Determination of Biomolecular Interdomain Motions using Nuclear Magnetic Resonance

by

Yang Qi

Department of Biochemistry
Duke University

Date: ______________________

Approved:

____
Terrence G. Oas, Supervisor

____
Hashim M. Al-Hashimi

____
Bruce R. Donald

____
Leonard D. Spicer

____
Pei Zhou

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Biochemistry
in the Graduate School of Duke University
2016

Abstract

Determination of Biomolecular Interdomain Motions using Nuclear Magnetic Resonance

by

Yang Qi

Department of Biochemistry
Duke University

Date: ____________________________

Approved:

____
Terrence G. Oas, Supervisor

____
Hashim M. Al-Hashimi

____
Bruce R. Donald

____
Leonard D. Spicer

____
Pei Zhou

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Biochemistry
in the Graduate School of Duke University
2016

Abstract

Biological macromolecules can rearrange interdomain orientations when binding to various partners. Interdomain dynamics serve as a molecular mechanism to guide the transitions between orientations. However, our understanding of interdomain dynamics is limited because a useful description of interdomain motions requires an estimate of the probabilities of interdomain conformations, increasing complexity of the problem.

Staphylococcal protein A (SpA) has five tandem protein-binding domains and four interdomain linkers. The domains enable *Staphylococcus aureus* to evade the host immune system by binding to multiple host proteins including antibodies. Here, I present a study of the interdomain motions of two adjacent domains in SpA. NMR spin relaxation experiments identified a 6-residue flexible interdomain linker and interdomain motions. To quantify the anisotropy of the distribution of interdomain orientations, we measured residual dipolar couplings (RDCs) from the two domains with multiple alignments. The N-terminal domain was directly aligned by a lanthanide ion and not influenced by interdomain motions, so it acted as a reference frame to achieve motional decoupling. We also applied *de novo* methods to extract spatial dynamic information from RDCs and represent interdomain motions as a continuous distribution on the 3D rotational space. Significant anisotropy was observed in the distribution, indicating the motion populates some interdomain orientations more than others. Statistical thermodynamic analysis of the observed orientational

distribution suggests that it is among the energetically most favorable orientational distributions for binding to antibodies. Thus, the affinity is enhanced by a pre-posed distribution of interdomain orientations while maintaining the flexibility required for function.

The protocol described above can be applied to other biological systems in general. Protein molecule calmodulin and RNA molecule transactivation response element (TAR) also have intensive interdomain motions with relative small intradomain dynamics. Their interdomain motions were studied using our method based on published RDC data. Our results were consistent with literature results in general. The differences could be due to previous studies' use of physical models, which contain assumptions about potential energy and thus introduced non-experimental information into the interpretations.

To my parents.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

CDIO	Continuous distribution of interdomain orientations
FAA	Frame axis-angle
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
DOF	Degree of freedom
NMR	Nuclear magnetic resonance
CSA	Chemical shift anisotropy
HetNOE	Heteronuclear Overhauser effect
RDC	Residual dipolar coupling
PCS	Pseudo-contact shift
PRE	Paramagnetic relaxation enhancement
FRET	Fluorescence resonance energy transfer
LRET	Luminescence/Lanthanide-based resonance energy transfer
SAXS	Small angle X-ray scattering
SpA-N	N terminal domains of Staphylococcal protein A
CaM	Calmodulin
CYPA	Cyclophilin A
DHFR	Dihydrofolate reductase
RUF	Rapid unfolding and folding

SVD	Singular value decomposition
OLC	Orthogonal linear combination
LBT	Lanthanide binding tag
TAR	Trans-activation response element
LTR	Long terminal repeat

Acknowledgements

I would like to thank my advisor, Dr. Terrence Oas for his guidance over the years. I also thank my committee members, Dr. Bruce Donald, Dr. Leonard Spicer, Dr. Hashim Al-Hashimi and Dr. Pei Zhou.

I would like to thank all the Oas lab members, including Dr. Andrew Hagarman, Dr. Pei-fen Liu, Danna Li, Dr. Jo-Anna Capp, Dr. Kyle Daniels, Pamela Mosley, Jonathan Li and Roy Hughes. I especially would like to thank Andrew, who had been my mentor for my first two years and helped me to kick-start the project. I also thank François Thálot, who is a joint undergraduate student of both the Oas lab and the Donald lab. In addition, I give my sincere gratitude to the Duke NMR center stuffs, Dr. Ronald Venters, Dr. Anthony Ribeiro and Mr. Donald Mika. I thank for their help and kindness when scheduling experiments, setting up experiments and trouble shooting.

Finally, I would like to thank my supportive friends and family. I wish to give my best regards to my friends Weiwei Li, Hui Kang, Dr. Yu Jiang, Xiao Yan, Ruo He, Shiwen Zhao, Rujie Yin and Dr. Tingran Gao, who share their life moments with me.

1

Introduction

1.1 Biomolecular Dynamics

1.1.1 Energy landscape and conformational distribution

Biomolecules are selected by evolution to perform biological functions in living cells. Living organisms are such complicated systems that they require an enormous amount of biomolecules with varying functions. Surprisingly, different biomolecules all share the same building blocks. Proteins are composed of amino acid residues. RNAs and DNAs are composed of nucleic acids. Evolution diversifies the biomolecules by varying the sequences of the building blocks. As a result, the function of a biomolecule is largely dictated by its sequence. Anfinsen's dogma postulates that the native protein structure is determined only by its amino acid sequence [1]. There is indeed a strong correlation between sequence and structure, verified by numerous studies [2, 3]. Since then, the structure-function relationship has been studied extensively. Thousands of biomolecule structure models have been determined in the hope that a unique structure model of the folded state can reveal the mechanism of the molecule on an atomic level. However, the dynamic nature of biomolecules is largely overlooked.

Living organisms are dynamic systems from the top to bottom levels. The migration of cells, the translation and rotation of molecules and the hopping of electrons are all needed for living organisms to function properly. On the molecular level, biomolecules are ‘screaming and kicking’ according to Weber [4]. Such a description, although intuitive, is not quantitative enough to accurately characterize the dynamic nature of biomolecules. A complete description of biomolecular dynamics requires a multi-dimensional energy landscape, which encompasses all possible conformations of a biomolecule and their associated state energies [5]. The concept of energy landscape is most familiar in the protein folding field, for example, the folding funnel hypothesis shown in Fig. 1.1 [6]. However, it can be generalized to folded biomolecules. From Fig. 1.1 and Fig. 1.2, we observe that the folding funnel and the energy landscape are basically the same concept. The only difference is that the energy landscape usually focuses on the low energy states instead of the whole. The surface of a energy landscape describes how the potential energy varies with the biomolecular geometry, which is defined by the conformational coordinates over all degrees of freedom (DOFs) of the molecule. In this multi-dimensional space, the energy landscape may have multiple local minima with significant populations. When the barriers between local minima are low, thermal fluctuations can drive the transition from one minimum to another minimum, resulting in motions of biomolecules. Because the energy landscape defines the energy differences between states as well as the energy barriers between them, it is comprehensive enough to cover both the thermodynamic and the kinetic behavior of a biomolecule. In addition, the energy landscape is a mapping of conformations to their corresponding state energies and thereby characterizes both the dynamic and the structural features of a biomolecule. In order to understand a biomolecule in action, the energy landscape needs to be studied in addition to the structures.

Considering the large number of DOFs available to a biomolecule, the energy landscape is high-dimensional and thus difficult to map. If every portion of the energy

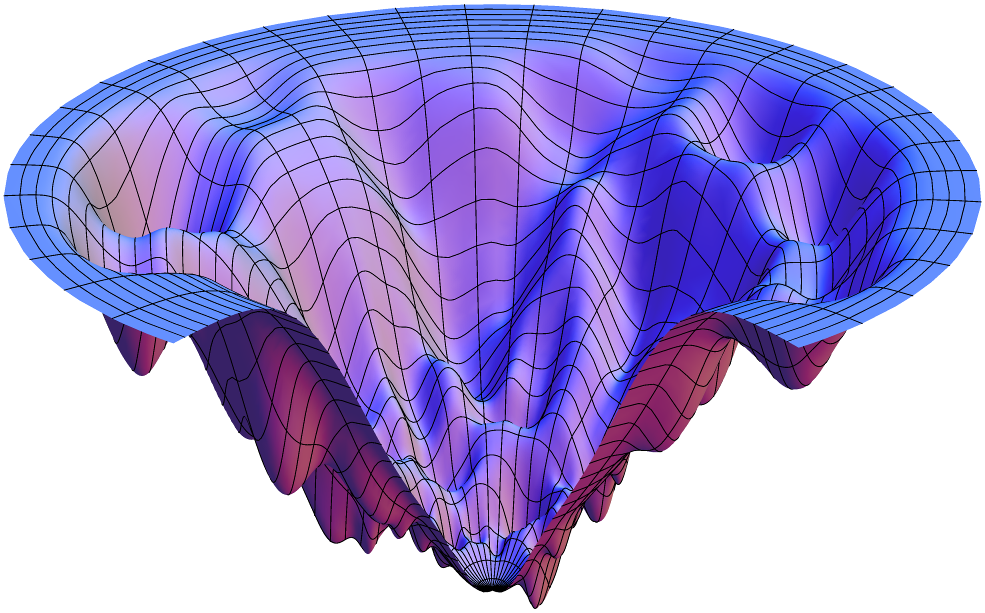


Figure 1.1: The folding funnel of a biomolecule. The xy-plane is an simplification of the conformational space which includes all possible conformations. The z-axis represents free energy.

landscape is physically accessible and the energy landscape provides no guidance that avoids having to sample all of conformational space, we will encounter the Levinthal's paradox [7]. The number of all possible conformations is astronomically large. It is impossible to enumerate all conformations in a reasonable time, let alone map the conformations to their corresponding state energies in an energy landscape. However, biomolecules, as well as their energy landscapes, are optimized by evolution. As a result, folded systems converge to the native state. Equilibrium fluctuations

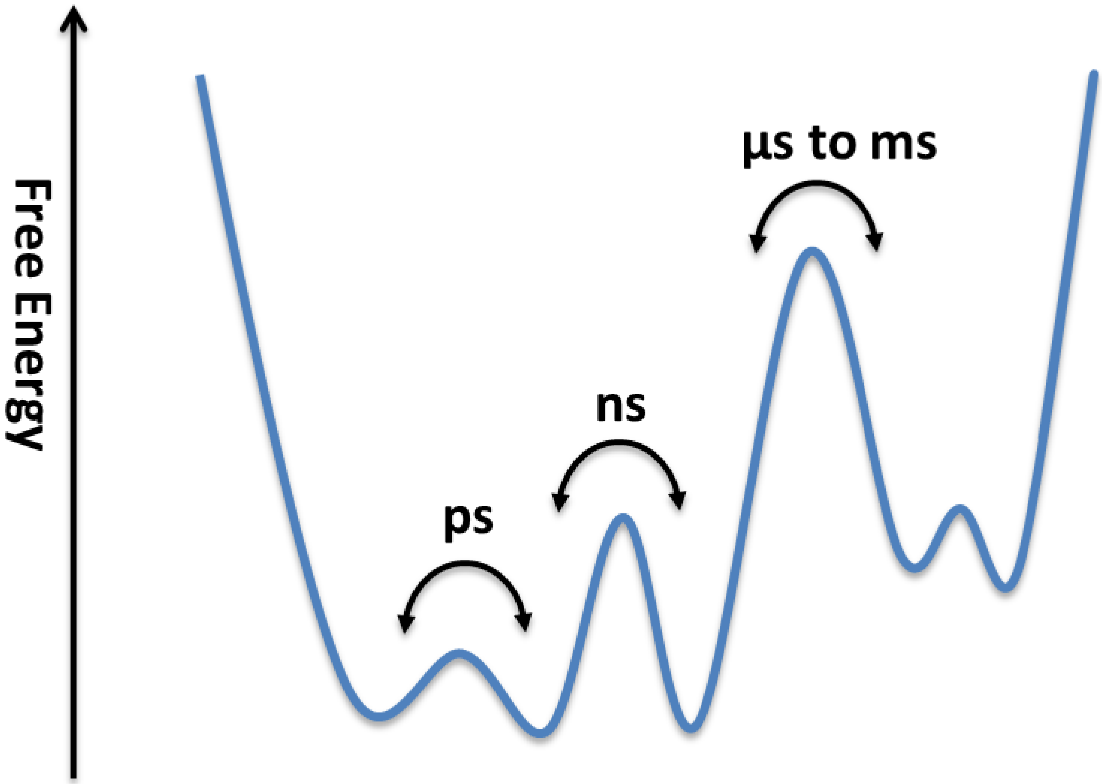


FIGURE 1.2: The energy landscape of a biomolecule. The x-axis is a simplification of the conformational space which includes all possible conformations. The y-axis represents free energy.

drive the folded systems to sample conformations close to the native state. In this case, the energy landscape in focus is the portion surrounding the native state and the sampled states only take a small portion of the high-dimensional space of an energy landscape. It is more difficult to characterize flexible systems because of the increase in dimensionality. Fortunately, different types of motions may not be coupled. Assuming decoupling between different motions, we can focus on the DOFs associated with a certain type of motion, thus reducing the complexity of the problem.

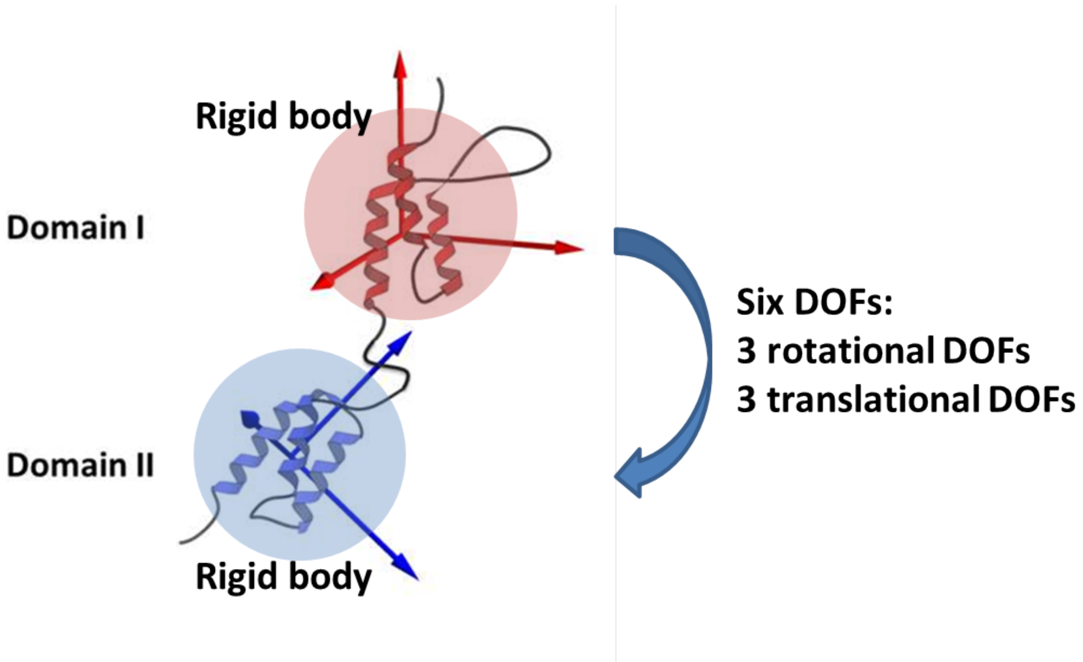


FIGURE 1.3: The domains are considered as rigid bodies by ignoring intradomain motions. Six degrees of freedom (DOFs), three translational DOFs and three rotational DOFs exist between the two domains.

For example, interdomain motions between two domains only have six DOFs when the intradomain dynamics are ignored (Fig. 1.3). In the worst-case scenario, when the motions of a flexible system depend heavily on each other, it is still possible to map the energy landscape on a low-dimensional projection. The choice of dimension includes the number of native contacts and the radius of gyration.

In the energy landscape, different conformations have different energies. The energy of each conformation determines the conformation's population. When the absolute energy is not of concern, information in the energy landscape can be represented by a distribution of populations. In the distribution, a probability is assigned to each conformation. The conformational distribution is an alternative way to char-

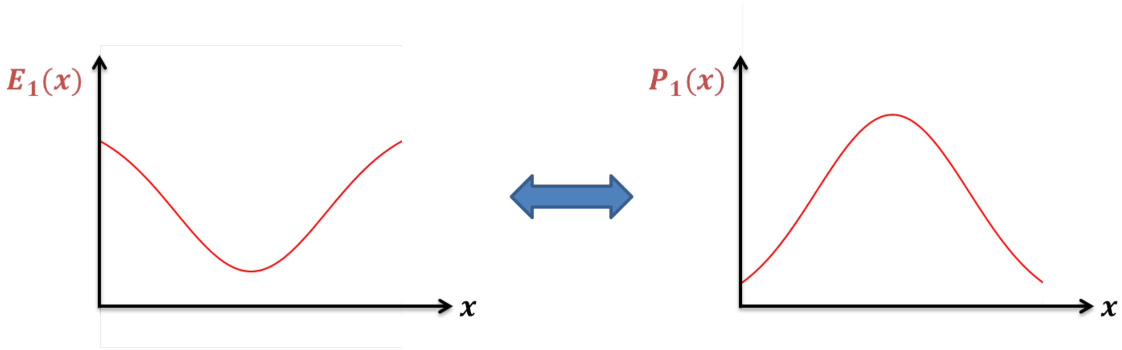


FIGURE 1.4: An energy landscape is equivalent to a conformational distribution. Energy landscapes and conformational distributions are both maps of the conformational space. Energy landscapes map conformations to free energies. Conformational distributions map conformations to populations.

acterize the energy landscape. Because there is a one-to-one correspondence between energy and population, the conformational distribution is as informative as an energy landscape (Fig. 1.4). In practice, the transition state between conformations has low population and thus it is hardly visible to most experimental observables. The reconstructed conformational distribution may only be a good representation of the major states rather than all the conformations. Nonetheless, the conformational distribution captures most features of the thermodynamic component of the energy landscape.

The above discussion of the energy landscape generally refers to the energy landscape of the native state of a ligand/substrate-free molecule. However, biomolecules perform their functions by binding to either to ligands or substrates, resulting in signal transduction or chemical reactions. When the dimension of the binding reaction or the catalysis reaction is taken into account, the picture of a biomolecule becomes even more complicated. Besides all the dimensions involved in one energy landscape, there is an extra dimension along which the energy landscape morphs. For a two state binding reaction, there is the energy landscape of the molecule when bound in addition to the energy landscape when the molecule is free. The two energy landscapes dictate the binding mechanism. There are two popular models to describe the binding mechanism, the induced fit model [8] and the conformation selection model [9, 10]. The two models differ in the sequential order of binding and conformation rearrangement. In the induced fit model, a molecule binds to its binding partner first and then goes through an induced conformation change [8]. In the conformation selection model, a molecule samples its binding conformation first and then binds to its binding partner [9, 10]. In reality, a molecule could adopt either binding mechanism or a partition of the two [11]. A detailed mapping of the energy landscapes can explain the binding mechanism. In some cases, the knowledge about the ligand-free energy landscape alone can help us infer the binding mechanism [12, 13, 14, 15, 16, 17, 18]. When the kinetic barrier in the ligand-free energy landscape matches the on-and-off rates of the binding reaction, we can infer that the conformation rearrangement in the absence of the binding partner is the rate limiting step and it is a necessary step in the binding reaction.

1.1.2 The time and spatial scale of biomolecular dynamics

Dynamics of biological macromolecules occur over various spatial and time scales (Fig. 1.5). On the top level, there are the translational motion and rotational motion of the whole molecule. Both of them are Brownian motions caused by random collisions mainly with solution molecules. Although the translational motion is important for the spatial distribution of biomolecules in a living cell, this kind of motion is out of the scope of our discussion. The rotational motion, or the global tumbling is observable in several NMR experiments. If the biomolecule is globular and roughly spherical, the global tumbling is isotropic in the sense that rotation of the molecule along any direction has the same rate. Otherwise, the global tumbling is anisotropic and it can not be described by a single rotation rate. Besides the shape of the molecule, other factors contribute to global tumbling. In the presence of a magnetic field, a molecule with significant diamagnetism or paramagnetism can interact with the field. This interaction results in a change in potential energy, thus affects orientation preference and possibly rotation rates. Although a non-spherical molecule may have anisotropic tumbling, it usually has no preference for a certain orientation because the energy potential around the molecule is uniform. Under a non-uniform energy potential field, the orientation distribution is anisotropic. Besides the global motions, several levels of dynamics exist within a molecule. If a molecule contains more than one domain, there may exist interdomain motions. Similar to the global motions, interdomain motions also have two components: the translation between adjacent domains and the relative orientation of the adjacent domains. In addition, intradomain motions such as backbone motions, sidechain rearrangements and bond vibrations are also important components of biomolecular dynamics. However, bond vibrations are usually decoupled from other motional modes because of their ultrafast timescale. As a result, bond vibrations are ignored when they are not directly involved in biological functions.

Dynamics can be separated into multiple tiers according to their timescales [19]. Tier-0 dynamics occur on the timescale of microseconds and slower. The energy barriers for tier-0 dynamics are on the order of kT, the product of the Boltzmann constant k and the absolute temperature T in Kelvin. Tier-1 dynamics occur on the timescale between nanoseconds and microseconds. Consequently, the energy barriers for tier-1 are smaller, usually in the range of less than 1 kT. Tier-2 dynamics occur on the timescale between picoseconds to nanoseconds. These energy barriers are even smaller. Higher tier dynamics also exist, such as the bond vibrations. However, as discussed earlier, higher tier dynamics are less involved in biological functions.

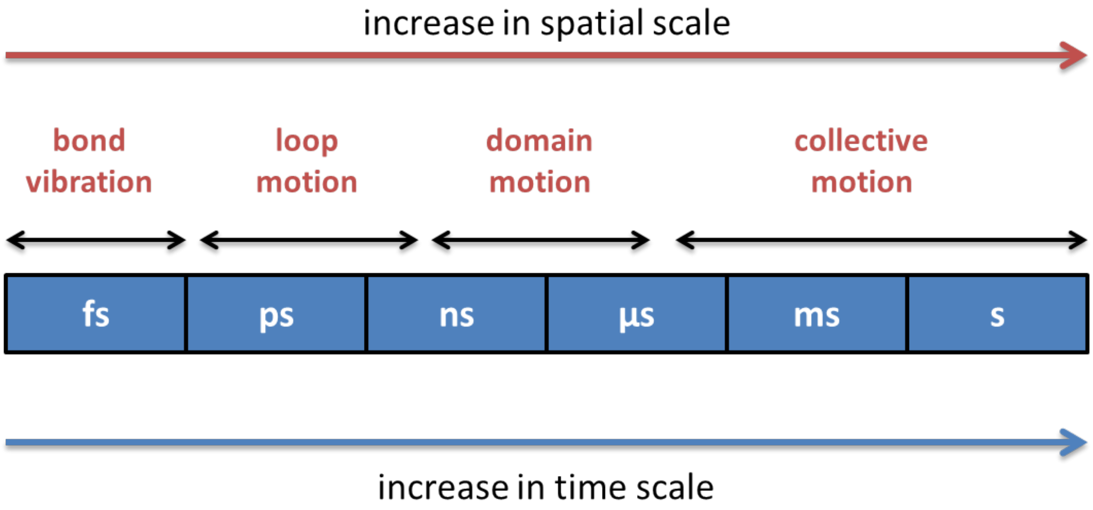


Figure 1.5: Spatial and time scales of biomolecular dynamics. Biomolecular dynamics occurs on varying spatial and time scales.

Tier-0 dynamics are considered to be slow timescale dynamics. The conformational states separated by tier-0 energy barriers have slow transitions between each other. The conformational change usually corresponds to large collective motions. On the other hand, tier-1 and tier-2 dynamics are considered fast timescale dynamics. Fast dynamics involves transitions between conformational substates. Although there is a strong correlation between the motion amplitude and the motion timescale, it should be noted that exceptions do exist. Large collective motions can be faster than small local motions. For example, rapid unfolding and folding (RUF) proteins can fold and unfold in milliseconds [20]. In contrast, proline isomerization may take seconds [21].

1.2 A Overview of Methods for Probing Dynamics

1.2.1 Nuclear magnetic resonance (NMR)

Nuclear magnetic resonance (NMR) was original developed by physicists to perturb the quantum states of nuclei [22, 23]. After going through several technical renovations, NMR has become a modern technique applied widely in physics, chemistry, biology and biomedical research. Although magnetic resonance imaging (MRI) is an important application of NMR theory for imaging, MRI is excluded from the following discussion. This brief overview focuses on the dynamic NMR techniques used to study the interplay between biomolecular structure, dynamics and function. In the following, I go through NMR techniques that are capable of directly observing dynamics of biomolecules in solution state and that are sensitive to a wide range of timescales. More importantly, dynamics are measured by manipulating and observing nuclei. The manipulation of the nuclei usually has small effects on dynamics of the observed molecule and thus maintains the integrity of the observed information. The dynamic information obtained by NMR is a beautiful complement to the structural information obtained by conventional X-ray crystallography.

NMR spin relaxation experiments can observe dynamics occurring in the picosecond to nanosecond timescale [24]. NMR can manipulate the magnetization of a nucleus and observe the change in magnetization over time. The change in magnetization depends on many factors, including the global tumbling of the molecule and the local motions of chemical bonds connecting two nuclei. These factors influence the fluctuations in local magnetic fields within the molecule. Because spins interact with the system through local magnetic fields, these factors determine the relaxation behavior of the spins, e.g. how spins relax to thermal equilibrium. The most commonly used spin pair is the ^{1}H-^{15}N spin pair. When there are motions, the bond vector of the ^{1}H-^{15}N spin pair reorients and thus changes the local magnetic field.

By monitoring the change of the ^{15}N magnetization, we obtain information about the orientational motion of the ^{1}H-^{15}N bond vector. For protein molecules, there is one H-N bond vector for each residue excepting proline. The H-N bond vector can serve as a probe to observe protein dynamics. As a result, we have uniformly distributed probes along the backbone of a protein molecule, with a little help from well-established isotope-labeling techniques. The rotational motion of a H-N bond vector is a mixture of multiple motional modes over various time scales. The motional modes include the global tumbling of the molecule, the local fluctuations of the backbone on the nanosecond to picosecond timescale and large collective motions occurring on slower time scales. However, spin relaxation experiments are not sensitive to motions on slower time scales. Spin relaxation experiments measure the relaxation of the ^{15}N magnetization. The relaxation happens most effectively when the frequency of the motion is close to the Larmor frquency of the nuclei. Because the Larmor frequencies of both ^{15}N and ^{1}H correspond to the nanosecond to picosecond timescale, the spin relaxation observables do not contain much information about faster motions and slower motions. Models of the motional modes were built in order to extract the information observed by spin relaxation experiments. The most widely used one is the Lipari-Szabo model. There are mainly two motional modes in the model, global tumbling and local bond reorientations. Effects resulting from other motional modes are pushed into a residual parameter, which represents slower motions. Both motional modes are in the nanosecond to picosecond timescale and thus are observable by spin relaxation experiments. For a globular biomolecule with a moderate molecular weight in the order of 1kD or 10kD, global tumbling usually occurs on the nanosecond timescale. In the Lipari-Szabo model, the global tumbling rate is described by one parameter if the motion is isotropic and it is described by three parameters if the motion is anisotropic. Local bond reorientations are faster than the global tumbling, occurring on the picoseconds timescale. One parameter is needed to describe the local motion frequency for each bond vector. In addition, because there are two components, we also need parameters to describe the weighting of the two components for each residue. The weighting coefficient is called the generalized order parameter. When the backbone is rigid, the global tumbling component is dominant and the order parameter is high. Otherwise, the local bond reorientation component is dominant and the order parameter is low. By calculating the order parameters, we can infer the flexibility of the backbone on the nanosecond to picosecond timescale. A detailed description of the spin relaxation technique can be found in section 2.1. Spin relaxation experiments are not limited to either protein systems or to the observation of backbone motions. Using other versions of the experiment, side-chain motions in protein systems and bond reorientations in DNA or RNA molecules can also be observed.

Relaxation dispersion experiments can observe the dynamics occurring in the microsecond to millisecond timescale [24]. The technique is built upon the spin relaxation experiments. After taking out the information of dynamics on the nanosecond to picosecond timescale, the residual information in spin relaxation experiments is about slower motions. For a single spin relaxation experiment, the information is very limited. In order to gain information about slower motions, relaxation dispersion experiments are used [24]. These slower motions are usually large collective motions accompanied by a dramatic change of chemical environment. The rate of the change in chemical environment is called the chemical exchange rate. Instead of varying the time interval to observe the decay of the magnetization, relaxation dispersion experiments vary the number of refocusing pulses and observe the change in line width. When the frequency of applied refocusing pulses matches the frequency of slower motions, the line broadening effect of chemical exchanges is most significant. Thus, by increasing the frequency of the applied refocusing pulses, the broadness of a peak changes. From the trend, we could infer when frequency of the applied refocusing pulses matches the chemical exchange rates. In addition to chemical exchange rates, relaxation dispersion experiments also provide information about the relative populations and the chemical environments of the two states. The observed timescale of relaxation dispersion experiments is limited to the microsecond to millisecond timescale. When the motion occurs slower than milliseconds, it does not cause line broadening. When the motion occurs faster than microseconds, the line broadening effect cannot be reduced because the frequency of applied refocusing pulses is limited by our physical equipment. However, when the motion indeed occurs in the microsecond to millisecond timescale, relaxation dispersion is a very powerful tool and it provides multiple types of useful information. In section 1.3, I will show several examples where the multiple pieces of information obtained by relaxation dispersion are used to fill gaps and solve puzzles.

Residual dipolar couplings (RDCs) can observe dynamics occurring from picoseconds to milliseconds [25, 26]. The measurement fill in the information about submicroseconds dynamics that are invisible to the previous two methods. It should be noted that information is encoded differently in RDCs. Spin relaxation and relaxation dispersion experiments measure the relaxation behavior of a spin by observing the change in either peak volume, intensity or width. However, as the name suggests, RDC experiments observe dipolar couplings. Dipolar coupling orginates from the dipole-dipole interaction in a magnetic field, which depends on the angle between the bond vector and the magnetic field. When the bond vector reorients itself due to either global or local motions, we observe an average of the dipolar couplings corresponding to each bond orientation. More precisely, the observed dipolar coupling is an average over time and all the molecules in solution. This average is called residual dipolar coupling. When a molecule tumbles isotropically in solution, the average becomes zero and we can not observe RDCs. In order to observe RDCs, the molecule must have a preference for certain orientations over others. This preference is achieved by alignment, which weakly restricts the tumbling of the molecule to a slightly anisotropic motion. When the RDCs are observable, they carry information about motions in the picosecond to millisecond timescale. It should be noted that there is a distinct difference between the dynamic information in RDCs and the dynamic information observed by relaxation experiments. Although RDCs observe a wide range of timescales, the averaging process that generates RDCs does not distinguish motions in different timescales. As a result, RDCs do not contain kinetic information in contrast to the relaxation methods. However, we can still separate the effects of different motional modes when they are not coupled with each other. When there are additional motional modes in a region of a biomolecule, the region tends to have smaller RDCs due to the additional averaging. When compared to another more rigid region, the local motion amplitude of the flexible region can be inferred. A detailed discussion of RDCs can be found in section 3.1.

Pseudo-contact shifts (PCSs) are very similar to RDCs. They both are temporal and spatial averages [27]. However, PCSs of a molecule are only observable when the molecule is aligned by a paramagnetic ion. The free electron in the paramagnetic ion has a large paramagnetism. In a magnetic field, the free electron spin is polarized by the external magnetic field and it creates an additional magnetic field parallel to the external magnetic field. Because the paramagnetism is anisotropic, the interaction with the magnetic field creates different potential energies for different orientations. As a result, paramagnetism leads to alignment. The additional magnetic field also changes the chemical environment of surrounding nuclei and perturbs their chemical shifts. The free electron perturbs the chemical shifts as if there is a contact between the free electron and the nuclei. Consequently, the perturbation in chemical shift is referred to as a pseudo-contact shift. Contrary to RDCs, PCSs are averages over both orientation and distance. PCSs also depend on the angle between a vector and the magnetic field. In this case, the interaction between the paramagnetic ion and a nucleus defines orientation and distance. On one hand, PCSs contain information invisible to RDCs because PCSs are distance dependent. On the other hand, information in PCSs is hard to utilize because the orientation information and the distance information are convolved with each other. Nonetheless, PCSs are very useful observables to determine dynamics.

Nuclear overhauser effects (NOEs) and paramagnetic relaxation enhancements (PREs) are also major NMR observables. Both of them contain motionally averaged information about distances. They are used more often in structure determination than the inference of dynamics because they provide limited temporal information. Both of them are proportional to r^{-6} where r is the distance between spins. Because of the r^{-6} dependence, conformations with small distances can dominate the signal intensity in the NOE case and dominate the relaxation rate in the PRE case. By taking advantage of the distance dependence, PREs can be used to detect minor conformations. When a minor conformation has small PRE distances, the r^{-6} weighting provided by the small distance can dramatically increase the contribution of the minor conformation. As a result, an ’invisible’ state can be observed. NOEs and PREs generally do not support the modeling of a structure ensemble with more than two members [28]. Because the distributional information content of these methods is limited, a large number of DOFs in the model can easily cause over-fitting. Consequently, they are usually used with other observables to constraint models of biomolecular dynamics.

1.2.2 X-ray crystallography

Since the discovery of myoglobin, X-ray crystallography has been a major technique to determine biomolecular structures. To produce an X-ray diffraction pattern, the molecules have to arrange themselves in a crystal. This crystallization usually locks the molecules into a single conformation and eliminates most of the dynamics in solution. However, when the amplitude of the motions is not large, we could still observe the dynamics in crystals through several advanced techniques. In the following, I review three X-ray crystallography techniques to determine dynamics in crystal.

Time-resolved X-ray crystallography is a technique to observe molecules in action and to take snapshots of the molecule when it perform its function. It can provide both detailed structural information at atomic resolution as well as conformation changes along a reaction path. In some cases, the technique is able to capture short-lived intermediate states [29, 30]. There are two ways to take snapshots of the conformational changes. One way is to trap the molecule into certain conformations. General trapping methods include changes in pH or temperature and the use of inhibitor. When the molecule is trapped into a conformation, the population of the conformation accumulates and the conformation becomes dominant in the crystal. The crystal with the trapped conformation can be used for X-ray diffraction and structure determination. Although this method is very useful, it can only be applied to trappable molecules. It should also be noted that artifacts can be generated during the trapping process. The other method is truly time resolved X-ray crystallography. In the experiments, the reaction is triggered in the crystal and X-ray diffraction patterns are collected for multiple time points. Several criteria need to be satisfied in order to perform the time resolved experiments. First, the biomolecule has to be active in the crystal. The molecule has to be capable of catalyzing its substrate or binding to its binding partner. Second, the reaction must be triggered at the same time in the crystal. In other words, the reactions of the all molecules in the crystal have to be synchronized. Third, the intermediate state has to be significantly populated. Otherwise, the X-ray diffraction pattern of the intermediate state is too weak to be detectable. Reaction triggering is a crucial part of the time-resolved experiments. The most commonly used trigger is laser, which is the fastest method to trigger reactions. A laser trigger is suitable for photosensitive molecules that undergo conformational changes after receiving photons. If a reaction is activated by certain chemicals instead of photons, we can trigger the reaction by applying laser pulses and releasing the caged compounds. With the advancement of Laue X-ray diffraction, time-resolved experiments only take a small number of crystal orientations to complete the data set, thus significantly increasing the data collection efficiency [31]. Using this method, a few X-ray diffraction patterns can generate data sets with microsecond to second resolution. Consequently, the method can provide both structural and dynamic information of a molecule. It can monitor the amplitude of a motion as well as the timescale of the motion.

In order for X-ray crystallography to work, the molecules in the crystal have to be very homogeneous. Otherwise, high-quality diffraction patterns can not be obtained. However, molecules are allowed to fluctuate to some extent in the crystal, thus creating a conformational distribution around the average structure. The heterogeneity created by the fluctuations leads to atomic displacement. The mean square atomic displacement is commonly known as the B factor or the temperature factor. As a result, B factors contain information about local dynamics [32]. However, B factors cannot be interpreted directly as the amplitude of local fluctuations because both intramolecular motions and lattice defects contribute to them. Nonetheless, super high resolution structures have been obtained in the absence of lattice defects. In these structures, B factors reflect the structural heterogeneity of the molecules. Although B factors do not contain information on the timescales of the structural fluctuations, they provide rich information about the structural changes in atomic level. In addition, anisotropic B factors can be modeled to represent non-uniform motions. The calculated anisotropic B factors provide information about both amplitude and directionality of motions. One caveat of the method is that it focuses on the local fluctuations and does not provide correlations between the motions.

1.2.3 Small angle X-ray scattering (SAXS)

Unlike the other X-ray techniques, small angle X-ray scattering (SAXS) does not require crystals to perform experiments. Macromolecules with length scales between 1 and 100 nanometers can be observed in the solution state. In a scattering experiment, macromolecules in the solution are subjected to an X-ray beam. The incident X-rays can be absorbed by a shell electron. The excited electron will emit a pulse in the same frequency as the incident X-ray, resulting in scattering. The intensity of the scattered radiation is recorded as a function of the angle between the incident radiation and the scattered radiation. Because molecules tumble isotropically in the solution, information obtained by SAXS experiments are orientation averaged. The scattering works as a ruler on the atomic level. By varying the angle between the incident radiation and the scattered radiation, we control the length of the ruler. The intensity of the scattered radiation is proportional to the number of atom pairs whose interatomic distance matches the ruler. When the angle is small, we observe averaged information on a large spatial scale. When the angle is large, we observe averaged information on a small spatial scale. The scattering intensity profile can be Fourier transformed. As it turns out, the transformed function represents the distance distribution function within the macromolecule. Although the distance distribution function does not inform on timescales, the direct observation of a distribution instead of an average provides us rich information about biomolecular dynamics. However, the SAXS observations are limited in certain ways. First, the information is only about the distance distribution. SAXS observations do not inform on the orientation distribution or the joint distribution. Although the distance distribution is a major component of the conformational distribution, it does not deliver the whole picture of biomolecular dynamics. Second, the distance distribution determined from SAXS experiments is the distribution of distances between all atom pairs. The convolution between a numerous number of atom pairs makes the distance distribution less informative and more degenerate. Labeling certain sites with heavy atoms may deconvolve the contributions. The scattering of the heavy atoms are strong compared to others. Consequently, the scattering profile of the two labels sites is observed and the distance distribution is determined. It should be noted that engineering heavy atoms into the biomolecule may perturb the system.

1.2.4 Single molecule fluorescence resonance energy transfer (smFRET)

Single molecule fluorescence resonance energy transfer (smFRET) is an implementation of the FRET technique. In smFRET, a single donor is excited and the acceptor within the molecule emits light after the energy transfer. As in the FRET technique, the distance between the donor and the acceptor can be calculated from the emission intensity. With only one donor-acceptor pair, the change in distance represents the conformation change in a single molecule, instead of an average for a collection of molecules. The time resolution of smFRET is on the millisecond timescale and the spatial resolution is on the order of one nanometer. The smFRET method can observe the distance between two labeled sites in real time. As a result, the technique provides valuable dynamic information.

1.3 Biomolecular Dynamics in Biology

Biomolecule with dynamics are selected by evolution because of several benefits offered by dynamics. First, dynamics provide structural plasticity to biomolecules. When a flexible system and a static system are compared in a binding reaction, the flexible system has to pay the entropy penalty in most cases because the flexible system is locked in a unique conformation in the bound state. The transition from a widely distributed conformational ensemble to a unique conformation significantly reduces the number of possible states and thus it is entropic unfavorable. However, a flexible system with significant dynamics may have more benefits to compensate for the entropic cost. The structural plasticity provided by molecular dynamics enables steric advantages. The binding process may involve an induced-fit step where multiple conformations are required. In this case, the transition between the conformations provides the access to reach a good complimentarity of the interface which can not be achieved by any one of the conformations. As a result, the dynamics of flexible molecules may increase the binding specificity. Second, dynamics can also offer evolutionary advantages. Evolution requires the biomolecule to switch from one function profile to a different function profile. The two function profiles can be so different that the transition from one profile to the other can not be achieved by one or two single mutations. When the transition is hardly possible, dynamics may enable the biomolecule to stop at a middle step by sampling both conformations for the two function profiles. In this way, evolution can happen in incremental steps rather than a hardly impossible leap. Last but not least, molecular dynamics can offer the structural fluctuations needed for catalysis. Catalysis requires the enzyme to go through multiple conformations when the substrate binds. Dynamics can prepare the needed conformations in the sense that the conformations have low state energies and high populations in the absence of the substrate. Consequently, dynamics are exploited by catalytic turnover.

