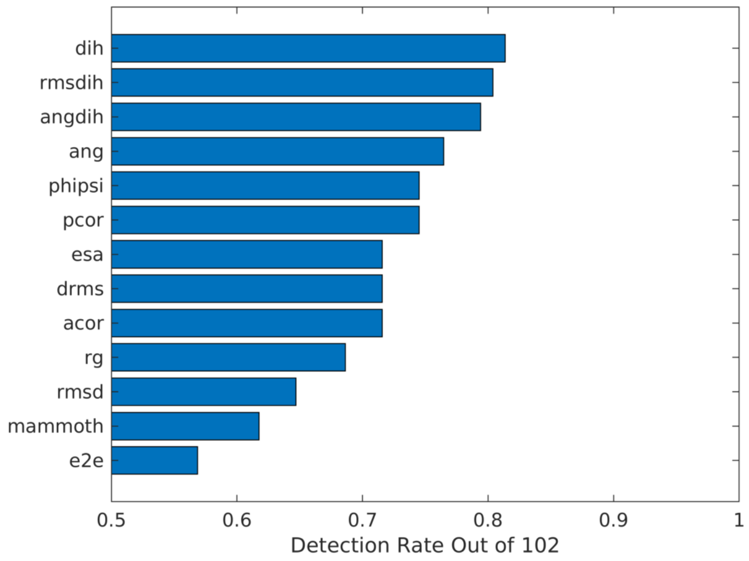


FIGURE 1.3: Wilcoxon Rank Sum tests were applied using the ranksum implementation in Matlab 2017. Each pair of 15 systems was tested for a total of 105 unique comparisons. Tests used sample sizes of 20 replicates for each system. Since it is known that each system is unique, there are no false positives. Results were compiled for the 102 comparisons where at least one ISDM made a positive detection. In the figure, detection rate is defined as the number of p-values where $p < (0.05/105)$ divided by the 102 compiled tests. ISDMs incorporating information about the backbone dihedrals made by C_α atoms–dih, rmsdih, and angdih–out-performed other ISDMs at this test.

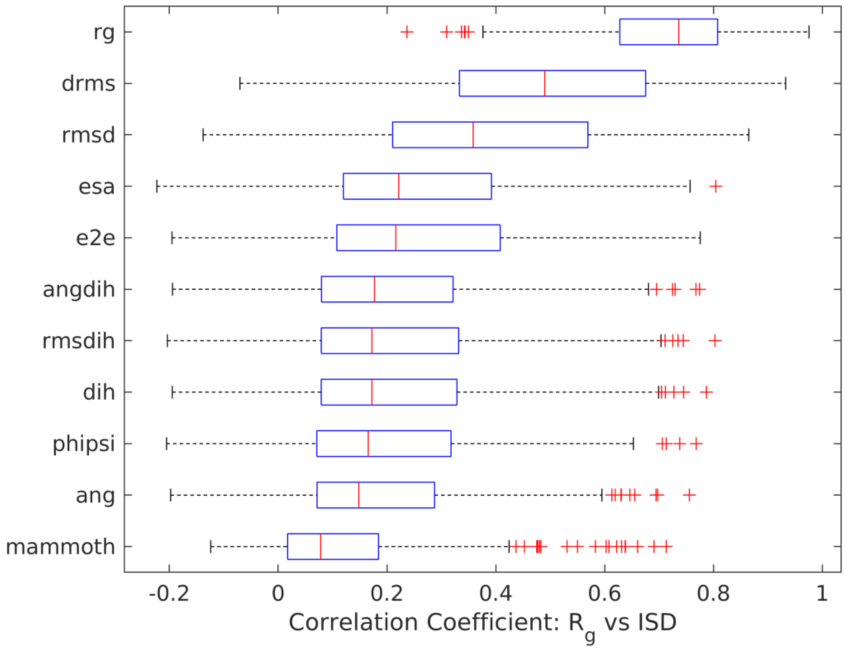


FIGURE 1.4: Size-dependence of ISDMs is displayed as the correlation coefficient between the R_g and ISD. The correlation coefficient was calculated for each simulation using the ISD matrix and a R_g matrix containing the greater of the two R_g for each pair of structures. Since there were 300 separate simulations, the results are displayed as a boxplot. The large median for the rg ISDM is not unexpected; however, the ISDMs drms and rmsd are also extremely size-dependent. MAMMOTH and ISDMs based on internal coordinates are mostly size-independent.

represented by the R_g. **Figure 1.4** is based on the correlation between computed ISD and the size of the larger structure being compared. Note that R_g here does not even take the size differences into account. Correlation coefficients were calculated for each replicate, so each boxplot contains the result of 300 correlations. The median correlation of over 0.7 is expected for the rg ISDM since it is entirely based on molecule size. However, the ISDMs drms and rmsd also give information that largely overlaps with simple size comparisons. The internal-coordinate-based ISDMs and MAMMOTH provide the most unique ISD information.

1.3.2 Measurement Saturation

One of the primary motivations for implementing a library of ISDMs was the observation that RMSD, a standard and well-known method of comparing structures, seemed to quickly saturate to a maximum value when applied to systems of IDPs (**Figure 1.1**).

RMSD is sensitive to small changes in structure which is useful when comparing conformations of folded proteins. However, when the reference structure is being compared to multiple dissimilar structures, it lacks the ability to differentiate between them.

Figure 1.5 illustrates the output of a scaled decorrelation (option -sdcr of the *g_isd* tool) for the synthetic polymers polyglutamine, polyalanine, and polyglycine. This simple decorrelation algorithm averages over all pairs of frames separated by time, Δt. The maximum Δt that can be calculated is $t_f/2$ where t_f is the total simulation time. The reasoning behind this is that decorrelations over longer periods of time will completely ignore portions of input data. The sampling rate of the input trajectory, t_d, was 0.5 ns in our analysis. The decorrelation value calculated before scaling is defined as the mean ISD over all pairs of structures separated by time Δt (**Equation 1.1**). Time step N represents the final simulation time, t_f. Note that due to integer math, the value of $\frac{N}{2}$ is replaced with $\frac{N}{2}+1$ for odd numbers of input time steps.

$$ISD_{\Delta t} \quad = \quad \frac{2}{N}\sum_{i}^{N/2} ISD_{t_i,t_j} \qquad t_j = t_i + \Delta t \tag{1.1}$$

As time Δt from the reference frame increases, all the ISDMs tend toward saturation. Since the units and scales of the various ISDMs are not directly comparable, each plot in **Figure 1.5** is divided by the maximum value of the plot to provide a normalized decorrelation value. A set of 1,000 ns MD simulation production runs was performed to provide sufficient time for all ISDMs to fully saturate. The *g_isd* tool incorporates several methods to analyze time decorrelation which reduce the output data by different amounts. In all cases, the tool produces output for no more than half of the time covered by the input trajectory. Therefore, we chose the range of our plots' x-axes to be 400 ns. The final 800 ns of trajectory data was used as the input to scaled decorrelation.

When applied to the most highly disordered and flexible molecules such as polyglycine (**Figure 1.5.c**), many of the ISDMs reach 80% of maximal decorrelation nearly immediately. When measuring less flexible systems such as polyglutamine (**Figure 1.5.a**), most ISDMs do not approach 80% of saturation until nearly 100 ns. The ISDMs rg and e2e become unreliable and noisy over a short period of time when applied to all three polymers. ISDMs based on internal coordinates approach saturation but retain a controlled slope such that on average, structures that are more distant in time can be differentiated from closer structures. These types of ISDMs appear to be preferable when studying long-term IDP dynamics. Note that for the highly disordered polyglycine, all ISDMs appear to achieve full saturation by 350 ns, but rmsd and drms retain the most stable slope for the longest period of time. When applied to polyalanine and polyglutamine, these ISDMs gave much less desirable results.

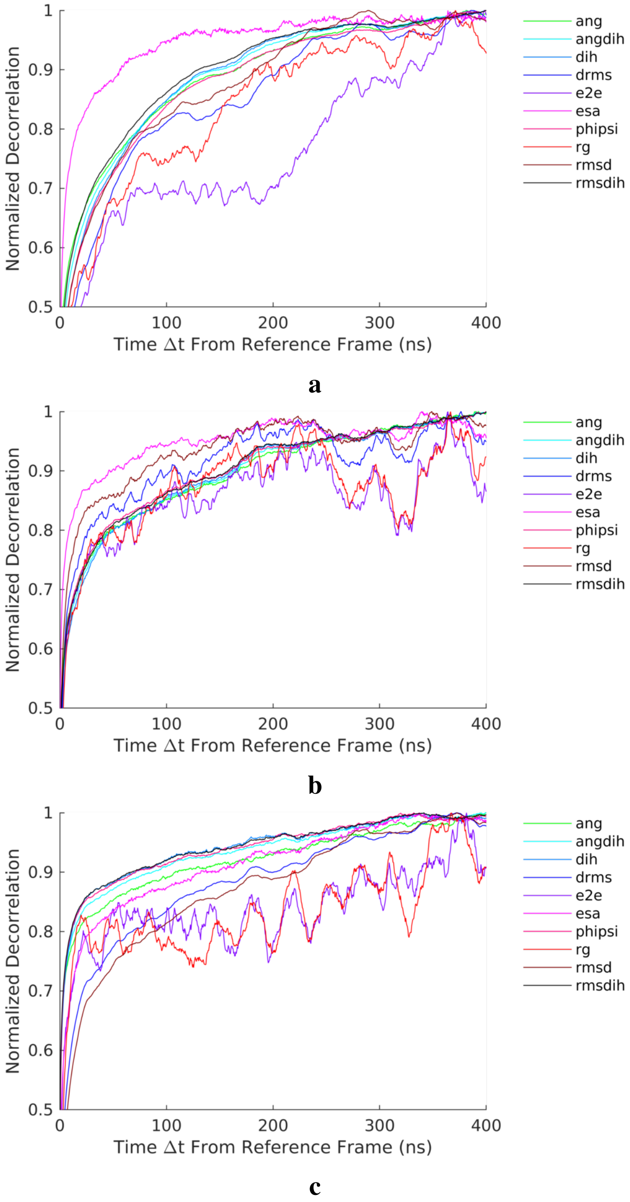


FIGURE 1.5: Simple decorrelation algorithm shows a tendency toward saturation of ISDMs as time, Δt from the reference frame increases. All ISDMs except the mirrored RMSD show a similar saturation, and the general trend is similar across multiple systems: (A) polyglutamine, (B) polyalanine, and (C) polyglycine. ISDMs saturate more slowly when measuring less disordered systems (polyglutamine) than when measuring highly disordered systems (polyglycine).

We are often interested in using a single structure as a reference frame and attempting to distinguish which of two other structures is more distant. We define this property of the ISDM as sensitivity, S (**Equation 1.2**). The *g_isd* tool runs this analysis with the -sens option. Note that the N_S in **Equation 1.2** is an arbitrary number of frames used to average the sensitivity. This can be set manually by the user with the -nsensitivity option. Higher values of N_S lead to more accurate and smoother results; however, the amount of output data from the sensitivity calculation is further reduced from that of the decorrelation output by $t_a = N_S \times t_d$ where t_a is the amount of simulation time used for averaging.

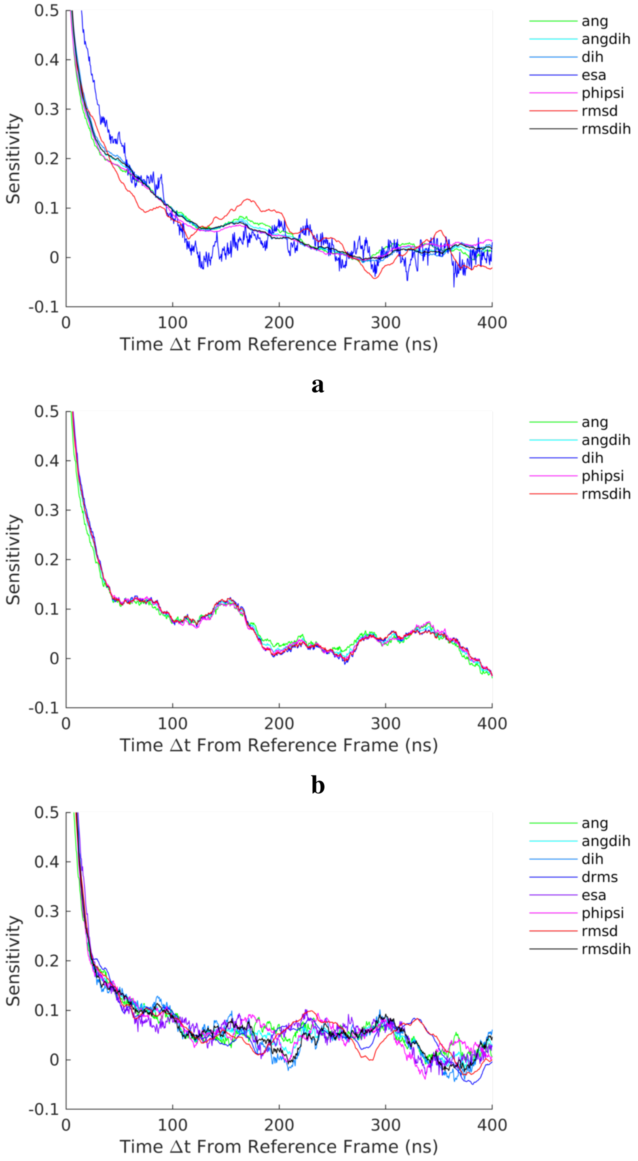
$$S_{t_i,t_j} = \frac{1}{N_S} \sum_{n_{jk}=1}^{N_S} (ISD_{t_i,t_k} - ISD_{t_i,t_j}) \qquad t_k = t_j + n_{jk} t_d \tag{1.2}$$

Several poorly behaved ISDMs were removed from the data presented in **Figure 1.6**. MAMMOTH outputs discrete probability values which did not lend itself to this analysis. However, the ISDM options that gave particularly noisy output are presented as a supplement in **Figure A.1**. In particular, sensitivity equal to zero and slightly negative is expected as the ISDMs approach saturation. However, highly negative sensitivity implied that the ISDM is giving unreliable results which see more similarity in distant structures. Note that ISDMs rmsd, drms, and esa performed poorly applied to some but not all simulations. All ISDMs based on internal polymer coordinates gave nearly identical results for this analysis.

1.3.3 Improving ISDMs With Rescaling

We were interested in whether size-dependent ISDMs could be improved with rescaling to a size-independent measure. The option to scale the output of these ISDMs is -scaled in the *g_isd* tool. Details of the rescaling implementations are discussed in **Section 6.1**. Of note, the scaled RMSD option uses a brute force approach that gives a good approximation of maximum possible RMSD for protein structures of a particular size. The rmsd and drms ISDMs show significant improvement at distinguishing structures based on disorder (**Figure 1.7**). The rmsd option with scaling enabled was nearly as effective at this as using the ISDMs based on internal coordinates. This may be an attractive option when one is looking at either portions of a protein or a non-protein system: any simulation where a polymer is not the subject.

Similarly, the scaled implementations of the rmsd and drms ISDMs are much less size-dependent. The scaled RMSD correlates with protein size approximately the same amount as internal coordinates (**Figure 1.8**). The scaling implementation is not as effective when applied to the drms option.



a

b

c

FIGURE 1.6: The sensitivity, S, of each ISDM to conformational change was calculated as described in **Equation 1.2**. Most ISDMs show a gradual decrease of sensitivity to conformational change as time from the reference frame increases. All three systems used the Amber ff03ws force field optimized for solute solvent energy balance to simulate IDPs. The homopolymers (A) polyglutamine, (B) polyalanine, and (C) polyglycine; were solvated with tip4p/2005 explicit solvent and run at 300 K.

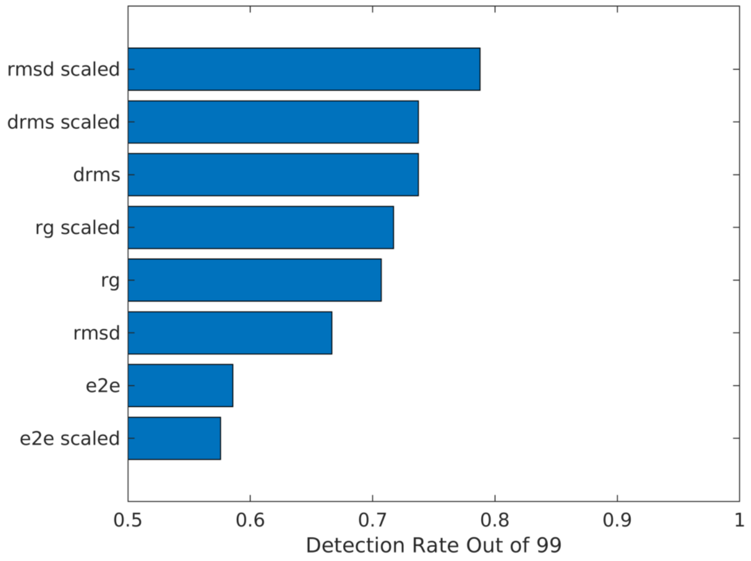


FIGURE 1.7: The ability to distinguish protein systems by quantified disorder is significantly improved among size-dependent ISDMs by rescaling the calculated ISD. The rmsd ISDM with scaling enabled performs nearly as well as ISDMs based on internal coordinates (**Figure 1.3**).

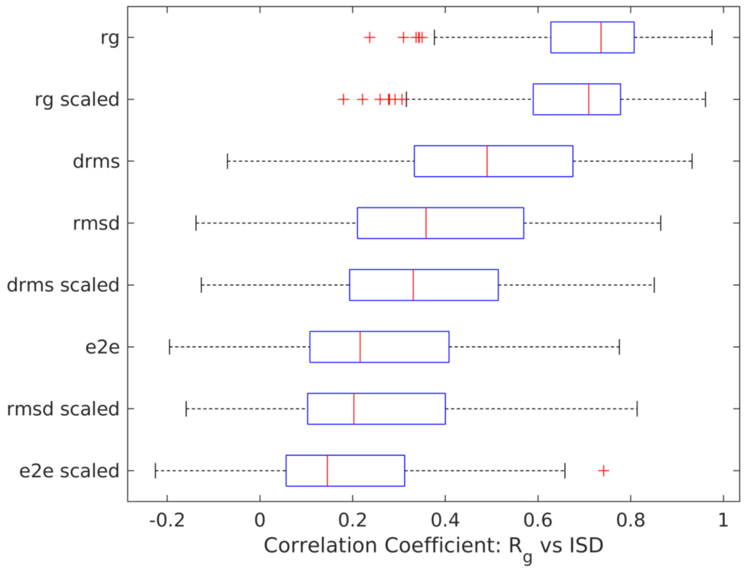


FIGURE 1.8: Several of the highly size-dependent ISDMs can be improved by rescaling. The ISDM option rmsd with scaling enabled is nearly as size-independent as ISDMs based on internal coordinates (**Figure 1.4**)

Figures A.2 and **A.3** show no significant differences in decorrelation or sensitivity based on rescaling.

1.3.4 Harmonic Vs Trig Sums Of Internal Coordinates

We were interested in whether different mathematical methods of summing the differences in internal coordinates may affect the characteristics of the ISDMs based on internal coordinates. These can be compiled as the arithmetic mean of the cosines of angle differences or as the root-mean-square of angle differences. This option can be switched in the *g_isd* tool with the -trig option. In short, **Figures A.4, A.5,** and **A.6** reveal no significant differences in decorrelation, sensitivity, or the ability to distinguish systems.

1.4 Conclusions

The comparison of ISDMs when applied to simulations of disordered proteins reveals that the optimal measure may change depending on the analysis desired or the system studied. ISDMs based on internal coordinates gave good overall performance in our comparisons; however, they are limited to continuous sections of protein backbone. They cannot generally be applied to disconnected amino acids. Several ISDMs saturate too quickly to be applied to highly disordered and flexible proteins. An interesting point is that using backbone angles based only on the C_α appears to give slight advantages in detection of differences between systems compared to the more commonly used $\phi - \psi$ angles representation. This small difference may be specific to the systems that we tested, but it is notable in that there are some systems where the $\phi - \psi$ angles cannot be calculated due to missing atoms. Using $\phi - \psi$ angles to calculate ISD is difficult to use in the analysis of most coarse-grained simulations and in all non-protein polymers. We theorize that perhaps enough information about the overall structure of the protein is retained by coordinates of the C_α atoms. The internal coordinate ISDM options based only on C_α atoms–ang, dih, angdih, and rmsdih–may also ignore some of the extraneous thermal fluctuations in individual bonds.

An important result is the demonstration that, with the exception of several intentionally poor ISDMs, RMSD is less capable of detecting differences between systems with varying levels of disorder than any other method of calculating ISD attempted in this study. Not only does the alignment of molecules becomes less meaningful as the dissimilarity of the two structures increases, RMSD is also highly size dependent. IDPs often exist as an ensemble of conformations including both highly extended conformations and more compact spherical and ellipsoidal structures. The scaled RMSD implemented in these analysis tools attempts to correct for the size dependence and shows significant improvements. Scaled RMSD would be highly useful in looking at segments of proteins which are not directly connected as polymers: for example, flexible binding sites.

While the esa ISDM showed poor overall performance in the aspects we tested by our comparisons and incurs high computational costs, a positive aspect of elastic shape analysis is that it resamples structures as curves in space. This allows it to compare structures that do not need to be homologous and do not necessarily even need to be similar in length.

MAMMOTH was originally designed to score attempts to solve protein folds with little structural information [39]. MAMMOTH would likely be useful in certain targeted and folding MD simulations which begin with highly dissimilar structures or when comparing an ensemble with different sequences.

The library of ISDM implementations has been released along with the source code for the analysis tool *g_isd* which was used to generate these results. The *g_isd* tool also employs this library of ISDMs to implement a variety of analyses: clustering algorithms, improved visualization through dimensionality reduction methods, and the identification local disorder and flexibility within folded proteins.

1.5 Acknowledgements

Dr. Joshua Phillips contributed advice on modifying Gromacs tools and provided sample source code. Implementation of the elastic shape analysis tool was completed by Robert Wang. The particular sequences of simulated FG-nucleoporins presented were chosen with input from Dr. Michael Rexach and Dr. Ed Lau. Ideas for implementation and analysis of the measures of interstructure distance came in part from Dr. Michael Colvin, Dr. Ajay Gopinathan, and Dr. Shawn Newsam.

Chapter 2

Identifying Local Disorder And Flexibility

2.1 Background

2.1.1 The Disorder/Order Spectrum

Proteins tend to exist along a continuum of order to disorder [32], and examples of functional unfolded proteins are widely acknowledged [70–72]. Intrinsically disordered proteins (IDPs) are involved in the regulation of cellular signaling networks and often show a propensity for binding-induced folding [72, 73]. The transition from a free disordered state to a folded state induced by binding has been studied through kinetics [74], real-time NMR spectroscopy [75], and folding simulations [76]. The repercussion of this duality of states between unfolded proteins and their transient folded state is that many IDPs may have local regions of secondary structure and order in the same way that many folded proteins have local regions of flexibility and disorder [77].

Protein flexibility is in fact advantageous when adaptability is required based on the target of binding. The continuum of folded proteins, partially disordered proteins (PDPs), and IDPs may fit into the relationship between structure and function [78]. DNA-binding proteins may first bind non-specifically to DNA before interacting with a specific target sequence [79] and molecular plasticity has been shown to aid in target recognition [80]. Disorder also tends to be prevalent in stress and shock response protein molecules which are most important in conditions where folded proteins may experience denaturing events [77, 81].

A simple, intuitive measure of order for a protein structure was proposed by Fisher and Stultz in 2011 which they refer to as the order parameter [32]. The order parameter defined by this algorithm is based on a scaled mean of the inter-structure distance (ISD) as calculated by root-mean-square deviation (RMSD) after fitting. This approach has the benefit of incorporating structural data into the prediction; however, the original algorithm does not produce results for individual amino acids as many sequence-based disorder predictors do [82–84]. This provides a novel way to analyze local flexibility and disorder in proteins rather than relying solely on root-mean-square-fluctuation calculations.

We attempt to quantify the continuous fold space of proteins [31] through the application of a library of measures of interstructure distance (ISDMs) optimized to study the conformations of IDPs in molecular dynamics (MD) simulations. We present the application *g_isd* which relies on the Gromacs software package [33–38].

The algorithm used by *g_isd* to calculate a universal quantity of disorder and detect local regions of flexibility and disorder is illustrated in **Figure 2.1**. The ISD is computed by comparing every pair of structures from the input ensemble or MD trajectory. Separate calculations are made for each individual amino acid using an ISDM which is independent of protein size and sequence length. The ISD is averaged and rescaled to a value between 0 and 1: these values correspond to an average folded protein and an entropic chain respectively.

We present a method to calculate a universal representation of disorder which allows the user to compare regions in different proteins on the same scale without dependence on the protein size or sequence. The order parameter will allow comparison along the entire continuum of order from stable proteins to flexible regions in folded proteins to PDP and IDPs.

2.1.2 Studied Protein Systems

We apply this approach, using the analysis tool *g_isd*, to both folded and disordered proteins. We observe local flexibility in cystine knots and one of the folded binding domains of the p53 tumor suppressor protein. Conversely, we observe local regions of relative stability in IDPs using MD simulations of nucleoporins and experimental ensembles of β-synuclein.

The cystine knot is a small structural motif defined by the presence of three stabilizing disulfide bonds. A family of plant cystine knot proteins, called cyclotides or cyclins, contain a disulfide bond connecting their ends which makes them cyclic. This property allows them to be one of the smallest known folded proteins [85]. Cystine knots provide a particularly difficult case for sequence-based disorder prediction tools since much of the stability of the folded protein is provided by disulfide bridges. With only 34 amino acids, the MCOTI-II protein lacks the large hydrophobic core necessary for most ordered proteins to fold. The trypsin inhibitor protein MCOTI-II is known from experimental evidence to be folded with a flexible loop 1 region [86].

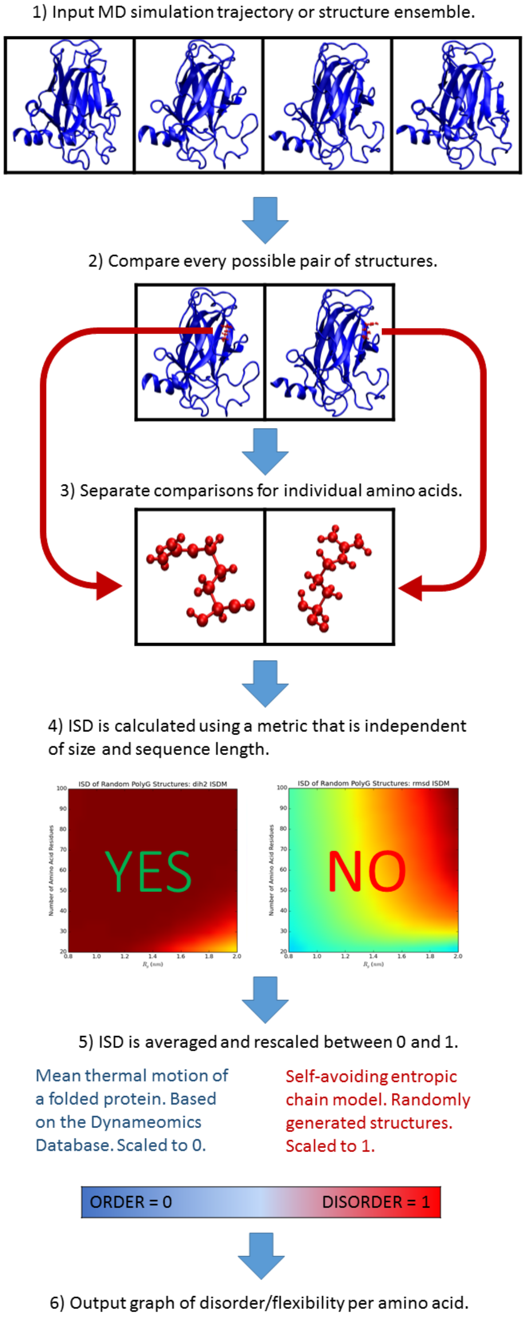

1) Input MD simulation trajectory or structure ensemble.
2) Compare every possible pair of structures.
3) Separate comparisons for individual amino acids.
4) ISD is calculated using a metric that is independent of size and sequence length.
ISD of Random PolyG Structures: dih2 ISDM
Number of Amino Acid Residues
R_g (nm)
YES
ISD of Random PolyG Structures: rmsd ISDM
Number of Amino Acid Residues
R_g (nm)
NO
5) ISD is averaged and rescaled between 0 and 1.
Mean thermal motion of a folded protein. Based on the Dynameomics Database. Scaled to 0.
Self-avoiding entropic chain model. Randomly generated structures. Scaled to 1.
ORDER = 0
DISORDER = 1
6) Output graph of disorder/flexibility per amino acid.

FIGURE 2.1: Analysis algorithm to calculate a universal disorder quantity using the *g_isd* analysis tool. (1) The command-line application accepts trajectories from MD simulations in Gromacs-compatible formats. (2) The ISD of every pair of structures is calculated using one of the implemented ISDMs to build an ISD matrix. (3) The ISD is calculated independently for individual amino acid residues. (4) Only ISDMs which compute ISDs that are independent of protein size and sequence length are appropriate candidates to calculate a universal measure of disorder. (5) The computed ISD is rescaled to a range from the expected variance of an average folded protein to a random entropic chain.

The DNA-binding domain of the tumor protein p53 is a folded sub-domain of a partially disordered protein. The p53 protein is an important part of cell-cycle regulation and triggers DNA repair [87]. Dysfunction of the p53 protein is heavily associated with a variety of cancers.

FG-nucleoporins (FG-nups) are functional IDPs containing numerous phenylalanine-glycine repeats. FG-nups are natively unfolded proteins [51] present is the nuclear pore complex and involved in the gating mechanism. [52]. FG-nups are an interesting model of IDPs as they do not fold but tend to be more compact than true entropic chains [53].

