Characterization of Intrinsically Disordered Protein Ensembles

by
Jacob C. Ezerski

A dissertation submitted to the Department of Physics,
College of Natural Sciences and Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Physics

Chair of Committee: Margaret Cheung

Committee Member: Greg Morrison

Committee Member: Claudia Ratti

Committee Member: James Briggs

Committee Member: Donna Stokes

University of Houston
May 2020

Acknowledgments

I would like to express my sincerest gratitude to Dr. Margaret Cheung for her tireless effort in mentoring me throughout my graduate career. Dr. Cheung has shown me what it means to be a scientist and has forever changed my interpretation of the objective pursuit of truth. Even outside of the scholastic environment, I find myself still benefiting from her lessons.

I would like to thank my thesis committee Dr. Greg Morrison, Dr. Jim Briggs, Dr. Donna Stokes, and Dr. Claudia Ratti for their amazing effort and guidance throughout my PhD career. I would like to thank my past and present group members, including Dr. Pengzhi Zhang, Jules Nde, Andre Gasic, Jacob Tinnin, Yossi Eliaz, Nathaniel Jennings, Rodney Helm. I would like to thank members of CTBP, including Dr. Qian Wang, Dr. Swarnendu Tripathi, Dr. Xingcheng Lin, Dr. Victor Tsai, Dr. Nick Schafer, Dr. Herbert Levine, and Dr. Peter Wolynes. I would like to thank Dr. Neal Waxham for his collaboration and Dr. Joan E. Shea for providing MD data for the Tau/R2 fragment.

I would like to thank and acknowledge the University of Houston Department

of Physics, the financial support from the National Science Foundation MCB 1412532 and ACI: 1531814, the training fellowship from the Gulf Coast Consortia, on the Houston Area Molecular Biophysics Program (NIGMS Grant T32GM008280)

I thank computational resources from Center for Advanced Computing and Data Systems at University of Houston, from Center for Theoretical and Biological Physics at Rice University.

I would like to express my love and gratitude to my mother, Olga Ezerski, for her constant support and incredible effort in bringing me where I am today in the face of adversity. I would like to thank Lindsay Layer for her affection and support while completing my PhD.

Finally, I would like to thank and remember my cat, Myla, who stayed by my side throughout the most difficult times on my path to obtaining my PhD, but passed away before its completion. Words can't express how much you mean to me.

Abstract

Intrinsically disordered proteins/peptides (IDPs) are a category of proteins that possess a poorly defined equilibrium structure. IDPs have been shown to play a central role in biological systems, however atomistic details about their binding mechanisms, selectivity, and specificity are poorly understood. The structures of IDPs are particularly challenging to study directly using experimental techniques due to their rapid inter-conversion of ensemble conformations. Similarly, theoretical techniques such as molecular dynamics simulations (MD), are typically parameterized via experimentally determined observables of stable proteins and therefore contain inherent biasing. Moreover, MD data refinement requires clustering of uniquely sampled structures in order to produce a non-biased ensemble. The work presented within this dissertation aims to remedy these IDP structure determination challenges by using circular dichroism (CD) to refine the conformations obtained from all-atom MD trajectories, and cluster the resulting conformations to remove degenerate structures in order to produce an unbiased ensemble with atomic resolution. The future application of the generated structure ensembles to-

wards IDP binding mechanism determination using a Markov State Model (MSM) is discussed and outlined.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Protein folding and dynamics

1.1.1 Proteins: the building block of living systems

In physics, we often describe phenomena in terms of relative size or time scales. On the smallest observable length scale, fundamental particles such as quarks, leptons, and bosons serve as the building blocks for atoms. The unique arrangement and interactions of these fundamental particles enable distinct elements such as oxygen, nitrogen and hydrogen to exist. Combining elements together to form molecules increases the system length scale and complexity. One set of molecules, known as amino acids, are of particular interest. Amino acids share a common structural feature (see figure 1.1): they posses both carboxylic acid and amino functional groups, which can react with each other to form polymer chains known as proteins (See figure

1.2). Upon formation of a protein, an individual amino acid is referred to as a residue to reflect the loss of water in peptide bond formation.

The physical properties of amino acids (and their subsequent classifica-

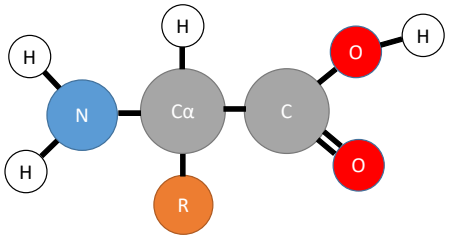

Figure 1.1: General structure of amino acids with substituted amino acid side chain labeled 'R'

tions) are dependent on the moiety of substituted side chain, which can vary in length, charge, and polarity. For example, non-polar side chains produce a hydrophobic effect and are insoluble in water. The side chains of alanine, valine, leucine, and isoleucine generate hydrophobic interactions within the protein and clump together. This effect increases the secondary structure stability of a given protein. Uncharged polar side chains including serine, threonine, cysteinem aspargine , and glutamine are able to form hydrogen bonds with water and are therefore hydrophillic. Finally, the most hydrophillic side chain groups are positively or negatively charged. At a biologically relevant pH of 7, lysine, arginine, and histidine are positively charged, while aspartate and glutamate are negatively charged.

Protein lengths can vary over several orders of magnitude, from two to several thousand residues, within biological systems. The hierarchy of size and

complexity of proteins is shown in figure 1.3. The description of amino acid sequence within a polypeptide is referred to as a primary structure. Due to the specific residue sequence present in the primary structure, portions of the protein will form secondary structures. The culmination of these structural subunits within the protein are refered to as the tertiary structure, which describes all aspects of the three-dimensional folding of a given protein. Finally, multiple protein subunits are assembled together to form a quaternary structure.

1.1.2 The protein folding problem

If an unfolded polypeptide chain has an extremely large number of conformational degrees of freedom, how is it able to reliably and repeatably fold into stable structures in a subsecond timescale? This questions was asked by Cyrus Levinthal in his 1969 paper as part of a thought experiment, which is commonly known as Levinthal's paradox[94, 177]. To illustrate the magni-

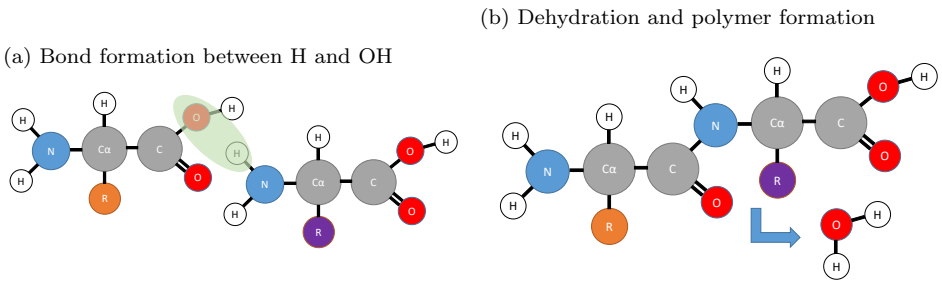

Figure 1.2: Illustration of a condensation reaction involving two amino acids. The amino (-NH2) moiety reacts with the carboxylic acid (-COOH) moiety to produce a peptide bond, liberating water in the process

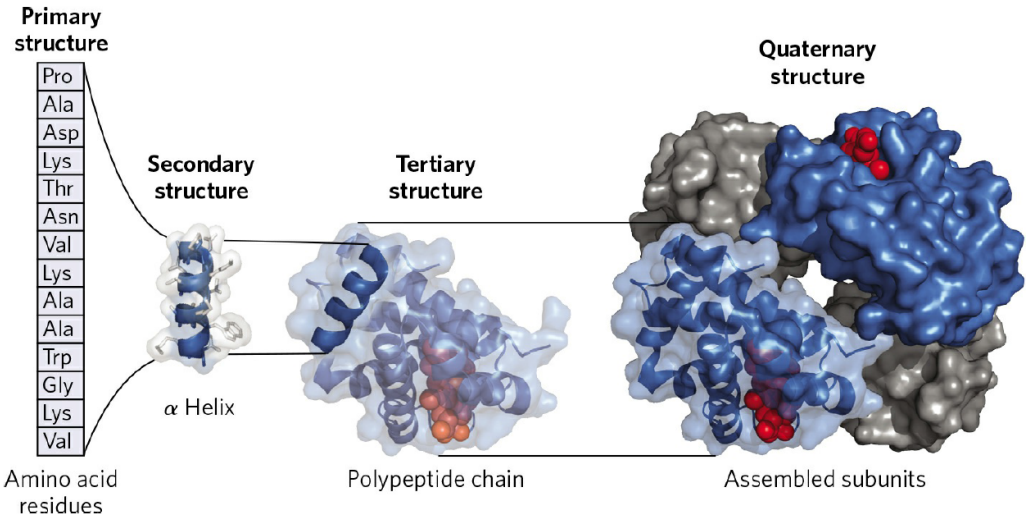

Figure 1.3: Levels of protein structure and complexity. A polypeptide chain is formed from amino acid condensation reactions, and is the primary structure of the macromolecule. In biological conditions, these polypeptides can form secondary structures, such as α-helices or β-strands. The tertiary structure of a protein includes all of the secondary structure subunits to form a monomer. Several individual monomers may exist together to form a quaternary structure, as seen in the CaMKII holoenzyme. (Image taken from Principles of Biochemistry [112])

tude of the problem, consider a protein composed of 101 amino-acids. If each of the 100 peptide bonds has the ability to sample 3 configurations, then this protein has $3^{100} = 5 \times 10^{47}$ total configurations available. If this protein was able to sample 10^{13} configurations per second, it would take approximately 10^{27} years to sample all of them[177]. Levinthal concluded that proteins do not randomly search for the natively folded configuration since proteins fold in the subsecond timescale. Levinthal's paradox highlights one of the main points behind the protein folding problem, which can be summarized by three main questions: (1) What is the relationship between 3D structure of a protein based on its linear amino acid sequence? (2) How are proteins

able to fold so fast despite having an extremely large conformation space of potential outcomes? (3) Can a protein's 3D structure be solved through theoretical approaches if only the amino acid sequence is known? [30]

Very few problems in physics are solvable in closed form. To put the protein folding problem in perspective, consider the solution to the Schrödinger equation for the hydrogen atom electron wave function. Although the hydrogen atom consists of only one electron and one proton, its solution required significant effort and was only obtainable because the hydrogen atom is a single-body problem. The closed form solution breaks down for many-body problems, such as the helium atom, and can only be approximated through numerical analysis. Now consider proteins, which consist of a large number of atoms on the order of hundreds or thousands. In addition to the atoms of the protein, one must also consider the solvent surrounding the protein of interest. Currently, a many-body problem of this magnitude cannot be solved numerically using quantum mechanical approaches. Fortunately, other theoretical and computational approaches exist to simplify the model in order to gain insight into protein behavior.

In constructing a theoretical model, we must first outline the forces affecting protein conformation. Although it appears relatively straight forward, a significant amount of approximations must be made before a computational model is feasible. If we consider the length scale of a single atom, one can easily identify the four fundamental forces at work: electromagnetic, strong, weak and gravitational. As the system becomes more complicated, as in the

case of a protein and its environment, many-body effects begin to emerge. Forces that appear to affect protein folding include hydrogen bonds, van der Waals interactions, backbone dihedral angles, electrostatic interactions, hydrophobic interactions and entropy. Generally, all of these phenomenon can be traced back to the four fundamental forces, however we must **coarse-grain** these complex many-body interactions in order to model and solve the problem.

Computational protein models have been an option since the advancement of computational technology. Prior to this, experimental methods were standard. In 1962, the Nobel Prize in Chemistry was awarded to Max Perutz and John Kendrew for their pioneering work in globular protein structure determination using X-ray analysis. This method and other experimental techniques such as nuclear magnetic resonance (NMR), cryo-electron microscopy (Cryo-EM), and circular dichroism spectroscopy(CD) are in use today for structural determination. Despite being able to generate high resolution structures using experimental techniques, we are still unable to answer the questions associated with the protein folding problem using experimental methods alone.

There appears to be a cyclic dependency between computational and experimental protein analysis methods. On the theoretical side, molecular dynamics (MD) force fields use Hamiltonian functions that are guided by the many-body forces described previously. These force fields are functions of intra-atomic distances within a given protein and take very similar forms.

The distinguishing features between them come from the parameterization of the Hamiltonian coefficients. Protein structures that are determined experimentally are used to refine the predictive power of MD through tuning. If the structure obtained through MD does not agree with the experimentally determined structure, an iterative approach is taken at modifying the Hamiltonian coefficients until a solution is reached.

1.1.3 Protein-protein interaction models

Proteins express their specific functions by interacting with other molecules. Protein-protein interactions that occur over relatively large distances (with respect to a protein's size) are typically governed by diffusion and electrostatic forces. These long-range interactions have been well studied, and several models have been proposed. The simplest model of the diffusion process governing bimolecular interaction is the Smoluchowski diffusion equation [147, 116]:

$$\nabla \cdot D(\mathbf{r}) \cdot [\nabla - \mathbf{F}(\mathbf{r})/k_b T]\, \rho(\mathbf{r}) = 0 \tag{1.1}$$

where $\rho(\mathbf{r})$ is the pairwise probability density at $\mathbf{r}$, $k_b T$ is the Boltzmann constant and absolute temperature (Kelvin), $\mathbf{F}(\mathbf{r})$ is the negative gradient of the bimolecular potential of mean force, and $D(\mathbf{r})$ is the diffusion tensor. Improvements to the Smoluchowski equation have been made to account for rate enhancements due to electrostatic interactions or rate reductions due to site-dependent binding. For example, Northrup et al. characterized several

modifications that take into account the possibility of proteins not binding upon initial encounter, and electrostatic rate enhancement [116].

Experimentally determined binding kinetics can provide insight into the equilibrium probability of two proteins binding (or not binding) at close ranges, however modeling the physics behind these probabilities is much more involved. Depending on the types of proteins being investigated, different binding models have been proposed.

1.1.3.1 Lock and key

One of the earliest models for protein-protein interactions was the "lock and key" binding mechanism, which was proposed by chemist Emil Fischer in 1894 [47]. Proteins that fold into stable structures at biologically relevant conditions, known as globular proteins, are thought to interact at complementary shape interfaces. The interaction between a receptor and ligand is strongest when the two "fit" together without any further structure modification. This type of interaction is typically observed within enzyme activity; a reaction only occurs between one or two possible substrates that match the enzyme active site. The kinetics of such a model could be described using hydrodynamics because the rate limiting reaction step is largely due to diffusion processes. This model assumes that both the enzyme and substrate do not require any conformational changes in order to bind. Realistically, there are few scenarios where this model can accurately describe protein binding. As a result, newer models that take into account flexibility have emerged.

1.1.3.2 Induced fit & conformational selection

The modern approach to modeling protein-protein interactions comes from the assumption that the target enzyme has a degree of flexibility, while the ligand is rigid. This is an important feature in recognition because a conformational change in the target must take place before a stable bound complex is formed. Figure 1.4 illustrates the differences between the induced fit[85] and conformational selection[109] protein binding mechanisms. The final bound complex resides at an energetic minimum, however before forming this complex, an energy barrier must be overcome. The differences in the two binding mechanisms stem from whether a conformational change takes place before or after the formation of an encounter complex. In the induced fit model, an energetically unfavorable encounter complex is first formed (the free energy barrier). The change in enzyme conformation upon encounter allows for the stabilized low energy complex to occur. Alternatively, the conformational selection model proposes that the enzyme will sample different conformations **before** an encounter complex is formed. When a conformation that is complementary to the rigid ligand is sampled, the stable bound complex in formed.

1.1.3.3 Mutual and Induced Conformational Fit

The mutual and induced conformational fit mechanism proposed by the Cheung group[162] revolves around flexibility within proteins, similar to the induced fit and conformational selection binding mechanisms. A key assump-

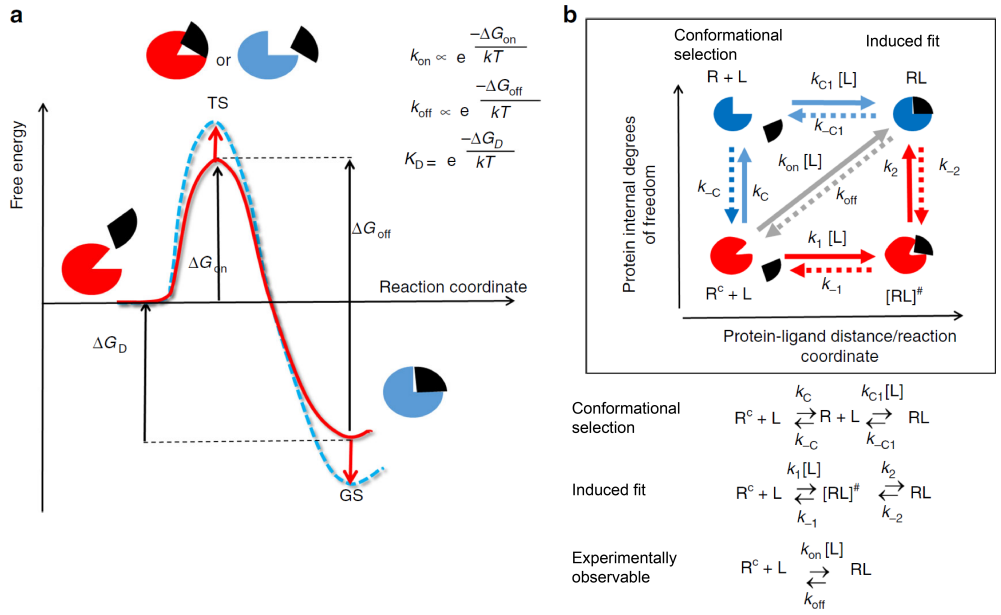

Figure 1.4: Comparison between conformational selection and induced fit mechanisms [2].

tion in the mutually induced mechanism is that both receptor and ligand proteins change conformation in order to form the final bound complex. This model is used to explain the experimentally observed conformational changes in calmodulin and its binding targets (CaM/CaMBT) in the free and unbound states (see figure 1.5). Since the structure of both the target and receptor proteins have changed after binding, this model is applicable towards proteins with unstructured regions and intrinsically disordered proteins.

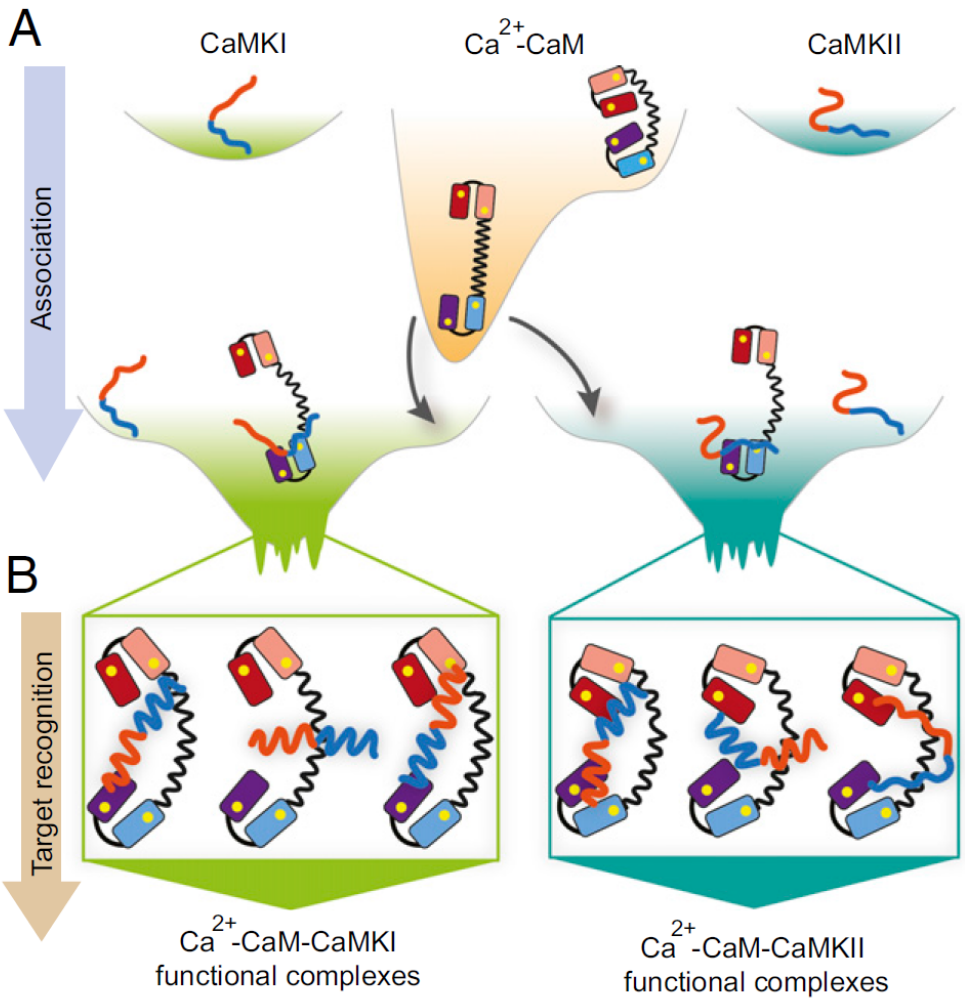

Figure 1.5: Case study on the mutually induced conformational changes that take place between CaM and its binding targets CaMKI and CaMKII peptides. [162]

1.1.3.4 Flycasting

On the opposite spectrum of the lock and key model exists the flycasting model [142]. The binding kinetics of disordered proteins and peptides is challenging to predict because identification of potential binding sites on a receptor is not dependent on its complementary structure. The flycasting mechanism asserts that binding kinetics of a disordered protein will be greater than a folded counterpart. Proteins with structure begin to bind at short distances (on the order of 1 Å), and require proper complementary surface alignment. This results in an energy barrier that must be overcome before the final complex is formed. On the other hand, fully denatured proteins will begin binding to a target if within a distance defined by the radius of gyration. A cooperative binding can occur because the disordered protein is able to reorient itself to better interact with the target site residues, as illustrated in figure 1.6.

1.1.4 Electrostatics of proteins in solution

Proteins exist in biological environments, which are typically aqueous in nature and contain a great number of other molecules or ions that have the ability to interact and change the behavior of proteins. For example, calmodulin has the ability to change conformations in the presence of Ca^{2+} ions, which is essential for its role in Ca^{2+} signal transduction. Another example of how co-solutes are able to modulate protein behavior is shown in chapter 2; the

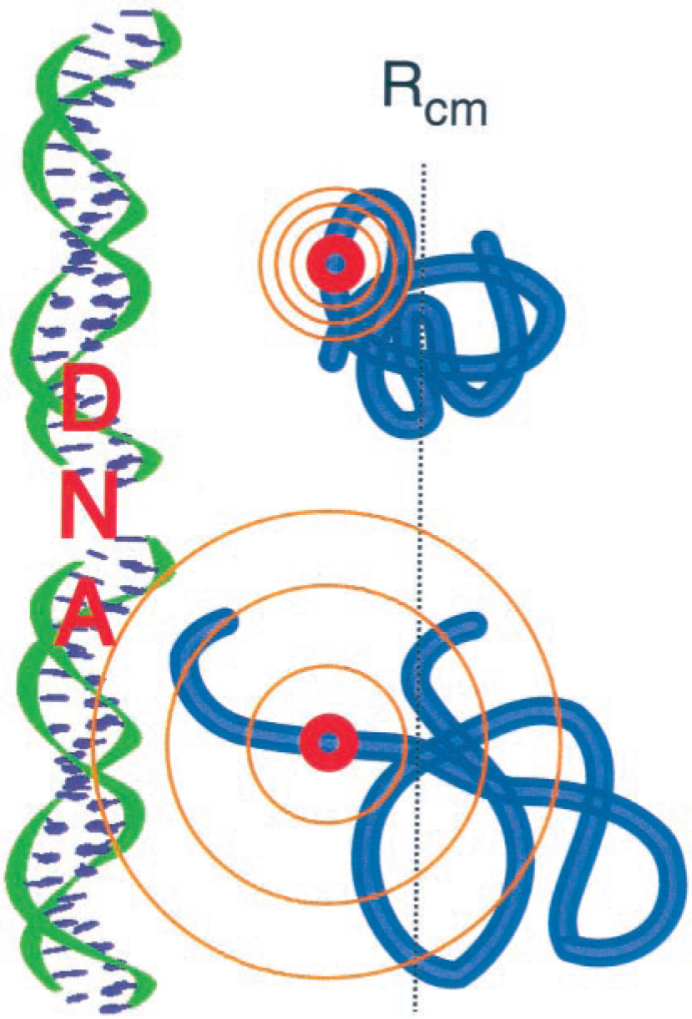

Figure 1.6: Cartoon depiction of how the fly-casting mechanism increases folding speed. At a distance R_{cm}, the partially folded ensemble is able to form a few initial contacts to the binding site, while the folded structure remains out of range due to the smaller number of conformational fluctuations in the folded state. Although the initial contacts are weak, they allow the protein to "reel" itself into the binding site to fold and bind simultaneously. [142]

effect of TMAO and urea osmolytes on the conformational ensemble of the Tau/R2 fragment is investigated through our analysis of MD simulation data from the Shea group.

These are complex examples of environmental factors that affect protein behavior. Explicit representation of ions or osmolytes in solution with proteins is challenging to model and lacks the generality needed for most applications. On the other hand, the effect of ions in solution on protein binding and folding can be generalized through its effect on electrostatics. Electrostatic rate enhancement is the increase in formation of the encounter complex due to Coulombic attraction of charged residues at long distances:

$$\vec{\mathbf{F}}_{ij} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}^2}\hat{\mathbf{r}}_{ij} \tag{1.2}$$

Conversely, the presence of ions in solution will decrease the effective force of long range electrostatics. Salts, such as NaCl or $CaCl_2$, ionize in solution. As such, they are not randomly distributed due to their own electrostatic properties interacting with the field generated by two charged residues or proteins at a distance r_{ij}. These mobile charges effectively dampen the electrostatic force and can be modeled through the Debye-Huckel equation:

$$\vec{\mathbf{F}}_{ij} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}^2} e^{-\frac{r_{ij}}{\lambda_D}}\hat{\mathbf{r}}_{ij} \tag{1.3}$$

where the ionic concentration, species, and temperature are taken into account in λ_D, which is known as the Debye length. Inclusion of this shielded in-

teraction between charged proteins is especially convenient for computational modeling. Additionally, the idea of electrostatic screening is particularly useful for determining factors that contribute to protein binding. In the first step of binding, proteins must form an encounter complex through diffusion. Once the encounter complex is formed, additional conformational changes must take place to form the final bound complex. From an experimental standpoint, only the overall equilibrium association rates can be determined. This includes a combination of electrostatic rate enhancements in the diffusion region and any conformational changes. Thus, additional information about the rate limiting steps can be found if electrostatic enhancements were shielded in ionic solution.

