Determination of Population-Weighted Dynamic Ensembles of RNA Using NMR Residual Dipolar Couplings

by

Shan Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in The University of Michigan
2014

Doctoral Committee:

Professor Hashim M. Al-Hashimi, Chair
Professor Charles L. Brooks III
Professor Heather A. Carlson
Professor Nils G. Walter

Acknowledgements

I would like to first thank Professor Hashim Al-Hashimi. His outstanding scientific insights remarkably broaden my vision of study of nucleic acids and his constant encouragements largely push me move forward to better scientific achievements. I would like to appreciate his great tolerance for my initially non-proficient English and kind guidance on teaching me scientific writing skills. I am grateful for my committee members: Professor Charles Brooks, Professor Heather Carlson and Professor Nils Walter, for their valuable comments on my projects that help me to improve my research to a better one.

I am also grateful for all the previous group members for setting a high standard that I have to follow and providing valuable assistance for me to overcome major difficulties encountered in my research. In particular, I would like to thank Dr. Elizabeth Dethoff and Dr. Catherine Eichhorn who taught me the initial knowledge of RDCs that is the main focus of my thesis; Dr. Qi Zhang for providing very useful and helpful information about tensor analysis that is of central importance in my project; Dr. Jameson Bothe for helping me quickly get used to the atmosphere of the group when I joined the group in 2009. I would like to thank all current members in Al-Hashimi lab especially Dr. Loic Salmon who is my collaborator in my first published paper; Dr. Yi Xue who provides me many useful advises in scientific programming; Dr. Bharathwaj Sathyamoorthy who taught me many necessary knowledge as well as skills on implementing NMR experiments.

I have to thank my whole family for supporting my decision of trying to get a phD degree in US. Without the supports and encouragements from my family, it is not possible for me to overcome difficulties in life and sustain a long-term focus and enthusiasm on my research. Finally, I want to thank my girlfriend, Qiong Gao, who enters my life during my lowest time and

provides endless love and strength of spirit that always encourage me to move forward toward my dream with courage and confidence.

Table of Contents

Acknowledgements ii

List of Tables vi

List of Figures vii

List of Appendices ix

Abstract x

Chapter 1 Introduction 1

1.1 RNA Dynamics 1

1.1.1 Roles of RNA Dynamics in Biology 1

1.1.2 Helix-Junction-Helix (HJH) Motif 3

1.2 Construction of Dynamic Ensembles of RNA 5

1.2.1 Introduction and Historical Perspective 5

1.2.2 Experimental Constraints 6

1.2.3 Ensemble Determination Methods 11

1.2.4 Assessing Accuracy of Dynamic Ensembles 15

1.3 Probing RNA Dynamics Using Residual Dipolar Couplings (RDCs) 17

1.3.1 Theory of RDCs 17

1.3.2 Measurement of RDCs 23

1.3.3 Dynamic Interpretation of RDCs 27

1.4 References 31

Chapter 2 Measuring Similarity Between Dynamic Ensembles 37

2.1 Introduction 37

2.2 Methods 38

2.2.1 Jensen-Shannon Divergence (Ω^2) and *S*-score 38

2.2.2 Sample and Select (SAS) approach 39

2.2.3 Evaluating quality of inter-helical ensembles determined with increasing input RDCs 40

2.2.4 Binning inter-helical orientations 41

2.2.5 Analysis of MD-trajectory-based ensembles 42

2.3 Results and Discussion 43

2.4 Conclusion 53

2.5 References 54

Chapter 3 Characterizing Uncertainty in Dynamic Ensmebles of Biomolecules Determined Using Residual Dipolar Coupings 56
3.1 Introduction 56
3.2 Methods 58
3.2.1 Constructing ensembles using Sample and Select (SAS) 58
3.2.2 Determining accuracy of predicted ensembles 59
3.3 Results and Discussion 59
3.3.1 Conformation pool 59
3.3.2 Experimental error based uncertainty 60
3.3.3 Ensemble size based uncertainty 62
3.3.4 Applications 67
3.4 Conclusions 72
3.5 References 73
Chapter 4 Preliminary Study of Dynamics of Exon Splicing Silencer 3 of HIV 1 RNA 75
4.1 Introduction 75
4.2 Materials and Methods 76
4.2.1 Preparation of ESS3 sample 76
4.2.2 NMR spectroscopy 77
4.2.3 Analysis of measured RDCs 77
4.3 Results and Discussion 78
4.4 Conclusion 85
4.5 References 86
Chapter 5 Conclusions and Future Directions 87
5.1 Conclusions and Future Directions 87
5.2 References 92
Appendices 94

List of Tables

Table 1.1 Elements of the second rank Wigner rotation matrix..... 29

Table 4.1 RDCs of ESS3 measured at 25^{o}C under pH=5.5 and 6.5 81

List of Figures

Figure 1.1 RNA conformation transitions..... 2
Figure 1.2 Secondary structure and schematic graph of helix-junction-helix motifs with different types of junctions. 4
Figure 1.3 Flowchart of SAS approach..... 15
Figure 1.4 Relative orientation between internuclear vector (CH bond vector as an example) and the magnetic field..... 19
Figure 1.5 Angular dependence of bond vector and magnetic field in molecular frame..... 21
Figure 1.6 Partially aligning biomolecules (RNA as an example) in solution 25
Figure 1.7 Measurement of RDCs 26
Figure 1.8 Angular dynamic information contained in RDCs. 30
Figure 2.1 Measuring population overlap and structural similarity between ensembles. 44
Figure 2.2 Measuring similarity between ensembles containing outliers 45
Figure 2.3 Prediction of ensembles using increasing RDC input sets. 46
Figure 2.4 Measuring similarity between ensembles using *S*-score (*S*) and χ^2.. 47
Figure 2.5 Comparing MD-generated and NMR-RDC selected ensembles of HIV-1 TAR.. 49
Figure 2.6 Comparison of experimentally measured and calculated RDCs using Anton MD trajectory, randomly selected ensemble, and SAS selected ensemble.. 50
Figure 2.7 Investigating the selection power of the SAS approach.. 53
Figure 3.1 Uncertainty in determined ensembles arising from experimental errors..... 62
Figure 3.2 Uncertainty in determined ensembles arising from ensemble size..... 64
Figure 3.3 Prediction of MD trajectory of HIV 1 TAR.. 66
Figure 3.4 Prediction of inter-helical dynamic ensemble of HIV 1 TAR. 70

Figure 3.5 Secondary structure and 1D distribution of predicted ensembles of ARG-bound and A22G-U40C HIV 1 TAR. 71
Figure 4.1 Secondary structure and resonance assignments of ESS3. 79
Figure 4.2 Correlation between RDCs measured under pH=5.5 and 6.5 82
Figure 4.3 RDCs measured at 25^{o}C under pH=5.5 and 6.5 as a function of the secondary structure of ESS3 83
Figure 4.4 Ensemble size test for constructing population-weighted ensemble of ESS3. 84

List of Appendices

Appendix 1 Sample and Select (SAS) Approach.....94

Appendix 2 REsemble Algorithm for Measuring Ensemble Similarity 107

Abstract

RNA undergoes large-scale conformational transitions in response to cellular cues, including changes in physiological conditions such as temperature and pH, recognition of proteins and ligand molecules, and RNA synthesis itself to perform a wide range of regulatory functions. A predictive understanding of how RNAs carry out their functions requires studies that go beyond static structure determination toward characterization of dynamic ensembles representing the broad RNA structure landscape. This thesis describes the development and application of Nuclear Magnetic Resonance techniques that rely on measurements of residual dipolar couplings (RDCs) for partially oriented RNAs in determination of dynamic ensembles.

The ability to assess methods for ensemble determination hinges on the ability to compare the similarity between two ensembles. We have developed a new method that successfully captures both population overlap and structural similarity that relies on measuring population overlap as a function of the bin size used to bin ensemble distributions. Using this new method, we showed fundamental limitations in conventional approaches for measuring ensemble similarity and also find unexpected similarities between ensembles determined for HIV-1 TAR RNA with the use of NMR RDCs and computational molecular dynamics simulations.

Using the new method for measuring ensemble similarity, we examined the accuracy with which ensembles can be determined with the use of RDCs under ideal conditions involving two domains and five perfectly orthogonal datasets. We found that the two factors resulting in uncertainty in determined dynamic ensembles of RNA are the experimental uncertainty of measured RDC as well as the ensemble size used to construct the ensemble. We developed an approach that takes into account these sources of uncertainty and applied it in the determination of ensembles for the bulge containing HIV-1 TAR in free state and ligand-bound states, for a

TAR mutant with distinct dynamics, and for the HIV-1 ESS3 RNA containing an AC wobble base pair.

Chapter 1

Introduction

1.1 RNA Dynamics

1.1.1 Roles of RNA Dynamics in Biology

RNA is conventionally categorized into three types: messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). These RNAs play essential roles in protein synthesis. However, over the past three decades, many new non-coding RNAs (ncRNAs) have been discovered such as riboswitches and microRNAs (miRNA)[1-5] that play essential roles in regulating the expression of genes. Such ncRNAs have significantly broadened the role of RNA in biology, spurring new biotechnological applications, including RNA-targeted drug discovery[6-10] and the design of RNA-based devices such as sensors[11-13].

Many coding and non-coding RNAs carry out functions through large changes in the RNA conformation that are typically trigged by cellular cues including recognition of protein and ligands, changes in physiological conditions such as temperature and pH or even RNA synthesis itself[14-22]. A quintessential example of the relationship between the conformation transition and the biological function of RNA is the *Tetrahymena* group I intron[20,23], which is a 400 nucleotides (nt) ribozyme that catalyzes its own excision from the corresponding pre-mRNA. In the self-splicing reaction, the 5'-splice site of the intron base pairs with its internal guidance sequence (IGS) to form the P1 helix, connecting to the rest of the intron through a single stranded junction J1/2. The self-splicing reaction initiates with the change of the orientation of P1 helix, which docks into the intron's conserved catalytic core through tertiary interactions, forming the "closed state" of the intron that allows the ligation of the exons as the following step. After the catalytic reaction, another undocking transition allows the P1 helix to return to the "open state" (the conformation without tertiary interactions with catalytic core)

(Figure 1.1). The rate of the interconversion between "open" and "closed" state is slow (~s^{-1}) and in some cases is the rate-limiting step of the entire self-splicing reaction[20].

Another well-characterized example of conformation changes in RNA is the hierarchical assembly of the central domain of 30S subunit of ribosome of prokaryote, which is a typical ribonucleoprotein (RNP) complex. The binding of ribosomal protein S15 to 16S rRNA induces the change of relative orientations of different helical domains that favor the subsequent binding of ribosomal protein S6 and S18 and therefore initiates the ordered assembly of the central domain of the 30S ribosomal subunit[24-30] (Figure 1.1). Premature binding of S6 and S18 to the unbound 16S rRNA may be disfavored partially because of the entropic penalty that has to be compensated due to the change of the inter-helical orientations in 16S rRNA[22].

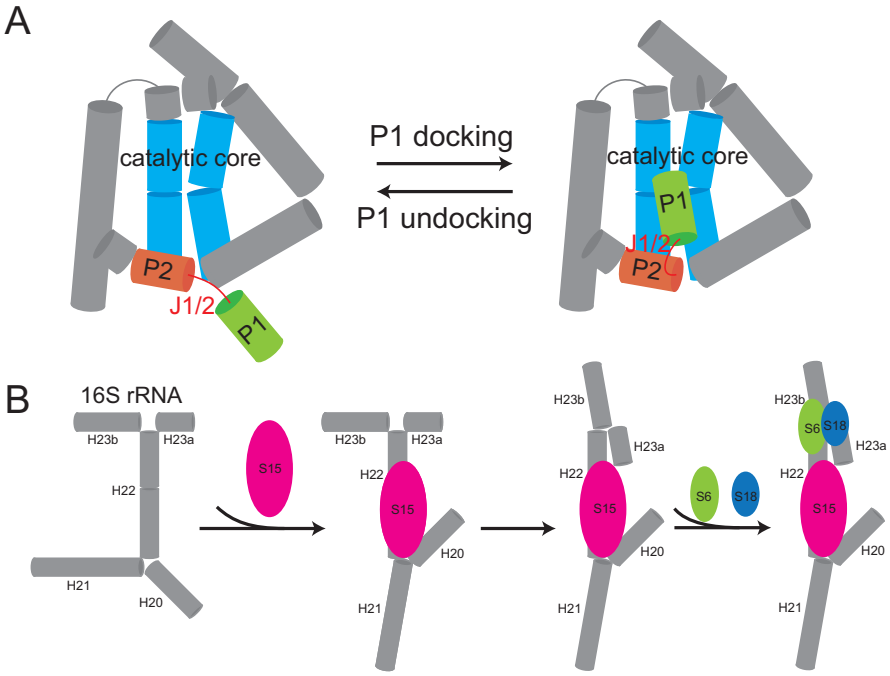

Figure 1.1 RNA conformation transitions. (A) Docking of P1 helix of *Tetrahemena* group I intron into catalytic core as the initial step of the self-splicing[13]; **(B)** hierarchical assembly of central domain of the 30S ribosomal subunit induced by ribosomal protein S15.

RNA dynamics can be complex, involving changes at the secondary and tertiary structure level as well as local changes in base-pairing and jittering dynamics[18,20,22,31-34]. However, studies have shown that RNA structure is very hierarchical composed of reoccurring motifs that often fold independently of tertiary context. Thus, RNA structure can be divided into different building blocks[35], allowing characterization of RNA structure in a divide-and-conquer manner. Analogously, the dynamics of these building blocks can be studied to obtain insights into the overall dynamic behavior of RNA and the properties that enable large conformational transitions to take place in a robust biologically specific manner[35-37]. Therefore characterization of the dynamics of different building blocks of RNA and the correlations between them is the basic and key step for understanding RNA dynamics.

1.1.2 Helix-Junction-Helix (HJH) Motif

In the hierarchical RNA structures, helix-junction-helix (HJH) motifs, composed of helices adjoined by intervening junctions, are fundamental and ubiquitous building blocks[36-39]. HJH motifs such as bulges, internal loops and higher order junctions[39] (Figure 1.2) direct the orientation of helices and changes in inter-helical orientation that are often observed in RNA directed functions[40-45] (Figure 1.1).

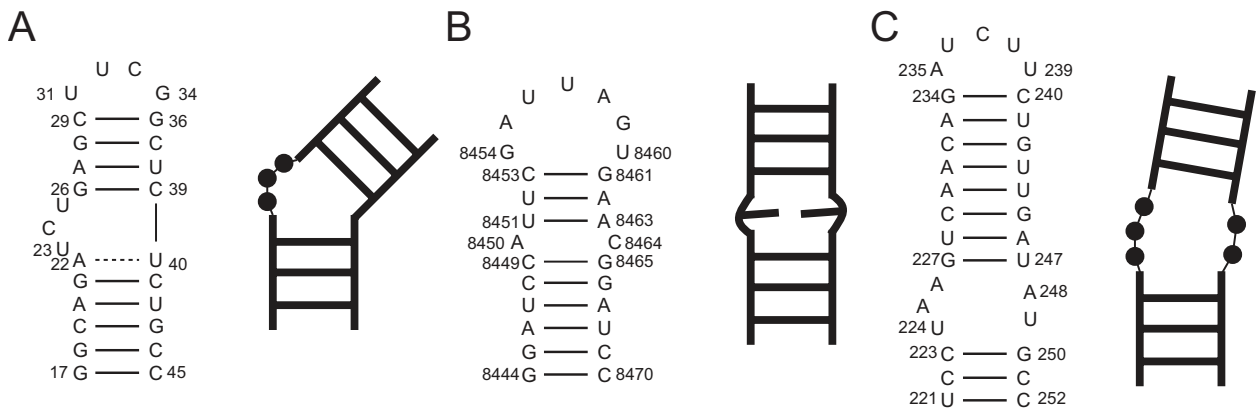

Figure 1.2 Secondary structure and schematic graph of helix-junction-helix motifs with different types of junctions. (A) HIV I TAR with bulge junction; **(B)** exon splicing silencer 3 (ESS3) with non-canonical base pair as the junction; **(C)** P6b domain of P4P6 segment in *Tetrahymena* group I intron with internal loop as the junction. For simplicity, the unpaired nucleotides in junction are shown as closed circles and only three base pairs are shown for each HJH motif in the corresponding schematic graph.

HJH motifs have been extensively studied using a variety of biophysical methods in order to obtain insights into junction topology and flexibility and their roles in defining RNA 3D structure and dynamics. By using electrophoretic gel mobility measurements, electron microscopy and simulation-aided transient electric birefringence (TEB) method, Lilley[44,45], Griffith[46] and Hagerman[40-42] demonstrated about two decades ago that junctions can induce directional bends in HJH motif in a sequence-context-dependent manner and quantitatively estimated the directional bend angles across junctions. Subsequent studies utilizing fluorescence resonance energy transfer (FRET)[47,48] and NMR spectroscopy[18,19,49] revealed that junctions do not induce static bends but a range of conformations of HJH motif with distinct inter-helical orientations. This inter-helical flexibility induced by junctions allows RNA to undergo conformational transition to perform its functions during a variety of biological processes such as catalysis[20,47], ligand binding[11,18], regulation of gene expression[23] and transcription[17,22].

More recently, NMR studies of HIV-1 TAR containing a 3-nt bulge HJH motif led to the discovery that the relative orientation of helices in two-way junctions is severely restricted by simple topological constraints – including steric constraints and connectivity constraints posed

due to finite length of connecting single-strands[37]. These constraints alone restrict the allowed orientation of helices, as described using three Euler angles, to <15% of the total allowed space[37-39]. These studies suggest that topological constraints encoded by the topology of HJH motifs strongly help define the range of inter-helical orientations that can be sampled. A coarse grain model has been developed recently that makes it possible to simulate the allowed inter-helical space for a given HJH motif[50]. However, in order to achieve a predictive understanding regarding RNA dynamic behavior and changes in inter-helical orientation, it is important to also measure the relative free energies of conformations within this allowed space; or alternatively, to determine dynamic ensembles of RNA describing the population-weighted distributions of inter-helical orientations.

1.2 Construction of Dynamic Ensembles of RNA

1.2.1 Introduction and Historical Perspective

Guided by the principle of "structure determines function", much effort has been directed toward characterizing structures of RNA. Although encouraging results have been reported, it was recognized in recent two decades that a single static structure is not sufficient for dissecting the biological functions of an RNA molecule because it requires large rearrangements of RNA conformation that can only be captured by an ensemble of conformations[18,19].

Many dynamic properties of RNA are encoded in the free energy landscape, which provides both thermodynamic and kinetic descriptions of the ensemble of conformations sampled by a RNA molecule in solution[51,52]. The population of a given conformation in an ensemble depends on its relative free energy, while the rates at which two conformations inter-convert is determined by their free energy barrier of separation. Cellular cues perturb the free energy landscape and trigger conformation transitions of RNA that lead to performance of biological functions. Insights into the free energy landscape can be obtained by detecting global and local motions of RNA over a broad range of timescales, which can provide direct information regarding the populations of different conformations and the rates of inter-conversion between them. Determining the entire free energy landscape and all details of the conformational ensemble is generally not feasible, because it is very challenging if not

impossible to measure the rates of transitions between many conformations exist in low abundance and/or for short periods of time that disallow experimental detection. Alternatively, studies have attempted to describe the distribution of the most populated conformations (typically >1%) representing the lowest energy minima along the free energy landscape. The focus has been primarily on the populations of these relatively highly populated conformations but less on the rates at which they interconvert that has proven to be challenging to detect experimentally.

Determining conformational ensembles for complex biomolecules or even their building blocks presents a significant challenge compared to characterization of high-resolution static structure using techniques such as X-ray diffraction. This is because: first, a much larger number of parameters need to be defined for an ensemble of conformations compared to a static structure, and measuring these parameters is in general challenging[53]; second, many conformations in a dynamic ensemble exist in low abundance and/or for very short periods of time, and therefore are challenging to detect experimentally[54]; finally, it can be very difficult to assess the accuracy and precision of a determined ensemble[55]. To overcome these challenges and furthermore accurately determine conformational ensembles of biomolecules, efficient ensemble determination methods utilizing proper experimental constraints and effective methods for evaluating the accuracy of the determined conformational ensemble are necessary.

1.2.2 Experimental Constraints

Experimental data used in ensemble determination as constraints have to meet three requirements. First, the data have to be sensitive to the structural degrees of freedom and time scale of the dynamics that are of interest. In general, the experimental parameters measured may be sensitive to different aspects of structure, e.g. global *versus* local structure, or rotational *versus* translational degrees of freedom. Construction of an accurate ensemble often requires the combination of different types of data, and this in it of itself, can be a challenge. Second, the experimental data must can be robustly computed for a given conformational ensemble. Depending on the type of data, it may be necessary to have additional information regarding for example constants that factor into the measurements. Finally, different types of data may be

sensitive to different timescales, complicating their combination in the ensemble determination. Several types of experimental data that are frequently used in ensemble determination are introduced below for which both advantages and disadvantages are discussed.

Small Angle X-ray Scattering

Small-Angle X-ray Scattering (SAXS) is an ideal tool to characterize global aspects of conformational ensembles of biomolecules. It is a technique in which the elastic scattering of X-rays is measured at very low angles (typically 0.1º-10º), thus providing information about the overall shape and size of biomolecules that are 5 nm to 25 nm in size, with lower scattering angles allowing larger dimensions to be resolved. Unlike X-ray crystallography, SAXS does not require a crystalline sample and can be performed under a variety of solution conditions[56]. Recent developments in SAXS by attaching nano-probes to specific segments of biomolecules, for example helices of nucleic acids, allow detection of conformation transition of biomolecules at sub-domain level, which dramatically enhance the spatial sensitivity of SAXS. However, in general owing to the random reorientation and vibration of molecules, ensemble averaging leads to structural information of lower resolution as compared to X-ray crystallography.

For a biomolecule interconverting between several conformations, the SAXS profiles will in principle represent the sum of contributions from each conformation in the sample, because the light matter interaction occurs at timescales much faster than the conformation changes. This renders SAXS insensitive to the precise timescales of the motion, allowing for easier interpretation in constructing ensembles[57]. However, due to its low resolution, for biomolecules that do not exhibit large conformation changes, the observed scattering profile can often be interpreted using a single conformation[56].

Chemical Shifts

The Nuclear Magnetic Spectroscopy (NMR) chemical shift (CS) is one of the simplest but most important parameters to measure using solution state NMR. The chemical shift is the resonant frequency of a nucleus of a biomolecule relative to a reference frequency. It is determined by the effective magnetic field experienced by a given nucleus. Besides the applied

magnetic field, a given nucleus also experiences local magnetic fields induced by currents arising from movements of electrons in molecular orbitals. For a given nucleus (^{1}H, ^{13}C, ^{15}N and ^{31}P), this electronic distribution and corresponding electronic movement depend on local aspects of the structure, including bond lengths, dihedral angles, hydrogen bonding as well as ring current effects arising due to circulation of π-electrons in the local conjugate groups (e.g. aromatic nucleobases in RNA), magnetic anisotropy, and other electrostatic effects[58]. For a dynamic ensemble of conformations, the observed CS of a given nucleus corresponds to a population-weighted average of the corresponding CS over all conformations given that the inter-conversion between conformations is faster than the corresponding difference between their chemical shifts (fast exchange limit). CSs are accurate probes of local structure and dynamics, but generally provide very limited information about long-range structure and dynamics.

Powerful methods for computing chemical shifts based on 3D structure have been well developed for determination of protein conformational ensembles[59,60], attributing to the growth in the database of protein structures with corresponding NMR ^{1}H, ^{13}C, and ^{15}N resonance assignments. In contrast, the database of nucleic acid structures with corresponding NMR resonance assignments is relatively small. Therefore, methods to compute CSs from a given nucleic acid structure remain challenging and difficult to test rigorously. Several approaches have been developed to compute ^{1}H[61-63] based on nucleic acid structure and the accuracy with which chemical shifts can be computed from structure based on these approaches is sufficiently high, allowing determination of 3D structure and characterization of motions in locally mobile regions[64]. One drawback in some of these approaches is that the experimental CS database is parameterized assuming single static structures, making it more challenging to identify the cases in which dynamics is not negligible.

Scalar Couplings

NMR scalar couplings arise from coupled interactions between the electrons and nuclear spins of two bonded nuclei, which result in the splitting of NMR resonances. The magnitude of scalar couplings depends on the nuclei involved, the number of bonds separating the nuclei and the intervening dihedral angle for three bond scalar couplings (^{3}J). Three bond scalar couplings

$^3J_{HH}$, $^3J_{HC}$, $^3J_{HP}$, and $^3J_{CP}$ have been used in probing dihedral angles involving the sugar, base, and phosphodiester backbone and are often used in NMR structure determination of RNA[65,66] using appropriately parameterized Karplus equations. Challenges in parameterizing Karplus relations using databases containing averaged scalar couplings, along with their limited structural resolution and inherent degeneracies have limited their widespread application. In general, scalar couplings exhibit similar sensitivities to the timescales of motional averaging (up to the millisecond) as RDCs (see below) and CSs.

RDCs and RCSAs

Many NMR interactions such as dipolar couplings and chemical shift anisotropy (CSA) are second rank interactions that depend on the orientation of dipolar and CSA tensors centered on nuclei of interest relative to the applied magnetic field. In solution NMR, a given nucleus *i* experiences the sum of the external magnetic field as well as the magnetic field generated by a directly bonded nucleus *j* and other nuclei in the vicinity. The latter contribution inversely proportional to the cube of the distance separating the two nuclei, which is the bond length for directly bonded spins, as well as on the angle, θ, between the inter-nuclear bond vector and the applied magnetic field, as described by the angular term $\frac{3cos^2\theta-1}{2}$. In isotropic solution, this magnetic dipole-dipole interaction, and in particular the angular term, averages to zero due to random Brownian rotational diffusion, and indeed, the loss of these otherwise very large interactions is one of the main reasons solution NMR exhibits high-resolution and sharp lines[53]. However, by introducing a small degree of alignment on a biomolecule, corresponding to alignment level of ~10^{-3}, one can break the isotropic averaging and re-introduce a small fraction of the dipolar interaction while retaining the high quality of solution state NMR spectra. This dipolar interaction manifests as an additional contribution to one-bond ^1J-couplings for two directly bonded nuclei and is referred to as a 'residual dipolar coupling' (RDC). RNA samples and in general nucleic acid samples can be aligned spontaneously upon applying external magnetic field due to interactions with the magnetic field itself[67,68], or by dissolution in an alignment medium such as filamentous bacteriophage[69,70]. The most commonly measured RDCs are usually the ones between two directly bonded nuclei[71]. As the distance between two directly

bonded nuclei can be well approximated as a constant[72,73], RDCs solely provide useful orientation information of the bond vector relative to the external magnetic field but are not sensitive to any translational degree of freedom. More detailed introduction about RDCs will be provided in section 1.3.

Partial alignment also leads to incomplete average of the anisotropic component of the chemical shift, allowing measurements of residual chemical shift anisotropies (RCSAs) as changes in chemical shift relative to the isotropic case[74,75]. These data report on the orientation of the CSA tensor, centered on the nucleus of interest, relative to the alignment tensor frame. For RNA molecules, RCSAs of backbone ^{31}P [76] and nucleobase ^{13}C [72,77] provide complementary information to RDCs but the interpretation of RCSAs generally requires accurate knowledge of the chemical shift anisotropy (CSA) tensor, which vary from site to site, and are very challenging if not impossible to determine *a priori*.

Other types of data

There are other types of data that have been used in structure but not in ensemble determination or in other biomolecules such as proteins but not in RNA that we expect will play important roles in the determination of RNA ensembles in the future. NMR data include measurements of ^{15}N and ^{13}C spin relaxation that report primarily on the dynamics of bond vectors in biomolecules occurring at picosecond to nanosecond timescales[78]; Paramagnetic Relaxation Enhancements (PREs) which depend on the distance between a given nucleus and an attached paramagnetic probe, and can report on low populated short-lived conformations in an ensemble[79,80]; and Nuclear Overhauser Effect based cross-relaxations (NOEs) that report on the network of proton-proton distances (and orientations for anisotropic overall diffusion) in a qualitative manner[81,82]. Such relaxation data not only depend on the distribution of conformations in the ensemble, but also have a complex correlation with the rates with which conformations interconvert and the timescales for overall rotational diffusion. Although this additional information could be involved in the future to extract timescale information, it currently complicates ensemble determination. In addition, NMR relaxation dispersion

techniques[83,84] allow for visualization of low-populated (<10%) and/or short-lived (lifetimes in the range of millisecond to microseconds) conformations in RNA[54].

Förster Resonance Energy Transfer (FRET)[32,85] and Electron Paramagnetic Resonance (EPR)[86] can be used to obtain distance information between fluorophores and spin labels respectively that are specifically attached to RNA. These data also depend on the orientation and dynamics of the fluorophores or spin labels, and approximations often have to be made to extract distance information[86-91]. Powerful single molecule approaches, such as single molecule FRET (smFRET) can be used to directly measure transitions within a single molecule and obtain information about the underlying conformations and their rates of inter-conversion that is difficult to obtain from ensembles[32,88,92].

1.2.3 Ensemble Determination Methods

Thus far, two approaches (see below) have been developed to construct conformational ensembles of biomolecules based on experimental measurements. In one approach, the experimental data is directly incorporated in generating the conformational ensemble while in the other approach, the experimental information is introduced as constraints in a second step after the generation of a conformation pool *a priori*. Both approaches heavily rely on proper parameterization in either computational modeling or simulation force fields. Although there are a growing number of studies showing that long-timescale or enhanced MD simulations quantitatively predict experimental data measured in proteins[93], nucleic acid force fields remain underdeveloped and poorly tested. The challenges include proper treatment of electrostatic effects and polarization involving the phosophodiester backbone and interactions with metal ions[94,95].

Restrained Molecular Dynamics

Restrained molecular dynamics provides a way to directly incorporate experimental data into molecular dynamic force fields. In this approach, experimental constraints are included in term of additional penalty functions or pseudo-energies terms in the default force field. Here, the data reproduction is only assessed on average, over the ensemble of conformations determined at

each step of the procedure, or sometimes over a time window in a trajectory. The expressions used for the penalty function vary depending on the type of experimental data, but they often assume a quadratic form[96]:

$$E_j = w_j \sum_i (\langle D_{calc} \rangle - D_{obs})^2 \quad (1.1)$$

where w_j represents the weight of a given data set j and D correspond to different experimental data points in this dataset. The simulation is therefore guided by the incorporated experimental data and results in an ensemble of conformations that can reproduce the experimental data possibly within experimental uncertainty.

By introducing an experimental pseudo-potential, this approach can direct the sampling towards conformations that may otherwise not be favored by the force field, but it also remains limited to the use of experimental data that can be efficiently computed at each step of the simulation. Moreover the number of degree of freedom in the calculated ensemble tend to be larger than the number of experimental constraints, leaving open the possibility of overfitting of experimental data, in which noises instead of experimental data are fitted. Therefore those procedures are optimal when there is sufficient experimental data for not only defining the number of degrees of freedom but also allowing for cross-validation. Another disadvantage is that the introduction of experimental pseudo-potential can potentially introduce perturbations without clear physical significance to the free energy landscape, which direct the simulation in an unpredictable way and prevent convincing interpretation of the resulting conformational ensemble[97].