The IDP β-synuclein is closely related to α-synuclein which is involved in the aggregation of plaques that cause Parkinson's disease [88]. However, β-synuclein is missing 11 residues from the portion of the protein that forms the core of amyloid plaques and is resistant to aggregation [88]. In fact, β-synuclein is known to inhibit α-synuclein aggregation both in vivo and in vitro [89].

In sum, we are attempting to cover a significant proportion of the continuum from order to disorder. The studied protein systems cover highly extended and flexible chains, compact but disordered molten globules, stable proteins with disordered sections, and fully folded proteins with some local flexibility.

2.2 Methods

Analysis tools optimized to study molecular dynamics simulations of IDPs are rare, so we discuss how *g_isd* was optimized to cover the continuum of order to disorder. We focused on improving the results of the Fisher-Stultz algorithm in several ways: (1) produce an overall value of disorder with individual scores for each amino acid, (2) easily plot the output in a format which is familiar to anyone who has used sequence-based disorder prediction tools, and (3) implement measures of ISD based on both internal and external protein coordinates. Note that not all ISDMs implemented in *g_isd*'s library are capable of calculated ISD for individual residues. We present only the subset of ISDMs that fulfill this condition.

2.2.1 Disorder Parameter Optimization

Optimization of the order parameter scaling constants comprises two primary goals: (1) an estimate of the maximum disorder of a random entropic chain and (2) an estimate of the structural variance expected in an average folded protein. Estimating these properties for a selection of ISDMs allows linear rescaling of the mean ISD to a globally representative estimate of disorder.

The upper theoretical limit of structural variance was calculated over a representative phase space by generating random polymer chain structures grouped by protein size (represented by the radius of gyration, R_g) and polymer sequence length. The random polymer chains were produced with a custom script written for the R environment [62] called random chains in R (RanchaR). Acceleration via parallelization was accomplished with assistance from external packages [63–65].

Polyglycine homopolymer chains with amino acid sequence lengths ranging from 20 to 400 were generated following a self-avoiding random-walk model (SARW). The generated chains were filtered and sorted by R_g into specific groups from 0.8 to 4.8 nm. Only chains within 10% maximum error of one of the target R_g values were kept. In total, several million chains needed to be generated to produce the 149 ensembles of 1,000 structures each representing the desired phase space of protein size and sequence length. Pulchra version 3.04 [66] was applied to convert the initial randomized polymer chains to optimized all-atom protein structures.

For the order parameter calculated by *g_isd* to be a universal representation of disorder, it must be independent of protein size and sequence length. To determine whether ISDM options fulfilled these requirements, plots of the computed ISD of phase space were created using the matplotlib package [90] from python 3. Since the chain generation algorithm allowed a range of R_g values for each ensemble, the target R_g was not used in the phase space plots. Instead, the mean R_g of the structures in the ensemble was used which is guaranteed to be close to the target value.

Structural variance explained by thermal noise in an average folded proteins was estimated using MD simulations sampled from the Dynameomics database [16]. The Dynameomics database uses a SQL format of properties rather than a set of raw trajectories. Therefore, individual properties necessary to create a trajectory were downloaded to text files representing 100 proteins from the database. These were converted to Gromacs-compatible gro format files using a python script, and the Gromacs tools trjcat and trjconv were use to compile the individual time step frames into MD trajectories.

Estimates of order and maximum disorder were carried out using the *g_isd* tool. ISD was averaged over all amino acid residues over all pairs of structures in each ensemble and trajectory described. The results were compiled with the assistance of GNU parallel [91].

2.2.2 MD Simulations

To validate our results and illustrate applications, we have completed MD simulations of several systems.

MD simulations were run on the trypsin inhibitor MCOTI-II using the wild type structure from the protein database (pdb) entry 1IB9 [86] and a two disulfide intermediate structure from pdb entry 2P08 [92] for 500 ns and contained approximately 25,000 atoms. The MD simulation of p53 uses the structure from pdb entry 2OCJ [87] and has a production run of 200 ns. The all atom simulations were carried out in explicit solvent using the tip3p water model [59] and were run with the Amber99SB-ildn force field [56–58] with 2 fs time steps in Gromacs 4.6.2 [33–38]. Production runs used the velocity rescaling thermostat [54] at 310 K and Parinello-Rahman pressure coupling [55].

All-atom MD simulations were run on a protein fragment of 50 amino acids in length using the velocity rescaling thermostat [54] and Parinello-Rahman pressure coupling [55]. The yeast FG-nucleoporin nup116 was simulated with the TIP4P/2005 explicit solvent model [61] using the Amber ff03ws force field [60] with 5 fs time steps for 1,000 ns production runs. To achieve stability with 5 fs time steps, the virtual sites and heavy hydrogens optimizations were used. The nup sequence of the MD simulations is given in **Table 1.1** and uses an uncapped sequences without the addition of acetamide or N-methyl groups. The initial configurations of coordinates were generated with RanchaR, and the system contained approximately 50,000 atoms.

2.2.3 Ensemble Of Experimental IDPs

An ensemble of conformations of the β-synuclein IDP derived from experimental NMR data [88] was downloaded from the Protein Ensemble Database [15].

2.2.4 Figure Abbreviations

Figures use abbreviations for the names of ISDMs: backbone angles (ang), backbone dihedral angles (dih), combined backbone angles and dihedrals (angdih), $\phi - \psi$ angles (phipsi), RMSD (rmsd), RMSD of backbone dihedrals (rmsdih), and the distance root mean square deviation (drms).

2.2.5 Further Details

Details of the individual ISDM implementations are discussed in the **Implementation Details For ISDMs** section.

2.3 Results And Discussion

2.3.1 Correcting For Size and Sequence Length Dependence

ISDMs based on internal protein coordinates are size-independent and sequence length-independent by default. The phipsi option has essentially no variance across the tested phase space in **Figure 2.2.a**. In these plots, single color in the output reveals the same calculated ISDs at all points along the phase space. The plotted area was heavily sampled with ensembles of SARW entropic chains and then interpolated with the function *griddata* from the scipy.interpolation package for python 3 [93]. Areas plotted in white are outside of the interpolated area and are not part of the sampled phase space. These areas are either highly extended or compact forms and are unlikely to represent the natural conformations of a truly entropic chain.

These ISDMs are good candidates to calculate a universal quantity of disorder. Furthermore, since the calculated ISDs have a flat response across the entire phase space, the ISD of fully disordered entropic chains can be estimated by combining the results across all 149 ensembles.

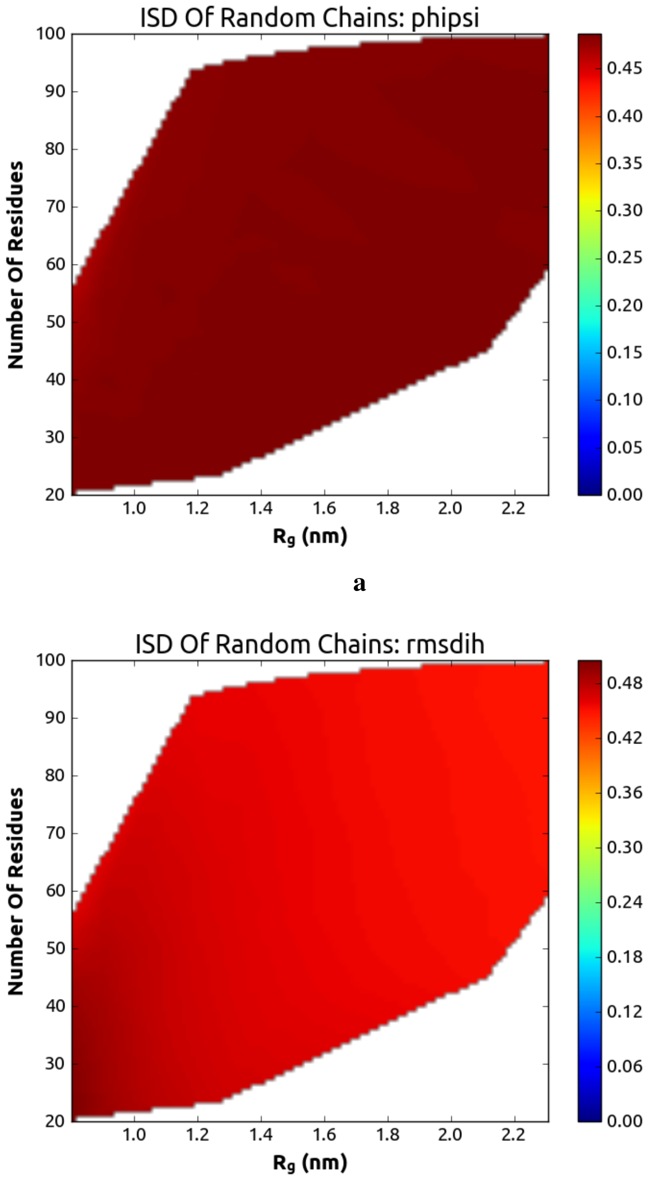
Conversely, the ISDM options rmsd (**Figure 2.3.a**) and drms (**Figure A.7.a**) are heavily dependent on protein size. The calculated ISDs increase continuously as the R_g increases. A method to scale rmsd and drms to make them less size-dependent, described in **Section 6.1**, was applied (**Figure 2.3.b** and **Figure A.7.b**). This improves the result in two ways. A larger portion of the phase space has a flat ISD, and the calculated ISDs no longer trend purely with size. Proteins with a longer sequence will also tend to take up more space, and the scaled rmsd and drms have a relatively flat trend diagonally on the phase space plot.

In general, these ISDMs are unreliable when comparing highly extended proteins (bottom right of the phase space) with compact ones (top left). Even with the scaling correction, the rmsd and drms options are not good candidates to calculate a universal quantity of disorder. However, there are numerous systems where it is impractical to use internal protein coordinates (binding sites, proteins with missing domains, and non-polymers).

The phase space was extended up to 400 amino acid residues and R_g of 4.8 nm with similar results. See supplemental figures in **Section A.2.2**.

2.3.2 A Universal Quantity of Disorder

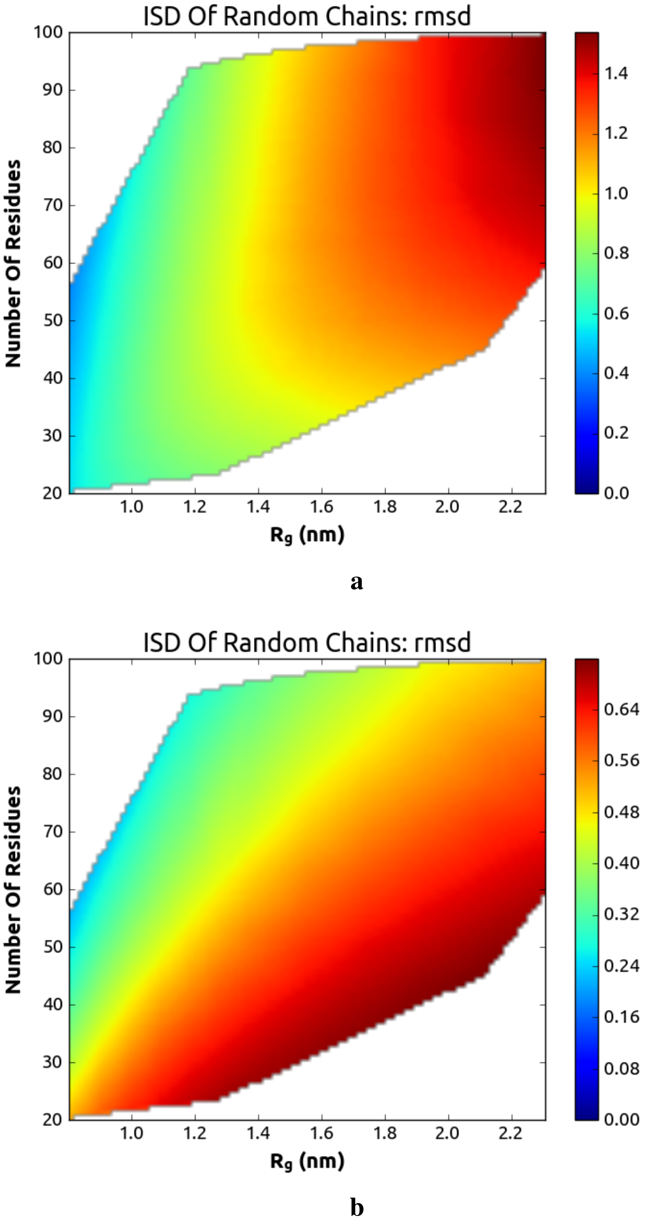
The calculated ISDs for the ensembles of random chains across the entire phase space were compiled to give a single quantity which represents the maximal disorder equivalent to a random entropic change. A similar quantity was calculated for the 100 folded



a

b

FIGURE 2.2: The ISDMs phipsi and rmsdih calculate ISDs which are not dependent on protein size or sequence length. The single output color reveals that the average ISD at each point in the phase space is the same. Areas plotted as white are outside of the interpolated area. This means that no ensemble of random chains was produced with the corresponding combination of sequence length and R_g.



a

b

FIGURE 2.3: (a) The dependence of ISD on size and sequence length is significant for the rmsd ISDM. Since proteins of different sizes cannot be compared directly, rmsd is not a suitable candidate for a universal measure of disorder. (b) Scaling improves the size-independence properties of rmsd significantly; however, conformations with relatively small R_g still record smaller ISD.

TABLE 2.1: For each ISDM, the parameterized values used to calculate a universe quantity of disorder are displayed. The **Order** value corresponds to the expected ISD of an averaged folded protein estimated by compiling results of 100 folded proteins from the dynameomics database. The **Disorder** value is the ISD of a fully disordered and random entropic chain.

ISDM	**Order**	**Disorder**
phipsi	0.11304	0.48496
ang	0.03634	0.17721
angdih	0.04714	0.17721
dih	0.08083	0.47773
rmsdih	0.14304	0.66732
rmsd	0.08091	0.52096
drms	0.04299	0.25439

TABLE 2.2: The effect of the -scaled option on the parameters of the rmsdih ISDM.

	Order	**Disorder**
Scaled	0.14304	0.66732
Unscaled	0.34429	0.46083

protein trajectories from the Dynameomics database. The medians of the compiled values for each ISDM are reported in **Table 2.1**. For ISDMs rmsd, drms, and rmsdih, only the parameters for the scaled versions of the ISDMs are reported since they have significant advantages. When quantifying disorder, the analysis tool *g_isd* linearly rescales the calculated ISD for each amino acid residue using these parameters. This analysis option is chosen -order command. The parameters may also be manually set by the user with the -zero and -one command-line options.

The -scaled command-line option is suggested for the rmsd, drms, and rmsdih ISDMs. This attempts to correct the size-dependent error associated with the rmsd and drms ISDMs. However, the option works differently for rmsdih. As described in **Section 6.1**, rmsdih is implemented to use local alignments of C_α backbone atoms. The result is a measure using internal protein coordinates that is expressed in normal distance units (nm). This is carried out using a shift to spherical coordinates; however, the r variable in the spherical coordinates corresponds to a degree of freedom that is frozen out in polymers due to peptide bonds. Unfortunately, there is still a significant amount of thermal noise present along these degrees of freedom. For the rmsdih ISDM, the -scaled option manually sets $r = 1$ which converts the ISD calculated by rmsdih to a unit-less quantity. The improvement to the rmsdih ISDM is reported in **Table 2.2**. The contrast range between the quantity reported as order versus disorder is increased more than threefold.

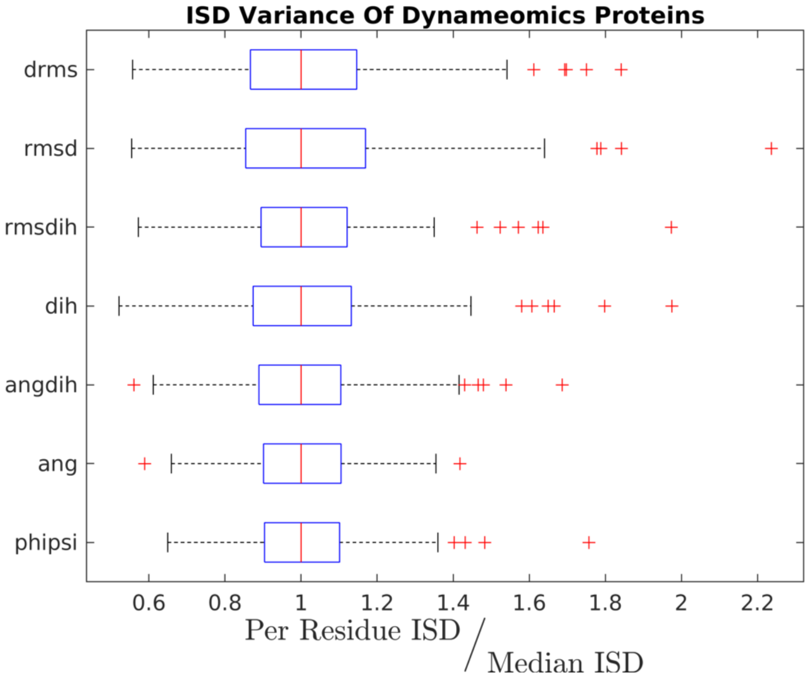


FIGURE 2.4: Comparison of ISD variance of folded proteins from the Dynameomics database. Each box plot has a sample size of 100 MD trajectories of folded proteins and shows several consistent outliers. Therefore, the median is a less biased approximation of the order constant than the mean.

During the estimation of the order and disorder constants, the means and medians of the compiled ISDs were similar. However, calculated ISDs of MD trajectories from the dynameomics database consistently revealed several more disordered outliers (**Figure 2.4**). Therefore, the median was chosen as a less biased estimate.

Figure 2.5 illustrates the compiled ISDs of the 149 ensembles of SARW entropic chains across the entire phase space of protein size and sequence length. The calculated per residue ISDs are divided by the median to display all values on a comparable scale. The box plots only include results for scaled versions of the rmsd, drms, and rmsdih ISDMs. Even after partially correcting for size-dependence, the rmsd and drms ISDM options show significantly more variance than ISDMs based on internal coordinates.

2.3.3 Applications: PDPs

The order parameter is implemented in the *g_isd* analysis tool with the -order option and outputs to the xvg format by default. A simple script, *xvg2bf.bash*, is able to

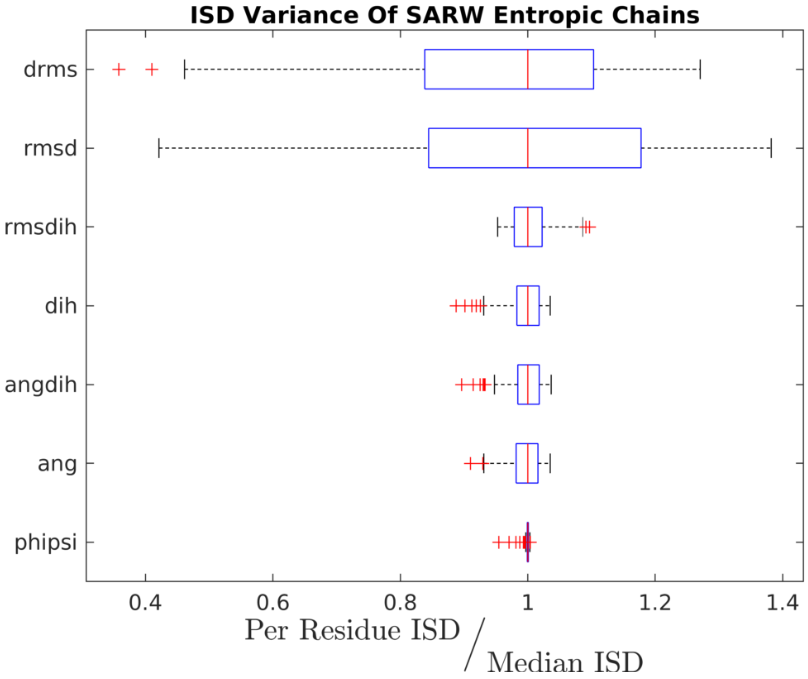


FIGURE 2.5: Comparison of ISD variance measured with ensembles of SARW entropic chains. Each box plot contains statistics for a sample size of 149 ensembles across the sampled phase space of protein size and sequence length. Calculated ISDs were normalized by dividing by the median. The drms and rmsd ISDM options show sizable variance because their size-dependence is only partially corrected by the -scaled option.

convert the calculated order parameter into a simple text file which the Gromacs tool editconf understands. The *editconf* tool is able to create pdb files where the beta factor field has been overwritten with the order parameter calculations. In the following figures, molecules were rendered in VMD with color set according to beta factors [94].

A set of functions for GNU Octave [95] or MATLAB are available to read in the calculated order parameter values and the protein sequence to display.

Proteins heavily reliant on disulfide bridges to maintain structure, such as cyclotides, are a difficult problem for sequence-based disorder prediction software. The sequence information is insufficient to determine the number of disulfide bonds in a protein structure as it is dependent on the protein's tertiary structure. We explore a more accurate disorder prediction approach employing MD simulation and calculate the regional disorder.

In **Figure 2.6**, the analysis of MD simulations of wild-type MCOTI-II (pdb entry 1IB9) protein detects the loop 1 region which has been observed to be disordered. The order parameter is calculated to be 1.0 for the serine and glycine residues 1 and 2 which are automatically colored red. The algorithm considers this terminal region to be as disordered as an entropic chain. The order parameter values calculated for the loop 1 region imply flexibility but not an equivalence to a fully disordered protein. The other sections of wild-type MCOTI-II are calculated to be as stable as an average folded protein.

Figure 2.7 contains the analysis carried out on the MCOTI-II intermediate with only two of three disulfide bonds formed. As is immediately obvious from the coloring of the protein structure, a significant portion of the loop 1 and loop 2 regions of the protein are highly flexible to fully disordered. In addition, significantly more amino acid residues are disordered in the intermediate form than in the wild-type. The beta factor illustration of flexibility facilitates the identification of the disordered regions within the tertiary structure of the protein rather than just along the sequence.

The p53 tumor suppressor protein contains 194 amino acids, so the sequence labels have been left off of **Figure 2.8.b**. The order analysis finds the greatest disorder at the terminal ends of the protein domain. However, these amino acids connect to the disordered linker regions of the full p53 protein, so this is not an unexpected result. The four peaks showing local flexibility are more interesting. From left to right, the four main peaks of flexibility in Figure 2.8.b correspond to the regions near residues 117, 185, 226, and 245. These happen to be near but not directly involved in DNA interaction sites. Since DNA is itself a flexible molecule, flexible regions near the DNA interaction sites allow the protein to either bond non-specifically to DNA to increase the probability of a binding event or reorient the interaction site to increase the chances of coming into direct contact with the binding site.

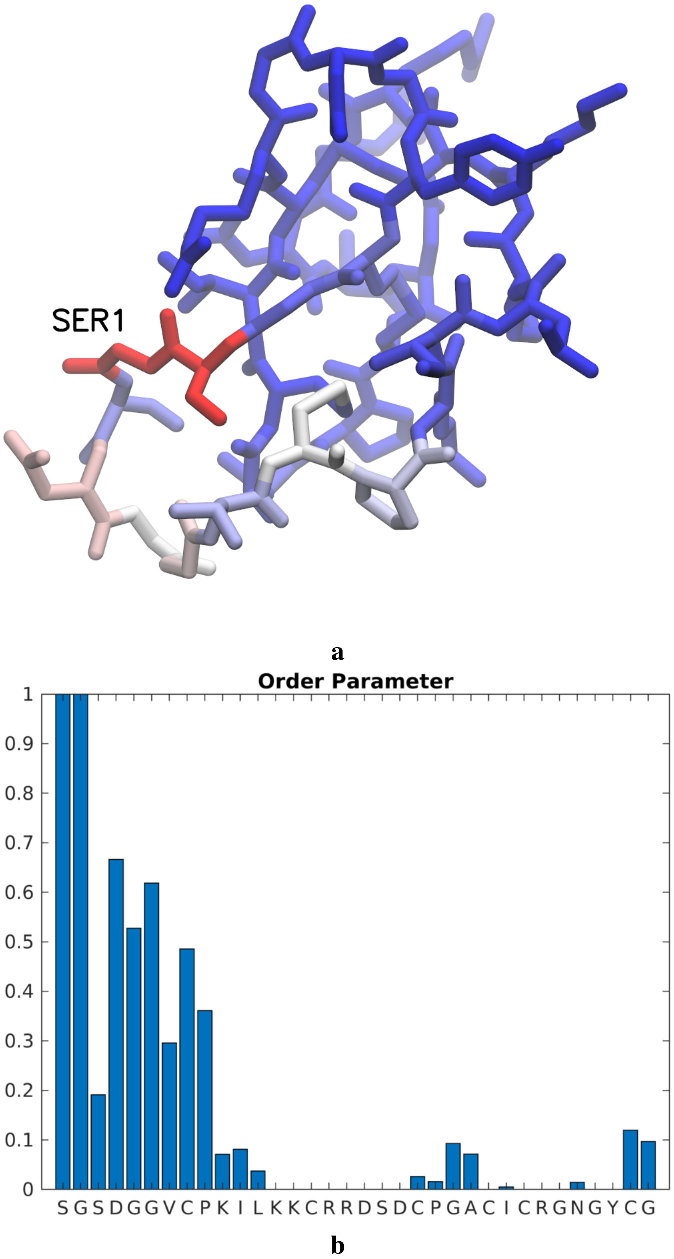
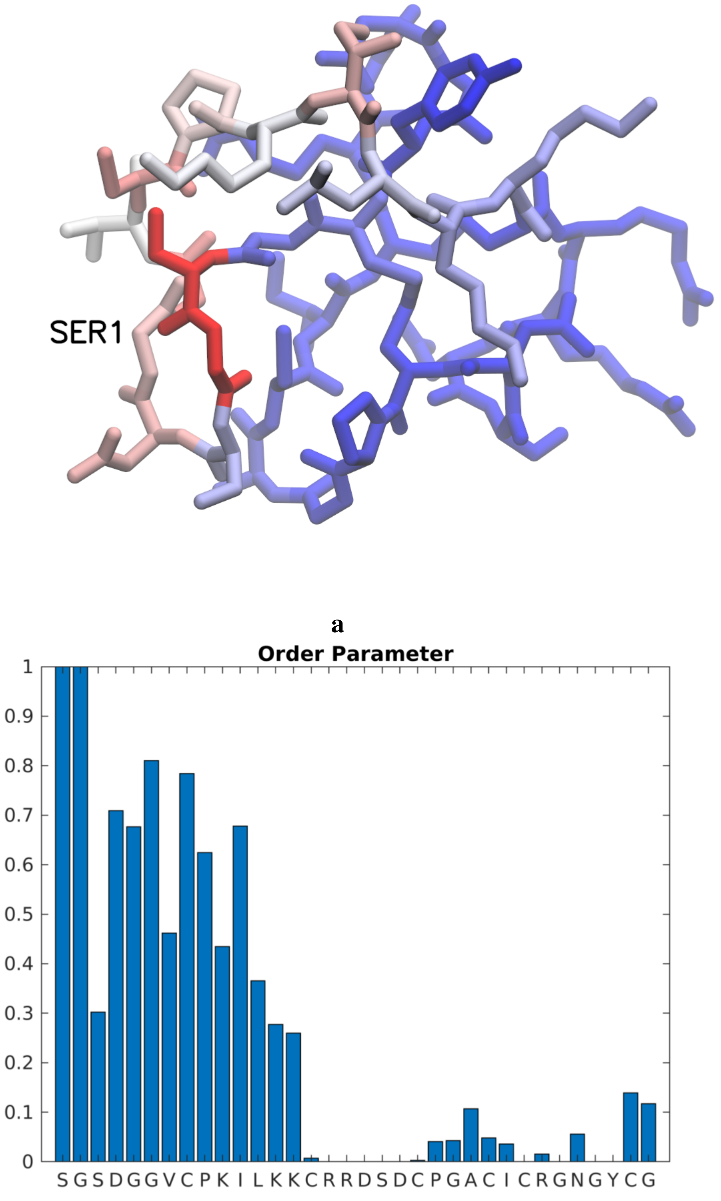


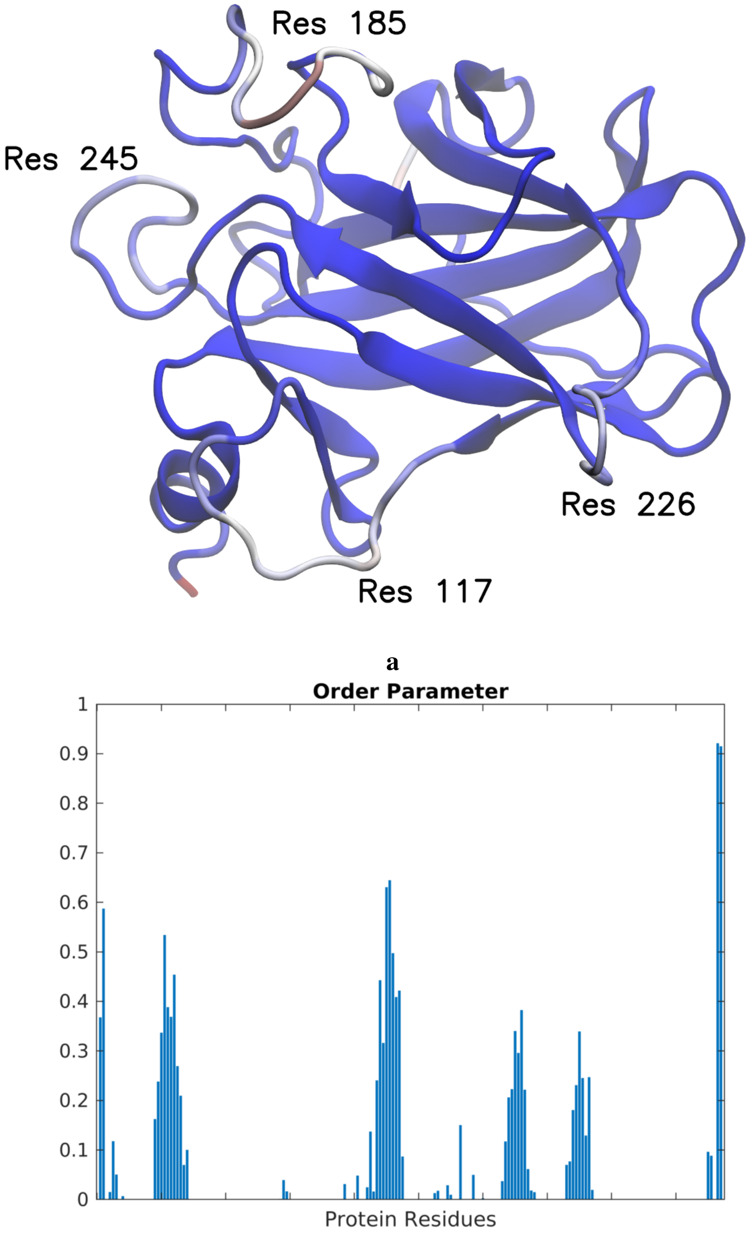
FIGURE 2.6: Folded structure of the cystine knot cyclotide protein, MCOTI-II. The loop 1 region has significant flexibility and is automatically colored red in VMD after applying the order parameter to the beta factor fields of the pdb file. The regions of the protein colored blue have been parameterized as ordered by the analysis of MD simulation trajectory with the *g_isd* tool.



a

b

FIGURE 2.7: The intermediate form of MCOTI-II protein has only two out of three formed disulfide bridges. MD simulations and order analysis reveal disorder in both the loop 1 and loop 2 regions of the protein. The automatic coloring of the protein assists in identifying the locations of disordered regions in the tertiary structure.



a

b

FIGURE 2.8: The order parameter analysis of *g_isd* identifies 4 regionals of local flexibility which are all near DNA interaction sites of this DNA-binding subdomain of the p53 tumor suppressor protein. The four identified peaks of disorder in (b) are near the four labeled regions. From left to right, the peaks are near residues 117, 185, 226, and 245.