Because of the above benefits, molecules with dynamics have been discovered to perform different functions, including enzyme catalysis, molecular recognition and signal transduction. Strictly speaking, signal transduction may involve biomolecules with catalytic or binding activity. As a result, signal transduction is not exclusive to enzyme catalysis and molecule recognition. Here, I review dynamic molecules with catalytic activities and binding activities. Some of the molecules are also important for signal transduction.

1.3.1 Enzyme catalysis

Enzyme catalysis is in no way a static process. Catalysis turns reactant molecules into product molecules, which is usually accompanied by conformational changes in the catalysis machines: enzymes. The conformational changes could either be induced by the binding or be an intrinsic property of the enzyme molecule. For the conformational changes of most enzyme, we probably don't have a yes or no answer. The conformational changes could be both induced and intrinsic. In the following discussion, I first show that distinct structures of enzymes are observed in a catalytic reaction. Next, I give several examples where the conformational dynamics is an intrinsic property encoded in the energy landscape of the enzymes. The studies do not rule out local induced-fits in the reactions, but they demonstrate dynamics can significantly contribute to catalytic reactions.

Cytochrome P450 superfamily contains hemoprotein members which catalyze the addition of oxygen molecules to hydrocarbons. The detailed mechanism of P450cam in the hydroxylation reaction of camphor was studied by X-ray crystallography with trapping techniques [33]. The primary obstacle for this study was the short half-time of the dioxy intermediate. To overcome the obstacle, conditions for simultaneously initiating reactions were found. By using the conditions, intermediates could be generated and observed in a high occupancy. A series of three intermediates in the bound complex were trapped and viewed as snapshots of the reactions. The determined structures show significant conformational changes in several residues and reveal the structural mechanism of the reaction.

Cyclophilin A (CypA) is a peptidylprolyl isomerase which belongs to the family of cyclophilins. The catalysis process switches between the cis and trans conformations of prolyl peptide bonds. Both cis and trans conformations of prolyl peptide bonds bound to CypA have been observed in the complex determined by X-ray crystallography. NMR relaxation experiments filled in detailed information about the dynamics involved in the catalysis reaction [12, 14]. R_2 rates measured by spin relaxation experiments show an increase in the region involved in catalysis. The increase is contributed from the R_{ex} component in R_2. The increase of R_{ex} could be due to either the binding step or the isomerization step in the catalysis reaction. The effects from the two steps was separated by increasing the substrate concentration and observing the response in the change of R_2. It turns out that the exchange dynamics observed for most residues are resulted from the binding step and only a few are resulted from the isomerization step. In addition, the rate constants for the conformational changes are in agreement with the rate constants of substrate interconversion, indicating the exchange dynamics may contribute to the substrate turnover. The k_{ex} rate constants for all residues were measured by relaxation dispersion. Besides a loop region, the k_{ex} rate constants for other residues are essentially the same, suggesting the existence of a collective motion within a dynamic network. The motion exists in the free enzyme without the substrate and thus the motion is an intrinsic feature of the enzyme. The collective dynamics contributing to the catalysis function could be a common feature for enzyme and may have evolved as an interplay between structure, dynamic and function.

Adenylate kinase is a phosphotranferase enzyme that catalyzes the reaction from one ATP moelcule and one AMP molecule to two ADP molecules. X-ray structures of free and substrate-bound adenylate kinase have been determined. Two nucleotide lids are observed in the structures, one for the binding of AMP and one for the binding of ATP. In the free state, the two lids are in an open conformation. In the bound state, the two lids are in a closed conformation. The closed conformation is required to exclude water from the active site and to position the substrates for phosphotransfer. The opening and closing of the two lids is the rate limiting step for the catalytic turnover observed by experiments [13, 15]. Crystallography studies captured three distinct open conformations which are in the trajectory leading to the closed conformation. The timescale for interconversion between the conformations is in the order of nanoseconds suggested by molecular dynamics (MD) simulations. However, NMR relaxation experiments observed collective motion within the molecule in milliseconds. The millisecond time scale motion was also confirmed by single-molecule FRET experiments. In addition, the distance between the two lids was estimated by NMR paramagnetic relaxation enhancement (PRE). Significant line broadening was observed for nuclei which is only close to the spin label in the closed conformation. The result suggests that the closed conformation is sampled by the dynamics of the substrate-free enzyme. It also indicates that the observed millisecond timescale motion corresponds to the open-close conformational change. Although the nanosecond timescale motion suggested by MD seems to contradicts the millisecond timescale motion observed by experiments, the two kinds of motions can be merged into a big picture with dynamics in multiple levels . The dynamics of adenylate kinase has a hierarchical structure of motions in multiple timescales. The collective motion in the millisecond timescale is driven by the fast timescale motions in nanoseconds.

Dihydrofolate reductase (DHFR) catalyzes the reduction of dihydrofolate (DHF) to tetrahydrofolate (THF) using the cofactor reduced nicotinamide adenine dinucleotide phosphate (NADPH) as electron donor. To complete the catalytic cycle, the enzyme DHFR has to go through five intermediate states, when the association and disassociation of either the substrate DHF, the product THF or the cofactor NADPH occur. The structures of the five intermediate states have all been determined by X-ray crystallography. Although the structures probably represents the lowest energy conformation of each state, they are not accurate representations of the five intermediate states as suggested by NMR relaxation dispersion experiments [16, 17]. Chemical exchanges were observed for the substrate binding region in the absence of the substrate. The observed exchanges, along with the calculated chemical shifts of the other conformational state, indicates the substrate-bound conformation is already sampled in the substrate-free states. Similarly, chemical exchanges were also observed for the cofactor binding region in cofactor-free states. Based on the result, each state not only has a significant population of the lowest energy conformation, it also contains observable populations of the preceding and following conformations. Here, dynamics-driven fluctuations populates minor conformations and prepares the enzyme for the next catalytic step. All the conformational exchanges occur on the timescale of microseconds to milliseconds, which is observable for the relaxation dispersion technique. The observed interconversion rate constants between the lowest energy state and the low energy state in equilibrium also agree with the kinetic rate constants between the two states. As a result, the binding and release of the substrate and the cofactor do not affect the energy barriers between the states significantly. The preferred kinetic path to the next state is already encoded in the energy landscape of the current state. In summary, conformations of the enzyme are redistributed in each step and the energy landscape of the current state prepares the kinetic path to the next state, so the catalytic function of the enzyme DHFR is coupled to its dynamics.

1.3.2 Molecular recognition

Biomolecules with extensive functional plasticity serve as hubs to bind and accommodate various binding partners. The functional plasticity is usually correlated with structural flexibility, which enables hub molecules to form complementary binding interfaces to their binding partners. Upon binding to each binding partner, the molecule could go through an induced-fit scheme or a conformational selection scheme. Here, I review several hub molecules and their binding strategies.

Ubiquitin is a small protein that regulates other proteins by ubiquitination. In ubiquitination, ubiquitin molecules bind to other proteins and form isopeptide bonds with lysine residues. After ubiquitination, ubiquitin molecules and their binding targets are covalently linked to each other. The molecule can be attached to one or multiple ubiquitin molecule. The attachment of one ubiquitin molecule is called monoubiquitination and the attachment of multiple ubiquitin molecules is called polyubiquitination. When a molecule is ubiquitinated, its most common fate is degradation through proteasome. However, ubiquitination can also serve as a regulation mechanism to change the activity, location or binding profile of the target molecule. As a regulator, ubiquitin demonstrates intensive functional plasticity and it can bind to a large number of proteins. The functional plasticity of ubiquitin comes from its structural flexibility. Because of its biological importance, ubiquitin has been crystallized in complex with dozens of its binding partners. The structures of these protein complexes have been determined. Although most regions of the ubiquitin only have small structural fluctuations, a loop region and the C-terminal end of the ubiquitin molecule demonstrate intensive motions [18]. Because of these motions, ubiquitin can form protein-protein interfaces with good complimentarity. Motions in the loop region and the C-terminal region was directly observed by spin relaxation experiments and quantified by order parameters. However, spin relaxation experiments are only sensitive to motions in sub-nanosecond timescales. Slower motions are be excluded from the observations. Indeed, the magnitude of motion measured by spin relaxation experiments does not agree very well with the magnitude inferred from the observed structural heterogeneity. Residual dipolar couplings (RDCs) were measured from ubiquitin molecules in its free form [18]. The reconstructed structural ensemble of ubiquitin covers the structural heterogeneity over a wide range of timescales, from picoseconds to milliseconds. The magnitude of motions determined from RDCs is larger than that suggested by the order parameters, indicating there exists structural dynamics between nanoseconds and milliseconds. In addition, the structural heterogeneity reconstructed from RDCs also agrees with that observed in crystal structures. As a result, the binding of ubiquitin to its binding partners is primarily a conformational selection and thus its intrinsic dynamics contributes significantly to its binding profile. Interestingly, the observed motions is mainly resulted from one motional mode, indicating the motion is concerted. The biological advantage of concerted motion is the reduced entropy cost upon binding. The dynamics of ubiquitin may be highly optimized by evolution to increase its functional plasticity without losing much affinity.

Calmodulin is a calcium-binding messenger protein important for signal transduction. It is also a hub molecule capable of binding to various binding partners. A detailed review of calmodulin biology is included in section 1.4. Calmodulin has two domains, an N-terminal domain and a C-terminal domain. The two domain are similar in structure and sequence. Each domain has two calcium binding sites formed by two EF hand motif. There is a short linker connecting the two domains. When the two domains are free of calcium, the linker is flexible [34, 35]. When the two domains are saturated with calcium ions, the linker adopts a helical structure. In the presence of a binding partner, the helix unwound and the two domains close to accommodate the binding partner. Because calmodulin has ~~both~~ binding sites for calcium ions and its protein binding partners, it has more than two major conformational states. There are calcium-free state and calcium-saturated state for each domain. In addition to the two states, calcium-saturated calmodulin forms an open-form in the absence of protein binding partners and forms a closed-form when bound to the protein binding partners. The conformational switch induced by calcium binding in one domain is a process independent of the other [34, 35]. The helix has intrinsic dynamics observed by spin relaxation experiments [36, 37]. Order parameters in the interdomain helical region suggests a considerable amount of motion in this region. Single-molecule FRET experiments also show that the interdomain distance has a wide-range distribution. Furthermore, structure ensemble of calcium-saturated calmodulin in the open state is constructed based on RDCs and SAXS data [38, 39, 40]. A detailed description of the construction method is included in section ??. The structure ensemble includes both conformations close to the open state and conformations close to the close state, demonstrating extensive structural heterogeneity. As in the ubiquitin case, binding partners of calmodulin selects preexisted conformations from the open-state conformational distribution. In addition to the interdomain dynamics, sidechain dynamics has also been altered upon binding to a protein binding partner. The entropy cost of the binding was estimated using sidechain order parameters as proxies. It should be noted that the entropy cost of the binding reaction includes multiple components, such as entropy contributions from solvent, ligand, backbone and sidechain. The proxy used in this method only estimates one component of the total change. In addition, correlation between sidechain motions are not accounted for in the order parameters. The entropy costs of a concerted motion and that of a set of independent motions is different. Nevertheless, the study indicated that the entropy is reduced upon binding, enacting a penalty on binding affinity. The penalty could a promise of evolution in order to increase the functional plasticity of calmodulin.

1.4 The Biology of Dynamic Biomolecules

1.4.1 Staphylococcal protein A (SpA)

Staphylococcus aureus was first described in abscess formation by Alexander Ogston in the late 19th century [41]. It is a Gram-positive spherical bacterium forming grape-like clusters. Carriers of S. aureus constitute 30% to 50% of healthy adult population and they are subjected to infections at a higher risk [42]. Syndromes caused by S. aureus range from minor infections such as boils and abscesses to severe diseases such as bacteremia, pneumonia, endocarditis, septic arthritis and osteomyelitis. Methicillin-resistant S. aureus (MRSA) have been populated and spread dramatically since 1970s

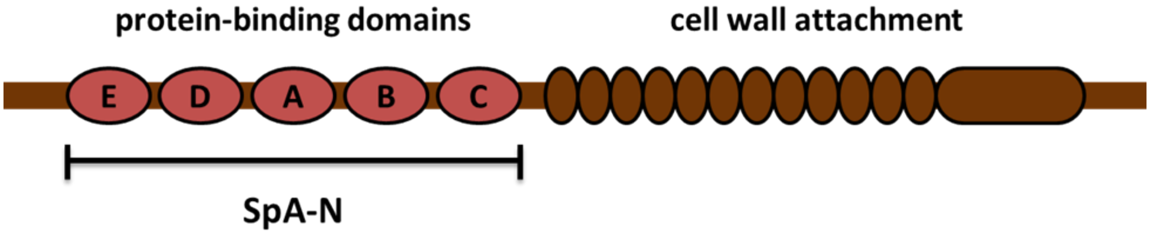


Figure 1.6: The domain sequence of Staphylococcal protein A.

in the United States, which has become a major public health concern [43]. More than 90% of either non-invasive or invasive S. aureus strains contain a virulent factor called Staphylococcal protein A (SpA) [44]. SpA contributes to biofilm formation, arthritis and septic death, leading to the Staphylococcal pathogenesis [45, 46]. SpA is a component of the cell wall [47, 48]. As shown in Fig. 1.6, full length SpA has 509 amino acids, including a signal peptide, an antibody binding region, a region X and a LPXTG motif arranged from N-terminus to C-terminus [49]. The signal peptide contains 35 amino acids. Upon the secretion of SpA across the cytoplasmic membrane, the signal peptide is removed by signal peptidase. After SpA has been secreted onto the membrane, the LPXTG motif is recognized and then cleaved by sortase [47]. The cleavage occurs between the threonine and the glycine in the LPETG sequence. Then, the N-terminal part of the cleavage product is covalently linked to the cell wall through a pentaglycine linkage. The region X contains two parts: a sequence-repetitive region Xr, followed by a non-repetitive region Xc. The sequence-repetitive region is composed of 12 repetitive octapeptides. The region X is thought to span the peptidoglycan layer of the cell wall and thus expose the antibody binding region of SpA [50].

The antibody binding region is named as SpA-N, the N-terminal half of Staphylococcal protein A. It contains five domains: E, D, A, B and C domains, which share a high sequence identity [49]. SpA-N can interact with multiple host proteins of the immune system, contributing most to the virulence of SpA [51, 52, 53, 54, 55, 56, 57]. Among the binding partners of SpA-N, two are of critical importance, the Fc region of antibodies and the tumor necrosis factor receptor 1(TNFR1). The binding between SpA and IgG Fc can block the C1q binding site, thus inhibiting the classical pathway of complement activation.(11) The complement system is a crucial part of the innate immunity. It can amplify the complement activation, clear pathogen cells by phagocytosis and induce cell lysis of pathogenic bacteria. The inhibition of the complement activation disables an important part of the innate immunity and contributes to the invasion of S. aureus [51]. The other one is the tumor necrosis factor receptor 1(TNFR1). SpA-N can induce TNFR1 signaling in epithelial cells and cause pneumonia [53, 54]. The SpA-N/TNFR1 binding recruits polymorphonuclear leukocytes. Although the recruitment of neutrophils is necessary to remove S. aureus, excessive inflammation in lung can block the airway and cause pneumonia. As SpA-N contributes to pathogenesis of S. aureus infections by interacting with multiple host proteins, studying the interactions between SpA-N and its binding partners is a key step to understand the pathogenesis of S. aureus infections and develop strategies to cure staphylococcal diseases.

1.4.2 Calmodulin

Calmodulin, as its name suggests, is one of the most important proteins to transduce calcium signals. Calmodulin molecules are widely distributed within a cell and with a micromolar concentration. Calmodulin has many binding partners, including transcription regulators, kinases and phosphatases [58, 59]. Many binding partners of calmodulin can not bind calcium themselves, so calcium bridge the gap between the binding partners and calcium ions by transducing calcium signals. Calmodulin belongs to the large family of EF-hand proteins. Calmodulin has two domains, each of them has two EF-hand motifs. The EF-hand motif is a special helix-loop-helix which has the ability to bind a calcium ion in the loop. The four calcium binding sites has different binding affinities to calcium ions. The two binding sites at the C-terminal domain have a binding affinity several folds higher than the binding affinity of the N-terminal sites [35]. Both binding affinities are on the order of micormolar, which matches the intracellular concentration of calcium ions. As a result, the binding profile of calmodulin makes it a perfect calcium sensor to respond to the fluctuations in calcium concentration. Although the N-terminal and the C-terminal domains in calmodulin are sequentially and structurally similar, their conformations in the absence of calcium ions are slightly different. In the absence of calcium ions, the N-terminal domain has a tightly packed conformation while the C-terminal domain partially expose a hydrophobic patch. The binding of calcium ions makes calmodulin adopt the open conformation, resulting in exposure of hydrophobic regions as well as the rigidification of the interdomain linker (Fig. 1.7). When the calcium-saturated calmodulin binds to its binding partners, the two domains come from the two ends and surround the binding partner in the middle. The helix connecting the two domains in the calcium-saturated conformation unwinds and becomes flexible linker. Calmodulin can also bind to some of its binding partners in the absence of calcium ions. The binding can raise the binding affinity of calmodulin to calcium ions.

1.4.3 Trans-activation response element (TAR)

The trans-activation response element (TAR) RNA comes from transcriptions of the long terminal repeat(LTR) region of human immunodeficiency virus type-1 (HIV-1). TAR is the binding target of Tat, which is a viral regulatory proteins important for the replication of HIV. The basal transcription of the LTR region produces spliced

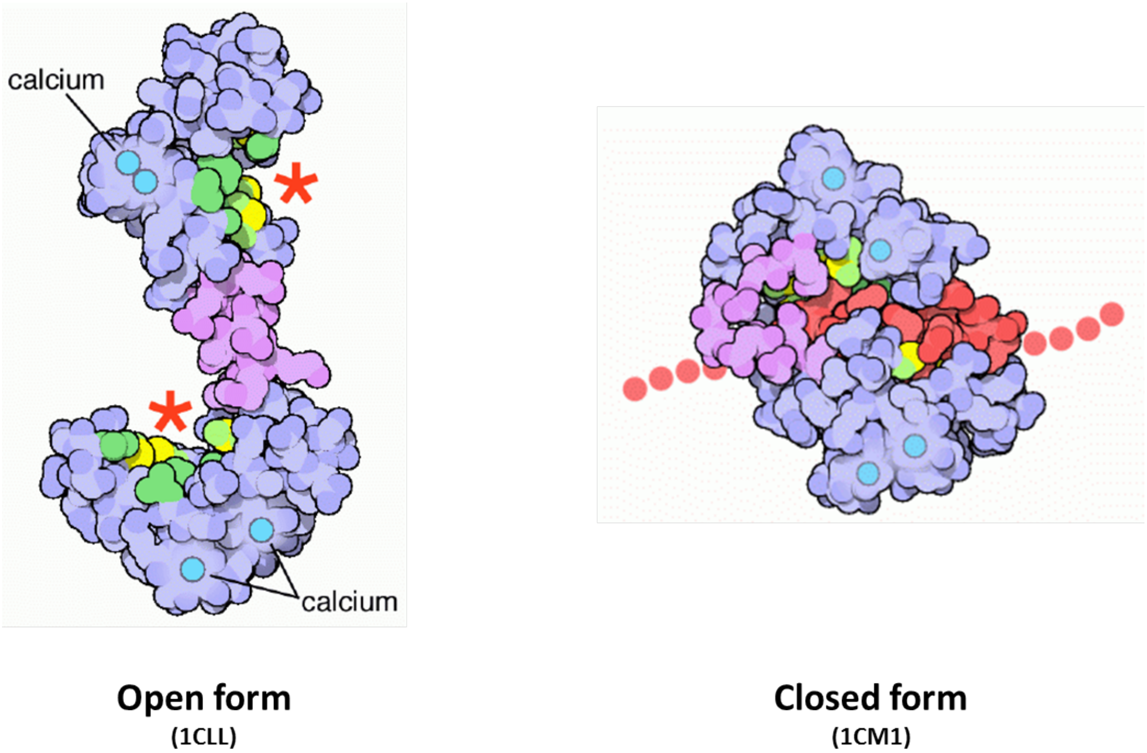


FIGURE 1.7: The open form and closed form of calmodulin.

mRNA product and leads to the production the viral regulatory protein Tat. Tat further stimulates the transcription by increasing the elongation efficiency. Without Tat, the HIV promoter still initiates the transcription at a high rate, but most transcripts end up incomplete. Tat facilitates the elongation step by binding to TAR RNA. The binding of Tat to TAR is essential for Tat function [60]. TAR is the transcription product of the LTR region. It forms a stable RNA hairpin at the 5' end of nascent transcripts of LTR [61]. The stable RNA hairpin formed by TAR recruits the Tat protein and makes Tat a gene specific elongation factor.

Tat protein is the major binding partner of TAR RNA. The RNA-binding domain

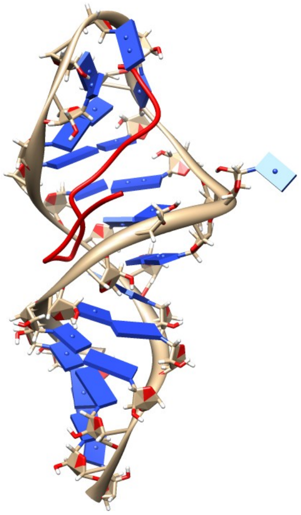


Figure 1.8: TAR RNA in complex with a peptide derivative of viral regulatory protein Tat. TAR RNA is shown with tan backbone and blue bases. The peptide derivative is shown as red.

of Tat is a flexible arginine-rich region [60]. The arginines and the positive charges are crucial for the binding. Even an arginine molecule as a free amino acid binds specifically to TAR RNA. The continuous A-form helix of TAR is interrupted by two bulge nucleotide (U10 and U12) [62]. U12 is disordered and does not involved in the binding. U10 is positioned in the major groove and is essential for the binding. As expected, U12 can be deleted with little effect on binding but the deletion U10 decreases the binding affinity dramatically. The two bulge nucleotides opens the major groove for binding. The RNA-binding domain of Tat forms a β-turn conformation and sits in the opened major groove.

2

Characterizing the flexibility of SpA-N using NMR spin relaxation

2.1 Introduction to NMR spin relaxation

2.1.1 Longitudinal and transverse relaxation

Particles with $\frac{1}{2}$-spin have two eigenstates, $|\alpha\rangle$ and $|\beta\rangle$. The two states are degenerate in the absence of an external magnetic field, but they have different energies in the presence of an external magnetic field. Nuclei observed by solution NMR usually have $\frac{1}{2}$-spin. Under the strong magnetic field generated by the NMR instrument, the population of the low energy eigenstate $|\alpha\rangle$ is slightly larger than the population of the other state $|\beta\rangle$. It should be noted that even a single nucleus can have both eigenstates with different proportions. Due to the population difference, a net magnetization along the magnetic field direction is generated. The direction is called the longitudinal direction. The plane perpendicular to the longitudinal direction is called the transverse plane. Most NMR experiments are based on manipulating and observing the net magnetization. When the magnetization is rotated to other directions, the state with the rotated magnetization has a higher energy than the ground state and it eventually relaxes to the ground state, which has the magnetization in the longitudinal direction. NMR spin relaxation experiments focuses on such relaxation process. When the magnetization is rotated to the transverse plane, the transverse magnetization decreases and the longitudinal magnetization increases over time. It turns out that the decrease in transverse magnetization is different from the increase in longitudinal magnetization. The process of recovering the longitudinal magnetization is the longitudinal relaxation and the process of losing the transverse magnetization is called the transverse relaxation.

The longitudinal relaxation is usually measured by an inversion recovery experiment, which observes the magnitude decrease of an inverted longitudinal magnetization (Fig. 2.1). The transverse relaxation is usually measured by observing the decay in the transverse plane (Fig. 2.2). Either the inverted longitudinal magnetization or the transverse magnetization has an exponential decay curve, which can be described by a rate constant R. The rate constants for the longitudinal and transverse relaxations are called R_1 and R_2 respectively. The transverse relaxation rate R_2 is always larger or equal to the longitude relaxation rate R_1. A portion of the decrease in transverse magnetization is due to the loss of coherence. The magnetization in the transverse plane is an ensemble sum. If each magnetization member in the ensemble is not in perfect pace with others. The scatter of magnetization in the transverse plane leads to the decrease in the total transverse magnetization. Although the total transverse magnetization decreases, the loss of coherence is an entropy driven process and the loss of coherence does not put the transverse magnetizations in ~~energy~~ more favorable orientations. The loss of coherence can be resulted from multiple factors, including inhomogeneity in the magnetic field and the motions of molecule. Usually, spin relaxation experiments are used to probe the molecular motions, so we would like to exclude the effects from field inhomogeneity. The most common strategy is to apply spin echo pulses when measuring the transverse relaxation rate

R_2. Spin echo pulses essentially invert the magnetizations on the transverse plane along a certain direction in the transverse plane. The spin echo pulses are applied in the middle point of a measurement. Assuming the rotation speeds of the transverse magnetizations remain the same before and after the inversion, magnetizations with different rotation speeds converge at the end. Because the field inhomogeneity does not vary during the experiment, the loss of coherence due to the inhomogeneity can be recovered.

2.1.2 Mechanism of relaxation

In a perfect static magnetic environment, the high energy state $|\beta\rangle$ does not fall back to the low energy state $|\alpha\rangle$ spontaneously. In order to make relaxation happen, there must be some mechanisms for energy transfers. In solution NMR, the main way for a nuclei to exchange energy is through fluctuations in its surrounding magnetic field. When the fluctuations have a frequency near the resonance frequency of a nucleus, there is a probability for the energy transfer to occur. In the absence of electromagnetic pulses, the external magnetic field is usually considered as static and homogeneous, so there is no fluctuations in the external field. However, the external field can induce small local magnetic fields. In the surrounding of a nuclei, there are other nuclei and electrons. Because of the diamagnetism of electron orbiting and the paramagnetism of nuclear spin and electron spin, the magnetic field felt by the nuclei is a sum of the external field and the induced magnetic fields. Although the external field is static, the induced local fields are fluctuating all the time because of motions on varying spatial and timescales. In the following, I discuss several mechanisms contributing to the constant fluctuations in the local magnetic field.

Because nuclear spins are paramagnetic, nuclear magnetic moments are generated under a magnetic field. The magnetic moment is mainly a magnetic dipole moment. The magnetic dipole alters the local magnetic field. The effect of the magnetic dipole

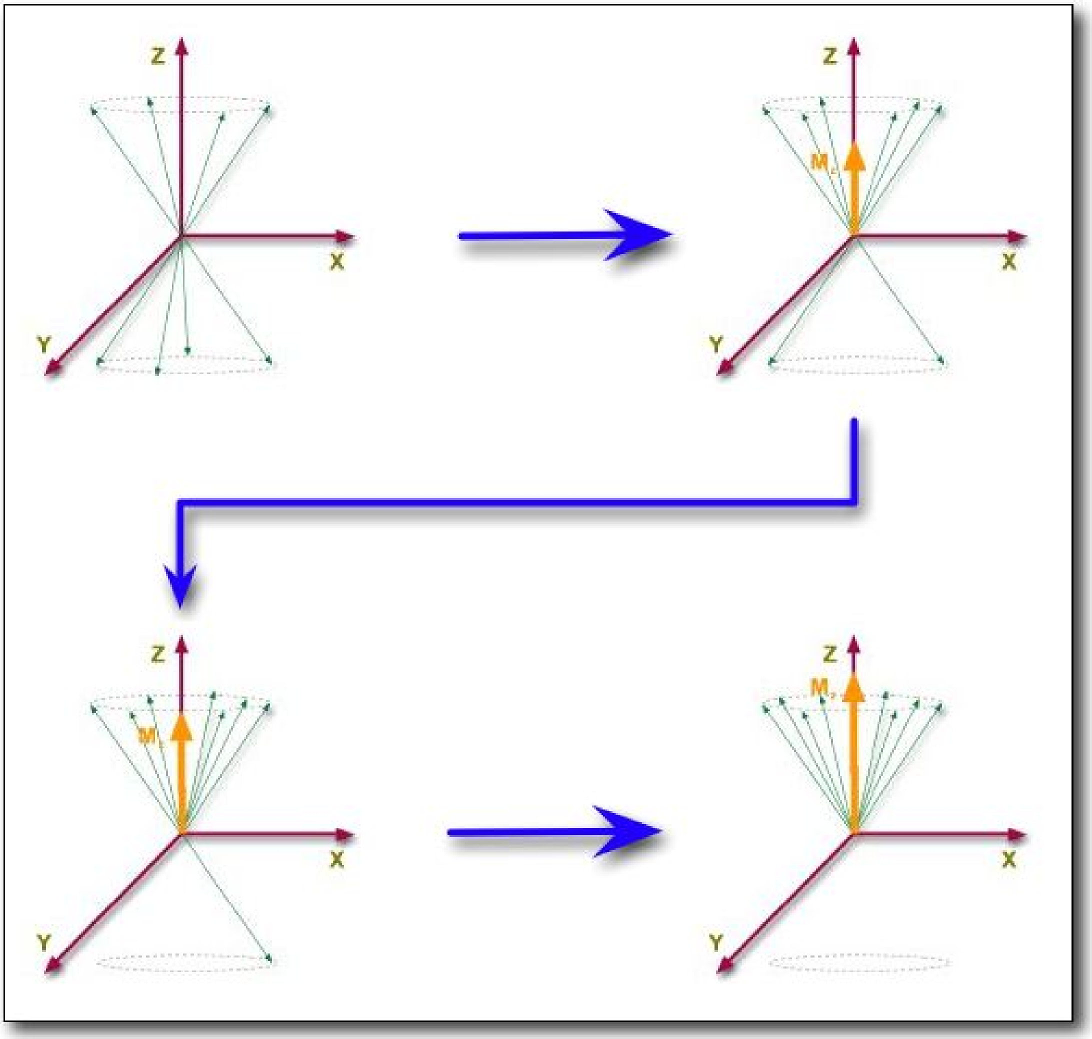


FIGURE 2.1: Longitudinal relaxation. In the longitudinal relaxation, the magnetization along the z-aixs M_z increases.

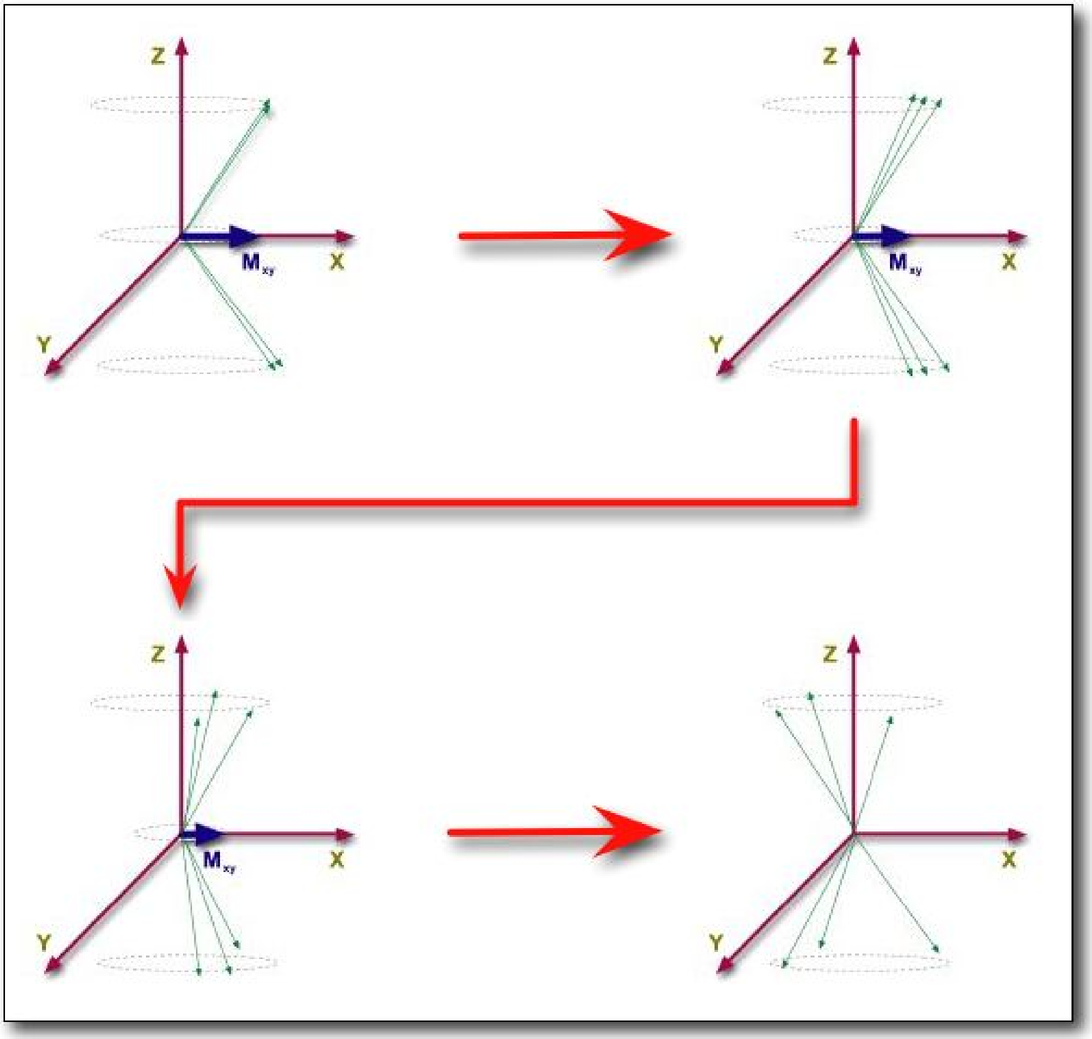


FIGURE 2.2: Dipole-dipole coupling between two adjacent nuclei. In the transverse relaxation, the magnetization on the xy-plane M_{xy} decreases.

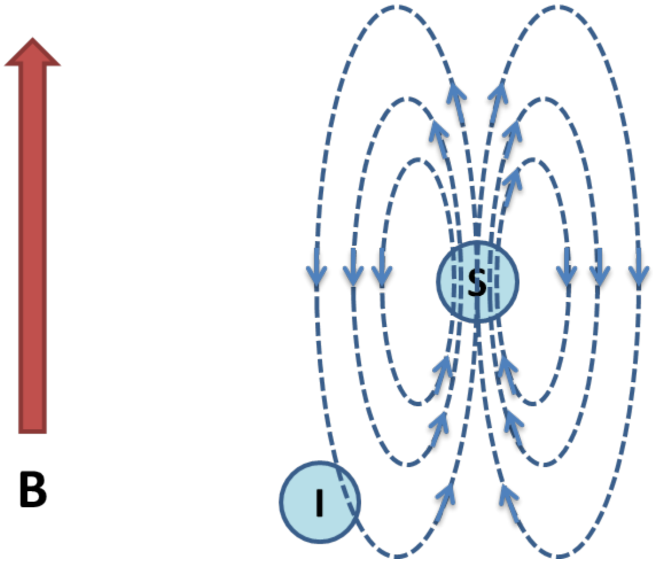


FIGURE 2.3: Dipole-dipole coupling between two adjacent nuclei.

of one nucleus on its adjacent nuclei is called the dipolar coupling (Fig. 2.3). Because the magnetic dipole is induce by the external magnetic field, the magnetic dipole is always parallel to the external magnetic field. However, there could be relative translational and orientational motions between the nucleus and its adjacent ones. Consequently, the local magnetic field felt by the adjacent nuclei varies when they move around the nuclei who generates the magnetic dipole. It should be noted that the motion around the nuclei is isotropic in solution and the average of the dipole coupling is zero in most cases. Nonetheless, fluctuations in the local magnetic field is real and the fluctuations contribute to relaxation.

Despite the surrounding nuclei, the surrounding electrons also contribute to the relaxation of a nuclear spin. Electrons has their intrinsic spins and they also circulate in their orbits. As nuclear spins, electron spins are also paramagnetic. The paramagnetism of electron spins is even much larger than the paramagnetism of the nuclear spins. However, in most cases, the effect of electron spins are ignored because paired electrons cancel out the their effects resulted from spins. A discussion of special cases when unpaired electron exists is included in subsection 3.1.4. The orbiting movement of electrons is diamagnetic. In the presence of an external magnetic field, the electron orbital generates a magnetic field against the external magnetic field. Nuclei affected by this local magnetic field experience a total magnetic field with a smaller magnitude. The decrease in the effective magnetic field is called shielding. Because magnetic induction lines are close loops, the diamagnetism of electron orbiting can also increase the effective magnetic field in certain part of the space and cause deshielding. Because of the shielding and deshielding. nuclei in different places have different effective magnetic fields. Even for the nuclei of the same type, their resonance frequencies are different due to the variations in their effective magnetic fields. The electromagnetic environment, or the chemical environment defines the chemical shift. In addition, the electron density around a nucleus can be anisotropic. As a result, when the molecule rotates, the total shielding effects on the nucleus varies and so does the local magnetic field. The variation in surrounding electron density leads to chemical shift anisotropy, which contributes to the local fluctuations of magnetic field and spin relaxation.

Chemical shift anisotropy contributes to the fluctuations of magnetic field in the same chemical environment. There is another mechanism can directly change the chemical environment and contribute to magnetic field fluctuations as well. When the molecule undergoes conformational changes, the chemical environment of a nucleus can change significantly. The process when a nucleus exchanges between two or more chemical environment is called chemical exchange. If the chemical exchange occurs much faster than the chemical shift difference, the nucleus experience an effective average of the two chemical environments. If the chemical exchange occurs much slower than the chemical shift difference, the slow exchange does not contribute much to fluctuations either. However, when the chemical exchange has a rate similar to the chemical shift difference, it contributes significantly to the loss of coherence. When a molecule has two or more distinct conformation states with medium exchange rates, chemical exchange leads to considerate increase in R_2.

2.1.3 Effects of motions on relaxation

The above three mechanisms contribute to magnetic field fluctuations and spin relaxations. In addition, spin relaxation rates depend on the fluctuation frequency, which is determined by motions on varying spatial and timescales. The motions can be roughly divided into three types, global tumbling, collective conformational exchange and local motions. Among them, collective conformational exchange usually occurs on the nanoseconds to milliseconds timescale. It drives chemical exchange and contributes mainly to the loss of coherence in the transverse relaxation. Global tumbling and local motions lead to both longitudinal and transverse relaxations. Global tumbling of a biomolecule is usually on the nanoseconds timescale and local motions are on the picoseconds to nanoseconds timescale. The global tumbling rate and the local motion frequency matches the resonance frequency of common used NMR nuclei. The resonance frequency is named as the Larmor frequency. For ^{1}H, ^{13}C and ^{15}N, the Larmor frequency is on the order of 100 MHz, which corresponds to the nanoseconds timescale. For the longitudinal relaxation, the relaxation rate R_1 heavily depends on the global tumbling rate. The inverse of the global tumbling rate is called the global correlation time τ_c, which defines the time it takes on average to rotate the molecule by one radian. When the global correlation time is on the order of 0.1 ns/radian, the global tumbling frequency matches the Larmor frequency and the longitudinal relaxation rate achieves its maximum. Below the threshold, the global tumbling is so fast that there is no significant transverse relaxation due to the loss of coherence. However, when the global correlation time is larger than the threshold, the longitudinal and transverse relaxation rates diverge.

The divergence is due to the dependence of the loss of coherence on random fluctuations. The angular deviations of transverse magnetizations from their mean can be treated as a random process (ref here: NMR primer). Consequently, the mean-square angular displacement can be written as:

$$\langle \theta^2 \rangle = l^2 N, \tag{2.1}$$

where l is the step size, which is proportional to the global correlation time τ_c. N is the number of steps, which is inversely proportional to the global correlation time τ_c. As a result, the mean-square angular displacement is proportional to the global correlation time τ_c. The mean-square angular displacement can be seen as a measure of the loss of coherence. So the portion of transverse relaxation due to the loss of coherence increases as the global correlation time increases. The schematic relationship between longitudinal/transverse relaxation and the global correlation time τ_c is plotted in the following Fig. **??**. (fig:R1/R2 vs tau-c)

2.1.4 ^{1}H-^{15}N heteronuclear relaxation experiments

~~Because~~ dipolar coupling and chemical shift anisotropy are two major mechanisms contributing to the relaxation rates, the relaxation behavior of spins surrounded by other dipoles or by an complicated electron environment could be very hard to predict. However, we could largely avoid the issue by performing ^{1}H-^{15}N heternuclear relaxation experiments on proteins. In these experiments, the ^{1}H-^{15}N spin pairs on the backbone do not have adjacent dipoles, so the spin pairs can be treated as isolated systems. Although dipolar coupling does occur between the pairs, the spins from different pairs are coupled over a large distance and their dipolar coupling can be ignored. In addition, backbone nitrogen atoms are all bonded to a carbonyl carbon, an alpha carbon and a hydrogen. The electron distributions around these nitrogen atoms are nearly the same. As a result, the dipolar coupling and the chemical shift anisotropy of the backbone nitrogens, as well as the relaxation behavior of these nitrogen atoms, can be easily modeled and predicted. Besides the ease of modeling, the ^{1}H-^{15}N spin pairs are also ~~nearly~~ uniformly distributed along the backbone. Except the first residue and proline residues, we have a ^{1}H-^{15}N spin pair for each residue. The uniformly distributed ^{1}H-^{15}N spin pairs probe the backbone dynamics of a protein molecule in a systematic way. At last but not least, the ^{1}H-^{15}N spectrum has a good dispersion. Consequently, the dynamics about most ^{1}H-^{15}N spin pairs can be extracted separately. The good dispersion reduces the degeneracy caused by signal overlapping.