Data guided selection of conformational ensemble from a pool

An alternative approach involves using the experimental data to guide the selection of conformations from a pool that is generated using computational methods, such as MD simulations or structure based exhaustive search[37-39,98]. The approach involves two steps: (1) generation of a pool of conformations that sufficiently sample the free energy landscape and (2) selection of a sub-ensemble that can reproduce experimental data from the conformation pool[81,94]. This approach is referred to as 'sample and select' (SAS)[94].

The success of SAS-based approaches highly depends on sampling of the starting conformation pool. For example, if a native conformation is not included in the starting pool, it will never be included in any determined ensemble. For RNA molecules, starting pools have been generated using standard MD simulations[19,99], replica exchange molecular dynamics simulations for enhancing sampling[100], and by performing an exhaustive grid-search when determining inter-helical ensembles involving a small number of structural degrees of freedom[18]. Monte-Carlo based approaches have been used in estimating RNA dynamic amplitudes but so far have not been used to construct conformation pools that can be used in SAS-based approaches for explicitly construction of conformational ensemble of RNA[101].

In the second step, sub-ensembles that can reproduce experimental data are selected from the conformation pool. The selection procedure can be accomplished using a variety of search algorithms including simulated annealing[18,19,94,99] and genetic algorithms[57,102,103]. For example, in the simulated annealing approach, N conformations are randomly selected from the conformation pool to generate trial sub-ensembles and the agreement between measured and predicted data is computed. Next, one of the conformations in the trial sub-ensemble is randomly chosen and replaced by another conformation randomly chosen from the remaining conformation pool, and the agreement with measured data is re-examined. The newly selected conformation is either accepted or rejected based on the Metropolis criteria and several iterations are carried out until convergence is reached, defined as achieving agreement with the measured data that is equal to or smaller than the experimental error. The ensemble size (N) is then incrementally increased in steps of 1 from N=1 until the convergence is reached and proper ensemble sizes are selected for constructing the ensemble (see Chapter 2 and 3). Given a selected ensemble size, this procedure can be repeated for sufficient iterations, with the conformations selected over all runs pooled together to obtain a final population-weighted conformational ensemble (Figure 1.3).

The SAS approach provides a natural means for evaluating nucleic acid force fields and for identifying potential pitfalls that can be addressed in future developments[19,99]. As nucleic acid force fields improve, we can anticipate that the SAS approach can be extended to include experimental data that are sensitive to broader timescales. The disadvantage is that, as stated

above, the quality of the determined ensemble strongly relies on the sampling of the conformation pool.

Other approaches for constructing ensembles

Although not yet implemented for nucleic acids, several approaches have been developed to improve sampling in MD simulations in characterizing protein dynamics. Enhanced sampling is particularly important when using NMR RDCs and SAXS data, which have timescale sensitivities that generally exceed those accessible by conventional MD simulations. These approaches include Accelerated Molecular Dynamics (AMD)[97,104], in which the rates of transition between distinct conformations is increased by adding a continuous non-negative bias potential to the energy surface and replica exchange molecular dynamics (REMD)[168], in which several simulations are run in parallel at different temperatures and allowed to exchange population or energy distribution according to certain algorithms. The resulting structural ensemble from both approaches can be used alone or as a relevant conformation pool for a SAS protocol.

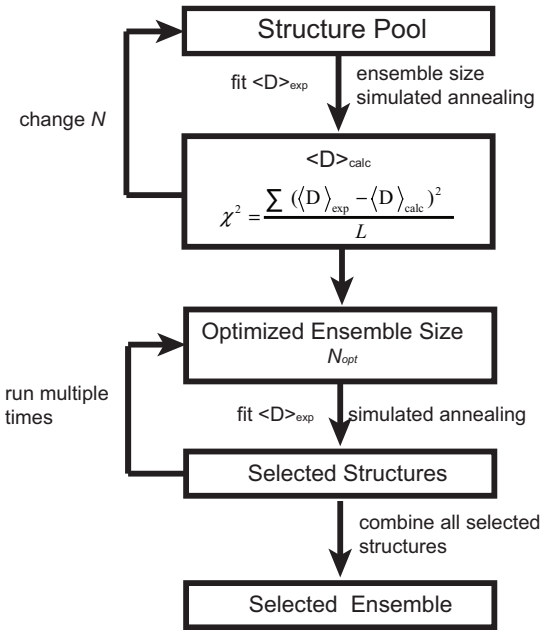

Figure 1.3 Flowchart of SAS approach.

1.2.4 Assessing Accuracy of Dynamic Ensembles

The construction of ensembles using experimental data represents an ill-defined problem because many different ensembles can reproduce the experimental observable. It is therefore important to assess accuracy and precision in the determined ensembles when interpreting the determined ensembles.

Cross-validation

Cross-validation is one of the most commonly used approaches for testing the quality of a determined ensemble. In this method, a subset (typically 10%) of the total experimental data is excluded from the ensemble determination process and the accuracy of the determined ensembles is assessed by how well it predicts the excluded data[105]. This provides a straightforward approach for identifying cases where the data is overfitted and for testing how well a given set of data can uniquely define an ensemble[99,103,106,107]. An important aspect of cross-validation is the choice of the excluded dataset. In general, the excluded data can either

correspond to data drawn randomly across all different types of data or correspond to one type of data among many. Regardless of the approach, the data used in ensemble determination have to carry the information needed to build a reasonable ensemble. On the other hand, choosing excluded data that is highly correlated to data used in ensemble determination should be avoided, as it may not provide stringent tests, e.g. two sets of highly correlated RDCs data sets (resulting from very similar alignments, see section 1.3.2).

A notable disadvantage of cross-validation is that it is probably unable to distinguish distinct ensembles that can reproduce the experimental data at a similar accuracy level. This is because cross-validation is an indirect test of the accuracy of the predicted ensembles, providing little direct information about the ensemble distributions and therefore can hardly avoid the degeneracy problem, which is very common in ensemble determination (see Chapter 3).

Tests on synthetic data

An alternate approach for assessing the ability of a given experimental data set to determine any aspect of an ensemble is to run simulations in which the synthetic data, corresponding to the same data that is measured experimentally, is used to reconstruct a known 'target ensemble'. The target ensemble should represent a reasonable challenge to the data. For example, target ensembles that are simply generated by randomly selecting conformations from a pool present a simpler sampling problem to ensemble determination methods as compared to selecting target ensembles that over emphasize low-populated regions in the conformation pool. In addition, the experimental data has to be properly noise corrupted to reflect experimental uncertainties and the noise-corrupted synthetic data is then evaluated for its ability to reproduce the target ensemble. A wide variety of approaches such as the *S*-matrix and Jensen-Shannon Divergence (JSD) have been used to quantitatively assess the similarity between two ensembles by comparing the histogram distribution of the degree of freedom of interest with specific bin sizes and therefore to assess how well a given set of data reproduces a target ensemble[55,108-111]. However it is shown by a recent study that these conventional metrics or methods do not fully capture ensemble similarities as they are insensitive to the magnitude of the structural differences in non-overlapping ensemble distributions, which can potentially result in wrong conclusions.

This problem can be largely resolved using recently developed REsemble method[112] in which the conventional metrics are calculated at systematically varied bin sizes instead of a arbitrarily chosen bin size (see Chapter 2).

Monte Carlo Analysis

Monte Carlo analysis is a very general procedure to indirectly assess the accuracy and the precision of a given model and can therefore be applied to the sampling of a conformational ensemble[106]. Here, an experimentally determined ensemble of conformations is typically treated as a target ensemble and used to generate noise-corrupted synthetic data. Next, the target ensemble is determined using several rounds of ensemble determination using synthetic data corresponding to the values calculated from the target ensemble and noise-corrupted independently for each simulations. The uncertainties of the degrees of freedom are then evaluated from the target and corresponding predicted values. Although this method can be computationally expensive and does not provide direct comparison of the distributions of the target and predicted parameters, Monte Carlo simulations can be applied for any determined ensemble to estimate the uncertainties in each determined structural parameter[99,106].

1.3 Probing RNA Dynamics Using Residual Dipolar Couplings (RDCs)

1.3.1 Theory of RDCs

Among biophysical techniques that have been developed and applied to study RNA dynamics[22,32,58,113], the measurement of residual dipolar couplings (RDCs) in partially aligned systems[67,114,115] is providing new insights into previously poorly understood aspects of RNA dynamics. There are several factors that make RDCs attractive probes of RNA dynamics. First, RDCs are sensitive to dynamics of a broad timescale ranging from picoseconds to milliseconds, which allows RDCs to capture both local structural motions as well as global conformation transitions and thereby provide very rich information of RNA dynamics[58]. Second, although one single RDC is only sensitive to the change of the angle between the internuclear vector and the

external magnetic field (see *Dipolar interactions*), collection of RDCs of various internuclear vectors in a RNA molecule can provide accurate information of long-range interhelical orientations of RNA, which is highly complementary to the measurements of NOE or PRE that provide relatively short-range distance information[18,67]. Third, RDCs can be measured in great abundance between nuclei in base, sugar and backbone moieties and can be straightforwardly computed based on just weak coupling assumption that can be applied in most dipolar interactions under partially aligned conditions[53]. Finally, by changing the alignment of a RNA molecule, more than one RDC data sets can be measured, which allows RDCs to give comprehensive and unbiased information of dynamics with high spatial resolution[99,116].

Dipolar Interactions

Analogous to a pair of bar magnets, nuclear dipole-dipole interactions originate from the through-space magnetic interaction between two nuclei, where the local magnetic field at a given nucleus is perturbed by the magnetic field of the other nucleus. Unlike *J*-coupling, dipolar interactions do not involve the interaction or correlation between the nucleus and the electrons. The Hamiltonian of dipolar interaction between spins *I* and *S* can be expressed as[117,118]:

$$\widehat{H}_{DD} = \left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_I\gamma_S h^2}{4\pi^2}\left(\frac{\hat{I}\cdot\hat{S}}{r_{IS}^3} - \frac{(\hat{I}\cdot\hat{r}_{IS})(\hat{S}\cdot\hat{r}_{IS})}{r_{IS}^5}\right) \quad (1.2)$$

where $\hat{I}$ and $\hat{S}$ are the spin angular momentum operators of spin *I* and *S* respectively; γ_I and γ_S are the gyromagnetic ratios of spin *I* and *S* respectively; μ_0 is the magnetic permittivity of vacuum; h is the Plank's constant; r_{IS} and $\hat{r}_{IS}$ are the separation between spin *I* and *S* and its corresponding position operator (Figure 1.2).

In solution NMR, heteronuclear dipolar couplings, in which case *I* and *S* are two distinct spins, are the most commonly and efficiently used dipolar couplings. It has been demonstrated that the weak coupling condition can be generally applied to most of such heteronucler dipolar interactions, in which case the eigenvectors of the Hamiltonian of dipolar interaction can be very well approximated by the eigenvectors of Hamiltonian of spin *I* and *S* without dipolar

interactions. This approximation is termed as "secular approximation" under which equation 1.2 for heteronuclear dipolar couplings can be simplified as shown in equation 1.3[117].

$$\widehat{H}_{DD} = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_I\gamma_S h^2}{2\pi^2 r_{IS}^3}\left(\frac{3cos^2\theta-1}{2}\right)\hat{I}_Z\hat{S}_Z \quad (1.3)$$

where θ represents the angle between the internuclear vector connecting spin I and S and the external magnetic field (Figure 1.2)[67].

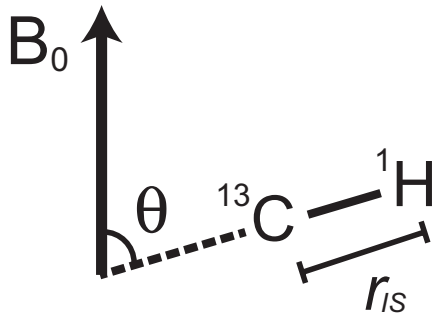

Figure 1.4 Relative orientation between internuclear vector (CH bond vector as an example) and the magnetic field.

Dynamics of the molecule leads to the change of the orientation of the dipolar coupling and therefore results in the averaged Hamiltonian of dipolar couplings as shown in equation 1.4[117].

$$\langle\widehat{H}_{DD}\rangle = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_I\gamma_S h^2}{2\pi^2 r_{IS}^3}\langle\frac{3cos^2\theta-1}{2}\rangle\hat{I}_Z\hat{S}_Z \quad (1.4)$$

This leads to the heteronuclear dipolar coupling between spin I and S that can be directly measured from NMR[119].

$$\langle D_{IS}\rangle = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_I\gamma_S h}{2\pi^2 r_{IS}^3}\langle\frac{3cos^2\theta-1}{2}\rangle \quad (1.5)$$

Under isotropic solutions, the averaged angular part in the equation 1.5 equal zero, which can be understood as an integral of the angular term over the entire orientation space (equation 1.6). Hence there are no net dipolar couplings in isotropic solutions.

$$\left\langle \frac{3cos^2\theta-1}{2} \right\rangle = \int_0^{2\pi} \int_0^{\pi} \frac{3cos^2\theta-1}{2} sin\theta d\theta d\varphi = 0 \quad (1.6)$$

However, in a weakly aligned and therefore anisotropic solution where the molecules cannot freely tumble but meanwhile the weak coupling condition can still be satisfied (see *Partial Alignment of Nucleic Acids*), the averaged angular term in equation 1.5 is non-zero and the values of these dipolar couplings in such anisotropic solutions are termed as residual dipolar couplings (RDCs)[67,68].

The Alignment Tensor

Although the angular dependence of dipolar couplings as shown in equation 1.5 has been known almost a century ago, it takes another half century for biophysicists to give the mathematical explanation of this angular dependence. It is first recognized by Saupe that the averaged angular term in equation 1.5 can be decomposed into a sum of the geometric terms describing the orientation of the internuclear vectors and the average tensor describing the nature of the alignment of the whole molecule in the magnetic field (equation 1.7)[120,121].

$$\left\langle \frac{3cos^2\theta-1}{2} \right\rangle = \sum_{kl=xyz} S_{kl} cos(\alpha_k)\, cos(\alpha_l) \quad (1.7)$$

where $\alpha_n$represents the angle between the corresponding internuclear vector and the n^{th} axis of the arbitrarily chosen molecular frame; S_{kl} represents the kl^{th} element of the 3 by 3 alignment tensor *S* describing the alignment of the molecule in the external magnetic field (Figure1.3) and can be calculated using equation 1.8.

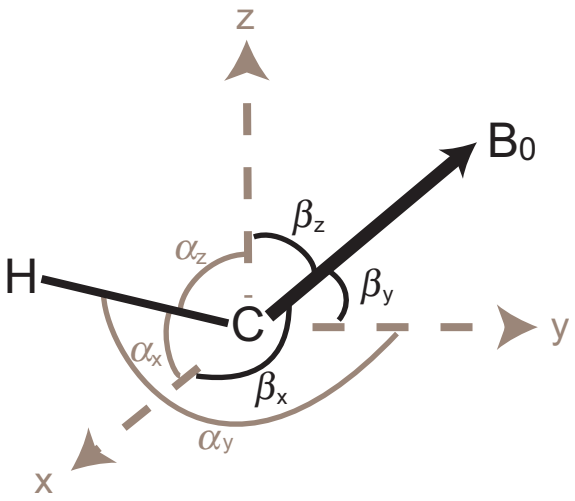

Figure 1.5 Angular dependence of bond vector and magnetic field in molecular frame. Interpretation of angular dependence of RDC in terms of the orientation of the internuclear vector (CH bond vector as an example) and the alignment tensor that can be calculated from the orientation of the external magnetic field (B_0) using equation 1.8. The frame (gray) is an arbitrarily chosen molecular frame.

$$S_{kl} = \langle \frac{3}{2} cos(\beta_k) cos(\beta_k) - \frac{1}{2}\delta_{kl} \rangle \quad (1.8)$$

where β_n represents the angle between the direction of the external magnetic field and the n^{th} axis of the arbitrarily chosen molecular frame; δ_{kl} is the Kronecker symbol that equals zero if and only if k=l. Here the whole alignment tensor S can be expressed in Cartesian representation as shown in equation 1.9[115,120,122,123].

$$S = \begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{pmatrix} \quad (1.9)$$

Because S is a real, symmetric (S_{kl}=S_{lk}) and traceless ($S_{xx} + S_{yy} + S_{zz}$=0) matrix, therefore there are only five independent elements in this matrix: a principal element S_{zz} (see below), an asymmetric parameter $\eta = \frac{|S_{yy} - S_{xx}|}{S_{zz}}$ and three off-diagonal elements S_{xy}, S_{yz} and S_{yz}[73]. Another

very useful parameter termed as generalized degree of order (GDO, ϑ)[122] that describes the amplitude of the motions encoded in the alignment tensor can be defined as shown below:

$$\vartheta = \sqrt{\frac{2}{3}\sum_{kl=xyz} S_{kl}^2} \tag{1.10}$$

and the amplitude of the relative dynamics between two segments or domains (e.g. domain 1 and 2) of a molecule can be defined as the internal generalized degree of order (GDOint, ϑ_{int}), which can be calculated using the expression below:

$$\vartheta_{int} = \frac{\vartheta_1}{\vartheta_2} \tag{1.11}$$

Because the alignment tensors are real symmetric matrices, therefore they can always be diagonalized through a linear transformation from the current molecular frame to its principal axis system (PAS) in which only the eigenvalues of the alignment tensor,$S_{xx}(PAS)$, $S_{yy}(PAS)$ and $S_{zz}(PAS)$ are non-zero[115].

$$S(PAS) = \begin{pmatrix} S_{xx}(PAS) & 0 & 0 \\ 0 & S_{yy}(PAS) & 0 \\ 0 & 0 & S_{zz}(PAS) \end{pmatrix} \tag{1.12}$$

The PAS and the eigenvalues of the alignment tensor *S* are independent of the molecular frame in which *S* is appearing and therefore provide a unique and robust way to assess the alignment of the molecule. It has to be noticed that the values of tensor elements $S_{xx}(PAS)$, $S_{yy}(PAS)$ and $S_{zz}(PAS)$ in equation 1.10 are probably different from the corresponding ones in equation 1.9 due to the fact that they are calculated in different frames but the traceless property of alignment tensor *S* is not affected. Additionally, it should not be misunderstood that although there are only two independent elements in the form of alignment tensor *S* expressed in PAS (equation 1.10), the other three independent elements are encoded in the linear transformation in terms of the three rotation angles that transform the alignment tensor *S* from an arbitrary molecular frame to its PAS. S_{zz} is defined as the principal element of the alignment tensor *S*, which has the largest

magnitude ($|S_{zz}(PAS)| \geq |S_{yy}(PAS)| \geq |S_{xx}(PAS)|$) in its PAS, and the direction of which is defined to be the principal direction of the alignment of the molecule in magnetic field.

As alignment tensor *S* plays the key role in connecting the orientation and RDC of every internuclear vector of a molecule, it is of central importance to accurately determine or calculate alignment tensors in analysis of experimentally measured RDCs. Several software for calculating the alignment tensor of biomolecule have been published and recent studies have demonstrated their robustness. For example, RAMAH developed by Al-Hashimi and co-workers[72] has been demonstrated to be very robust for calculating the experimental alignment tensor of biomolecules from measured RDCs using equation 1.7 and singular value decomposition (SVD) algorithm. Recently, two modeling-based software PALES[124] and PATL[125] have been developed to predict alignment tensor of biomolecule solely based on their structure and charge distribution without any experimental input. These methods can be applied as useful tools to predict the alignments and furthermore RDCs of a biomolecule in magnetic field. Encouraging results have been reported by a recent study[99] using these methods in determination of RNA dynamics, although the accuracy of the alignment tensors predicted using these methods could still be further improved.

1.3.2 Measurement of RDCs

Partial Alignment of Biomolecules in Magnetic Field

Measurement of RDCs under solution conditions requires introducing partial alignment of the biomolecule in solution[126], either by dissolving the biomolecule in an alignment medium[115] or in the case of nucleic acids and paramagnetic proteins, through direct interactions with the external magnetic field itself[67,127,128]. Using either aligning strategy, the alignment level of the biomolecule is of central importance in measurement of RDCs. Alignment levels higher than 10^{-2} (analogous to one out of one hundred molecules is completely aligned) give rise to extensive dipolar couplings that compromise the spectral resolution required for analyzing large biomolecules and possibly break the weak coupling condition that disables the expression of RDCs using equation 1.5 and results in unnecessary complexity in considering higher order

interactions that are otherwise negligible; alignment levels lower than 10^{-5} lead to RDCs that are too small compared to the width of the resonance peaks and therefore prevent precise measurements. In general, alignment levels ~10^{-3} is optimal[115,126], which lead to sufficiently large RDC values affording precise measurements with maintained appropriate spectral resolution.

Aligning biomolecules by dissolving them in an inert alignment medium is very straightforward and therefore very commonly used in measuring RDCs. The alignment level of the biomolecule of interest can easily achieve the optimal alignment level (~10^{-3}) and can be adjusted by simply changing the concentration of the alignment medium. Bax and co-workers first experimentally demonstrated the medium-induced alignment in solution using liquid-crystalline disc-shaped phospholipids "bicelles"[115]. Since then, several media or combinations of media have been introduced for partially aligning biomolecules in solution[71]. In particular, Pf1 phage, a 7.4kb rod-like shape single-stranded DNA genome with one coat protein per nucleotide, is the most favorable alignment medium for aligning nucleic acids[129]. This is because the identical coat proteins of Pf1 phage are negatively charged, largely reducing the undesired extensive attractive interaction between Pf1 and the nucleic acids[130,131]. Pf1 phage generally aligns nucleic acids with the principal direction of alignment tensor orientated along the long axis of the molecule. Unlike proteins, the alignment of which can be altered by changing the alignment media[132,133], alignment of nucleic acids can be hardly changed by altering the alignment media[134]. This is because the negative charges of nucleic acids primarily concentrate on the backbone phosphate atoms, resulting in a semi-uniform charge distribution on the surface of nucleic acids and giving rise to highly similar interactions with different alignment media that leads to highly correlated alignments of nucleic acids[130,135].

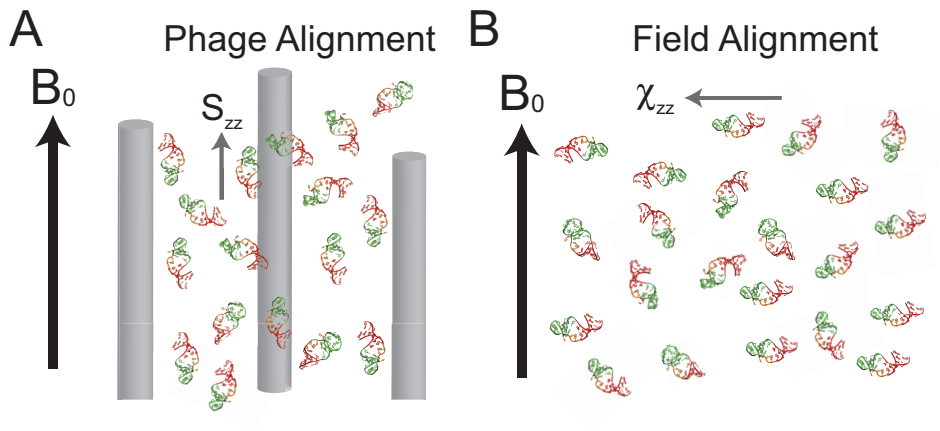

Figure 1.6 Partially aligning biomolecules (RNA as an example) in solution. (A) Medium-induced and **(B)** field-induced alignment of RNA in solution.

Another effective strategy for aligning biomolecule and in particular nucleic acids and paramagnetic proteins[67,136-139] is to induce spontaneous alignment by magnetic field itself, taking advantage of the interaction between the bulk external magnetic field and large magnetic susceptibility anisotropies ($\Delta\chi$) of the biomolecules[136,137]. The alignment level using this strategy is proportional to the square of the strength of the external magnetic field. In general, the alignment level induced by magnetic field is very small and therefore high magnetic field strength is more favorable for inducing alignment of nucleic acids and paramagnetic proteins. For example, the alignment level can reach ~10^{-4} at 800MHz but this is still one degree of magnitude lower than the optimal alignment level (~10^{-3}). Hence the resulting RDCs measured under field-induced alignment are relatively small (magnitude < 10Hz). However, field-induced alignment prevents any perturbations from alignment media. More importantly, the principal direction of field-induced alignment is approximately perpendicular to the magnetic field that is completely different from medium-induced alignment, of which the principal direction is approximately along the magnetic field (Figure 1.4). Therefore field-induced alignment strategy can provide a unique and independent alignment from medium-induced alignment for biomolecules and in particular for nucleic acids, which is very challenging if not impossible to achieve by changing or modifying alignment media[140,141].

NMR Experiments for Measuring RDCs

Several experimental strategies have been developed to measure a wide variety of RDCs of both proteins and nucleic acids. Most if not all of the experiments are subject to a two-step procedure: 1) measure *J* couplings alone in isotropic solution; 2) measure *J*+*D* couplings in partially alignment solution and then calculate RDCs by taking the difference between the values obtained from these two steps. In step one, the resonance peak of a spin-pair of interest splits into a doublet solely due to *J* coupling which can be measured by taking the difference of the frequencies between the doublet peaks; in step two, the splitting of the same spin-pair is due to both *J* coupling and dipolar coupling *D* arising from the partial alignment of the biomolecule (Figure 1.5). These two steps are implemented using the same NMR experiments and other solution conditions except the alignment level. Here only the NMR experiments used for measuring RDCs in nucleic acids will be discussed in detail and the experiments for measuring RDCs of proteins are designed based on the same principles and can be easily found in a series of review articles.

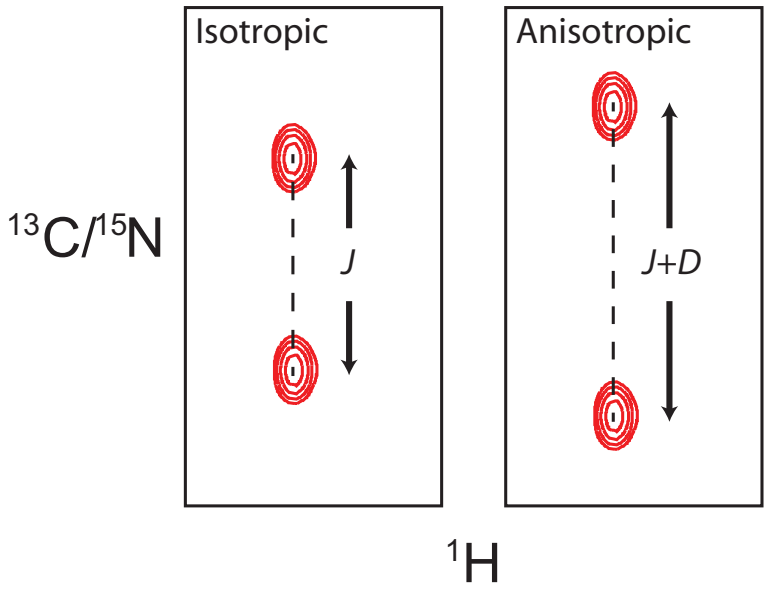

Figure 1.7 Measurement of *J* coupling and RDC. In step one, the *J* coupling is measured in isotropic solution (left); in step two, both *J* coupling and dipolar coupling (*D*) are measured in partially aligned anisotropic solution (right).

The most commonly measured RDCs in nucleic acids are those between directly bonded CH and NH nuclei (C2H2, C5H5, C6H6, C8H8, N1H1 and N3H3) in nucleobases and C1'H1' in sugar moieties. For small nucleic acids for which the resonance overlap is not serious, the RDCs can be measured using 2D HSQC–type experiments that employ inphase-antiphase (IPAP) scheme[142] to encode the individual components of the doublet along the ^{1}H, ^{13}C and ^{15}N dimensions. For large nucleic acids for which the short transverse relaxation time (T_2) causes much broader or even diminished resonance peaks, transverse relaxation optimized spectroscopy (TROSY) is the advantageous method for measuring RDCs due to the fact that relaxation interference between dipolar couplings of CH or NH spin pairs and sizable CSAs of ^{13}C or ^{15}N can effectively cancel each other and give rise to longer transverse relaxation time, resulting in remarkably narrower resonance peaks and much enhanced spectral sensitivity[143-150]. Multi-dimensional experiments that employ HCN or E-COSY methods with spin-state-selective excitation (S^3E) scheme can also be used to improve the spectral sensitivity as well as resolution especially for C6H6, C8H8 in nucleobases and C1'H1' in sugar moieties[151-157].

However, in general, the measurement of CH RDCs in sugar moieties (e.g. C2'H2', C3'H3', C4'H4', C5'H5', C5'H5") is dramatically more challenging because of severe spectral overlap in 2D CH experiments. Experiments are underdevelopment for exploiting better resolutions in measuring C2'H2' and C3'H3' RDCs[158]. Likewise, severe spectral overlap also complicates the measurement of RDCs between ^{31}P and sugar protons[159,160] which otherwise can provide unique information on backbone geometry. The spectral overlap for ^{31}P is due to the sizeable ^{31}P CSA relaxation that causes very short transverse relaxation time, yielding broad resonance peaks and low signal-to-noise ratio[72].

1.3.3 Dynamic Interpretation of RDCs

The utility of RDCs in studies of dynamics arises chiefly from the angular dependence in equation 1.5. To appreciate the full angular dynamic information contained within RDCs, it is useful and more convenient to use a spherical representation to express the measured time-averaged alignment tensors. For a single internuclear vector, due to the axial symmetry, the

direction of its PAS is always along the internuclear vector and only one of the five independent alignment tensor elements $\langle D_0^2 \rangle$ expressed using spherical representation in PAS is non-zero. This only non-zero element can be expressed in terms of the overall alignment tensor, O_m^2, of the entire molecule and five out of twenty-five time-averaged Wigner rotation elements, $\langle D_{n0}^2(\beta\gamma) \rangle$ that are functions of the Euler angles $(\beta\gamma)$ describing the orientation of the internuclear vector relative to the molecular frame (Figure 1.6)[18,49,161]:

$$\langle D_0^2 \rangle^l = \sum_{m=-2}^{2} \sum_{n=-2}^{2} O_m^2(PAS)^l \, D_{mn}^2(\theta_l) \langle D_{n0}^2(\beta\gamma) \rangle \quad (1.13)$$

Here, $O_m^2(PAS)^l$ are elements of the l^{th} overall alignment tensor describing averaging of the dipolar interaction due to overall motions (e.g. tumbling of molecule) expressed in the PAS of the tensor. $D_{mn}^2(\theta_l)$ are elements of a time-independent Wigner rotation matrix that transform the PAS of the l^{th} overall tensor into a common molecular frame. Importantly, equation 1.11 assumes that the internal and overall motions of the molecule are uncorrelated[18,116].