1.2 Intrinsically disordered proteins

Intrinsically disordered proteins /peptides (IDPs) are a category of proteins that possess a poorly defined equilibrium structure; they sample an ensemble of weakly ordered and unordered structures in solution [40, 34, 156, 159, 125]. IDPs have been shown to play a central role in biological systems through cellular signaling, regulation, and translation [159, 12, 169]. Additionally, behavioral changes of IDPs are associated with cancer [73] and neurodegenerative diseases [57, 110, 93] such as Alzheimer's disease. A distinguishing feature of IDPs is that they do not adhere to the classical structure-function paradigm, and typically form stable secondary or tertiary structures only

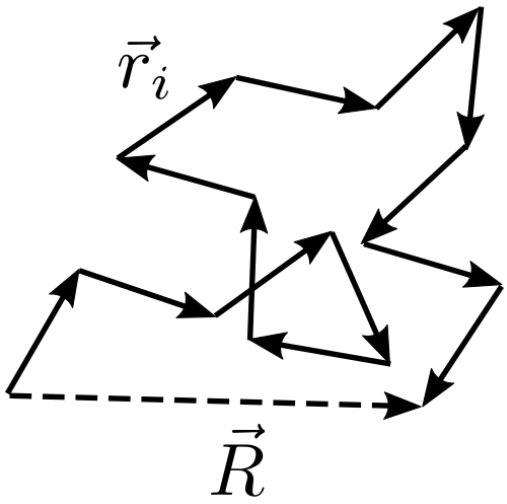

Figure 1.7: Illustration of the ideal polymer chain consisting of rigid rods with length l and orientation based on vector points $\vec{\mathbf{r}_i}$ Image credit: ThorinMuglindir

upon binding to target proteins [168, 51]. Unlike globular proteins that form folded structures at biologically relevant conditions, IDPs do not form single well-defined structure. As a result, IDPs are typically represented as an ensemble of possible structures that inter-convert rapidly. The lack of stable structures in the ensemble of unbound state [4, 49, 55, 97] enables binding to multiple targets on demand while maintaining a degree of selectivity and specificity due to their polymorphic properties [37].

1.2.1 Ideal polymer chain

A highly simplified version of a peptide can be modeled by an ideal polymer chain. The **freely jointed chain model** is the simplest ideal polymer chain, which has been well studied[136]. This model does not take into account

any molecular interactions, rather the dynamics of the ideal polymer chain proceeds through a random-walk process.

The model consists of rigid rods with length l, where the orientation of each rod is independent of neighboring rods. Since the monomers in this model are non-interacting, they are able to overlap. The total unfolded length of the polymer is given by $L = Nl$, where N is the number of monomer units in the chain and l is the length of each monomer. A vector drawn from one end of the chain to the other can be characterized as

$$\mathbf{R} = \sum_{i=1}^{N} \mathbf{r_i} \tag{1.4}$$

Consider a very large ideal chain, where $N \to \infty$. Since each monomer in the chain is an independent random variable, the central limit theorem is applicable. As a result, $\mathbf{R}$ and $\vec{\mathbf{r}_i}$ follow a Gaussian distribution for the random walk process. Hence, the end-to-end distance, $\mathbf{R}$, will fluctuate around an average point $\mu = 0$:

$$\langle \mathbf{R} \rangle = \sum_{i=1}^{N} \langle \mathbf{r_i} \rangle = 0 \tag{1.5}$$

The mean-square end-to-end distance is non-zero,

$$\langle \mathbf{R}^2 \rangle = \left\langle \left(\sum_{i=1}^{N} \mathbf{r_i} \right) \cdot \left(\sum_{j=1}^{N} \mathbf{r_j} \right) \right\rangle \tag{1.6}$$

Equation 1.6 can be combined and further reduced to

$$\left\langle \mathbf{R}^2 \right\rangle = l^2 \sum_{i=1}^{N} \sum_{n=1}^{N} \left\langle \cos\theta_{ij} \right\rangle \tag{1.7}$$

Since the monomers are not correlated, $\langle \cos\theta_{ij} \rangle = 0$ for $i \neq j$. Thus, the variance given as

$$\sigma^2 = \left\langle \mathbf{R}_x^2 \right\rangle = \left\langle \mathbf{R}_y^2 \right\rangle = \left\langle \mathbf{R}_z^2 \right\rangle = N\frac{l^2}{3} \tag{1.8}$$

Thus the average end-to-end distance of the polymer chain is given as $\sqrt{N}l$. Although this model is highly simplified, valuable insights into the physics behind peptides and proteins can be observed. In particular, the resulting average end-to-end distance from the above statistical arguments is not 0 or the total length of the chain L. Such a result can be counter intuitive since there are no forces acting on an individual monomer. Entropy and solvation effects are significant factors in IDP ensemble behavior due to hydrophobic/hydrophilic peptide residues. Globular proteins are typically composed of hydrophobic and uncharged residues, leading to a compact stable state.

1.2.2 The flat energy landscape of IDPs

A globular protein that folds into a stable conformation at biological conditions possesses local or global minima on its energy landscape. Thus, the folded conformation may be achieved regardless of folding pathway taken.

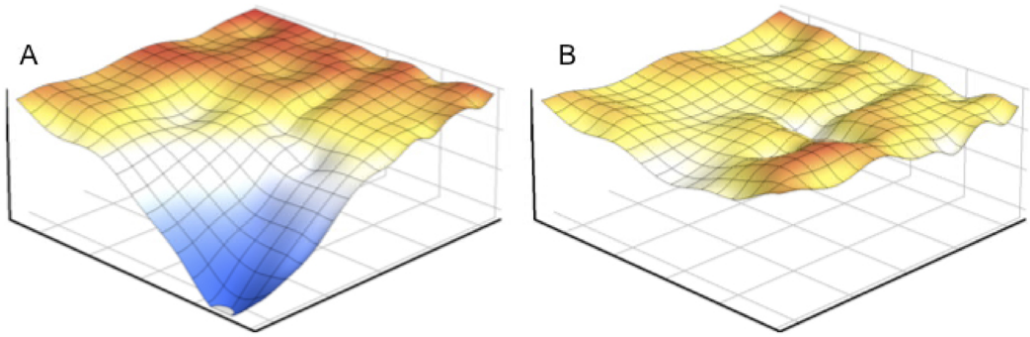

Figure 1.8: Example of typical energy landscapes for folded proteins and IDPs. The energy landscape of a folded protein (A) illustrates the funneled energy landscape. The folded state at biological conditions corresponds to the lowest energy point. Conversely, the flat landscape (B) illustrates the lack of a global energy minimum, which results in an ensemble of structures to be sampled at equilibrium. [49]

Conversely, IDPs do not possess a funneled energy landscape, as depicted in figure 1.8. The dimple-like landscape of IDPs illustrates that there are certain conformations that are relatively more stable than others, however the energy barrier between all possible conformations is less than the thermal energy in the system at biological conditions. Thus, IDPs sample an ensemble of structures at equilibrium.

1.2.3 Experimental analysis of IDPs

There are several methods that are useful in identifying IDPs, each possessing a different set of highlights and limitations. Because of the dynamic nature of IDPs, experimental methods are typically limited in their ability to distinguish individual conformations. Regardless of resolution, many experimental techniques are adequate in detecting whether a protein region or peptide is

disordered.

1.2.3.1 X-ray crystallography

X-ray crystallography is a popular method of protein structure resolution. Using X-ray scattering, the electron densities of a sample may be mapped and analyzed to determine structure. Early experiments observed that some sections of a crystallized protein had unresolvable electron densities in functional regions [71, 10]. There are other explanations given for the incoherence of scattering, such as the possibility of crystal defects, however the low electron density from scattering is typically a result of the dynamic motions of the disordered or unstructured regions of a protein [71]. Although the ensemble of structures produced by the disordered segments are not clearly resolved using x-ray crystallography, the lack of data enables for the identification of potentially disordered regions of a crystallized protein. Alternatively, nuclear magnetic resonance (NMR) spectroscopy became a competing experimental method for protein structure determination.

1.2.3.2 NMR spectroscopy

Nuclear magnetic resonance (NMR) consists of several parameters that are used in the study of IDPs, including: chemical shifts (CS), parmagnetic relaxation enhancements (PREs), residual dipolar couplings (RDCs), nuclear relaxation and relaxation dispersion [77, 84, 119]. Each one of these parameters is capable of producing data related to the construction of the IDP

ensemble. For example, in order to determine the local structure propensities of an IDP, CS data can be used to resolve localized secondary structures such as α-helix and β-strand conformations. PREs and RDCs are able to characterize and detect long-range interactions within the protein. Additionally, PREs can identify weakly populated states. Nuclear relaxation and relaxation dispersion is able to resolve timescale dynamics related to interconversion of structures within the ensemble or conformational changes that occur due to binding [75].

NMR parameters are averaged over all structures in the conformational ensemble [36]. In the construction of the energy landscape of the protein, steric restrictions are imposed so that a limited number of backbone dihedral angle combinations are possible, which reduces the complexity of ensemble generation. Due to the inherent flexibility of IDPs and rapid interconversion between multiple conformations, the chemical shift dispersion of most resonances is poor, and sequence-specific assignment of resonances is difficult[36]. A popular method of reconstructing the ensembles of IDPs uses computational techniques that limit the energy landscape sampled by MD through backbone restraints derived from NMR parameters.

1.2.3.3 Circular dichroism spectroscopy

Circular dichroism (CD) spectroscopy is a simple spectroscopy method of protein secondary structure determination. Like other experimental methods, CD describes the ensemble-averaged protein secondary structure only. Differ-

ences in the ensemble-averaged secondary structure are detected through the change in absorption between left and right polarized light. The absorption spectra between 190-240nm contains information on the particular secondary structure propensities within the ensemble [56, 100]. IDPs and denatured protein regions are typically identified through CD signals in the 195nm region [102]. Deconvolution of the CD spectra shares a similar difficulty to other specroscopy methods in that there is no analytical relationship between the experimentally observed spectra and protein secondary structure. Generally, the spectra for proteins with known structures are used to solve unknown experimental spectra for ensemble-averaged secondary structures.

1.2.4 Analytical methods

As discussed previously, the protein folding problem seeks to describe the 3D structure of a protein based on its sequence. Although resolution of the IDP structure ensemble based on sequence is still an unsolved problem, the magnitude of disorder as well as conformational propensities have been predicted for short proteins.

Uversky proposed a relationship between so-called "natively unfolded" proteins and sequence characteristics[158]. In addition to a correlation between hydrophobicity, net protein charge and disorder, Uversky shows that there exists a strong correlation between protein length and propensity for disorder (see figure 1.9). Ignoring dependencies on specific residue types, Uversky showed that shorter proteins and peptides appear to be less likely to fold into

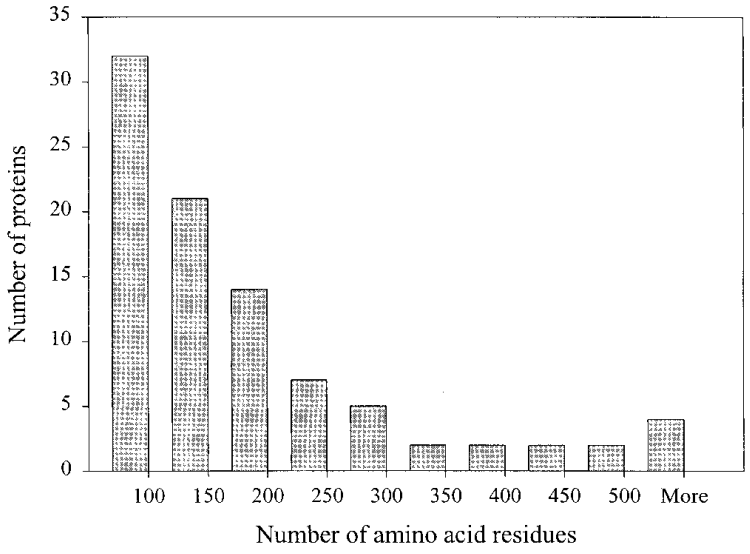

Figure 1.9: Histogram of the distribution of "natively unfolded" proteins with respect to protein length. Figure taken from Uversky et al. [158]

stable conformations than larger proteins.

1.2.4.1 Dependency on charge distribution

The Pappu group has shown a correlation between conformation and distribution of charged residues within peptides. In their 2013 study[24], the Pappu group was able to characterize the particular charged residue distribution through an order parameter κ. Their work illustrated the correlation of κ with the radius of gyration and compactness of a 50-residue peptide consisting of positively-charged lysine and negatively-charged glutamic acid residues. The effect of varying distributions of the charged residues within the peptide was investigated (see figure 1.10a). An order parameter related to the charge asymmetry is defined as $\sigma = \frac{(f_+ - f_-)^2}{(f_+ + f_-)}$, where f_+ and f_- are the

fractions of positively and negatively charged residues respectively. The κ parameter is calculated by partitioning the given sequence into N_{blob} overlapping residues of size g. The charge asymmetry for each segment g of blob i is calculated with $\sigma_i = \frac{(f_+ - f_-)_i^2}{(f_+ + f_-)_i}$. The deviation between asymmetries is defined as $\delta = \frac{\sum_{i=1}^{N_{blob}} (\sigma_i - \sigma)^2}{N_{blob}}$. Different sequence variants will result in different values of δ; upon determination of δ_{max}, the κ order parameter can be solved:

$$\kappa = \frac{\delta}{\delta_{max}} \tag{1.9}$$

The order parameter κ is correlated with the ensemble R_g, as shown in figure 1.10b. A generalized result of Pappu's study indicates that R_g decreases with increasing κ values.

1.2.4.2 Bioinformatic approach

There are several disorder prediction tools available, many of which use a residue correlation approach to predict the behavior of proteins based on sequence alone. This approach is used in the IDP prediction tool IUPred[33, 32]. Proteins have a great number of pairwise residue interactions, whose potential energies can be approximated through proteins with known structures. The particular sequence composition that results in stabilized or disordered proteins can be obtained from various experimental methods. Simon et al.[33, 32] characterize the total energy of a particular sequence through

(a)

(b)

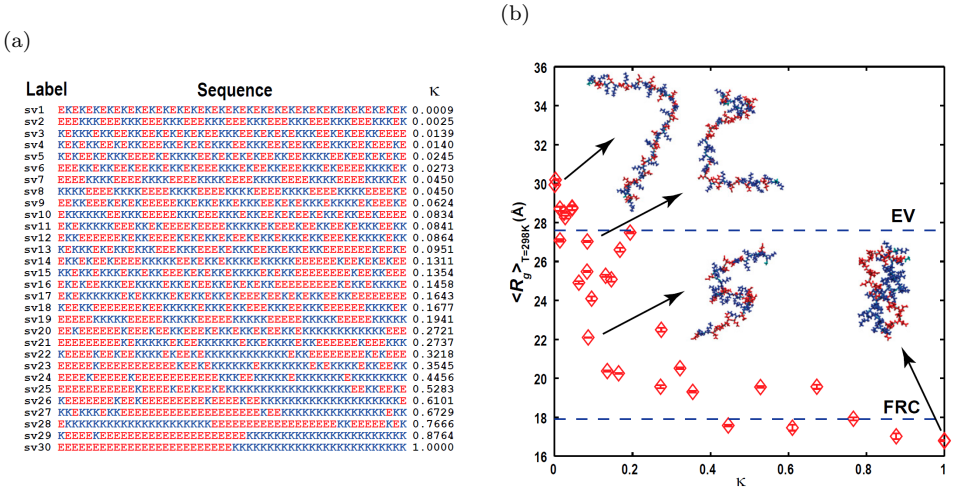

Figure 1.10: Conformational dependence on κ. (a)Thirty sequence variants were tested with different κ values. (b): Ensemble-averaged radii of gyration for sequence variants of the Glu-Lys system. Insets show representative conformations for four sequence variants. Side chains of Glu are shown in red, and side chains of Lys are shown in blue. The two dashed lines intersect the ordinate at Rg values expected for all sequence variants of the Glu-Lys system modeled in the excluded volume limit or as Flory random coils (FRCs)[24].

the individual pairwise interactions of each residue:

$$E = \sum_{ij}^{20} M_{ij} C_{ij} \tag{1.10}$$

where M_{ij} is the interaction energy of amino acid types i and j, and C_{ij} is the number of interactions between residue types i and j in a given protein conformation. However, if the conformation is not known, the energy per amino acid can be approximated:

$$\frac{E_{appx}}{L} = \sum_{ij}^{20} n_i P_{ij} n_j \tag{1.11}$$

where P_{ij} is a matrix that describes the energy correlation between residue i and j in a particular sequence through bioinformatically derived potentials, n_i and n_j describe the frequency of residue i and j in a sequence respectively. Thus, the propensity for stabilization or disorder can be approximated for unknown conformations of IDPs using equation 1.11. A drawback of this method is that the pairwise energy correlation matrices used in 1.10 and 1.11 rely on data obtained primarily from globular proteins. Despite the success of bioinformatic approaches to predict the degree of disorder in proteins and peptides, this analytical method does not provide any information regarding the actual conformations that can be sampled by the IDP ensemble. Thus, other approaches may be needed in order to obtain additional data for IDP modeling.

1.2.5 Current approaches and challenges in IDP modeling

Several methods for experimental structure determination of proteins, such as crystallography, cannot be used on IDPs due to this unstable nature. Theoretical approaches, such as MD simulations, largely rely on Hamiltonians whose coefficients are fitted to stable globular proteins, resulting in biased ensemble sampling. As such, the generation of IDP ensembles requires a combined computational and experimental approach. Variation of experimental and simulation methods can be seen across research groups that study IDPs, however the advantages of each approach is dependent on the particular system being studied. In a 2011 study [4], Head-Gordon et al. generate ensembles of the amyloid-β peptide using all-atom MD simulations with explicit water and the ff99SB force field, which are subsequently validated through the use of multiple NMR observables, such as chemical shift and scalar coupling constants. Structures were sampled from the MD equilibrium run at 1ns intervals, amounting to a total of fifty structures. Chemical shifts were calculated using SHIFTS [172], J-coupling was calculated using the Karplus equation [81]. This ensemble generation method does not use NMR observables to refine the generated ensemble; the comparison between ensemble structures and experimental observables is performed to simply validate the results. This places a strong dependence on force field accuracy and sufficient MD sampling rate. Brooks and Head-Gordon later

revised their ensemble generation model in a 2016 study [13] by perturbing the native protein structure about the phi/psi dihedral angles without biasing. The role of experimental NMR data is to define a Gaussian distribution, which exploits measurement uncertainty to define the distribution variance. Ensemble structures are ranked according to the location of back-calculated NMR observables on the experimentally generated Gaussian distribution. The generation of the structure pool using perturbation improves conformational sampling over force-field based trajectory data due to the lack of intrinsic force field bias, which was a significant obstacle throughout our method of ensemble generation discussed in chapter 3. Forman-Kay et al. published a similar approach to ensemble generation for the SH3 domain [20] as Head-Gordon et al with an alternative approach in the structure pool generation stage, which was produced by sampling structures in unfolding MD trajectories. Gong et al. [55] generate IDP structure ensemble by first using MD to generate a structure pool. The structures generated by MD are clustered, and chemical shift data for each cluster center is calculated and compared to experimental values of the IDPs of interest. Experimental chemical shifts here are taken from the BRMB databank. Although this study suggests that their model's main issue stems from structure degeneracy and weigh factors, they do not address the biasing issue in MD sampling. Similarly, Lindorff-Larsen et al. [97] use NMR to restrain computational simulations. They assert that a major difficulty in the model stems from the NMR experimental data; the NMR observables are the average of the struc-

ture ensembles. Their solution to structure generation is to match the average chemical shifts of MD trajectories to the experimental average. The resulting IDP ensembles are heavily influenced by back-calculations of observables for generated structures. NMR structure back calculations use a database that provides a relationship between known structures and chemical shifts. Unfortunately, the NMR chemical shift databases suffer a similar drawback to the CD databases; they consist of stable proteins with α-helix and β-sheet structures, instead of IDPs.

1.3 Computational protein modeling

Molecular dynamics simulations produce high resolution interaction and folding models of proteins, and allow for unparalleled insight into their interaction mechanisms. Unfortunately, there is always a trade-off between simulation time and system size. The number of calculations required in any MD simulation grows exponentially with system size for a single time step. Thus, a smaller system will be suited for more detailed calculations for a given simulation length. Conversely, a large system may require approximations in order to reduce the number of calculations required in a single time step. In chapters 2 and 3, the conformational ensembles of short peptides are investigated using all-atom MD. This was the natural choice for these systems due to their small sizes. On the other hand, chapter 4 discusses modeling the CaM/CaMKII binding kinetics through a coarse-grained simulation due

to the larger size and length scales associated with the system.

1.3.1 All-atom models

As the name suggests, all-atom molecular dynamics simulations take into account each atom of a protein. There are several all-atom MD simulation environments available, with Amber, Gromacs, and Lammps among the most popular. These simulators govern the motion of proteins through a Hamiltonian, which contains interaction terms that are bioinformatically derived.

1.3.1.1 Force field parameters

Amber uses the following potential energy function:

$$V_{Amber} = \sum_{bonds} k\left(r - r_{eq}\right)^2 + \sum_{angles} k\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left[1 + cos(n\phi - \gamma)\right] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] + \sum_{i<j}\left[\frac{q_i q_j}{\epsilon R_{ij}}\right] \tag{1.12}$$

The coefficient of each term in the potential is unique to the particular force field used. There are several different Amber specific force fields that contain alternative parameters for the potential function. The variance between force fields arises from different experimental fittings. Many of the fitted force field parameters come from globular proteins. This can pose a problem with structure biasing in equilibrium simulations since IDPs do not share the same stability. Best et al. attempt to remedy the over sampling of compact struc-

tures in the CHARM 22 force field by optimizing sidechain potentials as well as backbone potentials using NMR data for weakly structured peptides[9]. Refinement of the potential function parameters is a common method for continued improvements on simulation accuracy, however choice of solvent model can significantly affect the simulation outcome as well.

1.3.1.2 Explicit vs implicit solvent

The treatment of water in MD simulations can drastically change the sampled conformations. Best and Mittal [8] showed that changing the water model for the Amber ff03 force field from TIP3P to TIP4P resulted in an improved helix-coil equilibrium transition, compared to the original water model.

Implicit solvent is an alternative to using explicitly represented water molecules. It is an attractive option for all-atom simulations since explicit solvent simulations require an enormous amount of computational power to obtain adequate sampling. These implicit solvent models can approximate the free energy of solvent-solute interactions through the accessible surface areas. Zhou observed that the free energy landscape of implicit solvent MD was significantly different from the explicit MD counterpart in several Amber force fields, resulting in the over biasing of α-helical conformations [175]. Despite this drawback, an implicit solvent model was used in chapter 3 to generate a large sample space of the CaMKII peptides that would otherwise be unobtainable with an explicit solvent model.

1.3.2 Coarse-grained models

Experimental data fitting can be used to reduce the degrees of freedom of the system and smooth out the energy landscape [131]. Additionally, a coarse-grained model accounts for solvent effects through implicit representation of solvent within the force field [82]. These aspects of coarse-grained MD can significantly increase the time scale and size of a simulated protein[157]. In larger protein systems, such as the CaM-CaMKII binding model discussed in chapter 4, a larger time scale is required to describe the protein dynamics of interest. There is a complex transition between the folded and unfolded states of CaM [151], therefore increased sampling in addition to longer time scales is necessary to describe CaM dynamics.

There are several different coarse-grained models available, with the simplest one being the structure-based G$\bar{o}$ model [82], consisting of a single $C\alpha$ bead per amino acid that governs dynamics. The model discussed in chapter 4 requires special consideration, however, because of the inclusion of disordered peptides/regions. To resolve this issue, the coarse-grained model, AWSEM [170], is used instead. This model uses a 3-bead representation of an amino acid: $C\alpha$, $C\beta$, and O. Additionally, the parameters of the potential function shown in eq. **??** have been derived bioinformatically, and have the ability to include user-defined structure based potential through the "fragment memory" term (eq. **??**).

$$V_{total} = V_{backbone} + V_{contact} + V_{burrial} + V_{helical} + V_{FM} \tag{1.13}$$

with

$$V_{FM} = -\lambda_{FM} \sum_{m} e^{\left[\frac{(r_{ij}-r_{ij}^{m})^2}{2\sigma_{ij}^2}\right]} \tag{1.14}$$

Using the generated structure ensembles of the CaMKII peptides from chapter 3 to bias the energy landscape, a system consisting of IDPs and folding proteins may be modeled together with approximate all-atom accuracy.

1.4 Future work: Application to CaM/CaMKII

1.4.1 Functionality of CaM/CaMKII

Calcium ion signaling is an essential feature of neurological development, memory formation, and dendritic spine growth. Several hundred Ca^{2+} binding proteins have been identified and the majority of them share a particular Ca^{2+} binding motif. The transduction of calcium ion signals is mediated by calcium sensing proteins, such as calcineurin, troponin C and calmodulin (CaM). Calmodulin consists of four EF-Hand motifs connected by a flexible linker. The EF-Hand motif contains a helix-loop-helix topology that is able to interact with Ca^{2+} ions. In addition to the negatively charged amino acid residues present in the EF-hand motif, backbone carboxyl oxygens and water play a significant role in CaM's interaction with Ca^{2+} ions. CaM is of particular interest because it is the most ubiquitous Ca^{2+} sensing protein, and is present in all eukaryotic organisms. The bioligical significance of CaM can also be inferred from its highly conserved sequence. Where many of

the Ca^{2+} binding proteins serve to regulate free Ca^{2+} concentration, CaM uniquely interacts with Ca^{2+} as a secondary messenger to decode signals. This is due to the flexibility of CaM's linker and EF-hand motifs, enabling its target interactions to be modulated by its environment. Fluorescence experiments indicate that the rate of Ca^{2+} induced conformational change at the N-terminal are over an order of magnitude faster than that of the C-terminal [123].

1.4.2 CaM/CaMKII binding

One of the most abundant and evolutionarily conserved activation targets of CaM is calmodulin-depended protein kinase II (CaMKII)[67, 154]. CaMKII exists as a holoenzyme with mirrored 6 and 7 symmetric monomer ring structures that undergoes a complex activation/inactivation process involving both Ca^{2+} and CaM, allowing it to transduce Ca^{2+} ion signals into biological functions [135, 105]. CaMKII can be activated due to frequency dependent Ca^{2+} spikes, remain active independent of the Ca^{2+} or Ca^{2+} -CaM levels, and undergo molecular switch-like behavior that is critical for memory formation and learning [72]. Each of the monomer subunits must interact with a Ca^{2+} saturated CaM, inducing a change in conformation that exposes the catalytic domain [152]. From here, the exposed domain can be autophosphorylated by adjacent monomers that leads to a long term activated state known as CaM trapping that permits the monomer to remain active when Ca^{2+} levels return to the basal state[104]. CaM-trapping, in turn, is the fundamental

feature that permits CaMKII to act as a molecular decoder of the frequency and amplitude of intracellular Ca^{2+} -pulses ([27]).

