2015

Representing protein native states using weighted conformation ensembles

Vijay Vammi
Iowa State University

Representing protein native states using weighted conformation ensembles

by

Vijay Vammi

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Guang Song, Co-Major Professor
Robert Jernigan, Co-Major Professor
Amy Andreotti
Mark Hargrove
Zhijun Wu

Iowa State University

Ames, Iowa

2015

DEDICATION

I would like to dedicate this thesis to my family, friends and teachers who have shaped me to be the person who I am today. Without their constant support and guidance, this work would not have been possible.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

I would like to take this opportunity to convey my sincere thanks to everyone who have played an important role in conducting this research and writing this thesis. First and foremost, I would like to thank my PhD advisor, Dr. Guang Song, for his constant support and patience. I deeply admire his perseverance, critical thinking and kindness; qualities that he instilled in me during the time I worked with him. I would also like to thank him for being patient with me during my personal hardships. I am thankful to my co-major professor, Dr. Robert Jernigan, for helping me understand the importance of effective scientific communication time and again during group meetings and journal clubs. I am also grateful to my committee members, Dr. Amy Andreotti, Dr. Mark Hargrove and Dr. Zhijun Wu for their insightful comments and questions during the preparation of this work. A special thanks to Trish Stauble for her coordination, timely reminders and for her warmth, which made us feel right at home.

I would like to thank my parents, for all their effort and sacrifice endured to provide me the education I wanted and a special thanks to my brother and sister-in-law for their constant support and guidance. Friends, here and back in India, have played a very important role in my life and I am grateful to them for supporting me during my hardships and cheering me during my success. I would like to thank all my friends from India: Anand, Amar, Sushant, Arun, Arvind, Jagananth, Chaitanya, Anshu, Karun, Aswin, Satya, Rishav, Abhishek and Saikat for their encouragement and well wishes.

Graduate life is daunting and frustrating at times without the support of friends to provide the cushion to fall back on hard days. I would like to thank my friends from Ames: Abhijeet, Chitvan, Navjot, Nidhi, Srivani, Priya, Nalini, Sweta, Arun, Rohit,

Rohan, Omesh, Anu, Swati, Kinit, Teneti family, Hemant, Swaroop, Neevan, Rajat, Divya, Hyejin and the ever-growing members of Punk Pundits for providing me the distraction when I needed one and prioritizing my goals when I am too distracted.

I would also like to thank my lab members: Dr. Tu-Liang Lin and Hyuntae Na, and Dr. Abhijeet Kapoor, Chitvan Mittal, Divya Mistry for many stimulating interactions during my graduate life. Finally, I would like to thank Nidhi Shah for her support, kindness and inspiration during the last leg of graduate life.

Financial support from National Science Foundation (Career award, CCF-0953517) is gratefully acknowledged.

ABSTRACT

The important structural and functional roles played by proteins in the proper functioning of cellular processes cannot be overstated. To comprehensively understand their functional behaviors, structural models derived from experimental data have been developed and these models have played a significant role in explaining the functional mechanisms of proteins. The paradigm "structure drives function" had been active for many years until recent evidence suggested that the complex functions of proteins could not be fully explained by a single structure and dynamics played a very important role in deciphering their functions. To incorporate dynamics into structural representations, ensembles of conformations, instead of a single structure, are used frequently in recent literature and are found to be successful in explaining the functions of many proteins. The work described in this thesis focuses on methods used to construct such ensemble representations of proteins. A careful investigation of the issues and challenges in obtaining such ensembles is undertaken.

In the first part of the thesis, we focus on representing the native state of a given protein using a weighted ensemble representation, where relative populations (or Boltzmann weights) are assigned for individual members of the ensemble. This representation has the advantage of representing the dynamics using only a few conformational states, thereby minimizing the potential of over-fitting, while capturing the dynamics of the protein that a single average structure misses. Using Ubiquitin as an example, we show that determination of such a weighted ensemble representation is feasible when using RDCs as constraints. Moreover, the conformational states of the weighted ensemble are biologically relevant to the functional behaviors of the protein.

We then compare the quality of the weighted ensemble representation with other representations available for Ubiquitin and show that the weighted ensemble representation can successfully reproduce a series of experimental data (RDCs, Residual Chemical Shift Anisotropies, Amide Exchange reactivities and solution scattering profiles) equally well or even better than other representations and without over-fitting. We then extend this work and determine a weighted ensemble representation for Hen Egg White Lysozyme (HEWL). To establish the quality of this ensemble, we perform a series of rigorous cross-validation of this ensemble against extensive amount of experimental data available for HEWL. Lastly, we perform a series of NMR structure refinements under synthetic and controlled conditions to evaluate the structural quality of obtained solutions by various refinement protocols. Our results indicate that ensemble refinement protocols without using weights and good initial conformations may not result in better descriptions of protein native states even though they appear to fit experimental data better and even pass cross-validation test

CHAPTER 1. INTRODUCTION

1.1 Background and Literature Review

1.1.1 Protein Energy Landscape

The complex functions undertaken by proteins are best understood by their structure and dynamics. The energy landscape of protein folding is hypothesized to be rugged with many energy minima (1-4). This model of protein energy landscape can be used to understand the native states of a protein and the folding process. For more and more proteins, increasing evidence suggests that their functional behavior should be best understood not through one single structure but through the distribution and dynamic transition among a number of conformation states that form the native-state ensemble(5-9).

1.1.2 NMR Experimental Data

Per the latest statistics from the Protein Data Bank (PDB) (10), X-ray and Nuclear Magnetic Resonance (NMR) contribute to more than 99% of the deposited structures. This vast amount of structural data has significantly enhanced our understanding of the roles of structure and dynamics in the functions of many proteins. Structures resolved using X-ray crystallography have traditionally been represented by one single structure with regions exhibiting strong evidence of dynamics represented by multiple sub-states. The uncertainties in the positions of atoms are commonly represented by the thermal B-factors. But studies have shown that the obtained electron densities contain information

about the underlying dynamics of the protein and can be used to resolve ensembles (11-16), a view that moves away from the traditional "snap-shot" point of view of X-ray crystallography. With the rapid growth of PDB, protein structures are also becoming increasingly more available and for some well-studied proteins, tens and even hundreds of structures (of one same protein) have been determined. These structures have been shown to capture a representative subset of the native-state ensemble (17).

Nuclear Magnetic Resonance (NMR) studies the protein in the solution state and data collected from NMR naturally originates from the native state of the protein. A couple of relevant NMR data and the information content in them are:

i). Nuclear Over-hauser Effect (NOE): NOEs are observed for spatially proximal atoms and are used to characterize inter-atomic distances. Typically NOEs are used as distance constraints in determining the three dimensional structure of the protein (18, 19).

ii). Residual Dipolar Coupling (RDC): Residual dipolar coupling originates from the interaction of two nuclear spins (dipole-dipole) in the presence of the external magnetic field (20-22). Normally, the residual dipolar coupling reduces to zero because of isotropic tumbling. The anisotropic measurement can be obtained by the aid of various types of liquid crystalline media. RDCs encode the information about the relative orientation of the bond vectors and are used extensively in structure refinement (23-25).

Given the nature of protein native states, NMR data represents a time and ensemble average over all the possible conformations in the native state ensemble.