2.3.4 Applications: IDPs

Applying order analysis to IDPs serves in part as a reminder of the wide space covered by the order to disorder continuum. Many proteins may exist in a natively unfolded state without sharing the properties of an entropic chain. The FG-nucleoporin nup116 appears to remain in a somewhat compacted state as seen in **Figure 2.9.a**.

This is unlikely to be merely an artifact of the simulation parameters or forcefield. The AMBER03WS force field used to simulate nup116 has specifically been corrected to reduce the "stickiness" of most protein forcefields that induce compactness and transient secondary structure when applied to IDPs. In addition, experiments have verified that nup116 tends to exist in an ensemble of structures that is more compact on average than one would expect from a fully disordered entropic chain.

The phenylalanine-glycine (FG) repeats that exist throughout the protein causes interspersed regions to be hydrophobic within the primarily hydrophilic protein. This can cause local regions to collapse into more compact configurations which leads to constraints on the available degrees of freedom along the backbone. The reduced flexibility of the polymer chain in this constrained state is likely the reason the order parameter is calculated to fall between 0.3 and 0.7 in different regions of the IDP. This order parameter is less than one might expect for an intrinsically disordered protein.

Of particular interest from **Figure 2.9** is the ASSSGAK sequence of residues near the N-terminal region of the protein which appears to have extremely limited disorder.

On the other hand, the IDP β-synuclein shows a preference for a mostly extended conformation. The calculated order parameter is only slightly less than that of an entropic chain. The β-synuclein structures used in this analysis are an ensemble derived from NMR data. The only distinguishing feature from the ensemble of structures is the small region near the C-terminal end which shows significantly less disoder than the rest of the protein. This more ordered sequence of residues is the blue-colored region near the labeled proline amino acid.

2.4 Conclusions

When applied to MD simulations of folded proteins, order analysis applied by the *g_isd* tool enables the identification of local flexibility while functioning as a more accurate method to predict disorder in proteins. The analysis was able to produce interesting information about systems along the entire continuum from order to disorder.

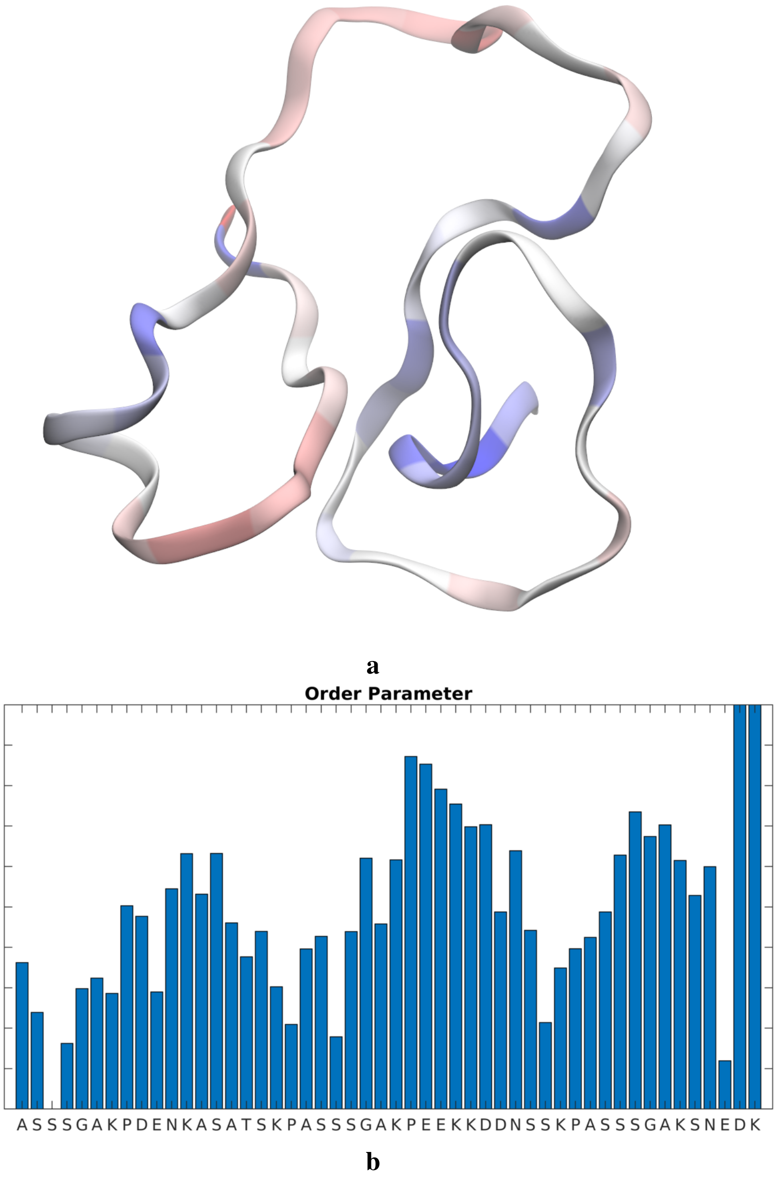


FIGURE 2.9: The FG-nucleoporin nup116 is an intrinsically disordered protein which favors a relatively compacted state (a). The tendency towards compacted conformations induces constraints on the protein backbone which results in a relatively low calculated order parameter (b). While the protein is natively unfolded, it is not as disordered as a true entropic chain.

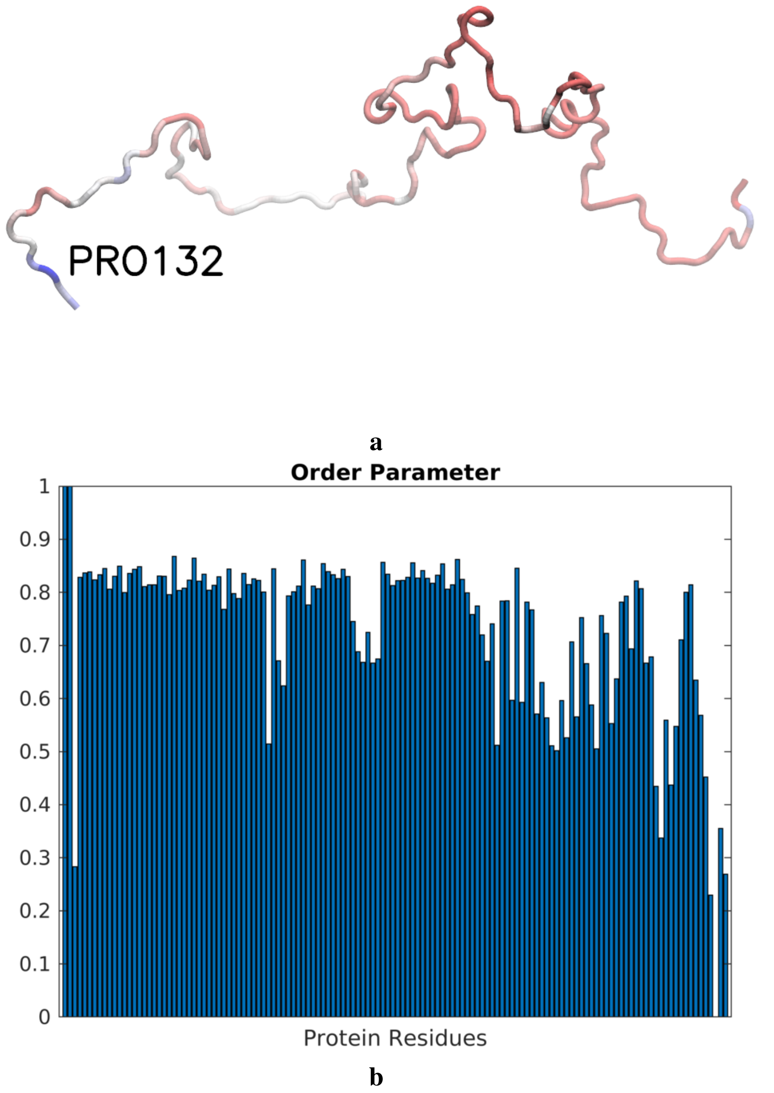


FIGURE 2.10: Ensemble of structures of the β-synuclein IDP. β-synuclein is closely related to α-synuclein which is involved in the formation of amyloid plaques seen in Parkinson's disease. Analysis of the β-synuclein IDP confirms that it is highly disordered. The order parameter values imply behavior that is nearly as disordered as an entropic chain.

The computation of order parameters for MD simulations that have already run takes a trivial amount of time from seconds to minutes. The analysis only requires a sample of hundreds to several thousands of structures to provide a useful estimate.

While the order and disorder parameter constants were estimated for several ISDM candidates, rmsdih with the -scaled correction is efficient and powerful. ISDMs such as rmsd and drms can not be fully corrected for their inherent size-dependence even with scaling which makes them a poor candidate to be used as a universal way to measure disorder. These options may still be useful when applied to systems where internal protein coordinates are not feasible. Looking at a small protein region with unconnected protein fragments such as a binding site may lend itself to analysis with rmsd or drms.

2.5 Acknowledgements

Figures illustrating the plotted phase space of entropic chains were created with Python source code based on the work of Dr. David Ando with support from Dr. Ajay Gopinathan. Structures used to simulate the cystine knot protein were suggested by Dr. Fabian Filipp. Implementation of the shape parameter analysis was completed with support from Dr. Michael Colvin.

Chapter 3

Hierarchical Spectral Clustering For MD Simulations

3.1 Background

As computational power has scaled upwards in scope and downwards in price, producing long molecular dynamics (MD) simulations of proteins on the order of μs to ms time scales has become more common. MD trajectories on this scale produce massive stores of information which present new challenges for analysis. Statistical means have been proposed to elucidate the important structural insights from our data sets without being inundated with high frequency thermal noise [96]. One common approach with a long history [97] has been to employ a wide variety of different clustering techniques [98, 99]. This allows the user to take a data-mining approach to the substantial amounts of data often produced in long MD simulations. Statistical clustering has seen widespread usage with both large-scale protein sequence data [100, 101] and individual protein trajectories [97–99, 102].

The application of spectral clustering methods to intrinsically disordered proteins (IDPs) has also been explored [21, 29]. Spectral clustering is able to reduce protein motions to a lower dimensional data set. Expressed in reduced dimensionality, unimportant degrees of freedom may be more easily differentiated from those central to protein motion.

We present an advanced clustering technique designed specifically to be applied to protein dynamics. Our algorithm utilizes a combination of spectral and hierarchical clustering. Spectral clustering broadly includes any form of clustering that utilizes reduced dimensionality. It is most commonly paired with k-means clustering, but hierarchical clustering is an agglomerative method that continuous combines clusters within a maximum distance. Our estimates of folded protein dynamics to calculate a universl order parameter provide an empirical basis for estimating this maximal cluster distance as well. The initial dimensionality reduction to carry out spectral clustering is implemented into the MD analysis tool *g_isd*. Visualization of results is provided through an external script designed to be used with GNU Octave [95] or MATLAB. The amplitude of structural variance that can be explained by thermal noised has been estimated using folded protein data from the Dynameomics database [16].

Our software implements a classical multidimensional scaling (CMDS) algorithm to reduce the dimensionality of protein motions within their fold space to improve the results of visualization and clustering. We have optimized our implementation of CMDS to use a metric which gives more robust and accurate performance than the more common RMSD. The dimensionality estimate computed by *g_isd* attempts to alleviate an issue with the analysis of both folded and disordered proteins. The variance observed in each essential dimension is comprised of both meaningful structural change and low amplitude thermal noise. Along the disorder to order continuum [32], the motions of highly ordered proteins see a proportionally increased contribution from thermal noise since few of their motions result in meaningful conformation change. Without optimizations to control for the magnitude of structural variance expected within folded proteins, dimensionality estimators detect thermal noise without structural change as important degrees of freedom for folded protein dynamics.

3.1.1 Homopolymer Polypeptides

We applied hierarchical spectral clustering (HSC) to MD simulations and experimentally derived ensembles along a known spectrum–from fully flexible with transient structure to highly disordered–to validate our results. Three systems of homopolymers were chosen for MD simulations as simple protein models: polyglutamine, polyglycine, and polyalanine.

Polyglutamine is disordered, but MD simulations reveal a tendency to form transient α-helix [50] and β-sheet [48, 49] secondary structure. MD simulations of polyalanine tend to produce small portions of temporary α-helical structure which has also been observed in experiments [44, 45]. Longer polyalanine chains, similar to the size of the homopolymers in our simulations, tended to have a more complex helix-turn-helix ensemble [44]. Polyglycine shows a conformational preference for extended conformations in experiment [41] and aggregation into an amyloid-like highly disordered material has been observed [42]. In MD simulations, polyglycine is compact yet highly flexible [43].

3.1.2 Optimized Protein Clustering

HSC, implemented through *g_isd*, was used to study protein systems from MD simulations and experimental ensembles of structures.

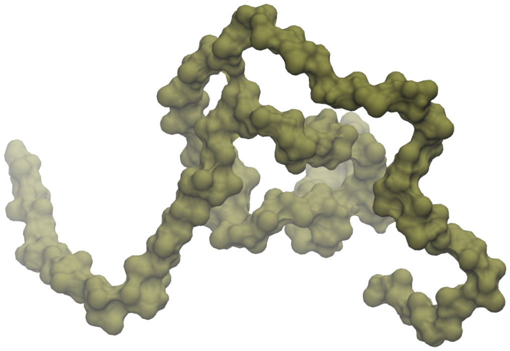


FIGURE 3.1: Tau protein structure derived from experimental NMR data. The protein was rendered in Visual Molecular Dynamics [94].

We have explored an experimentally derived ensemble of structures for the K18 domain of the Tau protein [103]. Tau plays a significant role in neurodegenerative diseases such as Alzheimer's and Parkinson's. While generally existing in a disordered state, Tau sometimes misfolds and self-aggregates into tangles of filaments [104]. Tau normally associates with tubulin to stabilize microtubules, but in its dysfunctional diseased state, it causes devastation to neuron cells.

FG-nucleopons (FG-nups) are IDPs associated with the nuclear core complex and functional in the gating machinism of the cellular nucleus [52]. They are rich in phenylalanine-glycine repeats and have a regional variance in sequence charge density. FG-nups show regional variations in their compactness: some areas express transient structure while others show true entropic chain behavior [53].

A family of plant proteins based on the cystine knot motif contain three disulfide bonds and a cyclic end-to-end attachment [85, 105]. The additional structural stability from these bonds make them one of the smallest folded proteins. These cystine knot proteins are considered interesting candidates as drug delivery scaffolds [106]. They contain a disordered loop 1 region which can generally be modified without affecting structure, so the loop can be replace with a targeted binding sequence.

3.2 Clustering Implementation

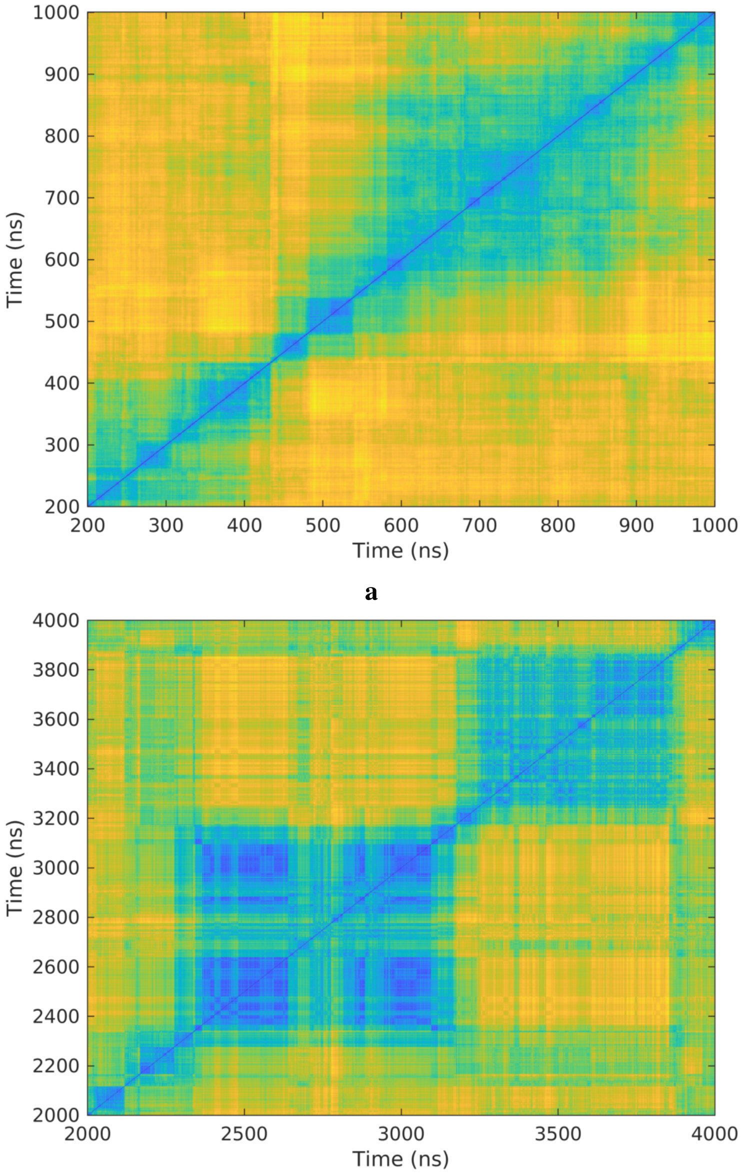
A simple method of clustering is to visualize the distances between structures as in **Figure 3.2**. The trajectory travels in time from the bottom left to the top right. Blocks of low interstructure distance (ISD) as seen in **Figure 3.2.b** are likely transient metastable states. This partially disordered cystine knot protein is likely shifting between several energetically favorable conformations with transition states in between. The fully disordered protein in **Figure 3.2.a**, nsp1, visits more states for shorter periods of time. This can be seen by the smaller size of the blocks along the diagonal time axis. Also note that the time scale of the simulation of the folded cystine knot is such that the longest-held metastable states last nearly as long as the entire nsp1 simulations.

Advanced forms of clustering often rely purely on big data and computational power to produce results. However, a useful idiom in this case is "Work Smarter, Not Harder!". An advanced and expensive version of clustering will likely not work as well as a simple one that takes into account several known aspects of MD trajectory data.

1. MD simulation trajectories contain time information. Therefore, sequential structures absolutely should be more likely to be clustered than non-sequential structures. In fact, for short time steps, it is essentially impossible for a protein to have a significant conformational change within a single time step. We take this into account with an adjustable averaging filter approach. While an averaging filter is potentially destructive when applied directly to MD simulation data, it is extremely simple to use on Euclidean distances in reduced dimensional space. This reduces thermal noise and random fluctuations without strongly affecting real changes in structure.

2. Clustering algorithms which attempt to find an "average" structure tend to create conformations that are not physically possible. Most proteins, even a significant portion of IDPs that do not show entropic chain behavior, exist in a discrete ensemble of preferred states. There are parts of fold space which are avoided because they are not energetically favorable. Average structures sometimes fall into these gaps. Therefore, our algorithm does not use average structures. We calculate the cluster centers as an average of the members; however, the single structure closest to the center is chosen to represent each cluster. All distances calculated between clusters are therefore also the distances between two actual protein conformations.

3. Most proteins, even most IDPs, tend to spend most of their time in local energy minima with short transitions between states. Some clustering algorithms can be tricked into collecting these transition states. We avoid this by always beginning node formations from the most similar protein structures among all comparisons. In this way, large clusters always form their cores in and around temporary structure or metastable states. This is particularly true when averaging filters are applied. Transition structures occur when a protein is changing between two or more conformations. Our



a

b

FIGURE 3.2: Color map display of the ISD of (a) the disordered protein FG-nucleoporin nsp1 and (b) a small folded cystine knot protein. Image rendered in MATLAB 2017b. Folded proteins tend to visit a small number of preferred conformations with short transitions between. The large dark blocks are local energy minima visited by the cystine knot. Disordered proteins spend less time in each conformation and transition between states more often.

hierarchical clustering is robust to cluster nodes being centered around these outliers. Transition structures tend to be either left out or collected by clusters of the metastable states they are transitioning to or from.

4. It is possible to estimate the amount of structural variance which can be explained by the random fluctuations caused by thermal noise. This can be used as a cutoff to estimate how much ISD between protein structures signifies a "real" differences between two conformers. This allows our algorithm to conduct unsupervised clustering since it knows when to stop. This also means that in instances of highly flexible and fully disordered proteins, some ensembles will simply not be clustered. Alternatively, manually setting the cutoff allows for a supervised version when forced clustering of highly disordered proteins is desired.

3.2.1 Spectral And Hierarchical Clustering

Spectral clustering requires first reducing the dimensionality of the input data. This was carried out using classical multidimensional scaling. In general, this is a difficult technique to apply to proteins because of the requirements on the underlying metric. CMDS utilizes a distance matrix for all comparisons. To function properly, the distances composing the matrix must be measured using a *metric* and the distances must be *Euclidean*. The requirements of these constraints are (1) symmetry, (2) the triangle inequality, and (3) the distances must not exist in a curved or reflected space.

In general, RMSD fails tests 1 and 2 because most molecular alignments use a heuristic approach which is not guaranteed to be a metric. Most measures based on internal coordinates fail test 3 because of the mirroring properties of angle space. Eventually, angles reach a maximal difference and begin to curve back around. For this reason, both can cause poorly behaved results in CMDS which generally results in the production of significant negative eigenvalues. The negative eigenvalues correspond to imaginary dimensions in reduced space where distances are also negative. This has an added side effect that for a given number of dimensions in the reduced space, additional dimensions may take away information rather than add it. All methods of calculating ISD have the potential to produce some negative eigenvalues due to limits of numerical precision. However, the metric we have chosen produces minute negative eigenvalues.

We have observed that RMSD is a particularly poor method of differentiating very different structures. Metrics based on internal protein coordinates are much more successful. We have based our calculation of ISD on an algorithm which calculates the RMSD after local alignments of the C_a dihedral angles along the protein's backbone. In essence, this provides the same information as comparing the internal coordinates of the polymer but converts the calculations into Euclidean distances to perform correctly with the CMDS algorithm.

Structural variance explained by thermal noise in an average folded proteins was estimated using MD simulations sampled from the Dynameomics database [16]. Protein data from Dynameomics needed to be converted to Gromacs-compatible formats using several custom scripts. MD trajectories representing 100 common protein folds were made available from the database. The *g_isd* tool calculated the average ISD (as calculated via the RMSD based on aligned backbone dihedrals) caused by thermal noise over all 100 protein simulations. This unitless value is 0.1841. Our metric intentionally freezes out the small degrees of freedom between amino acids along the backbone; however, since the average distance between amino acids in a polymer is fairly rigid at 0.38 *nm*, this is roughly equivalent to 0.0700 *nm* of local fluctuation. Since this is based on folded protein data, it is possibly too conservative of an estimation for IDPs.

3.2.2 Clustering Algorithm

The clustering algorithm follows these steps:

1. The ISD is calculated between all pairs of protein structures using our implementation of locally aligned RMSD along the backbone dihedral angles.
2. The ISD matrix is used as the input for CMDS. This implementation of CMDS can output all calculated dimensions in comma-separated values (CSV) format but also estimates the number of meaningful dimensions.
3. The dimensionally reduced data is given to the clustering algorithm where there are some preprocessing options. An averaging filter may be applied to the reduced data set. The original ISD matrix can optionally be used in place of the dimensionally reduced data. Obviously, using both of these options at the same time does not make sense.
4. All clusters of structures (including individual unclustered structures) are searched to find the overall minimum distance: the two most similar conformations. For clusters, a single representative structure is used to determine the clusters distances. In cases of ties, we have chosen not to break the tie randomly to make the result reproducible. The two closest structures are combined into a new cluster. If the two closest structures are already attached to a cluster then the clusters are combined.
5. The newly created cluster searches its members for a representative structure. First, the center is found in reduced dimensional space. Second, the member of the cluster which is closest to the center is chosen to represent the cluster for distance calculations. Future distances between structures in step 3 are all based on the representative structure. In cases of ties, we have chosen not to break the tie randomly to make the result reproducible.

TABLE 3.1: Polyglycine was too disordered to be clustered using the default cutoff from folded proteins. Several large structure clusters were identified for polyalanine, but several hundred structures (out of 1,000) were left unclustered. The results show polyglutamine to be much more rigid and structured with a smaller number of representative structures.

Polymer	**Dimensions**	**Clusters**
Polyglutamine	19	45
Polyalanine	29	380
Polyglycine	73	N/A

6. Steps 3 and 4 are repeated until the minimum distance between structures in step 3 is greater than the cutoff distance estimated from the Dynameomics database folded protein data.

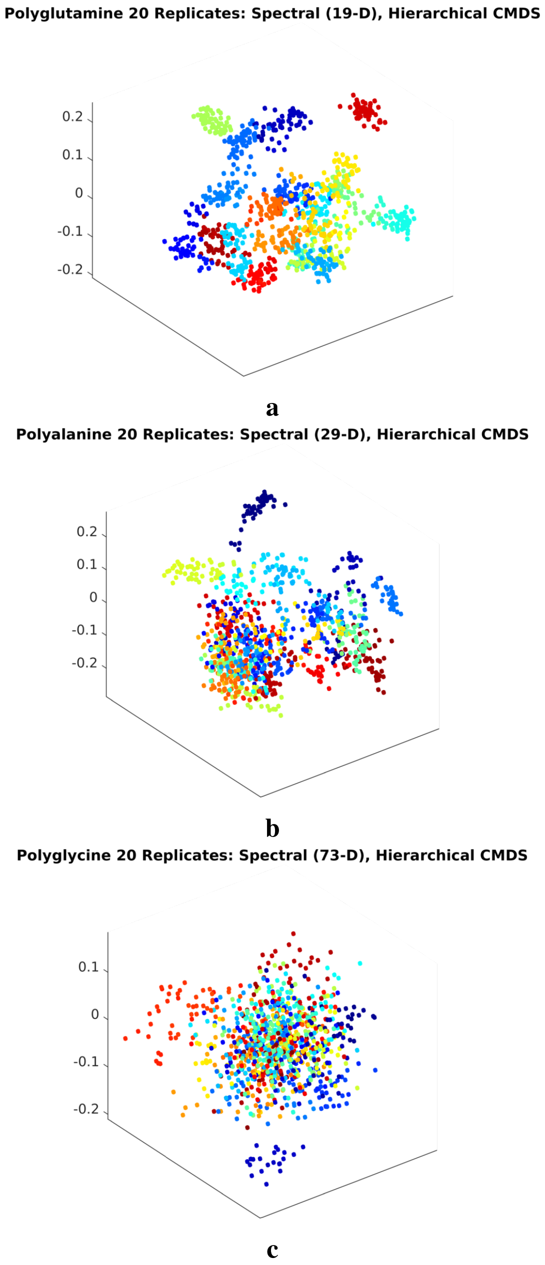
7. Output includes the numbers of clusters (singleton clusters are allowed), the number of structures in each cluster, and the index of the structure closest to the cluster center.

3.3 Results And Discussion

3.3.1 Validation With Homopolymer Models

HSC was applied to the model homopolymer trajectories. MD trajectories from all 20 replicates were combined into one and then sub-sampled. Clustering data is displayed in 3-dimensional space (**Figure 3.3**); however, the clustering utilizes a higher number of dimensions per the dimensionality estimator. Note that while distant structures in **Figure 3.3** are guaranteed to be distant, close structures are not guaranteed to actually be close due to the many hidden dimensions. The maximum number of theoretical degrees of freedom for a polymer of 50 amino acids is 98.

Numerical results are summarized in **Table 3.1**. Of note, the algorithm did not cluster polyglycine (there were a few exceptions) because nearly all the structures were considered to be too distant. This is the result that should be expected for an entropic chain where all structures are significantly different conformations. The polyglutamine ensemble shows a much more ordered appearance where portions of fold space are avoided and clusters appear around several relatively stable states. Polyalanine represents a hybrid with some compact clusters along with clouds of disordered transient states. The maximum number of structures in a single cluster for the polyglycine ensemble was 45.