Unfortunately, the mechanism of Ca^{2+} -CaM/CaMKII signal transduction is still unknown and debated. Using a Monte-Carlo based single molecule approach, Waxham and Kubota [87] showed that the N-terminal of CaM is highly sensitive to the spacial and temporal distribution of Ca^{2+} ions. The Ca^{2+} saturated N-terminal of CaM is able to partially activate and induce autophosphorylation of the CaMKII enzyme without Ca^{2+} saturation of the C-terminal. Ca^{2+} -CaM/CaMKII binding is thought to occur over a sequential series of steps beginning with N-CaM binding [50]. Alternatively, it has been shown[141] that the C-terminal of CaM has a dominant effect on binding to CaMKII targets.

1.4.3 CaMKII peptides

The molecular mechanism of the phosphorylated (trapped) and unphosphorylated state has been investigated with a family of CaMKII peptide models about the CaM binding domain of CaMKII (CaMKII 293-312). It is an excellent model for experimental and computational simulation due to its size and availability of the Ca^{2+} -CaM/peptide native bound complex[103]. When three charged residues of the CaMKII peptide (296-298; Arg-Arg-Lys) were mutated to alanine, there was a loss of binding affinity that was reflected in both an increased off-rate and a decreased on-rate. An atomistic understanding of the conformational transitions required during the binding

pathway from basal to the CaM-trapped state is unknown.

1.4.4 Markov state model of CaM/CaMKII peptides

The kinetics of equilibrium protein binding can be modeled using Markov State Model[122, 91, 86](MSM). The MSM is particularly useful in illustrating distinct binding pathways that are responsible for the observed protein kinetics. This offers a distinct advantage over modeling kinetics via free energy in complex protein interactions. In the case of the CaMKII peptides, a MSM can illustrate detailed changes in binding pathways between the wild-type and mutant peptides. Since IDPs with low stability are able to bind with high specificity and selectivity, modeling their complex kinetic pathways is paramount in understanding their behaviors. The Noe group is known for their work describing complex binding and folding kinetics using MSMs. An example of this model's use in characterizing binding with conformational dependencies is shown in figure 1.11. Small changes in a protein conformation are characterized as microstates, and the accumulation of changes between an initial and final state of interest can be described through macrostates.

1.4.4.1 Microstates

In order to generate a MSM, a set of microstates must be used to discretize the state space, Ω. These microstates are assumed to transition between each other over an infinite amount of time (ergotic)[127]. Thus, any clustering method that partitions the state space into non-overlapping clusters with low

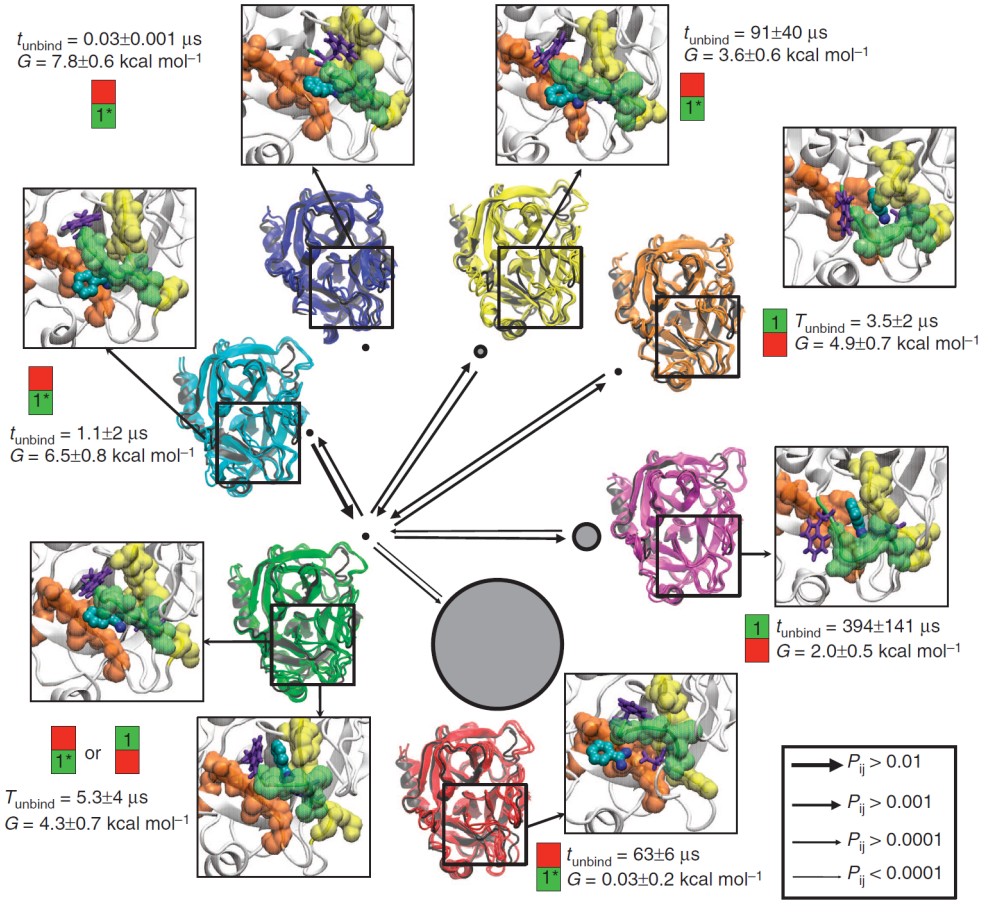

Figure 1.11: Trypsin conformations with Benzamidine-bound and the binding mode of Benzamidine. The seven conformational states shown are equal to six apo states, plus the yellow conformation that is only found with Benzamidine-bound. The binding pocket conformation is defined by three loops: the yellow loop (residues 187–194) with Asp189, the green loop (residues 215–221) with Trp215 and the orange loop (residues 225–230). The circles have an area proportional to the equilibrium probability of the respective conformation, given that Benzamidine is bound, π_i. Their respective relative free energies $G = -K_b T ln \pi_i$ and the unbinding times t_{unbind} (mean first passage time to unbinding) are given. The arrows indicate the transition probabilities for direct transitions between the different states. The binding mode (pocket 1 or 1*) is indicated by the green square with 1 or 1*.[126]

transition energy barriers will suffice for an MSM model. Despite the fact that trajectories may not show bidirectional transitions between microstates, we can filter out kinetically irrelevant microstates (microstates that trap the trajectory) and calculate the symmetric transition matrix.

1.4.4.2 Macrostates

Macrostates cluster sets of microstates that are related by some kinetically significant metric. For example, we can define several macrostates relating to the open/closed conformations of CaM, and the bound peptide location relative to CaM (N-terminal, C-terminal, linker). In the CaM-CaMKII binding model, we will use k-means clustering to partition the conformational state space according to single linkage RMSD matrix values of the coarse-grained trajectory $C\alpha$ atoms.

Generally, an MD simulation consisting of N frames is partitioned into k clusters. The number of clusters obtained is typically large (on the order of 10^3)[122, 115] because we require each state to be kinetically similar. The resulting microstates can be coarse-grained based on some similarity metric(eg. major conformational change, energy, etc.) to combine the initial microstate ensemble into a form that is qualitatively understandable.

1.4.5 Future work

Using the MSM with our work on clustering and ensemble generation of IDPs, a full binding model of the CaM/CaMKII peptide system can be generated.

In chapter 4, the technical aspects behind building and implementing the MSM are discussed.

Chapter 2

Combinatorial-averaged transient structure (CATS): A tool for clustering IDPs on a flat energy landscape

Published 2018 in JPCB[45]

2.1 Introduction

The data sets from molecular dynamics (MD) simulations can provide valuable insight into protein interactions and dynamics through clustering analysis [140]. There are several main categories of clustering methods: hierar-

chical (eg. single linkage) [54], vector quantization (eg. K-Means)[95, 79], neural network[41, 80], mixture density[176], and fuzzy [39, 44], to name a few [140, 176, 171, 83]. In addition to these, new methods routinely emerge that encompass characteristics from multiple algorithm categories, optimizing clustering of a specific set of data [140, 173, 52, 153]. Reduction of dimensionality is sometimes desired to reduce computational cost and necessary sample size; however the reduction of dimensions also can lead to lower separation resolution [153]. RMSD-based clustering methods are typically employed to find dominant protein structures due to their simplicity and low dimensionality, however this technique heavily relies on an appropriate cutoff between cluster groups to be effective. In RMSD-based clustering, choosing a separation cutoff that is too large will produce significant variation within the cluster and increase structure ambiguities, while a small cutoff may over-separate similar structures[88]. Due to the sensitivity of the input parameters and low separation dimensionality, an RMSD-based method may not be the best choice of analysis for proteins that sample a large conformation space. This is a defining characteristic of intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs), which represent a group of proteins that do not form a stable tertiary structure under physiological conditions and possess a large conformation space [159, 34]. A well-studied protein containing a disordered region is Tau[43], whose function regulates microtubule growth in the nervous system [29]. Significant interest in Tau has occurred due to its link with neurodegenerative diseases; notably Alzheimer's

disease [110, 58]. Describing the conformational dynamics of IDPs is difficult, therefore several approaches have been developed to characterize them that do not revolve around strict structure resolution (eg. sequence predictions) [161]. A popular experimental method for IDP structure analysis is NMR spectroscopy. Experimental methods like NMR resolve measured data through RMSD fitting to generate structure ensembles, and offer a macroscopic view of IDP dynamics, however these methods are typically unable to fully resolve the conformational substates due to high conformational variation [117, 76, 14]. Alternatively, MD simulations offer an atomistic level of resolution to address the problem of IDP interactions[138]. Regardless of method, structure resolution is essential for functional analysis of IDPs. Because of the inherent flexibility and large conformation space exhibited by IDPs [36], generation of dominant structures through the use of clustering is a useful tool in structural analysis. Here, we discuss the development of the Combinatorial Average Transient Structure (CATS) algorithm for determining structural ensembles of IDPs, and use it to analyze the trajectory data for the Tau/R2 fragment in urea and TMAO solutions published by the Shea group [93]. The clusters generated by CATS require no a priori information about the expected structure, which is advantageous for determining significant structures in MD trajectories where experimentally determined structures are unavailable. The choice of backbone phi and psi dihedral angle coordinates as structure descriptors was chosen for clustering due to their rotational and translational invariance. The CATS method generates clus-

ters through each descriptor coordinate distribution, therefore the phi and psi residue dihedral angles have the added benefit of specifying backbone orientation using two values instead of three values in a Cartesian coordinate system. Through our novel technique, we show that CATS is able to generate a robust set of primary structures highlighting the denaturing effect of urea and compacting effect of TMAO on the Tau/R2 fragment. Additionally, we compare our results to the clusters obtained with the GROMOS method and show that CATS produces more structurally unique clusters that correspond to a higher resolution energy landscape as described by a Fano factor[46] of 0.65 and 0.60 for the R2/TMAO and R2/urea trajectories respectively (compared to Fano factors of 0.36 and 0.54 produced by the GROMOS method). The drawback of this higher resolution energy landscape separation is that CATS requires a set of coordinate descriptors with multi-modal distributions that are well defined. CATS was tailored to separate the structures of proteins with assumed meta-stable conformational states, and as a result not all datasets are applicable.

2.2 Methods

2.2.1 Combinatorial averaged transient structure (CATS) clustering algorithm

CATS was created as a tool to identify highly probable ensemble structures within a trajectory for proteins that rapidly sample multiple conformations. Generally, any coordinate with a Gaussian-like distribution can be used for clustering. Our interest lies in conformational clustering from atomic coordinates; therefore we designed the process with dihedral coordinates as a descriptor. This descriptor is invariant to the translational or the rotational symmetry from the data set. Initially, the angular distribution of phi and psi dihedral angles for each residue is generated from the trajectory data. Assuming there are N residues, we obtain 2N individual coordinate distributions, where each of the 2N coordinate distributions contained between 1 and 3 Gaussian-like peaks (see figure 2.1). Each peak is fitted with a Gaussian curve, and the average angle and standard deviation of each curve is used to transform the dihedral coordinates from each trajectory frame into an index value. We obtain a set of 2N index values for every frame in the trajectory, which is used to generate clusters based on similarity between each frame's index set. Finally, the clusters are further refined by allowing members of low populated clusters to be absorbed by others. Figure 2.2 depicts the flow chart of CATS.

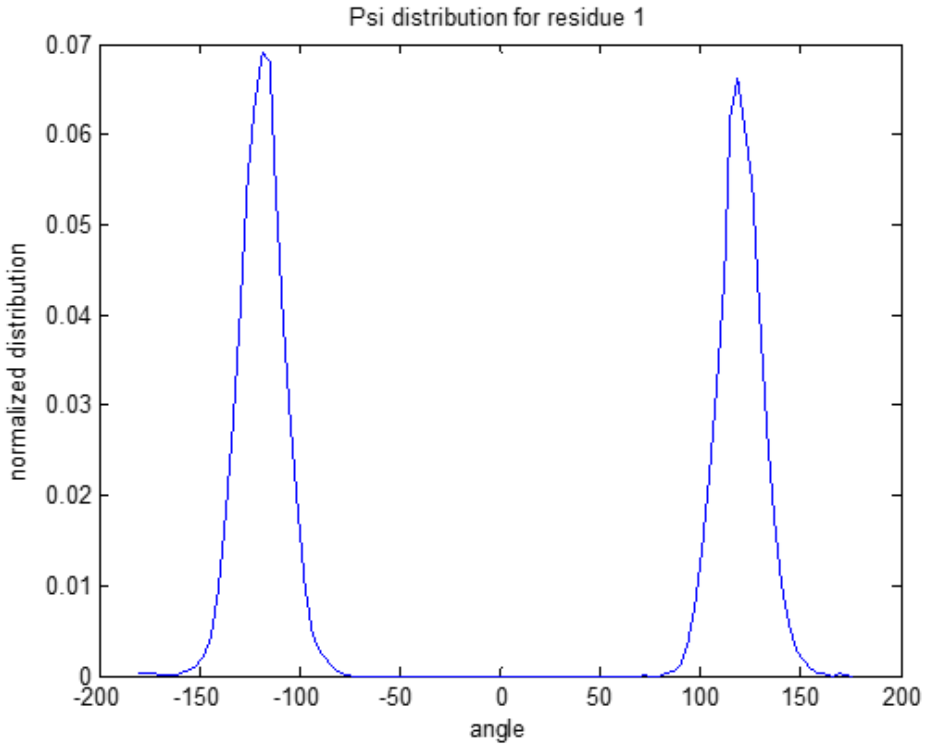

Figure 2.1: Example dihedral angle distribution of the Tau/R2 trajectory for a single residue. The multi-modal Gaussian distribution is a critical attribute of any coordinate descriptor used in the CATS clustering algorithm.

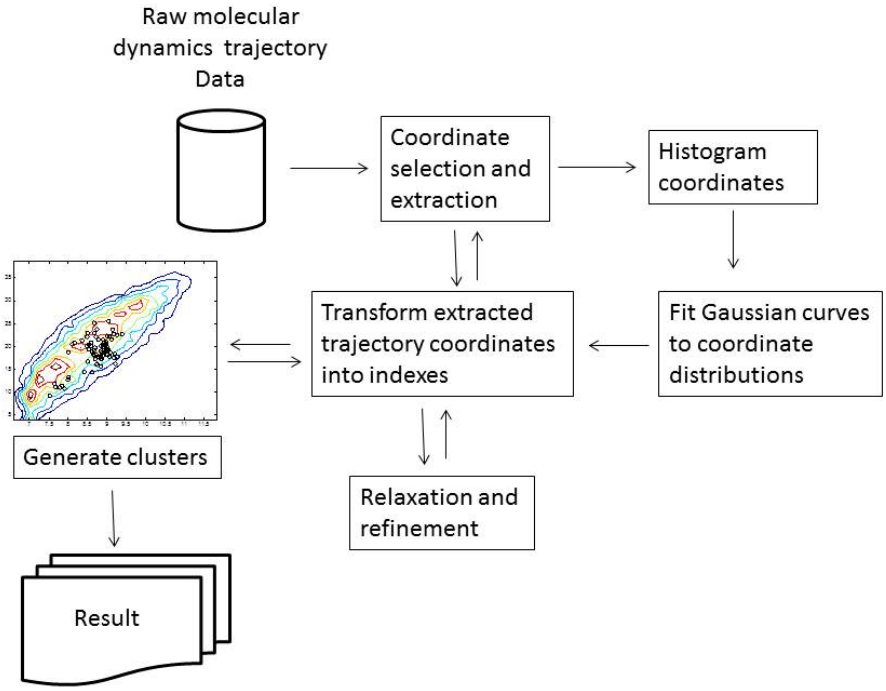

Figure 2.2: A flow chart of the CATS algorithm. Coordinates that exhibit Gaussian-like distributions are extracted from the trajectory and histogrammed. One or more Gaussian curves are fitted to the distribution of each coordinate, and each peak's average value and standard deviation is generated. These values are compared to the extracted trajectory frame coordinates and transformed into indices that label which Gaussian curve the coordinate belongs to within a given distribution. Clusters are initialized by grouping together trajectory frames that possess the same transformation indices. Refinement allows lower populated clusters to join higher populated ones if their indices have less than a user-defined number of mismatches.

2.2.1.1 Dihedral angle distribution analysis

The phi/psi dihedral coordinates for a given molecular dynamics trajectory of an N-residue protein is extracted for each of the N residues of every frame. We histogram these trajectory phi and psi dihedral angles using a bin size of 3.6° on a range of −180° to 180°. The dihedral angle distribution represented by the raw histogram data is smoothed and fitted with Gaussian curves of the form:

$$f(\theta) = \sum_{i=1}^{N} A_i e^{\frac{(\theta-\mu_i)^2}{2\sigma_i^2}} \tag{2.1}$$

with amplitude A, angle θ, dihedral average μ and standard deviation σ corresponding to peak i. The maximum number of peaks in the coordinate distribution, N, varies by coordinate and trajectory and generally ranges between 1 and 3. The standard deviation and amplitude in each coordinate distribution is used to calculate an approximate probability value for peak i given by $P_i = \frac{A_i\sigma_i}{\sum_j A_j\sigma_j}$. A maximum of 3 most probable peaks are tabulated, while statistically insignificant distribution peaks (less than 1% probability) are discarded.

2.2.1.2 Trajectory transformation

For a protein with N residues, every frame in the MD trajectory is transformed into a 2N-character string of numbers (0, 1, 2 or 3), which represent the specific angular distribution that each dihedral coordinate belongs to. This is accomplished by taking the dihedral coordinates from a given tra-

jectory frame and comparing it to the corresponding coordinate distribution peaks obtained from the previous step. Each of the frame's 2N dihedral coordinates will be assigned an indexing value of 0, 1, 2 or 3 if the coordinate value satisfies:

$$|\theta_i - \mu_i^n| < \epsilon\sigma_i^n \tag{2.2}$$

where θ_i is the trajectory frame's dihedral coordinate i, n is the coordinate distribution peak index, ϵ is a user defined constant, μ_i^n and σ_i^n are the distribution peak's average and standard deviation respectively. CATS assumes there is a maximum of 3 possible peaks for a given coordinate distribution, therefore the first peak (starting from $-180°$) is labeled 0, the second peak is labeled 1 and so on. If a dihedral coordinate from the trajectory frame is unable to be matched to any of the tabulated distribution peaks, it is given an index value of 3.

2.2.1.3 Initial cluster formation

Clusters are formed by grouping together the transformed trajectory strings. The first frame's indexing string defines the first cluster. The next frame's indexing string is compared to the first cluster by calculating the value of:

$$C = \sum_{i=1}^{2N} \Theta\left(\left|\lambda_i^k - \lambda_i\right|\right) \tag{2.3}$$

where Θ is the Heaviside step function, λ_i^k is the distribution index for the ith coordinate of cluster k and λ_i is the ith distribution index of a given

trajectory indexing string. If $C = 0$, the frame's indexing string matches the first cluster's indexing string, and the frame joins the cluster. If $C \neq 0$, the indexing string of the frame fails to match with the indexing string of the cluster and it forms a new cluster defined by that unique indexing string. Similarly, every subsequent frame's indexing string is compared to all previous clusters; if there is a match, it joints that cluster, otherwise a new cluster is formed as defined by the newest frame's indexing string. This process is repeated for all trajectory frames until all frames are clustered. There is an option in CATS to specify coordinates in that are to be completely ignored throughout cluster formation and relaxation; this effectively changes the original number of coordinate comparisons (2N) to a user dependent value.

2.2.1.4 Cluster representation and center structure

For a given cluster, each member is assigned an error factor that describes how close the member's dihedral coordinates are to the tabulated coordinate distributions defining the cluster. This error value, E, is calculated by:

$$E = \sum_{j=1}^{2N} \frac{\left|\theta_j - \mu_j^n\right|}{2N} \tag{2.4}$$

where θ_j is the cluster member's dihedral angle for coordinate j, μ_j^n is the cluster defined distribution peak average for peak index n and coordinate j. The trajectory frame number of the cluster member with the lowest error

value is used to represent the conformation of the cluster.

2.2.1.5 Cluster relaxation

Once a basis set of initial clusters are established, CATS attempts to reduce the number of clusters by combining lower populated clusters that are similar to higher populated clusters. Relaxation is based on two user-defined criteria: cluster population cutoff, α, and number of ignorable coordinates, β. For clusters with a population under the cutoff value, CATS will compare each member's indexing data to all other clusters, starting from the largest cluster and ending at the smallest cluster, excluding the current cluster. In the initial stage of clustering, all members of a specific cluster must contain the same set of indexing values (given by a string of $2N$ values), however in the relaxation stage, the relaxed cluster member may be absorbed into a different cluster as long as the indexing data between the relaxed member and new cluster does not differ by more than the specified number of ignorable coordinates β. In other words, each frame's indexing data is allowed to fail up to β index comparisons with a given cluster before indicating a failure to match. In the case that a cluster member cannot be absorbed elsewhere, it remains in its original cluster. Cluster centers are not recalculated after relaxation. Clusters that are relaxed are considered to be artifacts of simulation noise or transitions; initially, they are not grouped together with larger clusters due to a small fraction of coordinate index deviations. Essentially, relaxation serves to change the population rankings of major clusters by correcting the

deviation.

2.2.2 Implementation of CATS and GROMOS algorithms on the Tau/R2 fragment

2.2.2.1 Cluster setup using the GROMOS method

Following the setup from the original Tau/R2 study conducted by the Shea group, we use the GROMACS clustering package as a control [128, 6]. Clusters are formed using the algorithm described in Daura et al [25] using an RMSD cutoff value of 1.4Å.

2.2.2.2 Comparison of CATS clustering method

We focus our comparison between CATS clusters and the clusters obtained by the RMSD-based GROMOS method used in the original R2 fragment study [25]. The potential of mean force (PMF) was generated using the distributions of the radius of gyration (Rg) and end-to-end distance (Ree) using 25 evenly-spaced bins to histogram values.

2.2.2.3 All atom molecular dynamics simulation of the Tau/R2 fragment

The all-atom molecular dynamics simulations of the Tau/R2 fragment in TMAO/urea was generated by the Shea group from a previous study [93]. Details on the model and simulation setup for Tau/R2(273-284) fragment

can be found in the Shea group's publication [93], which will be briefly summarized here. The effect of TMAO and urea on the aggregation-prone (R2) region of the Tau protein is investigated using all-atom molecular dynamics (MD) simulations. Topologies were generated using GROMACS. Trajectories consisting of 6000 frames for the R2/Urea configuration and 7500 frames from the R2/TMAO configuration (totaling 120ns and 150ns respectively) are generated using a leap-frog algorithm to integrate the Newtonian equations of motion in the Nose-Hoover NVT environment at 300K using a time step of 2fs and sampling rate of 20ps.

2.3 Results

2.3.1 CATS clusters and their center structures represent local minima in the energy landscape

The dihedral phi and psi angles chosen for clustering produced a series of Gaussian-like peaks in their distributions. Each coordinate's peak describes a probability density, which associates the potential mean force (PMF) and free energy with the clustering coordinate. CATS transforms the trajectory into a 2N-dimensional energy landscape (where N represents the number of residues) and creates clusters based on the landscape's minima, as shown in figure 2.3.

CATS method produces a different analysis of the R2 trajectory than GRO-

MOS because GROMOS clusters by average RMSD. CATS uses the probability distribution analysis that is essentially its free energy landscape. Figure 2.4 illustrates cluster populations and center structures for the four highest populated clusters of CATS and GROMOS compared to the potential mean force (PMF) of the TMAO trajectory. Similarly, figure 2.5 illustrates the distribution of cluster populations for both methods compared to the PMF of the urea trajectory.

The top four cluster members in the R2/TMAO trajectory produced by CATS and GROMOS are similarly distributed around the center of the lowest point on the PMF contour map shown in figure 2.5. CATS and GROMOS produce the same center structure for the first cluster, and a β-hairpin structure for the second cluster. The third cluster of GROMOS is similar to the fourth cluster of CATS, however the third CATS cluster center and fourth GROMOS cluster center have completely different secondary structures. The distribution of cluster members from the third CATS cluster is closer to the center of the PMF map than the members of the fourth GROMOS cluster. Despite the difference in order between the top four GROMOS and CATS clusters, the fourth GROMOS cluster is still captured by the sixth CATS cluster. Similarly, the third CATS cluster is captured by the ninth GROMOS cluster. The similarities between some of the top structures indicates that the two methods do not completely produce mutually exclusive cluster groups.

The R2/urea trajectory produces a more diverse set of clusters. All GRO-

MOS cluster members are distributed around the minimum PMF in the neighborhood of Rg = 7.5Å, whereas CATS clusters have distributions around both of the PMF local minima. This results in noticeably different center clusters, with only CATS cluster 2 and GROMOS cluster 1 having a similar conformation. GROMOS clusters 2 and 3 appear to be similar, indicating the presence of degeneracies. The fourth GROMOS cluster is a beta-hairpin structure, which is significantly different from the other clustered structures. Despite GROMOS and CATS previously having captured similar structure centers at different ranks, CATS does not capture the hairpin conformation as its center structure throughout the top 20 CATS clusters due largely to the low probability of a positive phi dihedral angle coordinate (1%) of the 6th residue and poorly defined Gaussian-like peaks in the phi/psi angles of residues 5 and 7, which is responsible for the hairpin conformation (see figure 2.3C and supporting information for distribution). The low-probability phi peak in residue 6 was not taken into account during the distribution file setup, and therefore all structures containing this peak were flagged as non-clusterable. We re-ran CATS using a modified input file that explicitly includes the low probability phi coordinate of the 6th residue with no observable hairpin conformations (within the first 40 clusters). The frames captured by the 4th GROMOS cluster appeared to be spread over multiple low-ranked CATS clusters due to other dihedral coordinate variations. Despite similar RMSD values, the coordinate index values vary enough to eliminate the hairpin as a top ranking cluster.