1.1.3 Structural Modeling and Refinement Using Experimental Data as Constraints

The advancement of the experimental techniques and the increasing availability of experimental data have brought forth a number of exciting works that aim to model the underlying native energy landscape of the protein. Broadly speaking, these works could be classified into two schemes:

1.1.3.1 Refinements:

The experimental data observed and collected in NMR experiments often correspond to geometrical properties of proteins and can be used as constraints in modeling the structures. For example: NOEs encode distance information between protons and can be used as distance constraints. Scalar couplings reveal information about the torsional angles. Given these NMR data, structure refinement can be performed by running molecular dynamics simulations that minimize a preset pseudo-energy function that includes the experimental constraints along with some empirical potential terms (such as those that maintain covalent geometry) (26-28). The end result of such a minimization is a structural model that satisfies both the empirical potential and the experimental data. Several flavors of refinements have been attempted but the most prevalent are:

a). Single structure or average structure refinement: In this scheme, a single structure is used to satisfy both the experimental data constraints and the empirical potential. Since only one conformation is used, this model uses the least number of parameters to satisfy the constraints. For Ubiquitin, one of the most studied proteins, a

single structure has been shown to be sufficient in reproducing most experimental data(29, 30). But it was also pointed out that average structure representations, due to the lack of structural variance, cannot fully capture some of the underlying dynamics (31, 32). Average structure representation becomes less complete when the studied protein occupies multiple distinct sub-states, since the refinement protocol would be over-restrained (under-fitting) (33).

b). Ensemble refinement: In this scenario, instead of using a single structure, an ensemble of conformations is used to explain the experimental data. Consequently the number of parameters used in the model increases linearly with the number of conformations in the ensemble. In the case of Ubiquitin, there has been a number of recent work that aim at determining an ensemble of conformations for the protein, such as MUMO (33), EROS (23), and ERNST (34). The extent to which some of these ensembles represent the native states is however debatable since the ensembles, which contain over a hundred conformations, may be under-constrained by the experimental data (9, 35). As a matter of fact, since the experimental observations and data are macroscopic in nature and represent the ensemble and time averages of microscopic conformations, it may not be possible to verify the validity of each conformation individually in such ensembles. Indeed, the concern of most of these ensembles was mostly about representing the dynamics correctly, less about the validity of each individual conformation.

1.1.3.2 Sample and Select:

Unlike the refinement scheme that tries to obtain a solution satisfying both empirical constraints and experimental data simultaneously, Sample and Select (SAS) strategy solves it in two steps (36, 37). An initial broad pool of conformations is assumed to sufficiently sample the native energy landscape (sampling step) and a few conformations from this broad sample are selected to satisfy the experimental data (selection step).

The quality of the final solution in SAS strategy is heavily dependent upon the initial pool of conformations. Sampling that is insufficient or of poor quality would not be able to satisfy the experimental data better by the subsequent selection step and also could potentially result in wrong solution. Molecular dynamics simulations have been frequently been used to generate the starting pool (37, 38). The subsequent selection step could either be

a). Weighted selection, where along with the conformations, the weights are also adjusted to optimally satisfy the experimental data (36, 38, 39). Assigning weights to conformational states was considered using NOE (40) but elegant seminal work by Brunger's group had shown that regular NOE data alone was insufficient to determine the relative populations of a two-conformer ensemble (41). Most of the recent works using weighted ensembles are on intrinsically disordered proteins (IDPs).

b). Equal weighted selection, where all the conformations in the selection are given equal weight (37). An equal weighted selection can capture the relative population information to some extent, by including in the ensemble multiple copies of one similar conformation. The number of copies thus indirectly encodes the weight. Apart from work

by Dokholyan and co-workers (37), there has been minimal work done using equal weighted ensemble to satisfy experimental data.

1.2 Motivation and Aims of this Study

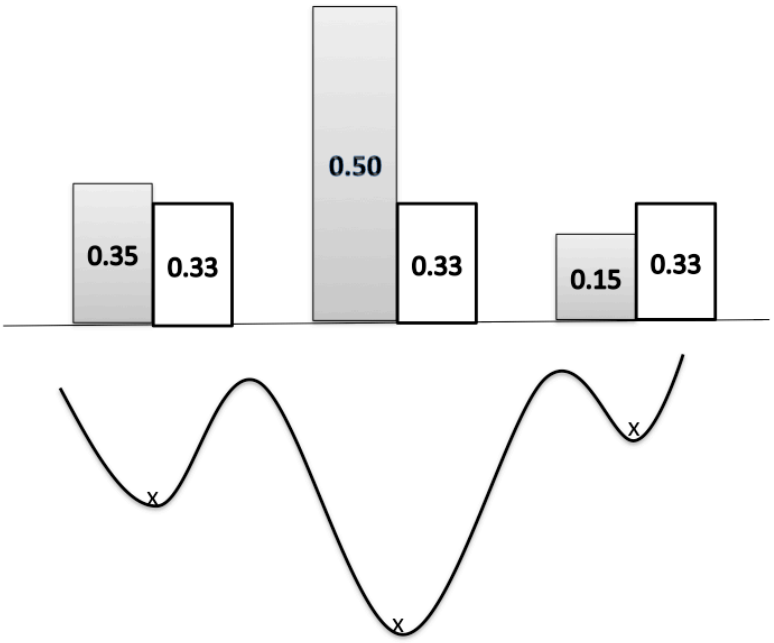

Figure 1.1: Pictorial representation of Boltzmann weights versus sampling weights. The 'x' marks represent conformations on a hypothetical energy landscape while the white bars represents sampling weights and shaded bars represent the Boltzmann weights.

For many a protein, the conformation space near its native states can be best represented by a number of inter-connected conformation states, each of which may have a different population, as illustrated in Figure 1.1. When an ensemble of conformations are used to represent the conformation space (shown as the cross marks in Figure 1.1), its quality in representing the conformation space is determined by three factors:

1) COMPLETENESS: Are all conformation states reached by at least one conformation?

2) COVERAGE: For each of the conformation states that are reached by some conformation(s), what is quality of the coverage? In other words, how well do the finite number of conformations that are in a given conformational state together represent that conformation state?

3) CONTRIBUTION: Is the number of conformations at each conformation state proportional to the ideal Boltzmann weight?

Ideally, we would like to have an ensemble that has an infinite number of conformations that cover all the conformation states according to the Boltzmann distribution. Such an ensemble would have perfect completeness, coverage and contribution. In reality, our ensembles are of finite sizes, having tens or possibly hundreds of conformations, which are relatively small comparing to the vast conformation space. Therefore, we do not have perfect completeness, coverage, or contribution.

Another key point to realize is that the matter of completeness and coverage are sampling issues. The conformations in an ensemble could have come from experiments, by structure determination methods such as X-ray crystallography, NMR, etc., or they could have been determined computationally. Whatever the source is, completeness and coverage are sampling issues. They reflect the sampling quality of a given ensemble.

However, how well an ensemble represents the conformation space near the native state and how well it can reproduce experimental data/observations are determined not solely by the ensemble's completeness or coverage. It depends also on the third factor - contribution. Without doubt an ensemble whose conformations are assigned a population (contribution) proportional to their actual Boltzmann weights would represent the conformation space the best, reaching the limit of that ensemble's ability in representing the conformation space. Therefore, an ensemble with a proper assignment of relative contributions given to its conformation states should do better than an ensemble without. As illustrated in Figure 1.1, the conformation space of a protein is represented by three conformations, which by default are given an equal weight of 1/3. However, the ensemble can be enhanced if the actual Boltzmann weights (represented by dark shaded blocks) can be determined somehow and assigned to the three conformations.