Polyglutamine 20 Replicates: Spectral (19-D), Hierarchical CMDS

a

Polyalanine 20 Replicates: Spectral (29-D), Hierarchical CMDS

b

Polyglycine 20 Replicates: Spectral (73-D), Hierarchical CMDS

c

FIGURE 3.3: HSC was performed on the set of model homopolymer using reduced dimensionality of 19, 29, and 73 for peptides polyglutamine, polyalanine, and polyglycine respectively. The homopolymer protein models exist on a continuum of disorder from somewhat rigid with transient secondary structure (a) to highly flexible and disordered (c).

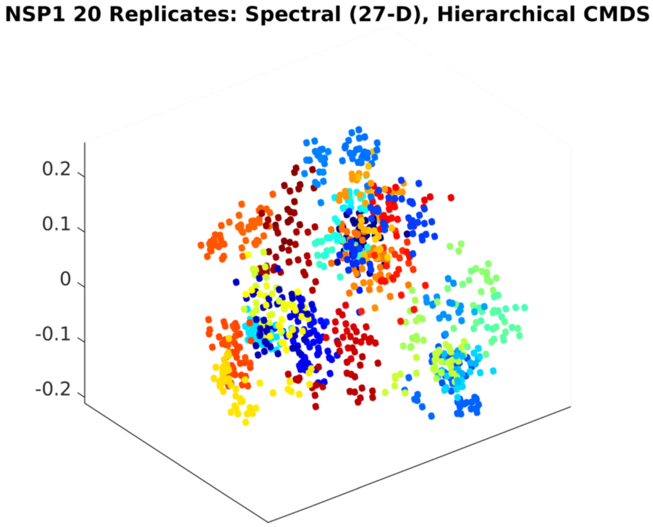


FIGURE 3.4: HSC was carried out on an ensemble of 20 replicates of the IDP, nsp1. The trajectory data were reduced to 27 dimensions before the clustering algorithm was applied. The ensemble appears most similar to the clustering results of the polyalanine homopolymer. The cluster sizes and shapes imply that nsp1 is more disordered than the nup116 FG-nup protein.

3.3.2 Applications To Proteins

The FG-nucleoporins are interesting as a family of natively unfolded proteins which provide vital functionality. However, the FG-nups' sizes are an important component of their function, and they sometimes respond to environmental stimuli by changing their properties of size or shape. This implies some form of constraint is placed on their available degrees of freedom. In fact, **Figures 3.4 and 2.9** reveals dimensionality estimation and clustering results more in line with polyalanine than the more fully disordered polyglycine. One reason to study these proteins is to attempt to distinguish between their preferred ensembles of conformations. The more randomized cloud-like appearance of the clustering results in **Figures 3.4** implies that nsp1 behaves as a more disordered and flexible protein than the FG-nup nup116 in **Figures 3.5**. Clustering results of nup116 reveal significantly more empty fold space and more compact clusters. This implies that a greater proportion of the possible fold space of nup116 is not energetically favorable, and the protein has a stronger preference for metastable states.

The final two proteins we have applied our clustering algorithm to exist on nearly opposite ends of the order/disorder continuum. The folded cystine knot protein in **Figure 3.6** contains only one disordered loop region and several small flexible residues. Our algorithm differentiates 6 preferred metastable states of the cystine knot protein along with 16

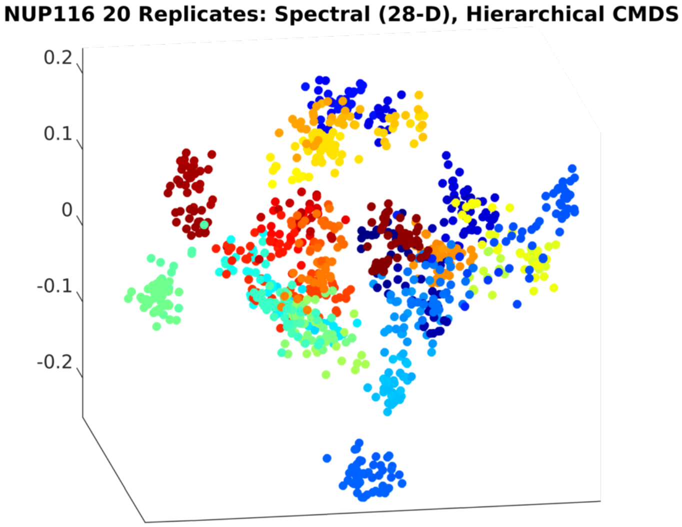


FIGURE 3.5: HSC applied to an ensemble of 20 replicates of the IDP nup116. The trajectory data were reduced to 28 dimensions. In contrast to the resulting clusters of nsp1, nup116 explores more tightly packed regions of conformational space and more of its fold space is avoided. This implies rigidity and constraints on its structure.

smaller clusters or individual structures. When the Gromacs tool *make_ndx* was employed to extract the disordered loop from the rest of the protein, HSC revealed a similar number of clusters. Therefore, most of the meaningful structural variance detected in the cystine knot protein comes from the disordered loop 1 region.

The Tau IDP (**Figure 3.7**) is a case of a real protein which is not clustered by our algorithm. Only approximately 10-20% of structures were combined with other structures. However, the largest cluster contained only 5 members. The ensemble of protein configurations does not appear to represent any significant metastable states. In part, this may be due to the algorithm which produces the structures from NMR data. These algorithms sometimes intentionally discard similar protein structures.

3.4 Conclusions

We present a method of clustering protein structures specifically designed for the dynamics of MD simulation trajectories. The algorithm is unsupervised by default, and uses a conservative estimate of folded protein dynamics to guarantees clusters which contain only variance that can be explained as thermal noise. The cluster outputs index numbers of representative protein structures for each cluster rather than averaged structure. HSC can

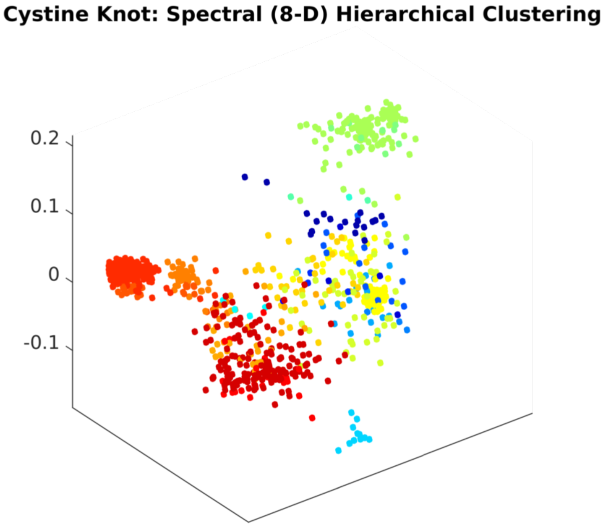


FIGURE 3.6: A folded cystine knot was differentiated into 22 clusters using 8 dimensions of data. Of those clusters, 6 contains most of the structures. Many of the clusters were small groups of outliers. An interesting note is that this small protein has a disordered loop region. Clustering of the disordered region alone resulted in a similar number of clusters. Therefore, most of the structural variation of the protein comes from that region.

reduce the large number of highly similar structures from MD trajectories into a handful. Due to its basis in physical quantities, the output structures are guaranteed to be minimally different from one another. The output structures are not artificial, and the algorithm is robust against clusters across transition states.

Since this method of clustering has a cutoff based on real physics rather than statistics, the number and size of clusters can be considered a rough estimate of the simulated system's disorder. Since the physical constants used are intended to be universal to proteins of all sizes and lengths, the results of unrelated proteins can generally be compared directly. Our algorithm is robust to common errors in the CMDS algorithm and accounts for thermal noise.

Using simulations, we have verified experimental results showing a slight difference in the level of disorder of the nsp1 and nup116 FG-nups. The nup116 protein appears to show a stronger preference for metastable states and spends less time transitioning between states. HSC verified that experimental NMR data of the K18 domain of Tau protein reveals behavior nearly as disordered as an entropic chain.

We suggest that one extremely elusive property of IDPs is the dynamics of fold space exploration. Quantifying the amount of fold space explored by IDPs is generally much simpler than quantifying how quickly they explore it. Reduced dimensional space

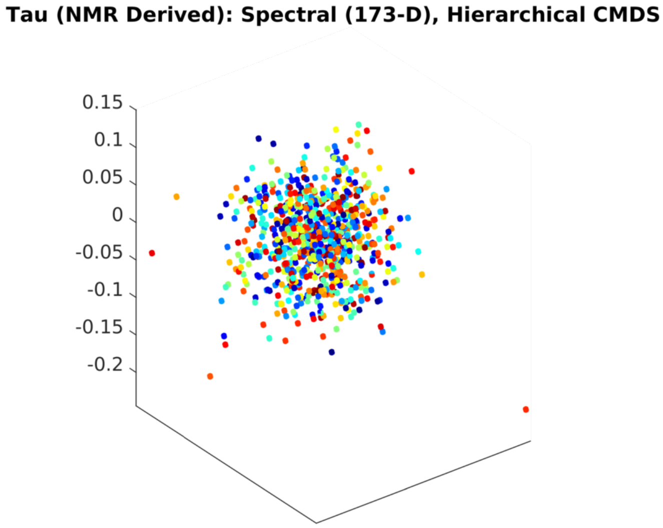


FIGURE 3.7: The clustering algorithm was applied to Tau protein and it determined that few of the structures are closely associated. It left nearly all structures unclustered. This is the type of behavior to be expected for an IDP that behaves like a random entropic chain. The dimensional estimator for Tau protein found 173 meaningful degrees of freedom out of 258 theoretically possible dimensions (this domain of Tau contains 130 amino acids).

provides a method to estimate a "speed limit" of sorts. Some proteins race through their conformational space while others slowly meander. Slight modifications of our approach would provide a universal estimate of the dynamics of disorder, and the estimates of thermal noise that we have presented would allow a means of universally scaling the speed of exploration. In this context, the thermal noise is a kT-like quantity that has been translated into units of protein dynamics.

3.5 MD Simulations

Homopolymer chains of 50 amino acids were simulated for the polypeptides polyglycine, polyalanine, and polyglutamine. FG-nup sequences for the proteins nsp1 and nup116 can be found in **Table 1.1**. Simulations of the cystine knot protein used pdb entry 2N8B [107] as an initial structure. The 2N8B structure contained a small error that required disulfide bonding to be handled manually. An ensemble of experimentally derived K18 domain Tau protein structures was downloaded from the Protein Ensemble Database [15, 103]. All MD simulations were run using Gromacs version 2016.3 [33–38].

All-atom MD simulations of IDPs were run on protein fragments of 50 amino acids in length using the velocity rescaling thermostat [54] and Parinello-Rahman pressure coupling [55]. MD simulations of IDPs in normal force fields are known to show erroneous secondary structure or unnatural compaction. Instead these simulations used the Amber ff03ws force field [60] with the TIP4P/2005 explicit solvent model [61] which has been specifically modified to improve the accuracy of IDP MD simulations. The 5 disordered proteins were each run in 20 replicates from initial configuration generated by a custom program called RanchaR (random chains in R). This relies on the R programming environment [62] and several packages for parallelization [63–65]. Pulchra version 3.04 [66] was used to convert the initial randomized polymer chains to optimized all-atom protein structures.

Since IDPs show a wide variance in size, MD simulations of them have issues with possible periodic boundary conditions errors. In order to estimate the necessary simulation box size, the RanchaR script was used to generate 100,000 randomized, self-avoiding entropic chains for each protein sequence. The maximum distance between residues was calculated for all generated structures to estimate the periodic boundary conditions large enough to fit more than 95% of the structures. Each replicate included a 50 ns equilibration followed by a 100 ns production run. Additonal results are presented based on a 1,000 nanosecond simulation of the FG-nucleoporin, nsp1. All simulations used uncapped sequences without the addition of acetamide or N-methyl groups. Explicit solvent replicates contained approximately 40,000 to 60,000 atoms depending on the polymer.

The IDP structures were energy minimized from their randomly generated initial configurations. Short MD simulations steps were carried to thermalize, simulate with position restraints, and gradually increase the time step from 1 to 5 fs. Several short NVT MD simulations were run with 1 fs, 2 fs, and 5 fs time steps for stability. Running MD simulations at 5 fs time steps required optimization settings in Gromacs: bond constraints, virtual site hydrogens, and heavy hydrogens in water.

The initial cystine knot structures were first energy minimized in vacuum and with solvent. The structure was then simulated by briefly thermalizing the protein to 310 K with position restraints. Simulations used the tip3p explicit water model [59] in a dodecahedral water box with the Amber ff99SB-ILDN force field [56–58]. All simulations after thermalization were run at 310 K. In order to approach a 5 fs time step, several short NVT MD simulations were employed at 1 fs, 2 fs, and 5 fs. As with the IDP simulations, 5 fs time steps required that we enable bond constraints, virtual hydrogen sites, and the heavy hydrogen atoms setting. A short constant pressure (NPT) simulation with Berendsen pressure coupling [67] was run followed by an equilibration simulation using Parinello-Rahman pressure coupling [55]. Finally, the cystine knot proteins were simulated for a total of 8 μs.

3.5.1 Further Details

To create the distance matrix required by the CMDS algorithm, ISD between strutures was calculated using a unitless version of locally aligned backbone dihedrals RMSD. This is chosen with the -rmsdih and -scaled options. The details of the implementations are discussed in the **Implementation Details For ISDMs** section.

3.6 Acknowledgements

The original plans for applying spectral clustering techniques to systems of disordered proteins came from Dr. Joshua Phillips and Dr. Shawn Newsam. Dr. Michael Colvin and Dr. Ajay Gopinathan were involved in designing testing procedures to validate various algorithms.

Chapter 4

Targeted Binding Of Cyclotides

4.1 Background

The cystine knot is a structural motif in proteins defined by the presence of three disulfide bonds which stabilize the loop. The motif was first observed in nerve growth factor [108] but is now known to be present in many inhibitor proteins. The motif is found in bone formation inhibitor [109] and insect serine protease inhibitor proteins [110].

A particular family of plant cystine knot proteins contains an end-to-end macrocyclic disulfide bond. This allows them to be one of the smallest folded micropeptides in nature [85]. Referred to as cyclotides or cyclins, these small proteins are defined by a single cystine knot motif and are only able to fold due to the stabilization provided by the disulfide bonds [105]. Many cyclotides have antimicrobial [85] or inhibitory properties.

The study of plant cyclotides has had widespread interest because the proteins are simple enough to be completely synthesized [111]. This has made the protein a viable scaffold for drug delivery and a candidate for other types of customized molecules [106]. Nearly all of the stable tertiary structure of cyclotides come from disulfide bridges. Therefore, the free loops in between bridges may be replaced with biologically active sequences without disturbing the structural stability of the knot. In particular, plant cyclotides contain a large disordered loop 1 region that may be modified with a targeted binding sequence.

We present a comparison of a modified and wild-type pair of cystine knot proteins with pdb entries 2N8B and 2N8C [107, 112]. The modified cystine knot was tagged with the addition of a 2-fluoropropanoic acid (2-FP) molecule covalently bonded to the N-terminal glycine residue. Molecular dynamics (MD) simulations of the pair of cyclotides were run to verify that the 2-FP label does not significantly affect the local structure. Due to the location of the protein attachment site, the label is in close proximity to the disordered loop 1 region. This cystine knot protein contains an integrin $\alpha v\beta 6$ cancer recognition site on this loop [107], so any induced folding altered dynamics of the disordered loop could impact the

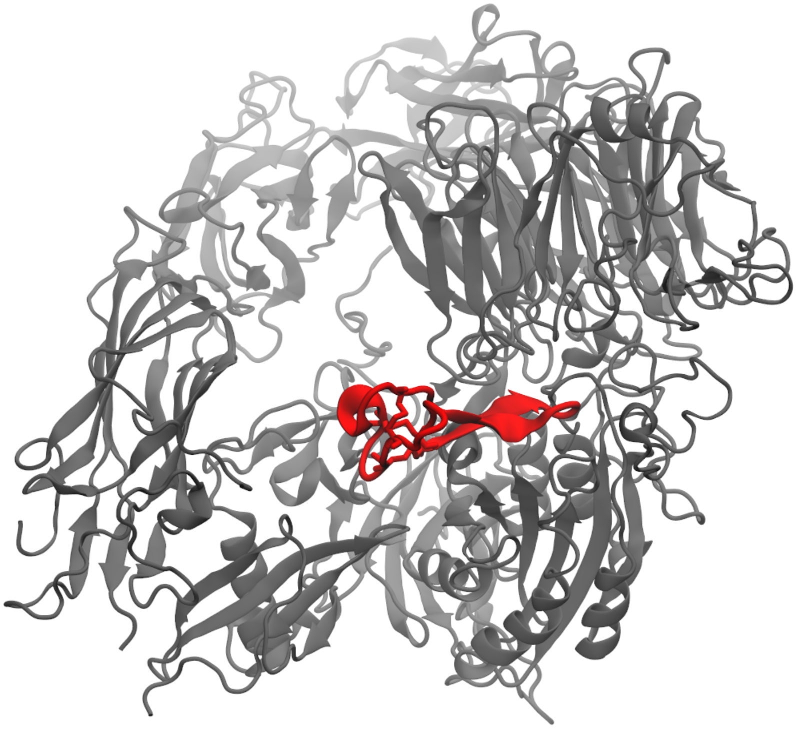


FIGURE 4.1: Illustration of the small cystine knot protein (red) docked to its target integrin molecule (gray). The ligand is docked near the binding site of a small peptide molecule with the sequence RGD. After docking and energy minization, the disordered loop 1 region of the cystine appears to form a β-sheet secondary structure.

effectiveness of the targeted protein. Integrin $\alpha v\beta 6$ shows increased expression in a wide variety of cancers [113].

The cystine knot ligand is presented in a final docked configuration in **Figure 4.1**. The Autodock software is a set of tools and software to dock small molecule ligands to target proteins [114–117]. It employs a force field for energy minimization with a genetic optimization algorithm and selective protein flexibility.

4.2 Methods

Before docking the cystine knot ligand to its target, we verified that the 2-FP modification did not significantly affect the dynamics of the loop 1 region of the cystine knot. A universal measure of disorder, the order parameter, was calculated using the results of MD simulations to compare the flexibility of the wild-type and modified cystine knots. The explored conformational space of the two conformers was compared and found to be overlapping.

The cystine knot is small for a folded protein at 36 amino acids in length. However, this is far larger than the small molecules normally docked by automated software. Therefore, we needed to implement a more complicated procedure.

In brief, the target sequence of the cystine knot protein was sampled in multiple configurations from MD simulations. The ensemble of 5 amino acid polypeptides were all docked using the Autodock Vina software [117]. Several of the best docking structures were kept, and the entire cystine knot protein was fit to the docked peptides using the Gromacs tool confrms. The fits were analyzed visually for significant steric clashes with the target protein, and the best were kept. Finally, an energy minimization and a short equilibration were run on the docked target-ligand complex using Gromacs software [33–38].

4.2.1 MD Simulations

MD simulations were performed on the cystine knot proteins with pdb entries 2N8B and 2N8C [107]. These structures were solved using NMR and the protein has significant unfolded regions. The resulting ensembles available from the protein database [112] had small structural errors. Therefore, the disulfide bonds needed to be handled manually.

MD simulations were run in Gromacs 2016.3 [33–38]. The initial cystine knot structures were first energy minimized in vacuum and with solvent. The structure was then simulated with position restraints following a short thermalization. Production runs of the cystine knot protein used the velocity rescaling thermostat set to 310 K. Simulations used the tip3p explicit water model [59] in a dodecahedral water box with the Amber ff99SB-ILDN force field [56–58]. Several short constant volume (NVT) MD simulations were employed to slowly increase the time step from 1 femtoseconds (fs) to 5 fs. In order to reach the targeted time steps, several optimizations were employed: bond constraints, virtual hydrogen sites, and the heavy hydrogen atoms setting. A short constant pressure (NPT) simulation with Berendsen pressure coupling [67] was run followed by an equilibration simulation using Parinello-Rahman pressure coupling [55]. Finally, the cystine knot proteins were simulated for 8 μs.

4.2.2 Ligand Docking

Due to the large sizes of both the ligand and target proteins being docked, we were unable to follow a simple and standard procedure to perform the docking. The integrin protein from pdb entry 1L5G [112, 118] was cleaned up and energy minimized using Gromacs [33–38] before the docking steps. MD simulations of the cystine knot were sampled every 2 microseconds to generate several initial conformations for docking. The target integrin protein has been solved while bound to a small peptide with an arginine-glycine-aspartic acid (RGD) sequence.

Our protein had two potential homologous sequences in the loop 1 region. An RTD sequence over residues 7-9 and a reversed ordered sequence of NGR over residues 5-7. Since it was not feasible to dock the entire cystine knot protein, we created two trimmed 5 amino acid peptide fragments around the target sequences: 4-8 and 6-10. Autodock tools and Autodock Vina were utilized to dock the 10 peptide fragments to the target binding site on the integrin protein. The search for docked conformations was set to an exhaustiveness of 40. Autodock Vina gives an ensemble of docked conformations; however, only 9 of the 90 docked structres were near the target binding site. Of these, the two structures with the more favorable binding energies were kept.

The MD simulations were sampled again, and a total of 8 structures of the full cystine knot protein were fitted to the docked target sequence. The fit was carried out using the Gromacs tool *confrms* [33–38]. Since the sequence on the cystine knot proteins only shared two amino acids with the true target sequence, the fit was performed using all heavy atoms from the matching amino acid residues and only backbone atoms from the others.

The resulting fitted cystine knots-integrin complexes were filtered visually. Since the fit procedure does not take steric clashes into account, many complexed structures were ruled out due to overlap between the two proteins that could not have been fixed with an energy minimization. All fitted complexes had at least some protein-protein overlap. The two best candidates were chosen based on their being the most physically possible. Energy minimizations and short MD equilibrations were run on two protein complexes.

4.3 Results And Discussion

4.3.1 Protein Analysis

The MD simulations of the cystine knots were compared against an ensemble of 800 structures produced based on NMR data using CYANA software [119–123]. This comparison was used to verify that the simulations were run for a long enough time to give a good representation of the protein dynamics. The wild-type and and modified cystine knot proteins were analyzed using the *g_isd* tool to quantify and compare the regional disorder and the sampled conformational space of the two proteins using a reduced dimensional representation. The *g_isd* tool calculated the order parameters for the two proteins and employed classical multi-dimensional scaling on their trajectories. Images of the sampled conformational space were produced in MATLAB 2017b. Protein renders of cystine knots with coloring based on the order parameter were produced using a chain of tools and scripts. The *g_isd* analysis tool calculates order parameter values, and a script *xvg2bf.bash* converts the file to a Gromacs-compatible format. The Gromacs *editconf* tool is able to overwrite the $beta$ factor values in pdb files, and VMD was used to display and render the final result [94].

The calculated values of the order parameter for the wild-type cystine knot protein are found in **Figure 4.2.b**. These values are overlayed onto the protein as a color key where the light region illustrates the local disorder of the loop 1 region.

An ensemble of structures was derived from the experimental NMR data that was used to produce the 2N8B and 2N8C pdb entries. The CYANA software produced 800 structures based on the experimental data, and this ensemble was used to verify that the MD simulations were long enough to sample the conformational space. Per **Figure 4.3**, the ensemble of structures derived from experimental data appears to explore less conformational space and be less disordered than the MD simulations would predict. The results indicate that the MD simulations are a decent representation of the full conformational space. The calculated disorder for the labeled cystine knot is nearly identical to the value calculated for the wild-type (**Figure 4.4**). Therefore we conclude that the 2-FP modification did not significantly hinder the dynamics of the cystine knot protein. The modification likely has no detrimental effect on the ligand's ability to bind to the target.

The MD trajectories of the pair of simulations for the modified and wild-type cystine knots were combined into a single file with the Gromacs tool, trjcat [33–38]. This allowed us to use the *g_isd* analysis tool to process both proteins in the same reduced dimensional space. The loop 1 region of the two conformers over residues 4 to 15 was used as the input to a classical multi-dimensional scaling (CMDS) algorithm implemented by *g_isd*. The dimensionality of the data was reduced to 3 to improve visualization. Note that the peptide fragment being used is small, so the calculated correlation between the original data and the reduced data is $r = 0.868591$. The comparison shows that the two proteins explored virtually identical conformational space in simulation. A small 10 ns averaging filter was applied to the results to reduce noise and focus on the two proteins' trajectories through their conformational space.

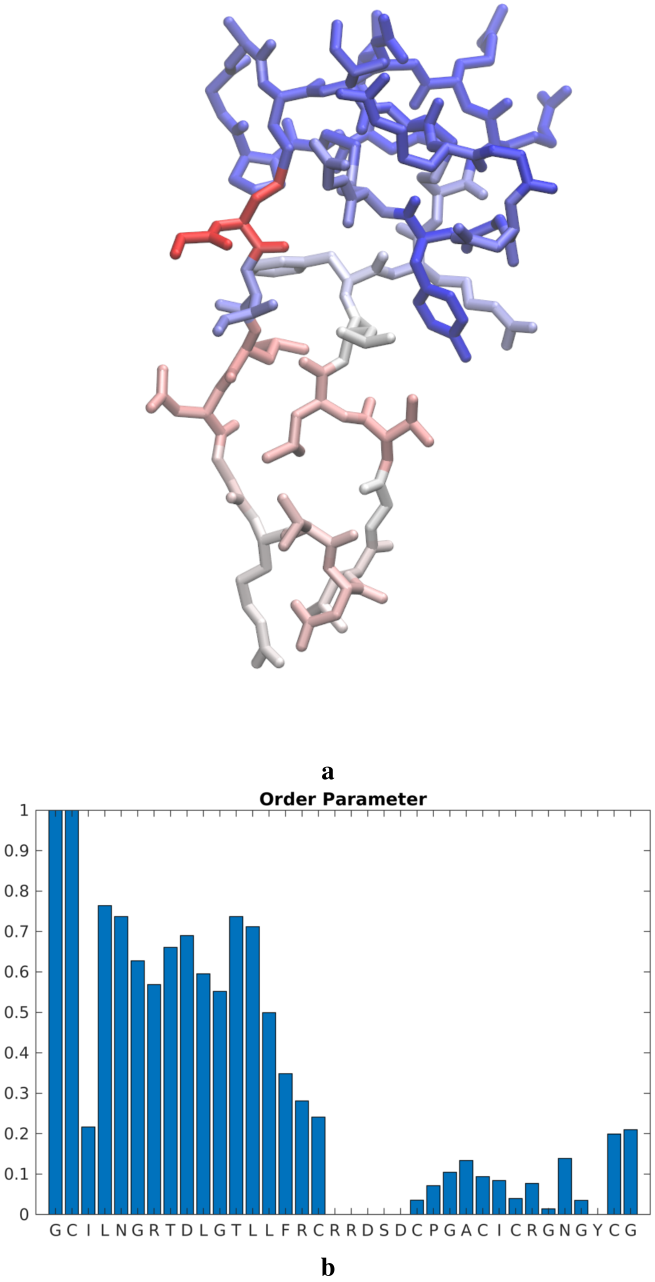


FIGURE 4.2: Detected regional disorder in the cystine knot protein. The order parameter for the protein was calculated by using the *g_isd* tool on the 8 μs MD trajectory.

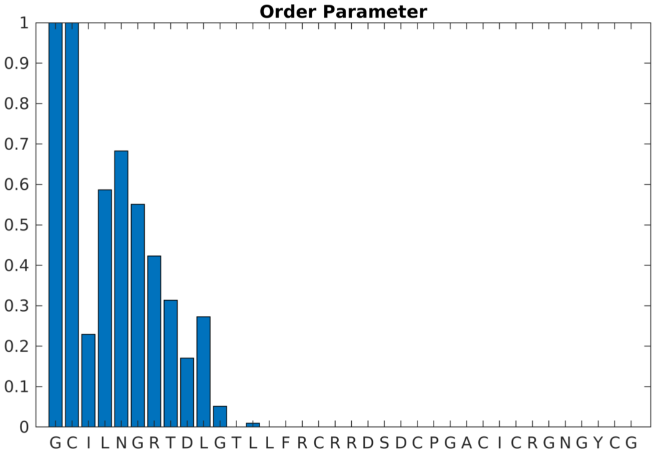


FIGURE 4.3: The order parameter calculated for the ensemble of structures based on NMR experimental data. Unexpectedly, the ensemble appears to explore less conformational space than the MD simulations.

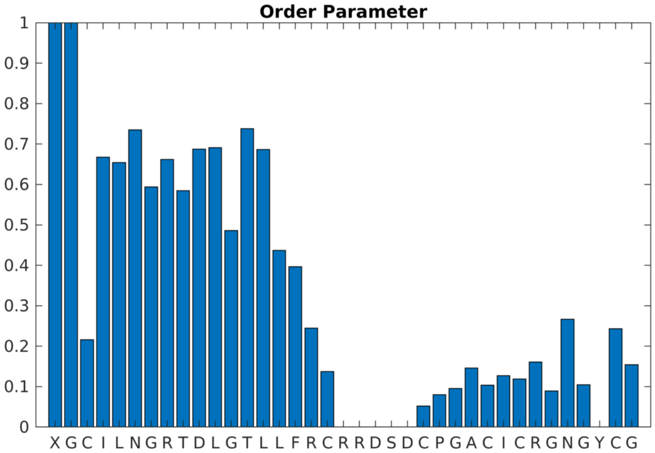


FIGURE 4.4: The disorder estimate for the modified cystine knot protein is nearly identical to that of the wild-type protein. From this, we infer that the 2-FP label modification did not affect the structure of dynamics of the disordered loop 1 region of the cystine knot.