As discussed previously, we could decompose the relaxation mechanisms into two components, dipolar coupling and chemical shift anisotropy~~. The statement is true~~ in the absence of chemical exchange. Consequently, the relaxation rates R_1 and R_2 can be calculated by summing up the dipolar coupling ~~contributed~~ component and the chemical shift anisotropy ~~contributed~~ component. Because the ^{1}H spin and the ^{15}N spin in the ^{1}H-^{15}N pair is coupled, there are four quantum states: $|\alpha_H\rangle|\alpha_N\rangle$, $|\alpha_H\rangle|\beta_N\rangle$, $|\beta_H\rangle|\alpha_N\rangle$ and $|\beta_H\rangle|\beta_N\rangle$. The transitions between the quantum states contribute to ~~the~~ relaxations. The efficiency of transitions is determined by the fraction of motions whose frequency matches the energy difference in the transitions. In other words, the rates of the transitions is proportional to the spectral density function evaluated at the corresponding frequency. The value of the spectral density function at certain frequency reflects the fraction of motions at this frequency. A more detailed discussion about spectral density function is included in the Lipari-Szabo model subsection. The longitudinal and transverse relaxation rates for the

^{15}N spin due to dipolar coupling can be calculated by:

$$R_1^{dd} = \frac{d^2}{10}(J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H + \omega_N)), \tag{2.2}$$

$$R_2^{dd} = \frac{d^2}{20}(4J(0) + J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H) + 6J(\omega_H + \omega_N)). \tag{2.3}$$

Here, d is the dipolar coupling constant:

$$d = \frac{\mu_0 \gamma_H \gamma_N \hbar}{4\pi^2 r_{HN}^3} \tag{2.4}$$

The presence of the $J(0)$ term reflects the contribution of slow motions to the loss of coherence.

In contrast to the dipolar coupling contributed relaxation, the ~~chemical shift anisotropy contributed~~ longitudinal relaxation only depends on the Larmor frequency of the observed spin. In our case, it is the Larmor frequency of the ^{15}N spin, ω_N. As in the dipolar coupling case, the transverse relaxation due to chemical shift anisotropy is also influenced by slow motions. As a result, the longitudinal and transverse relaxation rates for the the ^{15}N spin due to chemical shift can be calculated by:

$$R_1^{csa} = \frac{2c^2}{5} J(\omega_N), \tag{2.5}$$

$$R_2^{csa} = \frac{c^2}{15}(3J(\omega_N) + 4J(0)). \tag{2.6}$$

Here, c is a constant reflecting the magnitude of the anisotropy:

$$c = \frac{\omega_N \Delta\sigma}{\sqrt{3}}. \tag{2.7}$$

The $\Delta\sigma$ is ~~a constant~~ reflecting the anisotropy of electron distribution.

Although only longitudinal and transverse relaxations are discussed previously, the two experiments do not provide enough constraints to determine the spectral density function evaluated at five frequencies: 0, $J(\omega_N)$, $J(\omega_H - \omega_N)$, $J(\omega_H)$ and $J(\omega_H + \omega_N)$. In order to solve the problem, we first observe that $\omega_H - \omega_N$, ωH and $\omega_H + \omega_N$ have very similar values. So we could treat them as a single value in an approximation. In addition, the measurement of heteronuclear Overhauser effect(HetNOE) provides the third constraint we need after the approximation. The HetNOE is measured by irradiating the ^{1}H spin and observing the intensity of ^{15}N. As it turns out, the HetNOE intensity also depends on the motions. The ratio of the intensities of ^{15}N with and without pre-saturation in the ^{1}H spin is calculated by:

$$NOE = 1 + \frac{(d^2/10)(\gamma_H/\gamma_N)(6J(\omega_H + \omega_N) - J(\omega_H - \omega_N))}{(d^2/10)(J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H + \omega_N)) + (2c^2/5)J(\omega_N)} \quad (2.8)$$

2.1.5 Lipari-Szabo model

The Lipari-Szabo model takes minimal assumptions, the analysis method using the Lipari-Szabo model is usually called as a model-free method. However, the model assumes two kinds of motions in a biomolecule, the global tumbling and the internal motions. Although the internal motions may have more than one motional mode, the model treats the multiple modes as one and thus the method returns the average amplitude of the motional modes. This simplification can be seen more clearly following a mathematical treatment to the problem.

Although the motions of a biomolecule do not influence spin orientations, fluctuations of magnetic field are resulted from the change of orientations of bond vectors. In the ^{1}H-^{15}N spin relaxation experiment, the bond vector is the HN bond vector and its orientation change promotes the dipolar coupling and chemical shift anisotropy relaxation mechanisms. If we define the angle between the magnetic field and the bond vector as θ and plot the function of $f(t) = \cos\theta$ over time, we obtain a stochastic process driven by Brownian motion(fig). Although it is hard to see patterns in the $f(t) = \cos\theta$ function, the correlation between any two time points with a fixed

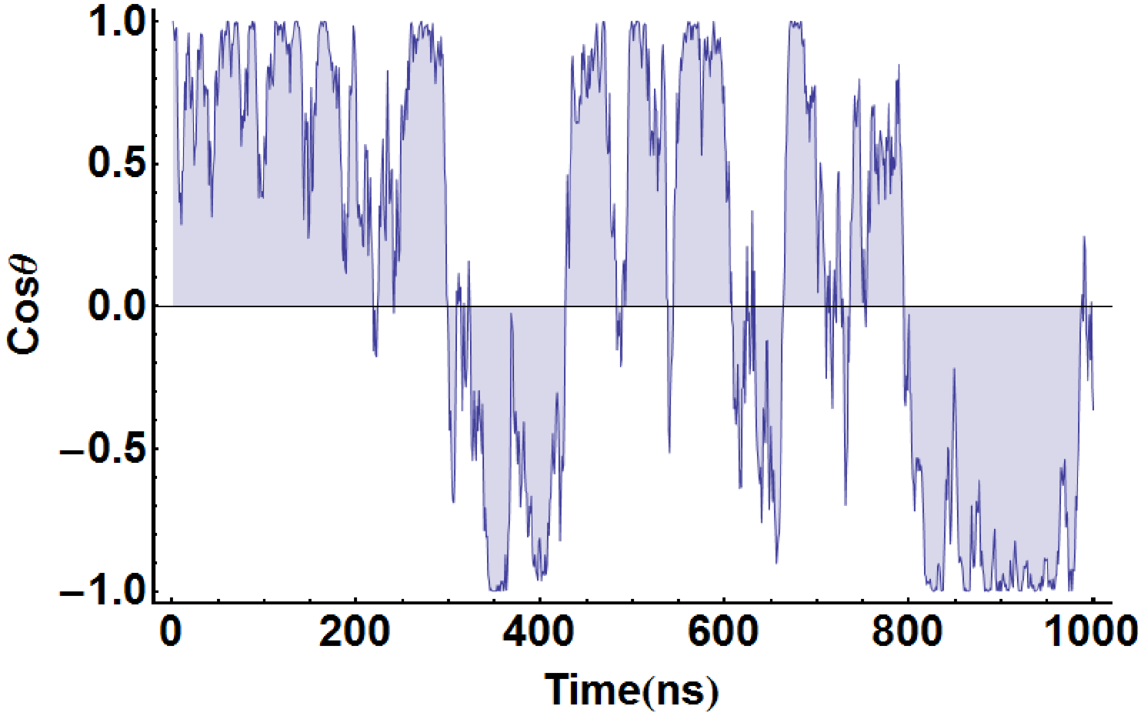


FIGURE 2.4: The function of $\cos\theta$ over time t in a stochastic process

interval can be analyzed. By varying the time interval, we obtain the autocorrelation function (Fig. 2.5):

$$g(\delta t) = \frac{\int f(t)f(t+\delta t)dt}{\int f(t)^2 dt}. \tag{2.9}$$

The autocorrelation function quantifies the timescale and intensity of the motions. If the motions are fast, the decay in correlation over time is also fast. If the motions are not restricted in any way, the autocorrelation time decays to zero. Otherwise, it reaches a plateau of a value in the range between 0 and 1. The value indicates how much restriction is imposed on the motions. The autocorrelation function usually

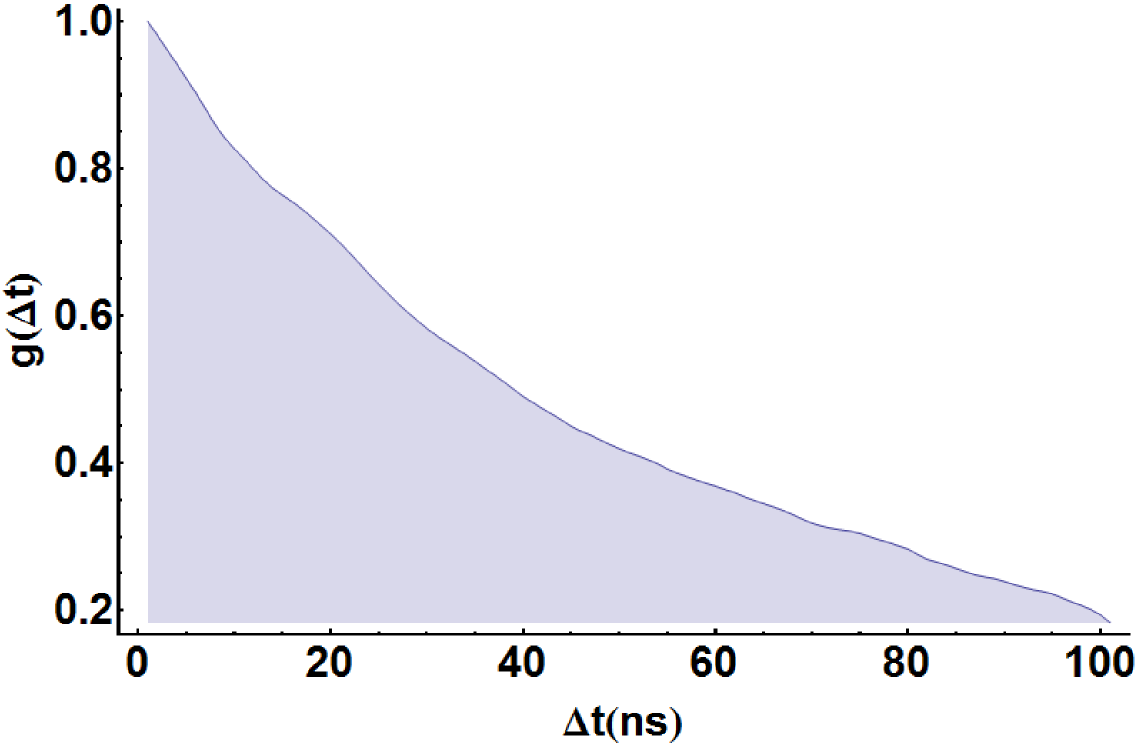


FIGURE 2.5: The autocorrelation function of the stochastic process.

takes the exponential form:

$$g(\delta t) = e^{-\delta t/\tau}, \tag{2.10}$$

where τ is the correlation time. The autocorrelation function is on the time domain and it can be Fourier transformed into the frequency domain. The Fourier transformed function is the spectral density function, which plots the contribution of motions at certain frequency to the overall motions. In reality, the spectral density function could be very complicated due to multiple motional modes. However, for most biomolecules, we can reasonably assume that only two major motional modes exist, the global tumbling motions and the internal motions. Besides the global tumbling, all other motional modes are considered as internal motions and summarized by one correlation time. As a result, we have the following autocorrelation function:

$$g(\delta t) = g_0(\delta t) g_I(\delta t), \tag{2.11}$$

where $g_0(\delta t) = e^{-\delta t/\tau_c}$ and $g_I(\delta t) = e^{-\delta t/\tau_e}$. τ_c is the global correlation time and the τ_e is the local correlation time. Although the model is an simplification, τ_e can be considered as the effective correlation time for an arbitrary correlation time and it provides a good intuition about internal motions even when multiple motional modes exist [63]. When the internal motions is restricted to certain degree, a generalized order parameter S^2 is used to specify the degree of restriction. As a result, we have the autocorrelation function in the following form:

$$\begin{aligned} g(t) &= g_0(t) g_I(t) \\ &= e^{-t/\tau_c}(S^2 + (1 - S^2)e^{-t/\tau_e}) \\ &= S^2 e^{-t/\tau_c} + (1 - S^2)e^{-t/\tau_c}e^{-t/\tau_e} \\ &= S^2 e^{-t/\tau_c} + (1 - S^2)e^{-t/\tau}, \end{aligned} \tag{2.12}$$

where $\tau = \frac{1}{1/\tau_c + 1/\tau_e}$. Although the order parameter S^2 is a good indication of the restriction on motions, it has limitations. The order parameter S^2 can vanish even if the motion is not isotropic. For the global motions, it should be noted that global tumbling can be anisotropic. When a molecule does not have a globular shape, the motions around different direction axes can have different rates. Depending on the shape of the molecule, its global tumbling motions may require up to three correlation times to describe [63]. The modeling and determination of the three correlation times is further discussed in the subsection 2.3.

There is a relationship between ~~the determined~~ spectral density values ~~and~~ ~~the~~ order parameter and ~~the~~ correlation times. By Fourier transforming Eq. 2.12, we have:

$$J(\omega) = S^2 \frac{\tau_c}{1+\omega^2\tau_c^2} + (1-S^2)\frac{\tau}{1+\omega^2\tau^2} \tag{2.13}$$

From previously determined spectral density function values at ω_N and $\omega_{0.87H}$, we could determine a global correlation time for all residues and an order parameter and an internal correlation time for each residues. However, a more intuitive way to visualize the result is through Lipari-Szabo mapping. We observe that the relationship in Eq. 2.12 is linear with respect to S^2. As a result, all $(J(\omega_N), J(\omega_{0.87H}))$ points determined for each residue on the line of the global tumbling point and its internal motion point (Fig. 2.6). The order parameter S^2 determines where the point is on the line. When S^2 is close to 1, its internal motions are very restricted and the point is close to the global tumbling point. When S^2 is close to 0, its internal motions are not restricted and the point is close to the internal motion point.

2.2 Measuring R1/R2 rates and NOE ratios of a SpA-N mimic

2.2.1 5B is a mimic of SpA-N

SpA-N has five tandem domains (E-D-A-B-C). The five domains share high sequence identity (Fig. 2.7). Each domain is a three-helix bundle. Their structures are also very similar. The high sequence identity and structural similarity make the spectral assignment of SpA-N very difficult. The residues at the same position of each domain and their associated spins experience only slightly different electromagnetic environment. Although the peaks on the ^{15}N-HSQC spectra are reasonably dispersed, the chemical shifts of C$_\alpha$, C$_\beta$ and CO are nearly the same for the residues at the same position of each domain. NMR assignment strategy links the previous and the following residue through matching the chemical shifts. When the chemical shifts are nearly equal, the assignment strategy fails. Fortunately, due to the high sequence identity and structural similarity, we can reasonably assume that the relaxation be-

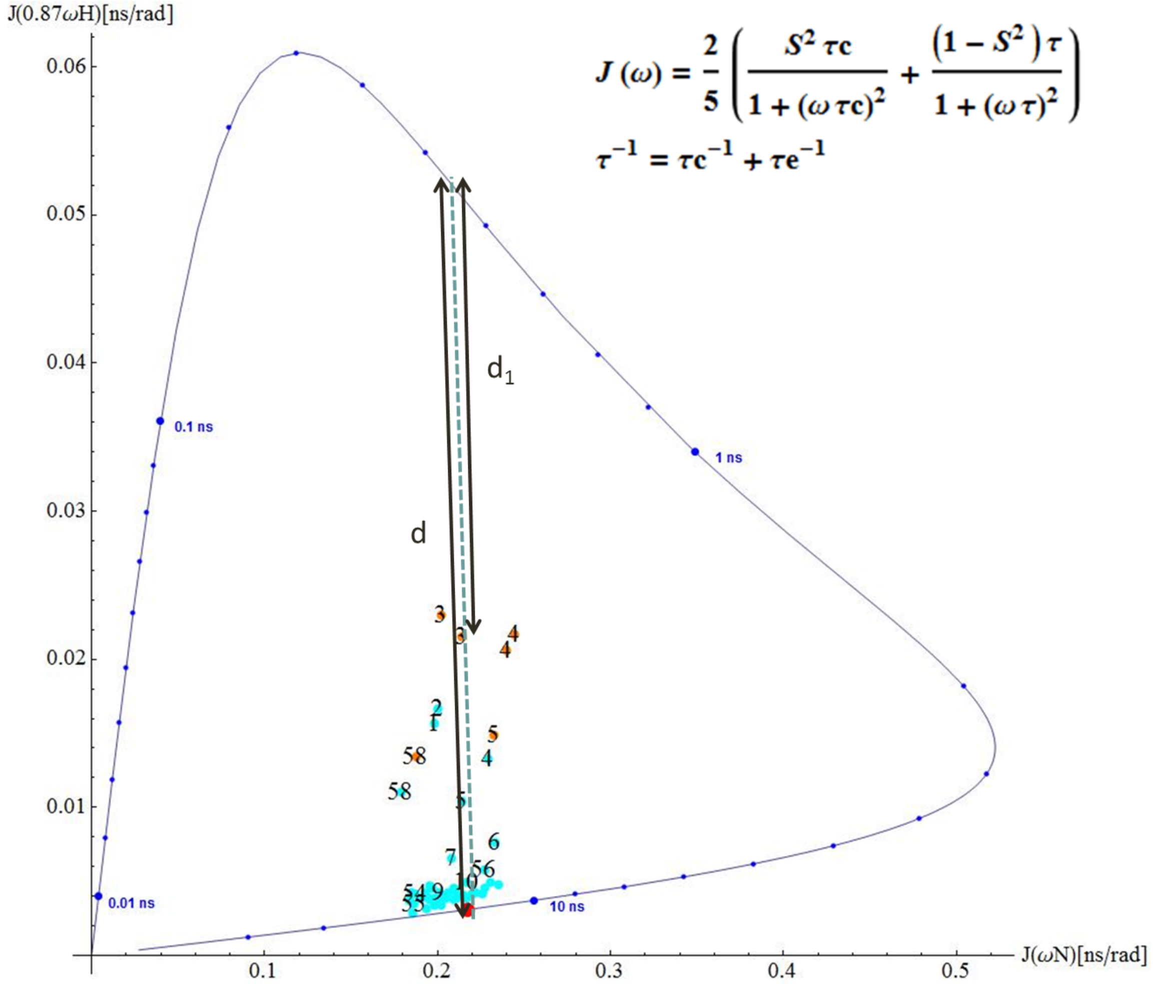


FIGURE 2.6: T graphical interpretation Lipari-Szabo mapping. In the Lipari-Szabo mapping, all $(J(\omega_N), J(\omega_{0.87H}))$ point on the line of the global tumbling point and its internal motion point. The order parameter S^2 determines where the point is on the line and it equals d_1/d.

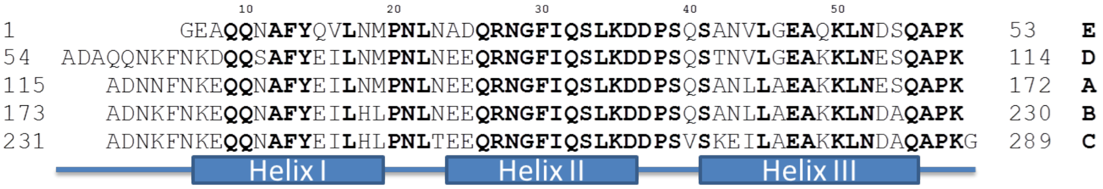

10
20
30
40
50
1 GEAQQNAFYQVLNMPNLNADQRNGFIQSLKDDPSQSANVLGEAQKLNDSQAPK 53 E
54 ADAQQNKFNKDQQSAFYEILNMPNLNEEQRNGFIQSLKDDPSQSTNVLGEAKKLNESQAPK 114 D
115 ADNNFNKEQQNAFYEILNMPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNESQAPK 172 A
173 ADNKFNKEQQNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNDAQAPK 230 B
231 ADNKFNKEQQNAFYEILHLPNLTEEQRNGFIQSLKDDPSVSKEILAEAKKLNDAQAPKG 289 C
Helix I
Helix II
Helix III

Figure 2.7: The high sequence identity among the five domains of SpA-N. The sequence of the five domains is listed. Sequence shared across all five domains are labeled as bold letters.

havior of the five domains are very similar. As a result, there is no need to distinguish the residues at the same position but from different domains. The averaged signal from the residues is a good representative of each one. Based on the idea, we replaced E/D/A/C domains with B domain and engineered a new construct called 5B, which has five B domains in a row (Fig. 2.8). The linker used in the 5B construct is also highly conserved. Because the sequence identity between domains is high (83 - 91%), this construct is a mimic of SpA-N structurally and functionally. Compared to SpA-N, the construct has a simpler ^{15}-HSQC spectrum. Except several residues in the N-terminus and in the C-terminus, all other residues at the same position of each domain should share nearly the same electromagnetic environment. Indeed, we observed the expected overlapping of HN peaks on a ^{15}N-HSQC spectrum (Fig. 2.9). The spectrum of 5B is nearly the same as the spectrum of a single B domain. For a single B domain, a total of 54 peaks are expected. Because we should also observe the first residue of each domain in the 5B spectrum, we only observed 6 extra peaks in addition to the expected 55 peaks.

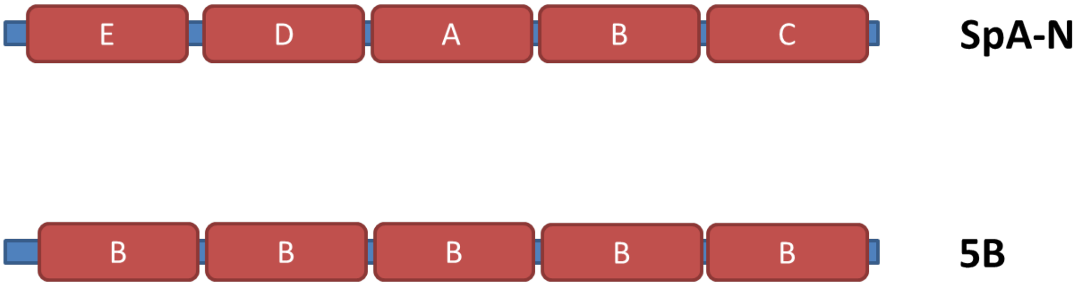


Figure 2.8: Schematic representation of SpA-N and the 5B construct. SpA-N has five domains with high sequence identity: E, D, A, B and C. The 5B construct has 5 B domains in a row. The 5B is a mimic of SpA-N both structurally and functionally.

2.2.2 Sequential assignment of 5B

Sequential assignment of 5B (BMRB Accession Number 18594) was achieved using two 3D NMR experiments, HNCACB [64, 65] and CBCA(CO)NH [36]. The magnetization transfers in the two experiments are shown in Fig. 2.10 and Fig. 2.11. Initial assignments were transferred from Karimi et al. [66] and Gouda et al. [67] The two experiments were performed with a ^{15}N, ^{13}C labeled 5B construct at 30C on a Varian INOVA 600MHz and a Varian INOVA 800MHz NMR spectrometer, respectively in 90%/10% H2O/D2O and 100mM NaCl, 50mM NaAc, pH5.5. The two 3D experiments were collected with 1024 complex points on the proton dimension, 66 increments on the carbon dimension and 70 increments on the nitrogen dimension.

Based on the assignment, the 55 overlapping ^{15}N-HSQC resonances correspond to the residues in all 5 domains of 5B. Six additional resonances were assigned to the first five residues at the N-terminus and the last residue at the C-terminus. Those terminal residues experience different electromagnetic environments from the corresponding interdomain linker residues because they have no adjacent domain.

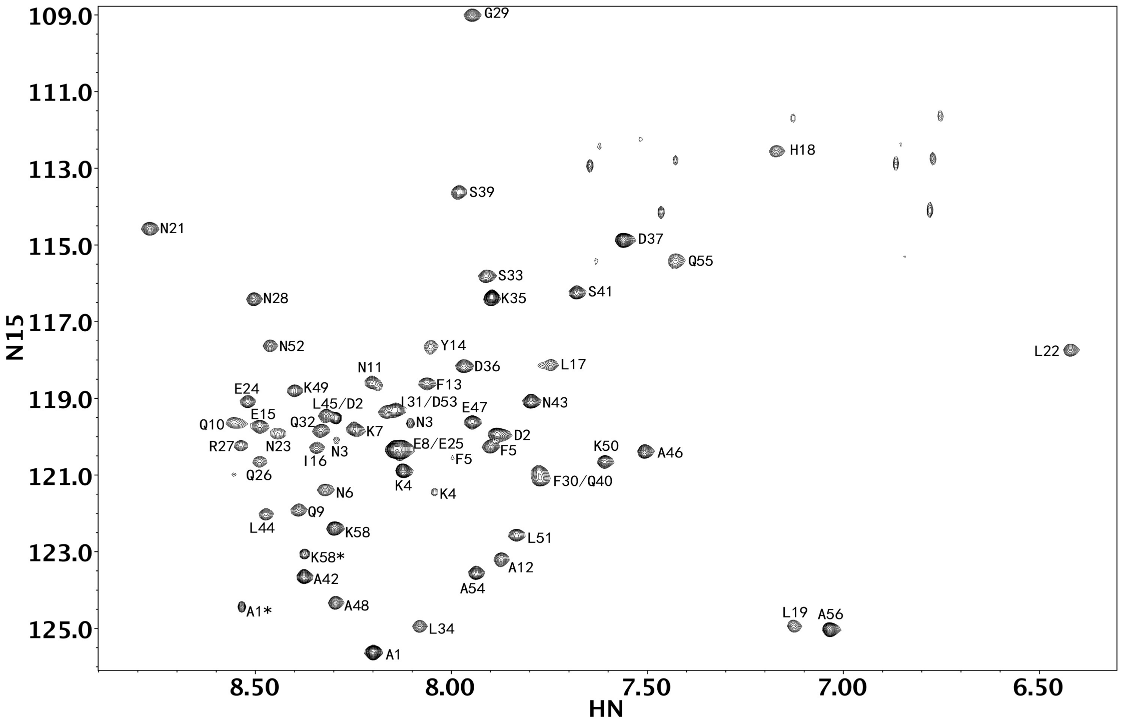


Figure 2.9: Assigned ^{15}N-HSQC spectrum of 5B.

2.2.3 Flexibility of 5B inferred from R1/R2 rates and NOE ratios

^{15}N T1, T2 measurements [68] and 15N NOE [64] experiments were performed with a ^{15}N labeled 5B construct at the same temperature on the Varian INOVA 600MHz NMR spectrometer. All spectra were collected with 1024 complex points on the proton dimension and 128 increments on the nitrogen dimension. The delay times used in T1 experiments were 10, 130, 250, 500, 750, 1000, 1500, 2000 and 3000ms and duplicate spectra were collected at 10, 750 and 3000ms delays for estimating errors. The delay times used in T2 experiments were 10, 30, 50, 70, 90, 110, 130, 170 and 250ms and duplicate spectra were collected at 10, 110 and 250ms delays for

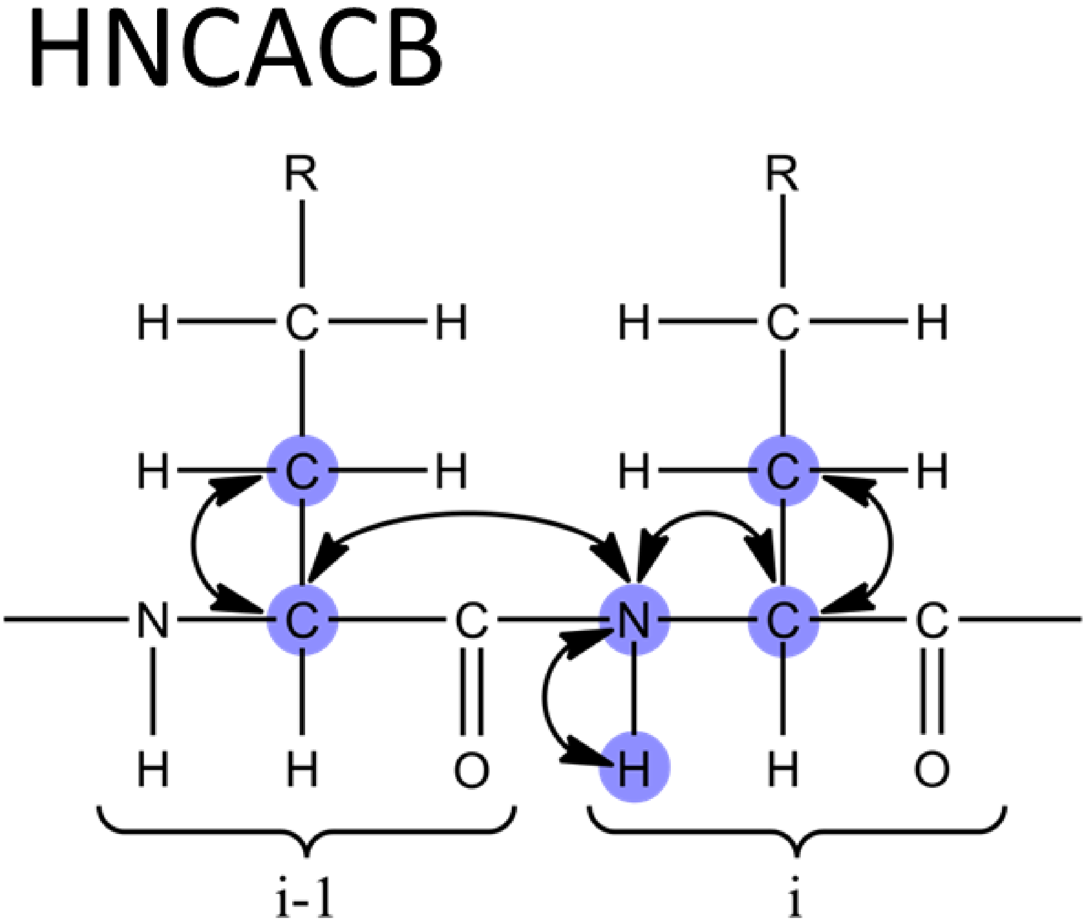


Figure 2.10: The magnetization transfer in HNCACB experiment. CA, CB and CA-1, CB-1 are all visible in the experiment.

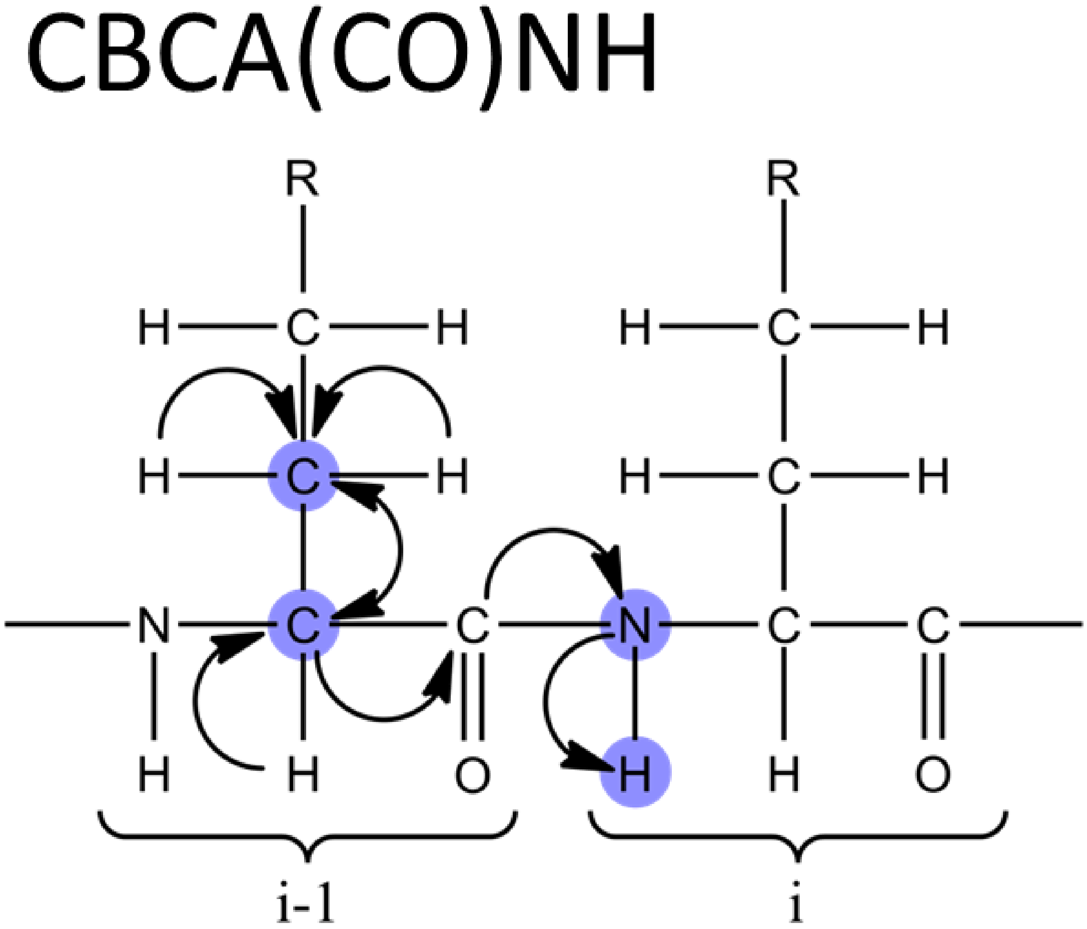


Figure 2.11: The magnetization transfer in CBCA(CO)NH experiment. Only CA-1 and CB-1 are visible in the experiment.

estimating errors. ^{15}N NOE experiments were performed by collecting two spectra either with or without a 2.5s presaturation delay. The relaxation data were analyzed using the software NMRViewJ [69]. Peak intensities on the 2D spectra were measured by integrating all point intensity within a ellipse centered on a peak. R1 and R2 rates were calculated by fitting the time points to an exponential function.

We derive some preliminary conclusions by just observing the R1/R2 rates and heteronuclear NOE ratios. A sharp drop in the heteronuclear NOE ratios was observed for the first five residues and the last one residue in each domain (Fig. 2.12). Although R1 rates do not very much in the R1 plot, R2 rates also drop for the first five residues and the last one residue in each domain. The drop in heteronuclear NOE ratios and R2 rates indicates the corresponding residues are flexible. Other parts of the protein have relatively large heternuclear NOE ratios and R2 rates, so they are fairly rigid and have limited amount of motion freedom. It should be noted that there are two groups of residues on the N-terminus and C-terminus. The first group of residues contains the five residues and the last one residue of domain 2-4. The second group of residues contains the five residues and the last one residue of domain 1-5. Because of the assignment problem mentioned earlier, we cannot distinguish one group from the other just from its chemical shifts. However, when we observe the relaxation behavior of the two groups, they are clearly different. One group of residues (represented as red points in Fig. 2.12) has smaller heteronuclear NOE ratios and R2 rates than the other one(represented as red points in Fig. 2.12). The second group of residues has no adjacent domains. Consequently, they should be more flexible and should have smaller heteronuclear NOE ratios and R2 rates. Based on this, we infer that the group of residues with smaller heteronuclear NOE ratios and R2 rates is the second group and that the group of residues with larger heteronuclear NOE ratios and R2 rates is the first group.

2.3 Analyzing the relaxation data with both symmetric and asymmetric tumbling models

2.3.1 Symmetric and asymmetric tumbling models

The global tumbling of a molecule can always be decomposed into three eigen rotations. The three rotations rotate the molecule around three perpendicular axes and may have different rates. Any other rotation of the molecule is a linear combination of the three eigen rotations. Because the global tumbling is driven by the Brownian motions, the rotation rates depend heavily on the shape of the molecule. When the molecule is globular, the rotation rates along the three directions are the same. Consequently, the global tumbling can be specified by a single correlation time. Based on Eq. 2.13, the spectral density function of isotropic tumbling takes the following form:

$$J_i(\omega) = S_i^2 \frac{\tau_c}{1+\omega^2\tau_c^2} + (1-S_i^2)\frac{\tau}{1+\omega^2\tau_i^2}, \tag{2.14}$$

where we have a global parameter τ_c and local parameters S_i^2, τ_i. When the global tumbling is anisotropic, the spectral density function still takes the form of Eq. 2.13. However, the parameter τ_c is not a fixed number and it depends on the orientation of the bond vector, in our case, the NH bond vector. As a result, we have the following equation for the anisotropic case (ref):

$$J_i(\omega) = S_i^2 \frac{\tau_{c,i}}{1+\omega^2\tau_{c,i}^2} + (1-S_i^2)\frac{\tau}{1+\omega^2\tau_i^2}, \tag{2.15}$$

where $(\tau_{c,i})^{-1} = \mathbf{v}^T D\mathbf{v}$. The matrix D is the diffusion tensor, which is a 3 × 3 symmetric matrix. By doing an SVD on the diffusion tensor, we could derive its eigen values D_1, D_2 and D_3. The relationships between the eigen values and the rotational diffusion constants D_{xx}, D_{yy} and D_{zz} are listed below:

$$D_{xx} = -D_1 + D_2 + D_3 \tag{2.16}$$

$$D_{yy} = D_1 - D_2 + D_3 \tag{2.17}$$

$$D_{zz} = D_1 + D_2 - D_3 \tag{2.18}$$

Given the rotational diffusion constants, the rotational correlation times can be calculated (ref):

$$\tau_{xx} = D_{xx}^{-1} \tag{2.19}$$

$$\tau_{yy} = D_{yy}^{-1} \tag{2.20}$$

$$\tau_{zz} = D_{zz}^{-1} \tag{2.21}$$

2.3.2 Lipari-Szabo mapping and analysis of the result

We first generated a Lipari-Szabo map by deriving spectral density estimates from the relaxation experiments (Fig. 2.13A). The Lipari-Szabo mapping method provides a graphical approach to estimate and display order parameters [70]. Because it is a graphical approach, it does not contain the explicit assumptions about either isotropic tumbling or anisotropic tumbling. In addition, in the pattern shown in Lipari-Szabo mapping, there are hints about the anisotropy of the global tumbling. The points in Figure 1A corresponding to the majority of residues form a cluster near the bottom of rigid limit curve. For these residues, reorientation of the backbone N-H bond is primarily due to global tumbling. However, 9 points are well separated from the cluster, indicating that the N-H bonds are much more flexible. Four of these points correspond to residues 3-5 and the C-terminal residue of 5B and are assigned to the termini because they have the lowest order parameter for the sequentially-assigned residue. Their low order parameters are typical for the termini of proteins. The remaining low order parameter N-H bonds correspond to residues 1-5 and 58 of each non-terminal B domain. The order parameters of these N-H bonds are nearly as low as those of the termini. The heteronuclear NOE ratios (Fig. 2.12) and the Lipari-Szabo mapping (Fig. 2.13A) are consistent with each other. Both of them show that there is a six-residue linker (KADNKF) between every pair of adjacent domains. While linkers between D-A-B-C domains of SpA-N have the same sequence as the linker in 5B, the linker between E and D domains has three additional residues. As a result, the E-D domain linker could be more flexible than others. The motions of each domain could deviate from symmetric tumbling because the motions are also affected by the motions of adjacent domains. Indeed, the Lipari-Szabo mapping plot does not have the appearance of one generated from a globular protein.

Based on the observation from the Lipari-Szabo mapping, we modeled the global tumbling as an anisotropic one and specified it using three correlation times. Using the asymmetric tumbling model, we estimated the three global correlation times. In order to estimate the three correlation times, we first projected the points corresponding to residues in the Lipari-Szabo model to the rigid limit curve. The projection performs a rough approximation of the global tumbling components of each residue. Using the global tumbling components, the rotational diffusion tensor D can be calculated straightforwardly using an SVD approach. Because the SVD approach performs a global fit, individual errors existed in the first-order approximations can be reduced by the SVD step. After the SVD step, we obtained the global correlation times and order parameters for each residue. The three correlation times are 11.02, 11.91 and 14.58 ns, which are much smaller than the correlation time of a globular protein of the size of SpA-N. The small correlation time implies there are interdomain motions due to the flexible linkers.