There are a couple of reasons why few work has been carried out to exploit the potential benefit of including these weights (or relative populations). First, an elegant seminal work by Brunger's group had shown earlier that regular NOE data alone was insufficient to determine the relative populations of a two-conformer ensemble (41). Thus it was not clear if there were enough experimental data to determine the populations meaningfully, even though the authors (41) were hopeful that relative populations could possibly be determined when other sources of experimental data were provided. Secondly, equal-weight conformations themselves can capture the relative population information to some extent, by including in the ensemble multiple copies of one similar conformation. The number of copies thus indirectly encodes the weight. However, it is an

insufficient way to represent the populations, as it requires more conformations to be in the ensemble and thus may worsen the potential problem of over-fitting.

Based on this intuition, we have narrowed our aims to be:

1. Develop methods that can assign relative populations to structure ensembles by using experimental RDC data as constraints, taking extensive care that the assigned weights are robust and not over-fitted.
2. Assess and validate the quality of so-determined weighted ensembles using a series of experimental data.
3. To make the method broadly applicable to many other proteins, determine also what the minimal requirement for experimental data is in assigning relative populations to ensembles.
4. Apply and analyze the potential role and benefit of relative populations in ensemble refinements.

1.3 Thesis Organization

Chapter 2 is a published paper detailing the method employed to assign relative populations to structural ensembles using RDC data. We carefully delineate the properties required by the structural ensemble to generate a reliable weighted ensemble. In chapter 3, we perform extensive cross validation of the 2 weighted ensembles of Ubiquitin, constructed in Chapter 2, using varied experimental data consisting of unused RDCs, Residual Chemical shift anisotropies, amide hydrogen reactivities and solution

scattering profiles. We also compare the 2 weighted ensembles against alternate representations of Ubiquitin available in literature. This chapter is a manuscript submitted for review. In Chapter 4, we extend the method employed in Chapter 2 to other proteins by identifying the minimal experimental data required to derive weighted ensembles. Further, we also construct a weighted ensemble of Hen Egg White Lysozyme (HEWL) using only NH RDC data and perform extensive cross-validations using Residual Chemical shift anisotropies and solution scatter profiles. Chapter 5 is a preliminary report of an ongoing work aimed to assess the structural quality of solutions generated by ensemble refinements using synthetic data.

CHAPTER 2. ENHANCING THE QUALITY OF PROTEIN CONFORMATION ENSEMBLES WITH RELATIVE POPULATIONS

A Paper published in Journal of Biomolecular NMR

Vijay Vammi, Tu-Liang Lin and Guang Song

2.1 Abstract

The function and dynamics of many proteins are best understood not from a single structure but from an ensemble. A high quality ensemble is necessary for accurately delineating protein dynamics. However, conformations in an ensemble are generally given equal weights. Few attempts were made to assign relative populations to the conformations, mainly due to the lack of right experimental data. Here we propose a method for assigning relative populations to ensembles using experimental residue dipolar couplings (RDC) as constraints, and show that relative populations can significantly enhance an ensemble's ability in representing the native states and dynamics. The method works by identifying conformation states within an ensemble and assigning appropriate relative populations to them. Each of these conformation states is represented by a sub-ensemble consisting of a subset of the conformations. Application to the ubiquitin X-ray ensemble clearly identifies two key conformation states, with relative populations in excellent agreement with previous work. We then apply the method to a reprotonated ERNST ensemble that is enhanced with a switched conformation, and show that as a result of population reweighting, not only the reproduction of RDCs is

significantly improved, but common conformational features (particularly the dihedral angle distributions of ϕ_{53} and ψ_{52}) also emerge for both the X-ray ensemble and the reprotonated ERNST ensemble.

2.2 Introduction

The functions of a protein are closely related to not only its structure but also its dynamics. For more and more proteins, it is becoming increasingly evident that their functional behavior is best understood not through one single structure but through the distribution and dynamic transition among a number of conformation states that form the native-state ensemble (5-9, 15, 42, 43). Such an ensemble representation is consistent with the energy landscape theory and the 'protein folding funnels' (4, 44, 45). With the rapidly growing Protein Data Bank (PDB) (10), protein structures are becoming increasingly more available and for some well-studied proteins, tens and even hundreds of structures (of one same protein) have been determined. These structures have been shown to capture a representative subset of the native-state ensemble (17).

On the other hand, the advancement of the experimental techniques and the increasing availability of experimental data has brought also a number of exciting recent works that aim to determine protein conformation ensembles instead of a single structure, using the experimental data as constraints (23, 24, 33, 46-48). The extent to which some of these ensembles represent the native states is debatable since the ensemble, which in some cases contains over a hundred conformations, may be under-constrained by the experimental data. As a matter of fact, since the experimental observations and data are macroscopic in nature and represent the ensemble and time averages of microscopic

conformations, it may not be possible to verify the validity of each conformation individually in such ensembles. Indeed, the concern of most of these ensembles was mostly about representing the dynamics correctly, less about the validity of each individual conformation.

For all of the above ensemble determination protocols, the conformations within the ensemble were given equal weights, i.e., $1/N_e$, where N_e is the size of the ensemble. While weights were listed out as part of the parameters in some of these methods, weights other than equal weights were not studied. Physically these weights represent relative populations of the conformations and thus their relative contributions to the ensemble.

There are a couple of reasons why few work has been carried out to exploit the potential benefit of including these weights (or relative populations). First, an elegant seminal work by Brunger's group had shown earlier that regular NOE data alone was insufficient to determine the relative populations of a two-conformer ensemble (41). Thus it was not clear if there were enough experimental data to determine the populations meaningfully, even though the authors (41) were hopeful that relative populations could possibly be determined when other sources of experimental data were provided. Secondly, equal-weight conformations themselves can capture the relative population information to some extent, by including in the ensemble multiple copies of one similar conformation. The number of copies thus indirectly encodes the weight. However, it is an insufficient way to represent the populations, as it requires more conformations to be in

the ensemble and thus may worsen the potential problem of over-fitting mentioned above.

Our hypothesis. In this work we propose that it is feasible to assign relative populations to ensembles by using experimental RDC data as constraints, and that adding relative populations should enhance an ensemble's ability in representing a protein's native states and its dynamics.

Assigning Relative Populations to An Ensemble: Background and Problem Definition.