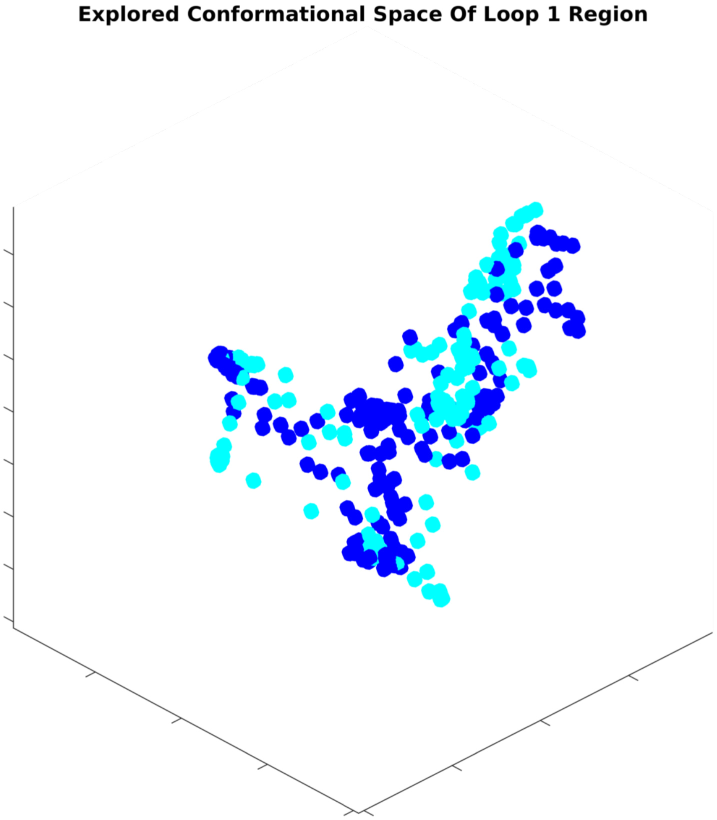


FIGURE 4.5: We display the paths of the loop 1 regions of modified and wild-type cystine knot proteins through conformational space in reduced dimensionality. Since the peptide fragments being analyzed are small, the three dimensions of reduced data highly correlate with the original information. The two variants explore virtually identical portions of the conformational space.

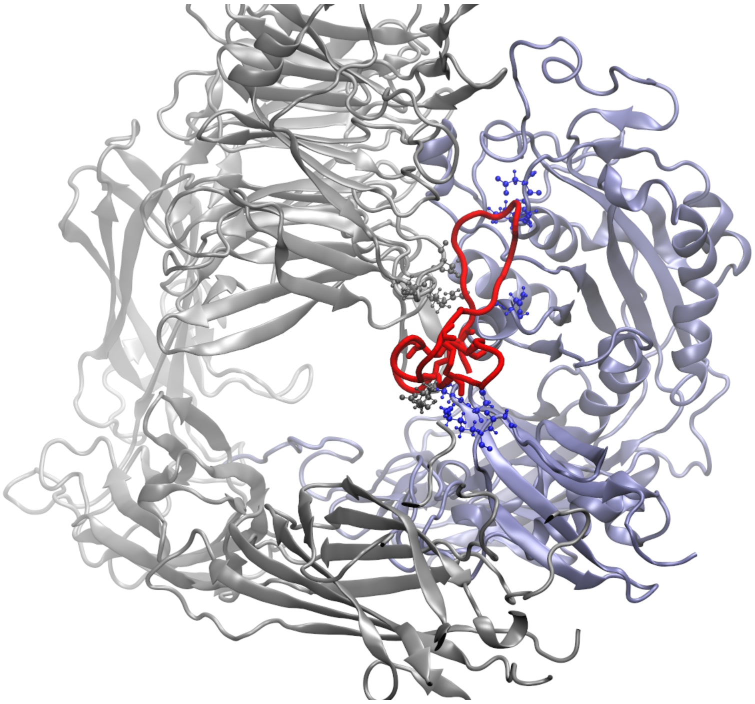


FIGURE 4.6: The two best fitted structures were energy minimized and equilibrated. One of the two ligands left the primary binding site, the space between the silver and purple domains just above the ligand's location (in red), during the equilibration.

4.3.2 The Docked Ligand–Target Complex

Autodock Vina calculated the affinity between ligand and target as approximately -5.5 kcal/mol for the most optimal docked structures. The best docked ligand fragments were chosen based on affinity and whether the docked ligand interacted with the targeted binding site. A variety of sampled cystine knot conformations (from the original MD simulatation trajectories) were fit to the docked peptide fragments. Most showed significant overlap with the target protein due to the tight spacing of the binding pocket. The two fitted proteins with the least overlap were run through an energy minimization and short equilibration. Finally, the best docked complex was easy to choose because during equilibration, one of the ligands popped out of the binding site (**Figure 4.6**). The optimally docked target–ligand complex is displayed in **Figure 4.1**.

4.4 Acknowledgements

This work was performed in collaboration with Dr. Fabian Filipp. Protein structures derived from experimental data were produced by his lab, and he offered direction and insights to the work. Dr. Michael Colvin assisted with advice and suggestions on setting up the MD simulations properly.

Chapter 5

Polymer Disorder Modulates Thermal Contraction

5.1 Background

The disorder-order spectrum exists in non-biological systems, and the tools applied to the study of biological molecules and systems also have applications to non-biological systems. In the following, we present a polymer system which presents the property of a giant thermal contraction mediated in part by the existence of local disorder [124].

Analysis tools to study local molecular disorder were applied to a polymer material containing repeats of the molecule S-dibenzocyclooctadiene (DBCOD). The DBCOD molecule contains an eight-member ring attached to two phenyl rings: one bonded on each end. The experimentally interesting property of this polymer material is a reversible thermal contraction that can be triggered using low-energy near-infrared light [124]. Materials with changes to shape or size that can be activated and reversed have seen applications in regenerative medicine [125, 126], medical drug delivery [127, 128], and robotics [129, 130]. In current medical and industrial applications, many materials with these properties rely on processes that have extreme ramifications at local molecular scales: phase transitions, molecular binding, and movement along the order-to-disorder spectrum [124].

Per the results of our local molecular analysis, the DBCOD thermal contraction appears to rely on a local conformational change which is low energy and does not involve severe changes to the local molecular environment such as the loss of covalent bonds or the addition of significant disorder. Furthermore, we show that properties of the thermal contraction of the DBCOD polymer material can be altered through the mixture of two polymer sub-types. One of the polymers contains 3,4’-oxydianiline (3,4’-ODA), pictured in **Figure 5.1.a**, which causes the polymer material to favor a locally disordered, globular conformation at the molecular level. This results in an amorphous material structure which is more kinetically favorable to the conformational change triggering thermal contraction

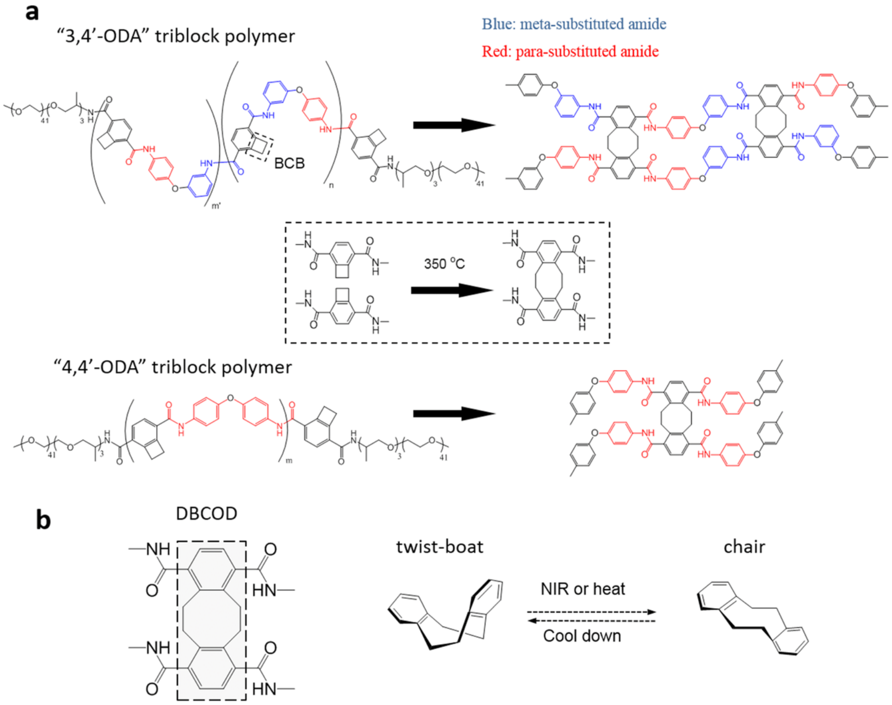

a
"3,4'-ODA" triblock polymer
Blue: meta-substituted amide
Red: para-substituted amide
BCB
350 °C
"4,4'-ODA" triblock polymer
b
DBCOD
twist-boat
chair
NIR or heat
Cool down

FIGURE 5.1: (a) Illustration of the two DBCOD sub-types: 3,4'-ODA polymer and 4,4'-ODA polymer The difference between the two polymer sub-types is the bond between the DBCOD monomers at the position farthest removed from the central cross-linked section highlighted in (b). The application of energy to the system, either heat or near infrared photons, causes a stochastic conformational change across the DBCOD cross-linked section favoring the chair over the twist-boat conformation. [124].

and expansion. The other polymer contains 4,4'-oxydianiline (4,4'-ODA), **Figure 5.1.a**, which tends to favor a more bundle-like, ordered conformation. The resulting macroscopic structure has the appearance of woven fibers and requires a higher kinetic energy input to produce the same degree of thermal contraction.

Per **Figure 5.2**, while both polymer types eventually reach similar degrees of contraction, the polymer containing less ordered 3,4'-ODA requires less energy to reach the maximal degree of contraction than the polymer containing more ordered 4,4'-ODA. The contraction per unit of temperature rise of the 3,4'-ODA polymer (-2350 ± 73 ppm/K) is approximately twice that of the 4,4'-ODA polymer (-1140 ± 64 ppm/K) [124].

The particular interest in the thermal contraction properties of the DBCOD polymer material is the low energy input requirement compared to other materials with a similar response to energy. Other studied polymers [131–133] tend to rely on the isomerization of a molecular subunit which generally requires higher energy UV radiation to trigger a

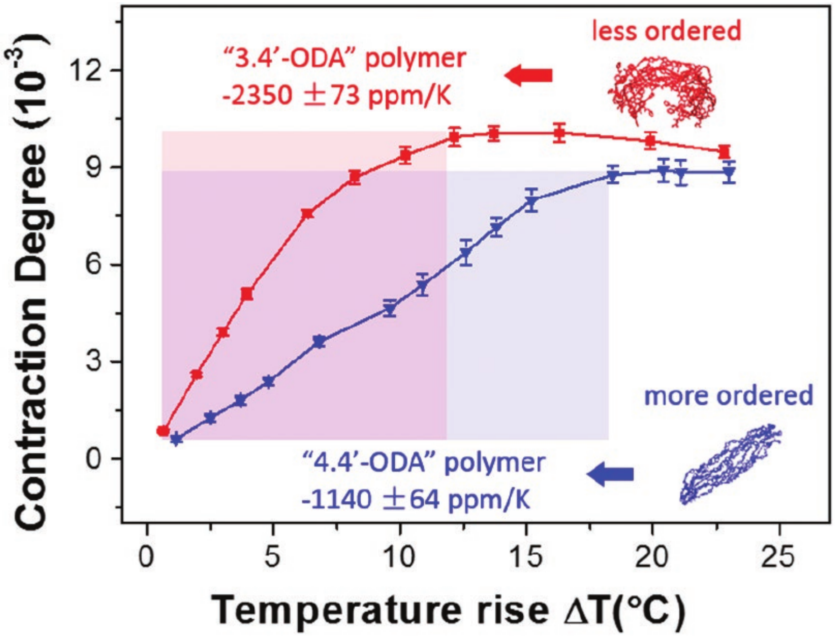


FIGURE 5.2: The less ordered 3,4'-ODA polymer approaches a total contraction of approximately -10,000 ppm (red) while the more ordered 4,4'-ODA polymer approaches a similar total contraction of approximately -9,000 ppm (blue). However, the polymer containing less ordered 3,4'-ODA requires far less energy to reach the maximal degree of contraction than the polymer containing more ordered 4,4'-ODA [124].

mechanoresponse. High-energy UV photons are necessary for the isomerization of the molecules, but high-energy photons may also break other incidental covalent bonds and results in unavoidable photobleaching of samples. Since UV light increases rates of cancer, these materials have limited biological and medical applications.

Experimental results show that the conformational change resulting in thermal contraction is a stochastic shift from twist-boat to chair in the central ring structure of the DBCOD monomer as illustrate in **Figure 5.1.b**. The low energy requirements of the DBCOD conformational change means that low energy infrared photons may be used instead, and a significant mechanical response occurs at near room temperature [124]. Furthermore, the mechanical response to input can be tuned using a mixture of the 3,4’-ODA and 4,4’-ODA polymers.

5.2 Experimental Methods

The experimental methods to synthesize the DBCOD molecule and measure the thermal contraction properties have been published [124].

In brief, several ODA precursor mixtures were combined with other ingredients to create polymer composite solutions [124]. This was poured onto a mold and evaporated to form films [124]. The films were annealed under a partial vacuum [124]. A 970 nm laser was used to heat the film [124]. A tensometer was used to measure the contraction force while an infrared camera monitored temperatures [124].

5.3 Simulation Methods

To supplement the experimental observations of the twist-boat to chair conformational change and the observed differences in disorder of the polymer material, molecular dynamics (MD) simulations were carried out to compare the two sub-types of the DBCOD polymer.

Since the DBCOD polymer is not closely related to a biological system, the simulation systems needed to be built *in silico*, and a custom set of force field parameters was generated. The system coordinates were built and bonding was controlled using a script in the R programming environment [62]. Invidual interactions of the force field were generated by the SwissParam web server [134]. MD simulations were run using Gromacs 4.6.5 [33–38]. All visual representations of molecules were created using Visual Molecular Dynamics (VMD) [94].

5.3.1 Building The System Coordinates

Initial coordinates for the simulation system containing 4,4’-ODA molecules were produced by a script written for the R programming environment [62]. The script builds up the simulation system from monomers as illustrated in **Figure 5.3**. The script writes out atomic coordinates for five *monomers* which are covalently bonded into strands of *polymers*. Sets of eight polymer strands are cross-linked into *sheets* across neighboring DBCOD molecules wherever they are in close proximity. Four sheets are placed in each *layer* without covalent bonds. The simulation *system* comprises 8 layers.

The simulation system contains 1,280 monomers as part of a solid-state system with no solvent. The algorithm to construct the initial coordinates for the system containing 3,4’-ODA molecules follows the same steps. However, the bond between monomers is randomly assigned to either type (A) or type (B) in **Figure 5.4**. The simulation systems used vacuum annealing rather than temperature annealing to reduce the initial energy barriers and randomize the subunit conformations. In order to facilitate the low pressure environment, the layer spacing was initialized to 4.0 nm to simulate a low density of 71.8 kg/m^3. [124]

5.3.2 Generating Force Field Parameters

The force field parameters for the DBCOD system were generated using the SwissParam web server [134]. However, the SwissParam software is optimized for the parameterization of small molecules; producing a custom force field for a system of tens of thousands of atoms is not feasible. Our approach to scale up the system size was first to parameterize a system containing all possible unique interactions contained within the total system. The tetramer in **Figure 5.5** contains both cross-linked and unlinked DBCOD subunits and contains both bonded and terminal 4,4’-ODA subunits. The structure was sketched in and exported from the software MarvinSketch in Marvin Beans version 14.11.3.0, 2014, ChemAxon (http://www.chemaxon.com). This is one of the methods suggested by the SwissParam web server to generate a compatible mol2 format file for upload.

After the MD force field files were generated by the SwissParam server, the individual interactions produced were harvested, reformatted, and incorporated into Gromacs topology files as a force field option. The improper dihedral potentials generated by the SwissParam server are harmonic which is a different format than the periodic ones with only one force constant used by most of the force fields in Gromacs. We replaced these improper dihedral parameters with periodic potentials derived from the Amber99SB-ILDN force field [56–58]. By inserting the SwissParam potentials in the standard format of a Gromacs topology file, the pdb2gmx tool was able to automatically generate topologies for systems of arbitrarily large sizes such as our DBCOD system with over 50,000 atoms.

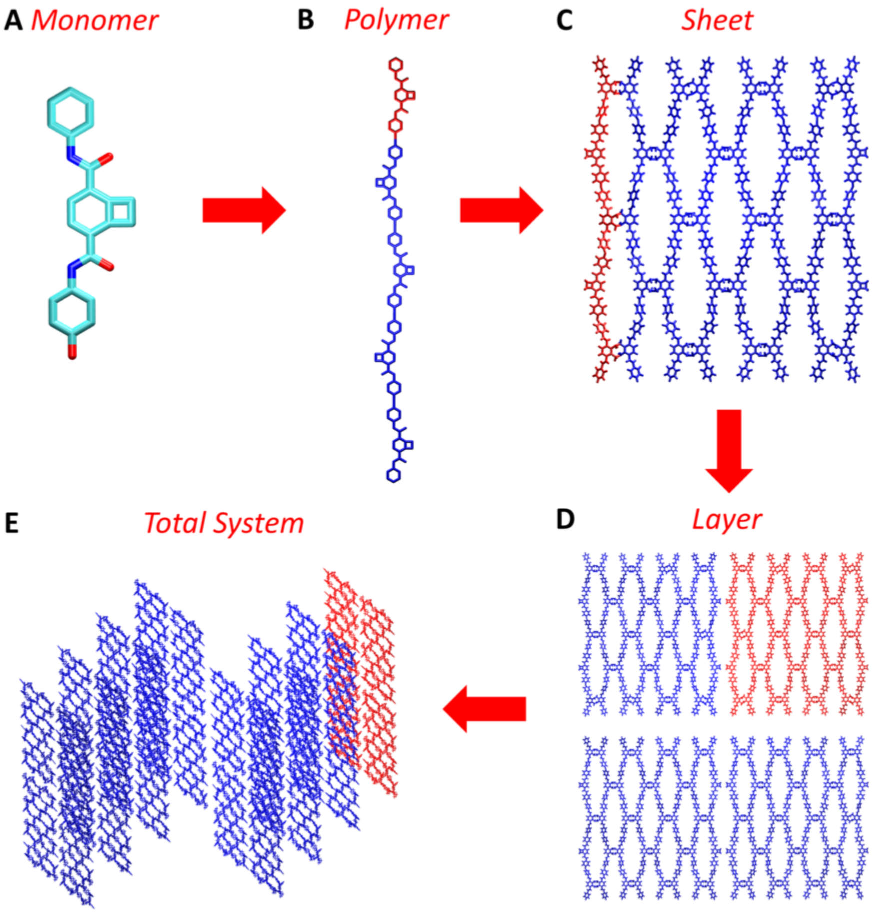


FIGURE 5.3: Illustrates the algorithm used by the R script to generate initial coordinates used to simulate the 4,4'-ODA molecular structure. The (e) *total system* is built from 8 (d) *layer*s. Each (d) *layer* contains 4 separate (c) *sheets* with 8 cross-linked (b) *polymer* strands of 5 (a) *monomers* (which are displayed without hydrogen atoms). The simulated system contains a total of 1,280 *monomers* and 55,296 atoms. At each structural step, one subunit of the previous step is highlighted in red. [124].

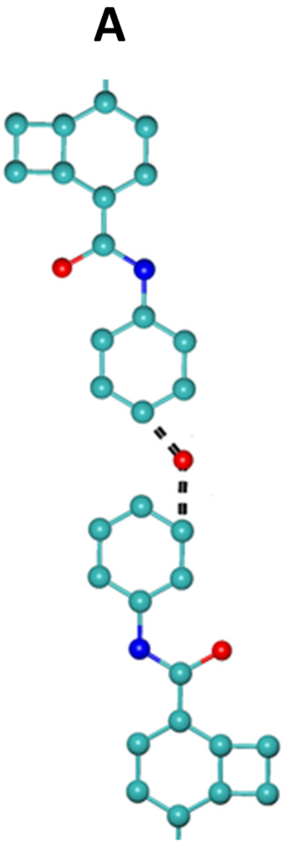


FIGURE 5.4: The algorithm used by the R script to generate initial coordinates for the 3,4'-ODA system follows the same steps as seen in **Figure 5.3**. However, the script must also randomly assign the 3,4'-ODA bond as one of two types. [124].

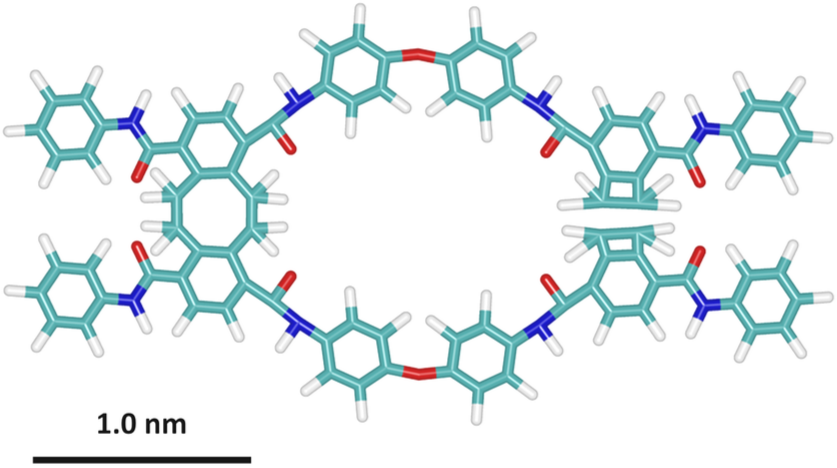


FIGURE 5.5: A tetramer of the 4,4'-ODA sub-type of the DBCOD-based polymer contains all the unique bonded interaction necessary to generate an MD force field for the simulated systems. [124].

5.3.3 MD Simulations Of DBCOD Molecule Systems

One simulation system was produced for each of the 3,4'-ODA and 4,4'-ODA polymers to compare the resulting local disorder. MD simulations were carried out using the Gromacs library of tools and software [33–38]. The systems were first energy minimized and then run through a short 100 ps constant volume (NVT) simulation with position restraints and the velocity rescaling temperature coupling method [54] at a temperature of 300 K. Next the system was annealed using partial vacuum with an NVT simulation of 4.0 ns at the initial system volume $(10,472nm^3)$ and density $(71.8kg/m^3)$.

A delayed collapse was simulated using constant pressure (NPT) settings. MD simulations of the systems were run under a constant 1.0 bar pressure using Berendsen pressure coupling [67]. Rather than doing a single collapse simulation which may have caused the molecules to be stuck in energetically unfavorable conformations, the collapse was halted at system volumes of approximately $4,000nm^3$ and $2,000nm^3$. Additional 4.0 ns NVT equilibration simulations were run at these system volumes before allowing the collapse to continue.

After the final NPT simulation where the system volumes were stabilized, a 2.0 ns NPT equilibration simulation was run using the more accurate Parrinello-Rahman pressure coupling method [55]. We verified that the final equilibration allowed the system to reach stable potential and kinetic energies (**Figure A.18**). The MD equilibration simulations of the two ODA systems were followed by 8.0 ns production runs.

Analysis of the MD simulation data used tools included in the Gromacs software library [33–38], the g_shape tool to compute shape parameters of molecular structures [18], and several custom scripts written for the R programming environment [62].

5.4 Results And Discussion

5.4.1 DBCOD Twist-Boat To Chair Conformational Change

Experimental results indicate that the cause of thermal contraction in DBCOD-based polymer is the energy-dependent, stochastic conformational change from twist-boat to chair configuration illustrated in **Figure 5.1.b**. To supplement the experimental data, production runs of the two polymer sub-types were analyzed to calculate the angle across cross-linked DBCOD subunits for each system. Computations were carried out using a script written for the R programming environment [62], and the calculated angles were compiled into histograms for each of the simulation systems.

In **Figure 5.6**, the two histogram peaks in the center of each figure correspond to acute angles made by the twist-boat configuration which is the global energy minimum of the DBCOD subunits. There are two histogram peaks towards the edges of each figure which represent a local energy minimum near $\pm\pi$ radians and originate from the extended angles across the the chair conformation of cross-linked DBCOD subunits. The simulation of 3,4'-ODA polymers contains a higher probability density of DBCOD subunits in the chair configuration relative to the simulation of 4,4'-ODA polymers. This agrees with experimental evidence that the 3,4'-ODA polymer material may shift from the twist-boat to the chair conformation at lower energy states than the 4,4'-ODA material. The stochastic conformational discrepancy between the ODA polymers provides a possible explanation at the molecular scale for the differences in thermal contraction observed at the macroscopic scale.

Note that experimental differences in the degree of contraction between the two polymers are extremely small relative to the simulation size, and only the degree of contraction has been recorded experimentally. Therefore, the volumes of the simulation systems cannot be compared directly.

5.4.2 Molecular Disorder Influences Larger-Scale Disorder

The microscopic appearance of the DBCOD-based polymer at scales of several hundred nanometers reveals ordered fiber-like structure for the 4,4'-ODA polymer which

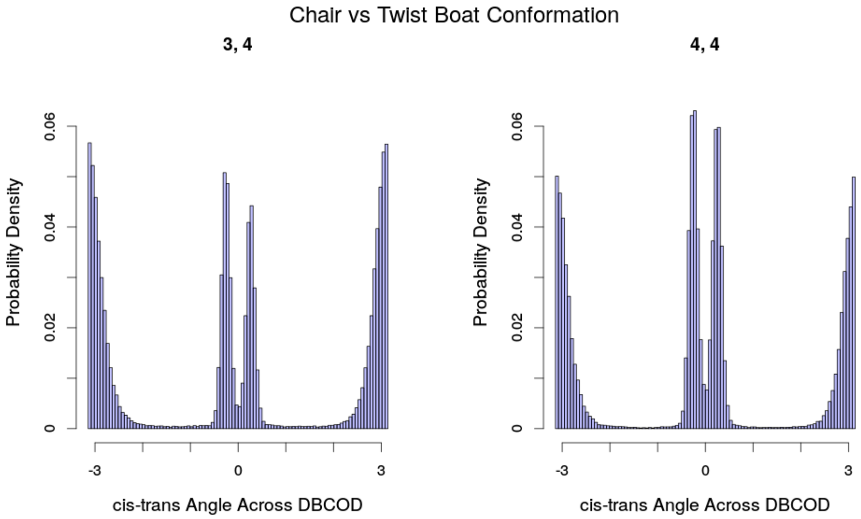


FIGURE 5.6: Histograms of the calculated angles across cross-linked DBCOD subunits illustrate the proportions of DBCOD polymers in the twist-boat and chair conformations. The angles are given in units of radians. The two central peaks correspond to the acute angles of the twist-boat configuration, and the two peaks near $\pm\pi$ radians correspond to the extended angles of the chair configuration. 3,4'-ODA polymer simulation contains a greater proportion of DBCOD subunits in the higher-energy chair conformation associated with thermal contraction relative to the 4,4'-ODA polymer simulation.

does not exist when the 3,4'-ODA polymer is incorporated into the material (**Figure 5.7**). MD simulations of the polymer material are not feasible on this scale, so analysis of the local molecular disorder was used to infer possible explanations.

Since system coordinates for MD simulations are built from 32 independent sheets which are not interconnected through covalent bonds (**Figure 5.3**), the MD simulations provided ensembles of independent structures for the 3,4'-ODA and 4,4'-ODA polymer systems. Several methods of shape and disorder were applied to the structural ensembles.

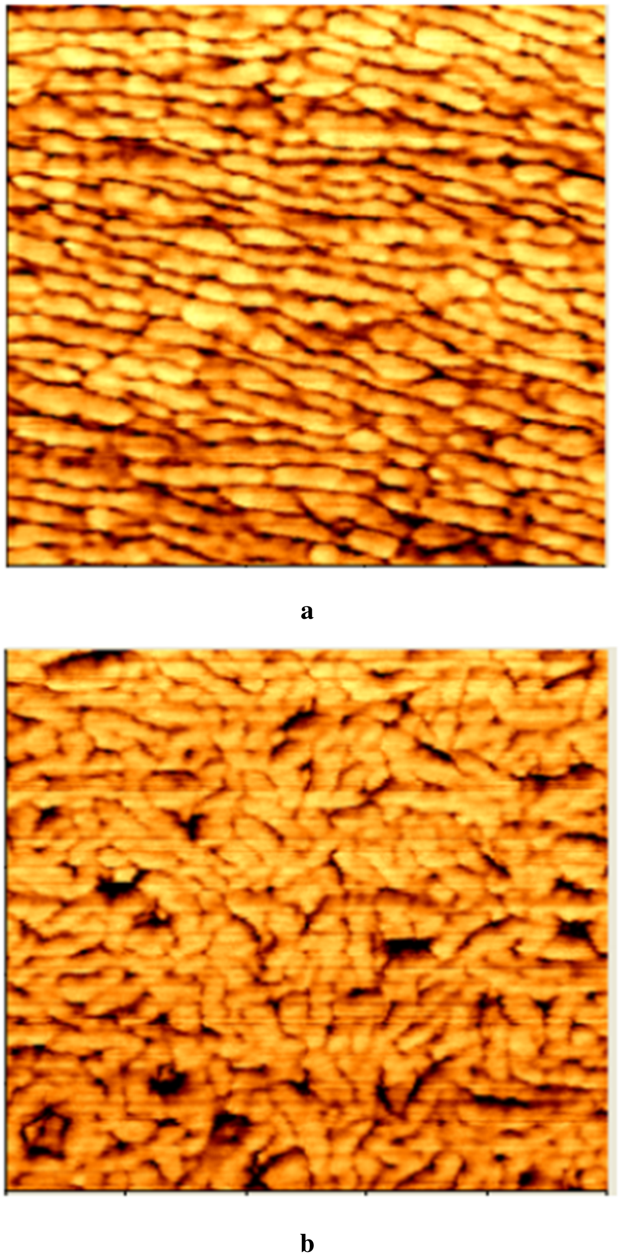
Qualitatively, sheets comprised of the 4,4'-ODA polymer show a preference for elongated bundle-like conformations instead of the irregular globular configurations representative of sheets containing 3,4'-ODA (**Figure 5.8**). To quantify this observation, we employed the shape analysis tool g_shape [18] to compute shape parameters for the ensembles of structures. The shape parameter is normalized such that cylindrical shapes have positive values, oblateness results in negative numbers, and a spherical shape should result in a parameter of approximately zero.