The information regarding internal motions is contained within the five time-average Wigner elements $\langle D_{n0}^2(\beta\gamma) \rangle$ ($\{n\}$=-2, -1, 0, 1, 2) which are functions of two Euler angles describing the orientation of the internuclear vector relative to the molecular frame (Table 1.1). The five time-averaged Wigner elements can be determined experimentally for each internuclear vector provided the measurement of RDCs under five linearly independent alignment conditions, as shown elegantly by Griesinger and Tolman[162,163]. These five parameters specify the average orientation of the internuclear vector relative to the molecular frame, the amplitude of any internal motions, as well as the extent and direction of motional asymmetry. Note that due to the inherent axial symmetry of the internuclear vector and thereby the alignment tensor of the internuclear vector, there is no sensitivity to internal motions that lead to rotations about the internuclear vector itself (which can be described by another Euler angle α), therefore limiting sensitivity to only two of the three Euler angles (Equation 1.13).

n/k	2	1	0	-1	-2
2	$e^{i2(\alpha+\gamma)}\cos^4\frac{\beta}{2}$	$e^{i(2\gamma+\alpha)}\sin\beta\cos^2\frac{\beta}{2}$	$e^{i2\gamma}\sqrt{\frac{3}{8}}\sin^2\beta$	$e^{i(2\gamma+\alpha)}\sin\beta\sin^2\frac{\beta}{2}$	$e^{i2(\gamma-\alpha)}\sin^4\frac{\beta}{2}$
1	$-e^{i(\gamma+2\alpha)}\sin\beta\cos^2\left(\frac{\beta}{2}\right)$	$e^{i(\gamma+\alpha)}\cos\frac{3\beta}{2}\cos\frac{\beta}{2}$	$e^{i\gamma}\sqrt{\frac{3}{8}}\sin 2\beta$	$e^{i(\gamma-\alpha)}\sin\frac{3\beta}{2}\sin\frac{\beta}{2}$	$e^{i(\gamma-2\alpha)}\sin\beta\sin^2\frac{\beta}{2}$
0	$e^{i2\alpha}\sqrt{\frac{3}{8}}\sin^2\beta$	$-e^{i\alpha}\sqrt{\frac{3}{8}}\sin 2\beta$	$\frac{1}{2}\left(3\cos^2\beta-1\right)$	$e^{-i\alpha}\sqrt{\frac{3}{8}}\sin 2\beta$	$e^{-i2\alpha}\sqrt{\frac{3}{8}}\sin^2\beta$
-1	$-e^{i(-\gamma+2\alpha)}\sin\beta\sin^2\left(\frac{\beta}{2}\right)$	$e^{i(-\gamma+\alpha)}\sin\frac{3\beta}{2}\sin\frac{\beta}{2}$	$-e^{-i\gamma}\sqrt{\frac{3}{8}}\sin 2\beta$	$e^{i(-\gamma-\alpha)}\cos\frac{3\beta}{2}\cos\frac{\beta}{2}$	$e^{i(-\gamma-2\alpha)}\sin\beta\cos^2\frac{\beta}{2}$
-2	$e^{i2(-\gamma+\alpha)}\sin^4\frac{\beta}{2}$	$-e^{i(-2\gamma+\alpha)}\sin\beta\sin^2\left(\frac{\beta}{2}\right)$	$e^{-i2\gamma}\sqrt{\frac{3}{8}}\sin^2\beta$	$-e^{i(-2\gamma-\alpha)}\sin\beta\cos^2\left(\frac{\beta}{2}\right)$	$e^{i2(-\gamma-\alpha)}\cos^4\frac{\beta}{2}$

Table 1.1 Elements of the second rank Wigner rotation matrix

The internuclear vector type analysis of RDCs has been successfully applied to proteins [121,132,139,162-164] but has not to be applied to nucleic acids. Such applications are challenging because of the difficulty in varying the overall alignment of nucleic acids as stated in section 1.3.2; in addition, it is generally more difficult to measure the required number of spatially independent RDCs to simultaneously determine both internal and overall tensor parameters. As mentioned above, this type of analysis also assumes that internal and overall motions are not correlated to one another, which is not generally applicable for highly flexible RNAs, although recently developed domain elongation approaches overcome this problem for simple two-domain RNAs[18,165].

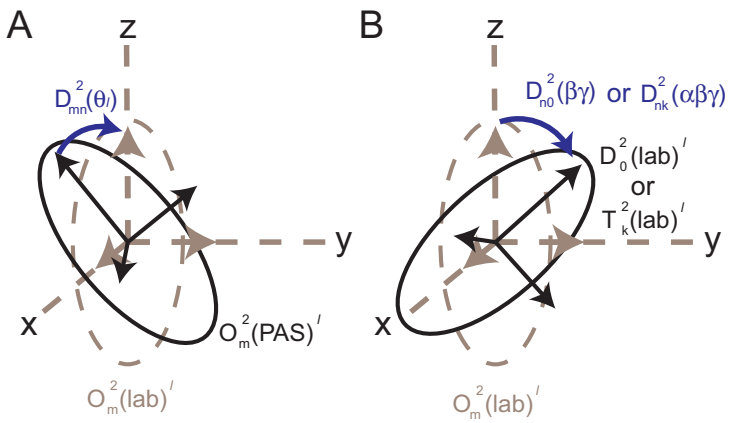

Figure 1.8 Angular dynamic information contained in RDCs. The dependence between the overall alignment tensor and the local alignment tensor can be decomposed into **(A)** transformation of overall alignment from its PAS to the molecular frame and **(B)** transformation of the transformed overall alignment from molecular frame to PAS of the internuclear vector or chiral domain.

In principle, much greater dynamic information can be obtained from analyzing collections of five or more spatially independent RDCs measured in a semi-rigid chiral fragment, such as an A-form helix in RNA[33,73]. Here, one can use the RDCs to determine all five elements of a time-averaged alignment tensor $\langle T^2_k \rangle^l$ (Figure 1.6) describing the alignment of a chiral fragment relative to the magnetic field, which can also be expressed in terms of the overall alignment tensor of the molecule $O^2_m(PAS)^l$ and time-averaged Wigner rotation elements, $\langle D^2_{nk}(\alpha\beta\gamma) \rangle$

$$\langle T^2_k \rangle^l = \sum^2_{m=-2} \sum^2_{n=-2} O^2_m(PAS)^l \, D^2_{mn}(\theta_l) \langle D^2_{nk}(\alpha\beta\gamma) \rangle \qquad (1.14)$$

where all twenty-five time-average Wigner elements $\langle D^2_{nk}(\alpha\beta\gamma) \rangle$ ($\{n,k\}$=-2, -1, 0, 1, 2) (Table 1.1) can theoretically be determined, provided the measurement of RDCs and $\langle T^2_k \rangle^l$under five linearly independent alignments[116]. These twenty-five time-averaged Wigner elements represent the theoretical maximum dynamic angular information due to internal motions that can be obtained from RDCs[116]. Here, the sensitivity extends to all three Euler angles, including α, as well as co-variations among the three Euler angles (Table 1.1), given the simultaneous

dependence of many Wigner terms on all three Euler angles[18,116]. The above approach is well-suited in analyzing RNA chiral helices and nine out of twenty-five Wigner elements have been experimentally determined in the HIV 1 TAR RNA system by using the domain elongation strategy[18]. The measurement of all twenty-five Wigner elements in RNA remains to be an important challenge for the future, which requires robust methods for varying alignment as discussed in section 1.3.2.

Some contents in this chapter are published in Annu. Rev. Phys. Chem. (Salmon L., Yang S. & Al-Hashimi H.M.)[53] and Recent Developments in Biomolecular NMR (Eichhorn C.D., Yang S. & Al-Hashimi H.M.)[166].

1.4 References

(1) Guttman, M.; Rinn, J. L. *Nature* **2012**, *482*, 339.
(2) Ponting, C. P.; Oliver, P. L.; Reik, W. *Cell* **2009**, *136*, 629.
(3) Wang, X. Q.; Crutchley, J. L.; Dostie, J. *Current Genomics* **2011**, *12*, 307.
(4) Bartel, D. P. *Cell* **2009**, *136*, 215.
(5) Bastet, L.; Dube, A.; Masse, E.; Lafontaine, D. A. *Mol. Microbiol.* **2011**, *80*, 1148.
(6) Shukla, G. C.; Haque, F.; Tor, Y.; Wilhelmsson, L. M.; Toulme, J. J.; Isambert, H.; Guo, P.; Rossi, J. J.; Tenenbaum, S. A.; Shapiro, B. A. *ACS Nano* **2011**, *5*, 3405.
(7) Deigan, K. E.; Ferre-D'Amare, A. R. *Acc. Chem. Res.* **2011**, *44*, 1329.
(8) Shay, J. W.; Keith, W. N. *Br. J. Cancer* **2008**, *98*, 677.
(9) Harley, C. B. *Nat. Rev. Cancer* **2008**, *8*, 167.
(10) Isaacs, F. J.; Dwyer, D. J.; Collins, J. J. *Nat. Biotechnol.* **2006**, *24*, 545.
(11) Stelzer, A. C.; Kratz, J. D.; Zhang, Q.; Al-Hashimi, H. M. *Angew. Chem., Int. Ed. Engl.* **2010**.
(12) Galloway, K. E.; Franco, E.; Smolke, C. D. *Science* **2013**, *341*, 1235005.
(13) Culler, S. J.; Hoff, K. G.; Smolke, C. D. *Science* **2010**, *330*, 1251.
(14) Williamson, J. R. *Nat. Struct. Mol. Biol.* **2000**, *7*, 834.
(15) Leulliot, N.; Varani, G. *Biochemistry* **2001**, *40*, 7947.
(16) Micura, R.; Hobartner, C. *ChemBioChem* **2003**, *4*, 984.
(17) Al-Hashimi, H. M. *ChemBioChem* **2005**, *6*, 1506.
(18) Zhang, Q.; Stelzer, A. C.; Fisher, C. K.; Al-Hashimi, H. M. *Nature* **2007**, *450*, 1263.
(19) Frank, A. T.; Stelzer, A. C.; Al-Hashimi, H. M.; Andricioaei, I. *Nucleic Acids Res.* **2009**, *37*, 3670.

(20) Shi, X.; Mollova, E. T.; Pljevaljcic, G.; Millar, D. P.; Herschlag, D. *J. Am. Chem. Soc.* **2009**, *131*, 9571.
(21) Levengood, J. D.; Rollins, C.; Mishler, C. H.; Johnson, C. A.; Miner, G.; Rajan, P.; Znosko, B. M.; Tolbert, B. S. *J. Mol. Biol.* **2012**, *415*, 680.
(22) Dethoff, E. A.; Chugh, J.; Mustoe, A. M.; Al-Hashimi, H. M. *Nature* **2012**, *482*, 322.
(23) Cech, T. R. *Annu. Rev. Biochem.* **1990**, *59*, 543.
(24) Berglund, H.; Rak, A.; Serganov, A.; Garber, M.; Hard, T. *Nat. Struct. Biol.* **1997**, *4*, 20.
(25) Orr, J. W.; Hagerman, P. J.; Williamson, J. R. *J. Mol. Biol.* **1998**, *275*, 453.
(26) Clemons, W. M., Jr.; Davies, C.; White, S. W.; Ramakrishnan, V. *Structure* **1998**, *6*, 429.
(27) Nikulin, A.; Serganov, A.; Ennifar, E.; Tishchenko, S.; Nevskaya, N.; Shepard, W.; Portier, C.; Garber, M.; Ehresmann, B.; Ehresmann, C.; Nikonov, S.; Dumas, P. *Nat. Struct. Biol.* **2000**, *7*, 273.
(28) Agalarov, S. C.; Prasad, G. S.; Funke, P. M.; Stout, C. D.; Williamson, J. R. *Science* **2000**, *288*, 107.
(29) Mulder, A. M.; Yoshioka, C.; Beck, A. H.; Bunner, A. E.; Milligan, R. A.; Potter, C. S.; Carragher, B.; Williamson, J. R. *Science* **2010**, *330*, 673.
(30) Adilakshmi, T.; Bellur, D. L.; Woodson, S. A. *Nature* **2008**, *455*, 1268.
(31) Nikolova, E. N.; Zhou, H.; Gottardo, F. L.; Alvey, H. S.; Kimsey, I. J.; Al-Hashimi, H. M. *Biopolymers* **2013**, *99*, 955.
(32) Al-Hashimi, H. M.; Walter, N. G. *Curr. Opin. Struct. Biol.* **2008**, *In Press*.
(33) Bailor, M. H.; Musselman, C.; Hansen, A. L.; Gulati, K.; Patel, D. J.; Al-Hashimi, H. M. *Nat. Protoc.* **2007**, *2*, 1536.
(34) Nikolova, E. N.; Kim, E.; Wise, A. A.; O'Brien, P. J.; Andricioaei, I.; Al-Hashimi, H. M. *Nature* **2011**, *470*, 498.
(35) Leontis, N. B.; Lescoute, A.; Westhof, E. *Curr. Opin. Struct. Biol.* **2006**, *16*, 279.
(36) Chu, V. B.; Lipfert, J.; Bai, Y.; Pande, V. S.; Doniach, S.; Herschlag, D. *RNA* **2009**, *15*, 2195.
(37) Bailor, M. H.; Sun, X.; Al-Hashimi, H. M. *Science* **2010**, *327*, 202.
(38) Bailor, M. H.; Mustoe, A. M.; Brooks, C. L., 3rd; Al-Hashimi, H. M. *Nat. Protoc.* **2011**, *6*, 1536.
(39) Mustoe, A. M.; Bailor, M. H.; Teixeira, R. M.; Brooks, C. L., 3rd; Al-Hashimi, H. M. *Nucleic Acids Res.* **2012**, *40*, 892.
(40) Zacharias, M.; Hagerman, P. J. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 6052.
(41) Zacharias, M.; Hagerman, P. J. *J. Mol. Biol.* **1995**, *247*, 486.
(42) Zacharias, M.; Hagerman, P. J. *J. Mol. Biol.* **1996**, *257*, 276.
(43) Riordan, F. A.; Bhattacharyya, A.; McAteer, S.; Lilley, D. M. *J. Mol. Biol.* **1992**, *226*, 305.
(44) Bhattacharyya, A.; Murchie, A. I.; Lilley, D. M. *Nature* **1990**, *343*, 484.
(45) Bhattacharyya, A.; Lilley, D. M. *J. Mol. Biol.* **1989**, *209*, 583.
(46) Wang, Y. H.; Griffith, J. *Biochemistry* **1991**, *30*, 1358.
(47) Zhuang, X.; Bartley, L. E.; Babcock, H. P.; Russell, R.; Ha, T.; Herschlag, D.; Chu, S. *Science* **2000**, *288*, 2048.
(48) Zhuang, X.; Kim, H.; Pereira, M. J.; Babcock, H. P.; Walter, N. G.; Chu, S. *Science* **2002**, *296*, 1473.

(49) Zhang, Q.; Al-Hashimi, H. M. *Nat. Methods* **2008**, *5*, 243.
(50) Mustoe, A. M.; Al-Hashimi, H. M.; Brooks, C. L., 3rd *J. Phys. Chem. B* **2014**, *118*, 2615.
(51) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.
(52) Cruz, J. A.; Westhof, E. *Cell* **2009**, *136*, 604.
(53) Salmon, L.; Yang, S.; Al-Hashimi, H. M. *Annu. Rev. Phys. Chem.* **2013**.
(54) Dethoff, E. A.; Petzold, K.; Chugh, J.; Casiano-Negroni, A.; Al-Hashimi, H. M. *Nature* **2012**, *491*, 724.
(55) Fisher, C. K.; Huang, A.; Stultz, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 14919.
(56) Rambo, R. P.; Tainer, J. A. *Curr. Opin. Struct. Biol.* **2010**, *20*, 128.
(57) Bernado, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. *J. Am. Chem. Soc.* **2007**, *129*, 5656.
(58) Bothe, J. R.; Nikolova, E. N.; Eichhorn, C. D.; Chugh, J.; Hansen, A. L.; Al-Hashimi, H. M. *Nat. Methods* **2011**, *8*, 919.
(59) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 1270.
(60) Robustelli, P.; Kohlhoff, K.; Cavalli, A.; Vendruscolo, M. *Structure* **2010**, *18*, 923.
(61) Cromsigt, J.; van Buuren, B.; Schleucher, J.; Wijmenga, S. In *Nuclear Magnetic Resonance of Biologica Macromolecules,* 2001; Vol. 338, p 371.
(62) Wijmenga, S. S.; Kruithof, M.; Hilbers, C. W. *J. Biomol. NMR* **1997**, *10*, 337.
(63) Barton, S.; Heng, X.; Johnson, B. A.; Summers, M. F. *J. Biomol. NMR* **2013**, *55*, 33.
(64) Frank, A. T.; Horowitz, S.; Andricioaei, I.; Al-Hashimi, H. M. *J. Phys.Chem. B* **2013**.
(65) Wijmenga, S. S.; van Buuren, B. N. M. *Prog. Nucl. Magn Reson. Spectrosc.* **1998**, *32*, 287.
(66) Furtig, B.; Richter, C.; Wohnert, J.; Schwalbe, H. *ChemBioChem* **2003**, *4*, 936.
(67) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 9279.
(68) Tjandra, N.; Garrett, D. S.; Gronenborn, A. M.; Bax, A.; Clore, G. M. *Nat. Struct. Mol. Biol.* **1997**, *4*, 443.
(69) Hansen, M. R.; Mueller, L.; Pardi, A. *Nat. Struct. Mol. Biol.* **1998**, *5*, 1065.
(70) Clore, G. M.; Starich, M. R.; Gronenborn, A. M. *J. Am. Chem. Soc.* **1998**, *120*, 10571.
(71) Getz, M.; Sun, X.; Casiano-Negroni, A.; Zhang, Q.; Al-Hashimi, H. M. *Biopolymers* **2007**, *86*, 384.
(72) Hansen, A. L.; Al-Hashimi, H. M. *J. Magn. Reson.* **2006**, *179*, 299.
(73) Musselman, C.; Pitt, S. W.; Gulati, K.; Foster, L. L.; Andricioaei, I.; Al-Hashimi, H. M. *J. Biomol. NMR* **2006**, *36*, 235.
(74) Ottiger, M.; Tjandra, N.; Bax, A. *J. Am. Chem. Soc.* **1997**, *119*, 9825.
(75) Cornilescu, G.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 10143.
(76) Wu, Z. R.; Tjandra, N.; Bax, A. *J. Am. Chem. Soc.* **2001**, *123*, 3617.
(77) Ying, J.; Grishaev, A.; Bryce, D. L.; Bax, A. *J. Am. Chem. Soc.* **2006**, *128*, 11443.
(78) Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128.
(79) Iwahara, J.; Clore, G. M. *Nature* **2006**, *440*, 1227.
(80) Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M. R.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 8407.

(81) Blackledge, M. J.; Bruschweiler, R.; Griesinger, C.; Schmidt, J. M.; Xu, P.; Ernst, R. R. *Biochemistry* **1993**, *32*, 10960.
(82) Vogeli, B.; Kazemi, S.; Guntert, P.; Riek, R. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1053.
(83) Palmer, A. G., 3rd; Massi, F. *Chem. Rev.* **2006**, *106*, 1700.
(84) Sekhar, A.; Vallurupalli, P.; Kay, L. E. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 11391.
(85) Walter, N. G.; Harris, D. A.; Pereira, M. J.; Rueda, D. *Biopolymers* **2001**, *61*, 224.
(86) Qin, P. Z.; Dieckmann, T. *Curr. Opin. Struct. Biol.* **2004**, *14*, 350.
(87) Bokinsky, G.; Zhuang, X. *Acc. Chem. Res.* **2005**, *38*, 566.
(88) Zhuang, X. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 399.
(89) Schiemann, O.; Piton, N.; Mu, Y.; Stock, G.; Engels, J. W.; Prisner, T. F. *J. Am. Chem. Soc.* **2004**, *126*, 5722.
(90) Schiemann, O.; Weber, A.; Edwards, T. E.; Prisner, T. F.; Sigurdsson, S. T. *J. Am. Chem. Soc.* **2003**, *125*, 3434.
(91) Weber, A.; Schiemann, O.; Bode, B.; Prisner, T. F. *J. Magn. Reson.* **2002**, *157*, 277.
(92) Solomatin, S. V.; Greenfeld, M.; Chu, S.; Herschlag, D. *Nature* **2010**, *463*, 681.
(93) Showalter, S. A.; Bruschweiler, R. *J. Am. Chem. Soc.* **2007**, *129*, 4158.
(94) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. *Biophys. J.* **2007**, *93*, 2300.
(95) Chu, V. B.; Herschlag, D. *Curr. Opin. Struct. Biol.* **2008**, *18*, 305.
(96) Nilges, M.; Gronenborn, A. M.; Brunger, A. T.; Clore, G. M. *Protein Engineering* **1988**, *2*, 27.
(97) Markwick, P. R.; Bouvignies, G.; Salmon, L.; McCammon, J. A.; Nilges, M.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 16968.
(98) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W.; Blackledge, M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17002.
(99) Salmon, L.; Bascom, G.; Andricioaei, I.; Al-Hashimi, H. M. *J. Am. Chem. Soc.* **2013**, *135*, 5457.
(100) Eichhorn, C. D.; Feng, J.; Suddala, K. C.; Walter, N. G.; Brooks, C. L., 3rd; Al-Hashimi, H. M. *Nucleic Acids Res.* **2012**, *40*, 1345.
(101) Bai, Y.; Chu, V. B.; Lipfert, J.; Pande, V. S.; Herschlag, D.; Doniach, S. *J. Am. Chem. Soc.* **2008**, *130*, 12334.
(102) Guerry, P.; Mollica, L.; Blackledge, M. *ChemPhysChem* **2013**, *14*, 3046.
(103) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908.
(104) Salmon, L.; Pierce, L.; Grimm, A.; Ortega Roldan, J. L.; Mollica, L.; Jensen, M. R.; van Nuland, N.; Markwick, P. R.; McCammon, J. A.; Blackledge, M. *Angew. Chem.* **2012**, *51*, 6103.
(105) Clore, G. M.; Garrett, D. S. *J. Am. Chem. Soc.* **1999**, *121*, 9008.
(106) Guerry, P.; Salmon, L.; Mollica, L.; Ortega Roldan, J. L.; Markwick, P.; van Nuland, N. A.; McCammon, J. A.; Blackledge, M. *Angew. Chem.* **2013**, *52*, 3181.
(107) Schwieters, C. D.; Clore, G. M. *Biochemistry* **2007**, *46*, 1152.
(108) Best, R. B.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 8090.
(109) De Simone, A.; Richter, B.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 3810.

(110) Richter, B.; Gsponer, J.; Varnai, P.; Salvatella, X.; Vendruscolo, M. *J. Biomol. NMR* **2007**, *37*, 117.
(111) Zhou, S. K.; Chellappa, R. *IEEE transactions on pattern analysis and machine intelligence* **2006**, *28*, 917.
(112) Yang, S.; Salmon, L.; Al-Hashimi, H. M. *Nat. Methods* **2014**.
(113) Rinnenthal, J.; Buck, J.; Ferner, J.; Wacker, A.; Furtig, B.; Schwalbe, H. *Acc. Chem. Res.* **2011**, *44*, 1292.
(114) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Nat. Struct. Mol. Biol.* **1997**, *4*, 292.
(115) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111.
(116) Fisher, C. K.; Zhang, Q.; Stelzer, A.; Al-Hashimi, H. M. *J. Phys.Chem. B* **2008**, *112*, 16815.
(117) Ernst, R. R.; Bodenhausen, G.; Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*; Clarendon Press: Oxford, 1987.
(118) Abragam, A. *Principles of Nuclear Magnetism*; Clarendon Press: Oxford, 1961.
(119) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
(120) Saupe, A. *Angew. Chem., Int. Ed. Engl.* **1968**, *7*, 97.
(121) Tolman, J. R.; Ruan, K. *Chem. Rev.* **2006**, *106*, 1720.
(122) Tolman, J. R.; Al-Hashimi, H. M.; Kay, L. E.; Prestegard, J. H. *J. Am. Chem. Soc.* **2001**, *123*, 1416.
(123) Losonczi, J. A.; Andrec, M.; Fischer, M. W. F.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334.
(124) Zweckstetter, M. *Nat. Protoc.* **2008**, *3*, 679.
(125) Berlin, K.; O'Leary, D. P.; Fushman, D. *J. Magn. Reson.* **2009**, *201*, 25.
(126) Tjandra, N. *Structure With Folding & Design* **1999**, *7*, R205.
(127) Bothner-By, A. A. In *Encyclopedia of Nuclear Magnetic Resonance*; Grant, D. M., Harris, R. K., Eds.; Wiley: Chichester, 1995, p 2932.
(128) Zhang, Q.; Throolin, R.; Pitt, S. W.; Serganov, A.; Al-Hashimi, H. M. *J. Am. Chem. Soc.* **2003**, *125*, 10530.
(129) Hansen, M. R.; Hanson, P.; Pardi, A. *Methods Enzymol.* **2000**, *317*, 220.
(130) Wu, B.; Petersen, M.; Girard, F.; Tessari, M.; Wijmenga, S. S. *J. Biomol. NMR* **2006**, *35*, 103.
(131) Zweckstetter, M.; Bax, A. *J. Biomol. NMR* **2001**, *20*, 365.
(132) Ramirez, B. E.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 9106.
(133) Al-Hashimi, H. M.; Valafar, H.; Terrell, M.; Zartler, E. R.; Eidsness, M. K.; Prestegard, J. H. *J. Magn. Reson.* **2000**, *143*, 402.
(134) Bondensgaard, K.; Mollova, E. T.; Pardi, A. *Biochemistry* **2002**, *41*, 11532.
(135) Zweckstetter, M.; Hummer, G.; Bax, A. *Biophys. J.* **2004**, *86*, 3444.
(136) Bastiaan, E. W.; Maclean, C.; Van Zilj, P. C. M.; Bothner-by, A. A. *Annu. Rep. NMR Spectrosc.* **1987**, *19*, 35.
(137) Tjandra, N.; Omichinski, J. G.; Gronenborn, A. M.; Clore, G. M.; Bax, A. *Nat. Struct. Mol. Biol.* **1997**, *4*, 732.

(138) Kung, H. C.; Wang, K. Y.; Goljer, I.; Bolton, P. H. *J. Magn. Reson., Series B* **1995**, *109*, 323.
(139) Prestegard, J. H.; Tolman, J. R.; Al-Hashimi, H. M.; Andrec, M. In *Biological Magnetic Resonance*; Krishna, N. R., Berliner, L. J., Eds.; Plenum: New York, 1999; Vol. 17, p 311.
(140) Latham, M. P.; Hanson, P.; Brown, D. J.; Pardi, A. *J. Biomol. NMR* **2008**, *40*, 83.
(141) Al-Hashimi, H. M.; Majumdar, A.; Gorin, A.; Kettani, A.; Skripkin, E.; Patel, D. J. *J. Am. Chem. Soc.* **2001**, *123*, 633.
(142) Ottiger, M.; Delaglio, F.; Bax, A. *J. Magn. Reson.* **1998**, *131*, 373.
(143) Zidek, L.; Wu, H.; Feigon, J.; Sklenar, V. *J. Biomol. NMR* **2001**, *21*, 153.
(144) Permi, P. *J. Biomol. NMR* **2002**, *22*, 27.
(145) Pitt, S. W.; Majumdar, A.; Serganov, A.; Patel, D. J.; Al-Hashimi, H. M. *J. Mol. Biol.* **2004**, *338*, 7.
(146) Brutscher, B.; Boisbouvier, J.; Pardi, A.; Marion, D.; Simorre, J. P. *J. Am. Chem. Soc.* **1998**, *120*, 11845.
(147) Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 12366.
(148) Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K. *J. Am. Chem. Soc.* **1998**, *120*, 6394.
(149) Fiala, R.; Czernek, J.; Sklenar, V. *J. Biomol. NMR* **2000**, *16*, 291.
(150) Ying, J.; Wang, J.; Grishaev, A.; Yu, P.; Wang, Y. X.; Bax, A. *J. Biomol. NMR* **2011**, *51*, 89.
(151) Boisbouvier, J.; Bryce, D. L.; O'Neil-Cabello, E.; Nikonowicz, E. P.; Bax, A. *J. Biomol. NMR* **2004**, *30*, 287.
(152) Yan, J.; Corpora, T.; Pradhan, P.; Bushweller, J. H. *J. Biomol. NMR* **2002**, *22*, 9.
(153) O'Neil-Cabello, E.; Bryce, D. L.; Nikonowicz, E. P.; Bax, A. *J. Am. Chem. Soc.* **2004**, *126*, 66.
(154) Miclet, E.; Boisbouvier, J.; Bax, A. *J. Biomol. NMR* **2005**, *31*, 201.
(155) Jaroniec, C. P.; Boisbouvier, J.; Tworowska, I.; Nikonowicz, E. P.; Bax, A. *J. Biomol. NMR* **2005**, *31*, 231.
(156) Pervushin, K. *Q. Rev. Biophy.* **2000**, *33*, 161.
(157) Zhu, G.; Yao, X. J. *Prog. Nucl. Magn Reson. Spectrosc.* **2008**, *52*, 49.
(158) Vallurupalli, P.; Moore, P. B. *J. Biomol. NMR* **2002**, *24*, 63.
(159) Wu, Z. R.; Tjandra, N.; Bax, A. *J. Biomol. NMR* **2001**, *19*, 367.
(160) Carlomagno, T.; Hennig, M.; Williamson, J. R. *J. Biomol. NMR* **2002**, *22*, 65.
(161) Meiler, J.; Prompers, J. J.; Peti, W.; Griesinger, C.; Bruschweiler, R. *J. Am. Chem. Soc.* **2001**, *123*, 6098.
(162) Peti, W.; Meiler, J.; Bruschweiler, R.; Griesinger, C. *J. Am. Chem. Soc.* **2002**, *124*, 5822.
(163) Tolman, J. R. *J. Am. Chem. Soc.* **2002**, *124*, 12020.
(164) Tolman, J. R.; Al-Hashimi, H. M. In *Annu. Rep. NMR Spectrosc*. Webb, G. A., Ed.; Academic Press: 2003; Vol. 51, p 105.
(165) Zhang, Q.; Sun, X.; Watt, E. D.; Al-Hashimi, H. M. *Science* **2006**, *311*, 653.
(166) Eichhorn, C.D.; Yang, S.; Al-Hashimi H.M. *Recent Developments in Biomolecular NMR* Royal Society of Chemistry Publishing, Cambridge, **2012**

Chapter 2

Measuring Similarity Between Dynamic Ensembles

2.1 Introduction

There is growing interest in moving beyond a static description of biomolecules towards a dynamic description in terms of conformational ensembles[1-8] in which a biomolecule is represented as a population-weighted distribution of many conformations. Studies indicate that biomolecules employ this broad pool of conformations during folding and when carrying out their biological functions[9]. An ensemble description of biomolecules can also help quantify thermodynamically important conformational entropy[10] and define a broad range of receptors that can be targeted in drug discovery[11].

Methods to assess similarity between static structures are well developed and widely used in classifying biomolecules, understanding evolutionary relationships between them, and in predicting their structures and functions[12,13]. New methods are needed to compare dynamic ensembles of biomolecules[14]. This is important not only for helping establish dynamics-function relationships[9], but also in assessing the quality of ensembles determined using experimental and computational methods[3,14]. Among many approaches for comparing probability distributions, the Jensen-Shannon Divergence (Ω^2)[2,14] and *S*-score (*S*)[15] have been used to compare dynamic ensembles of biomolecules. While these approaches provide quantitative information regarding ensemble similarity, particularly with regards to the population overlap between two distributions, they do not quantify the extent of structural similarity for non-overlapping conformations.