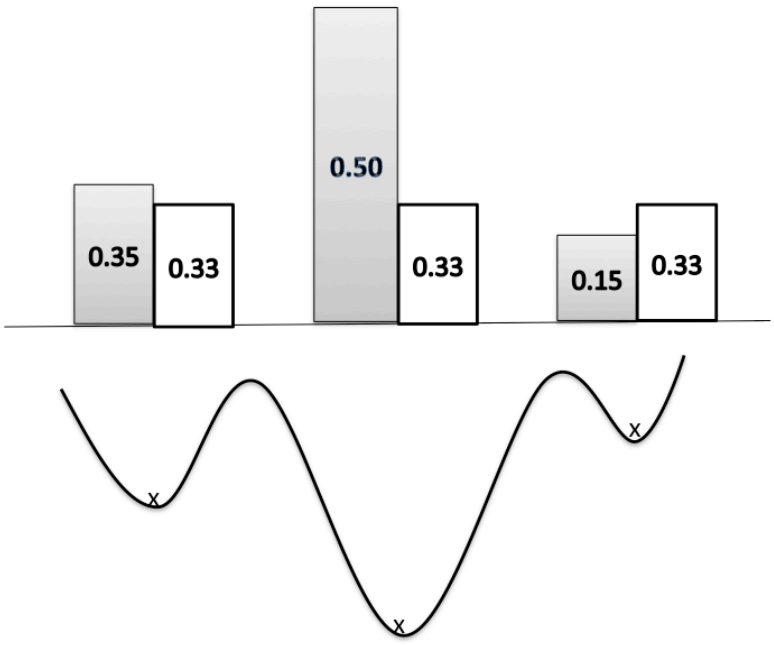

Figure 2.1: Pictorial representation of Boltzmann weights versus sampling weights. The 'x' marks represent conformations on a hypothetical energy landscape while the white bars represents sampling weights and shaded bars represent the Boltzmann weights.

For many a protein, the conformation space near its native states can be best represented by a number of inter-connected conformation states, each of which may have a different population, as illustrated in Figure 2.1. When an ensemble of conformations are used to represent the conformation space (shown as the cross marks in Figure 2.1), its quality in representing the conformation space is determined by three factors:

1) COMPLETENESS: Are all conformation states reached by at least one conformation?

2) COVERAGE: For each of the conformation states that are reached by some conformation(s), what is quality of the coverage? In other words, how well do the finite number of conformations that are in a given conformational state together represent that conformation state?

3) CONTRIBUTION: Is the number of conformations at each conformation state proportional to the ideal Boltzmann weight?

Ideally, we would like to have an ensemble that has an infinite number of conformations that cover all the conformation states according to the Boltzmann distribution. Such an ensemble would have perfect completeness, coverage and contribution. In reality, our ensembles are of finite sizes, having tens or possibly hundreds of conformations, which are relatively small comparing to the large conformation space. Therefore, we do not have perfect completeness, coverage or contribution.

Another key point to realize is that the matter of completeness and coverage are sampling issues. The conformations in an ensemble could have come from experiments, by structure determination methods such as X-ray crystallography, NMR, etc., or they could have been determined computationally. Whatever the source is, completeness and coverage are sampling issues. They reflect the sampling quality of a given ensemble.

However, how well an ensemble represents the conformation space near the native state and how well it can reproduce experimental data/observations are determined not solely by the ensemble's completeness or coverage. It depends also on the third factor - contribution. Without doubt an ensemble whose conformations are assigned a population (contribution) proportional to their actual Boltzmann weights would represent the conformation space the best, reaching the limit of that ensemble's ability in representing the conformation space. Therefore, an ensemble with a proper assignment of relative contributions given to its conformation states should do better than an ensemble without. As illustrated in Figure 2.1, the conformation space of a protein is represented by three conformations, which by default are given an equal weight of 1/3. However, the ensemble can be enhanced if the actual Boltzmann weights (represented by dark shaded blocks) can be determined somehow and assigned to the three conformations. Now, the immediate questions are: are relative contributions even determinable? And if so, how? And what is required to determine them?

In this work, our focus is on this third aspect -- contribution. Our hypothesis is that given an ensemble of reasonable quality in completeness and coverage, the relative contributions can be determined by using experimental RDC data as constraints.
We will apply a least-square fitting algorithm to determine the weights. To prevent over-fitting, careful cross-validation is applied. In the following Materials and Methods section, we present our approach in details.

2.3 Materials and Methods

Recall that the problem we want to address here is that, given a conformation ensemble and a sufficient amount of experimental RDC data, is it possible to assign meaningful populations to the conformations in the ensemble without incurring over-fitting? To what extent can we assign the populations? There are two extremes. One extreme is to assign each conformation with a population, which is physically unrealistic and generally cannot be achieved. The other is to assign the whole ensemble as a group with a (percentage) population of 1. This is equivalent to equal weights that have been used. Our hypothesis is that sufficient experimental data should allow weight assignment to clusters of conformations, or sub-ensembles, within the ensemble.

In this section, we present our method for assigning relative populations to clusters of conformations within an ensemble. The potential problem of over-fitting that often arises in such a process is carefully addressed. The significance of the assigned relative populations is further examined by cross-validation.

There are four major steps in our method, which are described in order in the following sections. Briefly, the first step, a pre-processing step, merges conformations in the ensemble into small conformation clusters. For ensembles whose sizes are small, this step is skipped. The second step takes the pre-processed ensemble and applies a least squares fitting algorithm to identify a subset of conformations/clusters that best represent the conformation states. Step three takes this subset as a whole and iteratively split it into smaller sets until right before over-fitting starts to occur. Lastly, the significance of the relative populations thus assigned is evaluated by cross-validation.

Step I: Pre-processing to reduce the dimensionality of the ensemble.

In cases where the ensemble size is large and it has more conformations than the number of experimental RDC data points, clustering (49, 50) is carried out to reduce the dimensionality of the ensemble. Here the dimensionality of an ensemble refers to the structural variety of the ensemble and is set to be the number of clusters in the ensemble. Initially each conformation in the ensemble forms its own cluster. Clustering structurally similar conformations into small clusters thus helps reduce the dimensionality and makes the ensemble manageable for the least square fitting procedure to be applied in the next step.

The distance between a pair of clusters is defined as the average of all the pairwise distances between the conformations in the two clusters. The distance between two conformations is defined by Q_{score}.

$$Q_{score} = \frac{\left(\sum_{\{i\ !=\ j\}} \exp\left(-\left(r^{A}_{\{i,j\}} - r^{B}_{\{i,j\}}\right)^2\right)\right)}{N(N-1)} \qquad (1)$$

where $r_{i,j}$ is the distance between atoms i and j in a conformation and N is the total number of atoms. Q_{score} value ranges from 0 to 1, 0 being the very dissimilar and 1 being perfectly similar (51).

Initially each conformation in the ensemble forms its own cluster. The following three steps are iterated. As a result, similar conformations will be bundled together into larger clusters, while the rest remains as singlet clusters.

1. Identify the closest pair of conformations in the ensemble. Merge them into a cluster if their distance is less than a threshold, D_{max}. Otherwise stop the procedure.
2. Grow the cluster formed in step 1 by repeatedly adding to it the next conformation whose average distance to the conformations in the cluster is the smallest and is less than D_{max}, otherwise stop adding.
3. Remove the cluster and go back to step 1.

Step II: Identify Representative Conformations by Least-Square Fitting to RDCs

Residual dipolar coupling comes from the interaction of two nuclear spins (dipole-dipole) in the presence of the external magnetic field and is defined as (20, 21, 52):

$$D_{\{ij\}} = -\frac{\mu h r_i r_j}{(2\pi r)^3} \left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle \qquad (2)$$

where r_i and r_j are the nuclear magnetogyric ratios of nuclei i and j respectively, h is Plank's constant, μ is permittivity of space, r is the internuclear distance between the two nuclei and Θ is the angle between the internuclear vector and the external magnetic field. The brackets represent the ensemble and time average. Normally, the residual dipolar coupling reduces to zero because of isotropic tumbling. The anisotropic measurement can be obtained by the aid of various types of liquid crystalline media.