2.3.2 Cluster probabilities based on RMSD derived populations can be misleading

The TMAO and urea PMF both resemble an oblong shape with one local minimum for TMAO and two local minima for urea. The lack of resolution of the PMF contour maps is also clear from the tendency of GROMOS and CATS clusters to not always center on a local minimum, which would occur for highly populated clusters. Conversely, the 2N-dimensional PMF landscapes shown in figure 2.3A-D contain higher resolution minima because each CATS cluster occupies a local minimum within the 2N-dimensional energy landscape. Additionally, the center structure produced by CATS for each cluster is the closest to the 2N-dimensional minima. Other members in the cluster vary within the energy landscape around this central point. The RMSD distribution of the GROMOS method illustrates that pairwise RMSD values have a single distribution curve and cannot take into account the PMF at high resolution. The implication of using an RMSD cutoff forces structures that may deviate slightly from the center of a local minimum into another cluster all together. In this case, probabilities derived by GROMOS based on populations of the cluster may not accurately describe stable states of high probability structures.

2.3.3 CATS clusters structures that GROMOS might have missed

GROMOS clusters have a strict clustering cutoff based on RMSD. Figure 2.6 illustrates two structures that would have normally been placed in separate clusters due to their large differences in Ree. Figure 2.6A illustrates the similarity between the dihedral angle coordinates of the clustered structures, despite different end-to-end distances. The two structures are members of the same cluster produced by CATS; one being the center structure, and one being a large deviation in terms of end to end distance. Since CATS produces the center structure at the closest possible minima point of the 2N-dimensional energy landscape for the cluster, any deviations in clustering coordinate would produce a structure at a higher PMF as calculated by the error function in eq. 2.4. The larger PMF due to backbone deviation does not move the conformation into another 2N-dimensional local minimum (at another cluster), but it does imply that the conformation of the peptide will shift to resemble the center structure under the influence of the PMF. Because CATS clusters form at local energy minima, cluster members can describe how the center structure conformation (given by eq. 2.4) is perturbed around the local minimum point.

2.3.4 Clusters produced by CATS have larger structure variations than clusters produced by GROMOS

Figure 2.7 illustrates the pairwise deviations between center structures produced by CATS and GROMOS methods. In both the TMAO and urea clusters, CATS tends to produce more unique centers than the GROMOS method. In the TMAO clusters, both methods appear to produce similarly varying cluster centers until cluster 12, were CATS begins to generate centers that have large RMSD values compared to previous structures. This trend can be seen by comparing the Fano factor [46] of the two methods. The Fano factor can be calculated using the following equation:

$$F = \frac{\sigma^2}{\mu} \tag{2.5}$$

where σ^2 is the variance and μ is the average of a random variable.

The top 20 TMAO structures for GROMOS and CATS produce a Fano factor of 0.36 and 0.65 respectively. Additionally, CATS produces structures with large comparative RMSD values as shown in figure 2.7. GROMOS appears to produce 3 centers (cluster 13, 19 and 20) that have large RMSD values compared to other centers. For the top 20 urea structures, GROMOS and CATS produce a Fano factor of 0.54 and 0.60 respectively. The Fano factor (eq. **??**) is a quantitative measurement of the noise in a given variable, which

in this case is the average RMSD between cluster pairs. Comparing the pairwise RMSD matrix in figure 2.7 and the Fano factors for the two methods suggests that CATS and GROMOS methods do not share the same top 20 cluster centers. Figures 2.4 and 2.5 illustrated how CATS and GROMOS methods produced similar structures with alternate rankings for the top 4 clusters. This trend does not appear to continue according to figure 2.7 A and B, which suggests that CATS clusters have larger pairwise RMSD values after the 4th top ranked cluster, meaning CATS and GROMOS methods have less agreement for center structures at lower ranked clusters.

2.4 Discussion

2.4.1 The cluster representation (center structure) produced by CATS captures energy landscape minima with better resolution than the center structure of RMSD-based algorithms

In our comparison of the top 20 cluster centers produced by the GROMOS and CATS methods (figure 2.7), we observed that CATS produces a greater number of unique center structures. Visually, the top 20 CATS cluster centers have a larger averaged RMSD than the top 20 GROMOS structures for both

TMAO and urea. We quantify this observation with the Fano factor, where CATS clusters possess a larger Fano factor in both cases. This behavior can be observed in the PMF plots for the urea and TMAO trajectories (figure 2.4 & 2.5); the cluster members were shown to sample a larger area on the PMF plot than the members of the GROMOS clusters. Since CATS produces distinct clusters using bimodal dihedral distributions, we expect each cluster group to have distinct features. The majority of the bimodal coordinate distributions have large differences in their average peak values, which are reflected through the large deviations between cluster centers. Since the GROMOS method produces a continuum of different clusters (depending on the cutoff value), there is no guarantee that the conformations of each cluster will be relatively unique. Therefore, the clusters produced by CATS will not sample from the same areas in an energy landscape defined by dihedral angles. This is a feature of RMSD minimization used by the GROMOS method. On the other hand, CATS captures the most probable conformation for IDPs, constrained by the dihedral coordinate energy landscape, where RMSD may not be small. Due to this phenomenon, CATS and GROMOS will assign different ranks/probabilities to similar structures, as observed in figures 2.4C-D and 2.5C-D. Thus, IDP-like proteins that have a larger RMSD but exist in the same energy state (as described by the dihedral PMF) may not be grouped together using the GROMOS method.

2.4.2 CATS captures alternate high-probability structures suitable for IDP simulations

The fourth largest cluster produced by the GROMOS method for the R2/urea trajectory, which is a beta-hairpin, was not captured by CATS (within the top 40 structures) despite its relatively large population. CATS appears to capture urea structures that mainly lack conformation. The discrepancy between the CATS and GROMOS top 4 cluster structures comes from the phi/psi coordinate distributions in residues 5-7. The phi dihedral angle PMF for the 6th residue (shown in figure 2.3C) contains a bifurcation in the distribution curve, where 1% of the phi distribution exists in the 100 degree region, while 99% exist in the -100 degree region. Not only does CATS reject low populated distributions (1% or lower by default), the construction of the distribution probability inputs also ignored the peak, therefore the hairpin conformation was deemed an "unclustered coordinate" because the dihedral angle values of the coordinate did not fit into any indexed distribution. CATS was re-run on the R2/urea trajectory with this peak accounted for, however the top 20 structures/clusters still did not contain the beta-hairpin conformation. The coordinate indexes of the CATS transformed trajectory frames that make up the hairpin cluster (GROMOS cluster 4) showed several "unclusterable" coordinates randomly throughout the cluster. We attribute this phenomenon to the hairpin structures belonging to distribution regions that poorly resemble a normal distribution. We observed several coordinate distri-

butions that do not converge to zero population between distribution peaks, which is where several hairpin coordinates reside. If the hairpin structures were a result of poor simulation sampling, and all distribution peaks obey the law of large numbers, then CATS could successfully have filtered out "noise" since at some limit the population of hairpin conformations would be minuscule compared to others. On the other hand, if the distributions never approach an ideal Gaussian distribution, then CATS would not be as apt for clustering (at least with the choice of dihedral coordinate distributions).

2.4.3 Poorly defined coordinate distributions and relaxation affect accuracy of CATS

CATS was developed to specifically analyze a trajectory that contains points of variation for the protein backbone. Coordinates with multiple distinct distribution peaks permit clustering. The limit of resolution in this scheme comes from the quality of the distribution peak and the number of coordinates used to describe the structure as a descriptor. The R2 fragment has 24 coordinates total (a phi and psi dihedral angle describing relative orientation for all 12 residues); therefore the maximum possible combination of coordinate distributions (assuming that each coordinate had two peaks in its distribution) is 2^{24}. This combinatorial maximum, however, does not occur due to backbone constraints imposed on the protein during the MD simulation. In the initial analysis of the MD trajectory (II.1.B), the de-

scriptor coordinate distributions are generated. The distributions alone do not characterize the constraints on the number of possible coordinate peak combinations. The backbone constraints, which forbid dihedral coordinate combinations that produce steric clashes or other unrealistic structures, are taken into account by clustering distribution peak combinations that exist in the trajectory. Once this analysis is complete and clusters are formed, the center structure of each cluster is generated based on the cluster member with the lowest error or deviation (see eq. 2.4) from each coordinate's distribution peak average. In this way, the representative center structure of a CATS cluster shares similarities to expectation-maximization (EM) based algorithms [171, 28], while the relaxation routine encompasses the idea of fuzzy clustering [39, 44, 26, 17, 62, 61, 53]. The concept behind CATS relaxation is that a protein will still perform a function similar to the cluster representative structure despite a lowered degree of membership. For example, a relaxation factor of 2 ignorable coordinates in the R2 fragment implies that all members of the cluster are at least 91% the same as with the representative structure.

CATS has the ability to cluster any generalized coordinate with separable Gaussian-like distributions. Depending on the data used for cluster analysis, this can be a positive or negative attribute. Coordinates with large standard deviations will produce clusters that reflect a large range of variation in such a coordinate. Conversely, narrow distributions will produce well defined clusters with low variation. These properties can be exacerbated based

on the coordinate chosen for clustering. In this study, we used the dihedral angle coordinates to form distributions of the R2 fragment in TMAO and Urea. We chose backbone dihedral angles phi and psi as our clustering coordinates due to their Gaussian-like distributions throughout the trajectory, independence from Cartesian coordinate positions (the R2 trajectories did not need to be aligned prior to clustering), and condensed representation of relative backbone orientation. In our comparison of CATS to GROMOS, we analyzed the top four clusters as functions of Rg and Ree. The clusters from both methods produced a high density population centered about a specific Rg and Ree region. The clusters produced by CATS periodically included members that appear to center about a different point on the density map, which was not observed in the GROMOS results. This feature is most likely due to a larger deviation of dihedral coordinates near the middle residues of the R2 fragment (residues 5, 6, or 7) as observed in figure 2.6. The distance travelled by a point on the N or C terminus due to a change in a central residue dihedral angle is proportional to the distance between the central residue and the terminus point. Since CATS is not optimized to minimize RMSD distances, clusters produced using dihedral coordinate distributions have the possibility of including members that seem out of place when analyzing member conformations as a function of Rg and Ree. As with all statistical methods, a larger sample size will improve accuracy. The larger Rg and Ree of CATS may not be improved by greater sample size since this is an artifact of the standard deviation exhibited by the coordinate distri-

bution. Coordinates with significantly large standard deviations may not be suited for CATS clustering since structure information cannot be partitioned out.

2.4.4 CATS is better suited for input parameter tuning than the GROMOS method

Determining the correct cutoff/input parameters for clustering proteins occurs on a case-by-case basis. For the GROMOS method, one must supply an RMSD cutoff that is both large enough to cluster a reasonable number of structures together while simultaneously filtering out dissimilar structures. Similarly, CATS requires the user to supply a standard deviation cutoff for the coordinate distributions so that the majority of structures can be grouped together. An important distinction between CATS and GROMOS methods with respect to these user-supplied cutoff values is that CATS will generally produce the same center structures regardless of standard deviation used; tuning the inputs for CATS will affect the relative populations of clusters only. On the other hand, tuning the RMSD cutoff values in the GROMOS method will affect both the center structure and populations of clusters. Determination of ignorable coordinates must also be made since volatile coordinates that have little significance on the overall structure (such as the dihedral angles of the first and last residues) should be ignored to avoid over constrained clusters. A practical consideration of the parameter tuning pro-

cess is the computational resources required for each method to complete the task. The RMSD-based clustering method in GROMOS requires significantly more computational time than CATS. Due to the pairwise RMSD calculations performed by GROMOS, the number of calculations required to generate the observed clusters is much greater than the number of calculations performed by CATS. In this study, we used the 'backbone' option for GROMOS, requiring 3 atom pairwise calculations per residue per frame. Each frame comparison requires a symmetric pairwise RMSD matrix to be formed. Therefore, the number of calculations required by GROMOS scales proportional to roughly O((3N)!) where N is the number of residues. CATS transforms the trajectory in a simple comparison step proportional to 2N, then clusters through iterative comparisons between transformed frames, which requires roughly O(2N) if using the dihedral coordinate descriptors from our studies. For larger proteins or larger trajectory sizes, the computational cost will increase dramatically. Although the CATS algorithm produces clusters significantly faster than the GROMOS method, there is a drawback of CATS when producing input parameters. We chose to neglect input parameter timing factors because there are multiple methods for input parameter generation, all of which are dependent on the trajectory. The automated fitting of Gaussian curves to the coordinate distributions based on our in-house numerical solver adds a trivial amount of time to CATS cluster production. This software, however, is in the early development stage; therefore its results can be inaccurate for complex distributions. The user

can produce input files using Gaussian fitting software from other sources, as well as a manual fitting tool, which was utilized here.

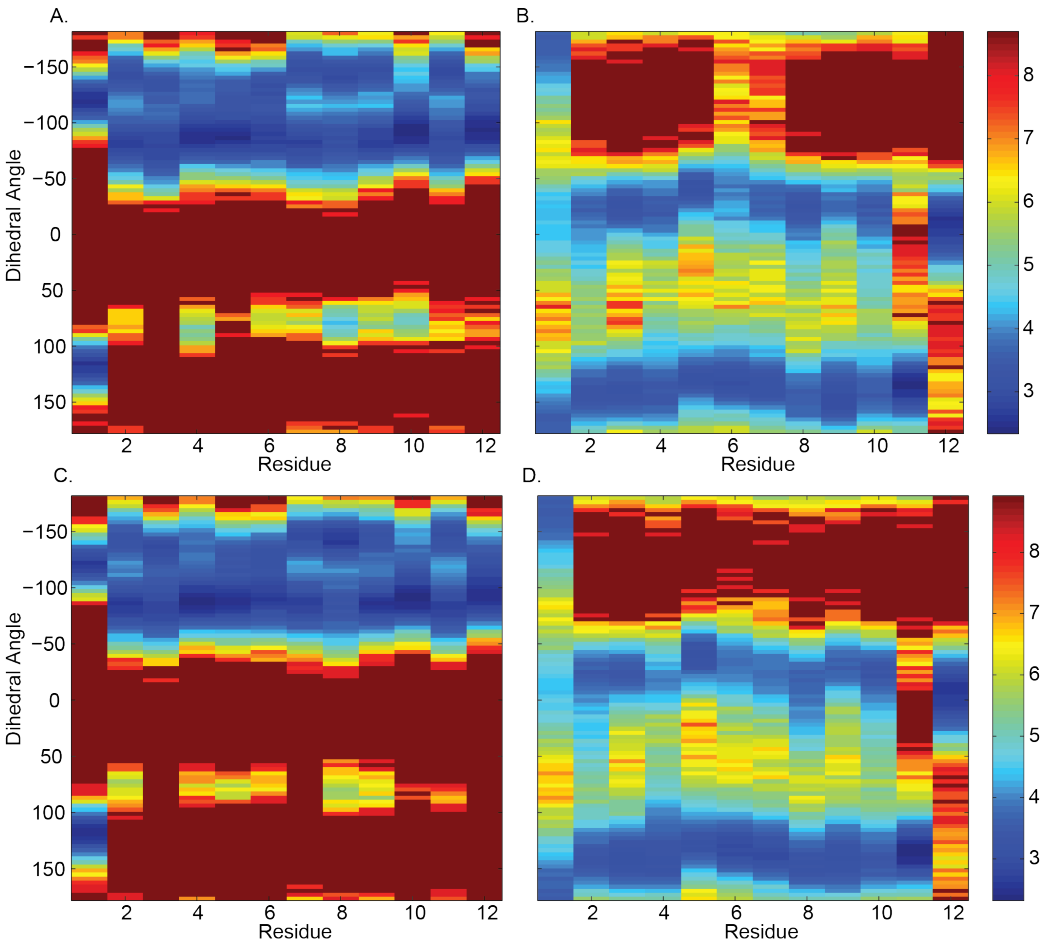

Figure 2.3: The R2/TMAO potential mean force in units of 1/kT is shown for (A) the phi dihedral angle distribution and (B) the psi dihedral angle distribution. The R2/urea potential mean force in units of 1/kT is shown for (C) the phi dihedral angle distribution and (D) the psi dihedral angle distribution.

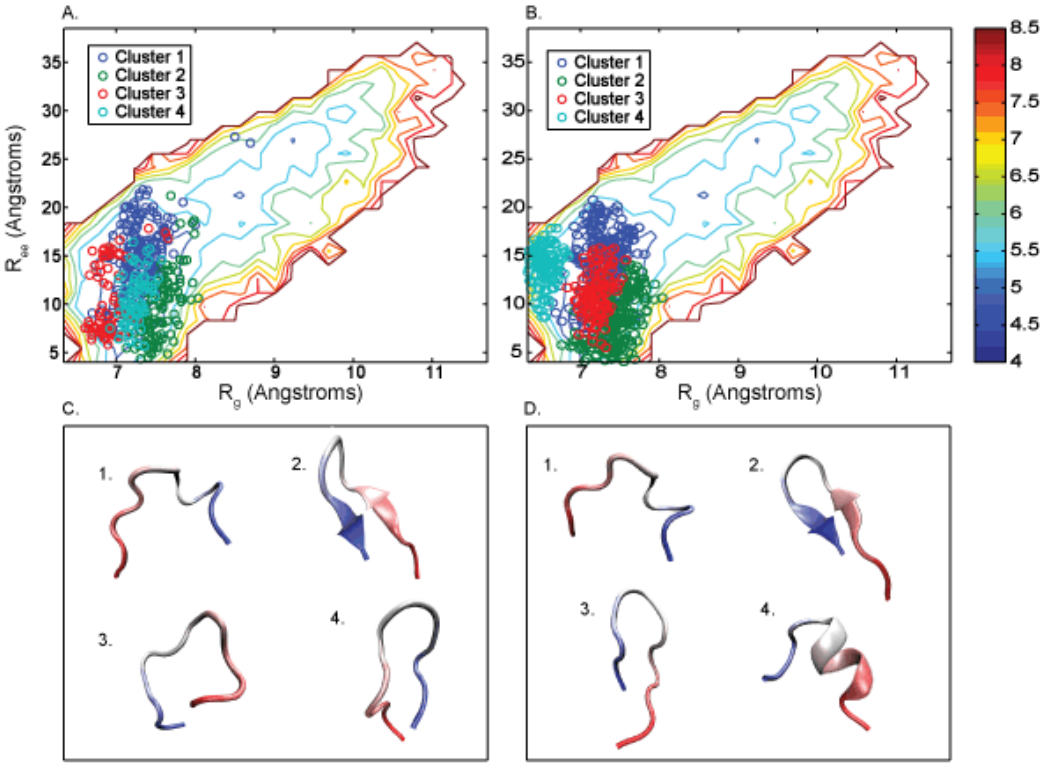

Figure 2.4: R2/TMAO top cluster member comparison. The top four most populated clusters of R2/TMAO produced by (A) CATS and (B) GROMOS methods are compared using the potential of mean force (in units of 1/kT) as a function of peptide conformation at room temperature over 100ns. The automated "center" structure (defined by eq. 2.4 for CATS and by the smallest average RMSD for GROMOS) of each cluster is shown for (C) CATS and (D) GROMOS.

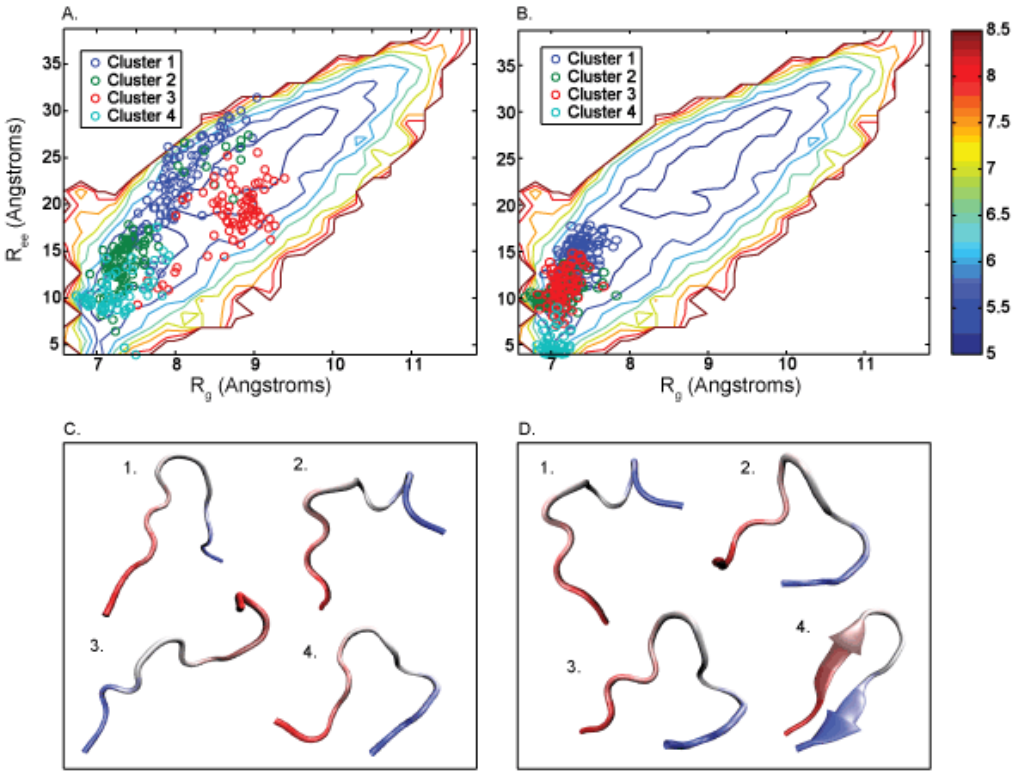

Figure 2.5: Urea top cluster member comparison. The top four most populated clusters of urea produced (A) CATS and (B) GROMOS methods are compared using the potential of mean force (in units of 1/kT) as a function of peptide conformations at room temperature over 150ns. The automated "center" structure (defined by eq. 2.4 for CATS and by the smallest average RMSD for GROMOS) of each cluster is shown for (C) CATS and (D) GROMOS.

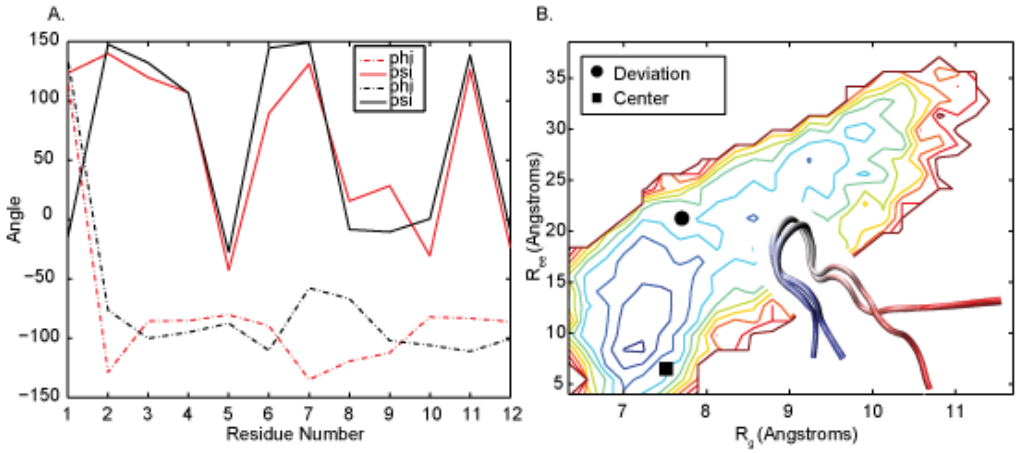

Figure 2.6: Using the first cluster generated by CATS for the R2/TMAO trajectory, the center structure and structure with the largest end-to-end distance is compared through (A) the phi and psi dihedral angles of the center structure (shown in black) and the deviation structure (shown in red) for each coordinate. (B) The end-to-end distance and radius of gyration for the two structures are shown in comparison to the R2/TMAO potential mean force map shown in figure 3 with the two structures overlaid to illustrate conformational differences.

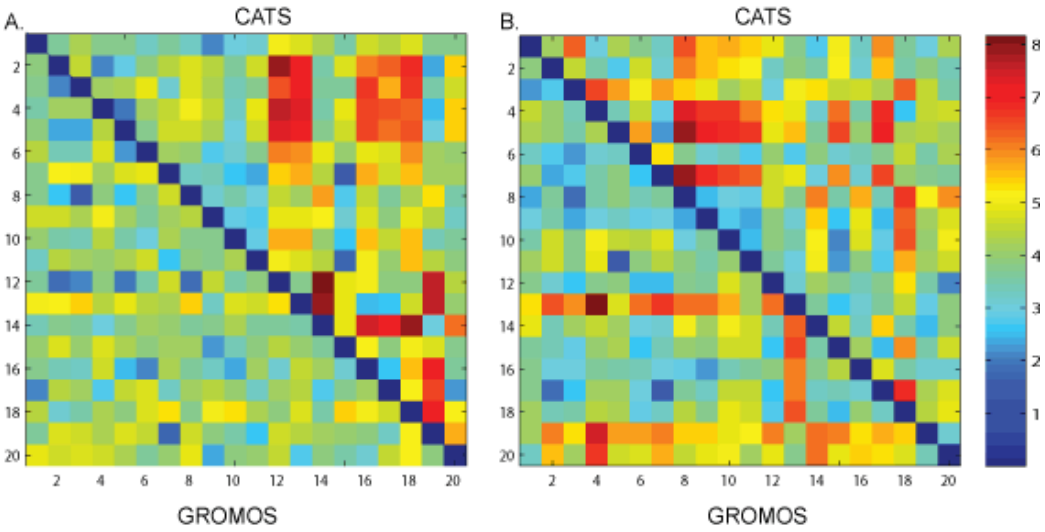

Figure 2.7: Pairwise backbone RMSD values (in angstroms) between center structures of the top 20 ranked clusters are shown for (A) R2/TMAO and (B) R2/urea. Cluster numbers are shown on the x and y axes, with 1 being the most probable cluster. The upper triangle illustrates the RMSD value between a given pair of cluster centers for CATS, and the lower triangle illustrates the RMSD value between a given pair of cluster centers for GROMOS.