For example, based on Ω^2 or *S*-score, two very similar yet non-overlapping conformational ensembles (gray and green in Figure 2.1a) are measured as having zero similarity. The same level of similarity is assigned to two conformational ensembles that differ

much more substantially (gray and magenta ensembles in Figure 2.1a). The underlying problem is that non-overlapping conformations in two distributions contribute to Ω^2 and *S* in manner independent of the extent of structural similarity (see **Methods**). Other common measures of similarity or distance between probability distributions suffer from the same limitation including the χ^2 and the Bhattacharyya distance. In addition, in application to ensembles, Ω^2 and *S*-score are typically reported for an arbitrarily chosen bin size used to describe a given structural variable. However, these measures of similarity are highly dependent on bin size or method used to cluster conformations in an ensemble[2,14]. Other approaches for comparing ensembles that involve computing the pairwise RMSD in atomic positions between every pair of conformations in two ensembles (eRMSD)[16] do not capture the population overlap, cannot be generally used to dissect individual structural degrees of freedom, and can be obscured by outliers.

We developed an approach for simultaneously quantifying population overlap and structural similarity between ensembles. Here, the overlap between two distributions is evaluated using methods such as Ω^2 and *S*-score as a function of increasing the bin size used to build the histogram describing a given structural variable, such as a torsion angle or distance. This approach captures improvements in the quality of ensembles determined using increasing input experimental data that go undetected using conventional methods and reveals unexpected similarities between RNA ensembles determined using NMR and molecular dynamics simulations.

2.2 Methods

2.2.1 Jensen-Shannon Divergence (Ω^2) and *S*-score

Mathematical expressions for the Jensen-Shannon Divergence (Ω^2) and *S*-score are given by Equations 1 and 2, respectively:

$$\Omega^2\left(w_i^T(m), w_i^P(m)\right) = S\left(\frac{w_i^T(m)+w_i^P(m)}{2}\right) - \frac{1}{2}\left[S\left(w_i^T(m)\right) + S\left(w_i^P(m)\right)\right] \quad (2.1)$$

$$S\left(w_i^T(m), w_i^P(m)\right) = \frac{1}{2}\sum_{i=1}^{N}\left|w_i^T(m) - w_i^P(m)\right| \quad (2.2)$$

in which $\{w_i^T(m)\}$ and $\{w_i^P(m)\}$ represent the population weights for the i^{th} bin in ensemble T and P, respectively for a given bin size, m. $S(w_i)=-\Sigma w_i(m)\log_2 w_i(m)$ in Equation 2 is the information entropy. Ω^2 and S vary between 0 and 1 for maximum and minimum similarity, and are equal to zero if and only if $\{w_i^T(m)\} = \{w_i^P(m)\}$. Equations 1 and 2 show that for non-overlapping regions in two distributions, defined as cases in which $\{w_i^T(m)\}=0$; $\{w_i^P(m)\}\neq 0$ or $\{w_i^T(m)\} \neq 0$; $\{w_i^P(m)\}=0$, the contribution to Ω^2 and S is independent of the extent of structural similarity.

The sum of population overlap over all bin sizes (K) normalized relative to values expected for worst predictions ($\Omega = 1$ for all bin sizes or random selection) provides a convenient single-value measure of population overlap and structural similarity which we refer to as $\Sigma_K\Omega(w^T, w^P)$ that ranges between 0 and 1 for perfect and zero similarity, respectively,

$$\sum_K \Omega(w^T, w^P) = \frac{\sum_m \Omega\left(w_i^T(m), w_i^P(m)\right)}{K} \tag{2.3}$$

Note that $\Sigma_K\Omega(w^T, w^P)$ is also a metric, and therefore symmetric $\Sigma_K\Omega(w^T, w^P) = \Sigma_K\Omega(w^P, w^T)$ and equal to zero if and only if two distributions are identical at all bin sizes or $\{w^T\} = \{w^P\}$.

2.2.2 Sample and Select (SAS) approach

In the SAS approach[18-20], experimental RDCs are used to guide construction of an ensemble by selecting N conformations from a conformational pool that minimize the following χ^2 function,

$$\chi^2 = \sum_{i=1}^{L}\left(D_i^{calc} - D_i^{exp}\right)^2/L \tag{2.4}$$

in which L is the total number of RDCs used in SAS, D_i^{calc} and D_i^{exp} are calculated and experimentally measured RDCs, respectively. In our implementation of SAS, first an initial ensemble of N conformations is randomly selected from the pool. Then at each step (k) of the selection procedure one conformation in the ensemble is randomly chosen and replaced by a conformation randomly selected from the rest of the pool. The change from step k to k+1 is accepted if $\chi^2(k+1) < \chi^2(k)$; if $\chi^2(k+1) \geq \chi^2(k)$ with a probability P=exp(($\chi^2(k)$-$\chi^2(k+1)$)/T), where

T is an effective temperature that is linearly decreased using a simulated-annealing scheme[18]. The initial effective temperature is set to sufficiently high so that >99% of the conformations can be replaced and slowly decreased until the acceptance probability is smaller than 10^{-5}. At each effective temperature, 200,000 steps were implemented followed by a decrease of effective temperature using $T_{i+1}=0.9T_i$. A MATLAB script (Appendix 1) was used to implement this SAS-based ensemble construction.

2.2.3 Evaluating quality of inter-helical ensembles determined with increasing input RDCs

The capability of RDCs to reconstruct inter-helical ensembles using the SAS approach was investigated using synthetic RDC data, using up to five RDC data sets corresponding to five perfectly orthogonal alignment tensors. In these simulations, a given conformation is represented using three inter-helical Euler angles (α_h, β_h, γ_h) describing the relative orientation of the two idealized A-form helices representing the TAR helices connected by a trinucleotide bulge (Figure 2.2a). The conformational pool necessary for the SAS selection was generated by using the corresponding topologically allowed space. This space corresponds to all possible inter-helical orientations that satisfy basic steric and connectivity restraints imposed by the bulge[21]. The pool was generated using a 5º-resolution grid (i.e. each conformation differs from its closest neighbor by a 5º change in one of the three Euler angles). For a trinucleotide bulge, the pool represents ~10% of the total possible inter-helical orientations. A target ensemble containing five distinct conformations (N=5) was then randomly selected from this topologically allowed pool. Five orthogonal alignment tensors arbitrarily fixed on the reference helix were then generated using the Gram-Schmidt procedure[22]. For each of the five alignment tensors, all possible one bond CH RDC were computed for the target ensemble. For each alignment tensor, the RDCs for the five conformations were averaged and error-corrupted assuming 2Hz RDC error.

The SAS approach was then implemented to select an ensemble of N=5 distinct conformations using one, two, three, four and five sets of input RDCs to guide selection. The target and the predicted ensemble were then compared using similarity measurements including Ω, S-score, χ^2 and Bhattacharyya distance at various bin sizes as described below. The same

process was repeated 50 times and the similarity between target and predicted ensembles were averaged over these 50 comparisons at each bin size. For the RDC cross-validation analysis, ensembles determined using one, two, three and four RDC data sets in the SAS selection were used to predict a fifth RDC data set that was not used in the selection. The resultant RMSD between the RDCs for this fifth data set and values back-calculated from the predicted ensemble was then computed.

2.2.4 Binning inter-helical orientations

The Cartesian distance in the Euler space, $((\alpha_{hA}- \alpha_{hB})^2 + (\beta_{hA}- \beta_{hB})^2 + (\gamma_{hA}- \gamma_{hB})^2)^{1/2}$ between two sets of Euler angles *A* and *B* defining two distinct inter-helical orientations does not provide a measure of structural similarity between the two conformations[21]. First, there are inherent degeneracies (α_h'=α_h+180, β_h'=-β_h, γ_h'=γ_h+180; α_h'=α_h-180, β_h'=-β_h, γ_h'=γ_h-180; α_h'=α_h+180, β_h'=-β_h, γ_h'=γ_h-180; α_h'=α_h-180, β_h'=-β_h, γ_h'=γ_h+180) that map several sets of distinct inter-helical Euler angles to the same conformation[21]. This problem was overcome by using a restricted grid of Euler angles devoid of any degeneracy[21]. Second, even after taking into account the above degeneracy, the Cartesian distance between two sets of Euler angles does not provide a faithful measurement of structural similarity. For example, the Cartesian distances between (0, 0, 0) and (5, 5, 5) is ~9^o in the Euler space whereas the two conformations differ by single axis rotation with amplitude ~11^o. Likewise, the conformations (5, 5, 0) and (170, -10, 170) differ by a Cartesian distance of ~237^o but the two conformations differ by a single axis rotation with amplitude ~25^o. More generally, the Cartesian distance between Euler angles can be smaller than, equal to or larger than the actual difference between two conformations. Therefore we used the amplitude of single axis rotation to bin inter-helical orientations together and measure similarity between ensembles[21] (see below).

The binning grid points are constructed by picking a binning origin, defined by minimum value of each of the three Euler angle in the two ensembles upon comparison, and then incrementing each Euler angles by an amount defined by the bin size to cover the entire non-degenerate 3D Euler space. Changing in the binning origin has minimal effects on the resulting analysis (data not shown). Next, the amplitude of a single axis rotation (ω) connecting a given

conformation in the ensemble defined by Euler angles (α_{h1}, β_{h1}, γ_{h1}) and a point on the grid (α_{h2}, β_{h2}, γ_{h2}) is computed,

$$R(\alpha_{h1}, \beta_{h1}, \gamma_{h1}) = O(x, y, z, \omega)R(\alpha_{h2}, \beta_{h2}, \gamma_{h2}) \tag{2.5}$$

in which $O(x, y, z, \omega)$ represents a single axis rotation about a unit vector (x, y, z) with amplitude (ω). $O(x, y, z, \omega)$ can also be expressed by a 3 by 3 matrix in terms of x, y, z and ω

$$O(x, y, z, \omega) = \begin{pmatrix} cos\omega + x^2(1 - cos\omega) & xy(1 - cos\omega) - zsin\omega & xz(1 - cos\omega) + ysin\omega \\ xy(1 - cos\omega) + zsin\omega & cos\omega + y^2(1 - cos\omega) & yz(1 - cos\omega) - xsin\omega \\ xz(1 - cos\omega) - ysin\omega & xy(1 - cos\omega) + xsin\omega & cos\omega + z^2(1 - cos\omega) \end{pmatrix}$$

(2.6)

And the rotation amplitude ω is given by,

$$\omega = arccos\left(\frac{O_{11} + O_{22} + O_{33} - 1}{2}\right) \tag{2.7}$$

in which O_{11}, O_{22} and O_{33} are the three diagonal elements of $O(x, y, z, \omega)$.

In this manner, the amplitude of the single axis rotation connecting a given conformation in an ensemble to every grid point is computed, and the conformation is binned to the grid point that leads to the minimum single axis rotation amplitude ω. The population of each grid point is then calculated to be the number of conformations binned divided by the total number of conformations in the ensemble. In our case, binning of the target and the predicted ensemble led to two population distributions on the same binning grid for a given bin size, and the value of $\boldsymbol{\Omega}$ between the two ensembles at the given bin size is then calculated using equation 2. This procedure was repeated as a function of increasing bin size. This analysis was performed using a MATLAB script (Appendix 2).

2.2.5 Analysis of MD-trajectory-based ensembles

An in-house perl script was used to compute inter-helical angles (α_h, β_h, γ_h) describing the relative orientation of two A-form helices[21]. All intra- and inter- base-pair parameters were computed using Curves+[23] and all the local torsion angles defining the sugar and backbone

geometry were computed using an in-house C script. The resulting inter-helical orientations defined by three Euler angles were binned and analyzed as described above. Distributions of base-pair parameters, sugar and backbone torsion angles were directly binned to a binning grid ranging between 0^o and 360^o with evenly distributed increments defined by the bin size. The value of Ω was calculated at each given bin size for each parameter/angle distribution using Equation 1 and the values of $\Sigma\Omega$ are calculated using Equation 3 for distributions of inter-helical orientation, base-pair parameter, sugar, and backbone torsion angles.

2.3 Results and Discussion

The results show that increasing the bin size effectively reduces the 'structural resolution' with which a given structural variable is defined, and thereby increases the probability of binning conformations in two ensembles into common bins (Figure 2.1). Ensembles that differ substantially in structural terms will require larger bin sizes to overlap. We assess overlap using the square root of Ω^2 because it provides several desirable properties, including being a proven metric[2]. The value of Ω comparing two ensembles either stays constant (barring statistical noise) or decreases with increasing bin size, and always plateaus at Ω=0 at some bin size cut-off. The plot of Ω versus bin size (REsemble) then provides a rich 2D description of ensemble similarity that simultaneously captures population overlap and structural similarity, with the latter encoded in the steepness with which Ω drops with bin size. The approach readily accommodates outliers, which result in long lasting near-zero Ω plateaus, without compromising the ability to detect similarity in other regions of the ensemble (Figure 2.2). Summing the values of Ω over K bin sizes and normalizing relative to values expected for zero overlap yields a single-value metric $\Sigma_K\Omega(w^T, w^P)$ which ranges between 0 and 1 for perfect and zero similarity, respectively (see Methods).

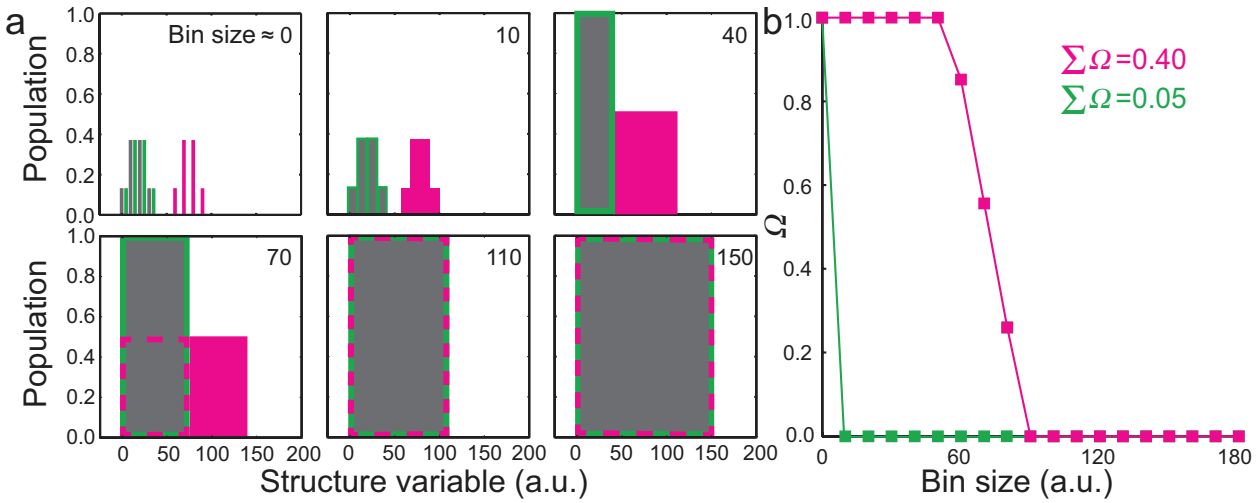

Figure 2.1 Measuring population overlap and structural similarity between ensembles. (a) Three discrete ensembles (gray, green, and magenta) described in terms of an arbitrary structural variable are shown as a function of increasing bin size used to build the histogram distribution. Dashed magenta and solid green boxes around the gray ensemble indicate the portion of magenta and green ensemble respectively that are binned together with the gray ensemble. **(b)** Plots of Ω as a function of increasing bin size comparing the gray vs. green (green line) and gray vs. magenta (magenta line) ensembles.

Applying this approach to our previous examples (Figure 2.1a), the structurally similar but non-overlapping ensembles (gray and green) start with $\Omega = 1$ for small bin sizes implying zero similarity, but Ω rapidly drops to zero with increasing bin size indicating strong structural similarity (Figure 2.1b). The drop in Ω with bin size is far less steep for the structurally more dissimilar ensembles (gray and magenta) (Figure 2.1b). $\Sigma\Omega$ is clearly different in the two cases (0.05 and 0.40, Figure 2.1b) and captures the structural differences between the two ensembles.

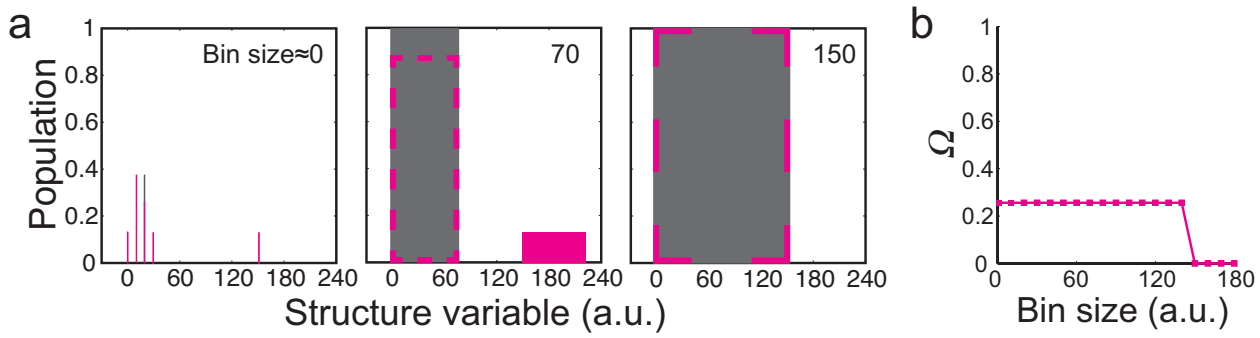

Figure 2.2 Measuring similarity between ensembles containing outliers. (a) Binning of two identical ensembles (gray and red) with the exception of a single outlier. **(b)** 2D Ω versus bin size plots measuring the similarity between the two ensembles. The relatively low Ω values at very small bin sizes accurately capture sharp similarities within the ensemble, the long lasting plateau captures the outlier and its structural dissimilarity, while the sharp drop in the Ω value to =0 at large bin size indicates that any outlier(s) are narrowly distributed.

Having the ability to measure ensemble similarity is fundamentally important for testing approaches currently under development for constructing ensembles of biomolecules using experimental data[3-8,18]. A common ensemble construction approach uses 'Sample and Select' (SAS)[18] (see Methods) or similar scheme[19] to guide selection of conformations from a computationally generated pool and construct ensembles that satisfy experimental data. Methods such as cross-validation[4,7,20] have been used to show that the quality of constructed ensembles generally improves with increasing input experimental data; however no study has directly quantified the extent or nature of the improvement.

We used our approach to measure the similarity between a known target ensemble (N=5) constructed by randomly selecting five conformations from a pool of ~40,000 conformations and ensembles reconstructed using SAS and up to five independent sets of synthetic residual dipolar couplings (RDCs)[24,25] (see Methods). For simplicity, we focused on determining ensembles describing the relative orientation of two chiral domains (in this case A-form RNA helices) as defined using three Euler angles (Figure 2.3a). Here, the conformational pool represents the topologically allowed orientations of two A-form helices linked by a trinucleotide bulge. As

described previously[21], we measure similarity in terms of the amplitude of single axis rotations (see **Methods**).

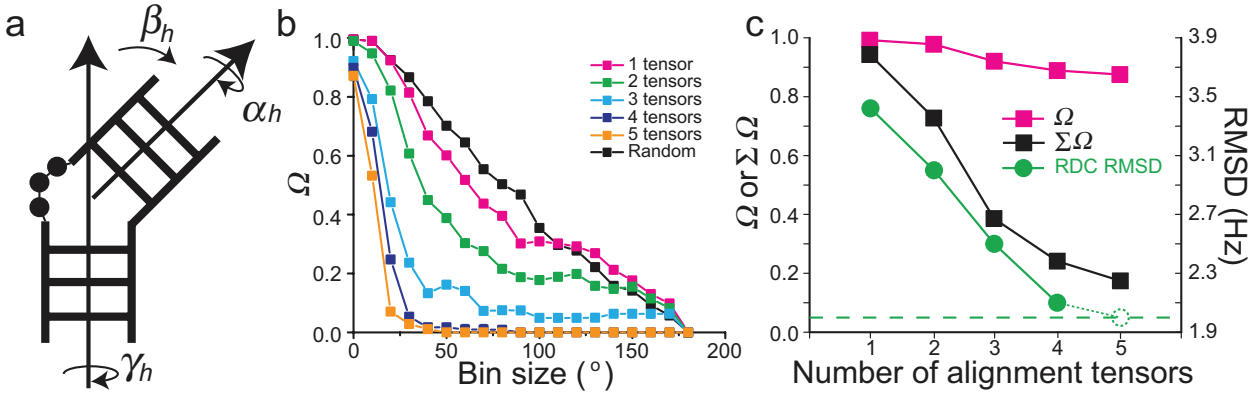

Figure 2.3 Prediction of ensembles using increasing RDC input sets. (a) The relative orientation of two helices (or domains) is defined using three Euler angles (α_h, β_h, γ_h). Shown are two RNA helices linked by a trinucleotide bulge. **(b)** Ω versus bin size comparing the inter-helical angle distributions about a trinucleotide bulge linker between a target ensemble (N=5) and ensembles (N=5) that are selected from the pool randomly (black) or using increasing number of input RDC data sets in SAS selections (color-coded, see inset). **(c)** The value of Ω at bin size ~0^o (magenta squares) and $\Sigma\Omega$ (black squares) as a function of number of RDC data sets used in ensemble reconstruction. Also shown is the root-mean-square-deviation (RMSD) in leave-out cross-validation in which a constructed ensemble is used to predict a common left out set of RDCs (green circles). The dashed circle represents the optimum RMSD when the left-out data set itself is included in the selection and the flat dashed line denotes the assigned 2-Hz RDC error.

The conventional Ω value computed between the target and SAS reconstructed ensemble at the smallest bin size of ≈0^o (see Methods) ranges between 0.87 and 0.99 (Figure 2.3b). This implies a very poor level of similarity that is comparable to that observed when comparing the target ensemble with an ensemble (N=5) constructed by randomly selecting conformations from the same pool without guidance from RDC data (Ω=0.99) (Figure 2.3b). Moreover, Ω changes insignificantly when increasing the number of RDC data sets used to reconstruct the ensemble (Figure 2.3c). Similar results are obtained using the S-score, χ^2 (Figure 4) and Bhattacharyya distance (data not shown). These results are at odds with cross-validation analysis (see Methods),

which shows substantial improvements in the quality of ensembles determined with increasing RDC data sets as judged based on their ability to predict a common fifth RDC data set that is left out from the ensemble construction. The root-mean-square-deviation (RMSD) between measured and predicted RDCs approaches the assigned RDC error when using four RDC data sets, implying strong similarity between the target and reconstructed ensembles (Figure 2.3c). This improvement in ensemble construction with increasing RDC data sets is perfectly captured when computing Ω as a function of increasing bin size. Ω decreases with increasing bin size and this reduction occurs more rapidly when a larger number of RDC data sets is used in the ensemble construction (Figure 2.3c). This decrease is much less steep for the randomly selected ensemble (Figure 2.3b) resulting in $\Sigma\Omega$ values that decrease with increasing input RDC data sets, in excellent agreement with the cross-validation results (Figure 2.3c).

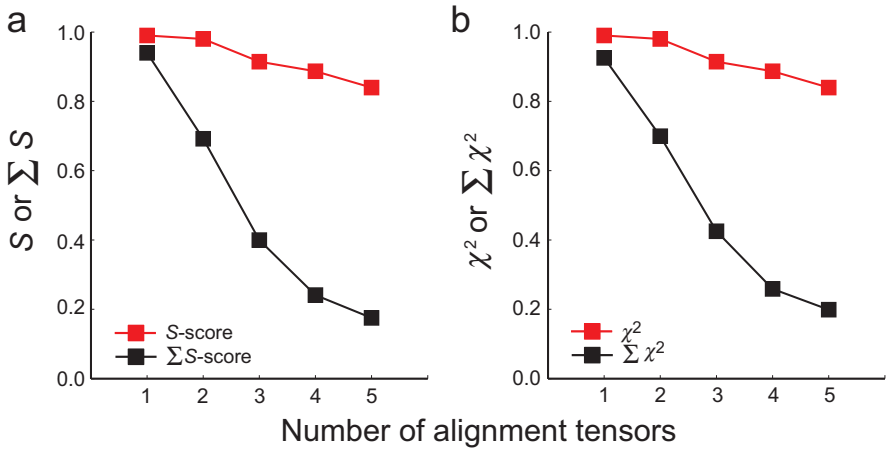

Figure 2.4 Measuring similarity between ensembles using *S*-score (*S*) and χ^2. Comparison is as in Figure 2.3c but using the normalized **(a)** *S*-score as measure of similarity between the target and the reconstructed ensemble. **(b)** Similar results are obtained when χ^2 is used as the measure of similarity between ensembles.

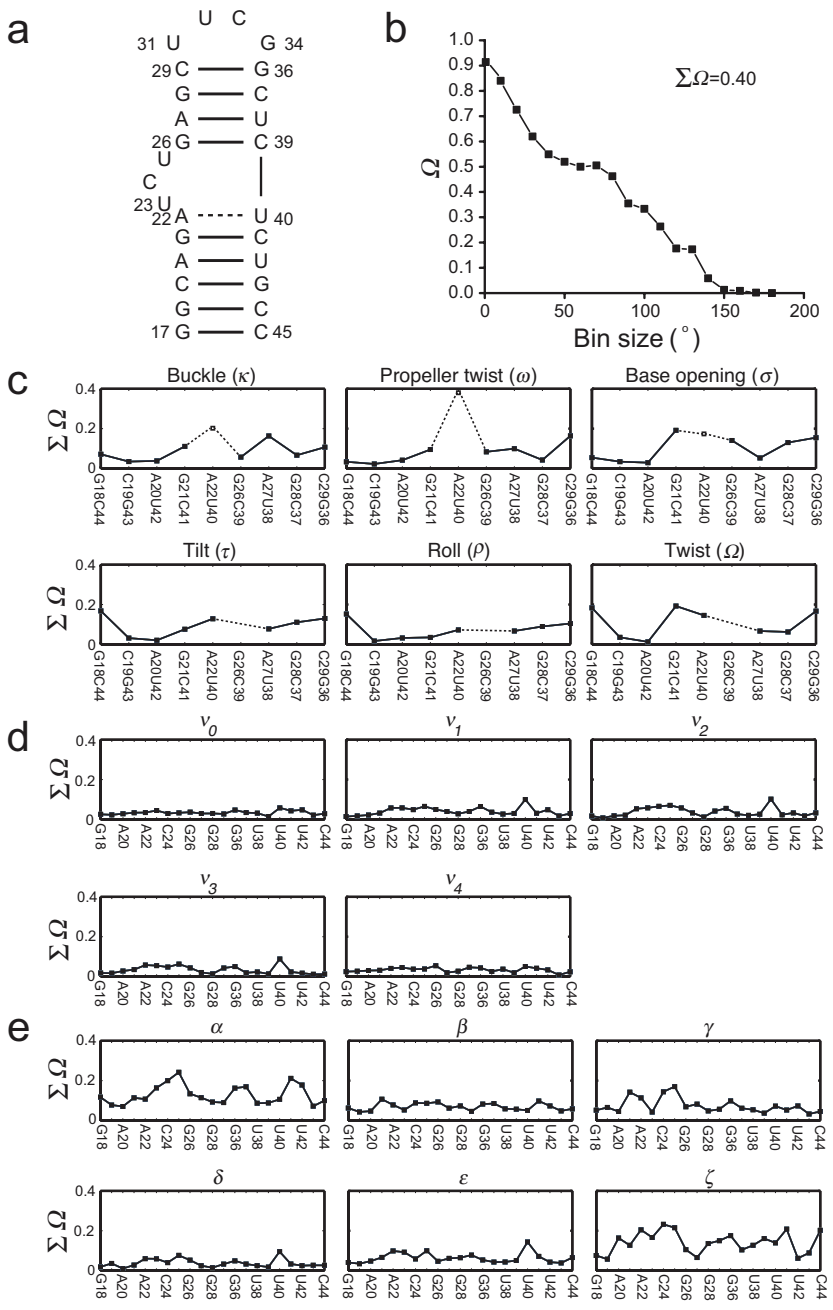
a
b
c
d
e
ΣΩ=0.40
Bin size (°)
Buckle (κ)
Propeller twist (ω)
Base opening (σ)
Tilt (τ)
Roll (ρ)
Twist (Ω)
ν0
ν1
ν2
ν3
ν4
α
β
γ
δ
ε
ζ

Figure 2.5 Comparing MD-generated and NMR-RDC selected ensembles of HIV-1 TAR. **(a)** Secondary structure of HIV-1 TAR RNA. The highly flexible junction A22-U40 base pair is indicated using a dashed line. **(b)** Ω versus bin size plots comparing the inter-helical angle distribution in the MD and RDC-selected (N=20) ensembles. The binning is performed in terms of single-axis rotation amplitudes (see Methods). (**c-e**) $\Sigma\Omega$ value comparing the distributions of **(c)** base-pair parameters, **(d)** sugar and **(e)** backbone torsion angles between the MD and the RDC selected ensemble. The intra-base-pair parameters for the flexible junction A22-U40 base pair are shown using open symbols and dashed lines and inter-base-pair parameters are not shown for the junction G26-C39 base pair because they are ill-defined due to presence of the bulge between G26-C39 and A22-U40.

We also used our approach to assess the quality of an ensemble determined for the transactivation response element (TAR) RNA (Figure 2.5a) from the human immunodeficiency virus type 1 (HIV-1) using molecular dynamics simulations. We previously reported[20] poor agreement (RMSD = 8.6 Hz; experimental uncertainty ~ 2 Hz) between four independent sets of RDCs measured in TAR (Figure 2.6) and RDCs predicted for a TAR ensemble obtained from an 8.2 μs MD simulation computed on Anton supercomputer using the CHARMM36 force field[20]. The specific degrees of structural freedom that underlie this disagreement remain unclear and are difficult to resolve given that RDCs report on both local and global aspects of structure[24,25].

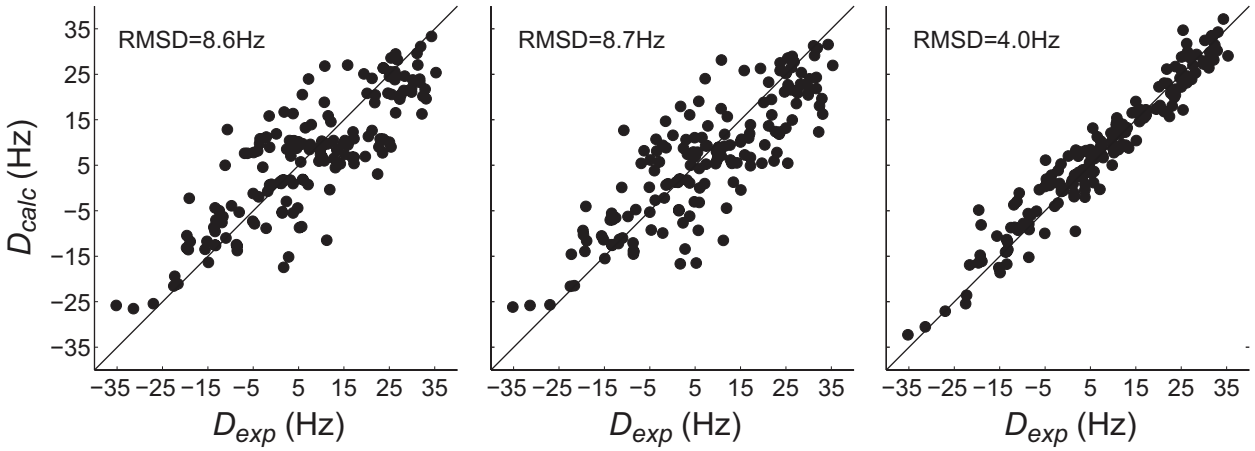

Figure 2.6 Comparison of experimentally measured and calculated RDCs using Anton MD trajectory, randomly selected ensemble, and SAS selected ensemble. The RDCs are calculated from and averaged over the entire 8.2 μs Anton MD trajectory (left panel), by combining 10 sets of 20 random conformers (middle panel) and by selecting 20 conformers using SAS approach (right panel) and compared with the experimentally measured RDCs.