For a protein with a number of distinct conformation states, the observed RDC data are best reproduced when the conformations close to these conformation states are present in the ensemble and given proper weighting. The conformations in a given ensemble may not all fall close to a conformation state. Here we use least square fitting to identify which conformations are needed and what relative populations should be given to them in order to best reproduce the experimental RDC data. By doing this, we can pick out key representative conformations from the ensemble. The relative populations assigned to them, however, are subject to the problem of over-fitting, due to the intrinsic

nature of least square fitting. However, measures will be taken to identify the onset of over-fitting and prevent it from affecting weight assignment, as addressed in the next section.

Appendix 2.7 describes how RDCs can be back calculated from a single conformation or an ensemble of conformations. In this process of back calculating, singular value decomposition is commonly used to obtain the least square solution for the alignment tensor. Here we apply the same technique iteratively to obtain the least square solution for the relative populations as well. First, equal weights ($\frac{1}{n}$) are given to all clusters (which are determined at step I) and Equation 13 (see Appendix 2.7) is used to obtain the optimal Saupe matrix, S. After S is obtained, it is used to determine $w_k' s$ by least squares fitting. The process is iterated until the weights have converged. In the end, each cluster has either positive or zero population, since the weights are derived under the nonnegative constraints (53). In the case where there are multiple RDC data sets, different alignment tensors are needed for different media. The optimal weight combination (the relative populations) is obtained by least squares fitting to all the RDC data sets. A detailed description of these iterative least squares fitting algorithms is given in Appendix 2.7.

The iterative least squares fitting of the conformations in the ensemble to multiple RDC datasets returns a list of clusters/conformations that have non-zero populations. The conformations in these clusters are recognized as representative conformations.

In cases where there are more conformational clusters than the experimental data points, representative clusters are identified through the following procedure.

1. From the pool of all available conformational clusters, randomly select N clusters, where N is the number of experimental data points.
2. Run the least squares fitting algorithm (Appendix 2.7) to determine cluster weights. Some clusters may have zero weights.
3. Repeat steps 1 and 2 many times and record the cluster weights at each iteration.
4. The top N clusters with the highest average weights are identified as representative clusters.

The representative clusters form the leaf nodes of a hierarchal clustering tree, built bottom up by merging the closest pair of clusters at each iteration.

Step III: Splitting and the Identification of Over-fitting

To avoid the potential problem of over-fitting that may take place in the process of assigning relative populations, we take steps to recognize the onset of over-fitting and prevent it from affecting the weight assignment. Recall that there are two extremes in assigning weights. One is to assign each conformation with a population. The other is to assign the whole ensemble as a group with a population of 1, which is equivalent to having equal weights. In our studies we have found that one may confidently move beyond equal weighting and assign relative (different) populations to sub-ensembles but not to the point that each conformation in the ensemble is given a weight. There exists a

limit where one cannot further divide the sub-ensembles into smaller pieces. This limit represents the extent to which relative populations can be assigned and it depends on the quality of the ensemble and the quality and quantity of the experimental data. In reality, the limit is determined through monitoring the onset of over-fitting.

In the following procedure, we iteratively split the ensemble, which is now made up of the representative conformations, into smaller and smaller clusters. The splitting process is the same as the inverse process of hierarchical clustering. At each iteration, only one cluster is spit into two, which corresponds to the merging of the closest pair of clusters in hierarchical clustering. Therefore there are k clusters at the k_{th}iteration. By applying the least squares fitting algorithms as described in Appendix B, we can assign relative populations (or weights) to these k clusters.

If we have N sets of experimental RDC data that are consistent with one other and contain random measurement noise within them, N sets of weights will be assigned to the k clusters. Now if the weight assignment is correct, we expect that these N sets of weights should strongly correlate with one another. The onset of overfitting is when such correlations start to greatly degrade. That is, it begins to fit to the noise. Since noise is random and uncorrelated in the different experimental data, the weights fitting to noise should also be uncorrelated. This recognition of the onset of over-fitting is even more sensitive when the correlations are computed using only the weights of the two newly birthed clusters at the k_{th} iteration. The idea is that, if the two newly birthed clusters belong to one conformation state and should not have been split, we expect the weights

assigned to them by different sets of experimental data should be ambiguous and lack consistency and thus low correlations. On the other hand, if these two clusters belong to difference conformation states and should be split, we expect to see consistent weight assignments from different experimental datasets and thus high correlations.

Replicate Experimental Data for Over-fitting Identification.

To identify over-fitting as outlined above, all the experimental data is duplicated to create N identical copies and then different random Gaussian noise are added to each of them. These N datasets are thus identical except for the noise in them.
A relatively large N is needed to have a high sensitivity to the onset of over-fitting. N is set to be 20 in this work. The standard deviation of the random Gaussian noise added to each replica is set to be 80% of the modeled experimental noise, which are bond-dependent and are set to be 0.26 Hz, 0.1 Hz, 0.5 Hz, 0.1 Hz and 0.1 Hz for NH, CaC, CaHa, CN and CHN datasets respectively as was done in (Clore and Schwieters, 2004a).

We use Q-factor to measure how well the weight assignments are correlated with one another. The definition of Q-factor is given in equation 3, where it is employed also to measure the similarity between experimental and computed RDC data. The maximum of the Q-factors between any two of the N weight assignments is denoted as MaxQ. A large MaxQ (above a certain threshold) indicates inconsistent weight assignments and

thus over-fitting for the two newly birthed clusters. A threshold value of 0.06 is used for MaxQ throughout all the cases investigated below. In summary, the procedure is:

1. Initially all the representative conformations belong to one single cluster.
2. Experimental data is replicated into N sets. $N = 20$.
3. Iteratively split the clusters (the exact inverse of a hierarchical clustering).
4. Assign sets of weights to clusters based on fitting to the experimental datasets.
5. Check if the weights assigned to the newly birthed clusters are significant (i.e., weight ≥ 0.01). If any weight is found to be insignificant, repeat the process by removing the insignificant cluster.
6. Compute the weight correlations and MaxQ for the two newly birthed clusters.
7. If the minimum of the weight correlations is negative and MaxQ is greater than a predefined threshold, it signifies that over-fitting has occurred. In this case, the two newly birthed clusters are merged back together and the cluster is marked ``final", indicating that it can no longer be split. Otherwise, continue and move on to the next iteration. Stop the procedure when there is no cluster left that can be split.

Step IV: Adding back other conformations

By the end of step III, we have partitioned the ensemble into a number of ``final" clusters, with N sets of weights assigned to each of them. Now compute the mean weight value and the standard deviation for each cluster. The clusters whose mean weight value

is less than its standard deviation are then removed, as they do not consistently have a positive weight.

Each of the remaining clusters is considered as representing an independent conformation state. Since it remains possible that the conformations that were excluded earlier at step I and step III may belong to one of the conformational states that these clusters are representing, adding some of them back to the clusters thus may possibly improve the quality of the ensemble. The sequence in which conformations are added back is arranged, in increasing order, by the minimum distance between a conformation and any of the clusters. A conformation is added to the cluster to which it is closest if including it decreases the overall Q-factor.