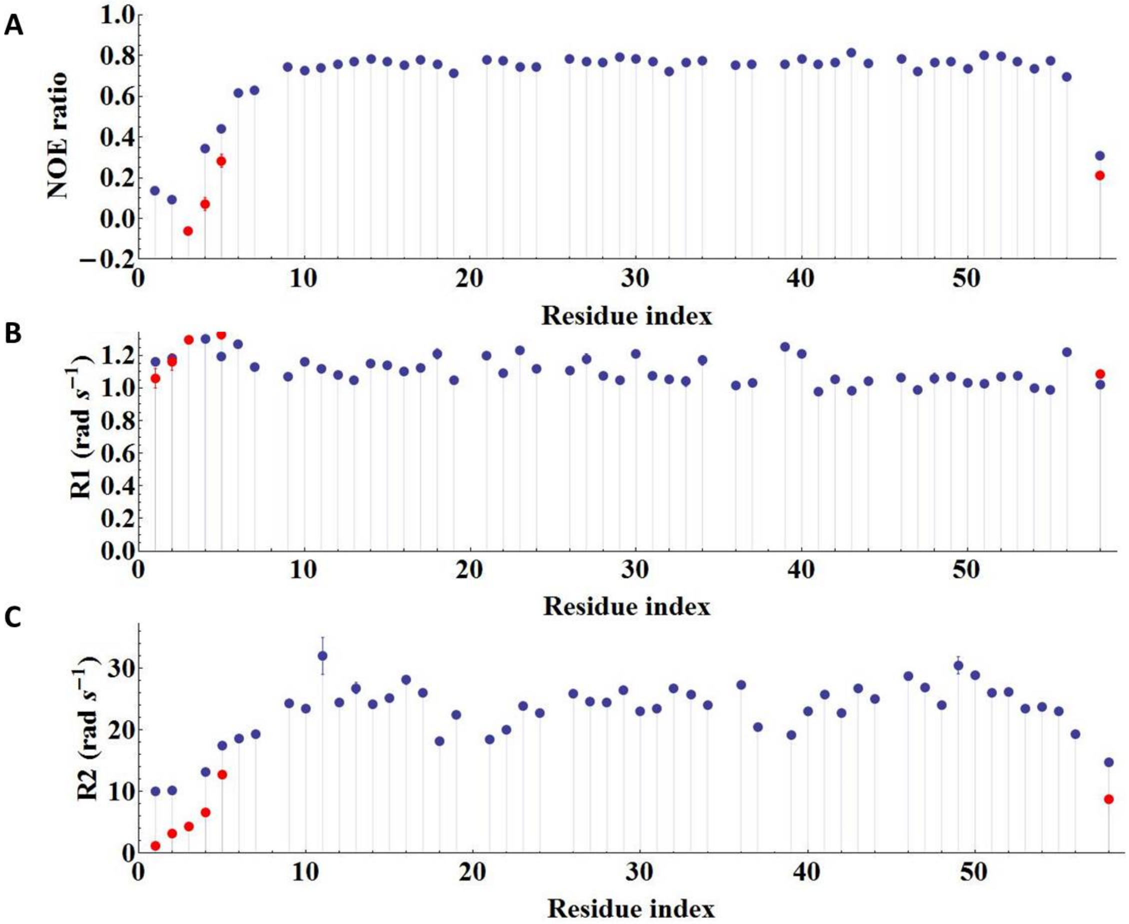


FIGURE 2.12: R1, R2 rates and HetNOE ratios of each residue in 5B. Red points correspond to terminal residues.

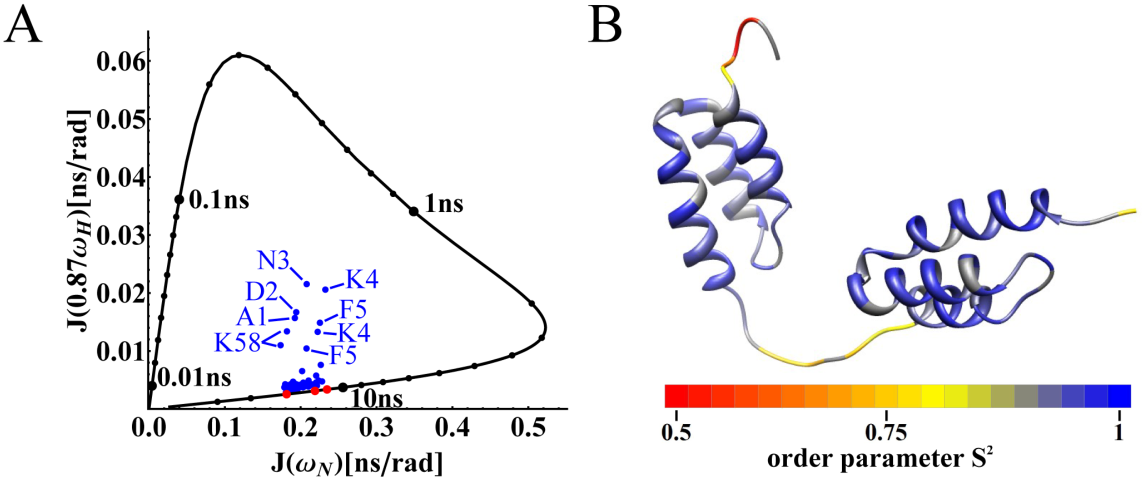


FIGURE 2.13: Relaxation properties of 5B. (A) Lipari-Szabo mapping of 5B. Residue resonances are plotted as blue points inside the rigid limit curve. Anisotropic correlation times are plotted as red points on the rigid limit curve. (B) order parameters are color coded onto a structure of double B domains. Residues are colored as gray when data is not available.

3

Measuring SpA-N inter-domain motions using residual dipolar couplings (RDCs)

3.1 Introduction to residudal dipolar coupling (RDC) and magnetic alignment

Nuclear magnetic resonance (NMR) observes the states of nuclei in the sample. Nuclei interact with the electromagnetic environments surrounding them. Each nucleus has a magnetic moment, which interacts with the surrounding magnetic field. When the surrounding magnetic field is altered by surrounding spins or electrons, NMR can observe the altered behavior of the nucleus and thus infer the change in its environment. Actually, the surrounding magnetic field has two components, the external magnetic field and the internal magnetic field. In an NMR experiment, the external magnetic field is supplied by a strong magnet. The intensive external magnetic field influenced all electrons and nuclei in the sample so all electrons and nuclei generate a small magnetic field around themselves. The sum of all the small magnetic fields are considered as the internal magnetic field. In the following subsections, I discuss about residual dipolar couplings(RDCs), which is about the interactions between two spins. The observation of RDCs depends on the alignment of the molecule, which can be a result of the interaction between an unpaired electron and the external magnetic field. The introduction also introduces the concept of motional decoupling, which is an experimental technique to use a dominating source of either paramagnetism or diamagnetism and control alignments. When the molecule is magnetically aligned, we could also observe the pseudo-contact shifts (PCSs), due to the interaction between the unpaired electron and ~~the~~ observed spin.

3.1.1 Mathematical formulations of residudal dipolar couplings (RDCs)

As discussed above, under an external magnetic field, nuclei with non-zero spins have induced magnetic dipole moments. The nuclear magnetic dipoles can interact with each other through two different mechanisms: one indirect mechanism and one direct mechanism. In the first mechanism, the two nuclear magnetic dipoles ~~can also~~ interact with each other indirectly when the two nuclei are covalently bonded. Because they are covalently bonded, the two atoms share electrons. The shared electrons are influenced by one magnetic dipole due to the electron-spin interaction. The influenced electrons, on the other hand, influences the dipole moment of the the other spin. The two nuclear magnetic dipoles interacts through the shared electrons. The indirect dipole-dipole interaction is called the scalar coupling (J coupling). In addition, nuclear magnetic dipoles can also interact with each other directly. The dipole moment generates a small magnetic field, which can be felt by adjacent spins. The direct interaction between magnetic dipoles of nuclei is called the dipolar coupling. The equation for dipolar couplings is listed below:

$$D_{IS} = -\frac{\mu_0 \gamma_I \gamma_S \hbar}{4\pi^2 \langle r_{IS}^3 \rangle} \frac{3\cos^2\theta - 1}{2}, \tag{3.1}$$

where μ_0 is the magnetic permeability of vacuum, γ_I and γ_S are gyromagnetic ratios of spin I and S respectively. r_{IS} is the internuclear distance between the two spins, θ

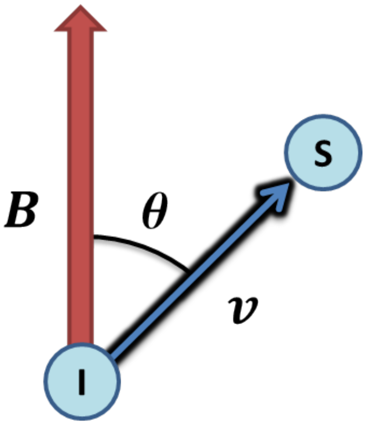


FIGURE 3.1: The angle between the bond vector and the direction of the magnetic field B_0. I and S are two spins forming the bond vector v.

is the angle between the magnetic field unit vector B_0 and the unit vector v between spin I and spin S in Fig. 3.1.

When a biomolecule is observed in solution by NMR, the molecule tumbles and may have internal motions. The unit vector between the two spins usually is not static as shown in Fig. 3.1. The resulted dipolar coupling is an average over all possible angle θ, which is called the residual dipolar coupling (RDC). The equation of RDC is listed below:

$$D_{IS} = -\frac{\mu_0 \gamma_I \gamma_S \hbar}{4\pi^2 \langle r_{IS}^3 \rangle} \langle \frac{3\cos^2\theta - 1}{2} \rangle. \tag{3.2}$$

Here, the bracket $\langle\rangle$ means an average on all possible angle θ. Observe $cos(\theta) = B_0^T v$, we have the following equation for RDC:

$$D = \frac{K}{2} \langle 3(B_0^T v)^2 - 1 \rangle, \tag{3.3}$$

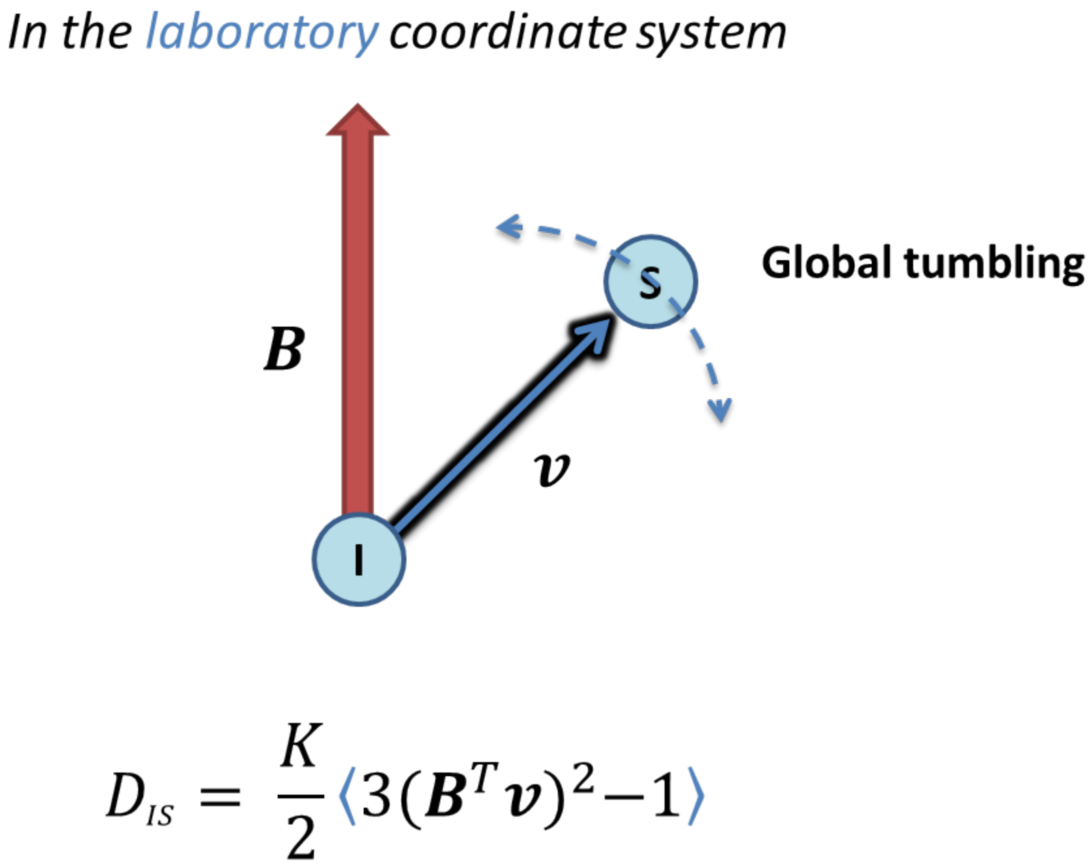
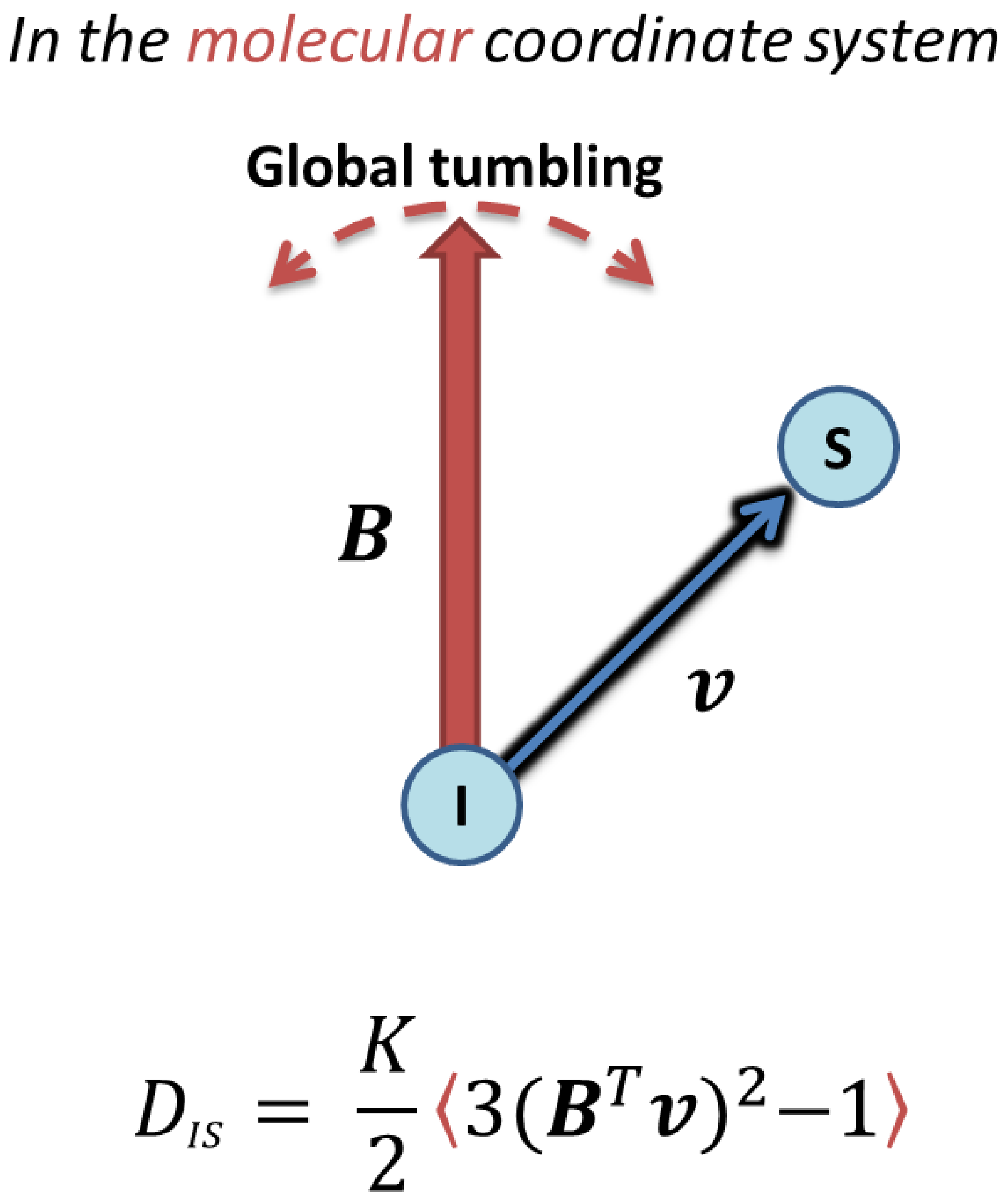

In the laboratory coordinate system
B
S
Global tumbling
v
I
$D_{IS} = \frac{K}{2}\langle 3(\boldsymbol{B}^T\boldsymbol{v})^2 - 1\rangle$

FIGURE 3.2: RDC is average over the vector v in the laboratory coordinate system.

where K is the dipolar coupling constant. K is calculated as:

$$K = -\frac{\mu_0 \gamma_I \gamma_S \hbar}{4\pi^2 \langle r_{IS}^3 \rangle} \tag{3.4}$$

In the laboratory coordinate system, the unit vector B_0 is static and the average is over vector v (Fig. 3.2). However, in the molecular coordinate system, the unit vector v is static and the average is over vector B_0(Fig. 3.3). The constant vector v can be taken out of the average, the resulted RDC equation is formulated in the



$$D_{IS} = \frac{K}{2}\langle 3(\boldsymbol{B}^{T}\boldsymbol{v})^{2}-1\rangle$$

FIGURE 3.3: RDC is average over the magnetic field vector B in the molecular coordinate system.

following way [71]:

$$D = \frac{K}{2}\mathbf{v}^T S \mathbf{v}, \tag{3.5}$$

where matrix S is a 3×3 traceless symmetric matrix in the following form:

$$\begin{aligned} S &= \langle 3BB^T - I \rangle \\ &= \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{bmatrix}. \end{aligned} \tag{3.6}$$

The matrix S is formally named as the Saupe tensor, which has 5 DOFs: axiality, rhombicity and orientation of the principle axes in a molecular frame. We observe that the Saupe tensor is an average over magnetic field vector B_0, so it contains information about how the magnetic field vector B_0 moves around the molecule in the molecular coordinate system. In other words, it contains information about how the molecule tumbles around the magnetic field vector in the laboratory coordinate system. When the molecule samples all orientations isotropically in solution, all elements in the Saupe tensor average to zeros. As a result, we do not observe RDCs and only observe scalar coupling (J coupling) in experiments. In order to observe RDCs, the molecule must have preferences over certain orientations. In this case, the Saupe tensor is not zero and the molecule is considered as aligned. The Saupe tensor is usually used to represent the alignment of the molecule. Saupe tensors can be calculated from RDCs by an SVD method [72]. Because the coupling we observe is always a sum of scalar coupling (J coupling) and dipolar coupling, we need to separate the dipolar coupling component from the sum. The answer is to perform two separate experiments. ~~The~~ molecules are aligned in one experiment and are not aligned in the other experiment. Because we only observe the scalar coupling (J coupling) in the unaligned experiment, we can abstract the scalar coupling (J coupling) component from the sum and obtain the dipolar coupling component.

3.1.2 Alignment methods

There are several experimental methods to align the molecule. The alignment methods can be divided into three categories ~~based on their principles~~: mechanical alignment, direct magnetic alignment and indirect magnetic alignment. The mechanical alignment does not depend on the external magnetic field. A dilute gel is incubated in a solution of biomolecules, so protein molecules can diffuse into the gel. The gel is compressed or stretched in a following step and mounted into the NMR tube. The stretch or compression of the gel at the macroscopic level provides mechanical forces to align the biomolecules at the microscopic level (Fig. 3.4). For this kind of alignment method, the generated mechanical force must be homogeneous across the sample. Otherwise, the observed RDCs may not agree on one unique value. Another disadvantage of the method is that the sample generally degrades over time. The degradation could be due to gel deformation and release of the mechanical force. The sample may not be used for multiple experiments with long time intervals and may not be suitable for one long time experiment.

The second alignment method is to align biomolecules directly through the magnetic interactions between the biomolecule and the external magnetic field. All biomolecules has their intrinsic paramagnetism and diamagnetism. The paramagnetism could come from spins of nuclei and electrons. Although the paramagnetism of electron spin is orders of magnitude larger than the paramagnetism of nuclei, only unpaired electron contribute to the net paramagnetism because the paramagnetism of paired electrons cancels out. The diamagnetism could come from the orbiting movement of electrons. The diamagnetism is especially large when electrons are circulating in a π-orbital. When either the paramagnetism or the diamagnetism of a molecule has large anisotropy, different orientations of the molecule correspond to different energies. In other words, the energy potential between the external magnetic

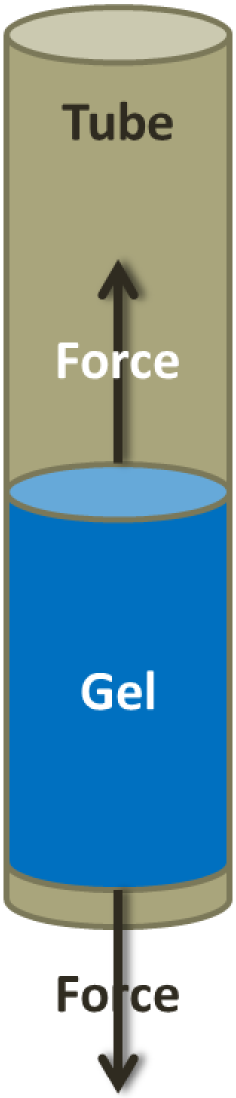


Figure 3.4: The experimental setting of a mechanic alignment. The gel is stretched at the macroscopic level. The stretch provides mechanical forces to align the biomolecules at the microscopic level.

field and the molecule is not uniform on all molecular orientations. Consequently, the molecule has a preference over certain orientations and it is aligned. The paramagnetism or diamagnetism anisotropy is usually small for protein molecules. In order to magnetically align protein molecules, an unpaired electron can be engineered into the molecule. The insertion of the unpaired electron could be through a covalent linkage to chemical agents with unpaired electrons, or through a coordination of ions with unpaired electrons. On the other hand, DNA and RNA molecules with idealized helix structures usually have large diamagnetism. The bases in DNA and RNA molecules are π-rings and their stacking contribute to large diamagnetism anisotropy.

The last alignment method is to align biomolecules indirectly through magnetic interactions (Fig. 3.5). As discussed above, DNA and RNA molecules with idealized helix structures have large diamagnetism. Phage has ~~nearly~~ parallel helices packed in its body and can be magnetically aligned very easily. When a high concentration of phage is added into a solution of biomolecules, the molecules interact with phage. Because phage is aligned, the interaction between the molecule and phage depends on the orientation of the molecule. Consequently, the molecule has a preference over certain orientations and it is aligned.

3.1.3 Motional decoupling

For the study of interdomain motions, the mechanical alignment and the indirect magnetic alignment are not the top choices. By definition, interdomain motions alter the shape of a molecule significantly. Mechanical alignment and ~~the~~ indirect magnetic alignment both depend on the shape of a molecule due to the use of alignment media, such as stretching gel or phage. Without losing generality, suppose the interaction between the alignment media is purely electrostatic. When the shape of the aligned molecule changes, the charge distribution on the molecule may also changes. If the charge distribution changes, the interaction between the molecule and the alignment

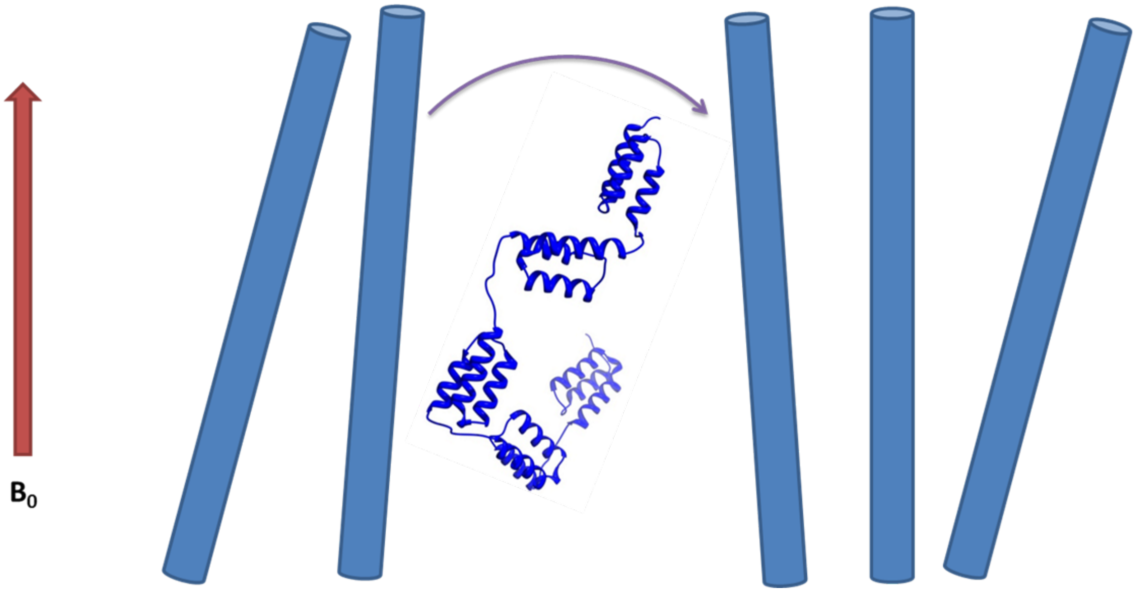


Figure 3.5: An indirect magnetic alignment. Phage is aligned and the alignment restricts the orientation of the molecule.

media changes. As a consequence, the alignment depends on both the global motions and the interdomain motions. The two motional modes are coupled in the RDCs obtained through mechanical alignment or ~~the~~ indirect magnetic alignment (Fig. 3.6). In a different perspective, ~~the~~ Saupe tensor depends on the conformation. The average observed by RDCs is weighted by ~~the~~ Saupe tensors. The true population weight for each conformation can not be constructed from the convolved observables.

On the other hand, when we use directly magnetic alignment, we could solve the above issue and achieve motional decoupling (Fig. 3.6). It should be noted that the alignment should be dominated by one domain. For example, the attachment of unpaired electron to only one domain in the protein case. Although the protein molecule has intrinsic diamagnetism and paramagnetism, the intrinsic magnetism

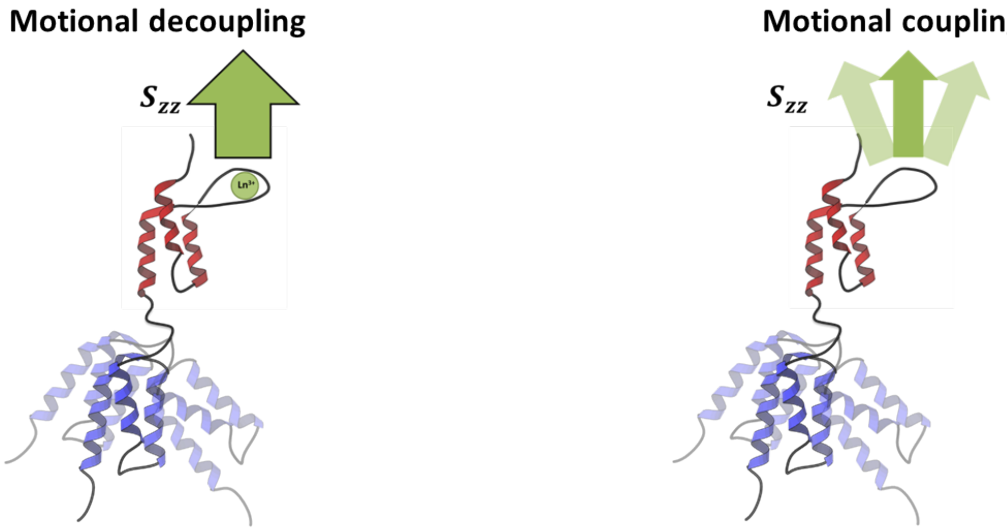


Figure 3.6: Motional coupling and motional decoupling. In the motional coupling scheme, global tumbling is affected by interdomain motions. In the motional decoupling scheme, global tumbling is independent of interdomain motions.

is very small compared to the paramagnetism of the unpaired electron. So the alignment of the domain with the unpaired electron reflects the global tumbling of the whole molecule. The internal motions of the molecule do not influence the alignment of the first domain and thus the global tumbling motion of the first domain is decoupled from the internal motions of the molecule.

3.1.4 The electromagnetic effects of unpaired electrons

As discussed before, the usual alignment method to achieve motional decoupling in a protein system is paramagnetic alignment. The paramagnetic alignment requires the attachment of a unpaired electron either through covalent linkage or coordination. In the presence of an unpaired electron, several NMR observables can be measured. The paramagnetism of the unpaired electron can be described by a magnetic susceptibility tensor, χ. Similar to the Saupe tensor, the χ tensor also has three eigen values along the three principle axes, χ_x, χ_y and χ_z. The χ tensor is isotropic if the the three eigen values are the same and the tensor is anisotropic if the eigen values are not the same. Metal ions such as Mn^{2+}, Gd^{3+} and Lu^{3+} have isotropic tensors. On the other hand, metal ions such as Fe^{3+}, Dy^{3+} and Tb^{3+} have anisotropic tensors. The electron g-tensor and the magnetic susceptibility χ tensor are closely related. Their relationship can be formulated as [27]:

$$\chi_{ii} = \frac{\mu_0 N_A \mu_B^2 S(S+1)}{3kT} g_{ii}^2, \tag{3.7}$$

where μ_0 is the magnetic permeability of vacuum, S is the electron spin quantum number, N_A is the Avogadro's number, μ_B is the magnetic moment of the free electron, k is the Boltzmann constant and T is the temperature. g_{ii} and χ_{ii} are an element of the g-tensor and an element of the χ-tensor respectively, where $i = \{x, y, z\}$.

The unpaired electron is paramagnetic no matter whether it is isotropic or anisotropic. In the presence of the unpaired electron, the relaxation of surrounding spins are enhanced due to the strong paramagnetisum. As a result, we observe paramagnetic relaxation enhancements for adjacent spins. When the spin is more close to the unpaired electron, the paramagnetic relaxation enhancement (PRE) effect is more significant. Actually, the enhanced relaxation rate is proportional to r^{-6}, where r is the distance between unpaired electron and an spin. In addition, when the unpaired electron has an anisotropic magnetic susceptibility χ-tensor, the interaction of the unpaired electron with the external magnetic field depends on the orientation of the unpaired electron, which is also the orientation of the molecule. Due to the preference of the molecule over certain orientations, the molecule is aligned. In this case, both RDCs and pseudo-contactr shifts (PCSs) do not average to zero and thus can be observed. The observed values of RDCs and PCSs depend on the χ-tensor. The equation for RDCs from a χ-tensor is the following:

$$D = \frac{B_0^2}{15\mu_0 kT}\frac{K}{2}\mathbf{v}^T\chi\mathbf{v}, \tag{3.8}$$

where B_0 is the magnitude of the external magnetic field. By comparing Eq. 3.5 and Eq. 3.8, we have the following relationship between the Saupe tensor and the χ tensor:

$$S = \frac{B_0^2}{15\mu_0 kT}\chi \tag{3.9}$$

The magnetic alignment originated from the unpaired electron also generates pseudo-contact shifts (PCSs). Unpaired electrons with anisotropic χ tensors perturb the chemical shifts of the surrounding spins. The perturbation in chemical shifts is called the paramagnetic shifts. The paramagnetic shifts result from both through-bond and through-space interactions. The through-bond interactions generate contact shifts and the through-space interactions generate pseudo-contact shifts. The bonded nuclei with contact shifts usually have fast relaxation due to the strong PRE effect. As a result, the pseudo-contact shifts are more useful observables. The equation for PCSs from a χ-tensor is the following:

$$\delta^{PCS} = \frac{1}{12\pi r^3}\mathbf{v}^T\chi\mathbf{v}, \tag{3.10}$$

where r is the distance between the unpaired electron and a nucleus. It should be noted that the above Eq. 3.10 only works when the distance r is a constant. When there are motions changing the distance r, PCSs are also averages over the distance r:

$$\delta^{PCS} = \langle\frac{1}{12\pi r^3}(3(B_0^T v)^2 - 1)\rangle. \tag{3.11}$$

As a result, PCSs contain more information than RDCs. PCSs are averaged by both rotational motions and translational motions. It is also possible that the extra information in PCSs can be used to reconstruct both rotational motions and translational motions.

3.2 Measuring binding affinities between di-domain mimics and lanthanide ions

Based on previous discussions, a good strategy to obtain motional decoupled RDCs is through engineering an unpaired electron system to a biomolecular system. For the SpA-N protein system, we chose lanthanide ions as our source of unpaired electron and paramagnetism. In order to incorporate the lanthanide ions into the system, we engineered a lanthanide binding tag into one of the domains. In this section, I demonstrate that the engineered constructs have high binding affinities to lanthanides. I start with the luminescence property of lanthanides, which is followed by a method to measure the binding affinity based on the luminescence property. At the end, I show the experiment results and calculate the binding affinities.

3.2.1 Luminescence property of lanthanides

Lanthanides are rare earth elements. The lanthanide series include 15 elements with atomic number from 57 through 71. All lanthanide elements have their 5d and 6s orbital filled. They vary in their 4f orbital, which is filled with 0 to 14 electrons. The lanthanides usually exist as trivalent ions with electrons in the 5d and 6s orbital deprived. The photophysical property of lanthanides comes from the transitions of the f-electrons, which gives long-lived luminescence and sharp absorption and emission lines. The f-f electronic transitions are forbidden, leading to low extinction coefficients. As a result, lanthanide ions has a poor ability to absorb light and thus direct photoexcitation of the lanthanide ions is difficult. In order to observe lanthanide luminescence, lanthanide ions are usually complexed with organic chromophores. Organic chromophores can absorb light more easily. With a proper nearby chromophore, ~~the~~ energy absorbed by the chromophore can be transferred to the lanthanide ion [73]. The excited lanthanide ion then emits luminescence at characteristic wavelengths (Fig. 3.7). In the above process, the organic chromophore absorbs light over a much broader spectral range. Intuitively, the organic chromophore works as an antenna and harvests light for lanthanide ions. Consequently, the organic chromophores are usually refered as antenna chromophores. It should be noted that the above organic-to-lanthanide energy transfer differs from fluorescence resonance energy transfer (FRET) and luminescence/lanthanide based energy transfer (LRET). For FRET, the energy transfer is through dipole-dipole interactions. The efficiency of energy transfer is proportional to the inverse sixth power of the distance between the donor and acceptor. In general, the distance is in the range of 10-80Å. The mechanism of FRET is very different from the mechanism of the organic-to-lanthanide energy transfer, which is direct transfer from the excited singlet state to the lanthanide ion. Consequently, the organic chromophore and the lanthanide ion must be close to each other. The LRET is a combination of both the organic-to-lanthanide energy transfer and FRET. A chromophore-lanthanide complex is used as the donor in LRET. For LRET, Donor transfer the energy to the acceptor in the same way as in FRET. The most commonly used luminescent lanthanide ion is Tb^{3+}, which has two characteristic emission wavelengths at 495 and 545 nm.

3.2.2 The di-domain mimics of SpA-N with lanthanide binding tags (LBTs)

We constructed a di-domain construct that links a B domain variant “Z domain” [74] and a C domain together using the conserved linker. Z domain differs from B domain by only two residues, A1V and G29A. Consequently, they have nearly indistinguishable structures [74, 75]. In order to align the molecule and achieve motional decoupling, we introduced a rigid lanthanide binding tag (LBT) into the loop between Helix II and Helix III of the Z domain to create the ZLBT domain in the

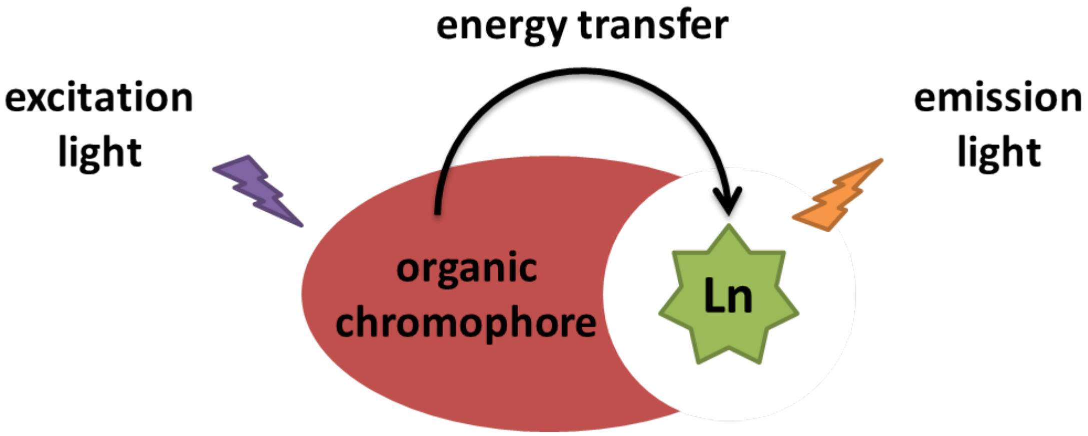


FIGURE 3.7: The energy transfer from an organic chromophore to a lanthanide ion. The organic chromophore works as an antenna and absorbs light from a broad spectrum. The absorbed energy is transfered to the lanthanide ion, leading to emission.

N-terminal domain, linked to C domain at the C-terminus (referred to hereafter as ZLBT-C). We also constructed a second version of lanthanide-binding ZLBT domain (NHis-ZLBT-C) by introducing a His-tag at its N-terminus (Fig. 3.8).

3.2.3 Measuring binding affinity between LBT and Tb^{3+} by fluorescence

TbCl$_3$ was titrated into 0.5 μM protein solution (ZLBT-C or NHis-ZLBT-C) to final concentrations ranging from 0.25 to 10 μM. The protein solution is in the buffer of 25 mM MOPS (pH7.2) and 100 mM KCl.The tyrosine residue in the middle of LBT was excited at 280 nm and the emission of the Tb^{3+} ion was observed at 495 and 545 nm. The emission intensities were fit to the following binding equation to obtain the dissociation constant K_D:

$$I_e = c_e \frac{[P] + [Tb^{3+}] + K_D - \sqrt{([P] + [Tb^{3+}] + K_D)^2 - 4[P][Tb^{3+}]}}{2[P]} \tag{3.12}$$

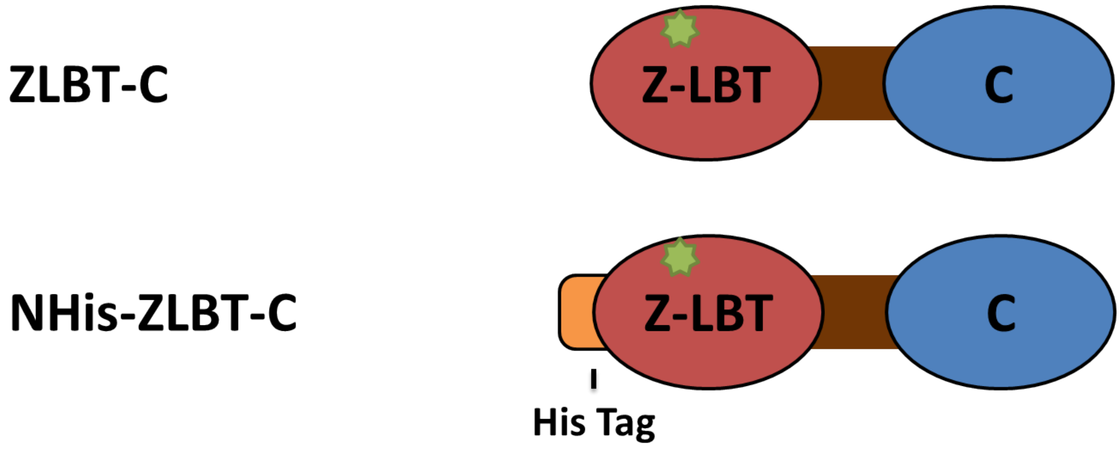


Figure 3.8: Schematic representation of the two di-domain mimics of SpA-N: ZLBT-C and NHis-ZLBT-C.

where I_e is the emission intensity, c_e is the emission coefficient, $[P]$ is the protein concentration and $[Tb^{3+}]$ is the concentration of Tb^{3+} ions.

The ZLBT-C construct and the NHis-ZLBT-C construct both have nanomolar lanthanide binding affinities (Fig. 3.9). The construct NHis-ZLBT-C has a tighter binding affinity, suggesting that the His-tag contributes to the coordination of lanthanide ion. Because the His-tag is in close proximity to the LBT, the tighter binding affinity could be due to a direct interaction between the His-tag and the lanthanide ion. Alternatively, the His-tag may favor tighter-binding-affinity conformation of LBT in some other way, indirectly enhancing the binding of the lanthanide ion to LBT.