Chapter 3

Molecular Dynamics Ensemble Refinement of Intrinsically Disordered Peptides According to Deconvoluted Spectra from Circular Dichroism

3.1 Introduction

It is challenging to determine the structural feature from an ensemble of IDPs. Popular methods for experimental structure determination of proteins, such as Cryo-EM or crystallography, are insufficient to independently determine

the structure of IDPs [168]. Solution experimental methods, such as nuclear magnetic resonance spectroscopy (NMR), are only able to produce an ensemble averaged structure, thus additional analysis must be performed in order to generate the structural ensemble [16]. Computational approaches, such as molecular dynamics (MD) simulations, are also used to generate IDP structures, however these methods largely rely on Hamiltonians whose coefficients are tuned using experimentally determined structures of stable proteins, resulting in overly biased structures [7, 68, 98]. To alleviate these drawbacks, combined computational and experimental approaches have also been used [55, 119, 13, 111]. A necessary feature of these combined approaches is the conversion between experimental observables and computationally generated structures. NMR structure back-calculations [172] use a database that provides a relationship between known structures and chemical shifts. Unfortunately, the NMR chemical shift databases consist of stable proteins with α-helix and β-sheet structures, instead of IDPs. The relationship between the spectroscopic observables and the distinguishing feature of a given protein is deconvoluted from the set of reference structures so that the features of proteins with unknown structures can be determined. This is the typical method for generation of computational models and force field refinement using spectroscopic methods [13, 172, 96, 19, 77, 163]. Despite the popularity of NMR analysis, there are several advantages to using CD spectroscopy for the analysis of IDPs in certain circumstances. CD measurements are of low cost, can be quickly performed and require a small amount of sam-

ple material [90, 166]; however, they cannot provide high-resolution (residue specific) structure approximations. We used several standard CD deconvolution algorithms in conjunction with the SDP-48 CD reference dataset, including SELCON3, CONTIN-LL, and CDSSTR, to analyze our IDP experimental CD spectra. The resulting conformational solutions did not converge, prompting us to speculate that there may be features within these standard algorithms that are incompatible with systems containing short IDPs. In this study, we have removed the bias in the deconvolution analysis of the CD spectra by developing a non-negative least squares (NN-LSQ) fitting method using a protein database that includes denatured proteins. We have produced the features of secondary structures from CD spectra and used them to extract an ensemble of structures from all-atomistic molecular dynamics simulations (MD). We applied this approach on refining the structures of a set of small disordered peptides derived from the calmodulin (CaM)-binding domain of calcium/CaM-dependent protein kinase II (CaMKII, 293-312) and its 1-amino-acid and 3-amino-acid mutants (see Table 3.1 for the amino-acid sequences). These peptides were chosen for detailed examination because they show significant differences (up to 6-fold) in association rates when interacting with CaM in solution at physiological ionic strength [164], with as yet no understanding of the underlying mechanism. With our combined approach of CD experiments and MD simulations, we have unexpectedly discovered that the increase of secondary structures in a particularly revealing peptide mutant (AAA) was due to the formation of a β-hairpin conforma-

tion that we propose leads to a frustrated state, limiting the overall encounter rate. Obtaining the structural ensembles of CaMKII peptides was a necessary and essential step towards a more accurate estimation of their binding rates for CaM and presently serves as a novel example for how secondary structure can be a barrier to productive protein-protein interactions. More broadly, the development of a new algorithm that is inclusive of disordered protein states for CD deconvolution should find widespread use.

3.2 Materials and Methods

3.2.1 Peptide synthesis and preparation

The three 20 amino-acid long peptides used in this study were modeled after the CaM-binding domain of CaMKII (residues 293-312; see amino acid sequences in Table 3.1) and were synthesized by LifeTein LLC. Their purity was greater than 95% and the composition of each peptide was validated by mass spectroscopy.

Table 3.1: CaMKII (293-312) peptide sequences are shown with mutated residues in red.

Peptide	Sequence
RRK (Wildtype)	FNARRKLKGAILTTMLATRN
RAK (Mutant - 1 site)	FNARAKLKGAILTTMLATRN
AAA (Mutant - 3 sites)	FNAAAALKGAILTTMLATRN

3.2.2 Measurement with CD spectroscopy

Far-UV CD spectra were collected on a JASCO-815 spectrophotometer controlled by Spectra Manager software. Suprasil cuvettes with a 1.0 mm path length were used for all experiments. The spectrometer parameters were typically set to the following unless noted otherwise: band width, 1 nm; response time, 1s and data pitch, 0.2 nm/min. A solution consisting of 100 μM peptide was made using 10 mM Tris buffer at pH 7.5 and measurements were taken by scanning the excitation wavelength between 190-260 nm with temperature controlled at 20°C. A total of 10 data accumulations for each run were made with a sweep rate of 100 nm/min. Data collection was repeated for each peptide a total of 3 times, using a freshly prepared sample in each run.Deconvolution of CD using a standard package: The CDPro software package suite [148, 149] was used to deconvolute the experimental CD spectra of the wildtype and mutant CaMKII peptides. We used the soluble and denatured protein (SDP-48) datasets in conjunction with the CDPro standard numerical fitting methods: CDSSTR, CONTIN/LL, and SELCON3 [148, 149]. Because CDPro gives reliable results with CD data in the range of wavelengths 190-240 nm when a large reference set is used (such as SDP-48) [148], we input our data in the same range in increment of 1 nm. The resulting structure approximation is presented as fractional values for six main secondary structure categories: helix (regular), helix (distorted), strand (regular), strand (distorted), turn, and unordered. We generalized the secondary structure codes into four main categories by consolidating the helix

Table 3.2: Consolidation of CDPro and Dictionary of Secondary Structure of Proteins (DSSP) structure annotations into generalized helix, strand, turn and unordered categories. We choose a consolidation scheme similar to Kardos et al [107], where the pi-helix secondary structure is counted as un-ordered due to its lack of distinction as a stable secondary structure. DSSP was implemented using the AMBERTOOLS trajectory analysis software CPPTRAJ, which contains an alternate set of structure codes despite using the DSSP algorithm.

Defined Structure Categories	CDPro Structures	DSSP Structures	CPPTRAJ implementation of DSSP
Helix	Helix(regular)	α-helix	α-helix
	Helix(distorted)	3-10 helix	3-10 helix
β-sheet	Strand(regular)	β-strand	Parallel β-sheet
	Strand(distorted)		Anti-parallel β-sheet
Turn	Turn	Turn	Turn
		Bend	
Unordered/other	Unordered	π-helix	π-helix
		β-bridge	None
		Irregular/loop	
		Turn (1-residue)	
		Bend (1-residue)	

(regular) and helix (distorted) into the helix category, and strand (regular) and strand (distorted) into the strand category for comparison of structure fractions produced by other analysis methods (see Table 3.2).

3.2.3 Deconvolution of CD data using Non Negative – Linear Square (NN-LSQ) fitting

We deconvoluted the experimentally determined CD spectra of the wildtype and mutant CaMKII peptides using a NN-LSQ fitting method. To do so, we used the reference dataset SDP-48, which consists of CD spectra of 48 soluble

and denatured proteins, as the basis spectra. We assumed that a linear combination of CD spectra from the reference protein database is sufficient to approximate the experimental spectra seen in RRK, RAK or AAA, and that the CD spectra of the reference proteins are linearly independent. The squared difference between any non-negative linear combination of the CD basis spectra and the experimental CD spectrum δ^2 is minimized by finding the optimal weight coefficients $\vec{x}$ as shown in Eqn. 3.1

$$\delta^2 = \left| \mathbf{C} \cdot \vec{x} - \vec{b} \right|^2 \tag{3.1}$$

where C is the 51x48 matrix representing the 51 CD spectrum points for all 48 reference proteins of SDP-48, $\vec{x}$ is a vector of the weight coefficients for the reference proteins, and $\vec{b}$ is the 51 by 1 vector of the experimentally measured CD values of the CaMKII peptide in the 190 to 240 nm wavelength range. The weight coefficients vector, $\vec{x}$, is determined by NN-LSQ fitting, and each coefficient is required to satisfy $x_i \geq 0$ to account for the possible differences in the signal amplitude in our experimental results and the reference database CD spectrum. We subsequently use the fitted weight coefficients of $\vec{x}$ to compute the secondary structure fractions given by Eqn. 3.2,

$$\vec{d} = \mathbf{A} \frac{\vec{x}}{|\vec{x}|} \tag{3.2}$$

Where $\mathbf{A}$ is the 6x48 matrix representing the 6 possible secondary structure fractions for each of the 48 reference proteins, and $\vec{d}$ is the 6x1 secondary

structure solution for the CaMKII peptide. The resulting structure approximation is presented as fractional values for six main secondary structure categories: helix (regular), helix (distorted), strand (regular), strand (distorted), turn, and unordered. We generalized the secondary structure codes into four main categories by consolidating the helix (regular) and helix (distorted) into the helix category, and strand (regular) and strand (distorted) into the strand category for comparison of secondary structure fractions produced by other analysis methods (see Table 3.2).

3.2.4 All-atom molecular dynamics (MD) simulations with implicit solvent of the peptides

3.2.4.1 MD setup and initialization

Because there is no high-resolution solved structure due to the disordered nature of the CaMKII peptides, we built the initial structures for molecular dynamics (MD) simulations using the LEaP module of AMBERTOOLS 14 [18] based only on the amino acid sequences (Table 3.1). The N and C terminals of these peptides were not capped or modified, consistent with the experimental study [164]. All molecular dynamics (MD) simulations were carried out using the package AMBER 14 with the ff99sb force field [68, 18]. We used an implicit solvent model with the Generalized Born [150, 118, 113] approximation and the modified Born radius parameter set mbondi2 [118]. We performed energy minimization on the initial structures using 1000 steps

of conjugate gradient followed by 1000 steps of steepest descent algorithms. The minimized structures were brought to the desired temperatures in two steps: heating each minimized structure to 277K, 285K, or 293K, followed by a simulated annealing cycle. Simulated annealing was carried out by heating structure coordinates obtained in the previous step to 400K over a period of 600ps, followed by cooling to the designated temperature over a period of 600ps with velocity randomization every 100ps. All setup runs used a time step of 2 fs. We restrained hydrogen dynamics by employing the SHAKE algorithm [137]. We used Langevin dynamics with a collision frequency of $2ps^{-1}$ to regulate the temperature (Langevin thermostat); periodically randomizing the velocity distributions was therefore necessary to avoid synchronization effects associated with Langevin thermostats [144].

3.2.4.2 MD production runs

We performed all-atomistic implicit solvent simulations for each CaMKII peptide at 277K, 285K, and 293K, replicating the operating temperature of the stop-flow kinetics experiment [164], a mid-point temperature, and the operating temperature of the CD measurements, respectively. The production run was performed at the designated temperature for a period of 80 ns with a 2fs time step. We sampled energy and trajectory data every 4ps, which was determined through correlation time analysis. All simulation steps from the setup and production runs were repeated an additional 14 times for every temperature/peptide combination, resulting in a total production run

simulation time of 2.4 μs (per peptide per temperature). Trajectories were tested for convergence using two approaches: Kullback-Liebler (KL) divergence [38, 89] between distributions of the potential energy in accumulated simulation time, and cluster analysis with respect to simulation time.

3.2.4.3 Data-guided extraction of all-atom peptide conformation ensembles

Determination of the secondary structure content in MD trajectories: The secondary structure content of the peptides was computed using the CPPTRAJ module of AMBERTOOLS[133], which calculates structure content based on the Dictionary of Secondary Structures of Proteins (DSSP)[78]. The results of our structure analysis generated 7 possible secondary structure categories per residue: α-helix, parallel β-sheet, anti-parallel β-sheet, 3-10 helix, π-helix, turn, and unordered. We consolidated the 7 secondary structure categories into 4 generalized secondary structure categories (see Table 3.2) and generated a histogram of the structure codes associated with each residue to produce the overall fractional secondary structure values in each trajectory frame.

3.2.4.4 Refinement of IDP ensemble structures from MD using CD deconvolution data

Using the secondary structure data for each frame of our MD trajectories, we selected pairs of trajectory frames that produce average secondary structure

fractions similar to those observed in the CD deconvolution data from our NN-LSQ fitting. For a given peptide trajectory, frames are extracted in pairs if the following equality is satisfied for each structure fraction:

$$\left| \frac{S_i^k + S_j^k}{2} - S_0^k \right| \leq \sigma \tag{3.3}$$

where S_i^k and S_j^k are the fractional values for the kth structure category (either helix, β-sheet, turn or unordered secondary structure categories) for frames i and j, and S_0^k is the structure fraction for category k derived from our NN-LSQ deconvolution results. Selected structures represent the refined IDP ensemble. Clustering is performed on the obtained ensembles to derive representative structures shown in Fig. 3.4.

3.2.4.5 Contact map analysis

CD-guided MD structures of the peptides from the CD-refined ensemble were used for contact map analysis. The definitions are described as follows: A contact between residue i and j (at least 4 residues away) is formed if any atom from residue i is within a cutoff distance of 4Å of any atom from the residue j. A backbone (sidechain) contact between residue i and j (at least 4 residues away) is formed if any backbone (sidechain) atom from the residue i is within a cutoff distance of 4Å of any backbone (sidechain) atom from the residue j. A single hydrogen atom from glycine is considered as its sidechain. A hydrogen bonding contact between residue i and j is formed if a donor

atom (D) from residue i is within a cutoff distance of 4Å of an acceptor atom (A) from residue j, and the D-H-A angle through a bonding hydrogen (H) is within a cutoff angle of 30°.

3.3 Results

The results of our CD spectra indicate a distinct secondary structure shift between RRK and AAA. The CD spectra presented in figure 3.1 shows the average secondary structure ensembles of RRK, RAK and AAA peptides. Typically, a negative CD signal at 220 nm indicates the presence of helical or strand structures, and a negative signal at 195 nm corresponds to denatured/disordered structures [56]. Here, the experimental data shows the existence of a secondary structure in AAA that does not exist in RRK or RAK. This data suggests that each charged residue mutation reduces the disordered content of the peptide's structure ensemble. Overall, the charged residue mutations result in a significant conformational change from the disorder in RRK to the more ordered structures in AAA. We first speculated the increased structures in AAA was due to helical secondary structure as alanine residues have the highest propensity to form α-helices [121]. However, the experimental CD spectrum for AAA displays only one negative peak at 222 nm but is missing a second smaller signal peak at 208 nm, which is a hallmark of alpha helical regions in CD spectra [56]. This indicates that there is a mixture of secondary structure components in the peptides. Therefore,

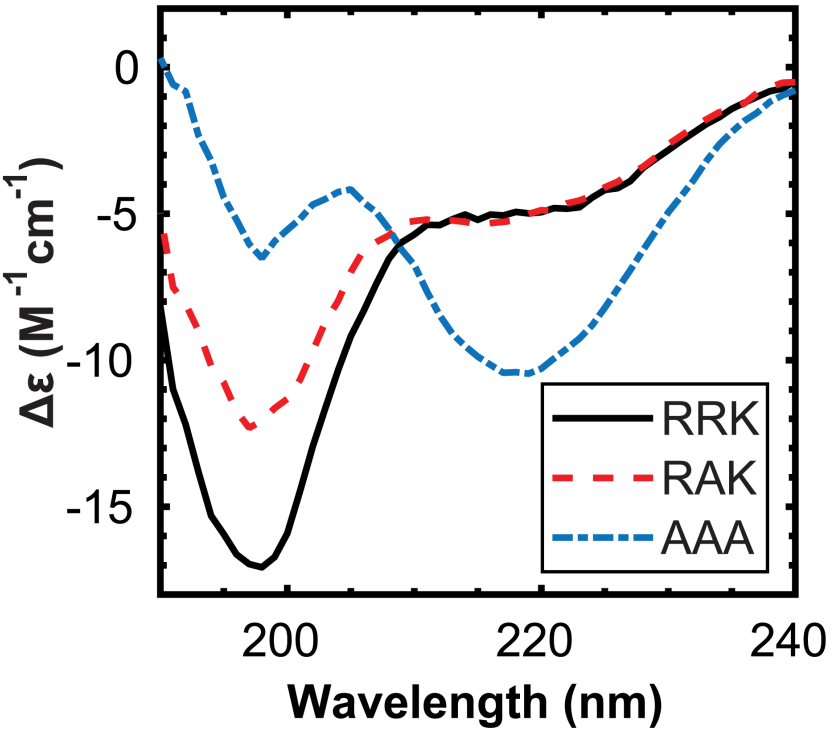

Figure 3.1: **Far-UV CD spectra of the CaMKII peptides.** The circular dichroism (CD) experimental spectra is shown for the RRK, RAK and AAA peptides. A solution consisting of 100 μM of each peptide was made in 10 mM Tris buffer at pH 7.5 and measurements were taken in a 1.0 mm quartz cuvette by scanning the excitation wavelength between 190-240 nm with temperature controlled at 20°C.

Table 3.3: Fractional secondary structure approximations are given for the CONTIN/LL, SELCON3, CDSSTR and NN-LSQ fit CD deconvolution methods. The approximate CD spectrum representing the CaMKII peptides is recreated from a linear combination of SDP-48 known conformation/spectra definitions that we developed. The unitless RMSD between the approximated and experimental spectrum ($\delta\epsilon$) is given. *SELCON3 was unable to converge for the RRK spectrum; the solution shown is the result of partial completion of the SELCON algorithm.

		Helix	Strand	Turn	Unordered	RMSD
RRK	*SELCON3	0.00	-0.06	-0.07	1.28	15.86
	CDSSTR	0.15	0.32	0.28	0.24	1.38
	CONTIN/LL	0.01	0.01	0.10	0.88	0.35
	NN-LSQ	0.04	0.15	0.09	0.72	0.26
RAK	SELCON3	0.04	0.03	0.01	0.94	4.72
	CDSSTR	0.17	0.29	0.23	0.31	0.84
	CONTIN/LL	0.03	0.02	0.08	0.87	0.40
	NN-LSQ	0.04	0.22	0.13	0.62	0.23
AAA	SELCON3	0.29	0.20	0.19	0.36	3.26
	CDSSTR	0.37	0.30	0.16	0.17	0.63
	CONTIN/LL	0.03	0.05	0.30	0.62	0.44
	NN-LSQ	0.04	0.34	0.17	0.46	0.36

we employed CDPro to deconvolute the CD spectra on the three peptides in the next section.

3.3.1 Standard CD deconvolution solvers produce inconsistent results on the content of secondary structures

We have used three standard CD deconvolution solvers, CDPro, CAPITO, and BESEL in attempts to analyze the structural information of the peptide spectra. We found that all of them produce inconsistent outcomes, show-

ing non-convergence and large RMSD compared to experimental spectra. (A) CDPro: We generalized the secondary structure codes used by CDPro into four main categories (see Table 3.2). The three standard deconvolution solvers (CDSSTR, SELCON3, and CONTIN/LL) from CDPro generate inconsistent fractions of secondary structures as shown in Table 3.3. The CONTIN/LL method shows that RRK contains mostly turn and unordered secondary structures, however, the CDSSTR method shows that RRK contains similar quantities of structured and unstructured regions. In the AAA deconvolution results, the CDSSTR and CONTIN/LL methods suggest opposing secondary structure content, with CDSSTR resulting in the increase of helical fractions and CONTIN/LL resulting in the increase of turn content. The SELCON3 methods appear to perform the worst among the three, giving large RMSD between the reconstructed CD spectra and the experimental data (figure 3.2) and producing unrealistic fractions of secondary structures. In summary, the deconvoluted data were inconclusive because of disagreement between the three solver methods (https://sites.bmb.colostate.edu/sreeram/CDPro/). (B) CAPTIO: Use of more recently developed tools for the analysis of CD spectra, shows either large RMSD or under-estimates the fraction of unordered secondary structure for proteins with rich disordered segments. Specifically, CAPITO [167], which uses basis spectra for each of α-helix, β-strand, and irregular secondary structures extracted from SP-175, was not able to produce a good fit for the CaMKII peptides. (C) BeStSel [106] : carries out a detailed secondary structure analysis providing information on eight

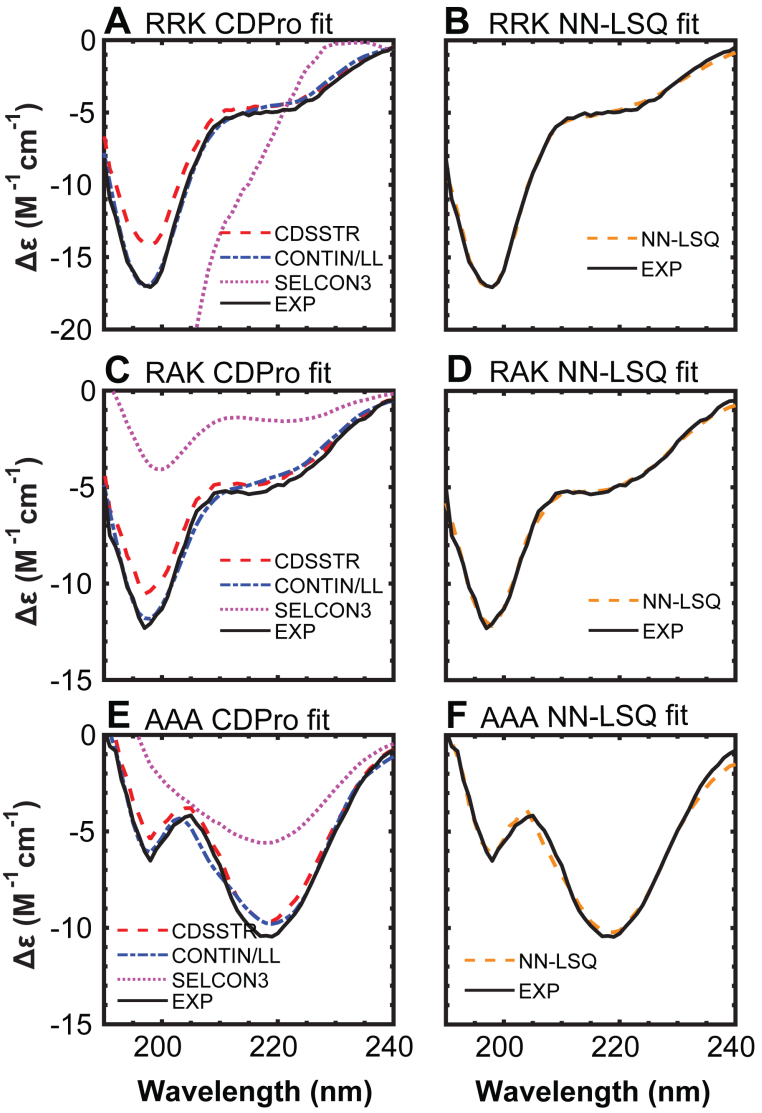

Figure 3.2: Comparison between the fitting of the CD spectra using the CDPro and NN-LSQ fitting. (A, C, E) The experimental CD spectrum is compared to the calculated CD spectrum derived from the CONTIN/LL, CDSSTR, and SELCON3 methods for RRK, RAK and AAA peptides, respectively. (B, D, F) The calculated CD spectrum using the NN-LSQ fitting method and SDP-48 database is compared to the experimental data for RRK, RAK and AAA peptides, respectively.

secondary structure components, and provides improved estimation of the β-strand content. Our analysis of the CaMKII peptides with BeStSel produced relatively large RMSD and reinforces that present CD analysis tools are not useful for this class of peptides.

3.3.2 CD deconvolution with NN-LSQ fitting indicates presence of β-hairpin secondary structure

The inconsistencies associated with the standard deconvolution models prompted us to review the fitting methods from the three standard deconvolution solvers. We noted that the current methods overly favor helical content by fitting the CD spectrum to a dataset of predominantly globular or membrane-bound proteins as well as by employing algorithms emphasizing the weights on helical structures. In order to avert these two issues, we chose to fit the CD spectrum with the data set of denatured proteins (SDP-48) and search for alternative fitting routines. We used NN-LSQ fitting because it simultaneously took into account the data from all protein structures in the SDP-48 reference set and made no a-priori assumptions about the secondary structure. Our NN-LSQ fit deconvolution results presented in Table 3.3 indicate that the primary effects of the mutation in the peptides emerge through an increase in the β-sheet category secondary structure (in red fonts in Table 3.3), while the helical content remains the same. The increase in secondary structure is naturally associated with a decrease in the disordered structure category,

where RRK has the highest disordered content with 72%, and AAA has the lowest disordered content with 46% (according to NN-LSQ fit deconvolution from Table 3.3).

3.3.3 All-atom MD simulations produce strongly biased structure ensembles

To generate an equilibrium ensemble of structures for the three peptides, we employed all-atom MD simulations with implicit solvent at three temperatures: 277K, 285K, and 293K. A total of 2.4μs of data sampled at 4ps intervals was collected for each peptide/temperature combination, and analyzed for their secondary structure content using DSSP. Data produced from this analysis was translated into a four-category generalized secondary structure scheme shown in Table 3.2. The secondary structure fractions for each trajectory were first averaged to illustrate the overall conformational trend produced in each simulation (figure 3.3). The analysis of the secondary structures shows that the MD simulations were incapable of generating an ensemble of structures that match with the CD analyses. More specifically, comparing to the deconvoluted secondary structure fractions from the CD data, the MD ensembles illustrate a significant bias towards helical content. The data for all peptides at all temperatures show the β-sheet content at less than 5% and the helical, turn, and unordered content in the range of 30-40%. Additionally, the DSSP analysis indicates no significant overall sec-

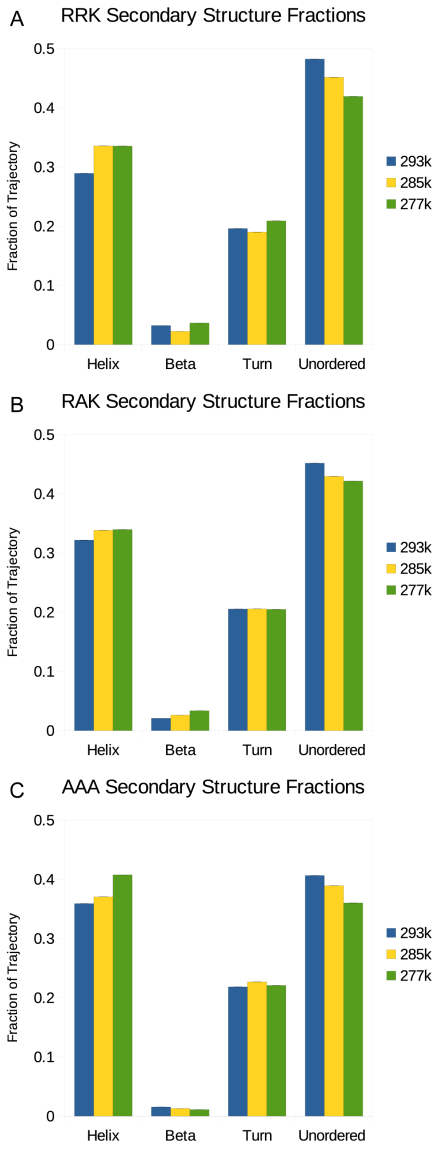

Figure 3.3: Average secondary structure fractions produced by the all-atom CaMKII peptide simulation. A histogram of the secondary structures produced by DSSP analysis for each frame of the CaMKII peptide trajectories are shown for (A)RRK, (B)RAK, and (C)AAA at 277K, 285K and 293K.

ondary structure shift between the wildtype and mutant peptides for all three temperatures. This finding is in contrast to the deconvolution data in NN-LSQ fitting, where the fraction of β-sheet content for RRK is 15% and increases to 22% and 34% for RAK and AAA respectively. In summary, the outcomes from the MD simulations do not appear to accurately represent the secondary structure shift that occurs between mutant peptides as indicated by our CD data.