Estimate the Uncertainty in Weight Assignments

After the conformation states (i.e., the clusters) have been identified and weights assigned to them, it is possible to estimate the uncertainty in the weight assignments, provided that there exist multiple sets of experimental data. This is because least squares fitting can be applied to fit each set of experimental data independently. If there are M sets of experimental data, this will result in M sets of weight assignments, or M weight assignments to each cluster. It is expected that the weight assignments for each cluster are in general not identical, since there is noise in the experimental data and the cluster representation for each conformation state is not perfect. The levels of uncertainty in the

weight assignments can be estimated by computing the standard deviation within the weight assignments for each cluster.

Cross-Validation

Q-factor is a commonly used measure of the agreement between the experimental and calculated RDCs and is defined as:

$$Q\text{-}factor = \frac{\sqrt{\sum\left(D_{calc} - D_{\exp}\right)^2}}{\sqrt{\sum\left(D_{\exp}\right)^2}} \qquad (3)$$

where D_{calc} is the calculated RDC and D_{exp} is the experimental RDC.

The introduction and assignment of relative populations to an ensemble improves the Q-factors. To assess the significance of such improvement, we leave out CaHa RDC from the experimental data when determining the weights. The CaHa dataset was then used for cross-validation. Lange et al (23) used CN vector for cross-validation. Given that the data used in refinement includes CaC, CHN, NH vector orientations, CN RDC might not be the best choice. CaHa vector, on the other hand, is not in the peptide plane and is thus independent of other bond vector orientations, making it a better cross-validation dataset. Cross-validation provides a way to check whether the better fitting gained by assigning relative populations is a fitting to the noise in experimental data or is a fitting to the true data. If the ensemble with relative population assignment does render a better

representation of the conformation space, we expect that the fitting to the leaving-out CaHa data should also improve.

All the conformations are stripped off its hydrogen atoms first and then re-protonated using Reduce (54) before computing Saupe matrices and back-calculating RDC values.

Ubiquitin ensembles and experimental RDC data sets

Ubiquitin has long been used as a model protein to probe protein dynamics and for which abundant experimental RDC datasets are available. A total of 62 RDC data sets, including NH, CN, CHN, CaC, CaHa and side chain methyl, were used to determine EROS ensemble (23). Since our procedure requires that the relative populations be determined by fitting to experimental RDC data, it is critical that the data has no significant errors. For this reason we have pruned the above dataset to remove any dataset whose data points are less than 40 and whose Q-factors are significantly higher when back-calculated using structure 1UBQ or 1D3Z (NMR ensemble).

Table 2.1: RDC datasets used for weighting Ubiquitin ensembles, coded according to (55)

Experimental data type	RDC data
NH	A1, A2, A4, A6, A7, A8, A9, A10, A11, A12, A13, A16, A21, A22, A23, A24, A25, A26, A27, A28, A29, A34, A36
NH, CN, CHN, CaC and CaHa	2 sets (56)

Table 2.1 lists the experimental datasets used in this work, using the code names given in Lakomek et al. (55). There exist a few other multi-vector datasets for Ubiquitin (57). However, they are not included here since they display relatively large Q-factor when applied to the NMR structure 1D3Z. For the same reason, NH datasets labeled A3, A5, A30, A31, A32, A33, and A34 (as in (55)) are not included either.

2.4 Results

In this section, we apply our method to assign relative populations to conformation ensembles of proteins. It is assumed here that the protein that an ensemble represents should have a small number of conformation states, and that some of the conformations in the ensemble, though sparse relative to the large conformation space, fall close to the protein's conformation states. These conformations may come from experimentally determined structures of the protein. Because of their scarcity, there is no expectation on these conformations that their distribution on the conformation space should be Boltzmann distribution. For such an ensemble, and using experimental RDC data as constraints, we will show to what extent one can meaningfully assign relative populations, or weights, to the ensemble. We aim to answer also, in order to assign meaningful relative populations, what is the minimum requirement on the ensemble. In the end, we apply the method to an ensemble of crystal structures of Ubiquitin.

Creating an Artificial Conformation Ensemble and Artificial RDC Data

To test our method, we first create an artificial energy landscape and a native state ensemble that will be used as a reference (33). We create also artificial RDC data based

on the ensemble composition. The advantage of using artificial ensembles and RDCs is that we have perfect control of their composition and their noise level.

Creating a Native State Ensemble. To create an artificial native state ensemble, five distinct conformations of protein ubiquitin are picked from an accelerated MD simulation (58). The conformations are chosen such that the minimum RMSD between any two conformations is greater than 2.5 Å. We assume that these five conformations represent the centers of all the (five) possible conformational states of the protein. We then sample more conformations around these centers and use them, together with the centers, to represent the conformation states. This is done using CONCOORD (59). CONCOORD, by default, can produce quite broad distributions of conformations. To ensure that each conformational state is tightly clustered, a damping coefficient of 0.3 is applied when generating the distance restraints from these five conformations. As a result, the average RMSD within any sub-ensemble is close to 1 Å. Thus, the conformations fall into five clearly separated clusters.

Next, we set the Boltzmann weight of each conformation state to be proportional to the number of conformations in its energy well (i.e., the sub-ensemble around each conformation state). The number of conformations sampled in each sub-ensemble and the associated Boltzmann weights are given in Table 2.2.

Table 2.2: Boltzmann weights of the five conformational states in the artificial ensemble.

Conformational State	One	Two	Three	Four	Five	Total
# of Conformations	100	200	350	500	700	1850
Boltzmann weight	0.054	0.108	0.189	0.27	0.378	1

Noise Conformations. Noise conformations are those that do not contribute to experimental observations. Strictly speaking though, every conformation in the ensemble contributes to the observations to some extent. But those conformations that are away from any of the protein's conformation states have so low a weight that they virtually do not contribute. We consider such conformations as noise conformations as contrast to those that do represent the protein's conformation states.

To create noise conformations, we use CONCOORD to sample around each conformational state without any damping. The average RMSD in this sampling is around 2.5 Å. To guarantee these conformations do represent noise, we remove from them any conformations that can give nearly the same Q-factor as the conformations representing the conformation states.

Generating Artificial RDC Data. Using all the conformations (1850 total, see Table II) of the ensemble, artificial RDC datasets matching the composition of the real experimental RDC data of Ubiquitin, are generated. The average *A* matrix of the ensemble is first calculated. Then for each of the experimental datasets listed in Table 2.1 the best-fit Saupe matrix is determined using 1D3Z NMR ensemble. An artificial RDC dataset is then created by multiplying the average *A* matrix with the Saupe matrix. At this point, these RDC datasets are noise-free. We will call them noise-free RDCs.

In reality, experimental data contains noise of about 0.5 to 1.0 Hz (24), we add Guassian noise to the artificially generated RDC data that are originally noise-free. The standard deviations of the noise are 0.26 Hz, 0.1 Hz, 0.5 Hz, 0.1 Hz and 0.1 Hz for NH, CaC, CaHa, CN and CHN datasets respectively as was done in (Clore and Schwieters, 2004a). Note that because of the way in which the artificial RDC data are generated, the given conformation ensemble can perfectly reproduce these RDC data prior to the adding of the noise, but not so after. In the rest of this article, unless explicitly noted, artificial RDCs refer to the ones that contain noise.

What Is Required of the Ensemble?