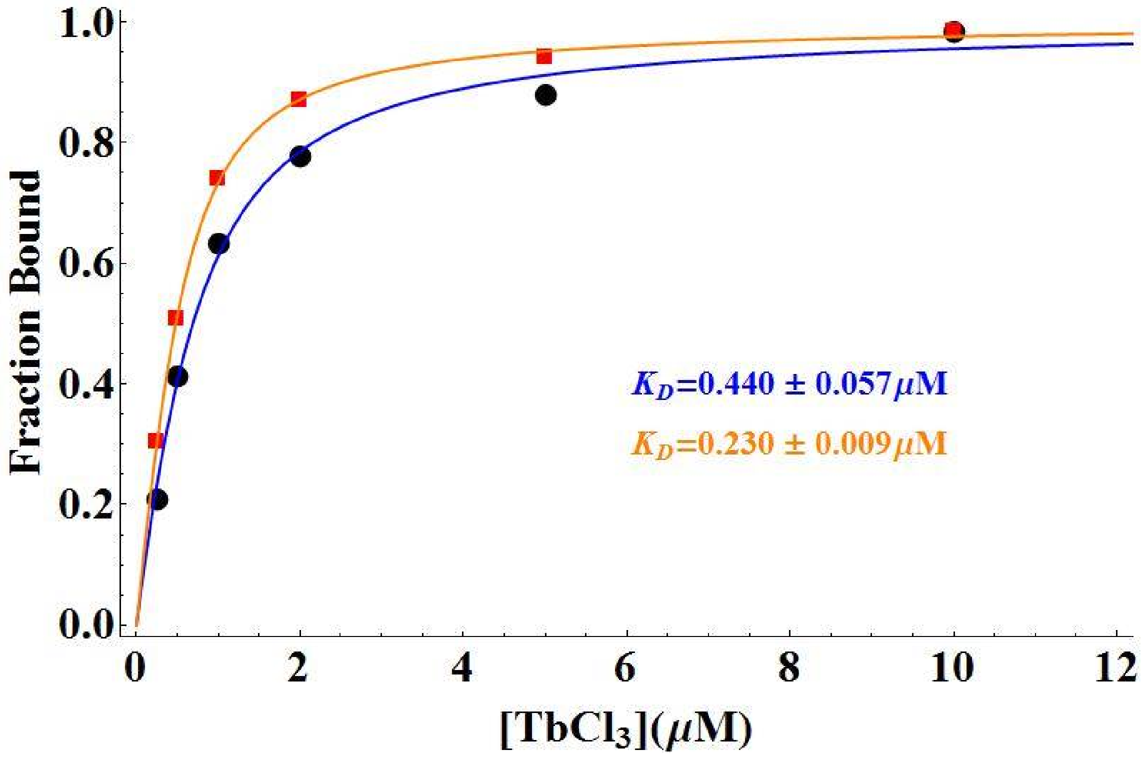


Figure 3.9: Binding affinities determined by fluorescence for the two di-domain mimics. Fluorescence data of ZLBT-C/Tb(black) and NHis-ZLBT-C/Tb(red). Fitted curves are shown as blue and orange respectively. The dissociation constants are shown followed by standard error.

3.3 Measuring RDCs with in-phase/anti-phase (IPAP) experiments

In the last section 3.2, I demonstrate that the constructs are able to bind lanthanide ions with a strong binding affinity. The constructs in coordination with lanthanide ions can be aligned in a strong magnetic field. With the alignment, we are able to observe and measure RDCs. In addition, ~~the~~ different lanthanide ions provides an inventory of alignment inducers. Because their different magnetic susceptibility, the alignment of a biomolecule can be controlled by using different lanthanide ions. In this section, I start with the paramagnetic properties of lanthanide ions. Then I demonstrate obtaining multiple alignments by using different combinations of lanthanide ions and protein constructs. At last, I introduce the NMR experiment to observe ~~the~~ RDCs and the procedure to measure ~~the~~ RDCs accurately.

3.3.1 Paramagnetism of lanthanides

As mentioned in subsection 3.2.1, lanthanide ions are trivalent ions with varying number of electrons in the 4f orbital. Most of them have unpaired electrons except La^{3+} and Lu^{3+}. The 4f orbital of La^{3+} has no electron and the 4f orbital of Lu^{3+} is full filled. Although the paramagnetic ions seem to be more useful, we do need the non-paramagnetic lanthanide ions. Because Lu^{3+} is not paramagnetic, it can be used as a diamagnetic reference in experiments. As discussed in subsection 3.1.1, an unaligned reference is necessary to extract the dipolar coupling component from the observed sum of both dipolar coupling and scalar coupling (J coupling). The Lu^{3+} ions can be used to prepare the unaligned sample and measure the scalar coupling (J coupling) component. All other lanthanide ions have unpaired electrons and significant paramagnetism. Among them, Gd^{3+} has an isotropic magnetic susceptibility χ tensor. Although Gd^{3+} is still paramagnetic and causes PRE, it does not lead to alignment in the magnetic field. Because RDC is our focus but not PRE, Gd^{3+} is not a suitable candidate.In addition, Promethium (Pm) is radioactive and thus also excluded from our further considerations. The remaining 12 lanthanide elements all have anisotropic magnetic susceptibility χ tensor and thus can be used to generate alignments. We observe that the 12 lanthanide elements have various magnetic susceptibility anisotropy, which is especially useful to us. For RDC experiment, one alignment only observes the information from a particular point of view and provides a fraction of information. As we can see in section 3.4, a maximum of five orthogonal alignments can be achieved. In other words, one alignment only provides one fifth of the maximal information which can be obtained by RDCs. Consequently, we would like to obtain as many different alignments as possible. The various magnetic susceptibility anisotropies of the 12 lanthanide elements offer a way to generate different alignments. As shown in subsection 3.1.1, the alignment is specified by the Saupe tensor and it has 5 degrees of freedom. Two of them are axiality and rhombicity. The different magnetic susceptibility anisotropy at least lead to different axiality and rhombicity. Using different lanthanide ions may also lead to different orientation of the principle axes. However, the orientation of the principle axes is difficult to control because it also depends on the coordinate of lanthanide ions. Nonetheless, the 12 lanthanide elements provide an inventory of alignments where the axiality and rhombicity of an alignment can be modified very easily. When choosing lanthanide ions from the inventory, one should be careful about the magnitude of the paramagnetism. For lanthanide ions with a large magnitude paramagnetism, they offer strong alignment, but they also give large PRE effect and wipes out signals from adjacent spins. On the other hand, lanthanide ions with a small magnitude of paramagnetism give small PRE effect but give weak alignment as well. Because there is an inventory of lanthanide ions, we could balance the benefits and the losses and strike a point in the middle.

3.3.2 Obtaining multiple alignments by using different combinations of lanthanides and protein constructs

We measured residual dipolar couplings (RDCs) of the two constructs, ZLBT-C and NHis-ZLBT-C. The introduced His-tag in the NHis-ZLBT-C construct could perturb the binding mode of lanthanide ions and potentially result in a different alignment. Indeed, alignments from the two constructs are different (Table 3.1 and Table 3.2). We also combined the two constructs with two types of lanthanide ions, Dy^{3+} and Tb^{3+}. The two lanthanides have different magnetic susceptibilities and behave differently in a magnetic field, so the combination gave us a total of four different alignments (Table 3.1 and Table 3.2). The RDCs in all four alignments were measured by ^{15}N IPAP HSQC experiments. The RDCs of C domain in all four alignments are significantly smaller than those of B domain, indicating extensive interdomain motions (Fig. 3.10). The small RDCs of C domain were carefully measured by fitting the peaks to a mixture model with both Gaussian and Lorentzian components ~~as described in Materials and Methods~~. We performed duplicate experiments to estimate the RDC error. The estimated error is 0.2 Hz, which agrees with literature estimates and the RMSDs in the correlation plots (Fig. 3.10) [76]. Because no significant intra-domain dynamics were observed within the picosecond to nanosecond timescale, our analysis treats the non-linker region of each domain as a rigid body. We fit Saupe alignment tensors to the RDCs from the non-linker residues of each domain in each alignment [72]. There is a good agreement between experimental data and back-calculated data (Fig. 3.10), which in turn confirms the validity of the rigid body assumption. Fit results are summarized in Table 3.1 and Table 3.2.

Table 3.1: Z domain Saupe tensors for the raw RDC alignments

Name	$\mathbf{S_{xx} \times 10^4}$	$\mathbf{S_{yy} \times 10^4}$	$\mathbf{S_{zz} \times 10^4}$	$\mathbf{S_{xy} \times 10^4}$	$\mathbf{S_{xz} \times 10^4}$	$\mathbf{S_{yz} \times 10^4}$
ZLBT-C/Dy	4.28001	-7.6780	3.39795	6.5597	0.4012	1.0613
ZLBT-C/Tb	3.52548	-8.3596	4.83416	13.082	-0.8591	-0.7281
NHis-ZLBT-C/Dy	7.34644	-6.4965	-0.84999	5.8968	2.5109	1.5852
NHis-ZLBT-C/Tb	3.97532	-7.9106	3.93530	13.305	-0.0841	-1.3921

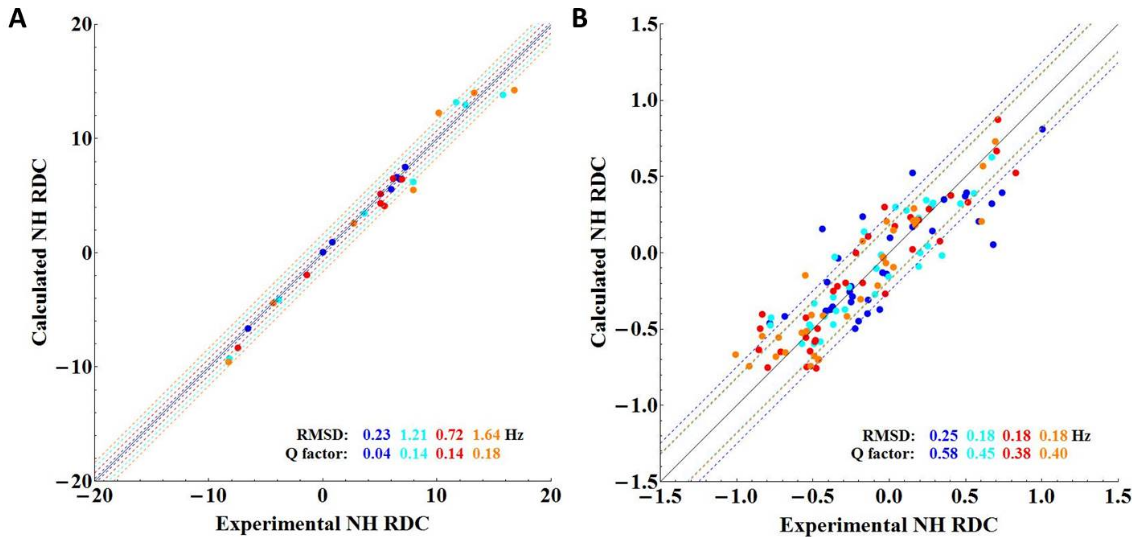


FIGURE 3.10: The correlation plot of all four RDC datasets of Z domain(A)/C domain(B) and the RDCs back-calculated from the fitted Saupe tensors.

Table 3.2: C domain Saupe tensors for the raw RDC alignments

Name	$\mathbf{S_{xx}} \times \mathbf{10^4}$	$\mathbf{S_{yy}} \times \mathbf{10^4}$	$\mathbf{S_{zz}} \times \mathbf{10^4}$	$\mathbf{S_{xy}} \times \mathbf{10^4}$	$\mathbf{S_{xz}} \times \mathbf{10^4}$	$\mathbf{S_{yz}} \times \mathbf{10^4}$
ZLBT-C/Dy	0.35884	-0.8921	0.53329	-0.2598	-0.2133	-0.3797
ZLBT-C/Tb	0.05846	-0.7146	0.65611	-0.4303	0.06909	-0.0526
NHis-ZLBT-C/Dy	0.44215	-0.9218	0.47967	-0.2330	0.00742	-0.5301
NHis-ZLBT-C/Tb	0.33124	-0.8024	0.47115	-0.5445	0.01350	-0.2809

3.4 Quantifying the information content in RDCs using the orthogonal linear combination (OLC) method

In the previous section 3.3, multiple alignments were obtained by using different combinations of lanthanide ions and protein constructs. Although the alignments are certainly different, we ~~still~~ don't know the similarity or difference between them. In this section, I introduce a way to measure the difference between different alignments and also a method to quantify the information content in the data set. At last, I show the result of the method and demonstrate that there is enough information content in our data.

3.4.1 The orthogonal linear combination (OLC) method

As mentioned in previous discussions (subsection 3.1.1), one alignment is specified by a Saupe tensor, which has 5 degrees of freedom (DOFs). As a result, one alignment can be represented as a 5-dimensional vector. The vector is a vectorized form of the corresponding Saupe tensor. Although the vectorized form is derived from harmonic functions in the original literature [26], I show a different derivation which gives the same result. A Saupe tensor is a 3 × 3 symmetric matrix. It has six different elements and thus lives in a six dimensional space. The vectorized Saupe tensor can be formatted as:

$$\mathbf{s} = [S_{xx}, S_{yy}, S_{zz}, S_{xy}, S_{xz}, S_{yz}]^T, \tag{3.13}$$

Because of the traceless constraint $S_{xx} + S_{yy} + S_{zz} = 0$, Saupe tensors actually only have five DOFs and live on a hyperplane within the six dimensional space. Variables S_{xx}, S_{yy} and S_{zz} live in a plane of the three dimensional space. Suppose a vector $\mathbf{v}$ on the plane is:

$$\mathbf{v} = S_{xx}\mathbf{i} + S_{yy}\mathbf{j} + S_{zz}\mathbf{k}, \tag{3.14}$$

where $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are three orthogonal base vectors. We can construct a new pair of base vectors:

$$\mathbf{i}' = \frac{\sqrt{2}}{2}(\mathbf{i} - \mathbf{j}), \tag{3.15}$$

$$\mathbf{j}' = \frac{\sqrt{6}}{6}(-\mathbf{i} - \mathbf{j} + 2\mathbf{k}). \tag{3.16}$$

The new base vectors are orthonormal and they are on the plane $S_{xx}+S_{yy}+S_{zz}=0$. If we write vector $\mathbf{v}$ regarding to the new bases, we have:

$$\begin{aligned}\mathbf{v} &= (\frac{\sqrt{2}}{2}S_{xx}-\frac{\sqrt{2}}{2}S_{yy})\frac{\sqrt{2}}{2}(\mathbf{i}-\mathbf{j})+\frac{\sqrt{6}}{2}S_{zz}\frac{\sqrt{6}}{6}(-\mathbf{i}-\mathbf{j}+2\mathbf{k}) \\ &= (\frac{\sqrt{2}}{2}S_{xx}-\frac{\sqrt{2}}{2}S_{yy})\mathbf{i}'+\frac{\sqrt{6}}{2}S_{zz}\mathbf{j}'.\end{aligned} \tag{3.17}$$

As a result, we could format the vectorized Saupe tensor in a different way:

$$\tilde{\mathbf{s}} = [\frac{\sqrt{6}}{2}S_{zz}, \frac{\sqrt{2}}{2}S_{xx}-\frac{\sqrt{2}}{2}S_{yy}, S_{xy}, S_{xz}, S_{yz}]^T, \tag{3.18}$$

The result agrees with the literature result and only differs by a constant [26]. In the following discussions, we only focus on the orthogonality between the vectors, but not their magnitudes. So it doesn't matter whether the vectors are scaled or out, as long as they are consistent with each other. From the result, we have vectors representing Saupe tensors and alignments. It is easy to measure the similarity or the difference between two vectors. One such measure would be the inner product of two vectors. When the two vectors are the same, the inner product is one. When the two vectors are orthogonal to each other, the inner product is zero. In addition, it is also possible to determine the span of the vectors. Even given more than five different vectors, they may not span the whole five dimensional space. They could reside on either a 4-dimensional hyperplane, 3-dimensional hyperplane, 2-dimensional plane or even a 1-dimensional point. The last scenario occurs when the vectors are essentially the same and only differ by some noise. One approach to determine the dimensionality of the spanned space is through singular value decomposition (SVD) [77]. By performing SVD on the set of vectors representing the alignments, we obtain the eigen vectors and their associated eigen values. Because the space for the vectors is a 5-dimensional space, we cannot obtain more than five nonzero singular values. However, even for the nonzero singular values, they could be small. When the singular values are smaller than the noise level, the corresponding dimension essentially contains no information and only noise. A better measure to quantify the information content in each dimension is to calculate the Q factor for each eigen vector. Through ~~the~~ SVD, we also obtain the linear combination coefficients for the data set. Using the coefficients, we can construct orthogonal linear combination (OLC) data sets corresponding to each eigen vector. The Q factor can be calculated for each OLC data sets by back-calculate the RDCs from the corresponding Saupe tensor, or the eigen vectors. When the Q factor is less than 0.3, the quality of the data set is generally considered as good. By calculating the Q factor for each OLC data sets, we can count the number of OLC data sets with good quality and quantify the information content in our data.

3.4.2 The information content of the RDC data correlates with the number of orthogonal alignments

For interdomain motions, the Saupe tensors summarize the information content in the RDC data sets [78]. We quantified the information content in the four sets of RDC data using the orthogonal linear combination (OLC) method [77]. The method projects different alignments in the five dimensional Saupe tensor space and finds a linear combination of these to yield a set of orthogonal alignment tensors. The information content of these orthogonal alignments can be quantified from the Q factors for each alignment. The number of orthogonal alignments with low Q factors correlates with the constraining power of the data. By quantifying the information content, we could build a model matching the constraining power of our data and avoid over-fitting. In addition, orthogonal alignments with low information content usually contain a considerable amount of noise. Filtering them out can reduce noise in the data.

Out of four different alignments, we obtained two orthogonal alignments with high information content (Fig. 3.11A). The singular value measures the amplitude of signal in the corresponding alignment while the Q factor measures the signal-to-noise ratio. The first orthogonal alignment has a large singular value of 34.4×10^{-4} and a low Q factor of 0.13, indicating high information content. Although the second orthogonal alignment has a much smaller singular value, its Q factor of 0.22 is still below 0.3, suggesting a much smaller noise and high information content. The third and fourth alignments have low singular values and high Q factors, so they have low information content. Consequently, the first two alignments were used for constraining the conformational model while the rest were discarded. With the two orthogonal alignments (Table 3.3 and Table 3.4), we have a total of 10 independent observables to constrain a model of the interdomain orientational distribution.

Table 3.3: Z domain Saupe tensors for the OLC RDC alignments

Name	$\mathbf{S_{xx} \times 10^4}$	$\mathbf{S_{yy} \times 10^4}$	$\mathbf{S_{zz} \times 10^4}$	$\mathbf{S_{xy} \times 10^4}$	$\mathbf{S_{xz} \times 10^4}$	$\mathbf{S_{yz} \times 10^4}$
OLC1	-8.6551	14.998	-6.3432	-20.564	-0.4184	0.3670
OLC2	5.04792	-2.4341	-2.6137	-1.5583	2.5678	2.2874
OLC3	0.35788	1.6480	-2.0059	0.80805	0.6185	-0.8079
OLC4	-0.0200	-0.1594	0.1794	-0.0918	0.2430	-0.2956

Table 3.4: C domain Saupe tensors for the OLC RDC alignments

Name	$\mathbf{S_{xx} \times 10^4}$	$\mathbf{S_{yy} \times 10^4}$	$\mathbf{S_{zz} \times 10^4}$	$\mathbf{S_{xy} \times 10^4}$	$\mathbf{S_{xz} \times 10^4}$	$\mathbf{S_{yz} \times 10^4}$
OLC1	-0.5211	1.5706	-1.0495	0.7714	0.0267	0.5242
OLC2	0.3574	-0.5473	0.1899	0.0369	-0.0785	-0.4544
OLC3	-0.0115	0.1544	-0.1429	-0.0471	0.1798	0.0169
OLC4	0.1947	-0.0993	-0.0954	-0.0693	-0.1063	-0.1589

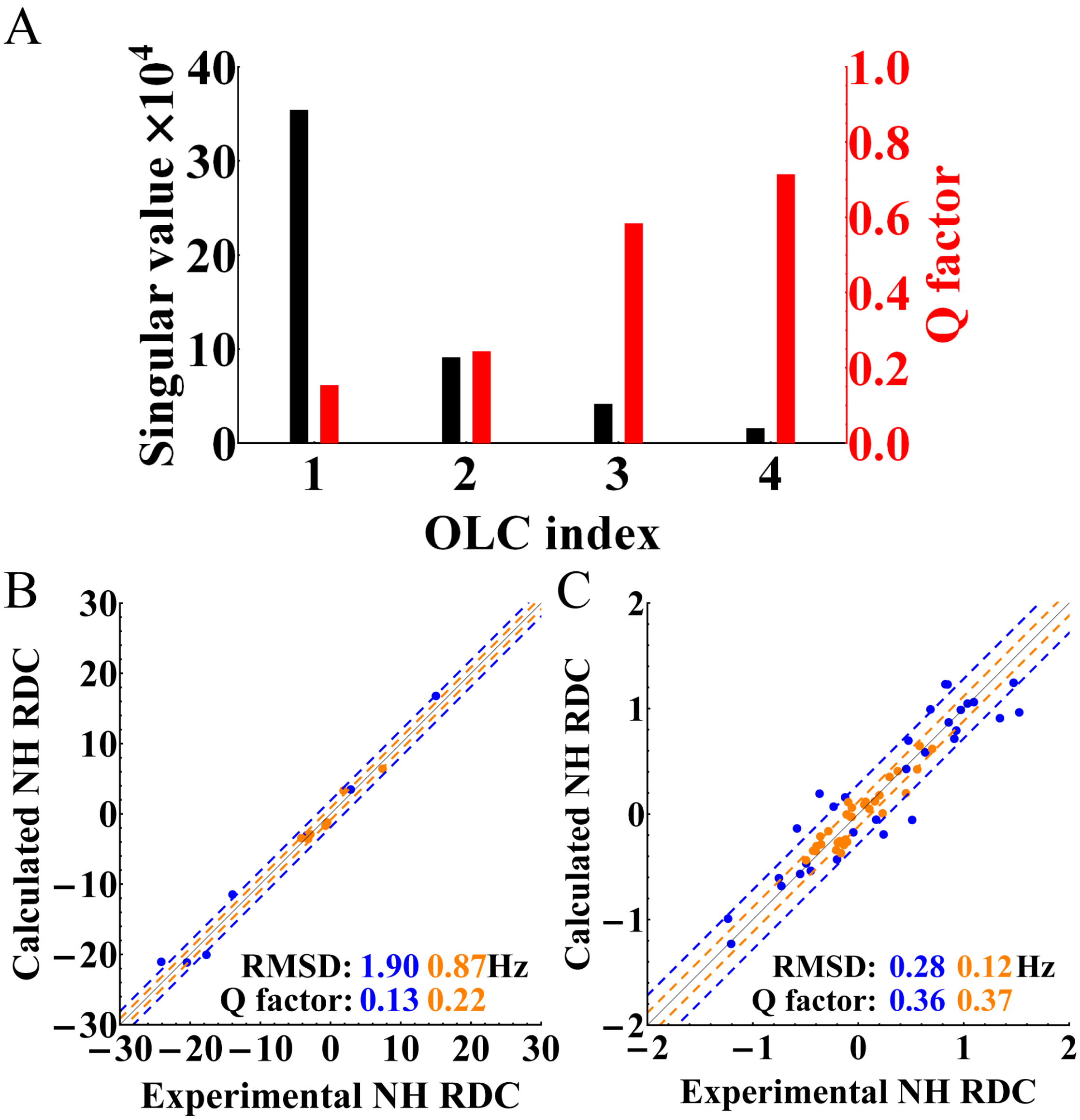


FIGURE 3.11: Orthogonal linear combination(OLC) analysis of the four RDC alignments of SpA-N. (A) Singular values(black bars) and Q factors(red bars) associated with each OLC alignment. (B) The correlation plot of the first OLC dataset(blue) and the second OLC dataset(yellow) of Z domain and the RDCs back-calculated from the two corresponding OLC Saupe tensors of the Z domain. (C) The correlation plot of the first OLC dataset(blue) and the second OLC dataset(yellow) of C domain and the RDCs back-calculated from the two corresponding OLC Saupe tensors of the C domain.

4

Determining interdomain motions of SpA-N as a continuous model

4.1 Modeling interdomain motions as a probability distribution

Based on the discussions in subsection 1.1.1, conformational distributions are good representations of biomolecular dynamics. For inter-domain motions, there are only 3 translational degrees of freedom (DOFs) and 3 orientational DOFs in the absence of significant intradomain dynamics. Because residual dipolar couplings (RDCs) only contain orientational information, we aim to reconstruct the orientational component of the interdomain motions. The obstacle in the reconstruction is the lack of information in RDCs. RDCs do not observe the underlying conformational distribution directly and only report average information, e.g. moments of the conformational distribution. The lack of information makes the reconstruction problem an ill-posed one, which may have multiple or an infinite number of solutions. In order to overcome the ill-pose nature of the problem, reasonable assumption or prior knowledge must be introduced to regularize the problem. The assumption is usually carried by a model. In this section, I summarize models in the previous studies and introduce a continuous model to describe the interdomain motions.

4.1.1 Discrete ensembles versus continuous distributions

Several methods have been developed to reconstruct interdomain motions, including maximum allowed probability (MAP) [39], sample-and-select (SAS) [79] and sparse ensemble selection (SES) [80]. All three methods use a discrete finite ensemble to describe interdomain motions. They select an ensemble of discrete conformers from a preconfigured conformational pool. When conformations in the pool are generated by molecular dynamics simulation or selected by an energy function, the conformations are restricted to regions where the empirical energy is favorable. Because of discrete nature and the use of energy function, all the methods suffer from three disadvantages. First, although the energy function offers regularization, it also introduces empirical assumptions into the result. Unfortunately, the energy function does not have a good ability to predict salient features of unfolded states or unstructured regions, raising doubts concerning the associated assumptions [81]. Second, for large amplitude interdomain motions, the number of discrete conformers required to represent the broad conformational distribution is enormous, which increases the risk of over-fitting. Last but not least, a discrete ensemble assigns certain probabilities to the conformations in the ensemble and assigns zero probability to the rest of the conformational space. Although the ensemble description captures several characteristics of biomolecular motions, the void of probability between structures is physically unreasonable.

On the other hand, continuous distributions can offer good regularizations with a smoothness assumption. The smoothness assumption is a reasonable assumption for many problems. In our case, because a biomolecular domain rotate through the space continuously, the probability distribution should also be continuous. In addition, the probability distribution is dictated by the molecule's surrounding force field. Given no evidence of abrupt change in the force field, the distribution should be smooth. As shown in the following, the smoothness assumption can effectively regularize the problem. The smoothness assumption eliminates the need to weight each conformation individually. Consequently, it reduces the number of parameters in the model and overcomes the over-fitting problem. At last, the smoothness assumption only introduces a small bias into the result, thus preserving the information content in the data. Even when the smoothness assumption is not reasonable, the bias is represented in a predictable manner, making the interpretation of the result more clearly.

4.1.2 Modeling inter-domain orientational motions as a Bingham distribution

The family of Bingham distributions [82] are widely used to describe circular distributions on the 2D sphere S^2 and 3D rotation space $SO(3)$ [83, 84]. Previous studies demonstrated the Bingham model's ability to represent salient features of a broad spectrum of orientational distributions [83, 84]. In the Kunze and Schaeben paper [83], the Bingham distribution was used in texture analysis. It was shown that the Bingham distribution can represent various crystallographic preferred orientation (CPO) patterns. The crystallographic preferred orientation is the relative orientation between a micro-crystal and its embedded mineral stone. The real distribution here is also a distribution of orientations between two rigid bodies, which is same as in our case. In the other paper [84], the Bingham distribution was used to represent oriented local features on an object. Both studies show that the Bingham distribution is general enough to describe a broad spectrum of distributions on the 3D rotation space $SO(3)$. The Bingham distributions are also the maximum entropy distributions on the hypersphere given the inertia matrix. On the other hand, the assumptions carried by the Bingham distribution on $SO(3)$ should be noted. The Bingham distribution is a unimodal distribution with certain ~~type of~~ curvatures. The uni-modality is one bias carried by the Bingham distribution model. In addition, because the Bingham distribution is also the maximum entropy distribution given the inertia matrix, it enforces the smoothness assumption. High frequency oscillations are excluded from the distribution. Nonetheless, we believe the Bingham distribution is a good starting point for modeling the interdomain motions and the assumptions are reasonable.

Consequently, we model the interdomain orientational distribution as the Bingham distribution on the 3D rotation space $SO(3)$, which takes the following form [82, 83]:

$$P(\tilde{\mathbf{q}}\,;\,X) = c^{-1}(X)\exp(\tilde{\mathbf{q}}^T X\tilde{\mathbf{q}}). \tag{4.1}$$

In Eq. (4.1), $c^{-1}(X)$ is the normalization factor. $P(\tilde{\mathbf{q}}\,;\,X)$ is the probability of $\tilde{\mathbf{q}}$ given X. $q \in SO(3)$ is a rotation and $\tilde{\mathbf{q}}$ is the 4D unit quaternion representation of q. X is a symmetric 4×4 matrix with a constant trace and thus 9 DOFs. It should be noted that the expression above is generally for a distribution on S^3 because $\mathbf{q} \in S^3$. However, because quaternions $\mathbf{q}$ and $-\mathbf{q}$ represent the same rotation and S^3 is a double cover of $SO(3)$, we can limit the domain of the distribution function to $SO(3)$ and normalize it to obtain the distribution on $SO(3)$. The Bingham distribution on S^3 is antipodally symmetric, so the Bingham distribution on $SO(3)$ is unimodal. Fig. **??** shows that the Bingham distribution on S^2 sphere is antipodally symmetric. The Bingham distribution on S^3 has the same feature but can not be directly visualized. The meaning of the 9 DOFs becomes more clear if we decompose X in the following way:

$$X = M^T \Lambda M. \tag{4.2}$$

Here, M is a rotation matrix belonging to the group $SO(4)$ and Λ is a diagonal matrix with a constant trace specifying the variances along the four principle directions. M contains 6 DOFs and Λ contains 3 DOFs. The rotation matrix $M \in SO(4)$ can be

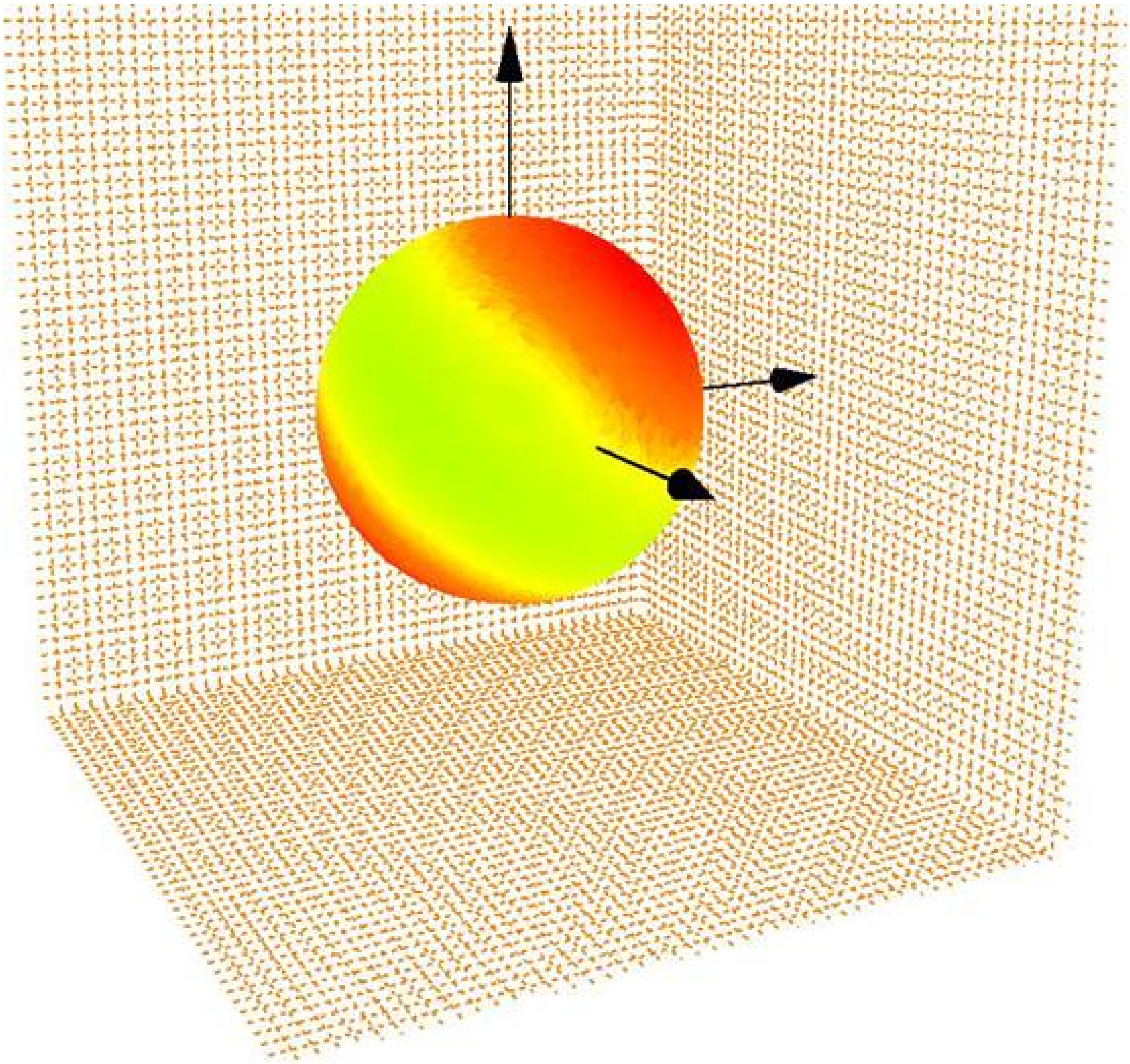


FIGURE 4.1: A Bingham distribution on S^2. The two modes on the opposite position of the sphere have relatively high probability and are colored red.

further decomposed into a left isoclinic rotation M^L and a right isoclinic rotation M^R:

$$M = M^L M^R. \tag{4.3}$$

The left isoclinic rotation M^L corresponds to a left quaternion rotation $q_{_L}$, the right isoclinic rotation M^R corresponds to a right quaternion rotation $q_{_R}$. Consequently, the 9 DOFs in the Bingham distribution can be separated into three matrices, Λ, M^L and M^R.

4.2 Using the branch-and-bound algorithm to find the optimal solution

Based on the Bingham model, our problem of determining the interdomain orientational motions becomes the problem of finding the best Bingham distribution satisfying the data. As shown in the previous subsection 4.1.2, the Bingham distribution has 9 DOFs. Finding the best Bingham distribution is a search problem on a 9 dimensional space. Because the evaluation cost for each point in the 9D space involves an integration, systematic sampling of the 9D space is too expensive. Instead, I apply a divide-and-conquer strategy by using a branch-and-bound algorithm. The section starts with an introduction to branch-and-bound algorithms. Then I give a bird's-eye view of the algorithm without going through the mathematical proofs and present the implemented software package of the algorithm. For readers who favors a mathematical description of the algorithm, the complete proofs can be found in my master's thesis [85]. At last, I present the best solutions for the RDC data of SpA-N.

4.2.1 Introduction to branch-and-bound algorithms

Branch-and-bound algorithms are usually applied to discrete and combinatorial optimization problems, but can be generalized for continuous optimization problems.

It employs the divide-and-conquer strategy and divides the original problem into smaller and smaller subproblems. The idea is that smaller problems can be solved more easily. The cost of solving all subproblems is less than the cost of solving the original problem. As the branch-and-bound name suggests, the algorithm has two steps, a branch step and a bound step. In the branch step, a problem is divided into smaller subproblems. In the bound step, the bounds of the subproblems are calculated. ~~The subproblems are usually optimization problems which search for the minimum or maximum of the original problem.~~ Without losing generality, we can assume the original problem is a minimization problem. In this case, when the lower bound of a subproblem is higher than the upper bound of the any other subproblem. The search region corresponding to the first subproblem is guaranteed to not contain the global minimum. The branch and bound steps are usually followed by a pruning step, which prunes the non-minimum-containing regions and excludes them from subsequent searches. The three steps are repeated for multiple rounds until that ~~only one point remains or that~~ the subproblem region is small enough to be considered as a point. The remaining point or region is the global minimum of the space. Branch-and-bound algorithms are provable in the sense that they guarantee to find the global minimum. They are also heuristic in the sense that the computation efficiency depends highly on the tightness of the bounds. In the worst case scenario, the time complexity of branch-and-bound algorithms is exponential. However, in practice, the algorithms usually excel an exponential algorithm. The speed-up comes from the bound and pruning steps. The calculation of bounds must be a fast computation compared to a systematic sampling in the region. When the bounds are tight, pruning is efficient and the search time on pruned regions is saved.

4.2.2 A brief description of the branch-and-bound algorithm for fitting RDC data

Because the magnetic alignment method provides motional decoupling (subsection), the alignment tensor of the first domain S_{I} is only resulted from global tumbling. The alignment transfers from the first domain through the flexible interdomain linker to the second domain. So the alignment tensor of the second domain S_{II} depends on both global tumbling and interdomain motions. The relationship between the vector form of the two Saupe tensors can be formulated as:

$$\mathbf{s}_{\text{II}} = \mathbf{E}[Q] \cdot \mathbf{s}_{\text{I}}. \tag{4.4}$$

The above equation decomposes the global tumbling and the interdomain motions into two parts. $\mathbf{s}_{\text{I}}$ contains information about global tumbling and $\mathbf{E}[Q]$ contains information about interdomain motions. When one or multiple pairs of $\mathbf{s}_{\text{I}}$ and $\mathbf{s}_{\text{II}}$ are present, the matrix $\mathbf{E}[Q]$ can be partial determined. With five orthogonal $\mathbf{s}_{\text{I}}$ and their corresponding $\mathbf{s}_{\text{II}}$, the matrix $\mathbf{E}[Q]$ can be fully determined. When the matrix $\mathbf{E}[Q]$ is either partially or fully determined, we obtain information about interdomain motions. The expression of the matrix $\mathbf{E}[Q]$ is the following:

$$\mathbf{E}_{R\sim p}[Q] = \begin{bmatrix} \langle x_1^2 - x_3^2 \rangle & \langle x_2^2 - x_3^2 \rangle & \langle 2x_1x_2 \rangle & \langle 2x_1x_3 \rangle & \langle 2x_2x_3 \rangle \\ \langle y_1^2 - y_3^2 \rangle & \langle y_2^2 - y_3^2 \rangle & \langle 2y_1y_2 \rangle & \langle 2y_1y_3 \rangle & \langle 2y_2y_3 \rangle \\ \langle x_1y_1 - x_3y_3 \rangle & \langle x_2y_2 - x_3y_3 \rangle & \langle x_1y_2 + x_2y_1 \rangle & \langle x_1y_3 + x_3y_1 \rangle & \langle x_2y_3 + x_3y_2 \rangle \\ \langle x_1z_1 - x_3z_3 \rangle & \langle x_2z_2 - x_3z_3 \rangle & \langle x_1z_2 + x_2z_1 \rangle & \langle x_1z_3 + x_3z_1 \rangle & \langle x_2z_3 + x_3z_2 \rangle \\ \langle y_1z_1 - y_3z_3 \rangle & \langle y_2z_2 - y_3z_3 \rangle & \langle y_1z_2 + y_2z_1 \rangle & \langle y_1z_3 + y_3z_1 \rangle & \langle y_2z_3 + y_3z_2 \rangle \end{bmatrix}, \tag{4.5}$$

where $R \in SO(3)$ has a distribution p and is parameterized as:

$$R = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix}. \tag{4.6}$$

Based on the discussions in subsection 4.1.2, the 9 parameters in the Bingham model can be separated into three matrices. Each one of the three matrices contains 3 parameters. We use the Bingham distribution to model the interdomain motions, so the matrix $\mathbf{E}[Q]$ is also a function of the 9 parameters. Inspired by the decomposition showing in Eq. 4.2 and Eq. 4.3, the 9 parameters in $\mathbf{E}[Q]$ can also be separated into three matrices by decomposing $\mathbf{E}[Q]$:

$$\mathbf{E}[Q] = Q(\tilde{\mathbf{q}_{\mathbf{R}}}^{-1}) \cdot \mathbf{E}[Q\,;\,\Lambda] \cdot Q(\tilde{\mathbf{q}_{\mathbf{L}}}^{-1}), \tag{4.7}$$

where $\tilde{\mathbf{q}_{\mathbf{R}}}^{-1}$ is the quaternion corresponding to M^R and $\tilde{\mathbf{q}_{\mathbf{L}}}^{-1}$ is the quaternion corresponding to M^L. The three matrices $Q(\tilde{\mathbf{q}_{\mathbf{R}}}^{-1})$, $\mathbf{E}[Q\,;\,\Lambda]$ and $Q(\tilde{\mathbf{q}_{\mathbf{L}}}^{-1})$ each contains three parameters. Here, $\Lambda \in \mathbb{R}^3$, $\tilde{\mathbf{q}_{\mathbf{R}}} \in S^3$ and $\tilde{\mathbf{q}_{\mathbf{L}}} \in S^3$. Furthermore, the product of the three matrices $\mathbf{E}[Q]$ transforms $\mathbf{s}_{\text{I}}$ into $\mathbf{s}_{\text{II}}$ and the three matrices can be considered as three operators which applied three consecutive transformations (Fig. 4.2). Indeed, we observe the matrices $Q(\tilde{\mathbf{q}_{\mathbf{R}}}^{-1})$ and $Q(\tilde{\mathbf{q}_{\mathbf{L}}}^{-1})$ are rotation operators which only alters the orientation of the input Saupe tensor. The other matrix $\mathbf{E}[Q\,;\,\Lambda]$ is a averaging operator which has both scaling and rotation effects. Based on the observations, I designed a branch-and-bound algorithm that focuses on the scaling effect of the transformation. The scaling of the Saupe tensor only happens in the averaging operation, but it also depends on the first rotation operation $Q(\tilde{\mathbf{q}_{\mathbf{L}}}^{-1})$. In the first step, the algorithm branches the $\mathbb{R}^3$ space of Λ and bounds the branched spaces (Fig. 4.3). The bounds are on the magnitude of the largest principle component of the transformed Saupe tensor. Because the scaling effect also depends on the first rotation operation $Q(\tilde{\mathbf{q}_{\mathbf{L}}}^{-1})$, the bounds are calculated by systematically sampling the S^3 space of $\tilde{\mathbf{q}_{\mathbf{L}}}$. The calculation of the bounds does not require a systematic sampling on $\mathbb{R}^3$ because of the monotonicity. We observe that the range between the upper bound and the lower bound of a region should cover the magnitude of the largest principle component of the second Saupe tensor S_{II}. Otherwise, the region must not contain the best solution and it is subsequently pruned. After the first step, a set of remaining regions in the $\mathbb{R}^3$ space of Λ are saved. Because the regions

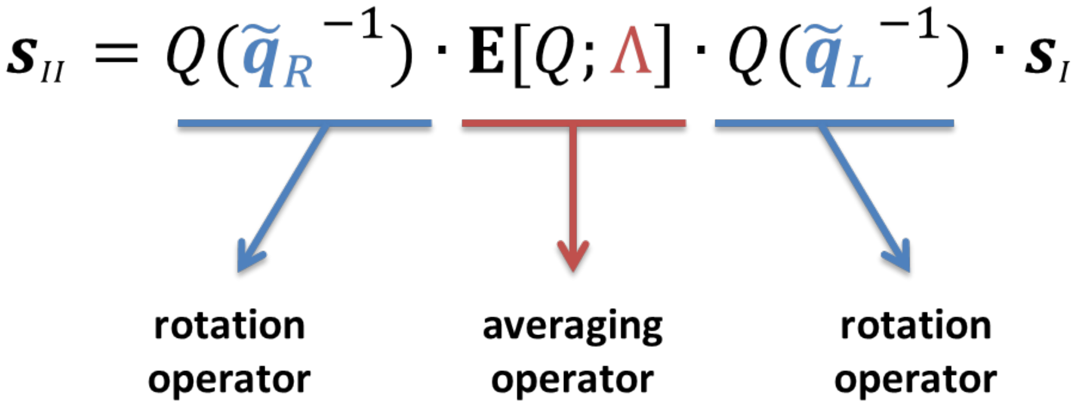

$s_{II} = Q(\tilde{q}_R{}^{-1}) \cdot \mathbf{E}[Q;\Lambda] \cdot Q(\tilde{q}_L{}^{-1}) \cdot s_I$
rotation operator
averaging operator
rotation operator

Figure 4.2: The three operators generated by matrix decomposition. The left and right operators are both rotation operators. The middle operator is an averaging operator. The three operators are applied on the Saupe tensor subsequently from right to left.

are small regions, the middle points of the regions are used to represent the regions.