3.3.4 Approximate structure ensemble of IDPs from all atom trajectories and CD deconvolution

To gain useful information from the MD simulations that agrees with our CD deconvolution data, we select pairs of trajectory frames from the production run with similar averaged secondary structure fractions as those observed in our NN-LSQ fit CD deconvolution data shown in Table 3.3. We analyzed peptide trajectories for the 293K production run using a σ value of 0.035 for each structure category (Eqn. 3.3). We obtained 11002 structures for RRK, 2410 structures for RAK and 130 structures for AAA. Deviations between the number of structures generated for each peptide appears to be correlated with the relative β-sheet content. All MD trajectories displayed poor sampling of β-sheet structures (figure 3.3), which may explain the decreasing number of extracted frames as the β-sheet content for each peptide increases. Although the structural ensemble derived from MD simulations

appeared to be biased, we assume that the force field still samples the correct peptide conformations in significantly smaller quantities. Since spectroscopic methods produce observables corresponding to the ensemble-averaged state, we only require that the extracted MD frames produce an ensemble whose average corresponds to the experimental CD data. Using the solutions obtained from CD deconvolution enables us to separate MD trajectory data that agrees with the experimental data from biased trajectory data.

A set of 10 structures representing each peptide ensemble was generated by clustering. Initially, the Hieragglo clustering method from CPPTRAJ with 10 total clusters was used. The results of this clustering method appear to be misleading due to the disproportionally large populations of the first clusters in RRK and RAK. Since the CaMKII peptides possess significant fractions of disordered content, it is likely that these large clusters have conformational variation within them, and are poor representations of the ensemble. To gain better resolution of the representative ensemble structures, a previously developed clustering algorithm was chosen to resolve the extracted structures. The Combinatorial Averaged Transient Structure (CATS) method has produced better structure resolution for IDPs than traditional clustering methods [45], and is therefore employed in the current study. The selected structures (figure 3.4) from CATS represent a set of highly probable conformations exhibited by the peptides in solution. Based on these representative structures, RRK and RAK display significant conformational variation compared to AAA, which forms compact β-sheet structures. In our NN-LSQ CD

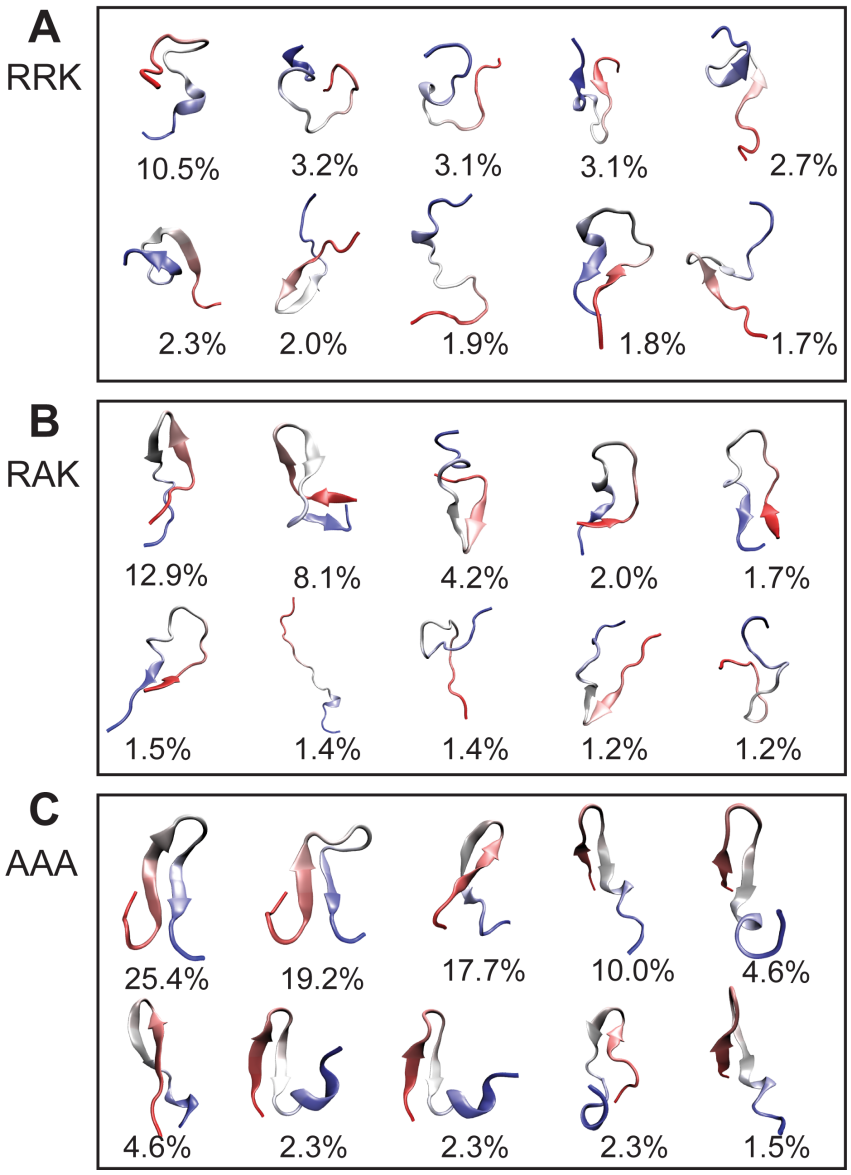

Figure 3.4: Sample conformations of generated ensembles. CaMKII peptide ensembles were generated by selecting MD trajectory frames from the 293K runs with secondary structure fractions that match the NN-LSQ CD deconvolution results. To illustrate structure features, the generated ensembles are clustered using an algorithm designed to cluster IDPs developed from our group: Combinatorial Averaged Transient Structure (CATS) [45]. Center structures from the 10 most populated clusters are shown for (A) RRK, (B) RAK, and (C). The peptides are colored according to atomic index, with the N-terminus shown in red and the C-terminus shown in blue.

deconvolution results, RRK and RAK present a high percentage of unordered structure at 72% and 62%, respectively. On the other hand, AAA possesses a lower degree of unordered structure at 46% (labeled in blue in Table 3.3). This result is consistent with the generated structure ensemble, which possess a maximum RMSD of 12.5Å for RRK, 12.4Å for RAK, and 10.5Å for AAA. The RMSD analysis of the generated ensemble also illustrates that AAA has the lowest standard deviation of RMSD values (0.986Å), compared to RRK and RAK (1.412Å and 1.737Å respectively). An increasing secondary structure content can be observed in figure 3.4 as a result of each sequential mutation of the RRK peptide. This observation is in agreement with the shift in ordered and disordered content predicted by CD deconvolution despite the apparent force field bias observed in the analysis of the complete trajectories (figure 3.3). We acknowledge that the precise quantitative shift in secondary structure fractions in each mutant may not be completely represented by the generated ensembles shown in figure 3.4, however they illustrate the approximate location of residual secondary structure. The set of RRK structures (figure 3.4A) contain relatively small regions of helical and β-hairpin regions. The N-terminus appears to be largely unstructured with the ability to participate in β-strand formation with C-terminus residues. On the other hand, the C-terminus appears to form turn/helical/hairpin structures more readily with other C-terminal or central residues. In the set of RAK structures (figure 3.4B), the presence of β-strand conformations is more prevalent in comparison to RRK. It can be observed that the N-terminus of RAK par-

ticipates in the majority of β-hairpin structure formation. Additionally, the number of structures with turn/helical regions in the C-terminus has decreased with respect to RRK, however this appears to be correlated with the increase in β-hairpin structure formation. Lastly, the set of AAA structures (figure 3.4C) all contain the β-hairpin secondary structure, however there appears to be two distinct variations of the hairpin: a symmetrical β-hairpin structure, and an asymmetrical hairpin-helix structure. The asymmetrical structures begin their hairpin motif closer to the N-terminus, and form a helical structure on the unbound C-terminal tail. Alternatively, the symmetrical structures start forming the hairpin motif in the central region of the peptide, with the N and C terminal binding instead. Contact map analysis shows AAA mutant adopts strong secondary structure formation. In order to gain more insight in the characteristics of the differential hairpin structures in the three peptides, we studied the contact formation of each peptide (figure 3.5). The CaMKII peptide can be broken down into three regions: N-terminus, C-terminus, and the center. The N-terminal region (293-298) contains positively charged residues in RRK/RAK, and neutral residues in AAA. The central region, or the CaM binding motif (L299-L308), is mainly composed by hydrophobic residues. The C-terminal region of each peptide (309-312) contains a charged arginine residue, which can potentially form hydrogen bonds or repel other positively charged residues in the N-terminus.In the wild-type peptide RRK, as seen in figure 3.5A, the probability of contact formation is generally low (less than 0.5), which suggests high variations in

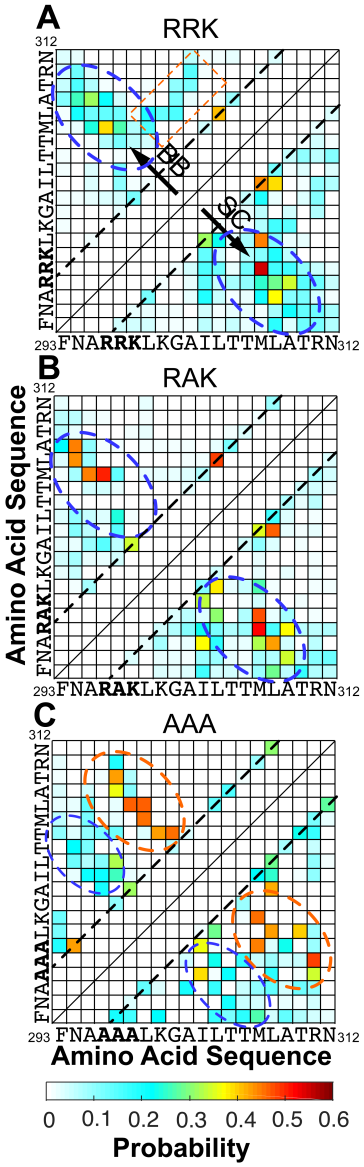

Figure 3.5: Contact probability map of the CD-refined MD structures. Probability of contact formation are plotted for peptide (A) RRK, (B) RAK, and (C) AAA. The upper triangle and lower triangle depict the probability of backbone to backbone (BB) and side-chain to side-chain (SC) contact formation, respectively. The amino acid sequences are provided as the axis labels. The blue/orange ellipses encircle anti-parallel β-sheet structures; the orange rectangle encircles parallel β-sheet structure; and the dotted straight lines mark the contacts in the α-helical structures. The criteria of contact formation are defined in the Method section. Figure generated by Pengzhi Zhang and Nathaniel Jennings.

the conformations adopted by the peptide. Secondary structures such as β-sheets can be formed with low probability. More specifically, the N-terminus and the C-terminus can possibly form anti-parallel β-sheet, suggested by the interactions in the cross-diagonal region of the contact map (blue ellipses), especially between sidechains of M307 and the middle basic residue (R297); the central region of the peptide can form parallel β-sheets, suggested by the low-probability interactions in the region of the contact map that are parallel to the diagonal (orange ellipse); more likely, the central region can form α-helix, indicated by the sparsely distributed higher probability contacts parallel to the diagonal (residues separated by 4 residues, dotted lines parallel to the diagonal in figure 3.5A), such as the backbone to backbone contact between L304 and L308, and the side-chain to side-chain contacts between L299 and I303, between I303 and M307. Upon mutation of R297A, in figure 3.5B, the interactions are sparser but mostly of higher probabilities. Comparing to the wild type, there is a higher probability of forming an anti-parallel β-sheet between the N-terminus and the central region of the peptide (blue ellipses in figure 3.5 5B). The N-terminus are likely to form stable contacts with hydrophobic residues in the central region close to the C-terminus, especially between the residues around the mutation A297 and M307-L308. Comparing to the wild type, interactions in the center of the RAK peptide do not seem to form any parallel β-sheet structure (figure 3.5). In AAA, further compaction in the peptide structure (figure 3.4) and increase in the secondary structures are observed (figure 3.5C). In contrast to

RRK and RAK, there is a relatively high probability of forming anti-parallel β-sheet structures between the N-terminus and the central region (blue ellipses, figure 3.5C), and a low probability of forming anti-parallel β-sheet structures between the central region and the C-terminus (orange ellipses, figure 3.5C). Interestingly, the mutated residues play an essential role. There are stable backbone to backbone interactions between the hydrophobic region formed by the mutated residues and neighboring residues (A297-G301) and hydrophobic residues in the central region (M308-L309), as well as side-chain to side-chain interactions between the mutated residues and residues in the central region as well as the C-terminus. To note, the mutated residue A298 forms a side-chain to side-chain contact with charged residue R311 in high probability, which is prohibited in RRK or RAK because of the electrostatic repulsion. In summary, AAA peptide shows a high probability of adopting anti-parallel β-sheet (as shown in figure 3.4C) and the stabilizing hydrophobic interactions of the AAA mutant may interfere with helix formation, which is a necessary conformational adjustment that aligns the CaM-binding motif to residues in CaM, including residues L299, I303, and L308 (lack of interactions within the CaM-binding motif along the lines parallel to the diagonal in figure 3.5C).

Furthermore, we analyzed the hydrogen bonds within each peptide ensemble to investigate the role of charged residue distribution in each peptide's equilibrium conformation (figure 3.6). Our analysis reveals two diagonal hydrogen bonding patterns in AAA between N- and C- terminal residues that does

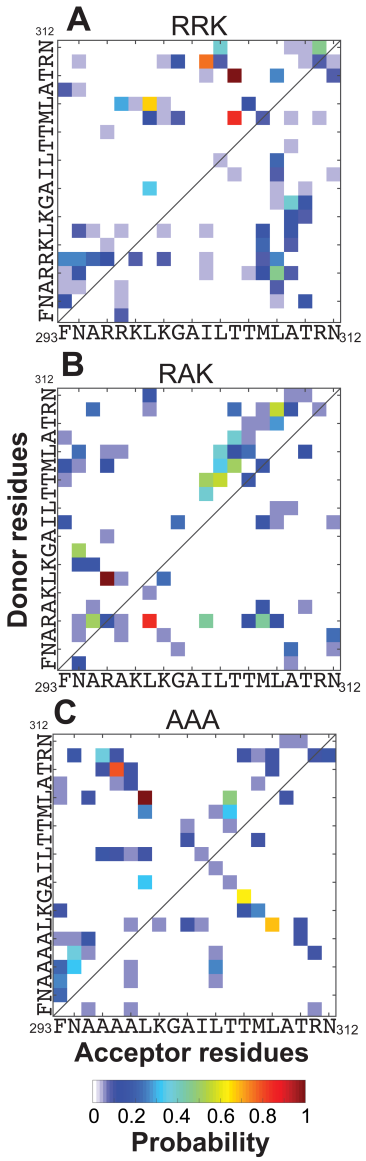

Figure 3.6: Hydrogen bond probability map. The relative probability of intramolecular hydrogen bond formation is shown for the ensemble of structures extracted from all atom MD simulations using the results from NN-LSQ CD deconvolution for (A) RRK, (B) RAK, and (C) AAA. Contacts are defined using a 30° angular cutoff and 4Å distance cutoff between hydrogen bond donor and acceptor residues. Contact probabilities are scaled such that the highest contact probability is 1.

not exist in RRK or RAK. Upon closer examination of AAA, we observe that the charged residue mutation sites form hydrogen bonds with the C-terminal region near R311. This binding pattern appears to contribute to the β-sheet secondary structure formation of AAA. On the other hand, the two highest probability bonds exist between T310-M307 and R311-L308, which may contribute to the formation of the C-terminal helical motif that is observed in several extracted AAA ensemble conformations (figure 3.4C). RRK and RAK show alternate hydrogen bond patterns on the diagonal that loosely resemble helical or turn conformations. RRK and RAK appear to form only one high probability hydrogen bond between M307 and L304, which is not formed by AAA. Although RRK and RAK both appear to form low probability hydrogen bonds with both N and C terminus residues, RAK possesses a high probability hydrogen bond between R296 and M307. Examination of the bonds formed by the charged residues of the N-termini in RRK and RAK illustrates a pattern of interactions with over half of the other residues, with RRK possessing a greater spread of low probability bonds than RAK. Direct binding pattern changes between RRK, RAK, and AAA at the mutation sites are expected, however many of the new hydrogen bonds do not appear to directly involve the charge residues at the mutation sites, implying the effect of charged residue mutations is not localized. This phenomenon is observed in the L299-L308 hydrogen bond: AAA has a high probability of forming this contact compared to RRK and RAK even though neither residue was mutated.

3.4 Discussion

Conformational ensemble of the CaMKII peptides are dependent on charged residue distribution. Our experimental CD measurements and CD deconvolution results indicate that the residual secondary structure of the 3-residue mutant, AAA, is hairpin-like. Additionally, our analysis revealed that the RRK and RAK peptides were composed of disordered and hairpin conformations, along with 4% residual helix structure (Table 3.3). The equilibrium conformational ensemble shift between the wildtype and mutant CaMKII peptides is directly correlated to solvation and electrostatic effects. Previously, the Pappu group has shown that the specific distribution of charged residues within a peptide will affect the equilibrium conformation [24, 101, 120]. To determine whether the conformational shift observed between RRK, RAK and AAA is attributed to changes in charge distribution, we analyzed the sequences of the CaMKII peptides using the IDP analysis tool CIDER [66]. Our analysis found that AAA is predicted to be in a compact/globular ensemble, while RRK is predicted to be in the most expanded conformation. This result was expected since RRK has the most heterogeneously distributed charges with respect to RAK or AAA. CIDER also predicted that RAK will be in a globular form based on the fraction of charged residues, however the similarity between the RRK and RAK CD spectra leads us to question the validity of this prediction.

3.4.1 CDPro overly emphasizes helical formation

Convergence for secondary structure fractions between each CDPro method is necessary for establishing confidence in the determination of secondary structures [148]. Our decision to employ an alternative CD deconvolution method came after a lack of convergence from the three major methods as outlined in Table 3.3. These algorithm features may be appropriate for globular proteins with stable secondary structures; however, we show that they are not applicable for the present set of CaMKII peptides. Due to this revealed incompatibility, we fitted the experimental CD spectra with the SDP-48 reference protein set. Although this reference set was also used with CDPro deconvolution algorithms, it did not produce the same results. Compared to the standard CD deconvolution methods, the fit including the SDP-48 database resulted in the lowest RMSD for all peptides (Table 3.2), signifying that, as expected, IDPs and other flexible proteins require alternate CD data analysis methods than those used for larger, globular proteins. A significant difference in our NN-LSQ fitting is that it considers all proteins within the SDP-48 dataset, whereas the CONTIN/LL, CDSSTR, and SELCON3 methods randomly select a subset of proteins from the reference set for solving the deconvolution problem. This may be an issue for peptides with disordered content since there are only five fully disordered proteins within the SDP-48 dataset. Hence, the best combination of proteins to fully represent CD spectra has, to date, not been obtained.

3.4.2 Force fields for molecular dynamics simulations favor helical formation

The Hamiltonian used in MD Force fields refine coefficients through experiments with larger globular proteins that are structured by nature [69, 92]. This effect has been demonstrated in our equilibrium peptide simulations, which were performed for all mutant variations and at different temperatures (see figure 3.3). The effect of temperature on the secondary structures of each peptide appears to be minimal. In each simulation, the helical conformation is over expressed regardless of temperature or even mutation. In all three peptide runs at each temperature, the same structural trend appears: helix, turn and unordered structure components are similarly distributed. Compared to the experimental CD results (figure 3.1), we expect the emergence of a dominant structure in AAA that does not appear in RRK or RAK. Since the trajectory data does not display this trend, the force field we used is assumed to contain conformational bias despite previous efforts to improve accuracy [98].

Newer force fields for molecular dynamics simulations that are designed for IDP and folded proteins are available [132, 70]. However, choosing the best model for our specific system was not a simple task. In addition, variations in the water model heavily affect the outcome of IDP simulations [8, 11]. There is a limitation to IDP force field development due to the lack of experimental data detailing the conformational ensemble of IDPs. Common

methods for experimentally refining force fields, such as SAXS, FRET and NMR, are only able to produce an average of the conformational ensemble [77, 15] and do not necessarily contain the observables needed to describe IDPs in silico. We elected to sample a larger set of data by implementing an implicit solvent model instead of focusing our efforts on finding the best MD parameterization. By combining simulation and experimental results, we are able to reveal and partially resolve the shortcomings in each method [48].

3.4.3 Conformations of unbound CaMKII peptide may be important to binding with CaM

The experimental study of the CaM-CaMKII binding kinetics between CaM and the CaMKII peptides illustrate an 6-fold increase in the association rate of RRK compared to AAA [164] in 150 mM ionic solution. This ionic strength effectively screens the electrostatic potential by a Debye length of 7.8Å. This screening effect can decrease the electrostatic rate enhancement for diffusion-limited binding kinetics [160, 1] of CaM and the CaMKII peptides, however the electrostatic potential is not completely screened over localized peptide regions. Comparison of the kinetic results to the conformational analysis in the present study resolves a finite set of possible binding mechanisms between CaM and the CaMKII peptides. We initially assumed that AAA would have a higher affinity for CaM because of the residual helical propensity induced by the alanine residues as these peptides are known to

adopt a helical conformation when they bind to CaM. It has been hypothesized that the presence of residual structure would reduce the energy barrier between unbound and bound states leading to increased association rates [130, 108]. Since the stopped-flow experimental results (decreased on-rate for AAA relative to RRK; [101]) disproved this hypothesis, we turned to our CD analysis for additional mechanisms, which has shown to offer a diverse range of secondary structures for other calmodulin binding target peptides [35].

The CD measurements indicate a distinct difference in the ensemble of RRK and AAA secondary structures. Our CD deconvolution results indicate that the secondary structure formed through each mutation is actually in the form of a hairpin structure. The apparent lack of helical structure in the peptide ensemble implies that the hypothesis that increased kinetics and peptide residual structure are positively correlated [23, 3, 74] is not applicable in modeling the CaM-CaMKII peptide binding. Moreover, a larger energy gap between the bound and unbound states may exist due to the presence of the stable hairpin structure in AAA [165, 64, 65]. In order for the mutual and induced conformational fit mechanism [162] to take place, the peptide must transition from the hairpin structure to the extended state in order to form productive and stable contacts with CaM. Our findings suggest that a significant conformational change must occur for the AAA peptide, reversing the hairpin structure to allow formation of the helical conformation upon formation of the CaM bound complex [103]. This provides a plausible mechanistic

explanation for the differences in association rates [101] and emphasizes that conformational frustration can be an important step in regulating the kinetics of protein-protein interactions.

Chapter 4

Investigation of CaMKII(293-312)/CaM binding kinetics

4.1 Introduction

Modeling the complex Ca^{2+} -CaM/CaMKII(293-312) kinetics has been an underlying goal of this dissertation. A major challenge in modeling this bimolecular interaction stems from the disordered nature of the CaMKII peptides and the flexible CaM linker. Equilibrium binding between CaM and the CaMKII peptides occurs in multiple steps:

$$[\mathrm{A}] + [\mathrm{B}] \underset{\mathrm{k}_{-1}}{\overset{\mathrm{k}_1}{\rightleftarrows}} [\mathrm{AB}^*] \underset{\mathrm{k}_{-2}}{\overset{\mathrm{k}_2}{\rightleftarrows}} [\mathrm{AB}] \tag{4.1}$$

where proteins A and B form the encounter complex, [AB*], and undergo conformational changes to produce the final bound canonical complex [AB]. The formation of the encounter complex is diffusion limited; the forward and backward kinetic rates, k_1 and k_{-1} respectively, can be found using the Smoluchowski equation. Unfortunately, the determination of k_2 and k_{-2} is not as straight forward from a theoretical perspective. CaM rapidly samples open and closed equilibrium conformations in the presence of Ca^{2+} ions, which enables it to form the canonical complex with its peptide targets [123]. In this system, the CaMKII peptide targets are IDPs. This increases the conformational degrees of freedom available to the transition pathway between canonically bound and unbound states. The kinetics involving disordered proteins appear to be unique to the system being studied. For example, Clarke et al. show that the coupled folding and binding of the cMyb petide with the KIX protein is much faster than the diffusion limited rates [139], and thus the upper-limit to kinetics can be approximated through the Smoluchowski equation with additional electrostatic rate enhancement considerations. Conversely, the conformational changes associated with coupled folding and binding for IDPs have also been shown to possess relatively slow kinetics compared to the diffusion limited region [31].

If the CaM/CaMKII interaction model possessed a fast binding and folding pathway, then diffusion would be the rate-limiting step and the lower charge on the mutant CaMKII peptide would explain the 6-fold decrease in k_{on} rate [164]. The stopped-flow fluorescence experiment was setup using a solution

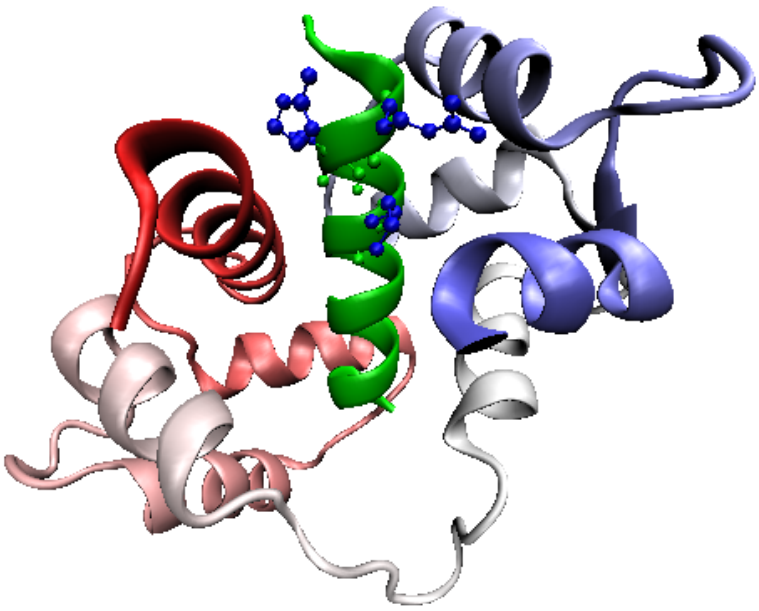

Figure 4.1: visualization of the crystal structure of the CaM/CaMKII(290-314) peptide complex. PDB ID: 1CM1. The charged residue mutation sites of RRK (CaMKII 296-298) are represented explicitly in blue. CaM is colored by atomic index, with red representing the N-terminus and blue representing the C-terminus.

with high ionic concentration in order to shield electrostatic rate enhancement effects. As a result, the observed kinetics are hypothesized to be due to conformational change of the binding partners. In chapter 3, our analysis of the experimental CD spectra for the wildtype and mutant CaMKII peptides illustrated that shift in the ensemble-averaged conformation had taken place. This result agreed with the hypothesis that the CaM/CaMKII kinetic rates are related to a conformational change of the binding partners. The

crystal structure of the CaM/CaMKII(293-312) canonical complex in figure 4.1 illustrates that RRK adopts an α-helical conformation once bound. The increase in secondary structure propensity of the AAA peptide (3-residue mutant) decreased stability of the bound complex despite having neutral or positive kinetic effects in other studies [134, 99].