A histogram of the shape parameters for the two ensembles reveals separate frequency peaks for the 3,4'-ODA and 4,4'-ODA polymers (**Figure 5.9**). The peak frequency for the shape parameter of 3,4'-ODA polymer occurs at approximately 0.3 which corresponds to a slightly elongated spherical shape. The 4,4'-ODA polymer's shape parameter peak at approximately 0.9 represents a more cylindrical shape. The distributions of the shape parameters for the two ensembles overlap to some extent, but there is a clear difference between the distributions.

An ordered bundle-like configuration should produce particle density peaks in the radial distribution function (RDF) computed by the g_rdf tool from the Gromacs software library [33–38]. The radial distribution function is often used to quantify ordered structure in solid-state materials [135], and the regular repeats in ordered molecules result in high particle density peaks that occur at the distance between repeats.

An ether oxygen atom exists between the ODA subunits of the synthesized polymer (**Figure 5.1**). Only one atom of this type exists per monomer, so it was the ideal reference atom chosen to calculate RDFs of the two ODA polymer sub-types. RDF plots of ether oxygen atoms for the 4,4'-ODA polymer ensemble shows strong particle density peaks at radial distances of approximately 0.6 and 1.8 nanometers (**Figure A.19**). The RDF plot for the 3,4'-ODA polymer ensemble has much weaker particle density peaks at approximately 0.6 and 1.7 nanometers. It can be inferred from the weaker particle density peaks that the 3,4'-ODA polymer is more disordered.

Peaks in the RDF plot originate from specific structural repeats in the polymer materials. To further examine the observed 0.6 and 1.8 nanometer peaks from the RDF plots, we theorized that the most uniform distances between ether oxygen atoms in the



a

b

FIGURE 5.7: At microscopic scales which are still much larger than what is feasible to explore with MD simulations, (a) the 4,4'-ODA sub-type of the DBCOD-based polymer exists in an ordered structure with a fiber-like appearance. (b) The 3,4'-ODA polymer exists in a more disordered environment without obvious fiber structures. Mixtures of the two polymer sub-types allow the amount of structural disorder to be tuned by controlling the ratio of polymer sub-types in the mixture. [124].

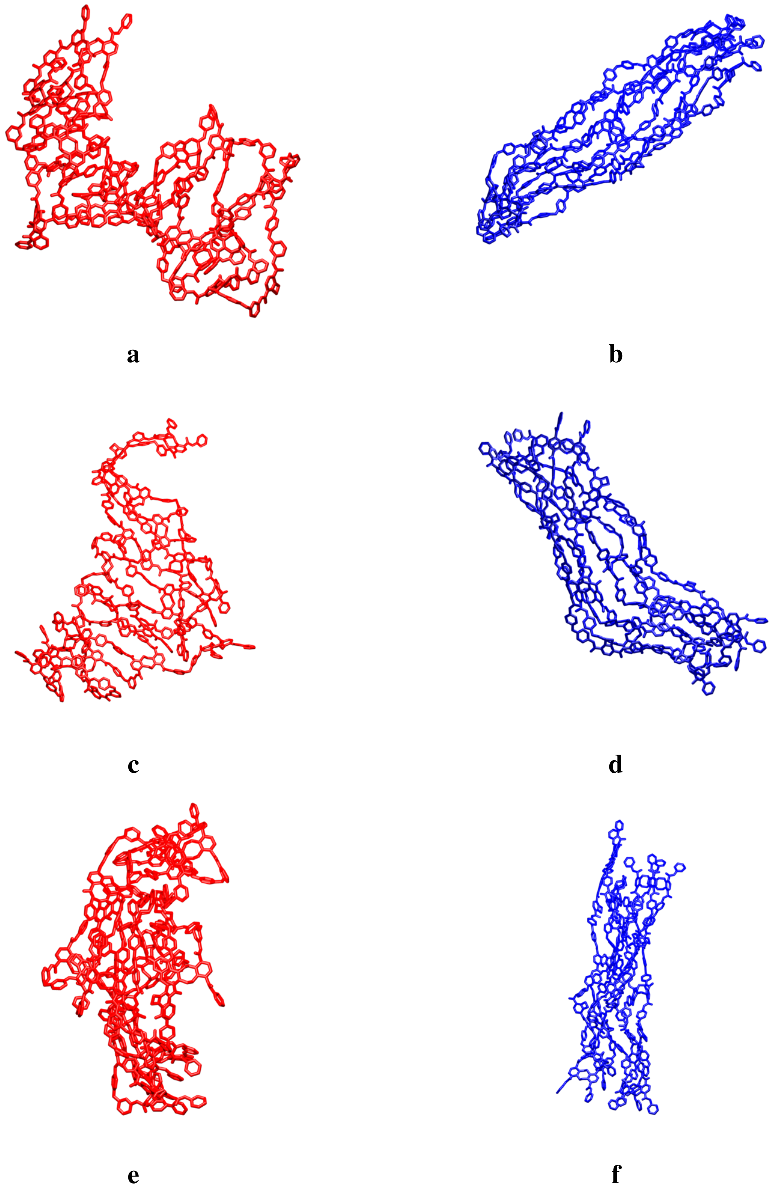


FIGURE 5.8: By sampling the ensemble of independent polymer conformations, the polymers containing 3,4'-ODA **(a,c,e)** have a preference toward irregular and globular conformations while the polymers with 4,4'-ODA **(b,d,f)** are primarily composed of ordered, bundle-like shapes. The shape of 4,4'-ODA polymer bundles may explain the preference for the ordered fiber-like structure appearing in the bulk material seen in **Figure 5.7**.

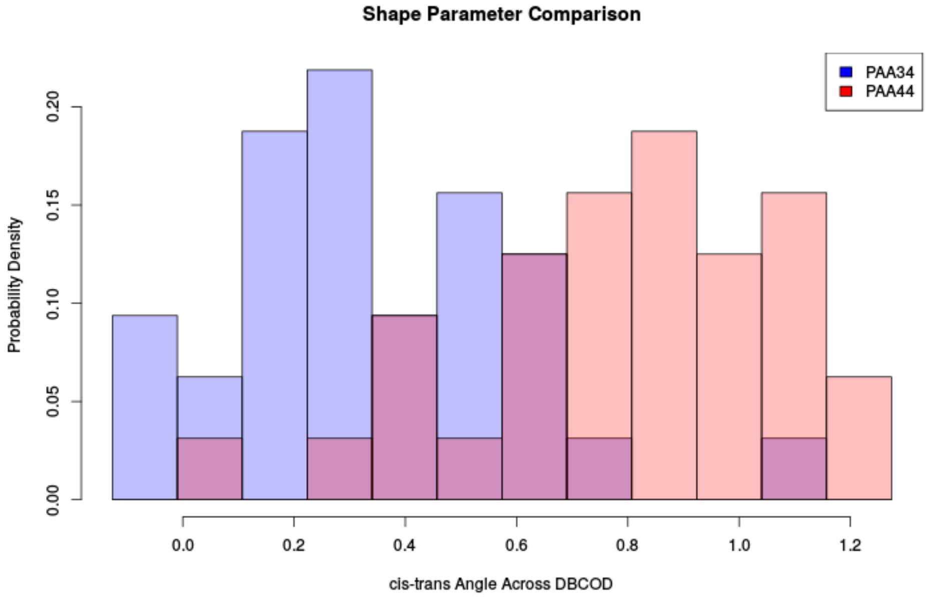


FIGURE 5.9: After partial vacuum annealing in simulation, 4,4'-ODA polymer sheets (red) tended to prefer extended bundle-like conformations rather than the more random, globular conformations favored by the 3,4'-ODA sheets (blue). The two simulations of DBCOD-based polymer each contained 32 independent structures. Using the analysis tool g_shape [18], shape parameters were calculated for the two structure ensembles. Histograms of the shape parameters have peaks at approximately 0.3 and 0.9 for the 3,4'-ODA and 4,4'-ODA polymer sheets respectively. Overlapping histogram bins are purple. Shape parameter values close to zero correspond with spherical shapes while positive shape parameters represent prolate spheroids. [124]

polymer sheets would be across DBCOD cross-links (**Figure 5.10.a Distance A**) and along the polymer chain between monomer subunits (**Figure 5.10.a Distance B**).

A script for the R programming environment [62] was applied to the final configuration of the two MD simulations to calculate distances between all neighboring ether oxygen atoms and compile them into histograms. The histograms peaks in **Figure 5.10** (b-e) occur at the same distances as the RDF plot peaks in **Figure A.19** which appears to confirm that the RDF peaks are derive from these distances. The stronger peaks in both the RDF plots and distance histograms for the 4,4'-ODA polymer verifies its preference for a more ordered ensemble of conformations than the 3,4'-ODA polymer.

5.5 Conclusions

Experimental results indicate that replacement of the 4,4'-ODA subunit of the synthesized DBCOD-based polymer with a 3,4'-ODA subunit reduces crystallinity and induces morphological disorder in the bulk material [124]. It was observed that the more amorphous environment created by the polymer containing 3,4'-ODA facilitated the DBCOD conversion and doubled the thermal contraction stress at energy input levels near room temperature [124]. This change results in more optimized control over its conformation reconfiguration at low energy levels. The synthesized material has thermal sensing properties due to its expansion and contraction at ambient temperature. The DBCOD-based polymer can be used to mechanically alter biological events through induced folding by the application of low energy infrared photons which are not toxic to biological materials and tissues.

MD simulations using a customized force field topology provided a possible mechanism by which the slight difference in bonds between the 3,4'-ODA and 4,4'-ODA induces a preference for more disordered structural conformations in the 3,4'-ODA polymer. The presence of 3,4'-ODA in the DBCOD-based polymer creates structural ensembles with a wider variety of configurations and a preference towards a globular shape over the bundle-like shapes prevalent in polymer containing 4,4'-ODA subunits. Though not simulated at the same scales of size as the bulk material observed under microscope, the ensembles of structures from the MD simulations provide a possible explanation for the observed divergence between the ordered, fiber-like appearance of the 4,4'-ODA polymer and the disordered, amorphous properties of the 3,4'-ODA polymer.

5.6 Acknowledgements

The work presented in this chapter was a collaboration with Dr. Xingyuan Shen and Dr. Jennifer Lu. The project was headed by Dr. Xingyuan Shen, and he performed the experimental portions of the work. Dr. Michael Colvin provided insights on the formatting

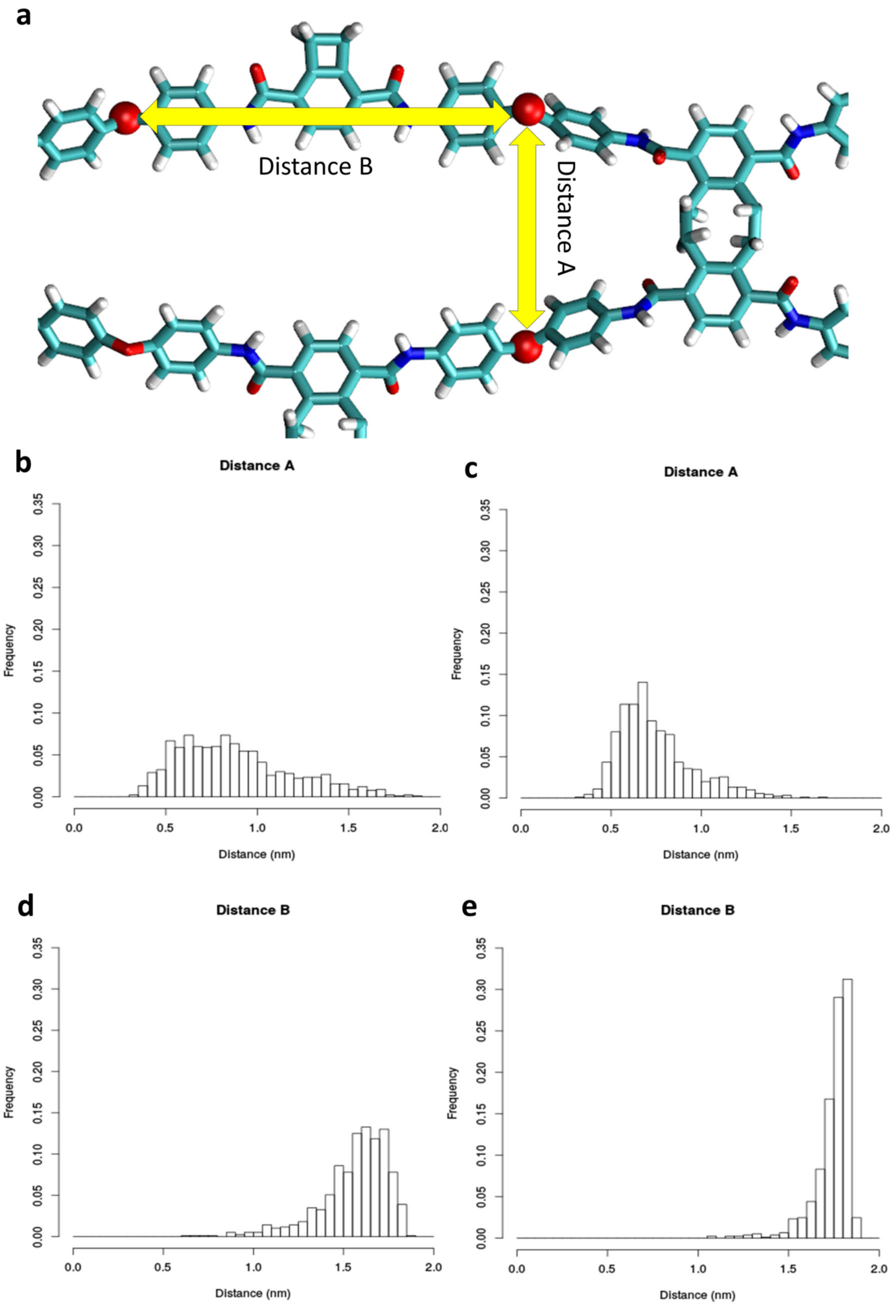


FIGURE 5.10: (a) The two peaks from the RDF plots in **Figure A.19** originate from Distance A and Distance B. Histograms of Distance A (b & c) and Distance B (d & e) between neighboring ether oxygen atoms contain peaks at the same distances as the RDF plots: approximately 0.6 and 1.7-1.8 nm. The 4,4'-ODA polymer histograms (c & e) have stronger peaks than the 3,4'-ODA histograms (b & d) suggesting the latter is more disordered. [124].

of Gromacs topology files. Dr. Joshua Phillips provided the original source code used to create the g_shape tool though it was heavily modified and updated for this work. My role was to complete and analyze the MD simulation portion of the project. An article focusing more heavily on the experimental results has been published [124]. This chapter expands on the computational aspects of the project.

Chapter 6

Reference For g_isd

6.1 Implementation Details For ISDMs

The analysis tool *g_isd* allows the user to choose between a variety of options to quantify the inter-structure distance (ISD) between compared structures. The individual options are referred to as measures of inter-structure distances (ISDMs). This section describes the implementation details for the available ISDMs as well as important user options.

6.1.1 Backbone Angles

The *g_isd* option (-ang) calculates the ISD as the root-mean-square of the differences between the backbone angles of the two structures being compared. The reference structure angle θ_{R_i} and the comparison structure angle θ_{S_i} is calculated for each contiguous set of three C_α atoms; therefore, for n amino acids, there are $n - 2$ total backbone angles. Each backbone angle is calculated using the gmx_angle function from the Gromacs library [33] with the two vectors $\overrightarrow{C_{\alpha_i}C_{\alpha_{i-1}}}$ and $\overrightarrow{C_{\alpha_i}C_{\alpha_{i+1}}}$ as inputs. This algorithm calculates θ_i using the equation:

$$\theta_i \quad = \quad \arctan \frac{\|\overrightarrow{C_{\alpha_i}C_{\alpha_{i-1}}} \times \overrightarrow{C_{\alpha_i}C_{\alpha_{i+1}}}\|}{\overrightarrow{C_{\alpha_i}C_{\alpha_{i-1}}} \cdot \overrightarrow{C_{\alpha_i}C_{\alpha_{i+1}}}} \tag{6.1}$$

Using this measure, the ISD is defined as the root-mean-square of the differences between backbone angles rescaled to return a value between zero for $\theta_{R_i} = \theta_{S_i}$ and a maximum of one. The -ang ISDM works with the -trig option which replaces the sum of the squared differences of the angles with the sum of the cosine differences of the angles.

$$ISD_{ang} = \frac{1}{\pi}\sqrt{\frac{1}{n-2}\sum_{i=2}^{n-1}(\theta_{R_i} - \theta_{S_i})^2} \tag{6.2}$$

6.1.2 Backbone Dihedrals

The *g_isd* option (-dih) calculates the ISD as the root-mean-square of the differences between the dihedrals of the two structures being compared. The backbone dihedral angle made by each set of four C_α atoms is determined by choosing two vectors normal to the planes formed by the two contiguous sets of three C_α atoms. The two normal vectors are calculated by taking the cross products of the atom coordinates $\overrightarrow{V_1} = \overrightarrow{C_{\alpha_i}C_{\alpha_{i-1}}} \times \overrightarrow{C_{\alpha_i}C_{\alpha_{i+1}}}$ and $\overrightarrow{V_2} = \overrightarrow{C_{\alpha_i}C_{\alpha_{i+1}}} \times \overrightarrow{C_{\alpha_{i+2}}C_{\alpha_{i+1}}}$. The dihedral angle θ_i between the resultant vectors is calculated using the Gromacs library function gmx_angle [33]. The dihedral angle θ_i is multiplied by the sign of $\overrightarrow{C_{\alpha_i}C_{\alpha_{i-1}}} \cdot \overrightarrow{V_2}$ which gives a consistent dihedral angle measure within a range of $[0, 2\pi]$.

Since θ_{R_i} and θ_{S_i} are both bounded by $[-\pi, \pi]$, the difference $\Delta\theta_i$ has a range of $[-2\pi, 2\pi]$. An adjustment is made by adding 2π for $\Delta\theta_i < -2\pi$ and subtracting 2π for $\Delta\theta_i > 2\pi$. Using this measure, the ISD is defined as the root-mean-square of the differences between the $n-3$ backbone dihedral angles. The -dih ISDM works with the -trig option which replaces the sum of the squared differences of the angles with the sum of the cosine differences of the angles.

$$ISD_{dih} = \frac{1}{2\pi}\sqrt{\frac{1}{n-3}\sum_{i=2}^{n-2}(\Delta\theta_i)^2} \tag{6.3}$$

6.1.3 Backbone Angles And Dihedrals

The backbone angles and dihedrals option (-angdih) defines the ISD as the geometric mean of the backbone angles and backbone dihedrals options. The geometric mean is used to guarantee equal contribution of both the angles and dihedrals even though the magnitudes of the two measures differ. For all of the options -ang, -dih, and -angdih, the final sum of root-mean-square differences of angles can be replaced with the cosine of the difference of angles by additionally using the -trig isdm option.

6.1.4 Phi–Psi Angles

The ϕ and ψ angles are calculated in the same way as backbone dihedrals described previously. However, all backbone atoms are used, and n amino acids produce $n-1$ angles of each type. This option defines the ISD as the root-mean-square differences between all $2n-2$ pairs of angles from the reference and comparison structures. The option -phipsi chooses this calculation and the option -trig isdm can be used in combination to replace the sum with the cosine of the differences of angles.

6.1.5 RMSD And Scaled RMSD

The RMSD implementation in *g_isd* (-rmsd) uses the Gromacs library functions reset_x and do_fit to perform the molecular alignment [33], but the distance sums are carried out with higher default precision. All RMSD results presented here use C_α atoms for alignment and distance calculations.

The -scaled option divides the standard RMSD by a scaling factor to provide a size-independent measure of ISD. The purpose is similar to previously described methods [136]. A similar mirrored approach to size-independent RMSD scaling can be chosen with the -mir option in *g_isd*. This is a computationally inexpensive approach where the reference structure is mirrored by multiplying by the negative identity matrix. However, this results in a reproduction of the structure which does not feature correct chirality. The scaling factor computed using this approach is too large.

The preferred implementation takes advantage of modern processing power to approximate the ISD of two unaligned molecules with the same size. The scaling factor is calculated by *g_isd* using the method described in the **Grid Search Rotation RMSD** section. The scaled RMSD calculation can be chosen in *g_isd* by using the options -rmsd and -scaled together.

6.1.6 Distance RMSD

Distance RMSD (-drms) is also referred to in some software packages as RMS Dist. For each of the two structures, a distance matrix is computed containing the Euclidean distances between every pair of atoms. The distance RMSD is based on the differences of distances between the two structures. Final ISD is calculated using **Equation 6.4**. When combined with the -scaled option, the ISD is divided by $2 \times R_g$ where R_g is the radius of gyration for the larger structure.

$$ISD_{drms} = \frac{1}{n}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\|\overrightarrow{R_iR_j}\| - \|\overrightarrow{S_iS_j}\|)^2} \tag{6.4}$$

6.1.7 Backbone Dihedral RMSD

The backbone dihedral RMSD option (-rmsdih) is a hybrid that uses internal coordinates but calculates a Euclidean distance with units of nanometers. The internal coordinates of the backbone angles and dihedrals are converted into a Euclidean measurement of distance by locally aligning each set of three consecutive C_α atoms and calculating the distance between a fourth C_α atom. We use a specific type of fit to vastly simplify alignment.

We use three pairs of C_α atoms for the alignment: $C_{\alpha_{i-2}}$, $C_{\alpha_{i-1}}$, and C_{α_i}. The RMSD calculation uses the distance between a fourth pair of C_α atoms, $C_{\alpha_{i+1}}$. A true alignment step is unnecessary if the first three atoms are instead aligned to a new coordinate system. The C_{α_i} atom is set as the origin in this coordinate system and the vector $V_z = \overrightarrow{C_{\alpha_{i-1}}C_{\alpha_i}}$ is aligned to the z axis. The first three C_α atoms create a plane, and the vector normal to the plane is aligned to the y axis.

In this new coordinate system, the spherical coordinates of the fourth C_α atom can be calculated using the methods described in the **backbone angles** and **backbone dihedrals** sections. The spherical coordinates of the fourth C_α atom can be calculated using the following equations:

$$\begin{aligned} r &= \|\overrightarrow{C_{\alpha_i}C_{\alpha_{i+1}}}\| \\ \phi &= \pi - ang(C_{\alpha_{i-1}}, C_{\alpha_i}, C_{\alpha_{i+1}}) \\ \theta &= dih(C_{\alpha_{i-2}}, C_{\alpha_{i-1}}, C_{\alpha_i}, C_{\alpha_{i+1}}) \end{aligned} \tag{6.5}$$

The *g_isd* tool converts the spherical coordinates to Cartesian and calculates the Euclidean distance between the $C_{\alpha_{i+1}}$ atoms of the two structures. Since the alignment employs coordinates from only one side of the atom used as the distance reference, the calculation is carried out originating from both the N-terminal and C-terminal ends. Each sweep of the coordinates computes $n-3$ distances due to the number of C_α atoms necessary for alignments. The final ISD is defined as root-mean-square of the $2n-6$ total calculated distances.

6.1.8 Backbone Angles And Position Correlation

The sample Pearson's correlation coefficient, r, is calculated between the cosine of the angles of comparison and reference structures (-acor). However, the ISD is defined as the mean over all backbone angles of a rescaled version of the correlation coefficient. All values less than or equal to zero are set to a value of 1.0 and values greater than zero are rescaled to $1.0 - r$.

The position correlation option uses the atomic coordinates on each axis to calculate three Pearson's correlation coefficients over all atoms (-pcor). After the comparison structure is aligned to the reference structure, the correlation coefficient is calculated for the x, y, and z components of the coordinates independently. The three correlations are combined by taking the arithmetic mean, and the result is rescaled in the same way as the backbone angle coefficient. This algorithm shares some aspects with a previously described method of structural comparison [137].

6.1.9 Elastic Shape Analysis

Our implementation of the elastic shape analysis method is a slightly modified port from Matlab to C of a previously described algorithm [40]. Briefly, elastic shape analysis resamples the compared structures as curves in space and attempts to quantify the effort of warping one curve onto the other. The algorithm is selected with the -esa option. The number of points used during curve resampling can be controlled by the user with the -esasamples option, but a sensible value is automatically chosen based on the number of C_α atoms in the input. The computational cost of elastic shape analysis scales proportionally to the number of samples chosen. This algorithm is computationally expensive making it impractical to run on trajectories with a large number of frames. The analysis requires frames to be sparsely sub-sampled.

6.1.10 MAMMOTH

We modified the source code of MAMMOTH, a structural comparison package [39], to interface with our tools. MAMMOTH was designed to compare different variants of related molecules and focuses on similarities in local secondary structure by comparing local sequences of six amino acids at a time. MAMMOTH approximates the probability of random structure sharing the amount of structural similarity observed and outputs a Z-score. The ISDM option converts the Z-score output to a p-value, and the ISD is defined as the inverted p-value: $ISD = 1 - P$.

6.1.11 Random Rotation RMSD

The RMSD between the comparison structure and a rotated reference structure is available as a negative control. The *g_isd* tool option -rrot implements a randomly rotated reference structure. After the comparison and reference structures are centered, the coordinates of the reference structure are multiplied by a randomly generated three-dimensional rotation matrix. The output is the computed RMSD of the resultant rotations. This is not the preferred method of computing the scaling factor due to inherent limitations of the random rotations. The randomized RMSD is subject to significant noise, and a good sampling of the rotational space requires many independent rotations. The number of rotations may be set by the user with the -nrotations option, but the tool's default value is 500. The total processing cost scales proportionally with the number of rotations, and a balanced sampling of the rotational space is not guaranteed.

6.1.12 Grid Search Rotation RMSD

The grid search rotation RMSD is conceptually similar to the **Random Rotation RMSD**. However, it guarantees a balanced sampling of the rotational space and therefore does not require as many rotations. This implementation first fits the compared structures, but then applies a rotation to one of the structures. By default, the algorithm that generates the rotations uses a grid search of 8 angles along each of x, y, and z axes for a total of $8^3 = 512$ possible rotations. The grid search density for rotation angles can be controlled by the user with the -griddensity option. For each set of three rotations, a single transformation matrix is calculated through matrix multiplication. The list of rotations are then checked for uniqueness. The default settings result in 208 unique rotation matrices from the original 512.

The algorithm to produce unique rotation matrices using a grid search of possible rotations was tested within the R programming environment [62] with a script applied to arbitrary unit vectors. After the 208 unique rotations, the sum of rotated vectors cancel out to the available numerical precision. The rotation matrices are applied to the reference structure to approximate an unaligned RMSD measurement. The results are averaged to estimate the RMSD of two structures which are unrelated but similar in size to the pair of structures being compared. Compared to the randomly rotated RMSD, the grid search rotation RMSD implementation is balanced, less prone to noise from random rotations, and requires less processing time by default. The scaling factor can be calculated separately from RMSD using the -grot option in the *g_isd* tool.

6.1.13 End-to-end Distance

The distance between terminal C_α atoms is calculated for the reference and comparison structure, and the difference of these distances is used to define the ISD. The scaled version divides this result by $2 \times R_g$, where R_g is the radius of gyration for the larger structure. The *g_isd* tool implements this option with the option -e2e and scaling is applied with the addition of the -scaled option.

6.1.14 Radius Of Gyration

The option -rg simply calculates the radii of gyration for the two structures being compared. The ISD is defined as the difference between the two values. The -scaled option divides the distance by R_g, the radius of gyration of the larger structure, to provide a size-independent distance. The scaled version of the computed ISD should be considered an approximate fraction of the maximum possible size difference between the two structures.

6.2 Acknowledgements

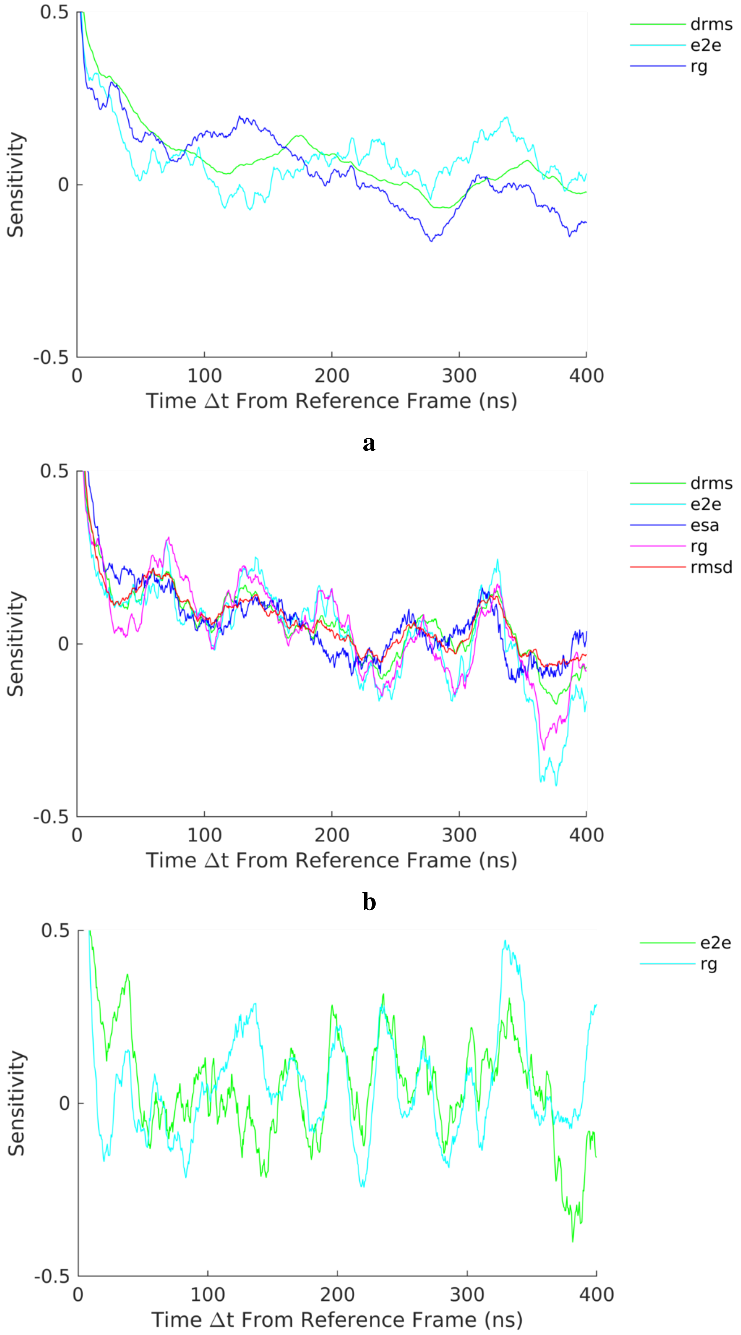
The elastic shape analysis was originally written in a combination of MATLAB and C code. The algorithm was ported into C by my collaborator, Robert Wang.