We previously showed[20] that using the SAS approach, a TAR ensemble that much better satisfies the four sets of RDCs could be constructed from the MD-generated pool (Figure 2.6). To assess the source of discrepancy between the MD simulation and measured RDCs, we used our approach to directly compare the MD trajectory and the SAS-based RDC-selected ensemble. We observed substantial differences ($\Sigma\Omega$= 0.40) in the inter-helical angle distributions between the two ensembles (Figure 2.5b). This discrepancy alone is expected to affect all RDCs measured in TAR because changes in inter-helical orientation lead to changes in the global structure and overall alignment of the molecule. The observed differences in inter-helical angle distributions are not surprising given that longer simulations are likely needed to properly sample conformational space, and that the TAR inter-helical orientation strongly depends on ionic strength[26].

In contrast, we observed much better agreement for local angle parameters, including base-pair parameters **(**Figure 2.5c**)**, sugar **(**Figure 2.5d**)** and phosphodiester backbone torsion angles **(**Figure 2.5e**)** where on average $\Sigma\Omega < 0.2$. Cases with $\Sigma\Omega > 0.2$ are rare and tend to be

concentrated in the junction A22-U40 base-pair and bulge residues which have previously been shown to be flexible by NMR spin relaxation[27], and the phosphodiester backbone torsion angles α and ζ which show broad distributions in the MD-ensemble. The deviations in α and ζ at the bulge linker, and in base-pair parameters for residues surrounding the bulge are likely linked to the deviations observed in the inter-helical angle distributions (Figure 2.5b). The ability of RDCs to define all the above angles during the SAS selection was confirmed by simulation tests (Figure 2.7). In the simulation tests, substantial improvement in the prediction of inter-helical orientation is observed for the SAS selected ensemble (for both N=20 or 100), leading to

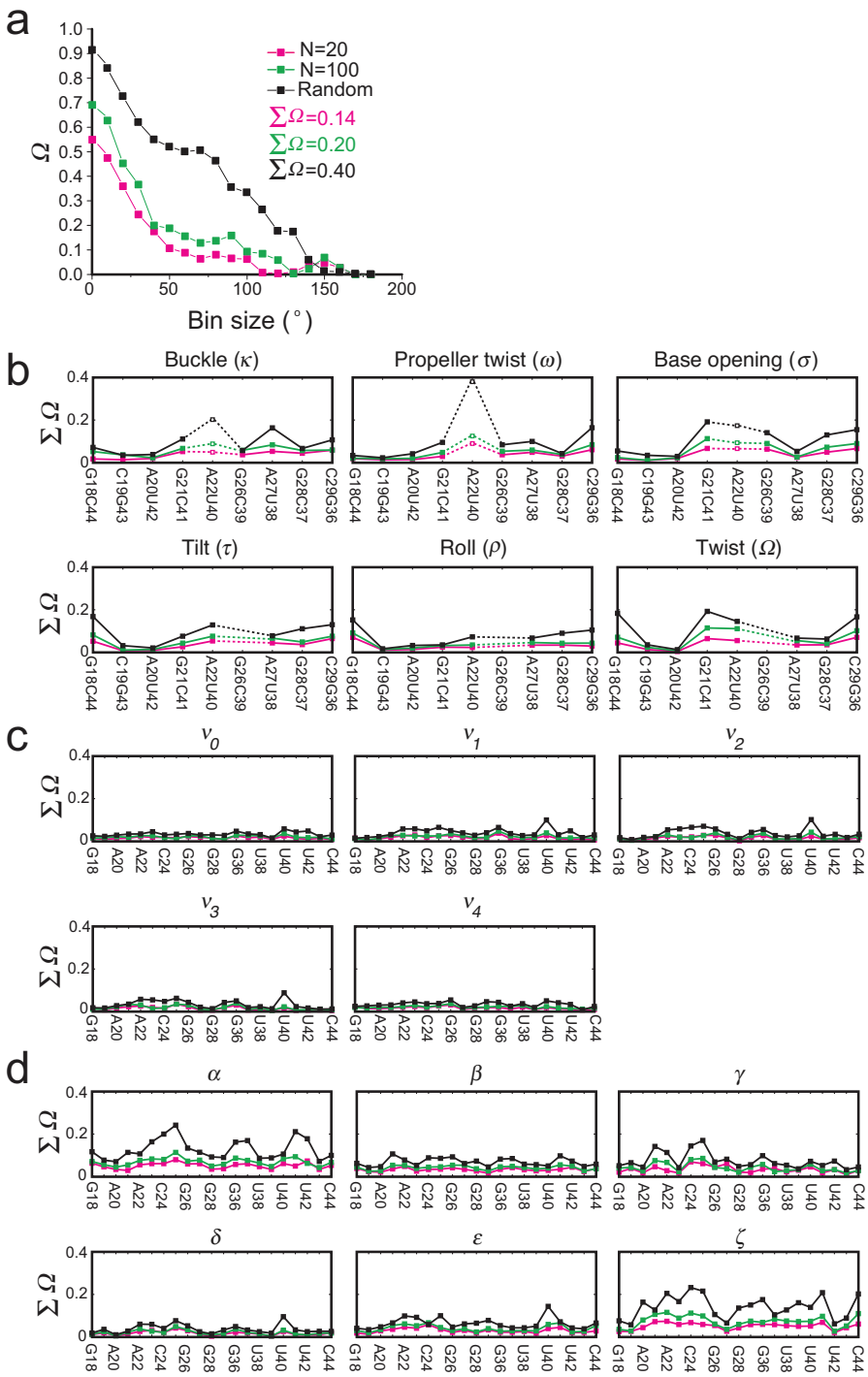
a
N=20
N=100
Random
∑Ω=0.14
∑Ω=0.20
∑Ω=0.40
Bin size (°)
b
Buckle (κ)
Propeller twist (ω)
Base opening (σ)
Tilt (τ)
Roll (ρ)
Twist (Ω)
c
d

Figure 2.7 Investigating the selection power of the SAS approach. A Monte-Carlo based scheme was used to investigate the selection power of SAS approach. The SAS selected TAR ensemble[20] is used as the target ensemble for which 163 independently noise corrupted RDC data sets are generated corresponding to the experimentally available RDC dataset. The SAS approach was implemented to predict the target ensemble using N=20 and N=100. A corresponding random selected ensemble is also presented. The comparison between the target versus RDC-selected (N=20 and N=100) and target versus randomly selected ensembles is shown in magenta, green and black respectively using Ω and $\Sigma\Omega$ for **(a)** inter-helical orientation and $\Sigma\Omega$ for **(b)** base-pair parameters, **(c)** sugar, and **(d)** backbone torsion angles. The intra-base-pair parameters for the flexible junction A22-U40 base pair are shown using open symbols and dashed lines and inter-base-pair parameters are not shown for the junction G26-C39 base-pair because they are ill-defined due to the presence of the bulge between G26-C39 and A22-U40.

corresponding $\Sigma\Omega$ values that indicate a good level of prediction (similar as local angles). The prediction of base-pair parameters (Figure 2.7b), sugar (Figure 2.7c) and backbone (Figure 2.7d) torsion angles consistently show that the SAS approach provides better predictions than the randomly selected ensemble. It is interesting to note that by defining inter-helical orientation and helical parameters, RDCs indirectly help define phosphodiester backbone torsion angles in and around the bulge. These results suggest that even though the MD trajectory yields poor agreements with RDCs measured throughout TAR, the main source of disagreement is the inter-helical angle distribution.

2.4 Conclusion

In conclusion, we have developed a simple and robust method, REsemble, to measure the similarity between dynamic ensembles that overcomes limitations in conventional methods that primarily capture population overlap at a single bin size and thereby fail to measure structural similarity. The approach can be used in conjunction with many other appropriate metrics for measuring ensemble similarity to compare any structural variable of interest. We anticipate many useful applications of this approach in dynamics-function studies.

This work is published in Nat. Methods[28]. The idea was conceived by Yang, S., Salmon L. and Al-Hashimi, H. M. Yang, S. and Salmon L. analyzed the data with help from Al-Hashimi H. M. Yang, S., Salmon L. wrote the scripts for analysis of the results.

2.5 References

(1) Shi, X.; Herschlag, D.; Harbury, P. A. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E1444.
(2) Fisher, C. K.; Huang, A.; Stultz, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 14919.
(3) Marsh, J. A.; Teichmann, S. A.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2012**, *22*, 643.
(4) Jensen, M. R.; Markwick, P. R.; Meier, S.; Griesinger, C.; Zweckstetter, M.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169.
(5) Showalter, S. A.; Johnson, E.; Rance, M.; Bruschweiler, R. *J. Am. Chem. Soc.* **2007**, *129*, 14146.
(6) Best, R. B.; Lindorff-Larsen, K.; DePristo, M. A.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 10901.
(7) Clore, G. M.; Schwieters, C. D. *Biochemistry* **2004**, *43*, 10678.
(8) Salmon, L.; Yang, S.; Al-Hashimi, H. M. *Annu. Rev. Phys. Chem.* **2013**.
(9) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789.
(10) Wand, A. J. *Curr. Opin. Struct. Biol.* **2013**, *23*, 75.
(11) Stelzer, A. C.; Frank, A. T.; Kratz, J. D.; Swanson, M. D.; Gonzalez-Hernandez, M. J.; Lee, J.; Andricioaei, I.; Markovitz, D. M.; Al-Hashimi, H. M. *Nat. Chem. Biol.* **2011**, *7*, 553.
(12) Richardson, J. S.; Richardson, D. C. *Annu. Rev. Biophys.* **2013**, *42*, 1.
(13) Erdin, S.; Lisewski, A. M.; Lichtarge, O. *Curr. Opin. Struct. Biol.* **2011**, *21*, 180.
(14) Lindorff-Larsen, K.; Ferkinghoff-Borg, J. *PLoS One* **2009**, *4*, e4203.
(15) De Simone, A.; Richter, B.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 3810.
(16) Bruschweiler, R. *Proteins* **2003**, *50*, 26.
(17) Cha, S.H. *International Journal of Mathematical Models and Methods in Applied Sciences.* **2007**, *1*, 300.
(18) Frank, A. T.; Stelzer, A. C.; Al-Hashimi, H. M.; Andricioaei, I. *Nucleic Acids Res.* **2009**, *37*, 3670.
(19) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. *Biophys. J.* **2007**, *93*, 2300.
(20) Salmon, L.; Bascom, G.; Andricioaei, I.; Al-Hashimi, H. M. *J. Am. Chem. Soc.* **2013**, *135*, 5457.
(21) Bailor, M. H.; Mustoe, A. M.; Brooks, C. L., 3rd; Al-Hashimi, H. M. *Nat. Protoc.* **2011**, *6*, 1536.
(22) Fisher, C. K.; Zhang, Q.; Stelzer, A.; Al-Hashimi, H. M. *J. Phys. Chem. B* **2008**, *112*, 16815.

(23) Lavery, R.; Sklenar, H. *J. Biomol. Struct. Dyn.* **1989**, *6*, 655.
(24) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 9279.
(25) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111.
(26) Casiano-Negroni, A.; Sun, X.; Al-Hashimi, H. M. *Biochemistry* **2007**, *46*, 6525.
(27) Zhang, Q.; Sun, X.; Watt, E. D.; Al-Hashimi, H. M. *Science* **2006**, *311*, 653.
(28) Yang, S.; Salmon, L.; Al-Hashimi, H. M. *Nat. Methods* **2014**, *Advance Online*

Chapter 3

Characterizing Uncertainty in Dynamic Ensembles of Biomolecules Determined Using Residual Dipolar Couplings

3.1 Introduction

An outstanding challenge in biophysics is to reconstruct the dynamics of biomolecules on the basis of experimental measurements[1-7]. Significant efforts have been directed in recent years towards constructing dynamic ensembles of biomolecules on the basis of ensemble-averaged experimental data, including several NMR interactions such as the Nuclear Overhauser effect (NOE)[8], paramagnetic relaxation enhancements (PRE)[9], chemical shifts (CS)[10,11] and residual dipolar couplings (RDC)[12-14] and small angle X-ray scattering (SAXS) data[15,16]. Here, one tries to solve for a conformational ensemble that accurately reproduces the measured averaged data in cases where a single static structure fails to reproduce the data within experimental error.

The determination of dynamic ensembles based on time- or ensemble-averaged data presents several challenges. The one we would like to examine in this study has to do with the fact that many distinct ensembles can often satisfy a given set of experimental data, which is an intrinsic limit in ensemble determination that can dramatically decrease the accuracy of the determined ensemble. While it can be trivial to find ensemble of conformations that satisfy experimental data within error[4], enumerating all possible ensembles that can equally satisfy the data and select the one that represents "reality" remains to be a challenging and largely unexplored problem.

Here, we seek to explore the accuracy with which conformational ensembles of biomolecules can be determined on the basis of ensemble-averaged experimental data and specifically NMR RDCs. There has been great interest in recent years in harnessing the broad time-scale and rich spatial sensitivity of RDCs in determining dynamic ensemble of

biomolecules, including globular proteins[17], intrinsically disordered proteins[18] and RNA[5,19]. We focus specifically on the problem of determining an ensemble of conformations describing the inter-helical orientation distribution of a simple RNA helix-junction-helix (HJH) motif, which can be defined and represented by only three Euler rotation angles[20,21]. Previous studies have shown that the helices composed of Watson-Crick base-pars can be very well approximated by a rigid idealized A-form geometry in which individual bond vectors experience uniform isotropic motions[22].

Determining the ensemble orientation of idealized helices represents an ideal ensemble determination problem for several reasons. First, one can pool a large number of RDCs measured on various vectors in a helix to characterize only three inter-helical Euler angles. Thus, there are much fewer degrees of freedom that have to be specified as compared to determining an atomic-resolution ensembles in which RDCs are used to determine distributions for both local and global degrees of freedom[3,4]. This provides a basis for illuminating the sources of uncertainty in ensemble determination. Second, the entire range of orientations that can be sampled by two helices can be defined *a priori* in an unbiased manner on the basis of topological constraints. Finally, it obviates the need to rely on conformational pools derived form other methods such as MD simulation in which correlations between degrees of freedom can affect data analysis. Besides providing a theoretical basis for assessing the accuracy of ensembles determined using RDCs, having the ability to determine ensembles of HJH motifs is very important for understanding RNA dynamics-function relationships.

Here we introduce a method for evaluating accuracy of determined ensemble of RNA inter-helical orientation using RDCs. The results reveal that although ensemble-averaged RDCs can be applied to reconstruct population-weighted ensembles that recover the underlying inter-helical distribution at a useful level of accuracy, even under ideal conditions, significant uncertainty remains. The uncertainty mainly results from experimental error and difficulties in establishing the optimal ensemble size used to construct the ensemble. The uncertainty arising from experimental error can be effectively suppressed by improving data collection/analysis schemes; however, the uncertainty arising from ensemble size used in ensemble determination is

very challenging if not impossible to suppress. Our results therefore suggest that dynamic ensembles of biomolecules should be determined using different ensemble sizes instead of only using the smallest ensemble size that can satisfy the experimental error (N_{min}) as commonly applied in most current ensemble determination strategies. The degenerate ensembles that cannot be distinguished by current experimental data should be further analyzed or validated using new experimental data in the future. Although the HJH model used in this study is RNA-based, our method is not model limited and can be generally applied to various biomolecules including DNA, protein and other biopolymers.

3.2 Methods

3.2.1 Constructing ensembles using Sample and Select (SAS)

We use the SAS approach[19,23] to construct population-weighted ensembles using RDCs. Here RDCs guide selection of conformations from a conformational pool containing thousands of conformations generated by molecular dynamics (MD) simulations or corresponding to an exhaustive set of allowed conformations (see Results and Discussion). In this approach, sub-ensembles with increasing size (number of distinct conformations) are constructed in an attempt to find number of distinct conformations (N) required to satisfy the measured RDCs. Here, N conformations are randomly selected from the pool and the agreement between measured and predicted RDCs is computed using equation 2.4 in Chapter 2.

Next, one of the conformations is replaced randomly by another conformation from the remaining conformations in the pool, and the agreement with measured RDCs is re-examined and the newly selected conformation is either accepted or rejected based on the Metropolis criteria: at each step (k) of the selection procedure, the change from step k to k+1 is accepted if $\chi^2(k+1) < \chi^2(k)$; if $\chi^2(k+1) \geq \chi^2(k)$ with a probability $P=\exp((\chi^2(k)-\chi^2(k+1))/T)$, where T is an effective temperature that is linearly decreased using a simulated-annealing scheme[19]. The initial effective temperature is set sufficiently high so that >99% of the conformations can be replaced and slowly decreased until the acceptance probability is smaller than 10^{-5}. At each effective temperature, 200,000 steps were implemented followed by a decrease of effective temperature using $T_{i+1}=0.9T_i$. Using such a simulated annealing based approach, many iterations are carried

out until the penalty function shown in equation 1 is minimized, defined as achieving the best agreement with the measured RDCs at ensemble size *N*.

The ensemble size is then incrementally increased in steps of 1 from *N*=1 until convergence is reached. Ensemble size *N* is chosen from the ensemble sizes that can reproduce the RDC within experimental error and the same SAS procedure was repeated using *N* for sufficient number of iterations until the population-weighted distribution reaches convergence. The population-weighted ensemble was then constructed by combining the sub-ensembles from all SAS iterations.

3.2.2 Determining accuracy of predicted ensembles

The accuracy of predicted ensemble is evaluated using recently developed REsemble[24] method as shown in Chapter 2 by calculating the similarity between the target and predicted ensembles at different bin sizes using the square root of Jensen-Shannon divergence (JSD) [4], Ω, as shown in equation 2.1 in Chapter 2. The similarity between target and predicted ensembles are measured using the same binning and comparison scheme as described in section 2.2.3 and 2.2.4 in Chapter 2.

3.3 Results and Discussion

3.3.1 Conformation pool

It has been emphasized in a recent review[1] that the accuracy of the dynamic ensemble determined using SAS approach highly depends on the conformation pool from which the conformations are selected. Many recent studies use conformation pools generated from molecular dynamics (MD) simulation[3,17,25]. Such MD-generated conformation pools, although provide encouraging results, could result in biased sampling of conformation space, which can possibly guide the prediction to degenerate ensembles that are far from target ensemble[26]. In this study, we used recently developed junction-topology allowed conformation space for rigid idealized two-way HJH motif of RNA in which the three Euler angles describing the inter-helical orientation are restricted and defined by imposing steric and connectivity constraints at the junction[20,21,27]. This junction-topology allowed space samples all possible inter-helical

orientations of the HJH motif in an unbiased manner and therefore largely avoids sampling imperfection that is commonly encountered in conformation space generated by MD simulations. Another advantage of this unbiased junction-topology allowed conformation space is that it only makes up <10% of all inter-helical orientations, which can remarkably improve the efficiency of computation.

3.3.2 Experimental error based uncertainty

To test the uncertainty in determined ensemble arising from experimental error, we constructed three different types of target ensembles: (1) ensembles in which all three Euler angles are defined by a Gaussian distribution with standard deviation equal to 10^{o}; (2) ensembles in which all three Euler angles are defined by a Gaussian distribution with standard deviation equal to 30^{o}; (3) ensembles randomly selected from the junction-topology allowed space. For each ensemble type, we generated eight different target ensembles containing 1, 2, 3, 4, 5, 10, 20 and 100 distinct and equally populated conformations. Five sets of linearly independent RDCs were then computed for each target ensemble assuming five orthogonal alignment tensors. Random RDC error of 0.5, 2, 4 and 8 Hz were assigned to the calculated RDCs of each target ensemble by assigning each RDC a number randomly selected from a normal distribution with standard deviation equal to the RDC error. SAS approach was then implemented to predict each target ensemble under these four different RDC uncertainties assuming number of distinct conformations (N) in target ensemble is known.

The target and predicted ensembles using SAS approach were then compared using REsemble[24]. The comparison between the target and the randomly selected ensemble $\Sigma\Omega r$ is used as the reference to normalize $\Sigma\Omega/\Sigma\Omega r$ value between target and predicted ensemble. The results revealed that for all target ensembles, the $\Sigma\Omega/\Sigma\Omega r$ value constantly increases with increasing RDC error indicating the fact that larger RDC errors lead to larger uncertainties in predicted ensembles as expected (Figure 3.1A-C). In particular, for single conformation target ensemble (N=1), only 8Hz RDC error results in $\Sigma\Omega/\Sigma\Omega r$ value slightly larger than 0. This is probably because the RDC RMSD resulted from two distinct conformations in junction-topology allowed space are larger than 4Hz and therefore the target conformation is the only conformation

that can satisfy the RDC RMSD in the junction-topology allowed space. However, there are some cases in which the RMSD between RDCs resulted from two distinct conformations are less than 8Hz, allowing these degenerate conformations to be selected in SAS approach that leads to $\Sigma\Omega/\Sigma\Omega r$ value larger than 0. This result suggests that smaller experimental error leads to smaller uncertainty in determined ensembles and therefore the uncertainty arising from experimental error can be effectively suppressed by improving experimental scheme leading to smaller experimental errors.

We also observed that as the width of the target ensemble increases (Gaussian distribution with standard deviation equal to from 10^{o} to infinitely large for random selection), the $\Sigma\Omega/\Sigma\Omega r$ value sharply increases (Figure 3.1A-C). This is because a broader region of conformation space from which the target ensemble is constructed contains more distinct conformations and therefore provides more combinations of conformations that can possibly form more degenerate predicted ensembles. This result reveals that a narrow ensemble distribution can likely be more accurately predicted than a broad ensemble distribution.

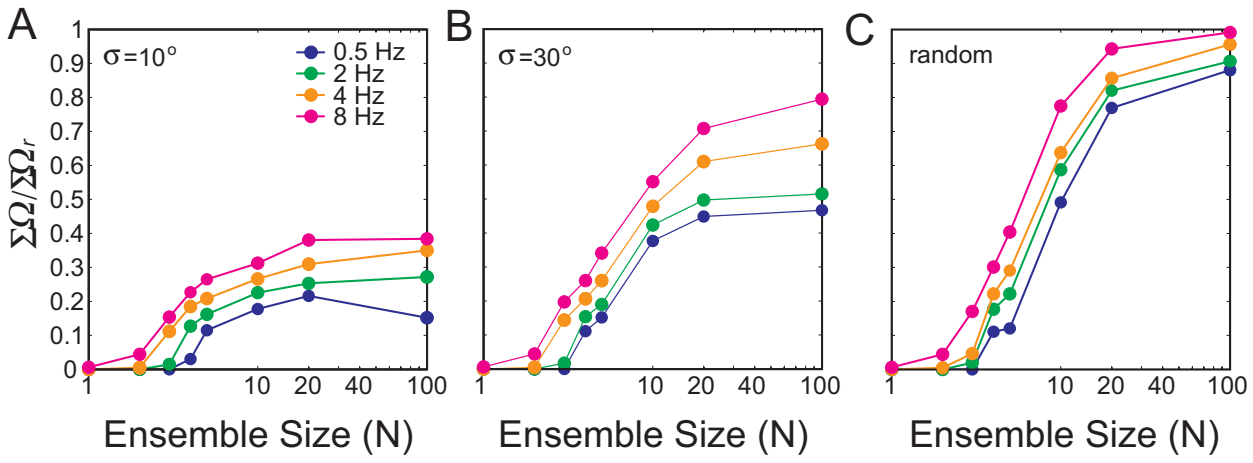

Figure 3.1 Uncertainty in determined ensembles arising from experimental errors. Predictions of target ensembles with all three Euler angles defined by Gaussian distribution with standard deviation equal to **(A)** 10°, **(B)** 30° and **(C)** randomly selected from junction-topology allowed space containing 1, 2, 3, 4, 5, 10, 20 and 100 conformations. The RDC errors mimicking experimental errors used in each target ensemble are 0.5Hz, 2Hz, 4Hz and 8Hz. All predictions were carried out assuming ensemble size *N* is known.

3.3.3 Ensemble size based uncertainty

In the analysis, we also observed that at a constant RDC error, $\Sigma\Omega/\Sigma\Omega r$ value increases with increasing ensemble size indicating that as the number of conformations in the target ensemble increases, there are more degenerate ensembles that can reproduce the RDCs within experimental error resulting in larger uncertainty in the determined ensemble. The only exception is the prediction of target ensemble containing 100 distinct conformations at 0.5Hz RDC error. For this prediction, the $\Sigma\Omega/\Sigma\Omega r$ value decreases compared with the one calculated from the prediction of the same type of target ensemble containing 20 distinct conformations (Figure 3.1A). This is likely because the number of distinct conformations in the narrow region of conformation space defined by the Gaussian distribution with standard deviation equal to 10° is around or smaller than 100 and 0.5Hz RDC error highly directs the SAS selection to this specific region with very low tolerance for conformations outside this region. Therefore this prediction selects most of the conformations in the same region from which the target ensemble is constructed and thereby has a better agreement with the target ensemble than the predicted

ensembles in the cases where the target ensemble makes up only a small portion of conformation space.

Although it is clear that larger ensemble size can result in more degeneracy and thus larger uncertainty in predicted ensembles from the prediction assuming ensemble sizes (N) is known as shown above, in practice the number of distinct conformations in a target or real dynamic ensemble cannot be known *a priori* and hence we can only construct the ensembles using an estimated optimal ensemble size. In current ensemble determination strategies, this optimal ensemble size N_{opt} is mostly chosen to be the smallest ensemble size that satisfies experimental error (N_{min}) and larger ensemble sizes are usually simply ignored due to the risk of overfitting of the experimental data. However, comparison between ensembles constructed using small and large ensemble sizes for which the risk of overfitting can be excluded has never been explicitly established and therefore whether N_{opt} is necessarily N_{min} is an open question. To address this question, we predict the target ensemble defined by Gaussian distribution with standard deviation equal to 10^{o} and 30^{o} and a random target ensemble containing 2, 5, 10, 100 conformations and 2Hz RDC error. Each target ensemble is predicted using N_{min} and an ensemble size N_{max} that is larger than N_{real} (N_{max}=500 for all predictions in this test). We then compared the resulting $\Sigma\Omega$ values from predictions using N_{min} and N_{max} with the corresponding $\Sigma\Omega$ values from prediction assuming N_{real} is known calculating the ratio $\Sigma\Omega/\Sigma\Omega(N_{real})$. The results reveal that for all predictions, N_{real} provide the most accurate predicted ensembles, as the ratio $\Sigma\Omega/\Sigma\Omega(N_{real})$ is equal to or larger than 1 for all predictions (Figure 3.2) although we cannot rule out the possibility that N_{min} and N_{max} can give more accurate predicted ensembles. For very small ensemble sizes, for example N=2, $\Sigma\Omega$ values resulted from N_{min} are the same as the ones from N_{real} likely because in these cases N_{min}=N_{real}, but $\Sigma\Omega$ values resulted from N_{max} are much larger because they give relative broad distributions instead of discrete ensembles. However for target ensembles with more distinct conformations, both N_{min} and N_{max} in general give larger $\Sigma\Omega$ values and hence worse predictions compared to the ensembles predicted using N_{real} (Figure 3.2). Interestingly, we observed that for target ensembles containing large number of distinct conformations (e.g. N_{real} = 100), N_{max} gives more accurate predicted ensembles compared to the

ones predicted using N_{min} regardless of the width of the target ensemble (Figure 3.2). This result implies that N_{min} can also possibly result in larger uncertainties in predicted ensemble and N_{opt} is not necessarily N_{min}. Therefore it suggests that dynamic ensembles of RNA or more general biomolecules should be determined using different ensemble sizes instead of N_{min} only.

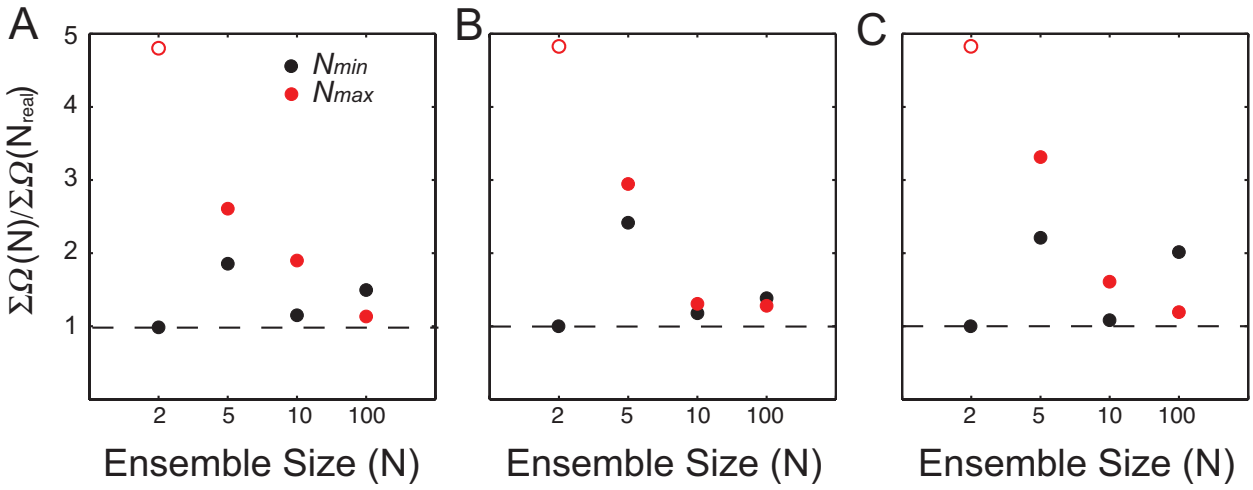

Figure 3.2 Uncertainty in determined ensembles arising from ensemble size. Prediction of target ensembles with all three Euler angles defined by Gaussian distribution with standard deviation equal to **(A)** 10°, **(B)** 30° and **(C)** randomly selected from junction-topology allowed space containing 2, 5, 10 and 100 distinct conformations. Each target ensemble was plotted using N_{min} (black), N_{real} and N_{max}=500 (red). The ratio $\Sigma\Omega/\Sigma\Omega(N_{real})$ for N=2 predicted using N_{max} (open circles) is not plotted according to the ratio because it approaches infinity in these predictions. All predictions were carried out using 2Hz RDC error.