The goal of the proposed future work in this chapter is to generate a model that is able to predict the kinetics of the CaM/CaMKII peptides and illustrate differences in binding mechanism due to the charged residue mutations. Bimolecular interactions involving a high degree of coupled conformational change require careful treatment of each possible binding pathway, since the culmination of pathways determine the transition between initial and final states. This chapter outlines previous work on bimolecular interaction models, and discusses the implementation of a MSM for the CaM/CaMKII system.

4.2 Overview of kinetics

Collins and Kimball [22] discuss diffusion-controlled reaction rates with some consideration of the partially absorbing boundary condition. The Solution to the Smoluchowski equation for the concentration of a diffusive species is

given in terms of the flux across the boundary:

$$\Phi = 4\pi R^2 D \left(\frac{\partial c}{\partial r}\right)_{r=R} \tag{4.2a}$$

$$\Phi = 4\pi R D c_0 \left(1 + \frac{R}{\sqrt{\pi D t}}\right) \tag{4.2b}$$

Collins and Kimball resolve a diffusion-limited reaction model that may not react after reaching the reactive boundary b. The generalized Smoluchowski equation is given as

$$4\pi R^2 D c'(R) = \kappa c(R) \tag{4.3}$$

which essentially states that the flux is proportional to the reactant concentration. The diffusion-controlled reaction is further controlled by the constant κ, where $\kappa = 0$ implies no reaction occurs, and $\kappa \to \infty$ implies that the reaction is fully diffusion limited.

Alternatively, in Northrup et. al. [116], particles achieving separation of b are modeled as either reacting or diffusing outward towards infinity by the product $k = k_D(b)p$, where the diffusion-controlled reaction can be approximated by the Smoluchowski result $k_D(b) = 4\pi D b$. Calculation of p is dependent on a number of factors within the reaction boundary, b. The quantity β_∞ is defined as the probability that a particle within the radius b will collide with the reactive surface of the receptor at least once rather than escape to infinite separation distance. The quantity α is defined as the intrinsic probability that a collision with the active site will react. If every collision with

the surface of the receptor causes a reaction, $p = \beta_\infty$ and $\alpha = 1$. The kinetic rate is trivial in this case:

$$k = k_D(b)\beta_\infty \tag{4.4}$$

where β_∞ only involves diffusion and can be calculated using Brownian Dynamics. In the case that not every collision results in a reaction, further expansion of α is required. In this case, subsequent diffusion away from an unsuccessful reaction attempt must be considered by including the quantity Δ_∞, which is the probability that another collision will occur after diffusing away following an initial failed reaction. The final reaction rate is given by

$$k = \frac{k_{D(b)}\beta_\infty\alpha}{1-(1-\alpha)\Delta_\infty} \tag{4.5}$$

The major issue with this calculation is that the collision with the reactive surface can have only two outcomes: react or diffuse away. In the case of CaM-CaMKII interaction, an encounter complex is generally formed, but the conformational changes that are required in order to form the bound complex cannot be described through simple probability factors. The encounter complex effectively sequesters CaMKII/CaM from a diffusion limited reaction model, therefore the α, β_∞ and Δ_∞ coefficients cannot describe this time dependent state, where CaM/CaMKII are neither reacted or free. Noe discusses the formation of a kinetic rate k_{AB} from a Markov state model

using the total probability flux [115]:

$$F = \sum_{i \in A} \sum_{i \notin A} \pi_i T_{ij} q_j^+ \tag{4.6a}$$

$$k_{AB} = \frac{F}{\tau \sum_{i=1}^{m} \pi_i q_i^-} \tag{4.6b}$$

The numerator in the equation for k_{AB} represents the total number of states that transition from an equilibrium state (π_i), which exist in the initial configuration (set A) to an intermediate set of states not in A, which ultimately form B instead of A. The denominator accounts for the timescale required for the transitions from A to B, and also removes over counting from the indirect transitions from A to B (bounces between multiple intermediate states). This equations shares a great similarity to the Northrup calculation [116], however this kinetic rate does not limit its usefulness to diffusive models.

Noe later expands this rate calculation to include diffusion [63]. The combined diffusion and bimolecular reaction rate was derived by Erban and Chapman in their 2009 paper [42]. The equation for k_{on} is given as

$$k_{on} = 4\pi D \left(r - \sqrt{\frac{D}{k_{AB}}} \tanh \left(r \frac{k_{AB}}{D} \right) \right) \tag{4.7}$$

In his paper, Noe mentions that this equation results in the same rate calculation as those produced in Northrup's paper [116]. Northrup's model does not appear to be applicable to protein reactions that involve time-dependent conformational change, therefore we must investigate further.

Taking a closer look at the work by Erban and Chapman [42], the derivation of the above equation makes use of the assumption that a reactive species within a radius b of the receptor will be removed by rate k_{AB}. At a distance less than the reaction radius, b, the concentration is assumed to take the form of the following differential equation:

$$\frac{d^2c}{dr^2} + \frac{2}{r}\frac{dc}{dr} - \frac{k_{AB}c}{D_A + D_B} = 0 \tag{4.8}$$

The solutions to this equation are in the form of radially dependent exponentials, which implies that the system concentration is heavily impacted by the rate of removal of species A by k_{AB}. Therefore there appears to be a bottleneck implied within this concentration equation: diffusion is the rate limiting step, and the bimolecular rate is comparatively fast.

HX Zhou discusses reaction rates that are diffusion limited versus rates that are limited by bimolecular interaction[174]. In a reaction scheme where an encounter complex forms: $A + B \underset{k_{D-}}{\overset{k_D}{\rightleftharpoons}} A \cdot B \underset{k_{c-}}{\overset{k_c}{\rightleftharpoons}} C$, the rate is given as

$$k_a = \frac{k_D k_c}{k_{D-} + k_c} \tag{4.9}$$

If $k_c >> K_{D-}$, then the diffusion step is rate limiting, and $k_a \approx k_D$. conversely, if there is a large conformational change between the bound and unbound states, this conformational step can be rate limiting. In this case, $k_a \approx \frac{k_D}{k_{D-}}\frac{k_c k_D}{k_{D-}}$. Hagen et al [59] model a peptide and enzyme interaction

where both diffusion and protein folding come into play. Their model is based on a two-step description; first the ligand and receptor diffuse together until they share a reaction volume, forming an encounter complex. At this point, it either dissociates at a rate k_{D-} or reacts at rate k_+ to form a complex. In the steady state approximation, the overall binding rate is given by k_{on}, with

$$\frac{1}{k_{on}} = \frac{1}{k_{D+}} + \frac{1}{Kk_+} \tag{4.10}$$

Where $K \equiv \frac{k_{D+}}{k_{D-}}$; the diffusion based kinetics. In the case of a reaction-limited model, $k_{on} \approx Kk_+$.

Gordon et al. [60] also discuss the case of a mutual and induced conformational binding mechanism as a rate limiting step. They suggest the use of pathway calculations similar to the concepts of Markov State Models to determine the rate constants.

4.3 Markov state model

4.3.1 Simple construction of the MSM

We begin our generation of the MSM model from a set of clusters, S, where the kth cluster is defined as a member of S

$$S = \{S_1, S_2, \dots S_k, \dots\} \tag{4.11}$$

Given a simulation of infinite time, each state $\mathbf{x} \in S$ will be visited an infinite number of times [127].

Elements in the transition matrix, $T_{ij}(\tau)$ represent the probability of a system to be in state j at time $t+\tau$ given that it was in state i at time t. If we define the transition count c_{ij} at the number of transitions from state i at time t to state j at time $t+\tau$, then we can generate an initial estimate of the transition matrix:

$$T_{ij}(\tau) = \frac{c_{ij}}{\sum_k c_{ik}} = \frac{c_{ij}}{c_i} \tag{4.12}$$

where $\sum_k c_{ik}$ represents the total number of transitions from state i. This trivial estimation assumes that the states are well distributed.[127]

4.3.2 Stationary states

In a perfectly sampled simulation over infinite time, the stationary density, $\mu(\mathbf{x})$ represents the density of time spent in state $\mathbf{x}$. $\mu(\mathbf{x})$ is invariant and related to the partition function by

$$\mu(\mathbf{x}) = \frac{e^{-\beta H(\mathbf{x})}}{Z(\beta)} \tag{4.13}$$

with the Hamiltonian given by $H(\mathbf{x})$ and β given by $\frac{1}{kT}$.

We define the stationary probability π_k as the probability to be in set S_k[127]

$$\pi_k = \int_{i \in S_k} \mu di \tag{4.14}$$

Where μ is the global stationary density. MSM models require that the transition matrix satisfy the detailed balance:

$$\pi_i T_{ij} = \pi_j T_{ji} \tag{4.15}$$

Stationary states define the equilibrium state density given an infinite relaxation time. A system approaches Markovian dynamics if the forwards and backwards transitions between two states is equal [115]. In building our MSM transition matrix, we can enforce the detailed balance by applying the maximum likelihood (Bayesian) estimator between forward and backward transitions.[143]

Since we are using a single memory for our AWSEM simulations and folding under annealing conditions, the trajectories are not in equilibrium, therefore the global stationary states cannot be calculated directly. Before approximating stationary states, we focus on the time dependent probability densities ($p_t(\mathbf{x})$) and the associated transition matrix.

4.3.3 Transition matrix

Noe[114] outlines the reversible element shift method, which describes how to modify the transition matrix such that it obeys the detailed balance while preserving the stationary states. We can also generate the symmetric count matrix by

$$T' = \frac{T^\dagger + T}{2} \tag{4.16}$$

4.3.4 Evolution of states

Clustering allows us to partition the conformation space Ω into k clusters. Given an initial probability distribution of states at time $t = 0$, the transition matrix generates the probability distribution of states after the lagtime τ:

$$\mathbf{p}(\tau) = T(\tau)\,\mathbf{p}(0) \tag{4.17}$$

It follows that the state distribution for any time $t = n\tau$ can be calculated if

$$T(n\tau) = [T(\tau)]^n \text{ given that } \tau \geq \tau_{eq} \tag{4.18}$$

where τ_{eq} is the Markov timescale of the data given. This quantity is the coarse-grained limit of sampling time, where lagtimes shorter than the sampling times provide no further information as it does not exist within the sampled simulation data.[114]

4.3.5 Piecing together data to approximate MSM lagtime

We do not know the ideal lag time (τ) needed to generate a transition matrix. A caveat of the lag time is that it must be shorter than the timescale of the process we are interested in measuring (canonical binding of the CaM-CaMKII complex)[114]. In the CaM-CaMKII binding model, no canonical

binding has been observed before the 10^6 Nstep mark. Therefore we can begin to iteratively estimate lagtimes by building a set of transition matrices at various lagtimes smaller than the folding event. These count matrices can be adjusted to satisfy the detailed balance, and the lagtime satisfying the condition

$$\mathbf{T}(n\tau) \approx [\mathbf{T}(\tau)]^n \tag{4.19}$$

We can use the associated lagtime to calculate time scales and stationary states, as well as estimate the quality of our model through error analysis.

4.3.6 Timescales and eigenstates

Given a microstate $\mathbf{x} \in S$, we have a probability distribution of the microstate at a time t given by $\rho_t(\mathbf{x})$. Then the transition matrix will describe the probability distribution of the state at time $t + \tau$:

$$\rho_{t+\tau}(\mathbf{x}) = \hat{T}(\tau)\,\rho_t(\mathbf{x}) \tag{4.20}$$

The eigenvalues of the transition matrix can be represented by

$$\lambda_i(\tau) = e^{\frac{\tau}{t_i}} \tag{4.21}$$

where the eigenvalue $\lambda_i(\tau)$ with timescale t_i is associated with eigenvector ϕ_i [129]. Then the probability distribution for a state y after time $t + \tau$ can

be computed with

$$\rho_{t+\tau}(\mathbf{y}) = \sum_{i=1}^{\infty} e^{\frac{\tau}{t_i}} \langle \psi_i | \rho_i \rangle \phi_i \tag{4.22}$$

The equilibrium solution to our system corresponds to a timescale of $t_1 \rightarrow \infty$, thus $\lambda_1 = 1$ and $\phi_1 = \mu$. For $i > 1$, we can calculate the weighted eigenfuctions $\psi_i = \frac{\phi_i}{\mu}$

Subsequent eigenstates and eigenvalues of the transition matrix represent finite timescales, with $t_2 > t_3 > t_4 \ldots$, hence, CaM-CaMKII binding would be observable under the solutions around t_2. With such solutions, kinetics can be calculated for the "slowest" timescale using the autocorrelation of the approximate eigenstates:

$$\langle \psi_i(\mathbf{x}_t) | \psi_i(\mathbf{x}_{t+\tau}) \rangle_t \leq e^{\frac{\tau}{t_i}} \tag{4.23}$$

4.4 Implementing the Markov State Model

The MSM has been useful in illustrating the mechanisms in folding and binding of a wide range of proteins, including those with large degrees of freedom [124, 126, 21]. Completion and accuracy of the MSM heavily depends on the data obtained from the CaM/CaMKII simulations.

4.4.1 CaM/CaMKII representation

An ongoing challenge surrounding the generation of the MSM for the CaM/CaMKII peptide system stems from accurate coarse-grained representation of CaM.

The behavior of CaM in solution is affected by its 4 EF-hand motifs that bind to Ca^{2+} at different rates, therefore its equilibrium conformations are transient [146, 145]. An all atom representation of CaM could be used to show conformational changes due to Ca^{2+} signaling, however the large system size makes an all-atom representation impractical for constructing an MSM. The coarse-grained representation of CaM and the CaMKII peptides is ideal for obtaining long trajectories, however the complete ensemble of structures must be sampled. In chapter 3, the ensemble of structures for the CaMKII peptides was generated. Since the AWSEM force field can be trained using this high resolution structure ensemble, the CaMKII peptide may be accurately represented. Assuming the CaMKII peptides possess a flat energy landscape, each of the ensemble structures have roughly an equal probability of sampled during MD. Unfortunately, the extended and collapsed equilibrium structure of CaM in solution is not equally sampled and is therefore not currently adequate. Coarse-grained representation of CaM in solution is currently an active research topic in the Cheung group.

4.4.2 Bimolecular diffusion

Formation of the encounter complex between CaM and the CaMKII peptides proceeds through diffusion. Since CaM and the CaMKII peptides both contain charged residues, consideration of electrostatic rate enhancement must be made. Originally, the AWSEM potential (see equation **??**) does not contain an electrostatic term. Our goal is to generate kinetics that follow the

trends observed in Waxham's original study [164], therefore we must consider electrostatic shielding. Experimentally, 150mM of $CaCl_2$ was used, which corresponds to a Debye length of 7.8Å. Thus, we can implicitly represent these ionic conditions in the AWSEM force field by including the Debye-Huckel equation (see equation 1.3).

I intend on placing the CaM/CaMKII peptides a distance of roughly 50Å apart, which enables both a diffusion-limited region with and without electrostatic enhancement to be simulated. Currently, it is not clear how the relative orientations of CaM and the CaMKII peptides would affect the encounter complex.

4.4.3 Binding and folding

The encounter complex is formed upon first contact between CaM and the CaMKII peptides. At this point, there are multiple pathways and conformational states that each binding partner can progress through in order to form the canonical bound complex. The equilibrium binding and unbinding behavior can be represented using the transition matrix. If we define macrostate **A** as the encounter complex and macrostate **B** as the canonically bound complex, then the eigenvalues and eigenvectors of the transition matrix can be solved for all forwards and backwards transitions between the **A** and **B** states.

4.4.4 Challenges

The transition matrix is ultimately dependent on the quality of simulation data used. As discussed earlier, an accurate coarse-grained equilibrium model of CaM in solution has not been found yet. Pilot simulations of CaM/RRK or CaM/AAA binding showed that once the canonical complex is formed, it does not unfold and return to the encounter complex state. A true Markov process requires both **A** and **B** macrostates to be sampled at equilibrium. Thus, the model is not physical if the MD simulation results do not contain transitions from **B** to **A**. Further development of the coarse-grained model must be taken.

4.4.5 Expected outcomes

Once issues regarding the coarse-grained representation of CaM are resolved, the MSM can provide excellent details on CaM/CaMKII peptide binding mechanisms. The bimolecular reaction between CaM and the CaMKII peptides involves both diffusion and folding.

4.4.5.1 Diffusion region

The effects of electrostatic rate enhancement on the diffusion limited region is expected to only affect the formation of the encounter complex at short distances due to ionic shielding. Using a large number of trajectories with unique initial conditions, the diffusive kinetic rate may be solved. A major unan-

swered question pertains to whether the formation of the CaM/CaMKII(293-312) canonical complex is limited by diffusion or conformational change. If the bimolecular reaction is indeed limited by diffusion, then the folding, binding and unbinding of CaM and the CaMKII peptides should take place before another peptide molecule has the opportunity to form the encounter complex with CaM. Additionally, the kinetic binding rates between RRK and AAA should be similar since the peptides are roughly the same size. From the experimental results of the Waxham group, we are aware that this scenario is not true and the kinetics are not governed by diffusion alone.

4.4.5.2 Binding and folding

If the kinetics of the CaM/CaMKII peptide interactions are dependent on conformational change, then the electrostatic rate enhancements in the diffusion region are inconsequential. The process of folding and binding is expected to occur over a larger timescale than the diffusion process, implying that the kinetic binding rates between CaM and the CaMKII peptides is non-linear.

CaM has a flexible linker, which enables it to wrap around its binding target to produce the canonical complex. The flexibility of the linker region implies that there are many folding pathways available for CaM. Likewise, the disordered nature of the CaMKII peptides implies the formation of contacts with CaM can occur through various different pathways. Since a MSM considers the culmination of all possible pathways between the encounter complex **A**

and final canonical complex **B**, the charged residue mutation of RRK is expected to remove or change some of these pathways.

Electrostatic effects are expected to play a significant role in binding once the encounter complex is formed because ionic shielding will be small at this length scale. RRK contains positively charged residues at the N-terminus and C-terminus, whereas AAA only contains a positively charged residue at the C-terminus. I hypothesize that the positively charged residues on RRK will interact with the positive and negative residues on the N-terminus and C-terminus of CaM. Once RRK is able to form simultaneous contacts with both CaM terminals, hydrophobic effects will increase the number of contacts between RRK and the CaM linker. This process will collapse both to form the bound canonical complex. Alternatively, AAA was shown in chapter 3 to possess a higher β-sheet secondary structure. Thus, more energy will be required to break this structure and form contacts with CaM instead. Additionally, the C-terminus of AAA may interact with negatively charged residues in either end of CaM, but the N-terminus of AAA may require more time to interact with CaM. This will increase the amount of time required to form enough native contacts with CaM to canonically bind.

Chapter 5

Conclusions

CATS brings together several aspects of different types of clustering algorithms in a unique way, exploiting distribution trends in the raw data to form clusters. The question we want to address is whether CATS out-performs other clustering algorithms for the purpose of conformational analysis of IDPs. The answer is not simply yes or no, as there are several factors that one must consider when comparing the results from CATS to those of an RMSD-based method, such as GROMOS. The difference between CATS and GROMOS lies in how one defines similarity between structures. RMSD similarity has the disadvantage of being based in Cartesian coordinate space; there is sensitivity to how the structures are aligned beforehand, and ignores energetic deviations within the backbone. CATS defines similar structures as those belonging to the same regions within the PMF space, thus backbone energetics has a significant effect on structure categorization. In recent

years, attention has been given to the development of machine learning (ML) algorithms for use in clustering proteins. CATS shares a large similarity with current ML algorithms [5]. With little modification, the CATS algorithm can correct for noise and irregularities discussed earlier that may hamper analysis and be fully capable of self-correction.

The importance of IDPs in biological function has become readily apparent in recent years. A major challenge in IDP modeling stems from experimental sampling of the structure ensemble. Popular methods such as NMR spectroscopy offer higher resolution, but are still limited in IDP ensemble determination. To overcome difficulties pertaining to experimental ensemble construction of IDPs, combined theoretical approaches are often used. Circular dichroism spectroscopy does not offer high resolution structure determination, however this drawback appears to be inconsequential for IDPs since MD simulation can be used to perturb the averaged structure to generate the IDP ensemble. In this study, we have used a combination of techniques to bridge the experimental data with theoretical data to generate a detailed picture of our CaMKII peptides despite the inherent inaccuracy of the MD simulation. Our resulting ensemble approximations illustrate how the residual secondary structure of the CaMKII peptides changes due to charged residue mutation. Our findings suggest that the AAA ensemble becomes stabilized through the formation of the hairpin secondary structure, which may explain the binding phenomenon observed in previous studies [164]. In addition to the free peptide ensemble, the observed structure shift may play a significant role

in complex stability post binding due to the formation (or lack thereof) of "fuzzy structures"[155].

The culmination of this work can be used in the generation of a Markov State model(MSM). Due to the complexity associated with protein-protein interactions involving IDPs, a MSM is the ideal method of characterizing binding pathways. Our future work aims to use the MSM in conduction with coarse-grained simulations to describe differences in the binding pathways between CaM and the wildtype/mutant CaMKII peptides. Since MSMs have been successful in describing complex binding pathways between flexible proteins[126, 21], it may be the ideal solution for explaining the phenomenon of CaM trapping. If we are able to successfully model the binding kinetics between CaM and the CaMKII peptides using coarse-grained MD, the next step is to expand the system to include multiple monomers of the full length CaMKII enzyme. Autonomous activation of the CaMKII holoenzyme occurs through a complex interaction of multiple Ca^{2+} CaM molecules and at least 2 CaMKII monomers. The our completed model will be able to illustrate the conformational changes that must take place in order to induce the autonomous CaMKII state. Contingent on the completion of coarse-grained representations of Ca^{2+} CaM and apoCaM, differences in CaMKII activation by various Ca^{2+} ion signals can be investigated. On a larger scale, the successful representation of IDPs and coarse-grained modeling is applicable to other protein interaction problems involving disordered regions. As the importance of IDPs in nature emerge, the complicated protein-protein inter-

actions involving them continue to be an active area of research.

Bibliography

[1] R. Alsallaq and H. X. Zhou. Electrostatic rate enhancement and transient complex of protein-protein association. Proteins, 71(1):320–35, 2008.

[2] M. Amaral, D. B. Kokh, J. Bomke, A. Wegener, H. P. Buchstaller, H. M. Eggenweiler, P. Matias, C. Sirrenberg, R. C. Wade, and M. Frech. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. Nat Commun, 8(1):2276, 2017.

[3] M. Arai, K. Sugase, H. Dyson, and P. Wright. Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. Proceedings of the National Academy of Sciences of the United States of America, 112(31):9614, 2015.

[4] K. A. Ball, A. H. Phillips, P. S. Nerenberg, N. L. Fawzi, D. E. Wemmer, and T. Head-Gordon. Homogeneous and heterogeneous tertiary structure ensembles of amyloid-β peptides. Biochemistry, 50(35):7612–7628, 2011.

[5] J. Behler. Perspective: Machine learning potentials for atomistic simulations. The Journal of Chemical Physics, 145(17):170901, 2016.

[6] H. J. C. Berendsen, D. Vanderspoel, and R. Vandrunen. Gromacs - a message-passing parallel molecular-dynamics implementation. Computer Physics Communications, 91(1-3):43–56, 1995.

[7] R. B. Best, N. V. Buchete, and G. Hummer. Are current molecular dynamics force fields too helical? Biophys J, 95(1):L07–9, 2008.

[8] R. B. Best and J. Mittal. Protein simulations with an optimized water model: Cooperative helix formation and temperature-induced unfolded state collapse. The Journal of Physical Chemistry B, 114(46):14916–14923, 2010.

[9] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. Journal of Chemical Theory and Computation, 8(9):3257–3273, 2012.

[10] W. Bode, P. Schwager, and R. Huber. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding: The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with ile-val at 1.9 Å resolution. Journal of Molecular Biology, 118(1):99–112, 1978.

[11] S. Boonstra, P. R. Onck, and E. van der Giessen. Charmm tip3p water model suppresses peptide folding by solvating the unfolded state. The Journal of Physical Chemistry B, 120(15):3692–3698, 2016.

[12] A. Borgia, M. B. Borgia, K. Bugge, V. M. Kissling, P. O. Heidarsson, C. B. Fernandes, A. Sottini, A. Soranno, K. J. Buholzer, D. Nettels, B. B. Kragelund, R. B. Best, and B. Schuler. Extreme disorder in an ultrahigh-affinity protein complex. Nature, 555(7694):61–66, 2018.

[13] D. H. Brookes and T. Head-Gordon. Experimental inferential structure determination of ensembles for intrinsically disordered proteins. Journal of the American Chemical Society, 138(13):4530–4538, 2016.

[14] D. H. Brookes and T. Head-Gordon. Experimental inferential structure determination of ensembles for intrinsically disordered proteins. Journal of the American Chemical Society, 138(13):4530–4538, 2016.

[15] M. Brucale, B. Schuler, and B. Samorì. Single-molecule studies of intrinsically disordered proteins. Chemical Reviews, 114(6):3281–3317, 2014.

[16] B. Brutscher, I. C. Felli, S. Gil-Caballero, T. Hošek, R. Kümmerle, A. Piai, R. Pierattelli, and Z. Sólyom. NMR Methods for the Study of Instrinsically Disordered Proteins Structure, Dynamics, and Interactions: General Overview and Practical Guidelines, pages 49–122. Springer International Publishing, Cham, 2015.

[17] G. A. Carpenter, S. Grossberg, and D. B. Rosen. Fuzzy art - fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks, 4(6):759–771, 1991.

[18] D. Case, I. Ben-Shalom, S. Brozell, D. Cerutti, T. Cheatham, III, V. Cruzeiro, T. Darden, R. Duke, D. Ghoreishi, M. Gilson, H. Gohlke, A. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. Walker, J. Wang, H. Wei, R. Wolf, X. Wu, L. Xiao, D. York, and P. Kollman. Amber 14, 2014.

[19] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo. Protein structure determination from nmr chemical shifts. Proceedings of the National Academy of Sciences, 104(23):9615–9620, 2007.

[20] W.-Y. Choy and J. D. Forman-Kay. Calculation of ensembles of structures representing the unfolded state of an sh3 domain. Journal of Molecular Biology, 308(5):1011–1032, 2001.

[21] A. P. Collins and P. C. Anderson. Complete coupled binding-folding pathway of the intrinsically disordered transcription factor protein brinker revealed by molecular dynamics simulations and markov state modeling. Biochemistry, 57(30):4404–4420, 2018.

[22] F. C. Collins and G. E. Kimball. Diffusion-controlled reaction rates. Journal of Colloid Science, 4(4):425 – 437, 1949.