In the second step, the S^3 space of $\tilde{\mathbf{q}}_{\mathbf{L}}$ are uniformly divided into small regions for each remaining point in the $\mathbb{R}^3$ space of Λ (Fig. 4.4). Again, bounds on the magnitude of the largest principle component of the transformed Saupe tensor are calculated. Here, the divided regions are small so the bounds are tight. As the first step, regions with bounds covering the target magnitude is saved. Other regions are pruned. The middle points of the regions are also used to represent the regions.

After the first two steps, we have a list of point pairs which can correctly transform the magnitude of the largest principle component. Each pair is composed of a point on the $\mathbb{R}^3$ space of Λ and a point on the S^3 space of $\tilde{\mathbf{q}}_{\mathbf{L}}$. From the two points, the orientation of the transformed Saupe tensor can be calculated. In the third step, the point on the S^3 space of $\tilde{\mathbf{q}}_{\mathbf{R}}$ can be directly calculated by comparing the orientation of the transformed Saupe tensor and the orientation of the target Saupe tensor, S_{II}

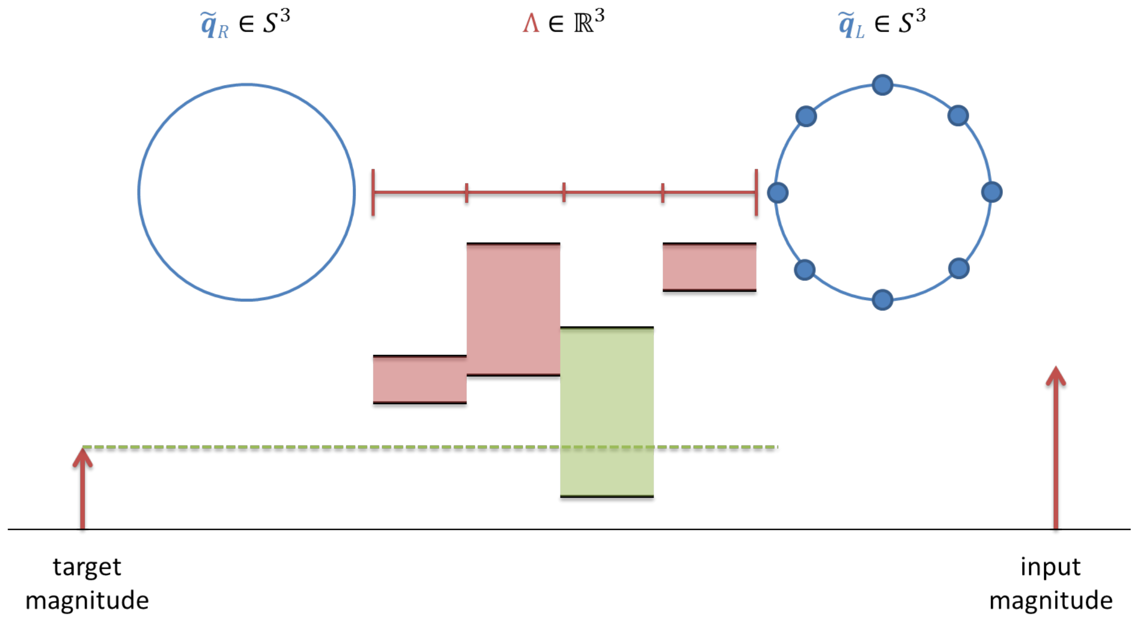


FIGURE 4.3: The first step in the branch-and-bound algorithm. For illustration, the $\mathbb{R}^3$ space is branched into 4 regions. The upper and lower bound for each region are calculated by sampling the right-side S^3 space. The red arrow on the right represents the magnitude of the largest principle component of the first Saupe tensor S_{I}. The red arrow on the left represents the magnitude of the largest principle component of the second Saupe tensor S_{II}. Regions (green) covering the target magnitude is saved for the next round of branch-and-bound. Other regions (red) are pruned.

(Fig. 4.5).

After the three steps, we have a list of point triples. All of them can correctly transform the magnitude of the largest principle component and the orientation of the Saupe tensor. However, Saupe tensors have 5 DOFs and the triples may not transform the magnitudes of other principle components. In order to fully evaluate the transformation and find the best fit triple, the following objective function is

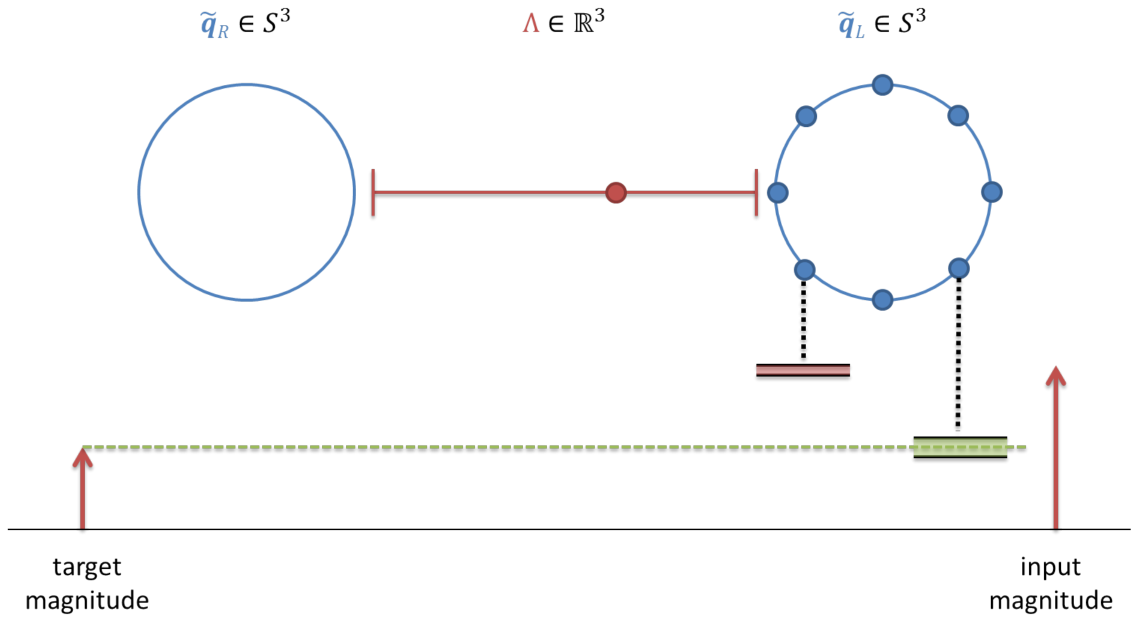


FIGURE 4.4: The second step in the branch-and-bound algorithm. For illustration, the S^3 space (on the right) is branched into 8 regions. The upper and lower bound for each region are calculated for each pair of points on the $\mathbb{R}^3$ space and the S^3 space . The red arrow on the right represents the magnitude of the largest principle component of the first Saupe tensor S_{I}. The red arrow on the left represents the magnitude of the largest principle component of the second Saupe tensor S_{II}. Regions (green) covering the target magnitude is saved for the next round of branch-and-bound. Other regions (red) are pruned.

used:

$$f(\mathbf{x}) = \sum_i (\mathbf{E}[Q|X] \cdot \mathbf{s}_{\text{I}_\text{i}} - \mathbf{s}_{\text{II}_\text{i}})^2. \tag{4.8}$$

The triples are sorted by their corresponding objective function values. The triple with the smallest value corresponds to the best-fit Bingham distribution.

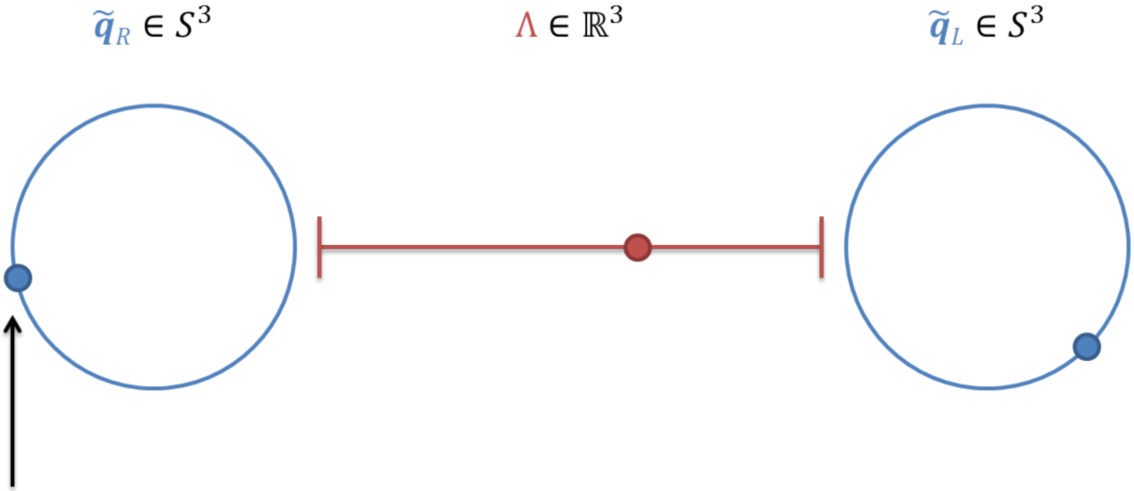


FIGURE 4.5: The third step in the branch-and-bound algorithm. The point on the S^3 space of $\tilde{\mathbf{q}_\mathbf{R}}$ can be directly calculated.

4.2.3 The CDIO software package

The above algorithm is implemented as a MATLAB software package for determination of continuous distribution of interdomain orientations (CDIO). Underlying the MATLAB code of CDIO, there are two key components, a continuous model that describes a broad spectrum of interdomain motions and a *de novo* branch-and-bound algorithm that guarantees to find the best-fit. CDIO is applicable to the study of interdomain motions in protein, RNA or DNA molecules. CDIO only requires two alignment tensors of each domain calculated either from residual dipolar couplings or pseudo-contact shifts to determine the best-fit interdomain orientation distribution.

The main function of the CDIO program is *BnBmain*. It depends on three helper functions: *r3prune*, *rsprune* and *enum*. For the first step, functions *branch* and *iBranch* branches the space of Λ and divides it into smaller regions, then *r3prune* calculates bounds for each region and prunes. For the second step, *rsprune* branches the space of $\tilde{\mathbf{q}}_{\mathbf{R}}$, bounds the branched regions and prunes. For the third step, *enum* both calculates the point on the space of $\tilde{\mathbf{q}}_{\mathbf{L}}$ for each remaining pair and evaluates the objective function 4.8 for the triple. The main function *BnBmain* sorts the values and outputs the parameters of the best-fit Bingham distribution.

CDIO requires MATLAB to run, but it does not require any MATLAB toolboxes. MATLAB R2013 or any later version should be sufficient. To install MATLAB, please follow the official instructions(`http://www.mathworks.com/help/install/ug/install-mathworks-software.html`).

In the CDIO distribution, the *mhg* function and its associated files are included. The copyright of the the *mhg* function belongs to Dr. Plamen Koev in San Jose State University [86]. You may want to recompile the *mhg.c* file before running. To compile the *mhg.c* file, type the following command at the MATLAB prompt:

```
mex mhg.c
```

The inputs to the main function *BnBmain* are two 2 × 5 matrices containing a total of four Saupe tensors. Given the Saupe tensor of the reference domain (Domain I) from two independent alignments, S^1_{I} and S^2_{I} and the Saupe tensors of the other domain (Domain II) from the two alignments, S^1_{II} and S^2_{II}. The Saupe tensors are formatted as in Eq. 3.6. The inputs **s1** and **s2** are formatted as:

$\mathbf{s1} = [S^1_{\mathbf{I}}\mathbf{(1,1)},\ S^1_{\mathbf{I}}\mathbf{(2,2)},\ S^1_{\mathbf{I}}\mathbf{(1,2)},\ S^1_{\mathbf{I}}\mathbf{(1,3)},\ S^1_{\mathbf{I}}\mathbf{(2,3)};S^2_{\mathbf{I}}\mathbf{(1,1)},\ S^2_{\mathbf{I}}\mathbf{(2,2)},\ S^2_{\mathbf{I}}\mathbf{(1,2)},\ S^2_{\mathbf{I}}\mathbf{(1,3)},\ S^2_{\mathbf{I}}\mathbf{(2,3)}];$

$\mathbf{s2} = [S^1_{\mathbf{II}}\mathbf{(1,1)},\ S^1_{\mathbf{II}}\mathbf{(2,2)},\ S^1_{\mathbf{II}}\mathbf{(1,2)},\ S^1_{\mathbf{II}}\mathbf{(1,3)},\ S^1_{\mathbf{II}}\mathbf{(2,3)};\ S^2_{\mathbf{II}}\mathbf{(1,1)},\ S^2_{\mathbf{II}}\mathbf{(2,2)},\ S^2_{\mathbf{II}}\mathbf{(1,2)},$

S^2_{II}**(1,3),** S^2_{II}**(2,3)];**

To run the program, type the following command at the MATLAB prompt:

BnBmain(s1,s2)

When no input arguments or incorrect arguments are supplied, the program sets the input as the SpA-N data automatically. If you only want to reproduce the SpA-N result, drop the arguments and simply type:

BnBmain

The CDIO software package is the main component of our continuous approach to reconstruct interdomain orientation distributions from RDCs. Before using the CDIO software package, calculation of Saupe tensors is required. Orthogonalization of the Saupe tensors using the OLC method [77] is highly recommended. Our complete approach is summarized in Fig. 4.6.

4.2.4 The best solutions for SpA-N RDC data

By applying the above method and using the CDIO software package, two solutions are found that can reproduce the RDC data nearly equally well. The two solutions have similar objective function values. In order to further validate the two solutions, RDC correlation plots are plotted for experimental RDC data and predicted RDC data of each solution. Good agreements are observed for both solutions, as shown in Fig. 4.7. The RMSD and Q factor of solution 1 for the first OLC dataset is 0.28 Hz and 0.36, respectively. The RMSD and Q factor of solution 1 for the second OLC dataset is 0.13 Hz and 0.40, respectively. The RMSD and Q factor of solution 2 for the first OLC dataset is 0.29 Hz and 0.37, respectively. The RMSD and Q factor of solution 2 for the second OLC dataset is 0.13 Hz and 0.42, respectively. In summary, the RMSDs and Q factors of solution 1 do not differ much from those of solution 2.

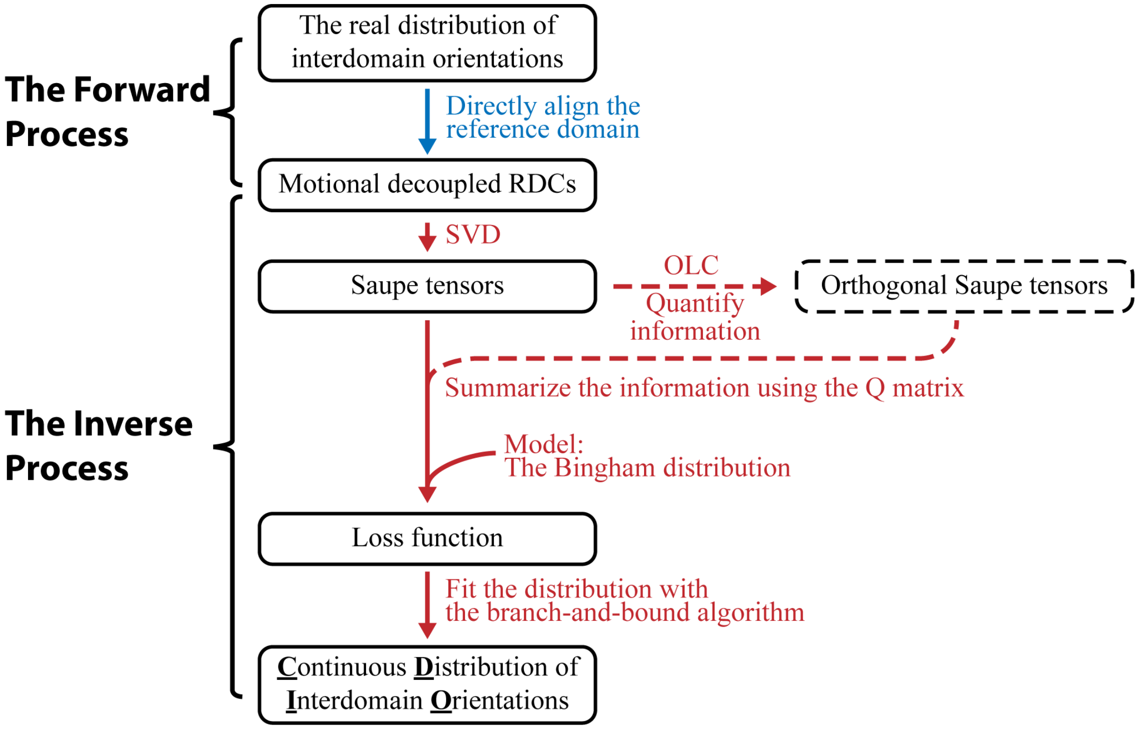

The Forward Process
The Inverse Process
The real distribution of interdomain orientations
Directly align the reference domain
Motional decoupled RDCs
SVD
Saupe tensors
OLC
Quantify information
Orthogonal Saupe tensors
Summarize the information using the Q matrix
Model: The Bingham distribution
Loss function
Fit the distribution with the branch-and-bound algorithm
Continuous Distribution of Interdomain Orientations

FIGURE 4.6: A flow chart of the experimental and computational steps used to calculate the continuous distribution of interdomain orientations. The loss function is the objective function in Eq. 4.8

The two solutions are two best solutions for the SpA-N RDC data.

4.3 Visualization methods

4.3.1 Visualization by projection

The challenge of visualizing orientation probability distributions is that they are joint in the three DOFs of $SO(3)$. Displaying the correlation between the DOFs requires a 4 dimensional representation. However, a direct visualization of a representation beyond 3D is not possible. One way to visualize the high dimensional distribution is to project it to a low dimensional space and visualize the low dimensional projection. It should be noted that information may lost or may be distorted during the projection. Fortunately, the rotation space $SO(3)$ only has three DOFs. It is essentially a

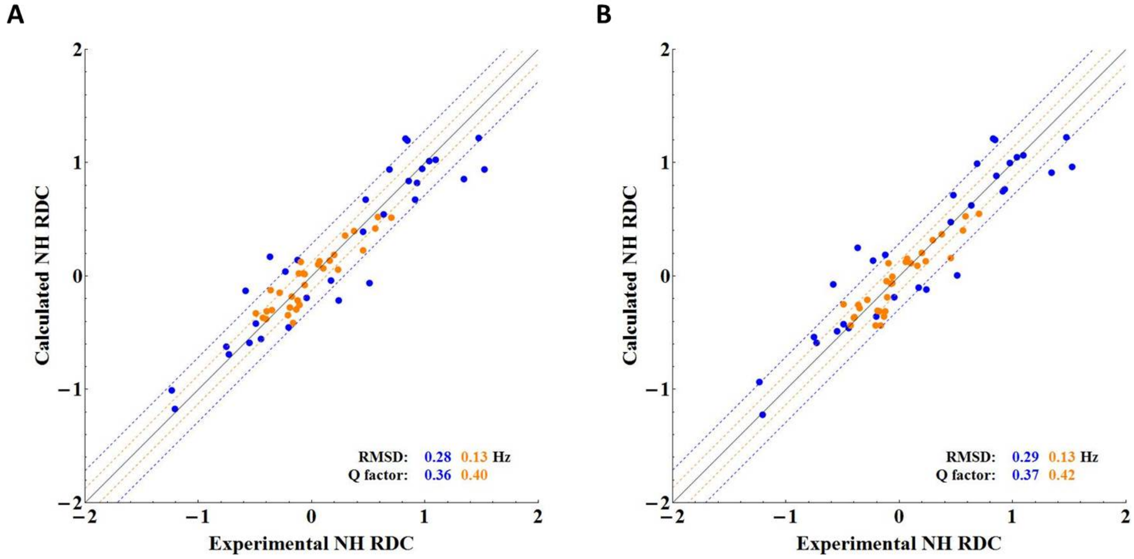


FIGURE 4.7: RDC correlation plots. The correlation plot of the first OLC dataset(blue) and the second OLC dataset(yellow) of Z domain and the RDCs back-calculated from solution 1(A) and solution 2(B).

3 dimensional circular surface which naturally resides in a 4D space. Consequently, the rotation space $SO(3)$ can be 'squeezed' into a 3D space by mapping the points in the rotation space $SO(3)$ to the points in the 3D space $\mathbb{R}^3$. It should be noted that no mapping can preserve the measure on the rotation space $SO(3)$ and that the 'squeezing' operation does distort the space to a certain extend. The goal of the following representation is not to eliminate, but to minimize distortions.

Projecting a high dimensional circular space into a low dimensional real space and preserving the features of the high dimensional space is not an easy problem. The most famous example for this kind of projection is to project the global surface of earth onto a flat map. The Mercator projection directly projects the latitude and longitude onto x and y dimension of a map. The projection conserves angles and preserves the shape of small objects on the map. However, it is also known for its distortion on object size. An object close to poles are represented much larger than the representation of the same object when it is near the equator. For example, the Greenland is represented larger than the Africa. In fact, the Africa is three times larger than the Greenland. Although a point on $SO(3)$ can be parameterized as three Euler angles. A direct mapping of the Euler angles to the $\mathbb{R}^3$ space poses the caveat as the Mercator projection. Instead, I propose two alternatives in the following. The two presentations, the frame axis-angle (FAA) representation and the disk-on-sphere (DoS) representation, minimize the distortion in the projection and preserve most features of a distribution on the rotation space $SO(3)$.

Frame axis-angle(FAA) representation.

In the frame axis-angle (FAA) representation [87], each inter-domain orientation can be specified by the direction **e** of the second domains z-axis and the rotation angle γ around this z-axis. A FAA vector is used to represent an inter-domain orientation with a direction **e** (given by α and β) and the magnitude as γ (see Fig. 4.8B). Because the heads of the vectors are set as the origin of the coordinate system, the tails of the vectors are inside or on the sphere with a radius of 2π. The tails are color coded to represent their associating probabilities. The transformation from a quaternion to its corresponding FAA vector in the frame axis-angle representation is achieved by:

$$v_{_{FAA}} = (\text{atan2}(\sin(\gamma), \cos(\gamma)) + \pi)[2q_1q_3 + 2q_2q_4, -2q_1q_2 + 2q_3q_4, q_1^2 - q_2^2 - q_3^2 + q_4^2] \quad (4.9)$$

where $v_{_{FAA}}$ is the FAA vector; atan2 is the four-quadrant inverse tangent, which yields values between $-\pi$ and π; the unit quaternion is $q = [q_1, q_2, q_3, q_4]$; and the expressions of $\cos(\gamma)$ and $\sin(\gamma)$ are listed below.

$$\cos(\gamma) = \frac{2q_1q_3 - 2q_2q_4}{\sqrt{1 - (q_1^2 - q_2^2 - q_3^2 + q_4^2)^2}} \quad (4.10)$$

$$\sin(\gamma) = \frac{2q_1q_2 + 2q_3q_4}{\sqrt{1 - (q_1^2 - q_2^2 - q_3^2 + q_4^2)^2}} \quad (4.11)$$

The FAA representation preserves the features of the rotation space $SO(3)$ better than an Euler angle representation. The joint α and β distribution at a given γ,

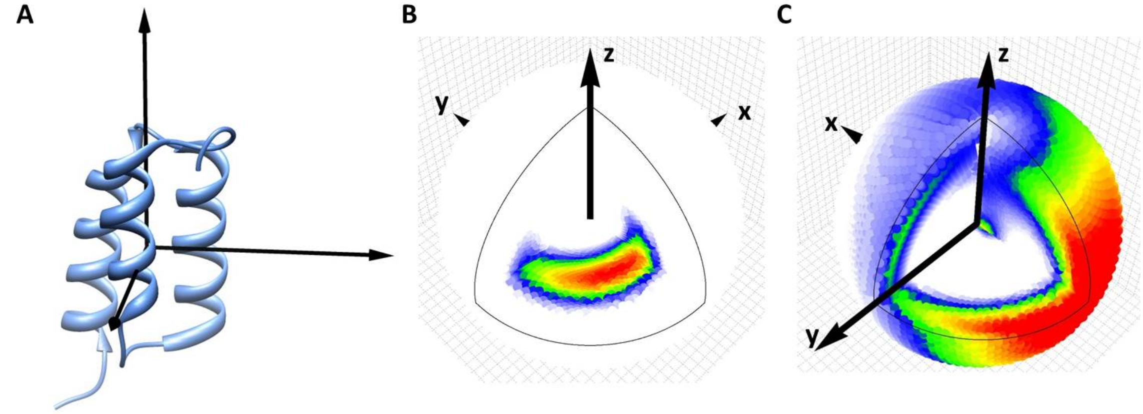


FIGURE 4.8: Calculated inter-domain orientation distribution. (A) The molecular frame of either domain in the presence of a single domain structure. The first(B) and the second(C) best calculated inter-domain orientation distributions in the FAA representation.

$P(\alpha, \beta \,|\, \gamma)$, can be plotted on a spherical surface without distortion. However, because the rotation space $SO(3)$ naturally resides in 4D space, any representation squeezing the rotation space $SO(3)$ into 3D space is inherently distorted. In the FAA representation, distortion occurs along the radial direction. Ideally, any inner sphere with a radius smaller than 2π should have the same surface area as the sphere with a radius of 2π, but of course this is impossible in a static 3D representation. Another distortion in the static representations shown in Fig. 4.8 is that the inner sphere with an infinitesimal radius ϵ should be adjacent to the sphere with a radius of 2π. Despite these distortions, the the FAA representation expresses most features of the rotation space $SO(3)$ and distributions in $SO(3)$ as well.

4.3.2 The disk-on-sphere(DoS) visualization method

In the disk-on-sphere representation, each interdomain orientation can be specified by the direction $\mathbf{e}_z$ of the second domain's z-axis and the direction $\mathbf{e}_x$ of the second domain's x-axis. Location of the center of a disk on a 2-sphere corresponds to $\mathbf{e}_z$, while the direction of a radial line on the disk corresponds to $\mathbf{e}_x$. The radial lines are color coded to represent their associated probabilities. The transformation from a quaternion to its corresponding vectors, $\mathbf{e}_z$ and $\mathbf{e}_x$, is achieved by:

$$\mathbf{e}_z = (2q_1q_3 + 2q_2q_4, \ -2q_1q_2 + 2q_3q_4, \ q_1^2 - q_2^2 - q_3^2 + q_4^2) \tag{4.12}$$

$$\mathbf{e}_x = (q_1^2 + q_2^2 - q_3^2 - q_4^2, \ 2q_2q_3 + 2q_1q_4, \ -2q_1q_3 + 2q_2q_4). \tag{4.13}$$

The disk-on-sphere representation preserves the features of the rotation space $SO(3)$ better than an Euler angle representation. The distribution of $\mathbf{e}_z$ can be plotted on a spherical surface without distortion. In our case to draw it, the distribution of $\mathbf{e}_z$ is discretized on the sphere to incorporate disks representing the conditional distribution $p(\mathbf{e}_x \,|\, \mathbf{e}_z)$, the disk-on-sphere representation expresses most features of not only the rotation space $SO(3)$ but also distributions on $SO(3)$. By zooming in, a finite discretization of $\mathbf{e}_z$ can be calculated and drawn from the underlying continuous distribution at any desired resolution.

4.4 The interdomain orientation distribution of SpA-N

To describe the interdomain orientation, it is first necessary to define the molecular frame of each domain. In the coordinate frame of either Z or C domain, the z-axis of the domain is parallel to the helices and points toward the N-termini of helices 1 and 3 , the y-axis is in the plane of helix II and III, and the x-axis is perpendicular to the plane (Fig. 4.9a,c).The 3D joint interdomain orientational distribution can be visualized in a *disk-on-sphere (DoS)* representation (Fig. 4.9). In this representation, the x, y, z-axes in Fig. 4.9 represent the coordinate frame of the reference domain (Z domain). The joint probability of a particular interdomain orientation, displayed as a line segment representing the x' axis of the C domain on a disk whose position on the sphere is determined by the orientation of the z' axis, is represented by the color of the line segment. The color scale for the probability, in terms of percentile, is shown by the legend at the bottom of the figure. Note that the line segment color on each disk represents the joint probability of both the z' axis orientation and a particular rotation around it.

With the two orthogonal alignments generated by the OLC method, we applied the algorithm to the SpA-N di-domain system and found two CDIO model solutions that can reproduce the RDC data nearly equally well (Fig. 4.7). A priori, it is possible that any linear combination of the two solutions could also fit the data. However, when we simulated a structural ensemble of the B-C di-domain molecule in the absence of any binding partner using the RANCH component of the EOM package [88], a strong disagreement was found between the first solution described above and the structural ensembles generated by RANCH. The DoS plot of the simulation shows a large void of probability in the white area (Fig. 4.10), suggesting that most conformers in this area have steric clashes. The distribution mode of the first solution largely coincides with this low probability region and 37.3% of its probability falls into the region (Fig. 4.10). Quantitatively, we defined a clash score based on the probability of a distribution in the low probability region and calculated clash scores for both solutions. Given a low probability region defined by a simulation, the probability of the region is defined as the floor probability p_{floor}. The floor probability represents the reasonable amount of probability existing in the region. In addition, define a ceil probability p_{ceil} as the largest probability drawn from an equal volume region in the simulated distribution. The clash score for a distribution is calculated as following:

$$score_{dist} = \frac{p_{dist} - p_{floor}}{p_{ceil} - p_{floor}} \tag{4.14}$$

where p_{dist} is the probability of a distribution falling within the low probability region. The clash score of the first solution is more than two fold higher than that of the second solution. By thermodynamic criteria, the simulated ensemble is also in better agreement with the second solution than the first solution or any linear combination of the two. These results suggest that the first solution is a so-called "ghost" solution [39], but further studies with additional orthogonal alignments are necessary to definitively rule it out.

In the second solution, the most probable interdomain orientation has the z'-axis of C domain close to the minus x-axis of Z domain and the x'-axis of C domain close to the minus z-axis of Z domain. It should be noted that the marginal distribution of the z'-axis is relatively broad but the marginal distribution of the x'-axis is narrow. In Fig. 4.9c, the region bounded by the green color takes 9.7% of the entire interdomain orientational space and it has a probability of 40%. The joint distribution suggests that the flexible linker enables the two domains to sample a relatively large range of the interdomain orientational space but somewhat restricts the orientation of the C domain's x'-axis.

4.5 Calculating conformational binding thermodynamics from continuous distributions

4.5.1 Theory

When a molecule binds its binding partner, its conformational distribution changes. Given the conformational probability distribution $P_1(x)$ in the absence of binding partners and the distribution $P_2(x)$ in the presence of a binding partner, thermodynamic parameters (ΔF, ΔU and ΔS) can be calculated assuming that a probability in the reference distribution $P_1(x)$ and its corresponding state energy $E_1(x)$ satisfies the Boltzmann distribution relationship:

$$P_1(x) \propto \exp(-\frac{E_1(x)}{RT}), \tag{4.15}$$

Here, we only focus on the intra-molecular energetic contributions to the thermodynamics (Fig. 4.11). The intra-molecular energetic contribution of a molecule comes from the switch from its pre-binding conformational distribution to its post-binding

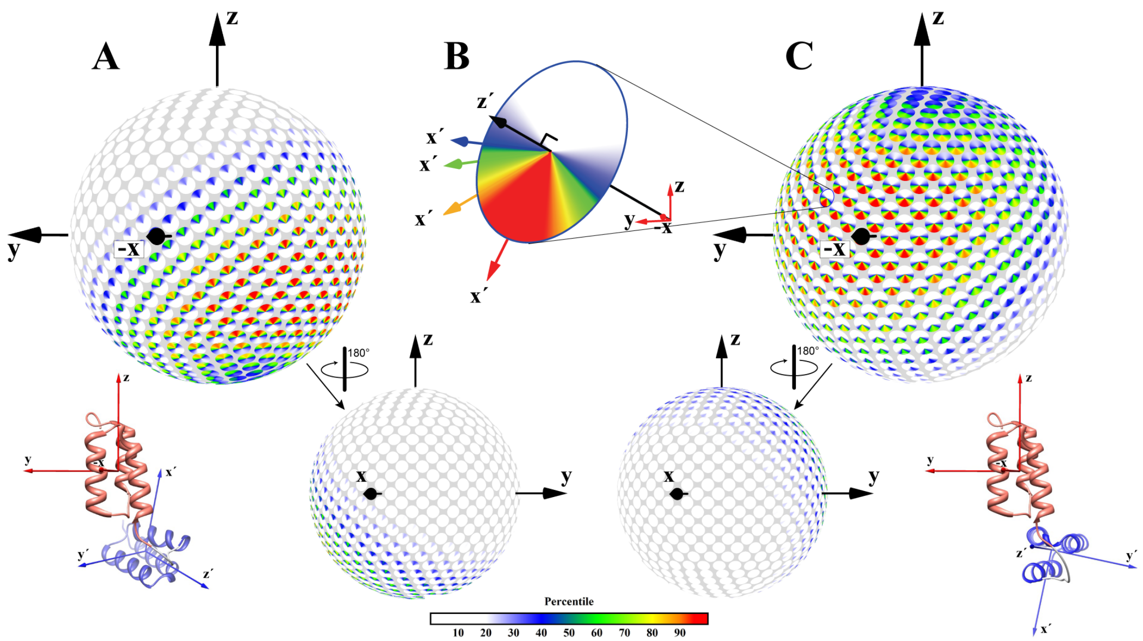


FIGURE 4.9: Continuous distribution of interdomain orientations models for ZLBT-C shown in disk-on-sphere (DoS) views. Two solutions (shown in panels A and C) give equivalently good fits to the data. Each panel shows the distribution from a front view (top) and a back view (bottom). Also shown are atomistic models whose interdomain orientation is the most probable in each solution. The linker conformation and interdomain distance are arbitrary. (B) An example disk showing the joint probabilities of 4 different different interdomain orientations, all with the same z' axis orientation but different rotations around z'.

conformational distribution. In other words, we aim to calculate the thermodynamics for posing the molecule into its post-binding conformational distribution. We define the system as the ensemble of the molecule without its binding partners, even when the molecule is in a complex. The internal energy of the system before binding is:

$$U_1 = \int P_1(x)E_1(x)dx. \tag{4.16}$$

The potential energy after binding can be separated into two components, an intramolecular component and an intermolecular component (Fig. 4.12). Only considering the energy contribution within the system, the internal energy of the system

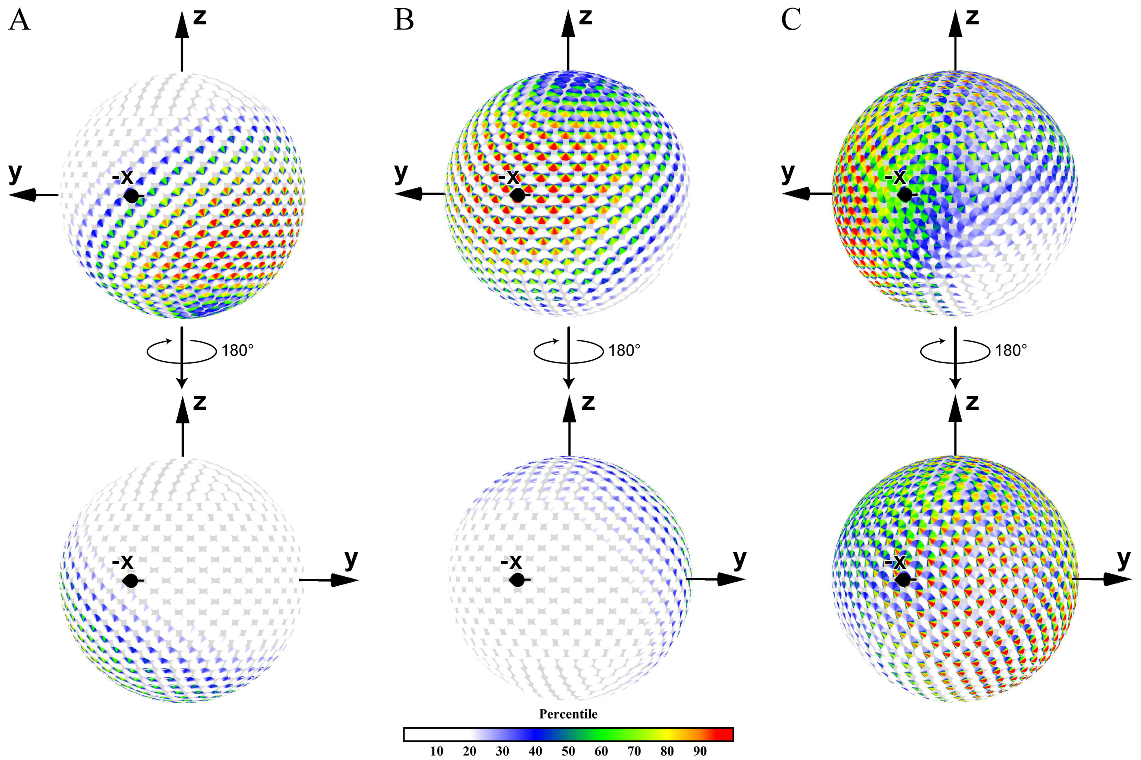


FIGURE 4.10: Comparing calculated interdomain orientational distribution with the simulated di-domain orientational distribution. The first(A) and the second(B) best calculated interdomain orientational distributions and the simulated interdomain orientational distribution(C) in the disk representation. The bottom of each panel shows the distribution from an opposite view. The color goes from red, yellow, green, blue to white as probability decreases.

after binding is:

$$U_2 = \int P_2(x) E_1(x) dx. \tag{4.17}$$

So the change of internal energy of the system in the binding reaction is:

$$\Delta U = U_2 - U_1 = \int (P_2(x) - P_1(x)) E_1(x) dx. \tag{4.18}$$

Because:

$$P_1(x) = \exp(-\frac{E_1(x)}{RT})/Z, \tag{4.19}$$

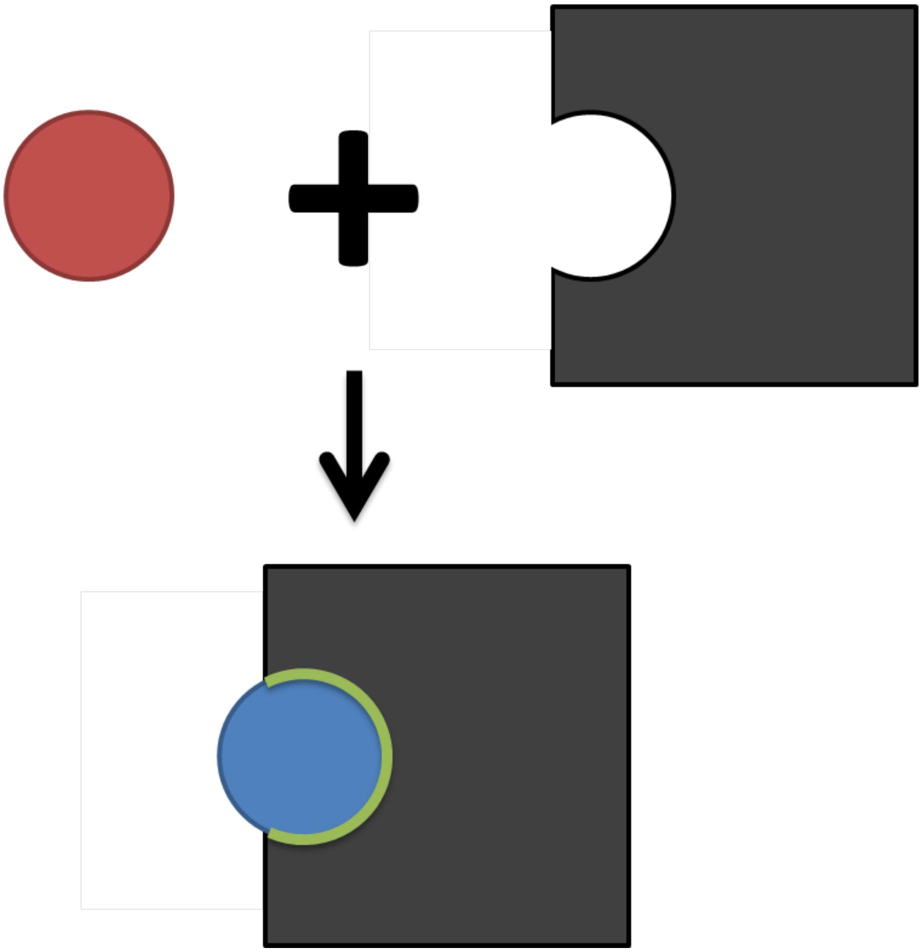


Figure 4.11: The energetic components of a binding reaction. When the circular molecule with a pre-binding conformational distribution (red) binds to its binding partner (black), the binding energetics can be separated into three components: the intra-molecular energetics of the circular molecule and its binding partner and the inter-molecular energetics between the two (green). The intra-molecular energetics of the circular molecule is a result of the switch from the pre-binding conformational distribution (red) to post-binding conformational distribution (blue).