To further explore the uncertainty in predicted ensemble arising from ensemble size, we used MD trajectory of HIV 1 TAR (8.2μs generated by Anton supercomputer and the CHARMM36 force field) as a more realistic target ensemble. We then used the inter-helical pool constructed from the junction-topology in the SAS approach and assumed RDC error of 2Hz. Once again, we find that the RDCs can be adequately back predicted in the SAS selection for $N \geq 4$ (Figure 3.3A). Thus we constructed ensembles assuming N_{min}=4 and an arbitrarily chosen N=100 as a second value of N. We find that the N=100 ensemble is a slightly better reproduction ($\Sigma\Omega$=0.12) than N_{min}=4 ($\Sigma\Omega$=0.18) (Figure 3.3B). In particular, the Ω values for N=100 are

smaller than those for N_{min}=4 at the small bin sizes (<80^o) indicating that N=100 gives more accurate ensemble; at large ensemble bin sizes (>80^o), Ω given by N=100 are slightly larger than the ones given by N_{min}=4 likely due to that the ensemble predicted using N=100 includes some lowly populated outliers, which can be seen from 1D distribution (Figure 3.3D) and 2D correlations (Figure 3.3E-G) of the Euler angles. We also tested the reproducibility of the SAS-determined ensembles by repeating predictions using N_{min}=4 and N=100 respectively and measuring the similarity between the predicted ensemble and the repeatedly predicted ensemble. The results reveal very high similarity between the predicted and repeatedly predicted ensembles using both N_{min}=4 and N=100, demonstrating that the determined ensembles are stable and reproducible (Figure 3.3C).

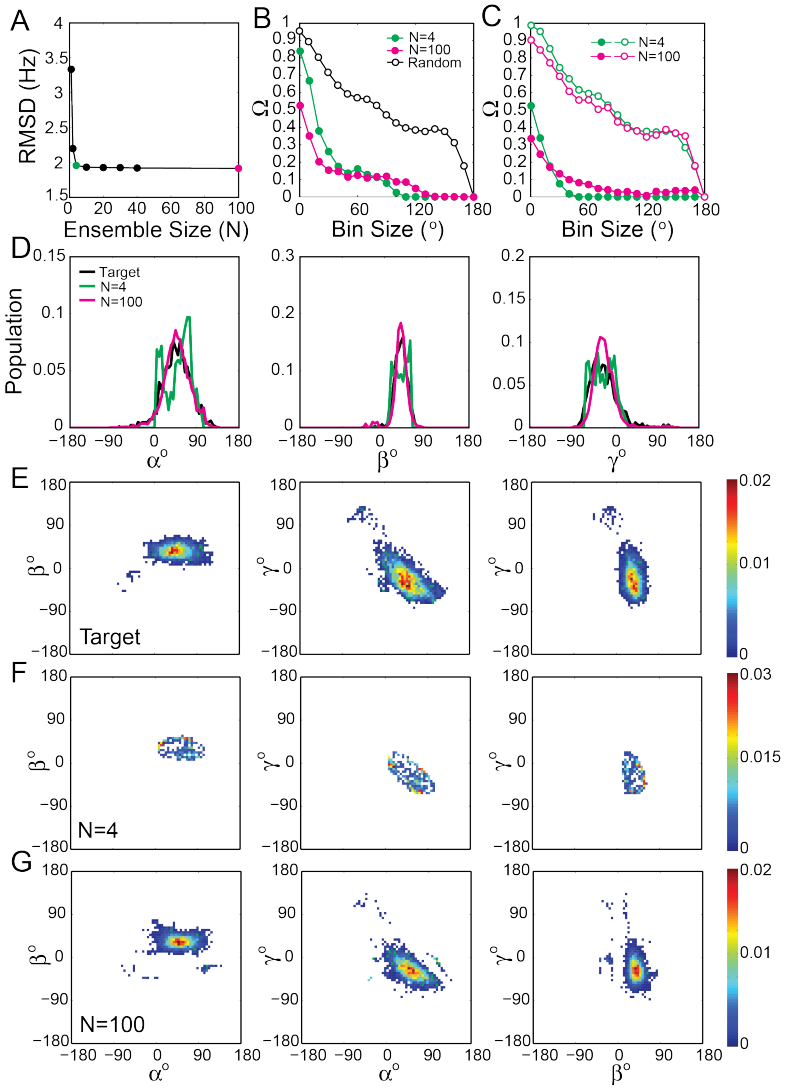

Figure 3.3 Prediction of MD trajectory of HIV 1 TAR. (A) RDC RMSD *versus* ensemble size (*N*); **(B)** Ω as a function of increasing bin size between the target ensemble and ensemble predicted using N_{min}=4 and N=100; **(C)** Test of reproducibility of predicted ensembles using N_{min}=4 (closed green circles) and N=100 (closed magenta circles). Random selected ensembles were plotted as references (open circles); **(D)** 1D distributions and **(E-G)** 2D correlations of three Euler angles of the target ensemble (MD trajectory of HIV 1 TAR) and predicted ensemble using N_{min}=4 and N=100.

3.3.4 Applications

Based on our benchmark studies, we developed a general approach for determining ensembles of inter-helical orientations for HJH motifs using RDCs and the topologically allowed inter-helical orientations. We accommodate the uncertainty in *N* by considering collections of ensembles determined with variable *N* that equally satisfy the RDC data. This approach is demonstrated in the determination of inter-helical ensembles for HIV-1 TAR in free and ligand bound forms.

HIV 1 TAR

RDCs of HIV 1 TAR were carefully measured in several recent studies[12-14,28]. Previous measured RDCs in domain 1 (EI-TAR) and domain 2 (EII-TAR) elongated TAR were used in combination to determine HIV 1 TAR dynamic ensemble from junction-topology allowed space. Here, RDCs belonging to junction base-pair A22-U40 were excluded due to high local flexibility[19,29]. The RDC RMSD *versus N* plots show that N_{min}=3 but that the RDCs can also be satisfied with much larger *N* value. We determined ensembles for N_{min}=3 and for *N*=100, both of which can reproduce the experimental RDCs with RMSD=2.8Hz (Figure 3.4A, B). To determine the dynamic ensemble of HIV 1 TAR, we repeated the SAS approach 300 times using N_{min}=3 and 9 times using *N*=100 and combined sub-ensembles selected in all SAS iterations in each prediction to form 900-conformation dynamic ensembles for both predictions. We found that the determination of N_{min}=3 and *N*=100 ensembles to be highly reproducible (Figure 3.4C). Both ensembles also pass two different cross-validations with similar accuracy: the leave-out cross-validation in which each RDC was omitted from the prediction of dynamic ensemble and then back-calculated from the resulting ensemble and compared to the corresponding measured RDC (Figure 3.4D); and the comparison between the measured RDCs of non-elongated HIV 1 TAR and RDCs calculated from the ensemble predicted using N_{min}=3 and *N*=100 and non-elongated HIV 1 TAR structure (Figure 3.4E). Therefore it is difficult to determine which ensemble more accurately captures the inter-helical orientation dynamics of HIV 1 TAR.

1D (Figure 3.4F) and 2D (Figure 3.4G, H) distributions of the three Euler angles clearly show that the N_{min}=3 and *N*=100 ensembles populate similar inter-helical Euler angles. Moreover, in

both cases, the RNA structure is not rigid, but rather samples a wide range of angles with small correlations observed between α and γ in the two cases (Figure 3.4G,H). Despite these similarities, the N=100 ensemble samples a relatively broader distribution of conformations as compared to the N_{min}=3 ensemble.

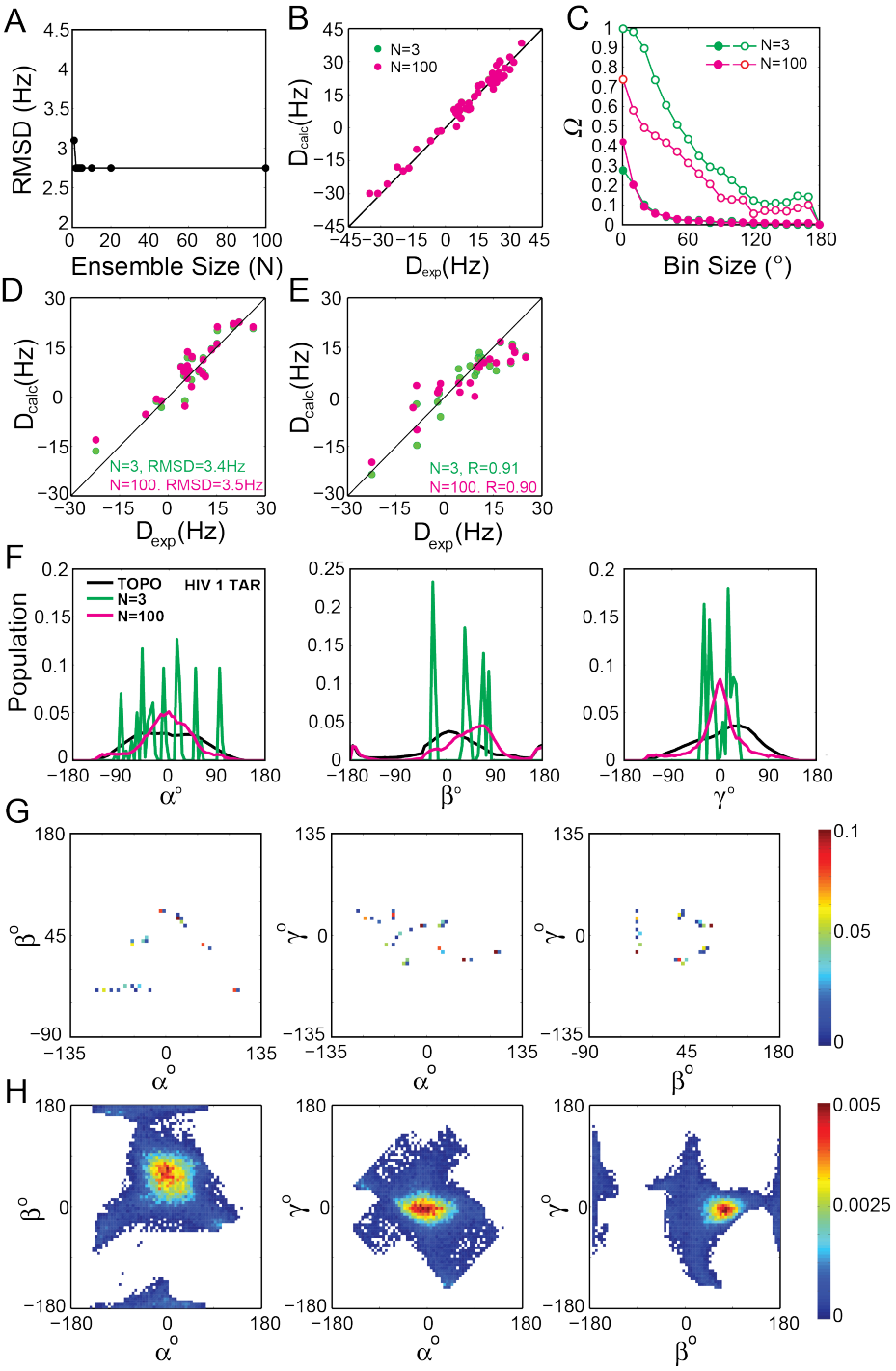
A
RMSD (Hz)
Ensemble Size (N)
B
N=3
N=100
D_calc(Hz)
D_exp(Hz)
C
N=3
N=100
Ω
Bin Size (°)
D
D_calc(Hz)
N=3, RMSD=3.4Hz
N=100. RMSD=3.5Hz
D_exp(Hz)
E
D_calc(Hz)
N=3, R=0.91
N=100. R=0.90
D_exp(Hz)
F
TOPO
N=3
N=100
HIV 1 TAR
Population
α°
β°
γ°
G
β°
α°
γ°
α°
γ°
β°
H
β°
α°
γ°
α°
γ°
β°

Figure 3.4 Prediction of inter-helical dynamic ensemble of HIV 1 TAR. (A) RMSD *versus* ensemble size (*N*); **(B)** correlation between experimental RDCs and RDCs calculated from ensembles predicted using N_{min}=3 (green) and *N*=100 (magenta); **(C)** test of reproducibility of ensemble predicted using N_{min}=3 (green, closed circle) and *N*=100 (magenta, closed circle). The comparison between randomly selected ensemble and ensemble predicted using N_{min}=3 (green, open circle) and *N*=100 (magenta, open circle) are shown as the references; **(D)** 1D distributions of three Euler angles predicted using N_{min}=3 (green) and *N*=100 (magenta). The distribution of three Euler angles in junction-topology allowed space is plotted as references; **(E)** 2D correlations of three Euler angles predicted using N_{min}=3; **(F)** 2D correlations of three Euler angles predicted using N_{min}=3; **(G)** leave-out cross-validation of ensemble predicted using N_{min}=3 (green) and *N*=100 (magenta); **(H)** correlation between measured RDCs of non-elongated HIV 1 TAR and corresponding calculated RDCs calculated from ensemble predicted using N_{min}=3 (green) and *N*=100 (magenta).

Argininamide bound HIV-1 TAR

Prior studies have shown that the amino acid argininamide (ARG) can be used as a ligand mimic of TAR's cognate protein target, the transactivator protein Tat. Prior NMR studies have shown that ARG bind to TAR, arrests inter-helical motions stabilizing a coaxial inter-helical conformation[12,28,30,31]. We applied the SAS approach to the previously reported 19 RDCs on ARG bound EI TAR[12] and the results of the analysis show that RMSD has no significant change after N_{min}=2. Therefore SAS was run 400 times using N_{min}=2 and 8 times using *N*=100 to form 800-conformational ensembles for both predictions. The RDCs calculated from both ensembles were in great agreement with measured RDCs (RMSD=2.2 Hz). Different from free state of HIV 1 TAR, ensembles determined using both ensemble sizes are very similar and in particular both ensembles give the identical narrow distribution of β around β=-10^o (Figure 3.5A), the magnitude of which is consistent with previously determined experimental results (β=8^o). The internal generalized degree of order (ϑ_{int}), which ranges from 0 to 1 indicating largest and smallest inter-helical flexibility respectively, were calculated from both predicted ensembles and the results for both predicted ensembles are similar (ϑ_{int} =0.99) and highly consistent with experimentally determined result (ϑ_{int}=1.09). Compared to determined ensembles of free state HIV 1 TAR, ensembles of ARG-bound state of HIV 1 TAR adopt much narrower distributions in all three Euler angles, indicating the fact that ARG arrests a rigid and coaxial conformation of HIV 1

TAR. These results also mirror the benchmark analysis, which shows that the uncertainty in N is less significant for narrower ensembles or ensembles containing relatively less distinct conformations.

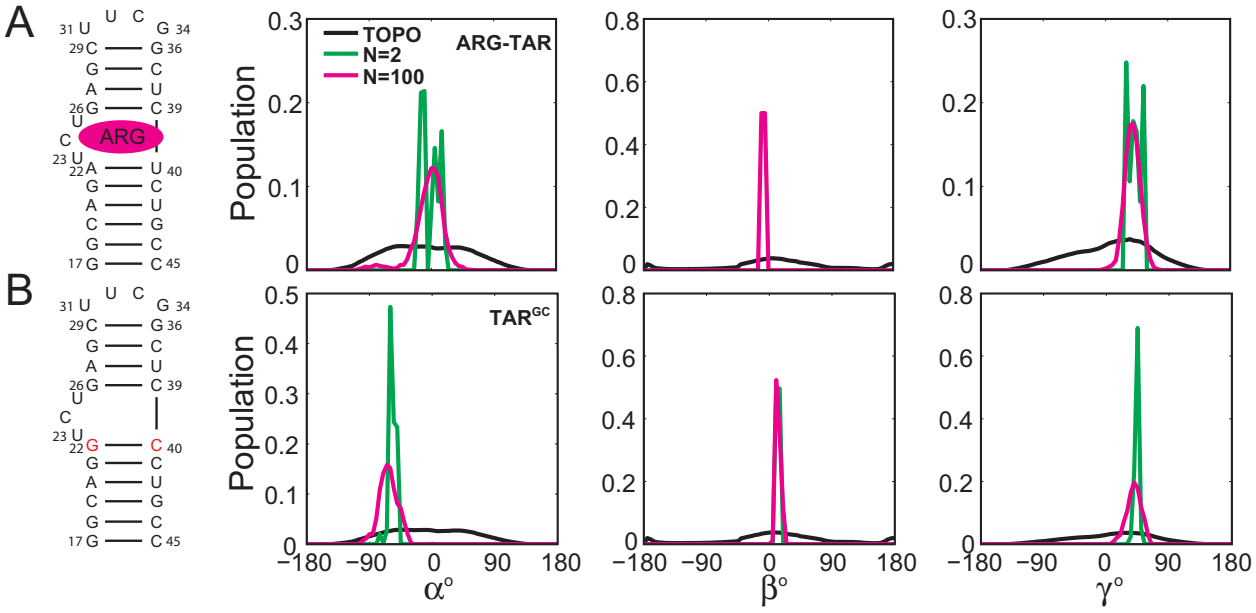

Figure 3.5 Secondary structure and 1D distributions of predicted ensembles of the three Euler angles of ARG bound and A22G-U40C HIV 1 TAR. (A) Argininamide (ARG) bound HIV 1 TAR; **(B)** A22G-U40C HIV 1 TAR. Each construct is predicted using N_{min}=2 (green) and arbitrarily chosen N=100 (magenta). The distributions of three Euler angels in the junction-topology allowed conformation space is plotted (black) as references.

A22G-U40C HIV-1 TAR

It has previously been shown that the flexible junction A22-U40 base pair of HIV 1 TAR plays important role activating inter-helical motions and that replacement of this base-pair with a more stable GC base pair results in an arrest of inter-helical motions and a coaxial conformation similar to that observed for the TAR-ARG complex. We therefore used our approach to determine the dynamic ensemble for the A22G-U40C mutant HIV-1 TAR (TAR^{GC}). The analysis of the 28 previously measured RDCs for TAR^{GC28} shows that N_{min}=2. Dynamic ensemble of TAR^{GC} was determined using N_{min}=2 and N=100, both of which yield the same

RDC RMSD = 4Hz. SAS approach was repeated for 400 and 8 times for N_{min}=2 and N=100 respectively to form 800-conformational ensembles. As expected, both ensembles reveal similar narrow distributions in all three Euler angles and in particular a very confined distribution of β around β=5^{o} (Figure 3.5B) that is consistent with experimental results (β=12^{o}). The inter-helical flexibility calculated from both predicted ensembles are also similar (ϑ_{int} =1.01) and consistent with experimentally determined result (ϑ_{int}=1.04), demonstrating that TAR^{GC} is highly stabilized by G22C40 base pair and adopts a rigid and coaxial inter-helical orientation that is similar to ARG-bound state of HIV 1 TAR.

3.4 Conclusions

In conclusion, we assessed the capability of RDCs as experimental constraints in ensemble determination by using the SAS approach utilizing rigid idealized A-form HJH model. As the rigid idealized A-form HJH can be described by simply three Euler rotation angles, it allows us to explicitly characterize the uncertainty of determined ensembles describing inter-helical orientation distributions using SAS approach and RDCs, which is prevented in studies that tried to characterize atomic-resolution ensembles due to astronomically large number of degrees of freedom. We demonstrated that experimental error and ensemble sizes used to determine dynamic ensembles are the two main factors that result in uncertainty in determined dynamic ensembles. The uncertainty arising from experimental error can be largely suppressed by lowering the experimental error; however uncertainty arising from ensemble size cannot be easily suppressed unless new experimental data are involved. Therefore the dynamic ensembles of RNA should be determined using different ensemble sizes and all the resulting ensembles that cannot be distinguished by current experimental data require further tests and validations using new experimental data in the future.

3.5 References

(1) Salmon, L.; Yang, S.; Al-Hashimi, H. M. *Annu. Rev. Phys. Chem.* **2013**.
(2) Clore, G. M.; Schwieters, C. D. *Biochemistry* **2004**, *43*, 10678.
(3) De Simone, A.; Richter, B.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 3810.
(4) Fisher, C. K.; Huang, A.; Stultz, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 14919.
(5) Salmon, L.; Bascom, G.; Andricioaei, I.; Al-Hashimi, H. M. *J. Am. Chem. Soc.* **2013**, *135*, 5457.
(6) Showalter, S. A.; Johnson, E.; Rance, M.; Bruschweiler, R. *J. Am. Chem. Soc.* **2007**, *129*, 14146.
(7) Jensen, M. R.; Markwick, P. R.; Meier, S.; Griesinger, C.; Zweckstetter, M.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169.
(8) Blackledge, M. J.; Bruschweiler, R.; Griesinger, C.; Schmidt, J. M.; Xu, P.; Ernst, R. R. *Biochemistry* **1993**, *32*, 10960.
(9) Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M. R.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 8407.
(10) Frank, A. T.; Horowitz, S.; Andricioaei, I.; Al-Hashimi, H. M. *J. Phys.Chem. B* **2013**.
(11) Sripakdeevong, P.; Cevec, M.; Chang, A. T.; Erat, M. C.; Ziegeler, M.; Zhao, Q.; Fox, G. E.; Gao, X.; Kennedy, S. D.; Kierzek, R.; Nikonowicz, E. P.; Schwalbe, H.; Sigel, R. K.; Turner, D. H.; Das, R. *Nat. Methods* **2014**.
(12) Zhang, Q.; Stelzer, A. C.; Fisher, C. K.; Al-Hashimi, H. M. *Nature* **2007**, *450*, 1263.
(13) Casiano-Negroni, A.; Sun, X.; Al-Hashimi, H. M. *Biochemistry* **2007**, *46*, 6525.
(14) Zhang, Q.; Al-Hashimi, H. M. *Nat. Methods* **2008**, *5*, 243.
(15) Fang, X.; Wang, J.; O'Carroll, I. P.; Mitchell, M.; Zuo, X.; Wang, Y.; Yu, P.; Liu, Y.; Rausch, J. W.; Dyba, M. A.; Kjems, J.; Schwieters, C. D.; Seifert, S.; Winans, R. E.; Watts, N. R.; Stahl, S. J.; Wingfield, P. T.; Byrd, R. A.; Le Grice, S. F.; Rein, A.; Wang, Y. X. *Cell* **2013**, *155*, 594.
(16) Shi, X.; Herschlag, D.; Harbury, P. A. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E1444.
(17) Lange, O. F.; Lakomek, N. A.; Fares, C.; Schroder, G. F.; Walter, K. F.; Becker, S.; Meiler, J.; Grubmuller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008**, *320*, 1471.
(18) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908.
(19) Frank, A. T.; Stelzer, A. C.; Al-Hashimi, H. M.; Andricioaei, I. *Nucleic Acids Res.* **2009**, *37*, 3670.
(20) Bailor, M. H.; Sun, X.; Al-Hashimi, H. M. *Science* **2010**, *327*, 202.
(21) Bailor, M. H.; Mustoe, A. M.; Brooks, C. L., 3rd; Al-Hashimi, H. M. *Nat. Protoc.* **2011**, *6*, 1536.
(22) Musselman, C.; Pitt, S. W.; Gulati, K.; Foster, L. L.; Andricioaei, I.; Al-Hashimi, H. M. *J. Biomol. NMR* **2006**, *36*, 235.
(23) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. *Biophys. J.* **2007**, *93*, 2300.
(24) Yang, S.; Salmon, L.; Al-Hashimi, H. M. *Nat. Methods* **2014**.
(25) Allison, J. R.; Varnai, P.; Dobson, C. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 18314.

(26) Markwick, P. R.; Bouvignies, G.; Salmon, L.; McCammon, J. A.; Nilges, M.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 16968.
(27) Mustoe, A. M.; Bailor, M. H.; Teixeira, R. M.; Brooks, C. L., 3rd; Al-Hashimi, H. M. *Nucleic Acids Res.* **2012**, *40*, 892.
(28) Stelzer, A. C.; Kratz, J. D.; Zhang, Q.; Al-Hashimi, H. M. *Angew. Chem., Int. Ed. Engl.* **2010**.
(29) Zhang, Q.; Sun, X.; Watt, E. D.; Al-Hashimi, H. M. *Science* **2006**, *311*, 653.
(30) Puglisi, J. D.; Tan, R.; Calnan, B. J.; Frankel, A. D.; Williamson, J. R. *Science* **1992**, *257*, 76.
(31) Aboul-ela, F.; Karn, J.; Varani, G. *Nucleic Acids Res.* **1996**, *24*, 3974.

Chapter 4

Preliminary NMR Study of Dynamics of exon splicing silencer 3 of HIV 1 RNA

4.1 Introduction

The human immunodeficientcy virus type 1 (HIV 1) requires balanced expression of nine regulatory proteins from the polycistronic RNA for replication[1]. The gene activity of HIV 1 is highly regulated during transcription initiation or posttranscriptional processing[2-4]. Successful production of virus requires well-regulated splicing pattern of HIV 1 RNA, which lead to excision of non-coding introns and ligation of coding exons from HIV 1 RNA[5]. Studies show that the splicing pattern of HIV 1 RNA is retained through a combination of non-conservative cores and splicing regulatory elements (SREs)[6,7], which activate or repress the splicing through either splicing enhancers or splicing silencers. Depending on the function of the sites, splicing enhancer and silencer can be categorized as exon splicing enhancer (ESE), exon splicing silencer (ESS) and intron splicing silencer (ISE)[8]. As a general mechanism, SREs arrest host factors, typically heterogeneous nuclear ribonucleoproteins (hnRNPs), which function as *trans* activators of splicing by stabilizing the components of the spliceosome at the non-consensus cores[6-8]. Previous studies have confirmed that hnRNP A1 protein can effectively attenuate the splicing activity at 3' splice site A2, A3 and A7 of HIV 1 RNA[9-14]. In particular, A7 site located at the terminal of 3' splice site of HIV 1 RNA where its activity is of central importance for regulating the excision of the intron and ligation of the exons. A complex of regulation elements including an ISS, ESS (ESS3) and ESE (ESE3) systematically establish the splicing pattern at A7 of HIV 1 RNA[15,16]. It has been shown that hnRNP A1 protein has the highest binding affinity to the ESS3 element of A7 splice site of HIV 1 RNA, which is a 25-nucleotide stem loop consisting of a 7-nucleotide apical loop and a 9-base-pair helix.

The three dimensional structure of ESS3 in solution has been determined by Tolbert and co-workers using NMR spectroscopy[1]. A key feature of determined ESS3 solution structure is a non-canonical AC wobble base pair disrupting the helix of ESS3, which is pH sensitive[17] and thereby can give rise to distinct dynamics of ESS3 under different pH conditions[1]. It is well known that under sufficiently low pH conditions, A^+C forms a wobble base pair in which adenine is protonated. However, at high pH, deprotonation of the adenine can break the hydrogen bond in A^+C base pair, distort the wobble base pair and them more like a internal loop. The purpose of this study is to use NMR RDCs to characterize the dynamic properties of ESS3 at different pH conditions as a starting point for applying ensemble-based screening approaches to search for small molecules that modulate HIV 1 RNA splicing.

4.2 Materials and Methods

4.2.1 Preparation of ESS3 sample

Uniformly 13C/15N labeled ESS3 samples for NMR studies was prepared by *in vitro* transcription utilizing T7 RNA polymerase (Takara Mirus Bio, Inc.), synthetic DNA templates with 5'-TTAATACGACTCACTATA-3' promoter (complementary promoter sequence was included in the complementary DNA sequence) (Integrated DNA Technologies, Inc.) and 15mL reaction volumes containing uniformly $^{13}C/^{15}N$ labeled ribonucleotide triphosphates (Cambridge Isotope Labs). Following the synthesis, ESS3 was purified on 20% denaturing polyacrylamide gel electrophoresis containing 8 M urea and 1XTBE (89mM Tris-borate, 89mM boric acid and 2mM EDTA), excised from the gel and electroeluted followed by overnight ethanol-precipitation. The RNA pellet was dissolved in water, annealed by heating to 95^oC for 5min, rapid cooling on ice and repeatedly exchanged into NMR buffer (15 mM sodium phosphate, 25 mM sodium chloride, 0.1 mM EDTA, and pH ~ 6.4) using a Centricon Ultra-4 concentrator (Millipore Corp.). The concentration of NMR sample (~1mM) was measured using Nanodrop 2000c spectrometer (Thermo Scientific Inc.). 10% D_2O was added into NMR sample after the measurement of NMR sample concentration.

4.2.2 NMR spectroscopy

All NMR experiments were performed on Agilent 600MHz NMR spectrometer equipped with a 5mm triple-resonance cryogenic probe at 298K. All NMR spectra were processed using NMRPipe/NMRDraw, analyzed using NMRDraw and overlaid using Sparky. Resonance assignments for ESS3 (pH=5.5) were obtained from previous study[1] and re-measured using non-constant time ^{1}H-^{13}C and ^{1}H-^{15}N HSQC experiments. The *J* couplings in nucleobase and sugar ^{1}H–^{13}C(^{15}N) were measured from the difference between the upfield and downfield components of the ^{1}H–^{13}C(^{15}N) doublet along the ^{1}H dimension using the narrow transverse relaxation-optimized spectroscopy (TROSY) component in the ^{13}C(^{15}N) dimension as implemented in 2D ^{1}H–^{13}C(^{15}N) S^{3}CT-heteronuclear single-quantum correlation (HSQC) experiments for sample dissolved in isotropic NMR buffer. The *J* coupling plus residual dipolar coupling (*D*) for each resonance were measured by repeating the measurements described above for sample dissolved in anisotropic NMR buffer with Pf1 phage (25mg/mL). The RDCs of each resonance were then calculated from the difference between the couplings measured in isotropic and anisotropic samples.

4.2.3 Analysis of measured RDCs

A in-house program RAMAH[18] was used to calculate the alignment tensor and corresponding predicted RDCs by fitting the measured RDCs to idealized A-form helices of ESS3 helix 1 and 2 (see Results and Discussion) and the Model 1 in the structural ensemble (PDB 2LDL) of ESS3 determined using NMR spectroscopy by Tolbert and co-workers[1]. The A-form helices were generated using 3DNA (version 2.1)[19] and the protons were added using REDUCE (version 3.23)[20]. SAS approach, as described in Chapter 1, was utilized to construct the dynamic ensembles of ESS3 using the measured RDCs. Two MD trajectories generated using AMBER 12 force field ff10 and Model 1 of the NMR structural ensemble of ESS3 as the starting coordinates were used as the conformation pool in determination of dynamic ensemble of ESS3: a 400ns trajectory without experimental constraints and a 535ns trajectory using previously measured NMR NOEs[1] as experimental constraints that contain 400 and 535 conformations respectively.

4.3 Results and Discussion

Prior NOE-based NMR studies of ESS3 indicated that at pH=5.5, A7 forms an $A7^{+}C21$ base pair while at higher pH = 6.5, this wobble base pair is deprotonated and this is coupled to the increased flexibility in the lower helix of ESS3 as measured by melting experiments[1]. For simplicity, we define the helix below A7C21 as helix 1 and the one above as helix 2 (Figure 4.1A). Here, we carried out detailed NMR chemical shift mapping and RDC experiments to more quantitatively characterize the structural dynamics of ESS3 and the impact of changing pH in altering dynamics of ESS3.

Chemical shift mapping experiments

We prepared uniformly $^{13}C/^{15}N$ labeled samples of ESS3 (Figure 4.1) and confirmed the assignments previously published through the measurements of NOESY and 2D HCN experiments. In Figure 4.1, we compare the 2D HSQC spectra of ESS3 at pH=5.5 and 6.5[1]. We observe the characteristic large changes in the chemical shift of A7-C2H2 upon lowering the pH that are consistent with protonation of adenine and formation of the $A7^{+}C21$ base pair (Figure 4.1). However, we also observed significant changes in chemical shifts for A7-C8H8, A7-C1'H1', C21-C5H5 and C21-C1'H1' (Figure 4.1). Interestingly, the chemical shift perturbations are not localized at the A7-C21 base pair. Rather, we also observe significant perturbations at neighboring residues U8 (C5H5 and N3H3) and G22 (C8H8 and N1H1), which likely reflect a conformation change that accompanies formation of the $A7^{+}C21$ wobble base pair and perturbations in the flexible apical loop adenines.