[23] P. Csermely, R. Palotai, and R. Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci, 35(10):539–46, 2010.

[24] R. K. Das and R. V. Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proceedings of the National Academy of Sciences, 110(33):13392–13397, 2013.

[25] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. Peptide folding: When simulation meets experiment. Angewandte Chemie-International Edition, 38(1-2):236–240, 1999.

[26] R. N. Dave and K. Bhaswan. Adaptive fuzzy-c-shells clustering and detection of ellipses. Ieee Transactions on Neural Networks, 3(5):643–662, 1992.

[27] P. De Koninck and H. Schulman. Sensitivity of cam kinase ii to the frequency of ca2+ oscillations. Science, 279(5348):227–30, 1998.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38, 1977.

[29] A. Desai and T. J. Mitchison. Microtubule polymerization dynamics. Annu Rev Cell Dev Biol, 13(1):83–117, 1997.

[30] K. A. Dill and J. L. Maccallum. The protein-folding problem, 50 years on. Science, 338:1042,1046, 2012.

[31] J. Dogan, T. Schmidt, X. Mu, Å. Engström, and P. Jemth. Fast association and slow transitions in the interaction between two intrinsically disordered protein domains. The Journal of biological chemistry, 287(41):34316–34324, 2012.

[32] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon. Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics, 21(16):3433–4, 2005.

[33] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol, 347(4):827–39, 2005.

[34] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered protein. Journal of Molecular Graphics and Modelling, 19(1):26–59, 2001.

[35] T. Dunlap, J. Kirk, P. E.A., M. Yoder, and T. Creamer. Thermodynamics of binding by calmodulin correlates with target peptide alpha-helical propensity. Proteins: Structure, Function, Bioinformatics, 81:607–612, 2013.

[36] H. J. Dyson and P. E. Wright. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. Adv Protein Chem, 62:311–40, 2002.

[37] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. Nature Reviews Molecular Cell Biology, 6(3):197–208, 2005.

[38] S. Eguchi and J. Copas. Interpreting kullback–leibler divergence with the neyman–pearson lemma. Journal of Multivariate Analysis, 97(9):2034–2040, 2006.

[39] Y. El-Sonbaty and M. A. Ismail. Fuzzy clustering for symbolic data. Ieee Transactions on Fuzzy Systems, 6(2):195–204, 1998.

[40] D. Eliezer. Biophysical characterization of intrinsically disordered proteins. Current opinion in structural biology, 19(1):23–30, 2009.

[41] T. Eltoft and R. J. P. deFigueiredo. A new neural network for cluster-detection-and-labeling. Ieee Transactions on Neural Networks, 9(5):1021–1035, 1998.

[42] R. Erban and S. J. Chapman. Stochastic modelling of reaction–diffusion processes: algorithms for bimolecular reactions. Physical Biology, 6(4):046001, aug 2009.

[43] N. A. Eschmann, T. D. Do, N. E. LaPointe, J.-E. Shea, S. C. Feinstein, M. T. Bowers, and S. Han. Tau aggregation propensity engrained in its solution state. The Journal of Physical Chemistry B, 119(45):14421–14432, 2015.

[44] S. Eschrich, J. W. Ke, L. O. Hall, and D. B. Goldgof. Fast accurate fuzzy clustering through data reduction. Ieee Transactions on Fuzzy Systems, 11(2):262–270, 2003.

[45] J. C. Ezerski and M. S. Cheung. Cats: A tool for clustering the ensemble of intrinsically disordered peptides on a flat energy landscape. J Phys Chem B, 122(49):11807–11816, 2018.

[46] U. Fano. Ionization yield of radiations. ii. the fluctuations of the number of ions. Physical Review, 72(1):26–29, 1947.

[47] E. Fischer. Einfluss der configuration auf die wirkung der enzyme. Ber Dtsch Chem Ges., 27:2984–2993, 1894.

[48] C. K. Fisher, A. Huang, and C. M. Stultz. Modeling intrinsically disordered proteins with bayesian statistics. Journal of the American Chemical Society, 132(42):14919–14927, 2010.

[49] C. K. Fisher and C. M. Stultz. Constructing ensembles for intrinsically disordered proteins. Curr Opin Struct Biol, 21(3):426–31, 2011.

[50] A. Forest, M. T. Swulius, J. K. Tse, J. M. Bradshaw, T. Gaertner, and M. N. Waxham. Role of the n- and c-lobes of calmodulin in the activation of ca(2+)/calmodulin-dependent protein kinase ii. Biochemistry, 47(40):10587–99, 2008.

[51] M. Fuxreiter. Fold or not to fold upon binding - does it really matter? Curr Opin Struct Biol, 54:19–25, 2019.

[52] B. Gabrys and A. Bargiela. General fuzzy min-max neural network for clustering and classification. Ieee Transactions on Neural Networks, 11(3):769–783, 2000.

[53] I. Gath and A. B. Gev. Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell., 11(7):773–780, 1989.

[54] A. B. Geva. Hierarchical unsupervised fuzzy clustering. IEEE Transactions on Fuzzy Systems, 7(6):723–733, 1999.

[55] H. Gong, S. Zhang, J. Wang, H. Gong, and J. Zeng. Constructing structure ensembles of intrinsically disordered proteins from chemical shift data. J Comput Biol, 23(5):300–10, 2016.

[56] N. J. Greenfield. Using circular dichroism spectra to estimate protein secondary structure. Nat Protoc, 1(6):2876–90, 2006.

[57] I. Grundke-Iqbal, K. Iqbal, Y. C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in alzheimer cytoskeletal pathology. Proceedings of the National Academy of Sciences, 83(13):4913–4917, 1986.

[58] I. Grundke-Iqbal, K. Iqbal, Y. C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder. Abnormal phosphorylation of the microtubule-associated protein tau (tau) in alzheimer cytoskeletal pathology. Proceedings of the National Academy of Sciences, 83(13):4913–4917, 1986.

[59] S. J. Hagen, J. Hofrichter, A. Szabo, and W. A. Eaton. Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding. Proceedings of the National Academy of Sciences, 93(21):11615–11617, 1996.

[60] G. G. Hammes, Y.-C. Chang, and T. G. Oas. Conformational selection or induced fit: a flux description of reaction mechanism. Proceedings of the National Academy of Sciences of the United States of America, 106(33):13737–13741, 2009.

[61] R. J. Hathaway and J. C. Bezdek. Fuzzy c-means clustering of incomplete data. Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, 31(5):735–744, 2001.

[62] R. J. Hathaway, J. C. Bezdek, and Y. K. Hu. Generalized fuzzy c-means clustering strategies using lp norm distances. Ieee Transactions on Fuzzy Systems, 8(5):576–582, 2000.

[63] M. Held and F. Noe. Calculating kinetics and pathways of protein-ligand association. Eur J Cell Biol, 91(4):357–64, 2012.

[64] J. Higo, N. Ito, M. Kuroda, S. Ono, N. Nakajima, and H. Nakamura. Energy landscape of a peptide consisting of alpha-helix, 310-helix, beta-turn, beta-hairpin, and other disordered conformations. Protein Science, 10(6):1160–1171, 2001.

[65] J. Higo, Y. Nishimura, and H. Nakamura. A free-energy landscape for coupled folding and binding of an intrinsically disordered protein in explicit solvent from detailed all-atom computations. Journal of the American Chemical Society, 133(27):10448–10458, 2011.

[66] A. S. Holehouse, J. Ahad, R. K. Das, and R. V. Pappu. Cider: Classification of intrinsically disordered ensemble regions. Biophysical Journal, 108(2):228a, 2015.

[67] S. S. Hook and A. R. Means. Ca(2+)/cam-dependent kinases: from activation to function. Annu Rev Pharmacol Toxicol, 41:471–505, 2001.

[68] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins, 65(3):712–25, 2006.

[69] J. Huang and A. D. MacKerell. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. Journal of Computational Chemistry, 34(25):2135–2145, 2013.

[70] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmuller, and A. D. MacKerell. Charmm36m: an improved force field for folded and instrinsically disordered proteins. Nature Methods, 14:71–73, 2017.

[71] R. Huber. Conformational flexibility and its functional significance in some protein molecules. Trends in Biochemical Sciences, 4(12):271–276, 1979.

[72] A. Hudmon and H. Schulman. Structure-function of the multifunctional ca2+/calmodulin-dependent protein kinase ii. Biochem J, 364(Pt 3):593–611, 2002.

[73] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic, and A. K. Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol, 323(3):573–84, 2002.

[74] V. Iešmantavičius, J. Dogan, P. Jemth, K. Teilum, and M. Kjaergaard. Helical propensity in an intrinsically disordered protein accelerates ligand binding. Angewandte Chemie International Edition, 53(6):1548–1551, 2014.

[75] M. R. Jensen, R. W. Ruigrok, and M. Blackledge. Describing intrinsically disordered proteins at atomic resolution by nmr. Curr Opin Struct Biol, 23(3):426–35, 2013.

[76] M. R. Jensen, R. W. H. Ruigrok, and M. Blackledge. Describing intrinsically disordered proteins at atomic resolution by nmr. Current Opinion in Structural Biology, 23(3):426–435, 2013.

[77] M. R. Jensen, M. Zweckstetter, J.-r. Huang, and M. Blackledge. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using nmr spectroscopy. Chemical Reviews, 114(13):6632–6660, 2014.

[78] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22(12):2577–637, 1983.

[79] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. Ieee Transactions on Pattern Analysis and Machine Intelligence, 24(7):881–892, 2002.

[80] J. Karhunen, E. Oja, L. Y. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. Ieee Transactions on Neural Networks, 8(3):486–504, 1997.

[81] M. Karplus and D. M. Grant. A criterion for orbital hybridization and charge distribution in chemical bonds. Proc Natl Acad Sci U S A, 45(8):1269–73, 1959.

[82] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski. Coarse-grained protein models and their applications. Chemical Reviews, 116(14):7898–7936, 2016.

[83] M. A. Koch, L. O. Wittenberg, S. Basu, D. A. Jeyaraj, E. Gourzoulidou, K. Reinecke, A. Odermatt, and H. Waldmann. Compound library development guided by protein structure similarity clustering and natural product structure. Proc Natl Acad Sci U S A, 101(48):16721–6, 2004.

[84] R. Konrat. Nmr contributions to structural dynamics studies of intrinsically disordered proteins. J Magn Reson, 241(100):74–85, 2014.

[85] D. Koshland. Application of a theory of enzyme specificity of protein synthesis. Proc Nat Acad Sci USA., 44:98–104, 1958.

[86] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden markov-models in computational biology - applications to protein modeling. Journal of Molecular Biology, 235(5):1501–1531, 1994.

[87] Y. Kubota and M. N. Waxham. Lobe specific ca2+-calmodulin nanodomain in neuronal spines: a single molecule level analysis. PLoS Comput Biol, 6(11):e1000987, 2010.

[88] I. Kufareva and R. Abagyan. Methods of protein structure comparison. Methods in molecular biology (Clifton, N.J.), 857:231–257, 2012.

[89] S. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Statist., 22(1):79–86, 1951.

[90] P. S. Kumagai, R. DeMarco, and J. L. S. Lopes. Advantages of synchrotron radiation circular dichroism spectroscopy to study intrinsically disordered proteins. Eur Biophys J, 46(7):599–606, 2017.

[91] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande. Markov state model reveals folding and functional dynamics in ultra-long md trajectories. J Am Chem Soc, 133(45):18413–9, 2011.

[92] O. F. Lange, D. van der Spoel, and B. L. de Groot. Scrutinizing molecular mechanics force fields on the submicrosecond timescale with nmr data. Biophysical Journal, 99(2):647–655, 2010.

[93] Z. A. Levine, L. Larini, N. E. LaPointe, S. C. Feinstein, and J. E. Shea. Regulation and aggregation of intrinsically disordered peptides. Proc Natl Acad Sci U S A, 112(9):2758–63, 2015.

[94] C. Levinthal. How to fold graciously. Mossbauer spectroscopy in biological systems, 67:22–24, 1969.

[95] A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. Pattern Recognition, 36(2):451–461, 2003.

[96] J. Lincoff, S. Sasmal, and T. Head-Gordon. The combined force field-sampling problem in simulations of disordered amyloid-beta peptides. Journal of Chemical Physics, 150(10):14, 2019.

[97] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen, and M. Vendruscolo. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. J Am Chem Soc, 126(10):3291–9, 2004.

[98] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. Proteins-Structure Function and Bioinformatics, 78(8):1950–1958, 2010.

[99] X. Liu, J. Chen, and J. Chen. Residual structure accelerates binding of intrinsically disordered actr by promoting efficient folding upon encounter. Journal of Molecular Biology, 431(2):422–432, 2019.

[100] J. L. Lopes, A. J. Miles, L. Whitmore, and B. A. Wallace. Distinct circular dichroism spectroscopic signatures of polyproline ii and unordered secondary structures: applications in secondary structure analyses. Protein Sci, 23(12):1765–72, 2014.

[101] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci U S A, 107(18):8183–8, 2010.

[102] K. Matsuo, Y. Sakurada, S. Tate, H. Namatame, M. Taniguchi, and K. Gekko. Secondary-structure analysis of alcohol-denatured proteins by vacuum-ultraviolet circular dichroism spectroscopy. Proteins, 80(1):281–93, 2012.

[103] W. E. Meador, A. R. Means, and F. A. Quiocho. Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures. Science, 262(5140):1718–21, 1993.

[104] T. Meyer, P. I. Hanson, L. Stryer, and H. Schulman. Calmodulin trapping by calcium-calmodulin-dependent protein kinase. Science, 256(5060):1199–202, 1992.

[105] P. J. Michalski. The delicate bistability of camkii. Biophys J, 105(3):794–806, 2013.

[106] A. Micsonai, F. Wien, E. Bulyaki, J. Kun, E. Moussong, Y. H. Lee, Y. Goto, M. Refregiers, and J. Kardos. Bestsel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. Nucleic Acids Res, 46(W1):W315–W322, 2018.

[107] A. Micsonai, F. Wien, L. Kernya, Y. H. Lee, Y. Goto, M. Refregiers, and J. Kardos. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. Proc Natl Acad Sci U S A, 112(24):E3095–103, 2015.

[108] A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. Cortese, A. K. Dunker, and V. N. Uversky. Analysis of molecular recognition features (morfs). J. Mol. Biol., 362:1043–1059, 2006.

[109] J. Monod, J. Wyman, and J. P. Changeux. On the nature of allosteric transitions: A plausible model. J Mol Biol, 12:88–118, 1965.

[110] A. Mudher and S. Lovestone. Alzheimer's disease – do tauists and baptists finally shake hands? Trends in Neurosciences, 25(1):22–26, 2002.

[111] C. Navarro-Retamal, A. Bremer, J. Alzate-Morales, J. Caballero, D. K. Hincha, W. Gonzalez, and A. Thalhammer. Molecular dynamics simulations and cd spectroscopy reveal hydration-induced unfolding of the intrinsically disordered lea proteins cor15a and cor15b from arabidopsis thaliana. Phys Chem Chem Phys, 18(37):25806–16, 2016.

[112] D. Nelson and M. Cox. Lehninger Principles of Biochemistry. Macmillan Learning, 7 edition, 2017.

[113] H. Nguyen, D. R. Roe, and C. Simmerling. Improved generalized born solvent model parameters for protein simulations. J Chem Theory Comput, 9(4):2020–2034, 2013.

[114] F. Noe. Probability distributions of molecular observables computed from markov models. J Chem Phys, 128(24):244103, 2008.

[115] F. Noe, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proceedings of the National Academy of Sciences, 106(45):19011–19016, 2009.

[116] S. H. Northrup, S. A. Allison, and J. A. McCammon. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. Journal of Chemical Physics, 80(4):1517–1526, 1984.

[117] C. J. Oldfield and A. K. Dunker. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem, 83:553–84, 2014.

[118] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins, 55(2):383–94, 2004.

[119] M. Ota, R. Koike, T. Amemiya, T. Tenno, P. R. Romero, H. Hiroaki, A. K. Dunker, and S. Fukuchi. An assignment of intrinsically disordered

regions of proteins based on nmr structures. Journal of Structural Biology, 181(1):29–36, 2013.

[120] C. N. Pace, R. W. Alston, and K. L. Shaw. Charge-charge interactions influence the denatured state ensemble and contribute to protein stability. Protein Sci, 9(7):1395–8, 2000.

[121] C. N. Pace and J. M. Scholtz. A helix propensity scale based on experimental studies of peptides and proteins. Biophys J, 75(1):422–7, 1998.

[122] V. S. Pande, K. Beauchamp, and G. R. Bowman. Everything you wanted to know about markov state models but were afraid to ask. Methods, 52(1):99–105, 2010.

[123] H. Y. Park, S. A. Kim, J. Korlach, E. Rhoades, L. W. Kwok, W. R. Zipfell, M. N. Waxham, W. W. Webb, and L. Pollack. Conformational changes of calmodulin upon ca2+ binding studied with a microfluidic mixer. Proceedings of the National Academy of Sciences of the United States of America, 105(2):542–547, 2008.

[124] F. Paul, F. Noe, and T. R. Weikl. Identifying conformational-selection and induced-fit aspects in the binding-induced folding of pmi from markov state modeling of atomistic simulations. Journal of Physical Chemistry B, 122(21):5649–5656, 2018.

[125] K. Pauwels, P. Lebrun, and P. Tompa. To be disordered or not to be disordered: is that still a question for proteins in the cell? Cell Mol Life Sci, 74(17):3185–3204, 2017.

[126] N. Plattner and F. Noe. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and markov models. Nat Commun, 6:7653, 2015.

[127] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: Generation and validation. The Journal of Chemical Physics, 134(17):174105, 2011.

[128] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics, 29(7):845–54, 2013.

[129] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé. Identification of slow molecular order parameters for markov model construction. The Journal of Chemical Physics, 139(1):015102, 2013.

[130] P. Radivojac, S. Vucetic, T. O'Connor, V. N. Uversky, Z. Obradovic, and A. K. Dunker. Calmodulin signaling: analysis and prediction

of a disorder-dependent molecular recognition. Proteins: Structure, Function, Bioinformatics, 63:398–410, 2006.

[131] K. M. Ravikumar, W. Huang, and S. Yang. Coarse-grained simulations of protein-protein association: an energy landscape perspective. Biophysical journal, 103(4):837–45, 2012.

[132] P. Robustelli, S. Piana, and D. E. Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. Proc Natl Acad Sci U S A, 115:E4758–E4766, 2018.

[133] D. R. Roe and r. Cheatham, T. E. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. J Chem Theory Comput, 9(7):3084–95, 2013.

[134] J. M. Rogers, C. T. Wong, and J. Clarke. Coupled folding and binding of the disordered protein puma does not require particular residual structure. Journal of the American Chemical Society, 136(14):5197–5200, 2014.

[135] O. S. Rosenberg, S. Deindl, L. R. Comolli, A. Hoelz, K. H. Downing, A. C. Nairn, and J. Kuriyan. Oligomerization states of the association domain and the holoenyzme of ca2+/cam kinase ii. FEBS J, 273(4):682–94, 2006.

[136] M. Rubinstein and R. H. Colby. Polymer Physics. Oxford University Press, 2003.

[137] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics, 23(3):327–341, 1977.

[138] R. D. Schaeffer, A. Fersht, and V. Daggett. Combining experiment and simulation in protein folding: closing the gap for small model systems. Current Opinion in Structural Biology, 18(1):4–9, 2008.

[139] S. L. Shammas, A. J. Travis, and J. Clarke. Remarkably fast coupled folding and binding of the intrinsically disordered transactivation domain of cmyb to cbp kix. The Journal of Physical Chemistry B, 117(42):13346–13356, 2013.

[140] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. J Chem Theory Comput, 3(6):2312–34, 2007.

[141] J. M. Shifman, M. H. Choi, S. Mihalas, S. L. Mayo, and M. B. Kennedy. Ca2+/calmodulin-dependent protein kinase ii (camkii) is activated by calmodulin with two bound calciums. Proc Natl Acad Sci U S A, 103(38):13968–73, 2006.

[142] B. A. Shoemaker, J. J. Portman, and P. G. Wolynes. Speeding molecular recognition by using the folding funnel: The fly-casting mechanism.

Proceedings of the National Academy of Sciences of the United States of America, 97(16):8868–+, 2000.

[143] D. Shukla, A. Peck, and V. S. Pande. Conformational heterogeneity of the calmodulin binding interface. Nature Communications, 7, 2016.

[144] D. J. Sindhikara, S. Kim, A. F. Voter, and A. E. Roitberg. Bad seeds sprout perilous dynamics: Stochastic thermostat induced trajectory synchronization in biomolecules. Journal of Chemical Theory and Computation, 5(6):1624–1631, 2009.

[145] B. Slaughter, R. J. Bieber-Urbauer, and C. Johnson. Single-molecule tracking of sub-milisecond domain motion in calmodulin. J. Phys. Chem. B., 109:12658–12662, 2005.

[146] B. D. Slaughter, M. W. Allen, J. R. Unruh, R. J. B. Urbauer, and C. K. Johnson. Single-molecule resonance energy transfer and fluorescence correlation spectroscopy of calmodulin in solution. Journal of Physical Chemistry B, 108(29):10388–10397, 2004.

[147] M. V. Smoluchowski. Drei vortrage uber diffusion, brownsche bewegung und koagulation von kolloidteilchen. Physik. Zeit., 17:557–585, 1916.

[148] N. Sreerama and R. W. Woody. Estimation of protein secondary structure from circular dichroism spectra: comparison of contin, selcon,

and cdsstr methods with an expanded reference set. Anal Biochem, 287(2):252–60, 2000.

[149] N. Sreerama and R. W. Woody. On the analysis of membrane protein circular dichroism spectra. Protein Sci, 13(1):100–12, 2004.

[150] J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case. Application of a pairwise generalized born model to proteins and nucleic acids: inclusion of salt effects. Theoretical Chemistry Accounts, 101(6):426–434, 1999.

[151] J. Stigler, F. Ziegler, A. Gieseke, J. C. M. Gebhardt, and M. Rief. The complex folding network of single calmodulin molecules. Science, 334(6055):512–516, 2011.

[152] M. Stratton, I.-H. Lee, M. Bhattacharyya, S. M. Christensen, L. H. Chao, H. Schulman, J. T. Groves, and J. Kuriyan. Activation-triggered subunit exchange between camkii holoenzymes facilitates the spread of kinase activity. eLife, 3:e01610 C1 – eLife 2014;3:e01610, 2014.

[153] M. S. Su and C. H. Chou. A modified version of the k-means algorithm with a distance based on cluster symmetry. Ieee Transactions on Pattern Analysis and Machine Intelligence, 23(6):674–680, 2001.

[154] R. M. Tombes, M. O. Faison, and J. M. Turbeville. Organization and evolution of multifunctional ca(2+)/cam-dependent protein kinase genes. Gene, 322:17–31, 2003.

[155] P. Tompa and M. Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. Trends in Biochemical Sciences, 33(1):2–8, 2008.

[156] H. T. Tran, A. Mao, and R. V. Pappu. Role of backbone-solvent interactions in determining conformational equilibria of intrinsically disordered proteins. J Am Chem Soc, 130(23):7380–92, 2008.

[157] J. Trylska, T. Tozzini, C.-e. Chang, and J. A. McCammon. Hiv-1 protease substrate binding and product release pathways released pathways explored with coarse-grained molecular dynamics. Biophysical journal, 92:4179–4187, 2007.

[158] V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins: Structure, Function, and Bioinformatics, 41(3):415–427, 2000.

[159] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Showing your id: intrinsic disorder as an id for recognition, regulation and cell signaling. J Mol Recognit, 18(5):343–84, 2005.

[160] M. Vijayakumar, K. Y. Wong, G. Schreiber, A. R. Fersht, A. Szabo, and H. X. Zhou. Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on barnase and barstar. J Mol Biol, 278(5):1015–24, 1998.

[161] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic. Flavors of protein disorder. Proteins, 52(4):573–84, 2003.

[162] Q. Wang, P. Z. Zhang, L. Hoffman, S. Tripathi, D. Homouz, Y. Liu, M. N. Waxham, and M. S. Cheung. Protein recognition and selection through conformational and mutually induced fit. Proceedings of the National Academy of Sciences of the United States of America, 110(51):20545–20550, 2013.

[163] W. Wang, W. Ye, C. Jiang, R. Luo, and H.-F. Chen. New force field on modeling intrinsically disordered proteins. Chemical Biology & Drug Design, 84(3):253–269, 2014.

[164] M. N. Waxham, A. L. Tsai, and J. A. Putkey. A mechanism for calmodulin (cam) trapping by cam-kinase ii defined by a family of cam-binding peptides. J Biol Chem, 273(28):17579–84, 1998.

[165] P. Weinkam, E. V. Pletneva, H. B. Gray, J. R. Winkler, and P. G. Wolynes. Electrostatic effects on funneled landscapes and structural diversity in denatured protein ensembles. Proc Natl Acad Sci U S A, 106(6):1796–801, 2009.

[166] L. Whitmore and B. A. Wallace. Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. Biopolymers, 89(5):392–400, 2008.

[167] C. Wiedemann, P. Bellstedt, and M. Gorlach. Capito–a web server-based analysis and plotting tool for circular dichroism data. Bioinformatics, 29(14):1750–7, 2013.

[168] P. E. Wright and H. J. Dyson. Linking folding and binding. Current opinion in structural biology, 19(1):31–8, 2009.

[169] P. E. Wright and H. J. Dyson. Intrinsically disordered proteins in cellular signalling and regulation. Nature Reviews Molecular Cell Biology, 16:18, 2014.

[170] H. Wu, P. G. Wolynes, and G. A. Papoian. Awsem-idp: A coarse-grained force field for intrinsically disordered proteins. J Phys Chem B, 2018.

[171] R. Xu and n. Wunsch, D. Survey of clustering algorithms. IEEE Trans Neural Netw, 16(3):645–78, 2005.

[172] X.-P. Xu and D. A. Case. Automated prediction of 15n, 13cα, 13cβ and 13c' chemical shifts in proteins using a density functional database. Journal of Biomolecular NMR, 21(4):321–333, 2001.

[173] W. Zhong, G. Altun, R. Harrison, P. C. Tai, and Y. Pan. Improved k-means clustering algorithm for exploring local protein sequence motifs representing common structural property. IEEE Transactions on Nanobioscience, 4(3):255–265, 2005.

[174] H.-X. Zhou and P. A. Bates. Modeling protein association mechanisms and kinetics. Current Opinion in Structural Biology, 23(6):887–893, 2013.

[175] R. Zhou. Free energy landscape of protein folding in water: explicit vs. implicit solvent. Proteins, 53(2):148–61, 2003.

[176] X. H. Zhuang, Y. Huang, K. Palaniappan, and Y. X. Zhao. Gaussian mixture density modeling, decomposition, and applications. Ieee Transactions on Image Processing, 5(9):1293–1302, 1996.

[177] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. Proceedings of the National Academy of Sciences, 89(1):20–22, 1992.