Appendix A

Supplementary Figures

A.1 Chapter 1 Supplement

The figures here include plots of time decorrelation, sensitivity, and ability to distinguish systems by relative disorder. Supplementary plots cover ISDMs which were poorly behaved and comparisons between sets of ISDMs with similar outputs. Differences in the method of summing the ISDM output for internal coordinate representations of proteins did not significantly affect any analysis tests. The simpler root-mean-square difference and the more complex mean of cosines methods gave similar results.

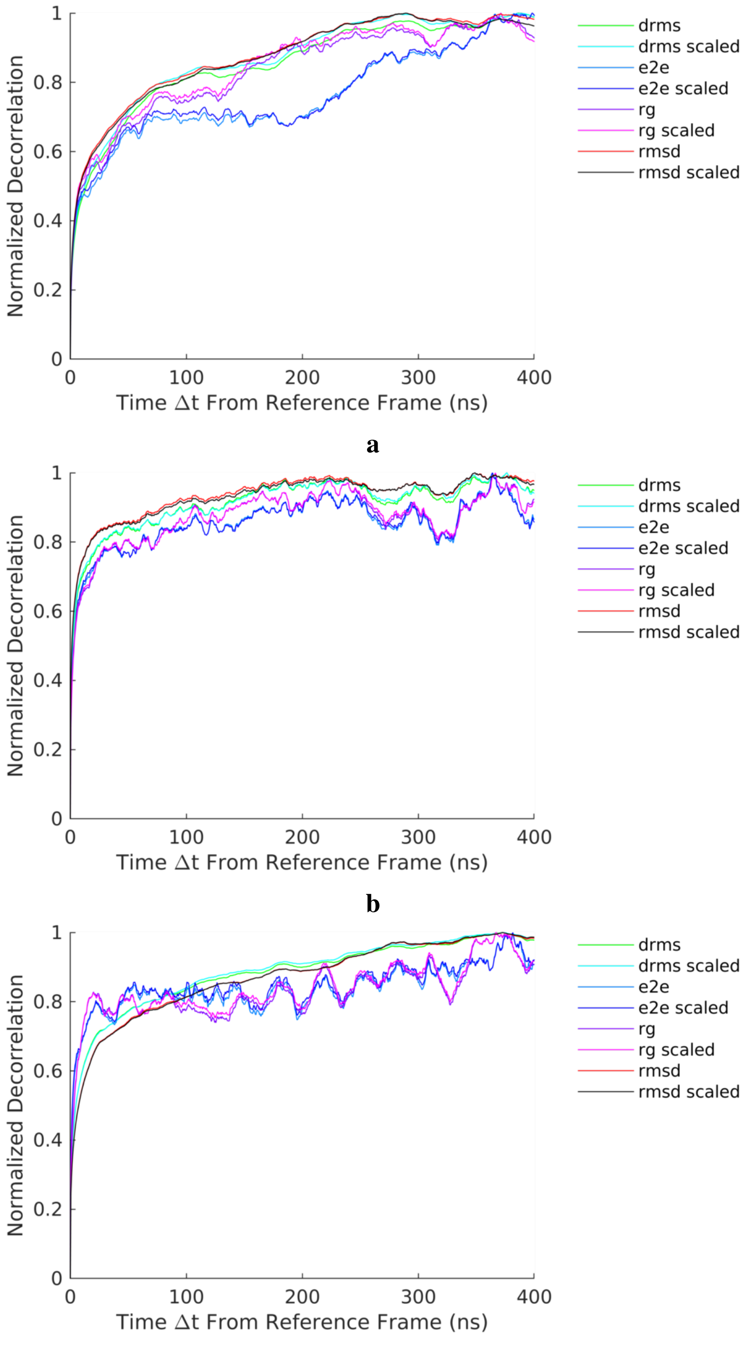


a

b

c

FIGURE A.1: Presents ISDMs which show significantly negative results when the sensitivity analysis is applied. Supplementary figure to **Figure 1.6**.

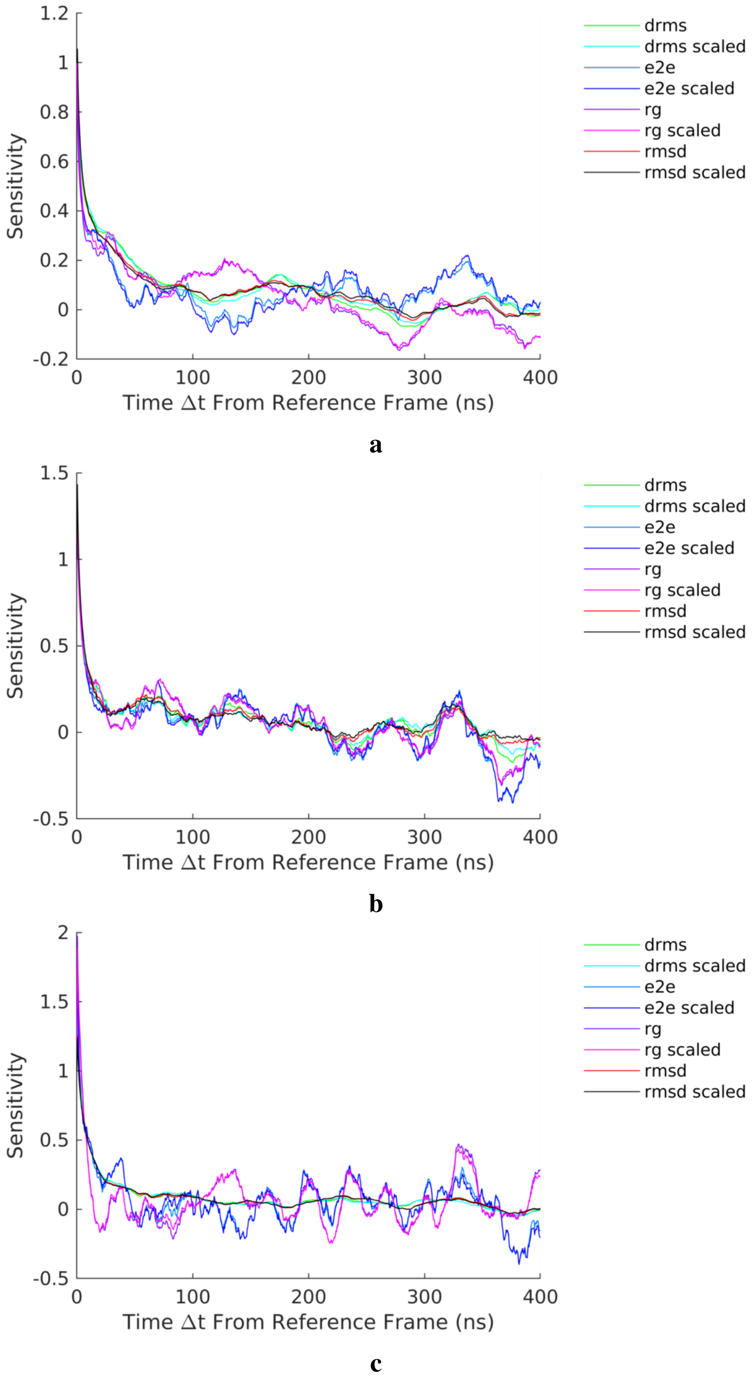


a

b

c

FIGURE A.2: Rescaling has no significant effect on the decorrelation analysis.

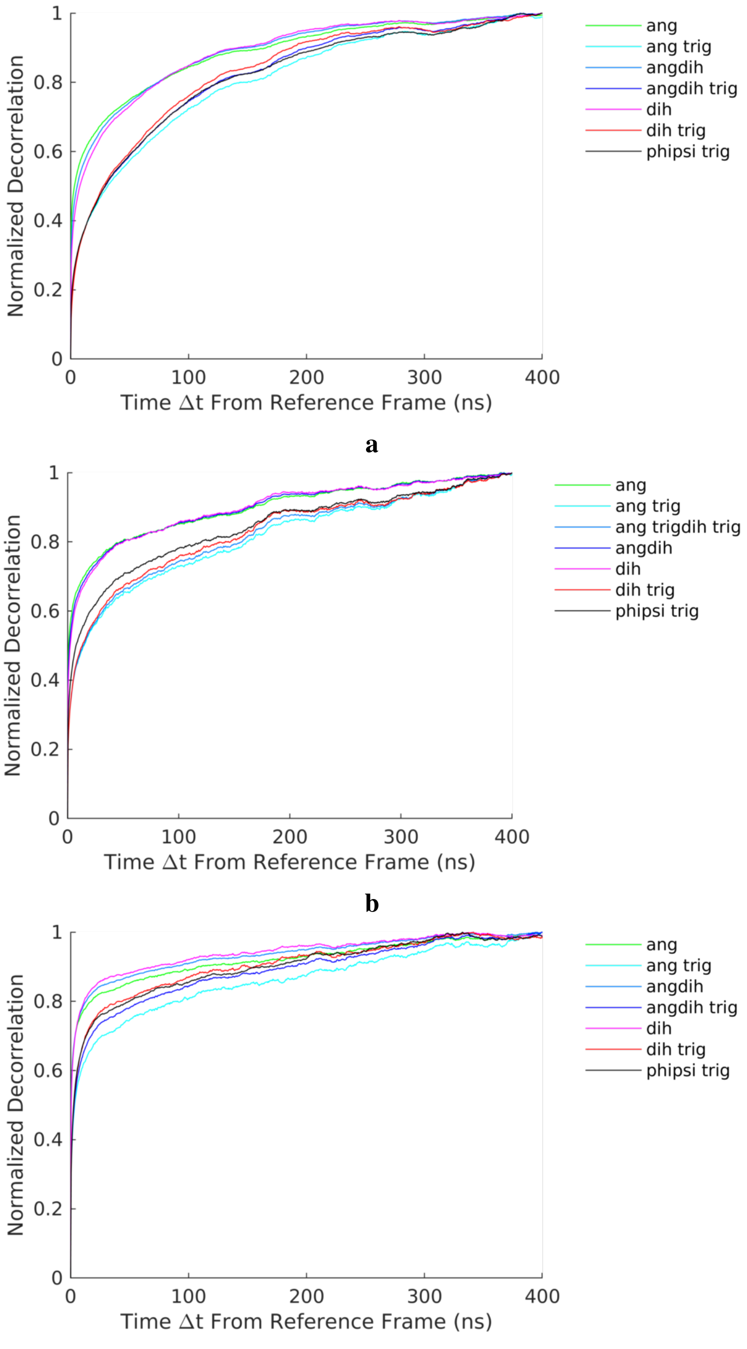


a

b

c

FIGURE A.3: Rescaling has no significant effect on the sensitivity analysis.

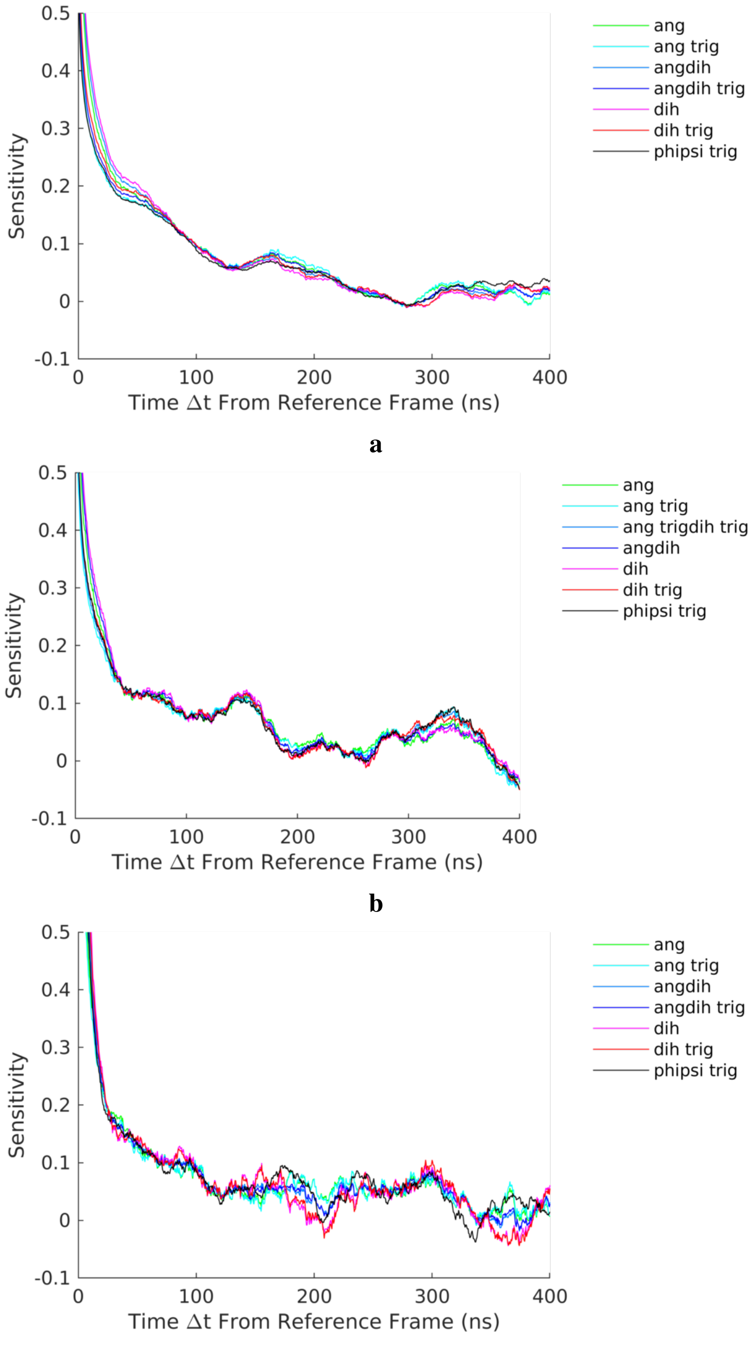


a

b

c

FIGURE A.4: The method of compiling sums has no significant effect on the decorrelation analysis.



a

b

c

FIGURE A.5: The method of compiling sums has no significant effect on the sensitivity analysis.

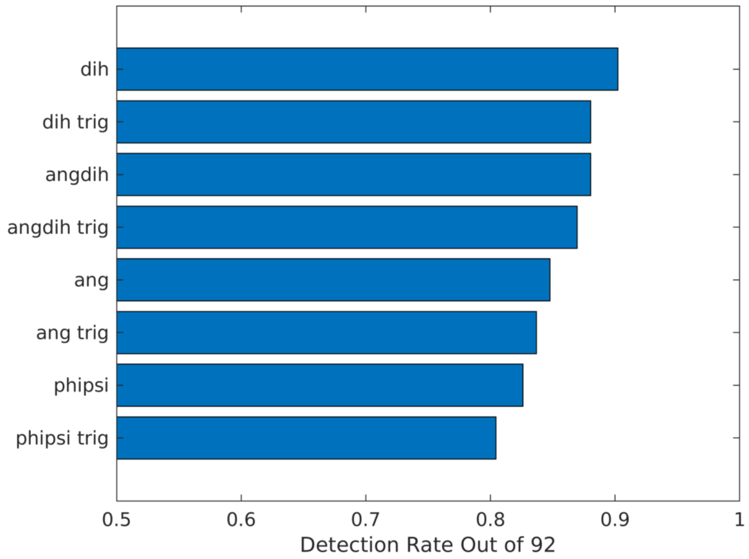
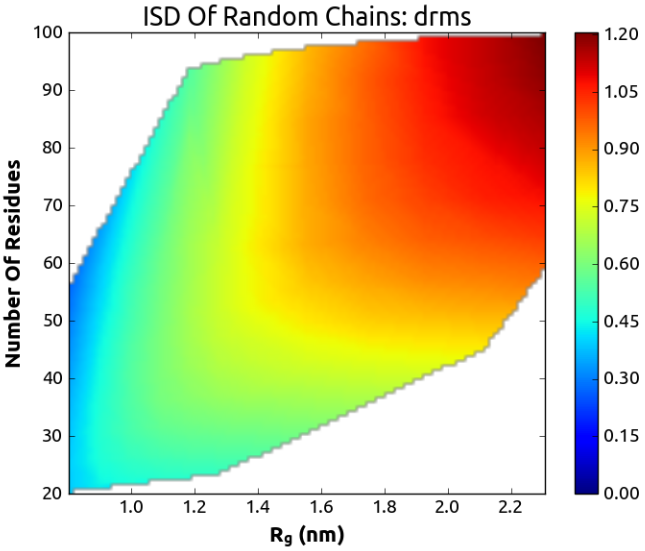


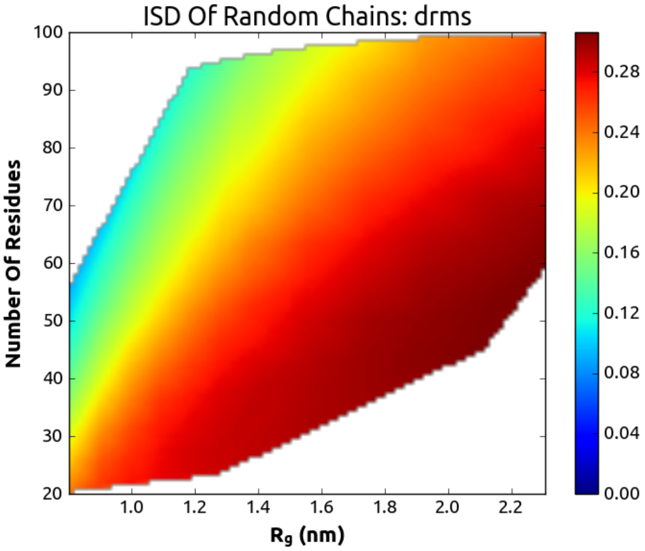
FIGURE A.6: The method of compiling sums has no significant effect on the ability to distinguish proteins based on quantified disorder.

A.2 Chapter 2 Supplement

Entropic chains were generated using a self-avoiding random walk model. Ensembles of chains sampled the likely conformational phase space of possible volumetric sizes and sequence lengths. Chains that are either too compact or too extended would be unphysical for realistic entropic chains. Size-independent ISDMs will compute approximately the same mean ISD regardless of size or sequence length. Size-dependent ISDMs will relate to these characteristics and therefore do not serve as a good universal measure of disorder. A large proteins may erroneously appear to be more disordered than a smaller protein using a poorly behaved ISDM.



a



b

FIGURE A.7: (a) The dependence of ISD on size and sequence length is significant for the drms ISDM. Since proteins of different sizes cannot be compared directly, drms is not a suitable candidate for a universal measure of disorder. (b) Scaling improves the size-independence properties of drms significantly; however, conformations with relatively small R_g still record smaller ISD.

A.2.1 Phase Space Up To 100 Residues

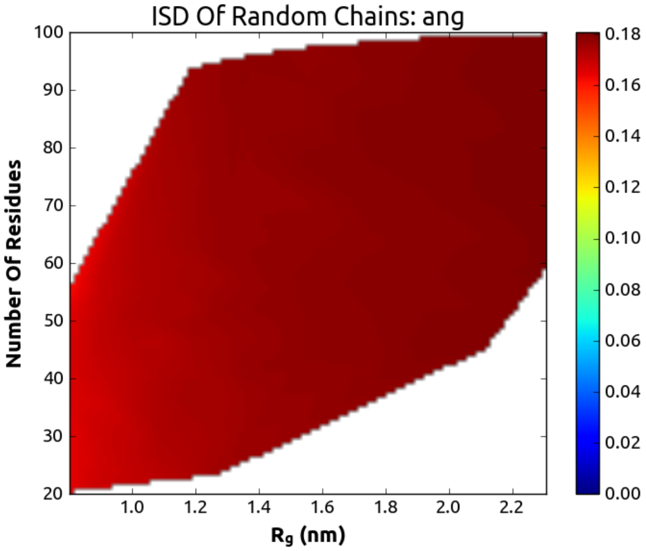


FIGURE A.8: Protein size and sequence length dependence.

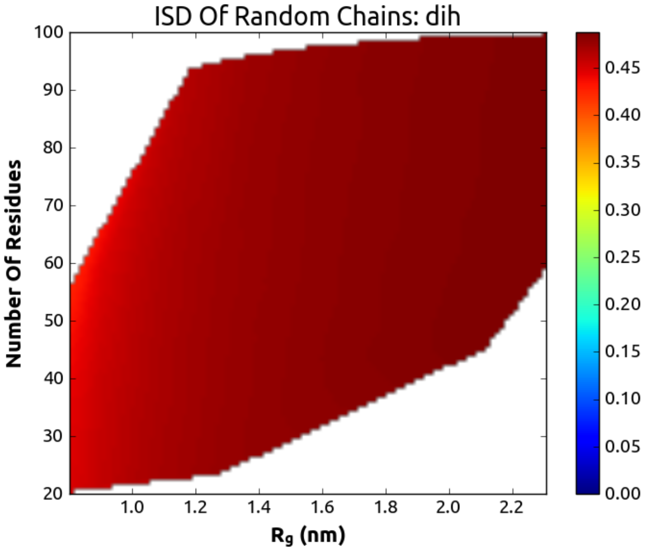


FIGURE A.9: Protein size and sequence length dependence.

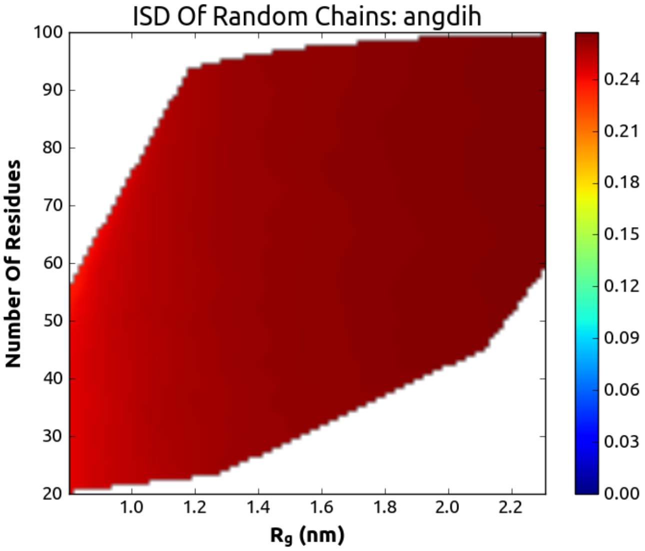


FIGURE A.10: Protein size and sequence length dependence.

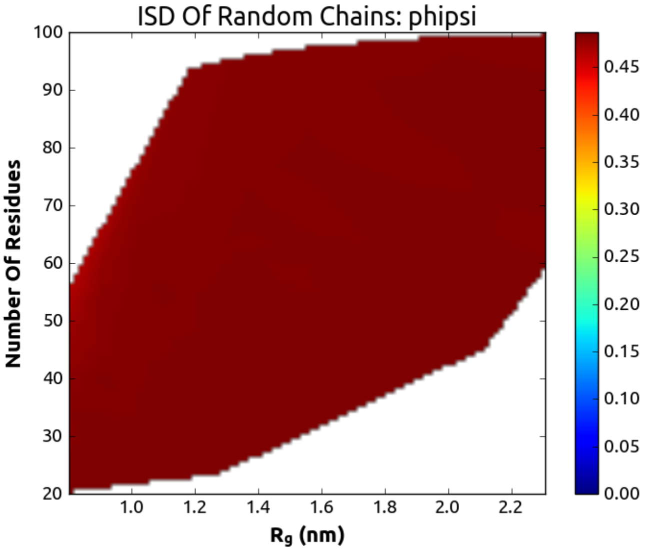


FIGURE A.11: Protein size and sequence length dependence.

A.2.2 Phase Space Up To 400 Residues

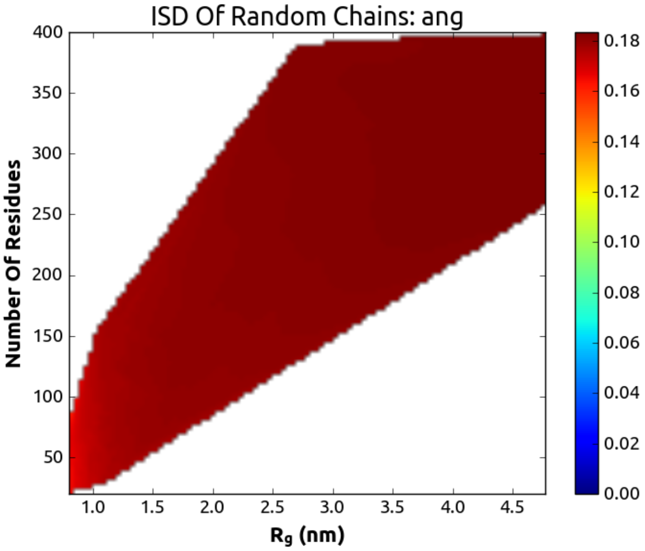


FIGURE A.12: Protein size and sequence length dependence.

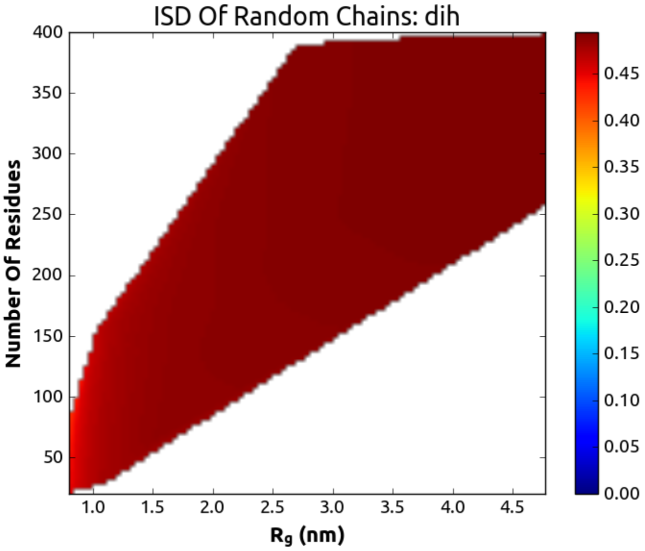


FIGURE A.13: Protein size and sequence length dependence.

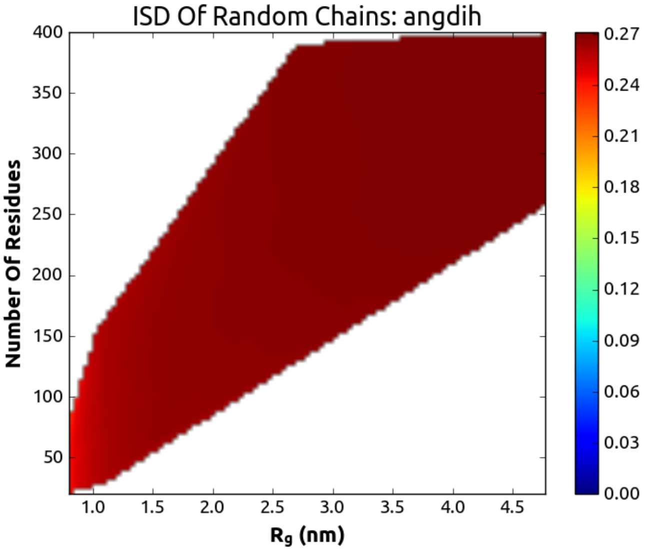


FIGURE A.14: Protein size and sequence length dependence.

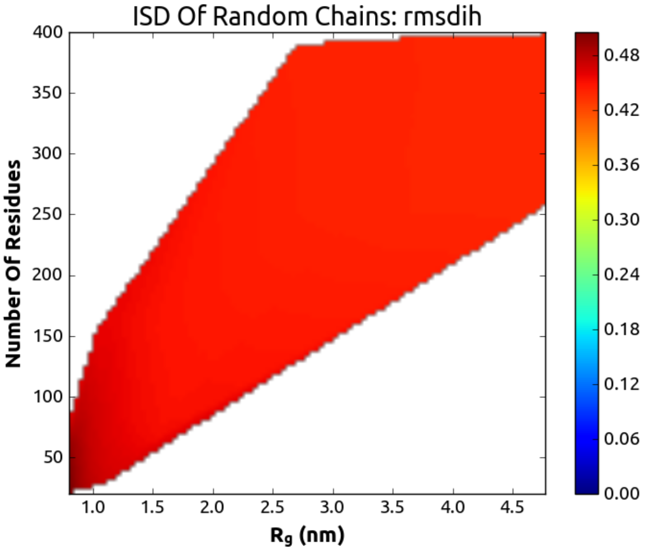


FIGURE A.15: Protein size and sequence length dependence.