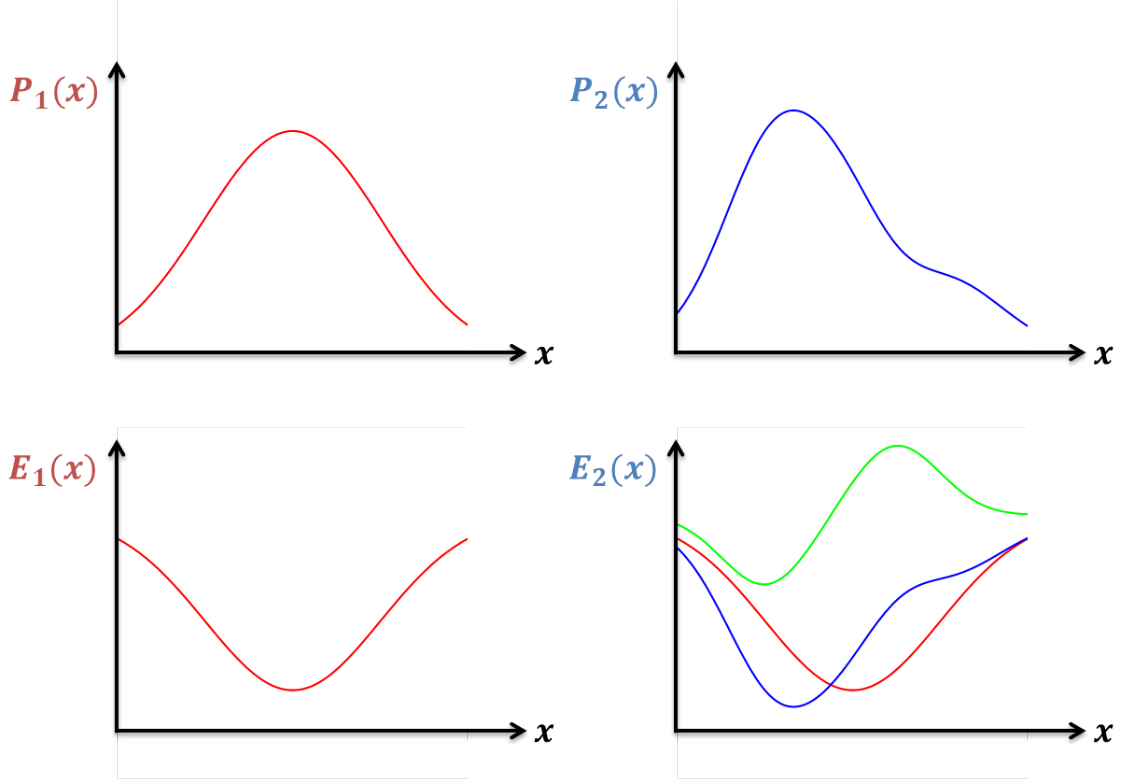


FIGURE 4.12: Illustration of the energy landscape and the conformational distribution before and after binding. The energy landscape and the probability density function of the conformational distribution before binding are shown as red curves on the left side. The energy landscape and the probability density function of the conformational distribution after binding are shown as blue curves on the right side. The potential energy after binding can be separated into two components, an intramolecular component (red) and an intermolecular component (green).

where Z is the partition function, we have:

$$E_1(x) = -RT \ln P_1(x) - RT \ln Z. \tag{4.20}$$

By substituting Eq. (4.20) into Eq. (4.18), we have:

$$\Delta U = -RT \int (P_2(x) - P_1(x)) \ln P_1(x)\ dx - RT \int (P_2(x) - P_1(x)) \ln Z\ dx. \tag{4.21}$$

The second term on the right-hand side in Eq. (4.21) is zero because both distributions $P_1(x)$ and $P_2(x)$ are normalized distributions and their integrals equal 1.

Consequently, we have:

$$\Delta U = -RT \int (P_2(x) - P_1(x)) \ln P_1(x) dx. \tag{4.22}$$

In addition, ΔS can be calculated using the following Eq.:

$$\Delta S = R \int (P_2(x) \ln P_2(x) - P_1(x) \ln P_1(x)) dx \tag{4.23}$$

Consequently, the Helmholtz free energy can be calculated from ΔU and ΔS as:

$$\Delta F = \Delta U - T\Delta S. \tag{4.24}$$

The above equations can be used to calculate ΔF, ΔU and ΔS from pre-binding and post-binding distributions.

4.5.2 The energetically favorable CDIO of SpA

The CDIO model allows a straightforward calculation of the difference in free energy, internal energy and entropy between two distributions. When used to compare CDIO models for free and bound proteins, this information can provide insights into the role of conformational reorientation to binding energetics. In the future, we plan to experimentally determine the CDIO for ZLBT-C bound to two IgG molecules. In place of this experimental data, we simulated the interdomain orientational distribution when binding to antibody by modeling the four Fab regions and the two single domain/Fc complex as rigid bodies and modeling the linkers between the rigid bodies as flexible chains. An ensemble of the double domain complex was generated using RANCH [88]. The simulation generated conformers by sampling the Ramachandran space of each flexible residue and rejected conformers with steric clashes. We then converted the resulting discrete ensemble into a CDIO model by convolution of Bingham kernels for each conformer (Fig. 4.13). Quantitative comparison of the simulated

ZLBT-C/IgG$_2$ CDIO to that of solution 2 gives a re-orientational Helmholtz free energy of binding (ΔF) of 0.8 and an internal energy of binding (ΔU) of 0.7 kcal/mol. To compare these values to those of other possible CDIOs, we also systematically calculated the ΔF and ΔU from 331776 maximum probability orientations with the same variance. The ΔF of solution 2 is in the 23rd percentile of all values, which range from 0.3 to 1.8 kcal/mol. The large difference between the energy cost of 0.8 kcal/mol and the maximum cost of 1.8 kcal/mol indicates that maximum probability orientation of solution 2 is posed in a favorable orientation to bind antibodies. It should be noted that due to lack of change in compressibility upon re-orientation, Helmholtz free energy is equivalent to Gibbs free energy ($\Delta F = \Delta G$) and internal energy is equivalent to enthalpy ($\Delta U = \Delta H$) in this case. Altering the maximum probability orientation could significantly increase the conformational energy cost of binding, so the interdomain orientational distribution we observed could have been evolutionarily optimized.

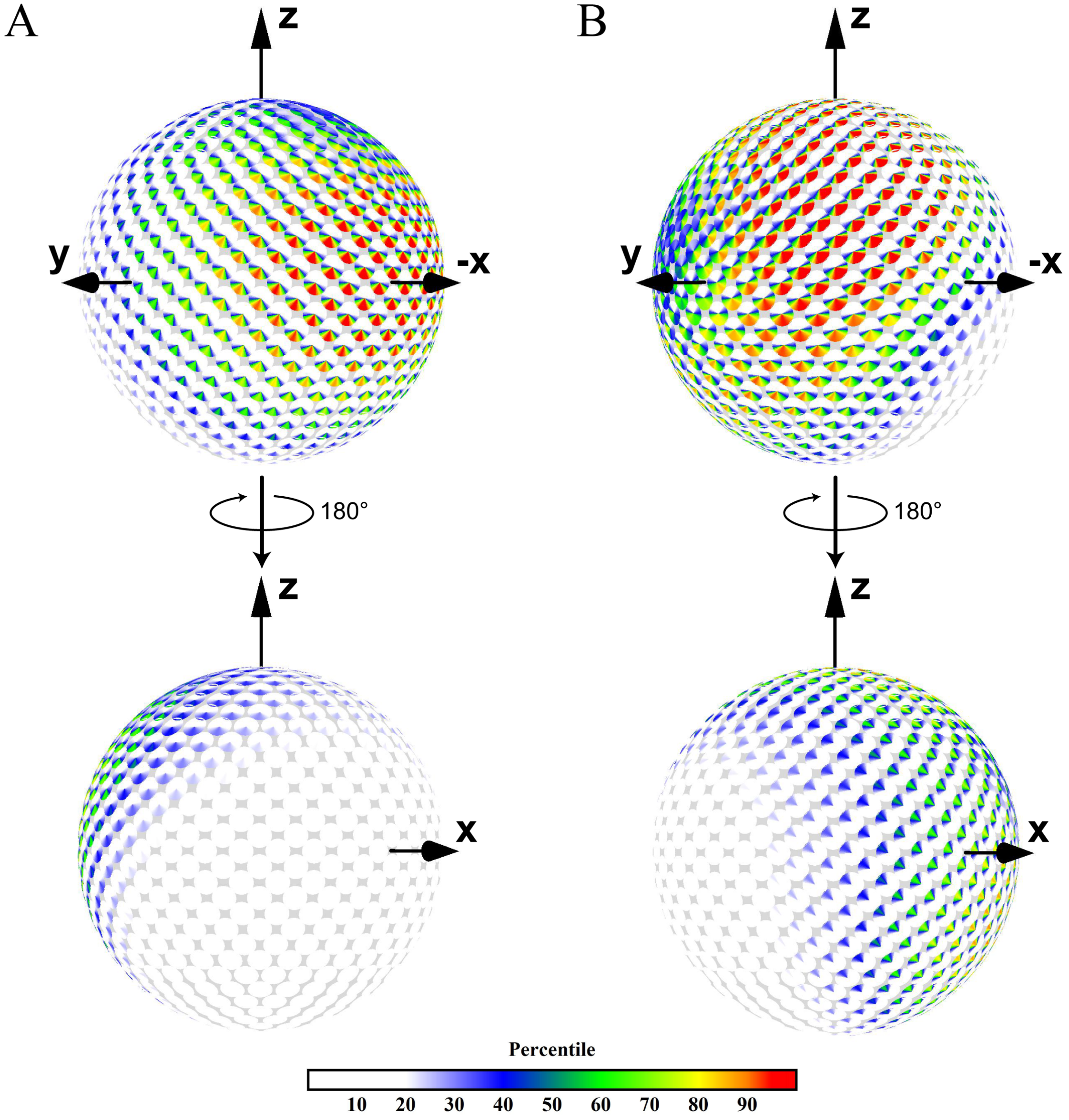


Figure 4.13: Comparing calculated interdomain orientational distribution with the simulated antibody-complex orientational distribution. The second best calculated interdomain orientational distributions(A) and the simulated antibody-complex interdomain orientational distribution(B) in the disk representation. The bottom of each panel shows the distribution from an opposite view. The color goes from red, yellow, green, blue to white as probability decreases.

5

Applications to other biomolecular systems

5.1 Application to calmodulin

Using the NMR data of calmodulin [39], I calculated the alignment tensors of both domains, which were subsequently fed into the branch-and-bound algorithm to search for optimal solutions. The first four optimal solutions were compared to the four conformations with highest maximum allowed probabilities(MAPs) in the 2007 JACS paper [39]. More specifically, the inter-domain orientation corresponding to the maximum point of each optimal distribution was used for comparison. The four highest-probability inter-domain orientations generally agree with the four high-MAP conformations, suggesting a consistency between the two methods.

5.1.1 The maximum allowed probability (MAP) method

In the 2007 JACS paper, an approach was presented to determine the maximum allowed probabilities(MAPs) of inter-domain orientations [39]. The MAP value is not the probability of an inter-domain orientation, instead, it represents the maximum probability of that orientation in any distribution that satisfies the experimental constraints. In other words, the MAP value is a measure that represents the information

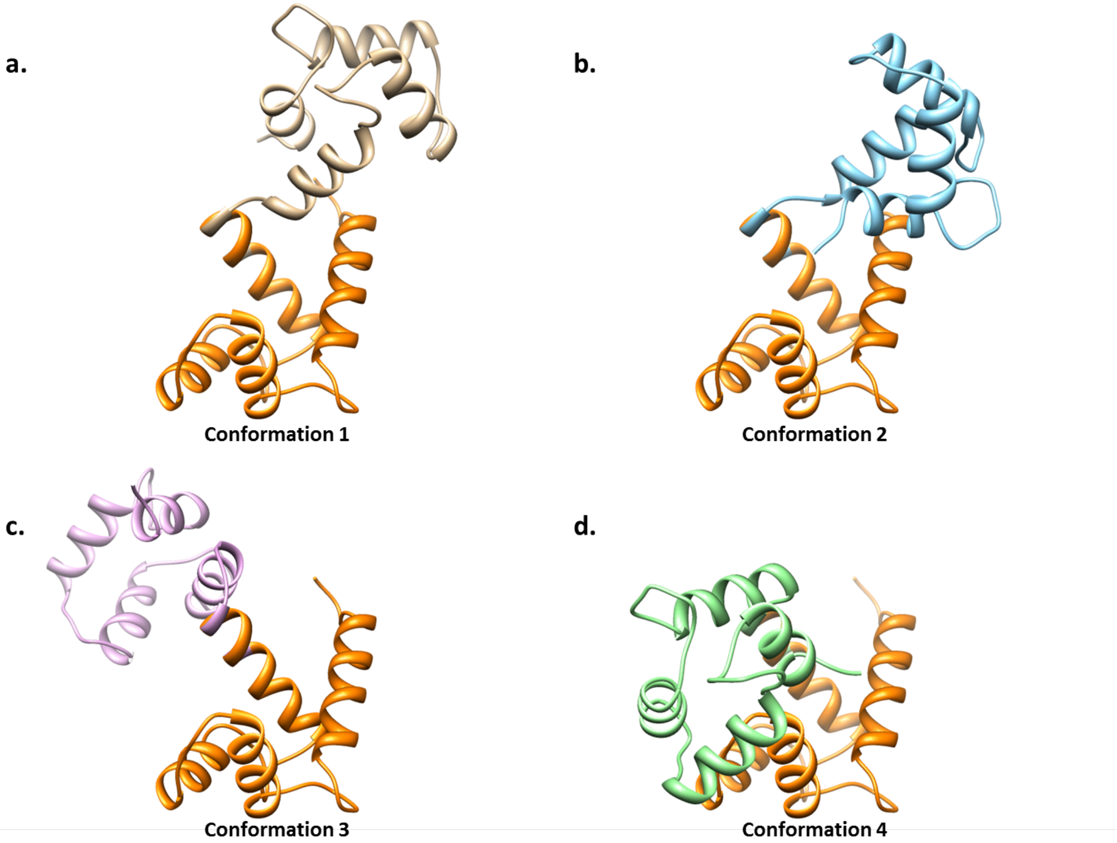


FIGURE 5.1: N-terminal domains of the four conformations are shown as orange ribbons. The C-terminal domains are shown as brown, cyan, purple and green ribbons, respectively. A-D. the four high-MAP conformations.

derived from experimental observables, but without a probabilistic interpretation. In the calmodulin case, the distribution of MAP value has four modes. The inter-domain conformations corresponding to local maxima of the four modes are shown in Fig.5.1.

5.1.2 Analysis of the RDC data of calmodulin using the CDIO approach

In the RDC and PCS dataset of calmodulin [39], PCSs were measured for the N-terminal domain and RDCs were measured for the C-terminal domain. Although PCS and RDC are different physical observables, both were induced by the magnetic alignment of lanthanide ions and thus they share the same alignment tensor. Alignment tensors of the N-terminal domain were calculated using the PCSs while align-

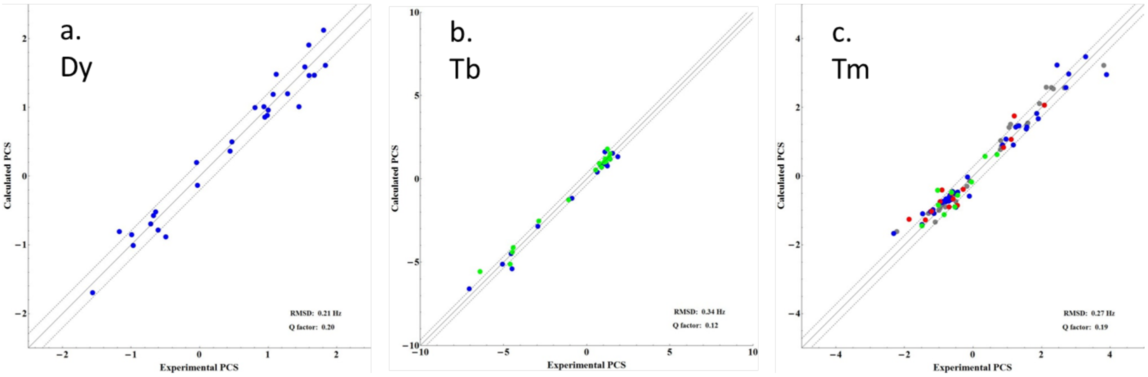


FIGURE 5.2: Experimental PCSs of the N-terminal domain are plotted against back-calculated PCSs in three alignments(Dy/Tb/Tm). RMSD and Q factor are reported in the bottom right corner.

ment tensors of the C-terminal domain were calculated using the RDCs. All tensors were calculated using a SVD method following the protocol described in [72]. Axial and rhombic components of the alignment tensors are summarized in Table 5.1 and Table 5.2. The determined alignment tensors are consistent with the reported alignment tensors in the 2004 PNAS paper and the 2007 JACS paper [38, 39]. Experimental RDCs/PCSs were plotted against the back-calculated ones in Fig.5.2 and Fig.5.3.

Table 5.1: Axial and rhombic components of N-terminal domain of calmodulin

	Dy^{3+}	Tb^{3+}	Tm^{3+}	
$\Delta\chi_{ax}$	34.78	35.54	24.75	$\times 10^{-32} m^3$
$\Delta\chi_{rh}$	-15.10	-12.42	-8.42	$\times 10^{-32} m^3$

Table 5.2: Axial and rhombic components of C-terminal domain of calmodulin

	Dy^{3+}	Tb^{3+}	Tm^{3+}	
$\Delta\chi_{ax}$	-2.05	-1.85	-3.74	$\times 10^{-32} m^3$
$\Delta\chi_{rh}$	0.85	0.93	2.46	$\times 10^{-32} m^3$

The calculated alignment tensors were subsequently fed into the fitting algorithm to find the optimal solution for the distribution. The objective function used in the fitting algorithm is Eq. 4.8.

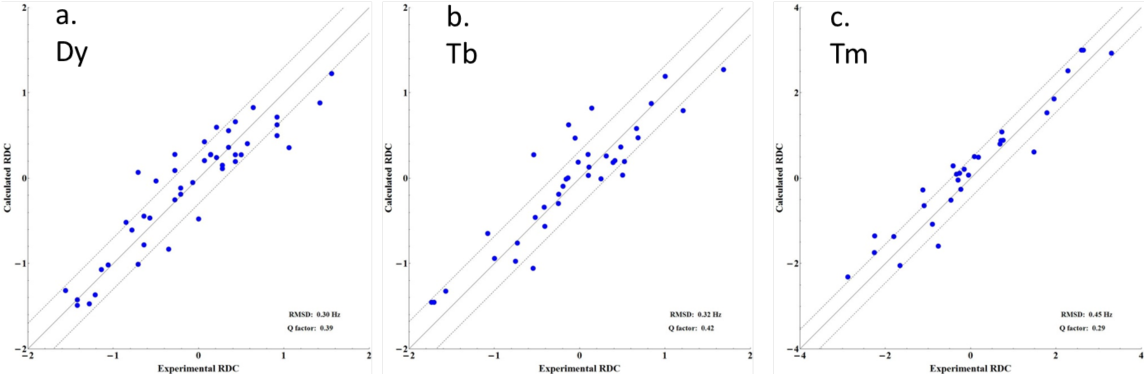


Figure 5.3: Experimental RDCs of the C-terminal domain are plotted against back-calculated RDCs in three alignments(Dy/Tb/Tm). RMSD and Q factor are reported in the bottom right corner.

The global minimum has a objective function value of 8.014. After analyzing all solutions with objective function values under 15, another three different solutions were found. They have objective function values of 8.017, 9.894 and 11.56, respectively. The value of the objective function does not mean much, but solutions with objective function values above 15 barely satisfy the experimental constraints. The four solutions are different from each other because the Jensen-Shannon divergences(JSD) between each pair of their corresponding distributions are above 0.2. The divergences between each pair are summarized in Table 5.3.

Table 5.3: JSD between the four solutions of CaM

JSD	Dist 1	Dist2	Dist3	Dist4
Dist1	—	0.55	0.94	0.95
Dist2		—	0.96	0.95
Dist3			—	0.55
Dist4				—

Inter-domain orientations are expressed as quaternions in the distribution. From the first four optimal solutions, I derived the quaternion with highest probability in each solution for distribution. From the four conformations shown in Fig.5.1, I also calculated the quaternion corresponding to each conformation. By comparing the two sets of quaternions, it can be shown easily that the highest probability orientation in solution 1 is nearly identical to conformation 3, the orientation in solution 2 is nearly identical to conformation 1, the orientation in solution 3 is nearly identical to conformation 4 and the orientation in solution 4 is nearly identical to conformation 2. The similarities between the highest probability orientations and conformations are first quantified by the inner product of their corresponding quaternions. The inner product range from 0 to 1, indicating least to highest similarity. The result is summarized in Table 5.4. I also applied another similarity measure called FAA percentile on the inter-domain orientations [87]. The percentile presents the fraction of all orientations that have a larger geometric difference from one orientation in the pair than the difference between the two in the pair. It ranges from 0 to 100, indicating least to highest similarity. In addition, inter-domain conformations corresponding to the quaternion pairs are shown in Fig. 5.4 to provide a more intuitive comparison.

Table 5.4: Similarity between the quaternions representing the maximum probability orientation of the four solutions of CaM

	orientation1 and conformer3	orientation2 and conformer1	orientation3 and conformer4	orientation4 and conformer2
Inner product	0.94	0.95	0.96	0.93
FAA percentile	98.2	99.1	98.9	98.6

The continuous probabilistic approach and the MAP method agree with each other in general. Four solutions were observed when fitting for distributions and four MAP maxima were observed in the 2007 JACS paper. The existence of multiple solutions or multiple modes could indicate a degeneracy in the RDC and PCS data of calmodulin, which does not have the information content to distinguish the solutions. Alternatively, all the modes could be real motional modes of calmodulin. The fact

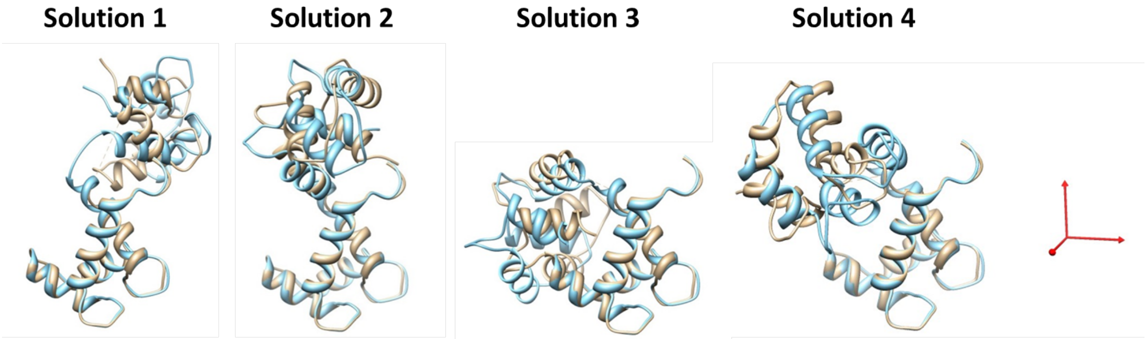


Figure 5.4: The N-terminal domains are in the bottom or the bottom-right corner of each panel. Cyan structures represent the MAP conformations. Brown structures represent the highest probability orientations calculated by the branch-and-bound method.

of finding multiple solutions indicates that the continuous probabilistic approach has the ability to detect multiple modes even though the algorithm uses a unimodal model for fitting.

5.2 Application to TAR

Using the motional decoupled RDC data of TAR [78], I calculated the alignment tensors of both helices for the two alignments. The two alignments use different helices as the reference helix. A modification was made to the algorithm in order to deal the reference-switched case. The alignments were subsequently fed into the modified branch-and-bound algorithm to search for optimal solutions. Because the interdomain motions of TAR is less intensive than that of SpA-N and that of calmodulin, the current parameter space searched by the algorithm does not seem to cover the optimal solution. Nonetheless, a preliminary result is presented here. The solution agrees with the one alignment of TAR, but does not agree with the other alignment very well. A more complete search on a larger parameter space can cover the estimated optimal solution, which should agree with both alignments.

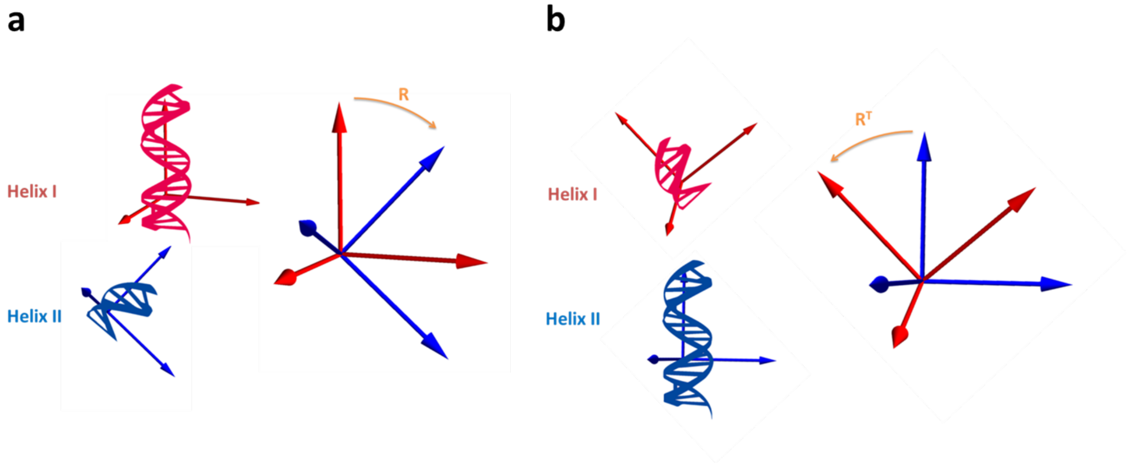


FIGURE 5.5: An inter helical orientation can be represented as a rotation. (a) When Helix I is the reference, the inter helical orientation is represented by rotation R. (b) When Helix II is the reference, the inter helical orientation is represented by rotation R^T.

5.2.1 *Generalization of the CDIO approach to the case when the reference domain is switched*

In the RDC dataset of TAR, two different alignments are obtained by inverting the reference helix [?]. Although the inter-helical motions are the same, parameterization of the distributions depends on the choice of reference helix. Nonetheless, a simple relationship between the two sets of parameters can be derived, which enables the branch and bound algorithm to deal with the RDC data of TAR.

As shown in Fig. ??, Helix I is elongated in the first alignment and Helix II is elongated in the second one. So Helix I is the reference helix in the first alignment while Helix II is the reference helix in the second alignment. Because the inter-helical motions are the same in the two alignments, the probabilities corresponding to one particular inter-helical orientation must equal each other. Assuming an inter-helical orientation is represented by a rotation R in the first scenario, the orientation is represented by its inverse rotation R^T in the second scenario, because of the change of reference domain, as shown in Fig. ??.

From Eq. 4.4 and Eq. 4.7, for the first alignment with elongated Helix I, we have the following relationship

$$s_2 = Q(\tilde{\mathbf{q}_\mathbf{R}}^{-1}) \cdot \mathbf{E}[Q\,;\,\Lambda] \cdot Q(\tilde{\mathbf{q}_\mathbf{L}}^{-1}) \cdot s_1 \tag{5.1}$$

For the second alignment with elongated Helix II, we also have a simple relationship between s_1' and s_2' by inverting the reference helix [85]. The relationship is the following:

$$s_1' = Q(\tilde{\mathbf{q}_\mathbf{L}}) \cdot \mathbf{E}[Q\,;\,\Lambda] \cdot Q(\tilde{\mathbf{q}_\mathbf{R}}^{-1}) \cdot s_2' \tag{5.2}$$

In the above two equations, s_1 and s_2 are vectorized Saupe tensor of Helix I and Helix II in the first alignment, s_1' and s_2' are vectorized Saupe tensor of Helix I and Helix II in the second alignment. Based on Eq. 5.2, we could derive

$$s_2' = Q(\tilde{\mathbf{q}_\mathbf{R}}^{-1}) \cdot \mathbf{E}[Q\,;\,\Lambda]^{-1} \cdot Q(\tilde{\mathbf{q}_\mathbf{L}}^{-1}) \cdot s_1' \tag{5.3}$$

From Eq. 5.1 and 5.3, the mathematical expression for switching the reference helix is almost the same as using a different alignment with the same reference helix. The only difference is the middle matrix $E[Q](\Lambda_I)^{-1}$, which is the inverse of the averaging matrix $E[Q](\Lambda_I)$. Consequently, the middle matrix acts like a concentrating operator and increases the eigenvalues of a Saupe tensor. The increase on eigenvalues is expected because s_1' is an averaging product of s_2' and s_2' should have larger eigenvalues. Additional function needs to be implemented to calculate the inverse $E[Q](\Lambda_I)^{-1}$. With the minor revision, the brand and bound algorithm will be able to find the best solution satisfying Eq. 5.1 and 5.3.

5.2.2 Analysis of the RDC data of TAR using the CDIO approach

Alignment tensors of both helices were calculated using a SVD method [72]. Asymmetry parameter of the alignment tensors are summarized in Table ?? and Table ??.

The determined alignment tensors are consistent with the reported alignment tensors [78]. Correlation plots of experimental RDCs from Helix I against the back-calculated ones are shown in Fig.5.6 and Fig.5.7. Correlation plots of experimental RDCs from Helix II against the back-calculated ones are shown in Fig.5.8 and Fig.5.9.

Table 5.5: Asymmetry parameters of Helix I and Helix II in two alignments with elongated Helix I (EI) or elongated Helix II (EII)

	HI-EI	HI-EII	HII-EI	HII-EII
η	0.067	0.570	0.436	0.088

The best interdomain orientation distribution was calculated using the CDIO approach. The best solution has an objective function value of 174. The large objective function value indicate that the solution may not be a good fit to the data. Indeed, when the RDCs are predicted by the solution, the predicted RDCs do not have a good agreement in both alignment. The predicted RDCs generally agrees with RDCs in the first alignment with elongated Helix I. The Q factor is 0.34. The predicted RDCs does not generally agree with RDCs in the second alignment with elongated Helix II. The Q factor is 0.58. The result is not very surprising because the parameter space searched by the algorithm is limited in a region. Distributions corresponding to this region do not include very sharp distributions. The interdomain motions is very limited in the TAR system, so the distribution describing the system should be a sharp one. The presented result is a preliminary one because only short runs of the algorithm can be completed by the time of writing this dissertation. Long time runs are in process and they allow a search on a larger parameter space.

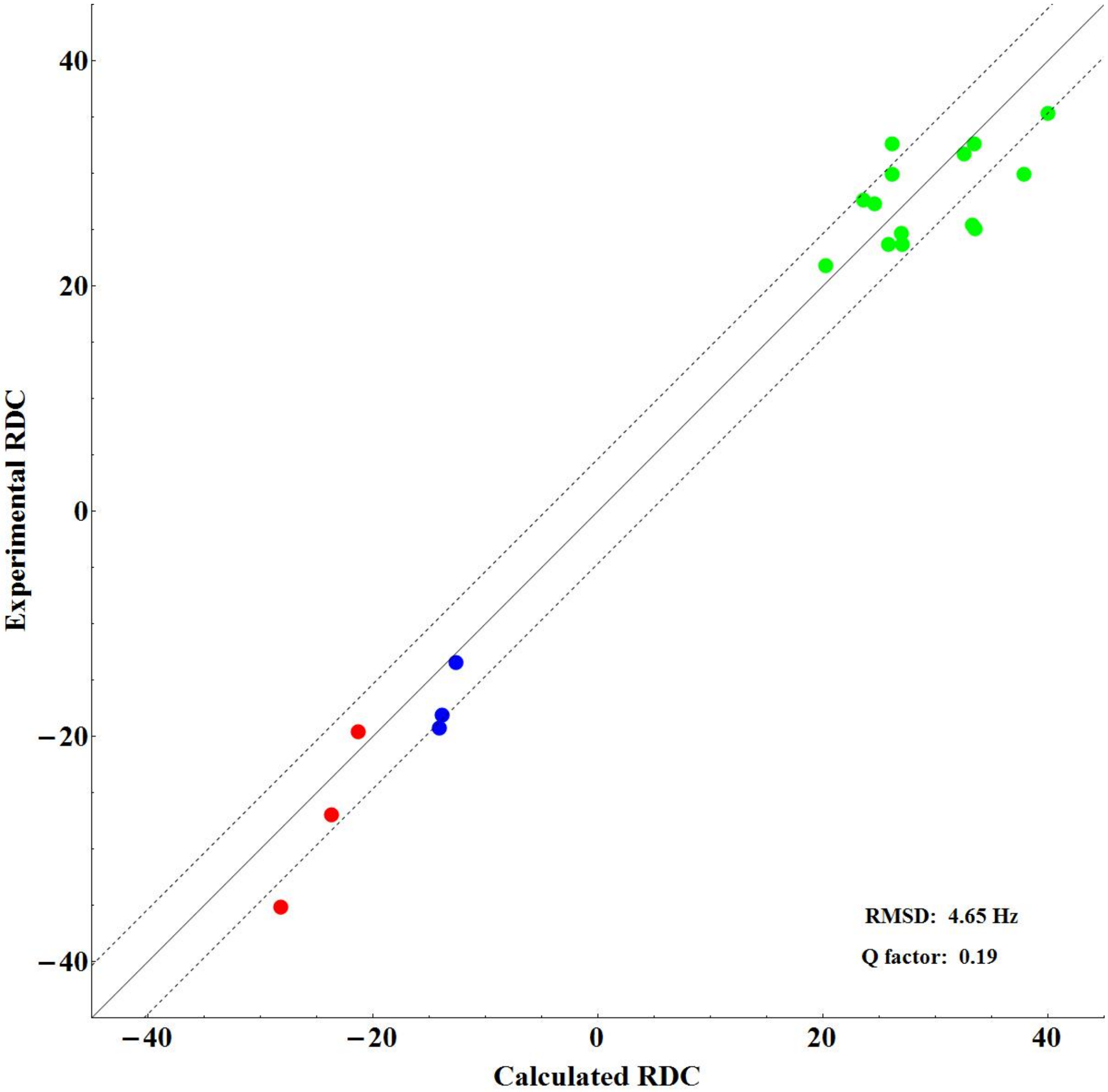


FIGURE 5.6: TAR Helix I RDC correlation plot with elongated Helix I. RMSD and Q factor are reported in the bottom right corner.

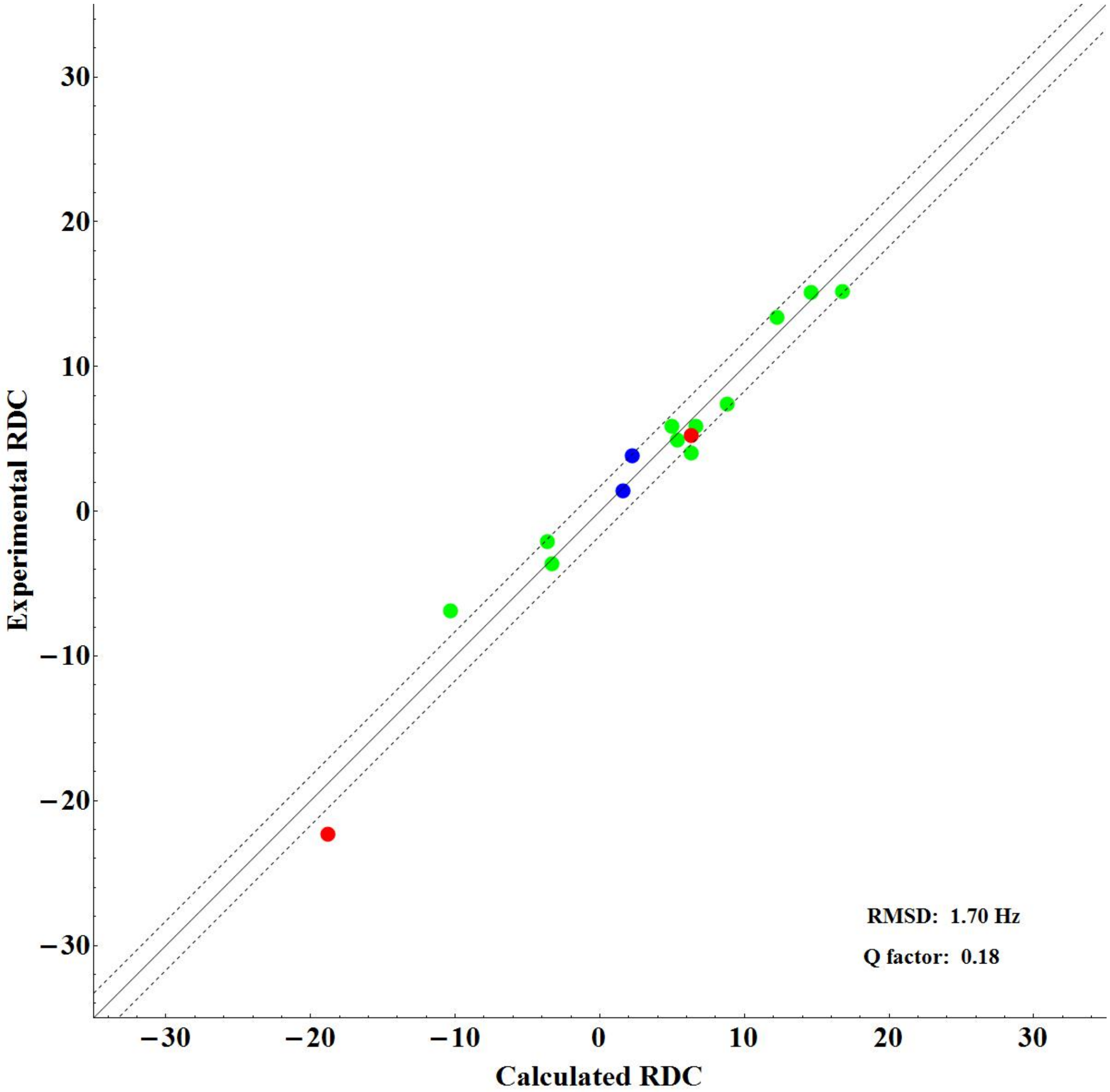


FIGURE 5.7: TAR Helix I RDC correlation plot with elongated Helix II. RMSD and Q factor are reported in the bottom right corner.