In the imino-HSQC spectrum, we observed two additional resonances in the guanine region at pH=6.5, indicating that new base pair(s) forms at pH=6.5 (Figure 4.1C). Because all the resonances of guanines in helix 1 and 2 are accounted for in the imino-HSQC spectrum, these new resonances must either reflect new new base pair(s) in the apical loop of ESS3 or a minor species involving helical guanine residues possibly involving an alternative secondary structure for the destabilized lower helix.

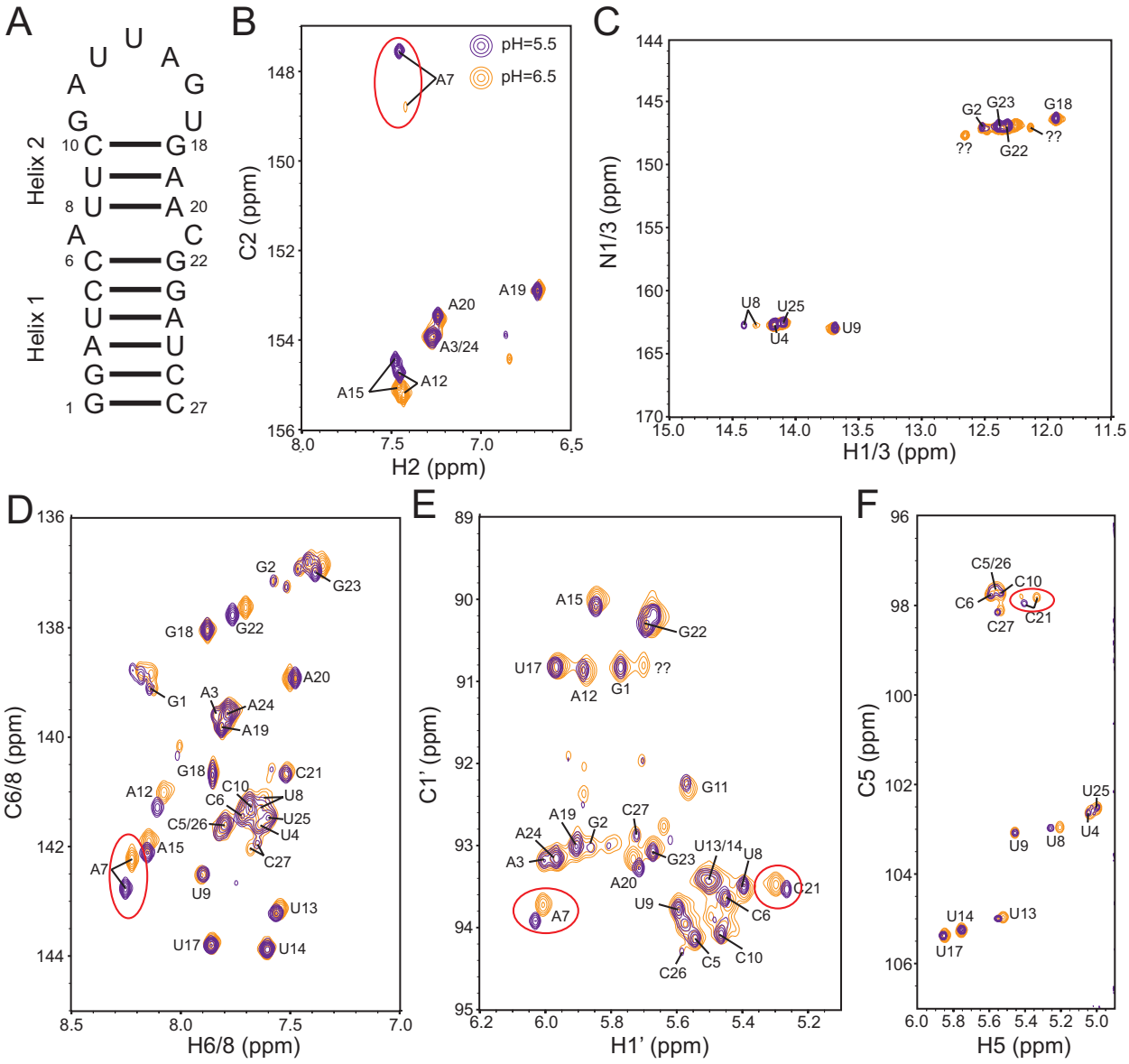

Figure 4.1 Secondary structure and resonance assignments of ESS3. (A) Secondary structure of ESS3; **(B)-(F)**: HSQC assignments of C2H2 **(B);** NH (imino) **(C)**; C6H6/C8H8 **(D)**; C1'H1' **(E)** and C5H5 **(F)** resonances. The questioning marks in **(C)** indicate the new guanine peaks in imino HSQC; the red circles indicate the significant changes of the chemical shift in A7 and C21.

Residue	pH=5.5	pH=6.5
G1(C8H8)	26.1	--

G1(C1'H1')	--	3.5
G2(C8H8)	14.1	--
G2(N1H1)	-1.9	--
U4(N3H3)	-5.5	-9.6
A7(C8H8)	38.1	32.8
A7(C2H2)	35.9	--
A7(C1'H1')	-39.6	-44.8
U8(C5H5)	22.8	24.4
U8(N3H3)	-12.8	--
U9(C6H6)	26.1	21.6
U9(C5H5)	21.5	24.0
U9(N3H3)	-9.7	-13.5
G11(C1'H1')	0.6	--
A12(C8H8)	11.0	9.2
A12(C2H2)	4.3	3.9
A12(C1'H1')	-12.0	-11.0
U13(C6H6)	--	4.8
U13(C5H5)	11.9	11.7
U14(C6H6)	6.3	5.4
U14(C5H5)	5.0	2.2
A15(C8H8)	13.1	11.1
A15(C2H2)	11.8	10.3
A15(C1'H1')	-1.5	-0.7
G16(C8H8)	15.6	14.3
U17(C6H6)	9.1	8.1
U17(C1'H1')	-7.4	-6.3
U17(C5H5)	-1.7	-1.1
G18(C8H8)	28.4	24.7
G18(N1H1)	--	-13.1
A19(C2H2)	28.0	24.4
A20(C8H8)	30.2	26.7
A20(C2H2)	30.9	26.7
C21(C6H6)	38.5	29.8
C21(C1'H1')	-8.1	-7.4
C21(C5H5)	22.1	--
G22(C8H8)	36.7	33.2
G23(C8H8)	40.9	34.2
G23(C1'H1')	--	-7.3
G23(N1H1)	-13.8	--

U25(N3H3)	--	-8.9

Table 4.1 RDCs of ESS3 measured at 25^{o}C under pH=5.5 and 6.5.

RDC measurements

To gain further insights into the structural dynamics of ESS3, we measured RDCs at pH = 5.5 and 6.5 using Pf1-phage as the ordering medium. In Figure 4.2, we compared the RDCs measured at the two pH conditions. We find very good agreement implying that formation of the A7^{+}C21 base pair does not significantly affect the global structure and dynamics of ESS3. Nevertheless, we do note that the RDCs measured for residues A7 and C21 differ by >2 Hz possibly indicating a pH-dependent change in structure and/or dynamics at this wobble base pair.

Examination of the measured RDCs at pH =5.5 and 6.5 as a function of the ESS3 secondary structure reveals that large RDCs are consistently observed for the long helix 1. Relative to helix 1, RDCs measured in helix 2 are slightly smaller in magnitude, and substantially smaller for the apical loop residues at both pH=5.5 and 6.5 (Table 4.1 and Figure 4.3). Surprisingly the magnitude of RDCs measured at A7 are inconsistently larger than counterparts measured in neighboring residues under both pH conditions, implying a unique local conformation of A7 in the A7-C21 base-pair. By contrast, the RDCs measured for C21 are consistent with counterparts measured in neighboring residues under both pH conditions, implying that C21 is less structurally perturbed relative to Watson-Crick geometry.

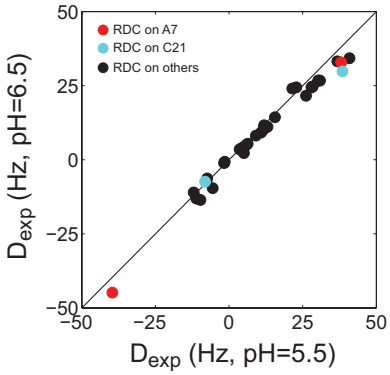

Figure 4.2 Correlation between RDCs measured under pH=5.5 and 6.5. Only the RDCs measured under both pH conditions are shown.

To gain further insights into how the AC wobble base pair modifies or perturbs the A-form helix of ESS3, we subjected all RDCs measured in the helix (excluding two terminal end base pairs) to an order tensor analysis utilizing an assumed idealized A-form helix. Prior studies have shown that the Watson-Crick base pairs surrounded by Watson-Crick base pairs can be accurately modeled using A-form helix geometry. Strikingly, for both pH conditions, we observe very poor correlation (RMSD = 5-8 Hz) between measured RDCs and values back-calculated using the best-fit order tensor implying that the $A7^{+}C21$ wobble base pair induces structure and/or dynamic perturbations to the ESS3 helix. We therefore repeated the fit excluding RDCs measured for base pairs surrounding the $A7^{+}C21$ base pair which deviate from A-form geometry due to the $A7^{+}C21$ base pair as suggested by chemical shift mapping data. This leads to much better agreement with RMSD = 2.0 and 1.7 Hz at pH=5.5 and 6.5 respectively. This suggests that the non-terminal helical base pairs largely adopt A-form geometry and the $A7^{+}C21$ wobble base pair induces perturbations to the A-form helix in a manner that is not strongly dependent upon pH. In principle, one can subject the upper and lower helix to independent order tensor analyses and thereby derive information about inter-helical structure and dynamics. However, severe overlap particular in the sugar moiety precluded measurement of sufficient RDCs for the short upper helix.

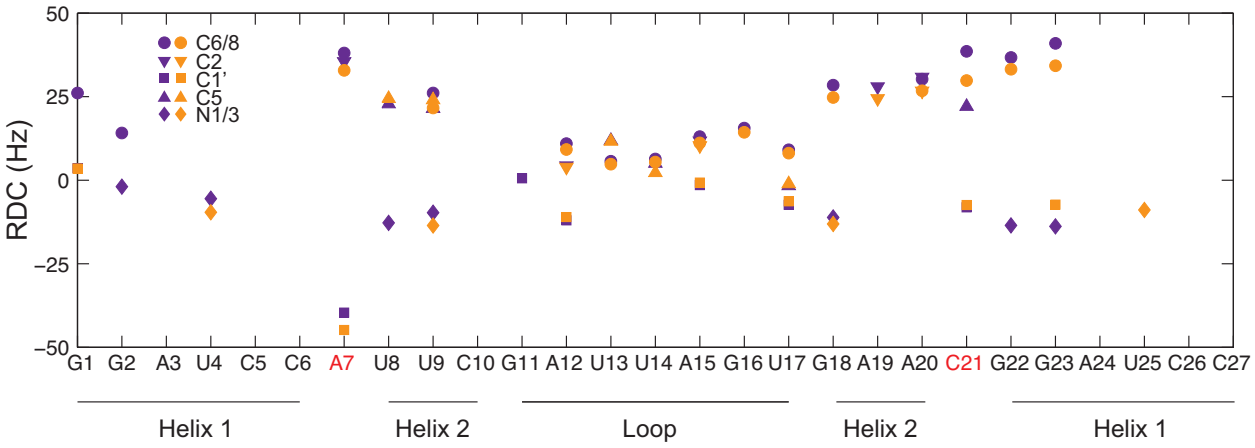

Figure 4.3 RDCs measured under pH=5.5 (purple) and 6.5 (orange) as a function of secondary structure of ESS3.

To gain insights into the structural dynamics of the apical loop, we also carried out an order tensor analysis by fitting the RDCs measured in the upper helix and apical loop to the available NMR structure. Not too surprisingly, we observed a poor fit (RMSD=12.0Hz and 10.6Hz for pH=5.5 and 6.5 respectively) most likely due to dynamics in the apical loop, which leads to attenuation of the measured RDCs. However, it was surprising that we obtained a poor fit to all RDCs against the NMR ESS3 structure even when excluding the apical loop (RMSD=8.7Hz and 10.0Hz for pH=5.5 and 6.5 respectively). These results suggest that there may be uncertainty in the NMR structure, including possibly dynamic contributions to the RDCs that need to be taken into account.

Constructing ensembles of ESS3

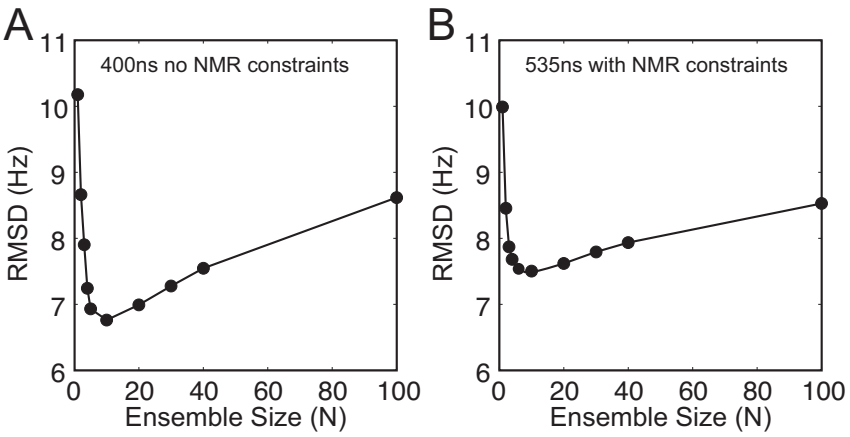

Figure 4.4 Ensemble size test for constructing population-weighted ensemble of ESS3. The r.m.s deviations between the measured and calculated RDCs averaged from ensembles constructed from **(A)** 400ns MD trajectory without experimental constraints and **(B)** 535ns MD trajectory using NMR NOEs as experimental constraints are plot as a function of ensemble sizes. Ensemble sizes *N*=1, 2, 3, 4, 5, 10, 20, 30, 40 and 100 are used in both tests.

To characterize the reason for the large deviation in the fitting of RDCs to NMR structure, we implemented a series of SAS calculations using varied ensemble sizes (*N* ranges from 1 to 100) to select the optimized ensembles of distinct conformations from the two MD trajectories that minimize the deviation between the measured and calculated RDCs. This simulation is only implemented for RDCs measured under pH=5.5 at which is NMR structure of ESS3 is determined. The results reveal that for both MD trajectories, *N*=1 gives very large RDC deviations (>10Hz) and N~10 gives the smallest deviation (Figure 4.4), indicating that ESS3 under pH=5.5 is nor certain rigid and cannot be described by a single static structure although the $A7^{+}C21$ wobble base pair is protonated and stabilized under this pH condition. However, the smallest RDC deviation (~7Hz, Figure 4.4) remained substantially larger than the RDC uncertainty. These results strongly suggest that the starting NOE-based NMR structure is inaccurate and highlight the importance of having an accurate average structure in the ensemble determination process.

4.4 Conclusion

We measured RDCs of ESS3 under pH=5.5 and 6.5 for characterizing its dynamics at both global and local levels. Our chemical shift mapping results clearly show the change of chemical shift of A7 and C21 involved in the pH-sensitive wobble base pair that is consistent with the prior study[1]. Great agreement was observed between the RDCs measured under pH=5.5 and 6.5, indicating that the structure of ESS3 does not change significantly upon the increase of the pH although the $A7^{+}C21$ wobble base pair is protonated and stabilized.

Fitting the measured RDCs to idealized A-form geometry of ESS3 suggests that the Watson-Crick base pairs surrounded by other Watson-Crick base pairs largely adopt the idealized A-form geometry; while, however, the $A7^{+}C21$ wobble base pair and its neighboring base pairs C6G22 and U8A20 deviate from idealized A-form geometry, which is consistent with the chemical shift mapping results. Fitting the measured RDCs to the NMR NOE-based structure of ESS3, either the whole structure of ESS3 including apical loop and all non-idealized base pairs or the idealized helical base pairs only, leads to very pool agreement between measured and back predicted RDCs, indicating that the determined NOE-based structure of ESS3 is inaccurate and needs to be refined using more experimental constraints e.g. both NOEs and RDCs.

We attempted to construct population-weighted dynamic ensemble of ESS3 using SAS approach and measured RDCs from MD trajectory generated with and without NMR NOEs. The results suggest that even under pH=5.5 in which case the $A7^{+}C21$ wobble base pair is protonated and stabilized, ESS3 is not certainly rigid and cannot be described by one single static structure. In contrast, a larger ensemble size (N~10) yield the smallest RDC RMSD which is however still substantially larger than experimental uncertainty of RDCs and therefore disallows accurate construction of dynamic ensembles of ESS3. This is likely due to that the NMR NOE-based structure of ESS3 is inaccurate and thereby results in incorrect sampling of the MD simulation. Hence, to accurately construct the dynamic ensemble of ESS3, the three dimensional solution structure of ESS3 has to be first refined using both NOEs and RDCs that provides a better starting structure of MD simulation. A new MD trajectory should then be generated using this

more accurate structure of ESS3 followed by the construction of dynamic ensemble using SAS approach and measured RDCs.

4.5 References

(1) Levengood, J. D.; Rollins, C.; Mishler, C. H.; Johnson, C. A.; Miner, G.; Rajan, P.; Znosko, B. M.; Tolbert, B. S. *J. Mol. Biol.* **2012**, *415*, 680.
(2) Cullen, B. R. *AIDS* **1995**, *9 Suppl A*, S19.
(3) Frankel, A. D.; Young, J. A. *Annu. Rev. Biochem.* **1998**, *67*, 1.
(4) McLaren, M.; Marsh, K.; Cochrane, A. *Front. Biosci.* **2008**, *13*, 5693.
(5) Cech, T. R. *Annu. Rev. Biochem.* **1990**, *59*, 543.
(6) Stoltzfus, C. M.; Madsen, J. M. *Curr. HIV Res.* **2006**, *4*, 43.
(7) Stoltzfus, C. M. *Adv. in Virus Res.* **2009**, *74*, 1.
(8) Saliou, J. M.; Bourgeois, C. F.; Ayadi-Ben Mena, L.; Ropers, D.; Jacquenet, S.; Marchand, V.; Stevenin, J.; Branlant, C. *Front. Biosci.* **2009**, *14*, 2714.
(9) Bilodeau, P. S.; Domsic, J. K.; Mayeda, A.; Krainer, A. R.; Stoltzfus, C. M. *J. Virol.* **2001**, *75*, 8487.
(10) Domsic, J. K.; Wang, Y.; Mayeda, A.; Krainer, A. R.; Stoltzfus, C. M. *Mol.Cell. Biol.* **2003**, *23*, 8762.
(11) Madsen, J. M.; Stoltzfus, C. M. *J. Virol.* **2005**, *79*, 10478.
(12) Hallay, H.; Locker, N.; Ayadi, L.; Ropers, D.; Guittet, E.; Branlant, C. *J. Biol. Chem.* **2006**, *281*, 37159.
(13) Zhu, J.; Mayeda, A.; Krainer, A. R. *Mol. Cell* **2001**, *8*, 1351.
(14) Asai, K.; Platt, C.; Cochrane, A. *Virology* **2003**, *314*, 229.
(15) Damgaard, C. K.; Tange, T. O.; Kjems, J. *RNA* **2002**, *8*, 1401.
(16) Marchand, V.; Mereau, A.; Jacquenet, S.; Thomas, D.; Mougin, A.; Gattoni, R.; Stevenin, J.; Branlant, C. *J. Mol. Biol.* **2002**, *323*, 629.
(17) Ravindranathan, S.; Butcher, S. E.; Feigon, J. *Biochemistry* **2000**, *39*, 16026.
(18) Hansen, A. L.; Al-Hashimi, H. M. *J. Magn. Reson.* **2006**, *179*, 299.
(19) Lu, X. J.; Olson, W. K. *Nat. Protoc.* **2008**, *3*, 1213.
(20) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol.* **1999**, *285*, 1735.

Chapter 5

Conclusions and Future Directions

5.1 Conclusions and future directions

Construction of population-weighted ensemble is an effective and feasible way to represent RNA dynamics[1-3]. The constructed ensemble of RNA not only provides a straightforward way to display the population distribution of various degrees of freedom of interest, but also provides a pool of structures that can be used for other studies such as RNA-small-molecule docking[4] or estimation of thermodynamic properties of RNA[5].

Successful construction of population-weighted ensemble of RNA requires at least three key factor[3]: (1) an efficient ensemble determination method; (2) properly selected experimental constraints that cover the timescale of the dynamics of interest; (3) a powerful method to evaluate the determined ensemble. This study utilizes the "sample and select" (SAS) approach[1,2,6], which has proven to be sufficiently efficient for construction of population-weighted ensembles for RNA. The conformation pools used in our SAS approach from which the ensemble is selected are generated using either molecular dynamics (MD) simulation or RNA junction-topology constraints. MD generated conformation pools provide both global and local structural details but may suffer sampling imperfection[2], which could disallow the native conformations to be selected in the SAS approach; junction-topology allowed space overcomes the sampling imperfection in MD generated conformation pool, samples each conformation in an equally-weighted and thereby unbiased manner and on average makes up only a small portion of all possible conformations (<10%), but so far it can only provide global structural information (e.g. inter-helical orientation distribution)[7-9]. Therefore the selection of conformation pool highly depends on the degrees of freedom that are of interest and sampling imperfection should be carefully tested no matter which conformation pool is used[3]. RDCs were incorporated in the SAS

approach as the experimental constraints for construction of dynamic ensembles of RNA not only because RDCs can capture dynamics of a broad timescale[10] but also due to the fact that RDCs are particularly sensitive to both inter-helical orientations[11-13] and local dihedral angles[2,14,15], which are the focus of this study.

However, to evaluate the predicted ensemble, although multiple metrics or methods have been developed including cross-validation[16], Jensen-Shannon divergence[17,18] and *S*-score[19-21], they either cannot distinguish degenerate ensembles that reproduce RDCs yielding similar RMSD or cannot fully capture the structural similarities between ensembles. In Chapter 2, a new metric REsemble[22] is developed, which can effectively capture the structural similarity between two ensembles by comparing the histogram distributions of specific degrees of freedom at systematically varied bin sizes. Different from conventional RMSD used for comparing two single static structures or ensemble RMSD (eRMSD) for comparing two conformational ensembles, REsemble metric does not pursue a pair-wise comparison of coordinates of atoms in terms of a single-value RMSD, but carries out the explicit comparisons of the distributions of degrees of freedom that the experimental constraints are sensitive to. In this case, REsemble can in fact rigorously evaluate the accuracy of each specific degree of freedom from the predicted ensemble. These evaluations can collectively report the overall similarity between target and predicted conformational ensembles and thereby the accuracy of the predicted dynamic ensemble. Although in this study, REsemble was used for comparing ensembles of orientation distributions only, it is generally applicable for comparing ensemble distributions of any degree of freedom.

By using REsemble metric developed in Chapter 2, we successfully characterized two main factors that result in uncertainties in determination of dynamic ensembles of RNA in Chapter 3: experimental error and ensemble size used for constructing dynamic ensembles. In particular, our results suggest that dynamic ensembles should be determined using different ensemble sizes, instead of only using smallest ensemble size that reproduce the experimental data within errors as commonly used in previous studies. Although the ensembles predicted using different ensemble sizes probably have similar features, they may encode distinct

information of thermodynamic or kinetic properties due to the distinct sampling schemes yielded from different ensemble sizes (e.g. small ensemble sizes give more discrete ensemble distributions while large ensemble sizes result in more continuous ensemble distributions). Thereby, it is important and necessary to further distinguish the ensembles predicted using different ensemble sizes using other experimental data. Although in Chapter 3, we focused on the analysis of a simplified helix-junction-helix (HJH) model of RNA[7-9], the conclusions drawn above are generally applicable to any biomolecules.

In Chapter 4, we carefully measured RDCs of ESS3 under pH=5.5 and 6.5 and attempted to determine the atomic-resolution dynamic ensemble of exon splicing silencer 3 (ESS3) of HIV 1 RNA using the SAS approach, the measured RDCs and different ensemble sizes as suggested in Chapter 3. However the fitting of measured RDCs to previously determined solution structure of ESS3[23] reveals very poor agreement likely due to the inaccuracies in the starting NOE-based NMR structure. Therefore, to more accurately determine the atomic-resolution dynamic ensemble of ESS3, refinement of the NMR structure of ESS3 using combined NOEs and RDCs has to be implemented and the resulting structure should be used for MD simulations to generate a better-sampled conformational pool for SAS analysis.

Future studies should focus on addressing remaining limitations in determination and evaluation of dynamic ensembles of RNA. Although we have focused on NMR RDCs, there are many other sources of data that can be used for determination of dynamic ensembles of biomolecules. For example, Al-Hashimi and co-workers have attempted to determine the dynamic ensemble of apical loop of HIV 1 TAR using chemical shifts (CS)[20]; Clore and co-worker reported determined dynamic ensembles of several proteins using paramagnetic relaxation enhancement data (PREs)[24,25]; Wang and Herschlag were able to construct dynamic ensemble of RNA and DNA using conventional and recently developed Au- small angle X-ray scattering data (SAXS) respectively[26-28]; and as a recent major breakthrough, Cheng and co-workers developed a new electron detection technique as well as a new image processing algorithm for Cryo-electron microscopy (Cryo-EM) method[29-31] that allow the determination of structures of small membrane proteins at near atomic-resolution, which can be potentially used to

determine dynamic ensembles of proteins and nucleic acids by simply "counting" flash frozen conformations. Although the biophysical techniques mentioned above have yielded many encouraging results, they were mostly used as the only experimental constraint in determination of dynamic ensembles, which can easily lead to degeneracies that are hard to eliminate. This is because a single type of experimental constraint is only sensitive to a specific type of degree of freedom or dynamics within narrow time scales, which could severely bias the selection of ensemble conformations. Hence combinatorial use of different biophysical techniques is critical for construction of comprehensive dynamic ensembles of biomolecules that can represent a variety of degrees of freedom as well as dynamics of broader time scales. Two straightforward combinations of experimental constraints that are worth attempting in the SAS approach are RDCs + CS and RDCs + Au-SAXS: the former combination allows the predicted ensembles to be sensitive to both global dynamics probed by RDCs and local dynamics probed by CS; the latter combination allows the predicted ensembles to be sensitive to both orientations probed by RDCs and translations probed by Au-SAXS data, which can dramatically enhance the spatial resolution of the predicted ensemble.

The REsemble approach for evaluating predicted ensembles can also be further improved especially for angular degrees of freedom (e.g. dihedral angles), which may encounter the boundary problem. The angles were binned from 0 to 360 degrees using systematically varied bin sizes in current study, however it should be noted that angles around the boundary (e.g. 0 and 360 degrees) could represent very similar or identical conformations that is indistinguishable in the current binning scheme. A possible way to overcome this boundary problem is to represent the distributions of angles using systematically varied origins (and thereby boundaries), apply REsemble to each representation of the distributions and adopt the representation that reports the highest similarity. A more straightforward way that can possibly overcome the boundary problem is to bin the angles in a circular manner if feasible, in which the boundaries are connected and thereby the similarity between these boundary values can be directly captured.

Another important future direction is to calculate the thermodynamic properties of RNA, especially conformational entropy, from determined population-weighted dynamic ensembles.

Although previous studies by Wand and co-workers have rigorously shown the relationship between order parameters and conformational entropy[32] that allows estimation of entropy of protein-protein recognition process, this method is still largely empirical and requires many theoretical assumptions. As the ensemble determined from our study is population-weighted, it is possible for us to develop a more rigorous method to calculate or estimate thermodynamic properties from the population distribution of conformations using information theory. Direct calculation of free energy from population-weighted ensembles is expected to be challenging, because the reference zero energy level of a RNA molecule is uncertain and it is also very difficult if not impossible to estimate the reference levels for different RNA molecules involved in one process or reaction. However calculation of the conformational entropy of each RNA molecule from the corresponding population-weighted ensembles using information theory is easier and more practical, because it does not involve the reference zero energy level. To validate the calculated entropy, it can be compared to results of isothermal titration calorimetry (ITC) experiments, which can simultaneously measure change of entropy, enthalpy and Gibbs free energy of a chemical reaction. Understanding conformational entropy of RNA in a predictable way from accurately determined population-weighted ensembles will greatly aid our understanding of how RNA functions especially for the processes that are hard to detect from experiments.

Finally, the solution structure of ESS3 has to be re-defined using both NOEs and RDCs. MD simulations should be carried out using the new solution structure of ESS3 as the starting coordinates. The same procedure for testing the accuracy of the structure of ESS3 and the sampling of MD trajectory as shown in Chapter 4 should be implemented before using the SAS approach to determine the population-weighted ensembles of ESS3. Thus far, RDCs were only measured under pH=5.5 and 6.5 and we plan to measure RDCs of ESS3 at pH=7.5 as well to probe the dynamic behavior of ESS3 at a pH value that is higher than the pKa of adenine in $A7^{+}C21$ wobble base pair, ensuring the deprotonation of A7 and thereby opening of $A7^{+}C21$ wobble base pair. This can be used as a control to clearly show the change of structure and dynamics of ESS3 from low to high pH conditions. Population-weighted dynamic ensembles of ESS3 under different pH conditions should be determined and compared to expose the change of

both inter-helical and local dynamics of ESS3. Specific focus should be on the local structure and dynamics of $A7^{+}C21$ wobble base pair and its neighboring base pairs upon change of pH conditions, which likely trigger the change of ESS3 dynamics. Using the determined dynamic ensembles of ESS3 as a structural pool, it will be possible to carry out *in silico* drug screening targeting at ESS3, which can potentially aid the development of HIV gene therapy in the future.

5.2 References

(1) Salmon, L.; Bascom, G.; Andricioaei, I.; Al-Hashimi, H. M. *J. Am. Chem. Soc.* **2013**, *135*, 5457.
(2) Frank, A. T.; Stelzer, A. C.; Al-Hashimi, H. M.; Andricioaei, I. *Nucleic Acids Res.* **2009**, *37*, 3670.
(3) Salmon, L.; Yang, S.; Al-Hashimi, H. M. *Annu. Rev. Phys. Chem.* **2013**.
(4) Stelzer, A. C.; Frank, A. T.; Kratz, J. D.; Swanson, M. D.; Gonzalez-Hernandez, M. J.; Lee, J.; Andricioaei, I.; Markovitz, D. M.; Al-Hashimi, H. M. *Nat. Chem. Biol.* **2011**, *7*, 553.
(5) Wand, A. J. *Curr. Opin. Struc. Biol.* **2013**, *23*, 75.
(6) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. *Biophys. J.* **2007**, *93*, 2300.
(7) Bailor, M. H.; Sun, X.; Al-Hashimi, H. M. *Science* **2010**, *327*, 202.
(8) Bailor, M. H.; Mustoe, A. M.; Brooks, C. L., 3rd; Al-Hashimi, H. M. *Nat. Protoc.* **2011**, *6*, 1536.
(9) Mustoe, A. M.; Bailor, M. H.; Teixeira, R. M.; Brooks, C. L., 3rd; Al-Hashimi, H. M. *Nucleic Acids Res.* **2012**, *40*, 892.
(10) Bothe, J. R.; Nikolova, E. N.; Eichhorn, C. D.; Chugh, J.; Hansen, A. L.; Al-Hashimi, H. M. *Nat. Methods* **2011**, *8*, 919.
(11) Zhang, Q.; Stelzer, A. C.; Fisher, C. K.; Al-Hashimi, H. M. *Nature* **2007**, *450*, 1263.
(12) Fisher, C. K.; Zhang, Q.; Stelzer, A.; Al-Hashimi, H. M. *J. Phys. Chem. B* **2008**, *112*, 16815.
(13) Zhang, Q.; Al-Hashimi, H. M. *Nat. Methods* **2008**, *5*, 243.
(14) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 9279.
(15) Jensen, M. R.; Markwick, P. R.; Meier, S.; Griesinger, C.; Zweckstetter, M.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169.
(16) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908.
(17) Lindorff-Larsen, K.; Ferkinghoff-Borg, J. *PLoS One* **2009**, *4*, e4203.
(18) Fisher, C. K.; Huang, A.; Stultz, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 14919.
(19) De Simone, A.; Richter, B.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 3810.