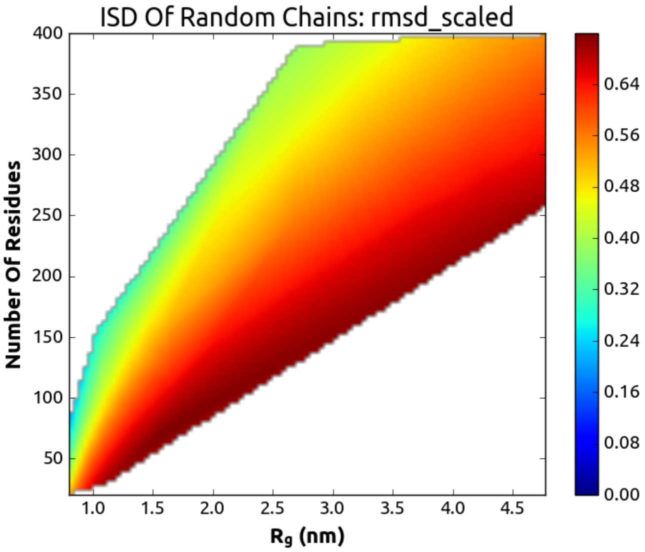


FIGURE A.16: Protein size and sequence length dependence.

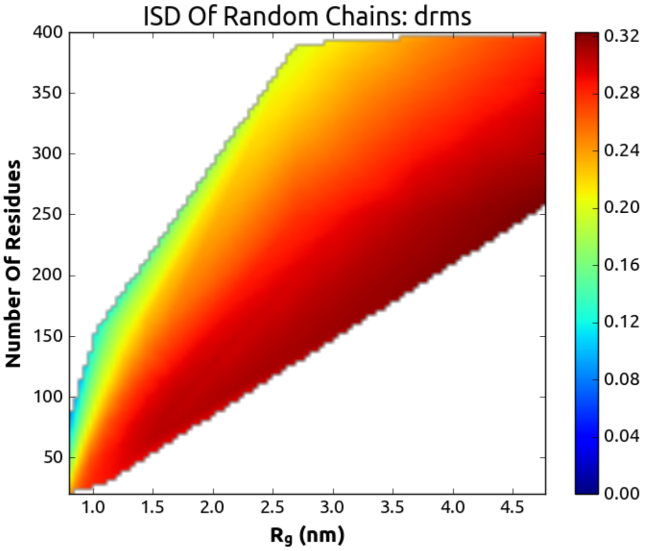


FIGURE A.17: Protein size and sequence length dependence.

A.3 Chapter 5 Supplement

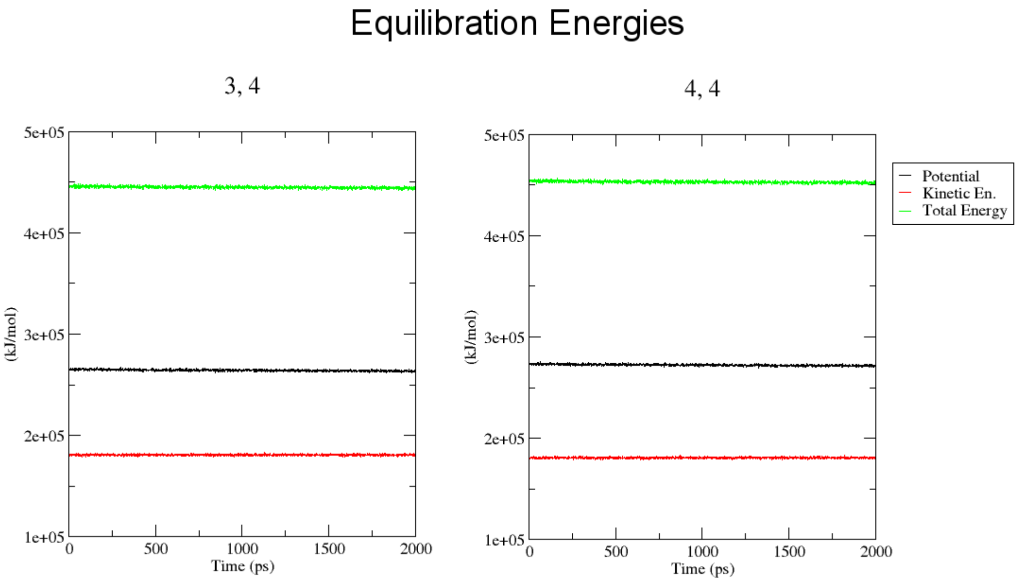


FIGURE A.18: Simulations of the DBCOD sub-types, "3,4'-ODA" and "4,4'-ODA", were initialized under low density conditions to perform partial vacuum annealing. The system was allowed to collapse under atmospheric pressure with NPT setting. A final equilibration was performed to verify that the system was able to reach a constant energy. [124].

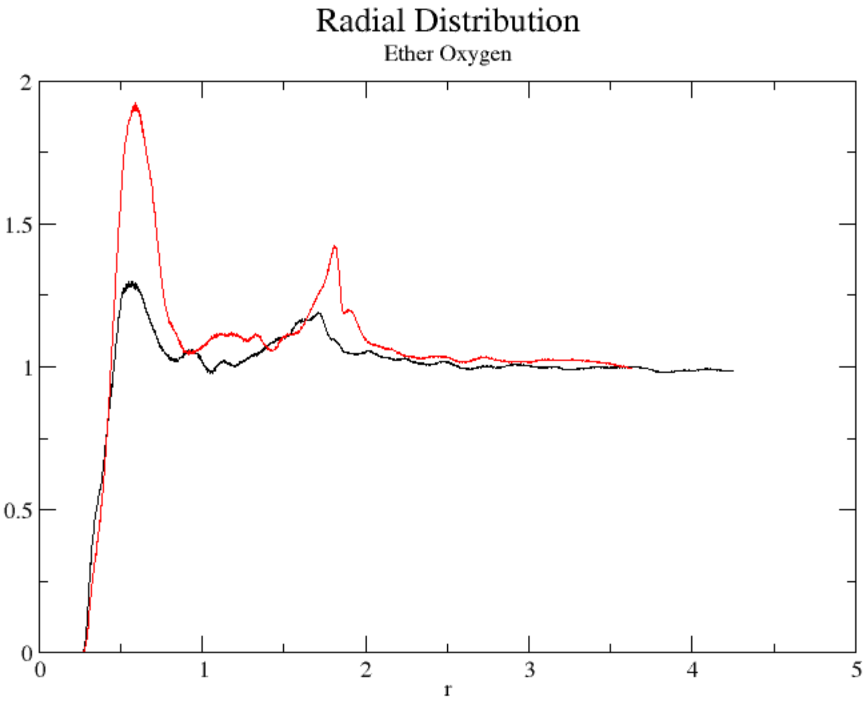


FIGURE A.19: The radial distribution function (RDF) as computed by the g_rdf tool. The RDF of the 4,4'-ODA polymer has normalized particle density peaks at distances of approximately 0.6 nanometers and 1.8 nanometers (red). The 3,4'-ODA polymer RDF has much weaker peaks at similar distances (black). Greater disorder in the 3-4'-ODA polymer sheets can be inferred from this result.

Bibliography

1. Andreeva, A. *et al.* SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research* **32,** D226–D229 (2004).
2. Koehl, P. Protein structure similarities. *Current Opinion in Structural Biology* **11,** 348 –353. ISSN: 0959-440X (2001).
3. Pearl, F. *et al.* The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Research* **33,** D247–D251 (2005).
4. Wallin, S., Farwer, J. & Bastolla, U. Testing similarity measures with continuous and discrete protein models. *Proteins: Structure, Function, and Bioinformatics* **50,** 144–157. ISSN: 1097-0134 (2003).
5. Chema, D. & Becker, O. M. A Method for Correlations Analysis of Coordinates: Applications for Molecular Conformations. *Journal of Chemical Information and Computer Sciences* **42,** 937–946 (2002).
6. Holm, L. & Sander, C. Protein Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology* **233,** 123 –138. ISSN: 0022-2836 (1993).
7. Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics* **17,** 412–425. ISSN: 1097-0134 (1993).
8. Daughdrill, G. W. *et al.* Understanding the structural ensembles of a highly extended disordered protein. *Mol. BioSyst.* **8,** 308–319 (2012).
9. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **41,** 6573–6582 (2002).
10. Ding, F., Jha, R. K. & Dokholyan, N. V. Scaling Behavior and Structure of Denatured Proteins. *Structure* **13,** 1047 –1054. ISSN: 0969-2126 (2005).
11. Harauz, G. *et al.* Myelin basic protein–diverse conformational states of an intrinsically unstructured protein and its roles in myelin assembly and multiple sclerosis. *Micron* **35,** 503 –542. ISSN: 0968-4328 (2004).
12. Nodet, G. *et al.* Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings. *Journal of the American Chemical Society* **131,** 17908–17918 (2009).
13. Radivojac, P. *et al.* Intrinsic Disorder and Functional Proteomics. *Biophysical Journal* **92,** 1439 –1456. ISSN: 0006-3495 (2007).

14. Sickmeier, M. *et al.* DisProt: the Database of Disordered Proteins. *Nucleic Acids Research* **35,** D786–D793 (2007).
15. Varadi, M. *et al.* pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Research* **42,** D326–D335 (2014).
16. Benson, N. C. & Daggett, V. Dynameomics: Large-scale assessment of native protein flexibility. *Protein Science* **17,** 2038–2050. ISSN: 1469-896X (2008).
17. Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S. & Shaw, D. E. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *Journal of the American Chemical Society* **134,** 3787–3791 (2012).
18. Yamada, J. *et al.* A Bimodal Distribution of Two Distinct Categories of Intrinsically Disordered Structures with Separate Functions in FG Nucleoporins. *Molecular & Cellular Proteomics* **9,** 2205–2224 (2010).
19. Lowry, D. F., Hausrath, A. C. & Daughdrill, G. W. A robust approach for analyzing a heterogeneous structural ensemble. *Proteins: Structure, Function, and Bioinformatics* **73,** 918–928. ISSN: 1097-0134 (2008).
20. Lyman, E. & Zuckerman, D. M. On the Structural Convergence of Biomolecular Simulations by Determination of the Effective Sample Size. *The Journal of Physical Chemistry B* **111,** 12876–12882 (2007).
21. Phillips, J., Colvin, M. & Newsam, S. Validating clustering of molecular dynamics simulations using polymer models. *BMC Bioinformatics* **12,** 445. ISSN: 1471-2105 (2011).
22. Ytreberg, F. M., Aroutiounian, S. K. & Zuckerman, D. M. Demonstrated Convergence of the Equilibrium Ensemble for a Fast United-Residue Protein Model. *Journal of Chemical Theory and Computation* **3,** 1860–1866 (2007).
23. Rauscher, S. & Pomès, R. Molecular simulations of protein disorder. *Biochemistry and Cell Biology* **88,** 269–290 (2010).
24. Dunker, A. K. *et al.* The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9,** S1. ISSN: 1471-2164 (2008).
25. Floriano, W. B., Domont, G. B. & Nascimento, M. A. C. A Molecular Dynamics Study of the Correlations between Solvent-Accessible Surface, Molecular Volume, and Folding State. *The Journal of Physical Chemistry B* **111,** 1893–1899 (2007).
26. Marsh, J. A. & Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophysical Journal* **98,** 2383 –2390. ISSN: 0006-3495 (2010).
27. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences* **107,** 8183–8188 (2010).
28. Pappu, R. V., Wang, X., Vitalis, A. & Crick, S. L. A polymer physics perspective on driving forces and mechanisms for protein aggregation. *Archives of Biochemistry and Biophysics* **469.** Highlight Issue: Protein Folding, 132 –141. ISSN: 0003-9861 (2008).

29. Phillips, J., Colvin, M., Lau, E. & Newsam, S. *Analyzing dynamical simulations of intrinsically disordered proteins using spectral clustering* in *Bioinformatics and Biomedicine Workshops, 2008. BIBMW 2008. IEEE International Conference on* (Nov. 2008), 17–24. doi:`10.1109/BIBMW.2008.4686204`.
30. Testa, L. *et al.* Electrospray ionization-mass spectrometry conformational analysis of isolated domains of an intrinsically disordered protein. *Biotechnology Journal* **6,** 96–100. ISSN: 1860-7314 (2011).
31. Kolodny, R., Petrey, D. & Honig, B. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology* **16,** 393 –398. ISSN: 0959-440X (2006).
32. Fisher, C. K. & Stultz, C. M. Protein Structure along the Order–Disorder Continuum. *Journal of the American Chemical Society* **133,** 10022–10025 (2011).
33. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2,** 19 –25. ISSN: 2352-7110 (2015).
34. Berendsen, H., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **91,** 43 –56. ISSN: 0010-4655 (1995).
35. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* **4,** 435–447 (2008).
36. Lindahl, E., Hess, B. & van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual* **7,** 306–317. ISSN: 0948-5023 (Aug. 2001).
37. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29,** 845–854 (2013).
38. Van Der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **26,** 1701–1718. ISSN: 1096-987X (2005).
39. Ortiz, A. R., Strauss, C. E. & Olmea, O. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science* **11,** 2606–2621. ISSN: 1469-896X (2002).
40. Liu, W., Srivastava, A. & Zhang, J. A Mathematical Framework for Protein Structure Comparison. *PLoS Comput Biol* **7,** e1001075 (2011).
41. Ohnishi, S., Kamikubo, H., Onitsuka, M., Kataoka, M. & Shortle, D. Conformational Preference of Polyglycine in Solution to Elongated Structure. *Journal of the American Chemical Society* **128,** 16338–16344 (2006).
42. Lorusso, M., Pepe, A., Ibris, N. & Bochicchio, B. Molecular and supramolecular studies on polyglycine and poly-l-proline. *Soft Matter* **7,** 6327–6336 (2011).
43. Tran, H. T., Mao, A. & Pappu, R. V. Role of Backbone-Solvent Interactions in Determining Conformational Equilibria of Intrinsically Disordered Proteins. *Journal of the American Chemical Society* **130,** 7380–7392 (2008).

44. Palenčár, P. & Bleha, T. Molecular dynamics simulations of the folding of polyalanine peptides. English. *Journal of Molecular Modeling* **17,** 2367–2374. ISSN: 1610-2940 (2011).
45. Raucci, R., Colonna, G., Castello, G. & Costantini, S. Peptide Folding Problem: A Molecular Dynamics Study on Polyalanines Using Different Force Fields. English. *International Journal of Peptide Research and Therapeutics* **19,** 117–123. ISSN: 1573-3149 (2013).
46. Singh, V. R. & Lapidus, L. J. The Intrinsic Stiffness of Polyglutamine Peptides. *The Journal of Physical Chemistry B* **112,** 13172–13176 (2008).
47. Vitalis, A., Wang, X. & Pappu, R. V. Quantitative Characterization of Intrinsic Disorder in Polyglutamine: Insights from Analysis Based on Polymer Theories. *Biophysical Journal* **93,** 1923 –1937. ISSN: 0006-3495 (2007).
48. Wang, X., Vitalis, A., Wyczalkowski, M. A. & Pappu, R. V. Characterizing the conformational ensemble of monomeric polyglutamine. *Proteins: Structure, Function, and Bioinformatics* **63,** 297–311. ISSN: 1097-0134 (2006).
49. Vitalis, A., Lyle, N. & Pappu, R. V. Thermodynamics of Beta-Sheet Formation in Polyglutamine. *Biophysical Journal* **97,** 303 –311. ISSN: 0006-3495 (2009).
50. Fedorov, M. V., Goodman, J. M. & Schumm, S. The effect of sodium chloride on poly-l-glutamate conformation. *Chem. Commun.* 896–898 (2009).
51. Denning, D. P., Patel, S. S., Uversky, V., Fink, A. L. & Rexach, M. Disorder in the nuclear pore complex: The FG repeat regions of nucleoporins are natively unfolded. *Proceedings of the National Academy of Sciences* **100,** 2450–2455 (2003).
52. Patel, S. S., Belmont, B. J., Sante, J. M. & Rexach, M. F. Natively Unfolded Nucleoporins Gate Protein Diffusion across the Nuclear Pore Complex. *Cell* **129,** 83 –96. ISSN: 0092-8674 (2007).
53. Milles, S. & Lemke, E. A. Single Molecule Study of the Intrinsically Disordered FG-Repeat Nucleoporin 153. *Biophysical Journal* **101,** 1710 –1719. ISSN: 0006-3495 (2011).
54. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126,** – (2007).
55. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52,** 7182–7190 (1981).
56. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65,** 712–725.
57. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78,** 1950–1958. ISSN: 0887-3585 (2010).
58. Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **21,** 1049–1074.

59. Jorgensen, W. L. Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *Journal of the American Chemical Society* **103,** 335–340. ISSN: 0002-7863 (1981).
60. Best, R. B., Zheng, W. & Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation* **10,** 5113–5124. ISSN: 1549-9618 (2014).
61. Abascal, J. L. F. & Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *The Journal of Chemical Physics* **123,** 234505 (2005).
62. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2015). <`https://www.R-project.org/`>.
63. Analytics, R. & Weston, S. *doMC: Foreach Parallel Adaptor for 'parallel'* R package version 1.3.5 (2017). <`https://CRAN.R-project.org/package=doMC`>.
64. Analytics, R. & Weston, S. *foreach: Provides Foreach Looping Construct for R* R package version 1.4.3 (2015). <`https://CRAN.R-project.org/package=foreach`>.
65. Analytics, R. & Weston, S. *iterators: Provides Iterator Construct for R* R package version 1.0.8 (2015). <`https://CRAN.R-project.org/package=iterators`>.
66. Rotkiewicz, P. & Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of Computational Chemistry* **29,** 1460–1465.
67. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81,** 3684–3690 (1984).
68. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1,** 80–83. ISSN: 00994987 (1945).
69. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist.* **18,** 50–60 (Mar. 1947).
70. Dyson, H. & Wright, P. E. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology* **12,** 54 –60. ISSN: 0959-440X (2002).
71. Tompa, P., Schad, E., Tantos, A. & Kalmar, L. Intrinsically disordered proteins: emerging interaction specialists. *Current Opinion in Structural Biology* **35.** Catalysis and regulation: Protein-protein interactions, 49 –59. ISSN: 0959-440X (2015).
72. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* **16.** Review Article, 18 (2014).
73. Arai, M., Sugase, K., Dyson, H. J. & Wright, P. E. Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proceedings of the National Academy of Sciences* **112,** 9614–9619. ISSN: 0027-8424 (2015).

74. Shammas, S. L., Crabtree, M. D., Dahal, L., Wicky, B. I. M. & Clarke, J. Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *Journal of Biological Chemistry* **291,** 6689–6695 (2016).
75. Ragavan, M., Iconaru, L. I., Park, C.-G., Kriwacki, R. W. & Hilty, C. Real-Time Analysis of Folding upon Binding of a Disordered Protein by Using Dissolution DNP NMR Spectroscopy. *Angewandte Chemie International Edition* **56,** 7070–7073.
76. Rogers, J. M. *et al.* Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proceedings of the National Academy of Sciences* **111,** 15420–15425. ISSN: 0027-8424 (2014).
77. Van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews* **114.** PMID: 24773235, 6589–6631 (2014).
78. Uversky, V. N. Dancing Protein Clouds: The Strange Biology and Chaotic Physics of Intrinsically Disordered Proteins. *Journal of Biological Chemistry* **291,** 6681–6688 (2016).
79. Kalodimos, C. G. *et al.* Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science* **305,** 386–389. ISSN: 0036-8075 (2004).
80. Levy, Y., Onuchic, J. N. & Wolynes, P. G. Fly-Casting in Protein–DNA Binding: Frustration between Protein Folding and Electrostatics Facilitates Target Recognition. *Journal of the American Chemical Society* **129.** PMID: 17243791, 738–739 (2007).
81. Hamdi, K. *et al.* Structural disorder and induced folding within two cereal, ABA stress and ripening (ASR) proteins. *Scientific Reports* **7,** 15544. ISSN: 2045-2322 (2017).
82. Linding, R. *et al.* Protein Disorder Prediction. *Structure* **11,** 1453–1459. ISSN: 0969-2126 (2003).
83. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7,** 208. ISSN: 1471-2105 (2006).
84. Ishida, T. & Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research* **35,** W460–W464 (2007).
85. Tam, J. P., Lu, Y.-A., Yang, J.-L. & Chiu, K.-W. An unusual structural motif of antimicrobial peptides containing end-to-end macrocycle and cystine-knot disulfides. *Proceedings of the National Academy of Sciences* **96,** 8913–8918. ISSN: 0027-8424 (1999).
86. Felizmenio-Quimio, M. E., Daly, N. L. & Craik, D. J. Circular Proteins in Plants: SOLUTION STRUCTURE OF A NOVEL MACROCYCLIC TRYPSIN INHIBITOR FROMMOMORDICA COCHINCHINENSIS. *Journal of Biological Chemistry* **276,** 22875–22882 (2001).
87. Wang, Y., Rosengarth, A. & Luecke, H. Structure of the human p53 core domain in the absence of DNA. *Acta Crystallographica Section D* **63,** 276–281 (2007).

88. Allison, J. R., Rivers, R. C., Christodoulou, J. C., Vendruscolo, M. & Dobson, C. M. A Relationship between the Transient Structure in the Monomeric State and the Aggregation Propensities of alpha-Synuclein and beta-Synuclein. *Biochemistry* **53.** PMID: 25389903, 7170–7183 (2014).
89. Beyer, K., Ispierto, L., Latorre, P., Tolosa, E. & Ariza, A. Alpha- and beta-synuclein expression in Parkinson disease with and without dementia. *Journal of the Neurological Sciences* **310,** 112–117. ISSN: 0022-510X (2011).
90. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* **9,** 90–95 (2007).
91. Tange, O. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine* (2011).
92. Cemazar, M., Joshi, A., Mark, A., Craik, D. & Daly, N. *The structure of a two-disulfide intermediate of MCoTI-II* 2007. doi:`10.2210/pdb2po8/pdb`.
93. Jones, E., Oliphant, T., Peterson, P., *et al. SciPy: Open source scientific tools for Python* 2001–. <`"http://www.scipy.org/"`>.
94. Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14,** 33–38 (1996).
95. Eaton, J. W., Bateman, D., Hauberg, S. & Wehbring, R. *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations* <`http://www.gnu.org/software/octave/doc/interpreter`> (2015).
96. Sessions, R. B., Dauber-Osguthorpe, P. & Osguthorpe, D. J. Filtering molecular dynamics trajectories to reveal low-frequency collective motions: Phospholipase A2. *Journal of Molecular Biology* **210,** 617 –633. ISSN: 0022-2836 (1989).
97. Karpen, M. E., Tobias, D. J. & Brooks, C. L. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2 ns trajectories of YPGDV. *Biochemistry* **32,** 412–420 (1993).
98. Torda, A. E. & van Gunsteren, W. F. Algorithms for clustering molecular dynamics configurations. *Journal of Computational Chemistry* **15,** 1331–1340 (1994).
99. Shao, J., Tanner, S. W., Thompson, N. & Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *Journal of Chemical Theory and Computation* **3.** PMID: 26636222, 2312–2334 (2007).
100. Paccanaro, A., Casbon, J. A. & Saqi, M. A. S. Spectral clustering of protein sequences. *Nucleic Acids Research* **34,** 1571–1580 (2006).
101. Weston, J. *et al.* Semi-supervised protein classification using cluster kernels. *Bioinformatics* **21,** 3241–3247 (2005).
102. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation* **9.** PMID: 26583988, 3084–3095 (2013).

103. Mukrasch, M. D. *et al.* Sites of Tau Important for Aggregation Populate β-Structure and Bind to Microtubules and Polyanions. *Journal of Biological Chemistry* **280,** 24978–24986 (2005).

104. Luo, Y., Ma, B., Nussinov, R. & Wei, G. Structural Insight into Tau Protein's Paradox of Intrinsically Disordered Behavior, Self-Acetylation Activity, and Aggregation. *J Phys Chem Lett* **5.** 25206938[pmid], 3026–3031. ISSN: 1948-7185 (2014).

105. Craik, D. J., Daly, N. L., Bond, T. & Waine, C. Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. *Journal of Molecular Biology* **294,** 1327 –1336. ISSN: 0022-2836 (1999).

106. Craik, D. J. & Du, J. Cyclotides as drug design scaffolds. *Current Opinion in Chemical Biology* **38.** Next Generation Therapeutics, 8 –16. ISSN: 1367-5931 (2017).

107. Filipp, F. & Tikole, S. Cystein knot with 2-FP label and integrin AvB6 cancer recognition site. To be published. doi:`10.2210/pdb2n8b/pdb`. <`http://www.rcsb.org/structure/2N8B`> (2015).

108. McDonald, N. Q. *et al.* New protein fold revealed by a 2.3-Å resolution crystal structure of nerve growth factor. *Nature* **354,** 411 (1991).

109. Robling, A. G. *et al.* Mechanical stimulation of bone in vivo reduces osteocyte expression of Sost/sclerostin. *Journal of Biological Chemistry.* doi:`10.1074/jbc.M705092200`. <`http://www.jbc.org/content/early/2007/12/17/jbc.M705092200.short`> (2007).

110. Kellenberger, C. *et al.* Serine Protease Inhibition by Insect Peptides Containing a Cysteine Knot and a Triple-stranded β-Sheet. *Journal of Biological Chemistry* **270,** 25514–25519 (1995).

111. Thongyoo, P., Tate, E. W. & Leatherbarrow, R. J. Total synthesis of the macrocyclic cysteine knot microprotein MCoTI-II. *Chem. Commun.* 2848–2850 (27 2006).

112. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242 (2000).

113. Bandyopadhyay, A & Raghavan, S. Defining the Role of Integrin $\alpha v\beta 6$ in Cancer. **10,** 645–52 (Aug. 2009).

114. Morris, G. M. *et al.* Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **19,** 1639–1662.

115. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **30,** 2785–2791.

116. Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry* **28,** 1145–1152.

117. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **31,** 455–461.

118. Xiong, J.-P. *et al.* Crystal Structure of the Extracellular Segment of Integrin $\alpha V\beta 3$ in Complex with an Arg-Gly-Asp Ligand. *Science* **296,** 151–155. ISSN: 0036-8075 (2002).
119. Gottstein, D., Kirchner, D. K. & Güntert, P. Simultaneous single-structure and bundle representation of protein NMR structures in torsion angle space. *Journal of Biomolecular NMR* **52,** 351–364. ISSN: 1573-5001 (2012).
120. Herrmann, T., Güntert, P. & Wüthrich, K. Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA. *Journal of Molecular Biology* **319,** 209–227. ISSN: 0022-2836 (2002).
121. Güntert, P. Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy* **43,** 105–125. ISSN: 0079-6565 (2003).
122. Güntert, P. in *Protein NMR Techniques* (ed Downing, A. K.) 353–378 (Humana Press, Totowa, NJ, 2004). ISBN: 978-1-59259-809-0. doi:`10.1385/1-59259-809-9:353`. <`https://doi.org/10.1385/1-59259-809-9:353`>.
123. Jee, J. & Güntert, P. Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *Journal of Structural and Functional Genomics* **4,** 179–189. ISSN: 1570-0267 (2003).
124. Shen, X. *et al.* Adjusting Local Molecular Environment for Giant Ambient Thermal Contraction. *Macromolecular Rapid Communications* **37,** 1904–1911. ISSN: 1521-3927 (2016).
125. Ebara, M., Uto, K., Idota, N., Hoffman, J. M. & Aoyagi, T. Shape-Memory Surface with Dynamically Tunable Nano-Geometry Activated by Body Heat. *Advanced Materials* **24,** 273–278.
126. Kloxin, A. M., Kasko, A. M., Salinas, C. N. & Anseth, K. S. Photodegradable Hydrogels for Dynamic Tuning of Physical and Chemical Properties. *Science* **324,** 59–63. ISSN: 0036-8075 (2009).
127. Mura, S., Nicolas, J. & Couvreur, P. Stimuli-responsive nanocarriers for drug delivery. *Nature Materials* **12,** 991 –1003 (2013).
128. Wu, G. *et al.* Remotely Triggered Liposome Release by Near-Infrared Light Absorption via Hollow Gold Nanoshells. *Journal of the American Chemical Society* **130,** 8175–8177 (2008).
129. Meng, H. & Li, G. A review of stimuli-responsive shape memory polymer composites. *Polymer* **54,** 2199 –2221. ISSN: 0032-3861 (2013).
130. Xie, T. Recent advances in polymer shape memory. *Polymer* **52,** 4985 –5000. ISSN: 0032-3861 (2011).
131. Henzl, J., Mehlhorn, M., Gawronski, H., Rieder, K.-H. & Morgenstern, K. Reversible cis–trans Isomerization of a Single Azobenzene Molecule. *Angewandte Chemie International Edition* **45,** 603–606.
132. Klajn, R. Spiropyran-based dynamic materials. *Chem. Soc. Rev.* **43,** 148–184 (1 2014).

133. Yokoyama, Y. Fulgides for Memories and Switches. *Chemical Reviews* **100.** PMID: 11777417, 1717–1740 (2000).

134. Zoete, V., Cuendet, M. A., Grosdidier, A. & Michielin, O. SwissParam: A fast force field generation tool for small organic molecules. *Journal of Computational Chemistry* **32,** 2359–2368.

135. Eder, F. R., Kotakoski, J., Kaiser, U. & Meyer, J. C. A journey from order to disorder – Atom by atom transformation from graphene to a 2D carbon glass. *Scientific Reports* **4,** 4060 EP – (2014).

136. Maiorov, V. N. & Crippen, G. M. Size-independent comparison of protein three-dimensional structures. *Proteins: Structure, Function, and Bioinformatics* **22,** 273–283. ISSN: 1097-0134 (1995).

137. Betancourt, M. R. & Skolnick, J. Universal similarity measure for comparing protein structures. *Biopolymers* **59,** 305–309. ISSN: 1097-0282 (2001).