In the section we aim to determine what is the requirement of the ensemble in order to have a meaningful weight assignment. We design four test cases to examine the applicability of the method. The purpose of these four cases is to show that neither under-sampling at each conformation state nor noise conformations hinder weight assignments.

Case I:

In this case we assume there is no noise conformations and the ensemble contains only conformations from the five conformation states. However, the number of conformations at each state is not proportional to its Boltzmann weight. 21, 60, 6, 7, 290 conformations are randomly selected from conformation state one, two, three, four, and five respectively and mixed together to form an ensemble. Our method is then applied to assign relative populations to this ensemble. Table 2.3 lists the clusters obtained in the

end, along with the composition of the clusters, weights assigned and expected weights of all the clusters.

Table 2.3: Final weights and cluster compositions for Case I. The convention used for the composition of a cluster is to enumerate in order the number of conformations belonging to five conformational states.

Cluster	Final weights± std	Composition	Belongs to	Expected Weight
Cluster1	0.072 ± 0.001	20,0,0,0,0	First state	0.054
Cluster2	0.097 ± 0.0004	0,8,0,0,0	Second state	0.108
Cluster3	0.183 ± 0.002	0,0,5,0,0	Third state	0.189
Cluster4	0.27 ± 0.002	0,0,0,7,0	Fourth state	0.27
Cluster5	0.376 ± 0.001	0,0,0,0,284	Fifth state	0.378

It is seen from Table 2.3 that the final weight obtained for each conformation cluster is highly similar to the expected Boltzmann weight and each cluster contains purely conformations that belong to that conformation state. Similar results are obtained when the same experiment is repeated with different replica noise.

Case II:

In this case, one of the conformation states (the third) was intentionally not included in the process of generating artificial experimental data. This is done to mimic the scenario where an ensemble contains a cluster of conformations from a state that does not belong to the native ensemble. While the purpose for the first case is to test if the method is able to assign right populations to the conformations contributing to the experimental observations, the purpose for this one is to test whether or not the method is able to assign no weight to conformations that do not contribute.

Table 2.4: New relative Boltzmann weights after the third cluster is excluded from artificial RDC data generation.

Conformational State	One	Two	Four	Five	Total
# of Conformations	100	200	500	700	1500
Boltzmann weight	0.066	0.133	0.333	0.467	1

The new relative Boltzmann weights are given in Table 2.4. The same conformation ensemble employed in case I, which includes conformations that do not contribute to the artificial RDC calculations, is used here. After applying our method, the resulting clusters, along with their compositions, assigned weights and the standard deviations, and expected weight are given in Table 2.5. From the results it is seen that, as with Case I, the weights obtained are highly similar to the expected values. Moreover, each cluster consists purely of conformations belonging to that cluster.

Table 2.5: Final weights and cluster compositions for Case II. The convention used for the composition of a cluster is the same as Table 2.3.

Cluster	Final weights± std	Composition	Belongs to	Expected Weight
Cluster1	0.087 ± 0.0001	9,0,0,0,0	First state	0.066
Cluster2	0.118 ± 0.004	0,60,0,0	Second state	0.133
Cluster3	0.322 ± 0.003	0,0,0,7,0	Fourth state	0.333
Cluster4	0.471 ± 0.001	0,0,0,0,280	Fifth state	0.467

Case III:

In the first two cases, the conformations in the ensemble are clearly separated into five distinct clusters. In reality, such distinction is often smeared by the presence of other conformations. These other conformations virtually do not contribute to the experimental

observations (the “noise” conformations). However, their presence makes it difficult to identify conformation states, or separate conformations representing a conformation state from those that do not. To mimic this reality, we introduce noise conformations into the ensemble.

The same conformations as used in Case I are used here (see Table 2.2). In addition, an equal number of noise conformations (see above on how they are generated) are added to each cluster so that they represent half of the total conformations in each cluster. As a result, the number of conformations in the ensemble is doubled and becomes 768, of which 384 are noise conformations. Clustering, as described in step I in the Materials and Methods section, results in 406 clusters, of which some are singlet clusters. Since the number of clusters is more than the number of unique experimental data points (around 200), ``representatives" conformations are identified by following step II (see Materials and Methods).

Table 2.6: Final weights and cluster compositions for Case III. The convention used for the composition of a cluster is the same as Table 2.3.

Cluster	Final weights± std	Composition	Belongs to	Expected Weight
Cluster1	0.069 ± 0.005	8,0,0,0,0	First state	0.054
Cluster2	0.099 ± 0.001	0,8,0,0,0	Second state	0.108
Cluster3	0.181 ± 0.003	0,0,5,0,0	Third state	0.189
Cluster4	0.273 ± 0.005	0,0,0,7,0	Fourth state	0.27
Cluster5	0.377 ± 0.002	0,0,0,0,281	Fifth state	0.378

Table 2.6 lists the results. There are five clusters, which are composed of 8, 8, 5, 7, and 281 conformations from the five conformational states respectively. All of the 384 noise conformations are successfully filtered out. From Table 2.6 it is seen that weight assignments for the clusters are highly similar to the expected values.

Case IV:

In all of the above cases, we have simulated full coverage of the conformational states by having each of the states represented by at least a few conformations. To assess the impact on the reproduction of the experimental data when one of the conformational states is missing all together, we apply our weighting algorithm again to the ensemble used in case III but this time each of the five clusters used to represent the five conformational states, in turn, is purposely left out. We want to see if the algorithm will produce RDC Q-factors with equal quality, while having substantially different conformational properties than the initial ensemble, by somehow rearranging the weights for the remaining clusters. Table 2.7 lists the results.

Table 2.7: New weight assignments and Q-factors when each of the five clusters, in turn, is purposely left out of the ensemble, as in case IV. CaHa is used for cross-validation. Note that the weights of the remaining four clusters do not add up to 1 in some cases. This happens when noise conformations form a new cluster(s) and are assigned a non-zero weight to compensate the missing cluster.

	Weights					NH	CaC	CaHa	CHN	CN
	W1	W2	W3	W4	W5					
With None missing	0.07	0.10	0.18	0.27	0.38	0.036	0.051	0.034	0.067	0.04
With State One missing	--	0.11	0.20	0.31	0.37	0.042	0.059	0.041	0.072	0.054
With State Two missing	0.10	--	0.26	0.26	0.36	0.047	0.056	0.054	0.074	0.052
With State Three missing	0.06	0.19	--	0.31	0.35	0.066	0.067	0.079	0.082	0.08
With State Four missing	0.23	0.0	0.26	--	0.27	0.08	0.074	0.084	0.102	0.091
With State Five missing	0.0	0.0	0.0	0.40	--	0.202	0.15	0.17	0.166	0.155

From Table 2.7, it is seen that the algorithm produces RDC Q-factors with nearly the same quality especially when the missing cluster has a low population, such as cluster one or two. Even with cluster three or four, its missing causes only a small deterioration in Q-factors. In all these cases, most of the contributions of the missing cluster are compensated by the weight adjustment of the remaining clusters or by assigning weight to a new cluster(s) that is formed by some noise conformations. However, when the missing cluster has an especially large population such as that of cluster five, the algorithm cannot recover the RDC Q-factors with nearly the same quality. The results of this test case thus clearly demonstrate the importance of having a full coverage of all the conformational states and that low Q-factors alone are not sufficient to provide full confidence in the completeness or correctness of an ensemble.