(20) Frank, A. T.; Horowitz, S.; Andricioaei, I.; Al-Hashimi, H. M. *J. Phys. Chem. B* **2013**.
(21) Camilloni, C.; Robustelli, P.; De Simone, A.; Cavalli, A.; Vendruscolo, M. *J. Am. Chem. Soc.* **2012**, *134*, 3968.
(22) Yang, S.; Salmon, L.; Al-Hashimi, H. M. *Nat. Methods* **2014**.
(23) Levengood, J. D.; Rollins, C.; Mishler, C. H.; Johnson, C. A.; Miner, G.; Rajan, P.; Znosko, B. M.; Tolbert, B. S. *J. Mol. Biol.* **2012**, *415*, 680.
(24) Tang, C.; Schwieters, C. D.; Clore, G. M. *Nature* **2007**, *449*, 1078.
(25) Iwahara, J.; Clore, G. M. *Nature* **2006**, *440*, 1227.
(26) Fang, X.; Wang, J.; O'Carroll, I. P.; Mitchell, M.; Zuo, X.; Wang, Y.; Yu, P.; Liu, Y.; Rausch, J. W.; Dyba, M. A.; Kjems, J.; Schwieters, C. D.; Seifert, S.; Winans, R. E.; Watts, N. R.; Stahl, S. J.; Wingfield, P. T.; Byrd, R. A.; Le Grice, S. F.; Rein, A.; Wang, Y. X. *Cell* **2013**, *155*, 594.
(27) Shi, X.; Beauchamp, K. A.; Harbury, P. B.; Herschlag, D. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, E1473.
(28) Shi, X.; Herschlag, D.; Harbury, P. A. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E1444.
(29) Cao, E.; Liao, M.; Cheng, Y.; Julius, D. *Nature* **2013**, *504*, 113.
(30) Liao, M.; Cao, E.; Julius, D.; Cheng, Y. *Nature* **2013**, *504*, 107.
(31) Li, X.; Mooney, P.; Zheng, S.; Booth, C. R.; Braunfeld, M. B.; Gubbens, S.; Agard, D. A.; Cheng, Y. *Nat. Methods* **2013**, *10*, 584.
(32) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J. *Nature* **2007**, *448*, 325.

Appendix 1

Sample and Select (SAS) Approach

```
%{

USAGE EXPLANATION

This is the main program for SAS.
pool: the structural pool
rdc_pool: the rdc pool corresponding to the structural pool
rdc_ave: input rdcs

%}
function run_run_RDC_fitting_Szz_off(pool,rdc_pool,rdc_ave)



   simu_test=input('Is this a simulation test? y/n [y]   ','s');
% input the type of calculation, either simulation test (y) or experimental determination (n)
   if isempty(simu_test)

      simu_test='y';

   end

   if simu_test=='y'   % Simulation test

      ens_size=input('What is the ensemble size (N)?   ');
      num_cycles=input('How many cycles will be run?   ');
      error_sigma=input('What is the RDC error (in Hz)?   ');

      disp('Simulation starts');

      info=run_RDC_fitting_Szz_off_flex_algn(pool,rdc_pool,ens_size,num_cycles,rdc_ave,error_sigma);
% implementation of SAS and retrieve the information from SAS
      ens_pred=info.ens;    % ens_pred is the predicted ensemble
      ens_serial_num=info.serial_num;
% ens_serial_num list the serial number of each conformation in ens_pred

      disp('Simulation done!');

      save output_ensemble.txt ens_pred -ASCII -TABS
% save predicted ensemble; output_ensemble.txt is the name of the saved predicted ensemble that upon change by
users
```

```
        save output_ensemble_serial_number.txt ens_serial_num -ASCII -TABS
% save serial number of each conformation; output_ensemble_serial_number.txt is the name of the saved serial
numbers that upon change by users

    else

        if simu_test=='n'  % Experimental determination

            ens_size=input('What is the ensemble size (N)?   ');
            num_cycles=input('How many cycles will be run?   ');

            disp('Prediction starts');

            info=run_RDC_fitting_Szz_off(pool,rdc_pool,ens_size,num_cycles,rdc_ave);
% implementation of SAS and retrieve the information from SAS
            ens_pred=info.ens;    % ens_pred is the predicted ensemble
            ens_serial_num=info.ens_serial_num;
% ens_serial_num list the serial number of each conformation in ens_pred

            disp('Prediction done!');

            save output_ensemble.txt ens_pred -ASCII -TABS
% save predicted ensemble; output_ensemble.txt is the name of the saved predicted ensemble that upon change by
users
            save output_ensemble_serial_number.txt ens_serial_num -ASCII -TABS
% save serial number of each conformation; output_ensemble_serial_number.txt is the name of the saved serial
numbers that upon change by users

        else

            disp('Calculation must be either simulation test (simu_test=y) or real prediction (simu_test=n)!');
% calculation must be either simulation test or experimental determination, otherwise error is reported

        end

    end

end
```

```
%{

USAGE EXPLANATION

pool: the structural pool
rdc_pool: the rdcs for each conformation in the entire pool
ens_size: number of conformation in each selected subset (N)
num_cycles: iterations the SAS will be run
rdc_noerr: averaged rdcs without assigning any uncertainty
error_sigma: standard deviation of distribution in error_file


rdcs are differetially error corrupted for each SAS iteration

%}



function info=run_RDC_fitting_Szz_off_flex_algn(pool,rdc_pool,ens_size,num_cycle,rdc_noerr,error_sigma)

   num_rdc=length(rdc_pool(:,1)); % number of rdcs for each conformation
   ens=zeros(1,length(pool(1,:))); % initialization of predicted ensemble
   serial_num=0; % initialization of serial number of conformations

   %-----assign error array from which the error values will be selected ------%

   error_file=normrnd(0,error_sigma,10000,1);
   [mu sigma]=normfit(error_file);

   while (abs(mu)>0.01)||(abs(sigma-error_sigma)>0.05)

      error_file=normrnd(0,error_sigma,10000,1);
      [mu sigma]=normfit(error_file);

   end

   %----error array is constructed----%


   %---- this for loop goes over all SAS iterations ----%
   for i=1:1:num_cycle

      rdc_err=err_corruption_flex_algn(rdc_noerr,num_rdc,error_file,error_sigma); % assign RDC error
      RMSD_threshold=sqrt(mean((rdc_noerr-rdc_err).^2)); % real rdc RMSD
      temp=RDC_fitting_Szz_off_flex_algn(pool,rdc_pool,ens_size,rdc_err);
% implementation of SAS and retrieve the information from SAS
      ens_temp=temp.ens; % predicted ensemble
      ens_rmsd=temp.RMSD; % rdc RMSD from selected ensemble subset
      ens_serial_num=temp.selected_states; % serial number of each conformation in ensemble
```

```
      %---- determine whether RMSD is smaller than or equal to real RMSD; 0.001 is added to real RMSD to
prevent digital error from MATLAB ----%
      if ens_rmsd<=RMSD_threshold+0.001

         ens=[ens;ens_temp];
         serial_num=[serial_num;ens_serial_num];

      end

      if mod(i,10)==0

         disp(i/10);

      end

   end

   ens(1,:)=[]; % finalization of predicted ensemble
   serial_num(1)=[]; % finalization of serial number of conformations

   %---- info is the output structure containing both predicted ensemble and serial numbers ----%
   info.ens=ens;
   info.serial_num=serial_num;

end
```

```
%{

USAGE EXPLANATION

pool: the structural pool
rdc_pool: the rdc pool corresponding to the structural pool
ens_size: number of conformation in each selected subset (N)
num_cycle: iterations the SAS will be run
rdc_ens: input rdcs
%}

function info=run_RDC_fitting_Szz_off(pool,rdc_pool,ens_size,num_cycle,rdc_ens)

   ens=zeros(1,3);  % initialization of predicted ensemble
   ens_serial_num=0; % initialization of serial number of conformations
   num_rdc=length(rdc_pool(:,1)); % number of rdcs for each conformation

   %---- this for loop goes over all SAS iterations ----%
   for i=1:1:num_cycle

      temp=RDC_fitting_Szz_off(pool,rdc_pool,ens_size,rdc_ens);
% implementation of SAS and retrieve the information from SAS
      ens_temp=temp.ens; % predicted ensemble
      ens_serialnum=temp.selected_states; % serial number of each conformation in ensemble
      ens=[ens;ens_temp];
      ens_serial_num=[ens_serial_num;ens_serialnum];

      if mod(i,10)==0

         disp(i/10);

      end

   end

   ens(1,:)=[];  % finalization of predicted ensemble
   ens_serial_num(1)=[]; % finalization of serial number of conformations

   %---- info is the output structure containing both predicted ensemble and serial numbers ----%
   info.ens=ens;
   info.ens_serial_num=ens_serial_num;

end
```

```
%{

USAGE EXPLANATION

pool: the structural pool
rdc_pool: the rdcs for each conformation in the entire pool
ens_size: number of conformation in each selected subset (N)
rdc_ens: input rdcs

%}

function info=RDC_fitting_Szz_off_flex_algn(pool,rdc_pool,ens_size,rdc_ens)

    tic;
    num_bond_vectors=length(rdc_pool(:,1)); % number of rdcs (bond_vectors) in input rdc vector
    total_snapshots=length(pool(:,1)); % number of conformations in structural pool
    ensemble_size=ens_size; % ensemble size
    T_effective_intial=1; % initial temperature of Simulated Annealing scheme
    MC_steps=2000; % Monte Carlo steps implemented at each effective temperature
    number_T_increments=100; % number of effective temperatures will be used in SA scheme
    C_pro=0.90; % scaling factor of effective temperature

    selected_array_old=floor(total_snapshots*rand(ensemble_size,1))+1; % randomly select initial ensemble subset

    D_ij_exp=rdc_ens;% input rdcs

    pre_D_ij_cal=zeros(num_bond_vectors,1);

    %---- this for loop calculate the rdcs of the initial ensemble subset ----%
    for jjj=1:1:ensemble_size

        selected_snapshot=selected_array_old(jjj);
        pre_D_ij_cal=pre_D_ij_cal+rdc_pool(:,selected_snapshot);

    end

    D_ij_cal=pre_D_ij_cal/ensemble_size;

    %---- calculate the panelty function between input and calculated rdcs ----%
    X2_old=sum((D_ij_cal-D_ij_exp).^2);
    X2_old=X2_old/(num_bond_vectors);

    T_effective=T_effective_intial;

    %---- the two-fold for loop implement SAS; mm loop goes over effective temperatures; m loop goes over MC
steps ----%

    for mm=1:1:number_T_increments
```

```
    for m=1:1:MC_steps

        %---- randomly select a snapshot in the ensemble to be changed ----%

        rand_int=floor(ensemble_size*rand())+1;
        old_snapshot=selected_array_old(rand_int);

        k=1;

        %---- this while loop replace one conformation in selected
        %subset by one conformation in the rest of the pool ----%

        while k==1

          rand_int_2=floor(total_snapshots*rand())+1; % randomly select a snapshot to replace

          %---- check whether the selected ensemble is already in ensemble or not ----%

          k=0;

          for j=1:1:ensemble_size

             random_test=selected_array_old(j);
             if rand_int_2==random_test

               k=1;
               break;

             end

          end

        end

        new_snapshot=rand_int_2;

        pre_D_ij_cal=pre_D_ij_cal-rdc_pool(:,old_snapshot)+rdc_pool(:,new_snapshot);
% calculate rdcs of new ensemble

        D_ij_cal_new=pre_D_ij_cal/ensemble_size;


        X2_new=sum((D_ij_cal_new-D_ij_exp).^2); % new scaling factor
        X2_new=X2_new/(num_bond_vectors);

        r=rand(); % random number between 0 and 1

        if X2_new<=X2_old % accept move if X2_new<X2_old
```

```
            X2_old=X2_new;

            selected_array_old(rand_int)=new_snapshot;

            D_ij_cal=D_ij_cal_new;

        else

            if exp((X2_old-X2_new)/T_effective)>r % if X2_new>X2_old, accept if MC_probability<r

               X2_old=X2_new;

               selected_array_old(rand_int)=new_snapshot;

               D_ij_cal=D_ij_cal_new;

            else

               pre_D_ij_cal=pre_D_ij_cal+rdc_pool(:,old_snapshot)-rdc_pool(:,new_snapshot);
% reject the move and resume the calculated rdcs

            end

        end

    end

    T_effective=T_effective*C_pro; % change the effective temperature by multiplying scaling factor to the current
temperature

    if mod(mm,10)==0

        fprintf('%d',mm/10);
        fprintf(1,'  ');

    end
  end
  disp('\n');


  selected_states=pool(selected_array_old,:);

  disp('X2= ');
  RMSD=sqrt(X2_old);
  disp(RMSD); % display rdc RMSD

  %---- display the ensemble when N is smaller than or equal to 10 ----%
```

```
    if ens_size<=10

        disp(selected_states);

    end

    %---- retrive the information of the predicted ensemble ----%
    info.RMSD=RMSD;
    info.ens=selected_states;
    info.selected_states=selected_array_old;
    info.D_ij_cal=D_ij_cal;

    toc;

end




%{

USAGE EXPLANATION

pool: the structural pool
rdc_pool: the rdcs for each conformation in the entire pool
ens_size: number of conformation in each selected subset (N)
rdc_ens: input rdcs

%}

function info=RDC_fitting_Szz_off(pool,rdc_pool,ens_size,rdc_ens)

    tic;

    num_bond_vectors=length(rdc_ens(:,1)); % number of rdcs (bond_vectors) in input rdc vector
    total_snapshots=length(pool(:,1));  % number of conformations in structural pool
    ensemble_size=ens_size; % ensemble size
    T_effective_intial=1; % initial temperature of Simulated Annealing scheme
    MC_steps=2000; % Monte Carlo steps implemented at each effective temperature
    number_T_increments=100; % number of effective temperatures will be used in SA scheme
    C_pro=0.90; % scaling factor of effective temperature

    D_ij_exp=rdc_ens; % input rdcs

    D_ij_pool=rdc_pool; % rdc pool
```

```
selected_array_old=floor(total_snapshots*rand(ensemble_size,1))+1; % randomly select initial ensemble subset

pre_D_ij_cal=zeros(num_bond_vectors,1);

%---- this for loop calculate the rdcs of the initial ensemble subset ----%
for jjj=1:1:ensemble_size

   selected_snapshot=selected_array_old(jjj);
   pre_D_ij_cal=pre_D_ij_cal+D_ij_pool(:,selected_snapshot);

end

D_ij_cal=pre_D_ij_cal/ensemble_size;

L_old=sum(D_ij_exp.*D_ij_cal)/sum(D_ij_cal.^2);   % scaling factor between input and calculated rdcs

%---- calculate the penalty function between input and calculated rdcs ----%
X2_old=sum((D_ij_cal*L_old-D_ij_exp).^2);
X2_old=X2_old/(num_bond_vectors);



T_effective=T_effective_intial;

%---- the two-fold for loop implement SAS; mm loop goes over effective temperatures; m loop goes over MC
steps ----%

for mm=1:1:number_T_increments

  for m=1:1:MC_steps

     %---- randomly select a snapshot in the ensemble to be changed ----%

     rand_int=floor(ensemble_size*rand())+1;
     old_snapshot=selected_array_old(rand_int);

     k=1;

     %---- this while loop replace one conformation in selected
     %subset by one conformation in the rest of the pool ----%

     while k==1

       rand_int_2=floor(total_snapshots*rand())+1; % randomly select a snapshot to replace

       %---- check whether the selected ensemble is already in ensemble or not ----%

       k=0;
```

```
        for j=1:1:ensemble_size

            random_test=selected_array_old(j);
            if rand_int_2==random_test

              k=1;
              break;

            end

        end

    end

    new_snapshot=rand_int_2;

    pre_D_ij_cal=pre_D_ij_cal-D_ij_pool(:,old_snapshot)+D_ij_pool(:,new_snapshot);
% calculate rdcs of new ensemble
    D_ij_cal_new=pre_D_ij_cal/ensemble_size;

    L_new=sum(D_ij_exp.*D_ij_cal_new)/sum(D_ij_cal_new.^2);   % new scaling factor

    X2_new=sum((D_ij_cal_new*L_new-D_ij_exp).^2); % new penalty function
    X2_new=X2_new/(num_bond_vectors);


    r=rand(); % random number between 0 and 1


    if X2_new<=X2_old % accept move if X2_new<X2_old

        X2_old=X2_new;
        selected_array_old(rand_int)=new_snapshot;
        D_ij_cal=D_ij_cal_new;
        L_old=L_new;

    else

        if exp((X2_old-X2_new)/T_effective)>r % if X2_new>X2_old, accept if MC_probability<r

            X2_old=X2_new;

            selected_array_old(rand_int)=new_snapshot;

            D_ij_cal=D_ij_cal_new;
            L_old=L_new;
```

```
            else

                pre_D_ij_cal=pre_D_ij_cal+D_ij_pool(:,old_snapshot)-D_ij_pool(:,new_snapshot); % reject the move and resume the calculated rdcs

            end

        end

    end

    T_effective=T_effective*C_pro; % change the effective temperature by multiplying scaling factor to the current temperature

    if mod(mm,10)==0

        fprintf('%d',mm/10);
        fprintf(1,'  ');

    end

end

disp('\n');

selected_states=pool(selected_array_old,:);
disp('X2= ');
RMSD=sqrt(X2_old); % display rdc RMSD
disp(RMSD);

%---- display the ensemble when N is smaller than or equal to 10 ----%
if ens_size<=10

     disp(selected_states);

end

%---- retrive the information of the predicted ensemble ----%
info.RMSD=RMSD;
info.ens=selected_states;
info.selected_states=selected_array_old;
info.D_ij_cal=D_ij_cal;
info.L=L_old;

toc;

end
```

```
%{

USAGE EXPLANATION

This program is for assigning uncertainty to input rdcs

rdcs_exp: input averaged rdcs
num_rdcs: number of rdcs
err_E1: input error pool
sigma: standard deviation of distribution in error_file


rdcs are differentially error corrupted for each SAS iteration

%}

function rdcs_exp_err=err_corruption_flex_algn(rdcs_exp,num_rdc,err_E1,sigma)

   num_bv_E1=num_rdc;

   num_err_E1=length(err_E1);


   error=zeros(num_bv_E1,1);  % initialization of assigned error


   error_E1=err_E1(floor(num_err_E1*rand(num_bv_E1,1))+1); % assign uncertainty

   %---- this while loop determines the accuracy of the assigned
   %uncertainty; generated uncertainty must have the standard deviation
   %differing from sigma by less than 0.05 Hz
   while std(error_E1,1)<(sigma-0.05)||std(error_E1,1)>(sigma+0.05)

      error_E1=err_E1(floor(num_err_E1*rand(num_bv_E1,1))+1);

   end

   error=[error error_E1];

   error(:,1)=[]; % finalization of assigned uncertainty

   rdcs_exp_err=rdcs_exp+error; % assign uncertainty to input rdcs
end
```

Appendix 2

REsemble Algorithm for Measuring Ensemble Similarity

```
%{

USAGE EXPLANATION

Below are REsemble algorithm for inter-helical orientaion
This is main program for REsemble algorithm for inter-helical orientation
ens_ref_0: reference ensemble of interhelical orientation
ens_pred_0: predicted ensemble of interhelical orientation

%}

function reso_record=resolution_calc(ens_ref_0,ens_pred_0)
   tic;
   reso_candid=[5 10 15 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180]; % bin width used for
binning
   length_reso_candid=length(reso_candid); % number of bin widths that will be used
   reso_index=1;
   current_JSD=2;
   ens_ref=bin_ensemble(ens_ref_0); % pre-bin the ensemble to the grids
   ens_pred=bin_ensemble(ens_pred_0);

   threshold_JSD=0;
   reso_record=[reso_candid' zeros(length_reso_candid,1)]; % initialization

   s=1;
   %---- this while loop goes over all bin widths until JSD is smaller than the threshold ----%
   while current_JSD>threshold_JSD

      current_reso=reso_candid(reso_index);

      current_JSD=ens_sax_jsd_bin(ens_ref,ens_pred,5,current_reso);
      reso_record(s,2)=current_JSD;
      reso_index=reso_index+1;
      if current_reso==180

         break;
```

```
        end
        s=s+1;

    end

    toc;

end


%{

USAGE EXPLANATION

ens_1, ens_2: two interhelical ensembles upon comparison
bd: smallest bin width used; the binning grid will be gone over using this bin width
sax_bd: bin width used for binning according to single axis rotation amplitude

%}

function JSD=ens_sax_jsd_bin(ens_1,ens_2,bd,sax_bd)
% this program will calculate JSD between ensemble 1 and 2 using bin width sax_bd in term of single axis rotation

    bin_width=bd;
    sax_bin_width=sax_bd;
    L_ens_1=length(ens_1(:,1));
    L_ens_2=length(ens_2(:,1));

    a_min=min([min(ens_1(:,1)),min(ens_2(:,1))]);
    a_max=max([max(ens_1(:,1)),max(ens_2(:,1))]);
    b_min=min([min(ens_1(:,2)),min(ens_2(:,2))]);
    b_max=max([max(ens_1(:,2)),max(ens_2(:,2))]);
    g_min=min([min(ens_1(:,3)),min(ens_2(:,3))]);
    g_max=max([max(ens_1(:,3)),max(ens_2(:,3))]);

    n_a=floor((a_max-a_min)/bin_width)+1;
    n_b=floor((b_max-b_min)/bin_width)+1;
    n_g=floor((g_max-g_min)/bin_width)+1;

    a_max=a_min+bin_width*n_a;
    b_max=b_min+bin_width*n_b;
    g_max=g_min+bin_width*n_g;

    a=a_min:bin_width:a_max;
    b=b_min:bin_width:b_max;
    g=g_min:bin_width:g_max;

    n_states=n_a*n_b*n_g;
    population=zeros(2,n_states);
```

```
j=1;
for i_a=1:1:n_a

   if isempty(ens_1)&&isempty(ens_2)

           break;

   else

      for i_b=1:1:n_b

         if isempty(ens_1)&&isempty(ens_2)

            break;

         else

            for i_g=1:1:n_g

               if isempty(ens_1)&&isempty(ens_2)

                  break;

               else

                  pre_p_1=strmatch([a(i_a) b(i_b) g(i_g)],ens_1);
                  p_temp_1=length(pre_p_1);

                  pre_p_2=strmatch([a(i_a) b(i_b) g(i_g)],ens_2);
                  p_temp_2=length(pre_p_2);

                  if p_temp_1==0&&p_temp_2==0

                     continue;

                  else

                     pre_p_1=sax_ens([a(i_a) b(i_b) g(i_g)],ens_1,sax_bin_width);
                  %all conformers within single axis bin width are included
                     pre_p_1_temp=length(pre_p_1);
                     p_temp_1=pre_p_1_temp;

                     ens_1(pre_p_1,:)=[]; % any counted conformers are deleted

                     pre_p_2=sax_ens([a(i_a) b(i_b) g(i_g)],ens_2,sax_bin_width);
                  %all conformers within single axis bin width are included
                     pre_p_2_temp=length(pre_p_2);
                     p_temp_2=pre_p_2_temp;
```

```
                    ens_2(pre_p_2,:)=[];   % any counted conformers are deleted

                    population(1,j)=p_temp_1;
                    population(2,j)=p_temp_2;

                end

                j=j+1;

            end

        end

    end

end

%---- calculation of JSD ----%
population(1,:)=population(1,:)/L_ens_1;
population(2,:)=population(2,:)/L_ens_2;

P_1=population(1,:);
P_2=population(2,:);
P_12=mean(population);
S_1=0;
S_2=0;
S_12=0;
for i=1:1:n_states

    if P_1(i)>0

        S_1=S_1-P_1(i)*log2(P_1(i));

    end

    if P_2(i)>0

        S_2=S_2-P_2(i)*log2(P_2(i));

    end

    if P_12(i)>0
```

```
        S_12=S_12-P_12(i)*log2(P_12(i));

      end

   end

   JSD=sqrt(S_12-0.5*S_1-0.5*S_2);

end




%{

USAGE EXPLANATION

ensemble: input ensemble

%}

function ens_new=bin_ensemble(ensemble)
% this program will put all off-grid conformations to the closest binning grid

   L_ens=length(ensemble(:,1));
   ens_new=[0 0 0];

   for i=1:1:L_ens

      current_state=ensemble(i,:);
      a=current_state(1);
      b=current_state(2);
      g=current_state(3);

      if mod(a,5)==0&&mod(b,5)==0&&mod(g,5)==0

         ens_new=[ens_new;a b g];
         continue;

      else

         a_min=5*floor(a/5);
         b_min=5*floor(b/5);
         g_min=5*floor(g/5);

         bin_cand=[a_min b_min g_min;a_min+5 b_min g_min;
            a_min b_min+5 g_min;a_min b_min g_min+5;
```

```
        a_min+5 b_min+5 g_min;a_min+5 b_min g_min+5;
        a_min b_min+5 g_min+5;a_min+5 b_min+5 g_min+5];

    single_axis=0;

    for j=1:1:8

        sax=single_axis_R(bin_cand(j,1),bin_cand(j,2),bin_cand(j,3),a,b,g);
        single_axis=[single_axis;sax];

    end

    single_axis(1)=[];
    single_axis_min=min(single_axis);

    bin_min=find(single_axis==single_axis_min);

    ens_new=[ens_new;bin_cand(bin_min(1),:)];

  end

end

ens_new(1,:)=[];

end




%{

USAGE EXPLANATION

a1, b1, g1: rotation angle of conformation 1 (a1, b1, g1)
a2, b2, g2: rotation angle of conformation 2 (a2, b2, g2)

%}

function theta=single_axis_R(a1,b1,g1,a2,b2,g2)
% output amplitude of single axis rotation between (a1, b1, g1) and (a2, b2, g2)

   R1_A=rotation(-g1,-b1,-a1);
   R2=rotation(a2,b2,g2);

   R_total=R1_A*R2;

   theta=acos(0.5*(R_total(1,1)+R_total(2,2)+R_total(3,3)-1))*180/pi;
```

```
end


%{

USAGE EXPLANATION

state: current binning grid
ensemble: the ensemble (distribution) that needs to be binned
bin_width: width used for binning

%}

function subset_match=sax_ens(state,ensemble,bin_width) % bin all the conformations in ensemble that have the amplitude of single axis rotation smaller than bin width to the current binning grid

   L_ens=length(ensemble(:,1));
   subset_match=0;

   for i=1:1:L_ens

      if strmatch([200 200 200],ensemble(i,:))

         continue;

      else

         theta=single_axis_R(state(1),state(2),state(3),ensemble(i,1),ensemble(i,2),ensemble(i,3));

         if theta<bin_width+0.1

            subset_match=[subset_match;i];

         end

      end

   end

   subset_match(1)=[];

end
```

```
%{

USAGE EXPLANATION

alpha, beta, gamma: Euler rotation angles

%}

function rot=rotation(alpha,beta,gamma) % construct Euler rotation matrix

    a=alpha/180*pi;
    b=beta/180*pi;
    g=gamma/180*pi;

    rot_a=[cos(a) -sin(a) 0;sin(a) cos(a) 0;0 0 1];
    rot_b=[cos(b) 0 sin(b);0 1 0;-sin(b) 0 cos(b)];
    rot_g=[cos(g) -sin(g) 0;sin(g) cos(g) 0;0 0 1];

    rot=rot_a*rot_b*rot_g;

end



%{

USAGE EXPLANATION

Below are REsemble algorithm for local dihedral orientation
This is main program for REsemble algorithm for local dihedral orientation
ens_ref: reference (target) ensemble
ens_pred: predicted ensemble

%}

function reso_local=run_ens_jsd_local(ens_ref,ens_pred)

    tic;
    bin_width=10:10:360; % bin width vector that should be chosen and changed by users
    bin_width_size=length(bin_width);
    num_angles=length(ens_ref(1,:));

    reso_local=[bin_width' zeros(bin_width_size,num_angles)];

    for j=1:1:num_angles

        for i=1:1:bin_width_size

            reso_local(i,j+1)=ens_jsd_local(ens_ref(:,j),ens_pred(:,j),bin_width(i));
```

```
        end

    end

    toc;

end
```

```
%{

USAGE EXPLANATION

ens_ref: reference (target) ensemble
ens_pred: predicted ensemble
bin_width: width used for binning two distributions

%}

function JSD=ens_jsd_local(ens_ref,ens_pred,bin_width)

    ens_1=ens_ref+180; % local angles are translated to range from 0 to 360
    ens_2=ens_pred+180;

    grid_angle=0:bin_width:bin_width*floor(360/bin_width); % construct binning grids
    grid_size=length(grid_angle); % number of binning grids

    population_size=grid_size;
    population_array=zeros(2,population_size); % initialization of population distributions under bin_width

    ens_1_size=length(ens_1); % number of conformations in reference ensemble

    %---- binning reference ensemble and construct population distributions ----%
    for i=1:1:ens_1_size

      current_state=ens_1(i);
      current_bin_grid=bin_width*floor(current_state/bin_width);
      ind_angle=find(grid_angle==current_bin_grid);
      population_array(1,ind_angle)=population_array(1,ind_angle)+1;

    end


    ens_2_size=length(ens_2); % number of conformations in predicted ensemble
```

```
  %---- binning predicted ensemble and construct population distributions ----%
  for i=1:1:ens_2_size

    current_state=ens_2(i);
    current_bin_grid=bin_width*floor(current_state/bin_width);
    ind_angle=find(grid_angle==current_bin_grid);
    population_array(2,ind_angle)=population_array(2,ind_angle)+1;

  end

  %---- calculation of jsd between reference and predicted ensembles ----%
  population_sum=sum(population_array);
  ind_empty=find(population_sum==0);

  population_array(:,ind_empty)=[];

  population(1,:)=population_array(1,:)/ens_1_size;
  population(2,:)=population_array(2,:)/ens_2_size;

  P_1=population(1,:);
  P_2=population(2,:);
  n_states=length(P_1);
  P_12=mean(population);
  S_1=0;
  S_2=0;
  S_12=0;

  for i=1:1:n_states

     if P_1(i)>0

        S_1=S_1-P_1(i)*log2(P_1(i));

     end

     if P_2(i)>0

        S_2=S_2-P_2(i)*log2(P_2(i));

     end

     S_12=S_12-P_12(i)*log2(P_12(i));

  end

  JSD=sqrt(S_12-0.5*S_1-0.5*S_2);

end
```