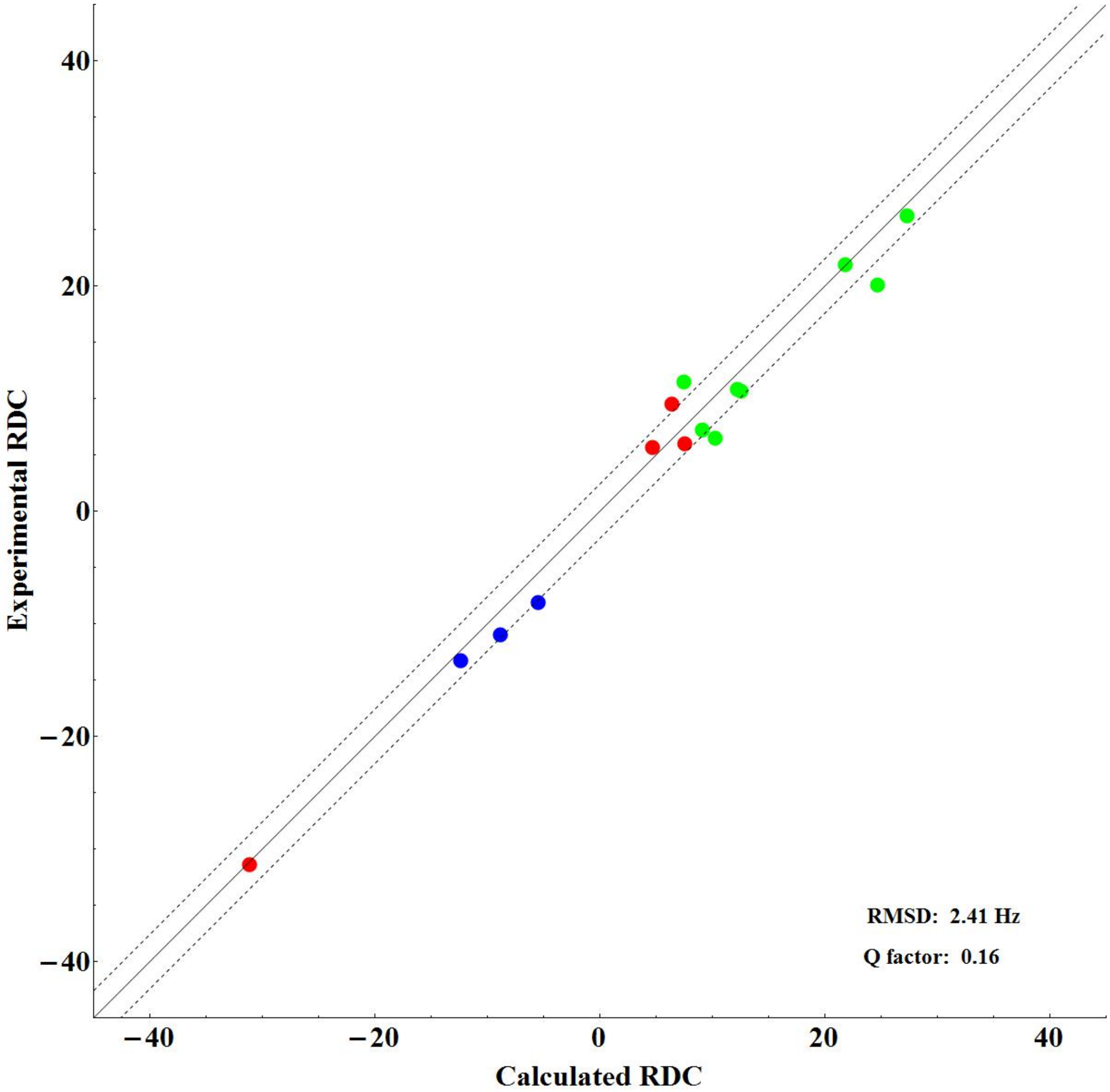


FIGURE 5.8: TAR Helix II RDC correlation plot with elongated Helix I. RMSD and Q factor are reported in the bottom right corner.

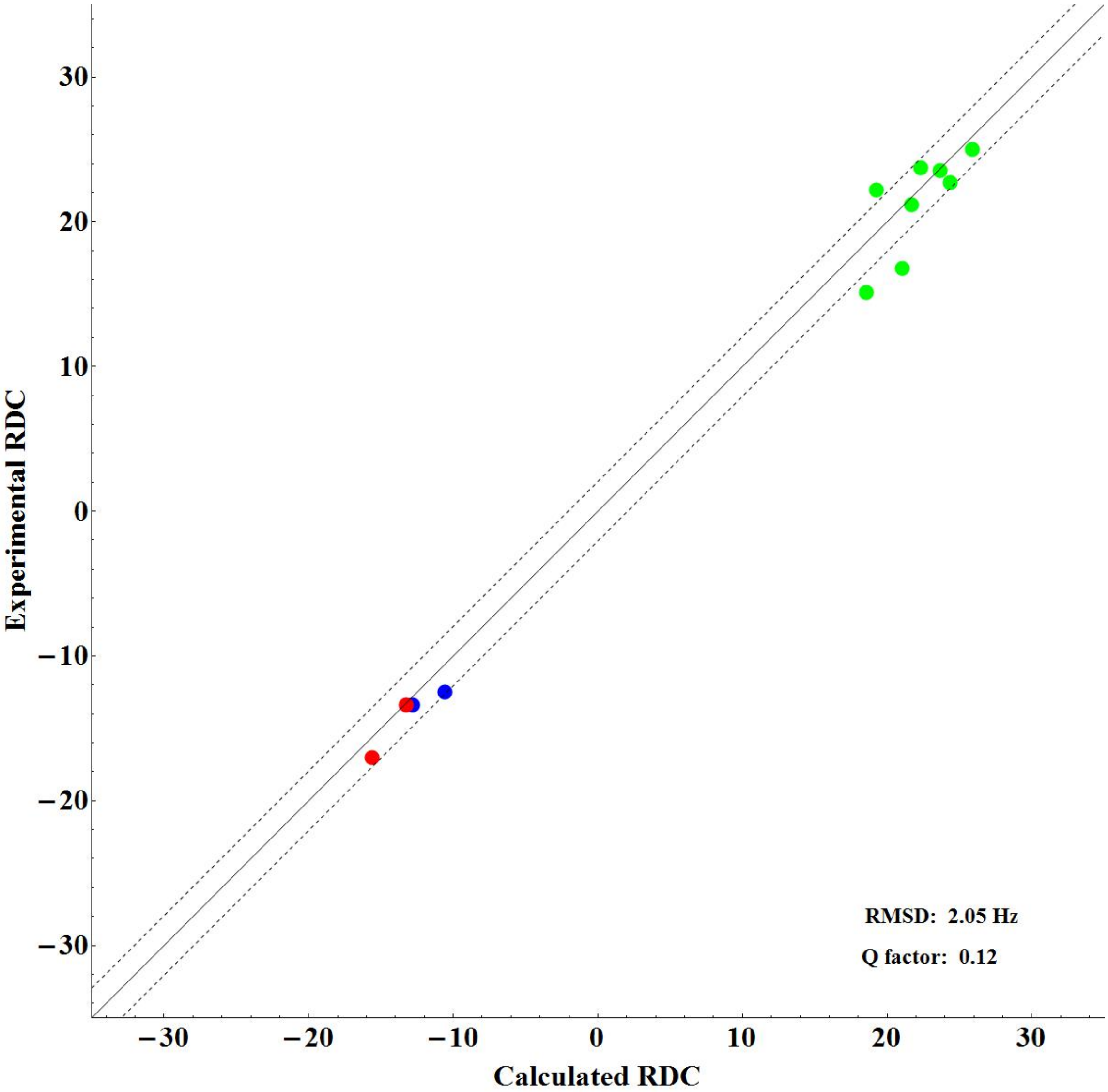


FIGURE 5.9: TAR Helix II RDC correlation plot with elongated Helix II. RMSD and Q factor are reported in the bottom right corner.

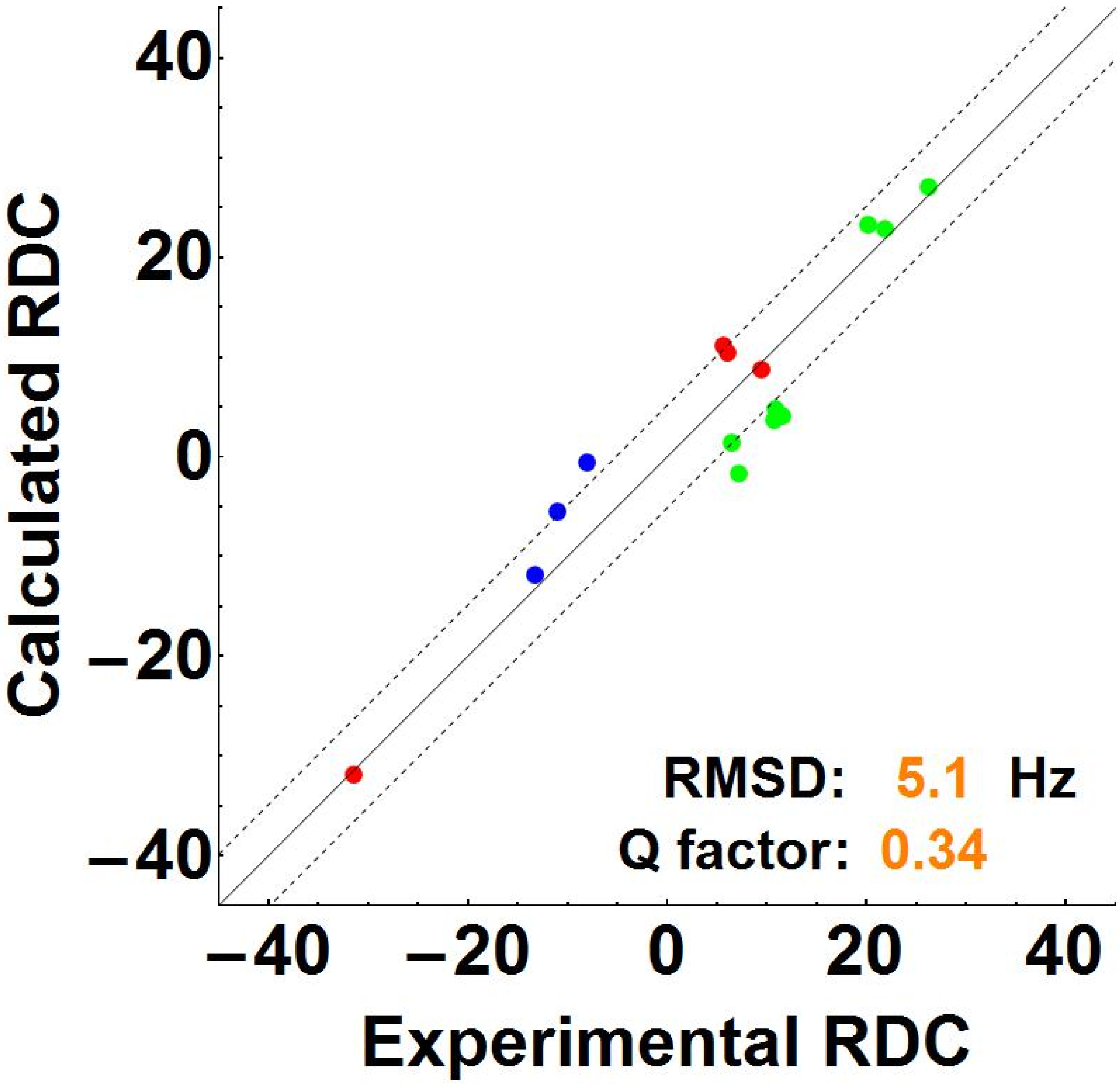


Figure 5.10: TAR EI RDC correlation plot predicted by the best solution. Only RDCs from Helix II are plotted. RMSD and Q factor are reported in the bottom right corner.

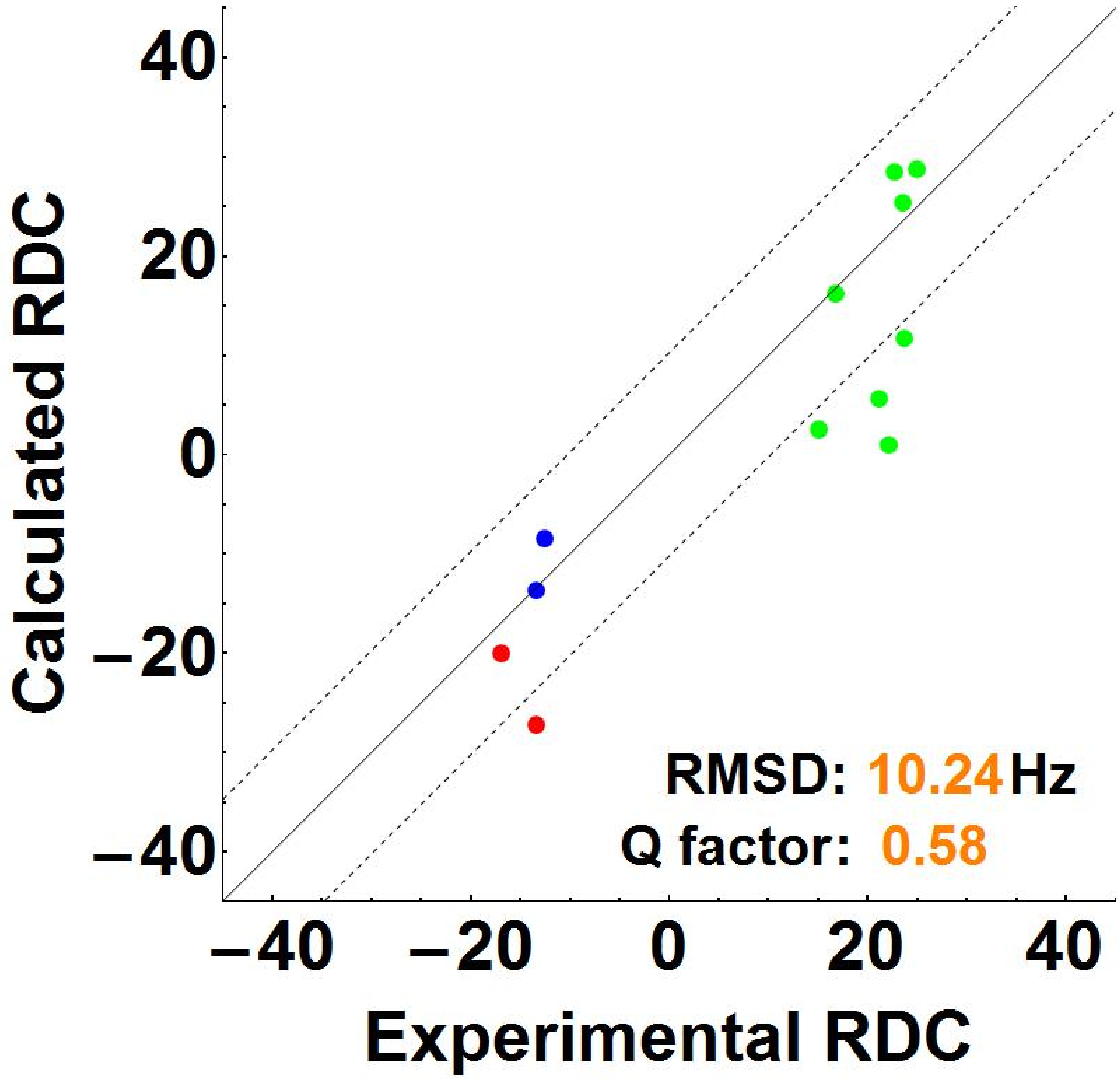


FIGURE 5.11: TAR EII RDC correlation plot predicted by the best solution. Only RDCs from Helix II are plotted. RMSD and Q factor are reported in the bottom right corner.

6

Conclusions

Biomolecules have dynamics over various spatial and time scales. The motions associated with these dynamics are usually convolved with each other. There are global tumbling and interdomain motions in the SpA-N system. Intradomain dynamics on the nanoseconds to picoseconds timescale has been quantified by NMR spin relaxation experiments. The order parameters suggest that the intradomain dynamics is not significant and hardly have influence on interdomain motions and global tumbling. The high order parameters in the core region of each domain also justify the calculation of a Saupe tensor for each domain. ~~flexible linkers.~~ In a retrospective view, the high quality fitting of RDC data using a single Saupe tensor for each domain also supports low intradomain dynamics. As a result, we only consider two motional modes in the SpA-N case. The two motional modes are usually convolved with each other without a special experimental design. In order to decouple the interdomain motions from global tumbling, we engineered a lanthanide binding tag into one domain of the di-domain mimics and aligned the constructs magnetically. One domain is aligned directly and the other domain is aligned indirectly. RDCs of the directly aligned domain contain information about global tumbling, e.g. how the alignment restricts its orientations. RDCs of the indirectly aligned domains contain information about both global tumbling and interdomain motions. Using a mathematical method based on the exterior tensor algebra [71], the global tumbling component from the indirectly aligned domain is separated from the interdomain motion component, leaving us the exclusive information about interdomain motions. This information was subsequently used to constrain a continuous orientational distribution on $SO(3)$ instead of a discrete structural ensemble, which describes a broad spectrum of interdomain motions.

Our study provides useful information about the motions between adjacent domains in SpA-N. The interdomain orientational distribution, revealed using our CDIO method, has several salient features that could be important for the function of SpA-N. The broad, albeit highly anisotropic, orientational distribution indicates that the molecule samples a large region of the interdomain orientational space, as a result of high structural flexibility. As previously noted, this structural flexibility may be crucial to the large functional plasticity displayed by this important *S. aureus* virulence factor [89]. The interdomain orientational distribution is certainly influenced, if not dominated, by the six-residue flexible linker. The linker sequences are nearly identical in the 4 linkers in SpA-N and among more than 50 *S. aureus* isolates. One possible reason for this conservation is to maintain the same interdomain orientational distribution. If this hypothesis is correct, the interdomain orientational distribution is selected by evolutionary pressure as a functional phenotype of the protein. Because SpA has been shown to bind to at least four other proteins as part of its function in addition to Fc of IgG, this orientational distribution is presumably optimized for binding to all of its binding partners.

Based on the magnitudes of RDCs in the first domain and in the second domain, the ZLBT-C construct has the most flexible interdomain linker compared to previously studied di-domain systems [78, 39]. Surprisingly, even with six residues and 12 degrees of freedom between the two adjacent domains, the interdomain orientational distribution is not close to a uniform distribution. In fact, the region bounded by blue color in Figure 4D takes only 35% of the entire interdomain orientational space but it has a total probability of 80%. Although the distribution is broad, it certainly shows a strong preference over a limited volume. This result has general implications to studies of flexible biomolecular systems. Our result suggests that a binary classification of biopolymers as flexible or inflexible is an over-simplification. Describing flexible systems requires quantifying not only secondary structure and packing of each conformer, but also the variances and correlations of the conformational distribution. These features could vary across different flexible systems, especially when they have multiple functions. For flexible systems with different parts working jointly, quantifying the conformational distribution is necessary to reveal the molecular mechanism underlying concerted motions and the relationship between motion and function.

It should be noted that the Bingham model is associated with assumptions. It is a unimodal model with a smoothness assumption. The uni-modality of the model may not be reasonable, but we observe the potential of our method to detect multiple modes. We believe the smoothness assumption is reasonable assumption for two reasons. First, smoothness implies continuity. the interdomain motions should be continuous motions, which leads to a continuous distribution of interdomain orientations. Second, the information content in RDCs is limited. It does not contain information about high frequency oscillations, so high frequency components in the distribution function can not be constrained. The smoothness assumption can exclude the high frequency components and regularize the problem effectively. However, we also believe that further minimizing assumptions is beneficial. Less assumptions means less bias in the result as long as the problem is still regularized. Minimizing assumptions will be a future direction of the project. In fact, we have already established a theoretical framework for a maximum entropy method, which only enforces the smoothness of the resulted distribution [85]. The maximum entropy method does not assume either the number of modes or curvature of the probability density function, so it is a more flexible and adaptive method.

Bibliography

[1] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[2] Brian Kuhlman and David Baker. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, 97(19):10383–10388, 2000.

[3] Chris Sander and Reinhard Schneider. Database of homologyderived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, 9(1):56–68, 1991.

[4] Gregorio Weber. Energetics of ligand binding to proteins. *Adv Protein Chem*, 29(1), 1975.

[5] Rv H Austin, KW Beeson, L Eisenstein, H Frauenfelder, and IC Gunsalus. Dynamics of ligand binding to myoglobin. *Biochemistry*, 14(24):5355–5373, 1975.

[6] Peter E Leopold, Mauricio Montal, and José N Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, 89(18):8721–8725, 1992.

[7] Cyrus Levinthal. How to fold graciously. *Mossbauer spectroscopy in biological systems*, pages 22–24, 1969.

[8] DE Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences*, 44(2):98–104, 1958.

[9] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *Journal of molecular biology*, 12(1):88–118, 1965.

[10] Merry M Rubin and Jean-Pierre Changeux. On the nature of allosteric transitions: implications of non-exclusive ligand binding. *Journal of molecular biology*, 21(2):265–274, 1966.

[11] Tomasz Wlodarski and Bojan Zagrovic. Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proceedings of the National Academy of Sciences*, 106(46):19346–19351, 2009.

[12] Elan Zohar Eisenmesser, Daryl A Bosco, Mikael Akke, and Dorothee Kern. Enzyme dynamics during catalysis. *Science*, 295(5559):1520–1523, 2002.

[13] Magnus Wolf-Watz, Vu Thai, Katherine Henzler-Wildman, Georgia Hadjipavlou, Elan Z Eisenmesser, and Dorothee Kern. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nature structural and molecular biology*, 11(10):945–949, 2004.

[14] Elan Z Eisenmesser, Oscar Millet, Wladimir Labeikovsky, Dmitry M Korzhnev, Magnus Wolf-Watz, Daryl A Bosco, Jack J Skalicky, Lewis E Kay, and Dorothee Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–121, 2005.

[15] Katherine A. Henzler-Wildman, Vu Thai, Ming Lei, Maria Ott, Magnus Wolf-Watz, Tim Fenn, Ed Pozharski, Mark A. Wilson, Gregory A. Petsko, Martin Karplus, Christian G. Hubner, and Dorothee Kern. Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–844, 2007. 10.1038/nature06410.

[16] Dan McElheny, Jason R Schnell, Jonathan C Lansing, H Jane Dyson, and Peter E Wright. Defining the role of active-site loop fluctuations in dihydrofolate reductase catalysis. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14):5032–5037, 2005.

[17] David D Boehr, Dan McElheny, H Jane Dyson, and Peter E Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *science*, 313(5793):1638–1642, 2006.

[18] Oliver F Lange, Nils-Alexander Lakomek, Christophe Fares, Gunnar F Schrder, Korvin FA Walter, Stefan Becker, Jens Meiler, Helmut Grubmüller, Christian Griesinger, and Bert L De Groot. Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *science*, 320(5882):1471–1475, 2008.

[19] Hans Frauenfelder, Stephen G Sligar, and Peter G Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[20] J. K. Myers and T. G. Oas. Preorganized secondary structure as an important determinant of fast protein folding. *Nat Struct Biol*, 8(6):552–8, 2001.

[21] John F Brandts, Herbert R Halvorson, and Maureen Brennan. Consideration of the possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues. *Biochemistry*, 14(22):4953–4963, 1975.

[22] F. Bloch, W. W. Hansen, and Martin Packard. Nuclear induction. *Physical Review*, 69(3-4):127–127, 1946. PR.

[23] E. M. Purcell, H. C. Torrey, and R. V. Pound. Resonance absorption by nuclear magnetic moments in a solid. *Physical Review*, 69(1-2):37–38, 1946. PR.

[24] Arthur G Palmer III. Nmr characterization of the dynamics of biomacromolecules. *Chemical reviews*, 104(8):3623–3640, 2004.

[25] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. Nmr evidence for slow collective motions in cyanometmyoglobin. *Nat Struct Mol Biol*, 4(4):292–297, 1997. 10.1038/nsb0497-292.

[26] Joel R Tolman and Ke Ruan. Nmr residual dipolar couplings as probes of biomolecular dynamics. *Chemical reviews*, 106(5):1720–1736, 2006.

[27] Ivano Bertini, Claudio Luchinat, and Giacomo Parigi. Magnetic susceptibility in paramagnetic nmr. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 40(3):249–273, 2002.

[28] Barbara Richter, Joerg Gsponer, Péter Várnai, Xavier Salvatella, and Michele Vendruscolo. The mumo (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *Journal of biomolecular NMR*, 37(2):117–135, 2007.

[29] Wolfgang Kabsch, Emil F Pai, Gregory A Petskoli, and Roger S Goody. Time-resolved x-ray crystallographic study of the conformational change in ha-ras p21 protein on gtp hydrolysis. *Nature*, 345:24, 1990.

[30] Friedrich Schotte, Manho Lim, Timothy A Jackson, Aleksandr V Smirnov, Jayashree Soman, John S Olson, George N Phillips, Michael Wulff, and Philip A Anfinrud. Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, 300(5627):1944–1947, 2003.

[31] Friedrich Schotte, Jayashree Soman, John S Olson, Michael Wulff, and Philip A Anfinrud. Picosecond time-resolved x-ray crystallography: probing protein function in real time. *Journal of structural biology*, 147(3):235–246, 2004.

[32] Ethan A Merritt. Expanding the model: anisotropic displacement parameters in protein structure refinement. *Acta Crystallographica Section D: Biological Crystallography*, 55(6):1109–1117, 1999.

[33] Ilme Schlichting, Joel Berendzen, Kelvin Chu, Ann M Stock, Shelley A Maves, David E Benson, Robert M Sweet, Dagmar Ringe, Gregory A Petsko, and Stephen G Sligar. The catalytic pathway of cytochrome p450cam at atomic resolution. *Science*, 287(5458):1615–1622, 2000.

[34] Mingjie Zhang, Toshiyuki Tanaka, and Mitsuhiko Ikura. Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. *Nature Structural and Molecular Biology*, 2(9):758–767, 1995.

[35] Mark B Swindells and Mitsuhiko Ikura. Pre-formation of the semi-open conformation by the apo-calmodulin c-terminal domain and implications for binding iq-motifs. *Nature Structural and Molecular Biology*, 3(6):501–504, 1996.

[36] Gaetano Barbato, Mitsuhiko Ikura, Lewis E Kay, Richard W Pastor, and Ad Bax. Backbone dynamics of calmodulin studied by nitrogen-15 relaxation using inverse detected two-dimensional nmr spectroscopy: the central helix is flexible. *Biochemistry*, 31(23):5269–5278, 1992.

[37] James L Baber, Attila Szabo, and Nico Tjandra. Analysis of slow interdomain motion of macromolecules using nmr relaxation data. *Journal of the American Chemical Society*, 123(17):3953–3959, 2001.

[38] Ivano Bertini, Cristina Del Bianco, Ioannis Gelis, Nikolaus Katsaros, Claudio Luchinat, Giacomo Parigi, Massimiliano Peana, Alessandro Provenzani, and Maria Antonietta Zoroddu. Experimentally exploring the conformational space sampled by domain reorientation in calmodulin. *Proceedings of the National Academy of Sciences of the United States of America*, 101(18):6841–6846, 2004.

[39] Ivano Bertini, Yogesh K Gupta, Claudio Luchinat, Giacomo Parigi, Massimiliano Peana, Luca Sgheri, and Jing Yuan. Paramagnetism-based nmr restraints provide maximum allowed probabilities for the different conformations of partially independent protein domains. *Journal of the American Chemical Society*, 129(42):12786–12794, 2007.

[40] Ivano Bertini, Andrea Giachetti, Claudio Luchinat, Giacomo Parigi, Maxim V Petoukhov, Roberta Pierattelli, Enrico Ravera, and Dmitri I Svergun. Conformational space of flexible biological macromolecules from average data. *Journal of the American Chemical Society*, 132(38):13553–13558, 2010.

[41] Alexander Ogston. On abscesses. *Review of Infectious Diseases*, 6(1):122–128, 1984.

[42] Franklin D Lowy. Staphylococcus aureus infections. *New England journal of medicine*, 339(8):520–532, 1998.

[43] Adelisa L Panlilio, David H Culver, Robert P Gaynes, Shailen Banerjee, Tonya S Henderson, James S Tolson, and William J Martone. Methicillin-resistant staphylococcus aureus in us hospitals, 19751991. *Infection Control and Hospital Epidemiology*, 13(10):582–586, 1992.

[44] Sharon J Peacock, Catrin E Moore, Anita Justice, Maria Kantzanou, Lisa Story, Kathryn Mackie, Gael O'Neill, and Nicholas PJ Day. Virulent combinations of adhesin and toxin genes in natural populations of staphylococcus aureus. *Infection and immunity*, 70(9):4987–4996, 2002.

[45] Niklas Palmqvist, Timothy Foster, Andrzej Tarkowski, and Elisabet Josefsson. Protein a is a virulence factor in staphylococcus aureus arthritis and septic death. *Microbial pathogenesis*, 33(5):239–249, 2002.

[46] Nekane Merino, Alejandro Toledo-Arana, Marta Vergara-Irigaray, Jaione Valle, Cristina Solano, Enrique Calvo, Juan Antonio Lopez, Timothy J Foster, Jos R Penads, and Inigo Lasa. Protein a-mediated multicellular behavior in staphylococcus aureus. *Journal of bacteriology*, 191(3):832–843, 2009.

[47] William Wiley Navarre and Olaf Schneewind. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiology and Molecular Biology Reviews*, 63(1):174–229, 1999.

[48] Arne Forsgren. Significance of protein a production by staphylococci. *Infection and immunity*, 2(5):672, 1970.

[49] Mathias Uhlen, B Guss, B Nilsson, Sten Gatenbeck, L Philipson, and M Lindberg. Complete sequence of the staphylococcal gene encoding protein a. a gene evolved through multiple duplications. *Journal of Biological Chemistry*, 259(3):1695–1702, 1984.

[50] Bengt Guss, Mathias Uhlén, Björn Nilsson, Martin Lindberg, John Sjöquist, and Jörgen Sjödahl. Region x, the cellwallattachment part of staphylococcal protein a. *European Journal of Biochemistry*, 138(2):413–420, 1984.

[51] John D Lambris, Daniel Ricklin, and Brian V Geisbrecht. Complement evasion by human pathogens. *Nature Reviews Microbiology*, 6(2):132–142, 2008.

[52] Carl S Goodyear and Gregg J Silverman. Death by a b cell superantigen in vivo vh-targeted apoptotic supraclonal b cell deletion by a staphylococcal toxin. *The Journal of experimental medicine*, 197(9):1125–1139, 2003.

[53] Marisa I Gómez, Aram Lee, Bharat Reddy, Amanda Muir, Grace Soong, Allyson Pitt, Ambrose Cheung, and Alice Prince. Staphylococcus aureus protein a induces airway epithelial inflammatory responses by activating tnfr1. *Nature medicine*, 10(8):842–848, 2004.

[54] Marisa I. Gómez, Maghnus O'Seaghdha, Mariah Magargee, Timothy J. Foster, and Alice S. Prince. Staphylococcus aureus protein a activates tnfr1 signaling through conserved igg binding domains. *Journal of Biological Chemistry*, 281(29):20190–20196, 2006.

[55] Marisa I Gómez, Maghnus O Seaghdha, and Alice S Prince. Staphylococcus aureus protein a activates tace through egfrdependent signaling. *The EMBO journal*, 26(3):701–709, 2007.

[56] J. Hartleib, N. Köhler, R. B. Dickinson, G. S. Chhatwal, J. J. Sixma, O. M. Hartford, T. J. Foster, G. Peters, B. E. Kehrel, and M. Herrmann. Protein a is the von willebrand factor binding protein on staphylococcus aureus. *Blood*, 96(6):2149–56, 2000. Hartleib, J Kohler, N Dickinson, R B Chhatwal, G S Sixma, J J Hartford, O M Foster, T J Peters, G Kehrel, B E Herrmann, M Research Support, Non-U.S. Gov't United states Blood Blood. 2000 Sep 15;96(6):2149-56.

[57] Truc Nguyen, Berhane Ghebrehiwet, and Ellinor IB Peerschke. Staphylococcus aureus protein a recognizes platelet gc1qr/p33: a novel mechanism for staphylococcal interactions with platelets. *Infection and immunity*, 68(4):2061–2068, 2000.

[58] Ernesto Carafoli. Calcium signaling: a tale for all seasons. *Proceedings of the National Academy of Sciences*, 99(3):1115–1122, 2002.

[59] David Chin and Anthony R Means. Calmodulin: a prototypical calcium sensor. *Trends in cell biology*, 10(8):322–328, 2000.

[60] Barbara J Calnan, Sara Biancalana, Derek Hudson, and Alan D Frankel. Analysis of arginine-rich peptides from the hiv tat protein reveals unusual features of rna-protein recognition. *Genes and Development*, 5(2):201–210, 1991.

[61] Ben Berkhout, Robert H Silverman, and Kuan-Teh Jeang. Tat trans-activates the human immunodeficiency virus through a nascent rna target. *Cell*, 59(2):273–282, 1989.

[62] Joseph D Puglisi, Lily Chen, Scott Blanchard, and Alan D Frankel. Solution structure of a bovine immunodeficiency virus tat-tar peptide-rna complex. *Science*, 270(5239):1200–1203, 1995.

[63] Giovanni Lipari and Attila Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *Journal of the American Chemical Society*, 104(17):4546–4559, 1982.

[64] Neil A Farrow, Ranjith Muhandiram, Alex U Singer, Steven M Pascal, Cyril M Kay, Gerry Gish, Steven E Shoelson, Tony Pawson, Julie D Forman-Kay, and Lewis E Kay. Backbone dynamics of a free and a phosphopeptide-complexed src homology 2 domain studied by 15n nmr relaxation. *Biochemistry*, 33(19):5984–6003, 1994.

[65] MICHAEL Wittekind and LUCIANO Mueller. Hncacb, a high-sensitivity 3d nmr experiment to correlate amide-proton and nitrogen resonances with the alpha-and beta-carbon resonances in proteins. *Journal of Magnetic Resonance, Series B*, 101(2):201–205, 1993.

[66] Afshin Karimi, Masazumi Matsumura, Peter E Wright, and H Jane Dyson. Characterization of monomeric and dimeric b domain of staphylococcal protein a. *The Journal of peptide research*, 54(4):344–352, 1999.

[67] Hiroaki Gouda, Hidetaka Torigoe, Akiko Saito, Moriyuki Sato, Yoji Arata, and Ichio Shimada. Three-dimensional solution structure of the b domain of staphylococcal protein a: comparisons of the solution and crystal structures. *Biochemistry*, 31(40):9665–9672, 1992.

[68] Lewis Kay, Paul Keifer, and Tim Saarinen. Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *Journal of the American Chemical Society*, 114(26):10663–10665, 1992.

[69] Bruce A Johnson and Richard A Blevins. Nmr view: A computer program for the visualization and analysis of nmr data. *Journal of biomolecular NMR*, 4(5):603–614, 1994.

[70] Michael Andrec, Gaetano T. Montelione, and Ronald M. Levy. Lipariszabo mapping: A graphical approach to lipariszabo analysis of nmr relaxation data using reduced spectral density mapping. *Journal of Biomolecular NMR*, 18(2):83–100, 2000.

[71] Bruce R Donald. *Algorithms in structural molecular biology.* MIT Press Cambridge, MA:, 2011.

[72] Judit A Losonczi, Michael Andrec, Mark WF Fischer, and James H Prestegard. Order matrix analysis of residual dipolar couplings using singular value decomposition. *Journal of Magnetic Resonance*, 138(2):334–342, 1999.

[73] Martinus HV Werts. Making sense of lanthanide luminescence. *Science Progress (1933-)*, pages 101–131, 2005.

[74] D. Zheng, J. M. Aramini, and G. T. Montelione. Validation of helical tilt angles in the solution nmr structure of the z domain of staphylococcal protein a by combined analysis of residual dipolar coupling and noe data. *Protein Sci*, 13(2):549–54, 2004.

[75] Lindsay N Deis, Charles W Pemble, Yang Qi, Andrew Hagarman, David C Richardson, Jane S Richardson, and Terrence G Oas. Multiscale conformational heterogeneity in staphylococcal protein a: Possible determinant of functional plasticity. *Structure*, 22(10):1467–1477, 2014.

[76] Lishan Yao, Jinfa Ying, and Ad Bax. Improved accuracy of 15n1h scalar and residual dipolar couplings from gradient-enhanced ipap-hsqc experiments on protonated proteins. *Journal of biomolecular NMR*, 43(3):161–170, 2009.

[77] Ke Ruan and Joel R Tolman. Composite alignment media for the measurement of independent sets of nmr residual dipolar couplings. *Journal of the American Chemical Society*, 127(43):15032–15033, 2005.

[78] Qi Zhang, Andrew C Stelzer, Charles K Fisher, and Hashim M Al-Hashimi. Visualizing spatially correlated dynamics that directs rna conformational transitions. *Nature*, 450(7173):1263–1267, 2007.

[79] Yiwen Chen, Sharon L Campbell, and Nikolay V Dokholyan. Deciphering protein dynamics from nmr data using explicit structure sampling and selection. *Biophysical journal*, 93(7):2300–2306, 2007.

[80] Konstantin Berlin, Carlos A Castaneda, Dina Schneidman-Duhovny, Andrej Sali, Alfredo Nava-Tudela, and David Fushman. Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *Journal of the American Chemical Society*, 135(44):16595–16609, 2013.

[81] João Henriques, Carolina Cragnell, and Marie Skepö. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *Journal of Chemical Theory and Computation*, 11(7):3420–3431, 2015.

[82] Christopher Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, 2(6):1201–1225, 1974.

[83] Karsten Kunze and Helmut Schaeben. The bingham distribution of quaternions and its spherical radon transform in texture analysis. *Mathematical Geology*, 36(8):917–943, 2004.

[84] Jared Glover, Gary Bradski, and Radu Bogdan Rusu. Monte carlo pose estimation with quaternion kernels and the bingham distribution. In *Robotics: science and systems*, volume 7, page 97.

[85] Yang Qi. On provable algorithms for determination of continuous protein interdomain motions from residual dipolar couplings. 2016.

[86] Plamen Koev and Alan Edelman. The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation*, 75(254):833–846, 2006.

[87] Anthony K Yan, Christopher J Langmead, and Bruce Randall Donald. A probability-based similarity measure for saupe alignment tensors with applications to residual dipolar couplings in nmr structural biology. *The International Journal of Robotics Research*, 24(2-3):165–182, 2005.

[88] Pau Bernadó, Efstratios Mylonas, Maxim V Petoukhov, Martin Blackledge, and Dmitri I Svergun. Structural characterization of flexible proteins using small-angle x-ray scattering. *Journal of the American Chemical Society*, 129(17):5656–5664, 2007.

[89] Jo A Capp, Andrew Hagarman, David C Richardson, and Terrence G Oas. The statistical conformation of a highly flexible protein: Small-angle x-ray scattering of s. aureus protein a. *Structure*, 22(8):1184–1195, 2014.

Biography

Yang Qi was born on Sep 24th, 1988 in Xinxiang, P.R. China. He received his bachelor's degree in Fundamental Sciences of Biology and Chemistry from Tsinghua University in 2010. He earned his master's degree in Computer Science and his doctoral degree in Biochemistry both from Duke University in 2016. His master advisor was Professor Bruce R. Donald and his Ph.D. advisor was Professor Terrence G. Oas. His research is focused on determining the continuous motions of biomolecular interdomain dynamics from NMR.