X-ray Ensemble and Experimental data

X-ray structures of the same protein but solved under different conditions are hypothesized to form a native state ensemble of that proteins (17). 68 X-ray structures of Ubiquitin with 100% sequence identity are taken from PDB. After considering the fact that multiple chains exist in some of the structures, a total of 143 different conformations are identified and used to form the Ubiquitin conformation ensemble. Table 2.8 lists all the PDB-ids along with their chain identifiers. To partition this ensemble into proper sub-ensembles and determine their relative populations, we follow the procedure described in the Materials and Methods section and find that 18 out of 143 crystal structures have a significant weight and are chosen as representative conformations. This

new ensemble of 18 crystal structures was then subjected to the splitting procedure and as a result, six clusters are identified. The rest of the 125 structures, one by one, are then tried to be merged into one of six existing clusters.

Table 2.8: PDB ids as well as chain identifiers of the 143 Ubiquitin X-ray conformations used in this work to form the Ubiquitin X-ray ensemble.

1AAR-A, 1AAR-B, 1CMX-B, 1F9J-A, 1F9J-B, 1NBF-C, 1NBF-D, 1OGW- A, 1P3Q-U, 1P3Q-V, 1S1Q-B, 1S1Q-D, 1TBE-A, 1TBE-B, 1UBI-A, 1UBQ- A, 1UZX-B, 1WR6-E, 1WR6-F, 1WR6-G, 1WR6-H, 1WRD-B, 1XD3-B, 1XD3-D, 1YD8-U, 1YD8-V, 2AYO-B, 2C7M-B, 2C7N-B, 2C7N-D, 2C7N- F, 2C7N-H, 2C7N-J, 2C7N-L, 2D3G-A, 2D3G-B, 2DX5-B, 2FID-A, 2FIF- A, 2FIF-C, 2FIF-E, 2G45-B, 2G45-E, 2GMI-C, 2HD5-B, 2HTH-A, 2IBI- B, 2J7Q-B, 2J7Q-D, 2JF5-A, 2JF5-B, 2O6V-A, 2O6V-C, 2O6V-E, 2O6V- G, 2OOB-B, 2QHO-A, 2QHO-C, 2QHO-E, 2QHO-G, 2WDT-B, 2WDT- D, 2WWZ-A, 2WWZ-B, 2WX0-A, 2WX0-B, 2WX0-E, 2WX0-F, 2WX1- A, 2XEW-A, 2XEW-B, 2XEW-C, 2XEW-D, 2XEW-E, 2XEW-F, 2XEW- G, 2XEW-H, 2XEW-I, 2XEW-J, 2XEW-K, 2XEW-L, 2XK5-A, 2ZCC-C, 2ZNV-C, 3A1Q-A, 3A1Q-D, 3A33-B, 3A9J-B, 3A9K-B, 3ALB-A, 3ALB- B, 3ALB-C, 3ALB-D, 3BY4-B, 3C0R-B, 3C0R-D, 3EEC-A, 3EEC-B, 3EFU-A, 3EHV-B, 3EHV-C, 3H1U-A, 3H1U-B, 3H7P-B, 3H7S-A, 3H7S-B, 3HM3-A, 3HM3-B, 3HM3-C, 3HM3-D, 3I3T-B, 3I3T-D, 3I3T-F, 3I3T-H, 3IFW-B, 3IHP-C, 3IHP-D, 3JSV-B, 3JVZ-X, 3JVZ-Y, 3JW0-X, 3JW0-Y, 3K9P-B, 3KVF-B, 3KW5-B, 3LDZ-E, 3LDZ-F, 3LDZ-G, 3M3J-A, 3M3J-B, 3M3J-C, 3M3J-D, 3M3J-E, 3M3J-F, 3MHS-D, 3NHE-B, 3NOB-B, 3NOB- C, 3NOB-D, 3NOB-E, 3NOB-F, 3NOB-G, 3NOB-H

The resulting conformation clusters along with their weights are given in Table 2.9. The cluster that contains the unbounded conformation of ubiquitin, 1UBQ, is found to have the largest weight of ~55%, while the second clusters, consisting exclusively of ubiquitin structures in complex with deubiquitinating enzymes, has the second largest relative population of ~29%.

Table 2.9: The six conformational clusters and their weights of the weighted X-ray ensemble. The conformations included in each cluster are listed by their PDB ids as well as chain identifiers.

Cluster	Final weight ± std	Composition
Cluster1	0.55 ± 0.03	1AAR-B, 1UBQ-A, 2C7M-B, 2C7N-H, 2QHO-A, 3EHV-C, 3M3J-A, 3M3J-E
Cluster2	0.29 ± 0.03	2G45-B, 2G45-E, 2HD5-B
Cluster3	0.064 ± 0.001	2DX5-B, 3KW5-B
Cluster4	0.043 ± 0.002	1YD8-V
Cluster5	0.027 ± 0.004	3HIU-A
Cluster6	0.026 + 0.001	1TBE-A

While we were working on this manuscript, one work was published in an early edition of PNAS (60). The work studied the native equilibrium dynamics of Ubiquitin and reported that the protein conformation was exceptionally stable with ∼70% of populated states about 0.5 Å RMSD away from the native state 1UBQ while ∼20% of the populated states showed a conformational switch in Asp52/Gly53/Glu24 residues, referred to as ``switched'' conformer and the remaining ∼10% had partially frayed alpha helix at the C-terminus (60). Our results as shown in Table 2.9 agree with their findings extremely well. In addition, another recent study of conformational states of ubiquitin found the presence of an alternative conformer in complex with deubiquitinating enzymes. In the authors' own words, ”This alternative conformer is likely to have functional significance, because the Asp52/Gly53/Glu24 switched conformer is also found in structures of ubiquitin, ubiquitin aldehyde, or diubiquitin in complex with deubiquitinating enzymes (e.g., PDB entries 2G45, 2HD5, 2IBI, 1NBF, 3I3T, 3IHP, 3NHE, 3MHS, and proximal ubiquitin of 2ZNV, which are all discussed further below). In contrast, the un-switched conformer is seen in essentially all other ubiquitin structures, including the previous structures for monomeric ubiquitin, di- and tetra-ubiquitin, and

complexes with other kinds of enzymes" (61). Our method not only identifies this special conformation state of ubiquitin (the 2nd cluster in Table 2.9), but also assigns it an accurate relative population. Many of the PDB entries for ubiquitin in complex with deubiquitinating enzymes are selected and grouped together by our algorithm to form cluster 2, a cluster consisting exclusively of ubiquitin structures in complex with deubiquitinating enzymes.

The remaining four clusters contain the following structures. Cluster 3 consists of 2DX5 and 3KW5. 2DX5 is a structure of ubiquitin in complex with mouse EAP45-GLUE domain. 3KW5 contains a structure of ubiquitin in complex with ubiquitin carboxy terminal hydrolase L1. Cluster 4 contains 1YD8, a structure of ubiquitin in complex with human GGA3 GAT domain. Cluster 5 contains 3H1U, a structure of ubiquitin in complex with cadmium ion. Lastly, cluster 6 contains 1TBE, a structure of ubiquitin in the form of tetraubiquitin. These four clusters all together have a relative population of about 15%. Figure 2.2 shows the final structure ensemble (center) as well as individually, a representative conformation from each cluster (panels a to e).

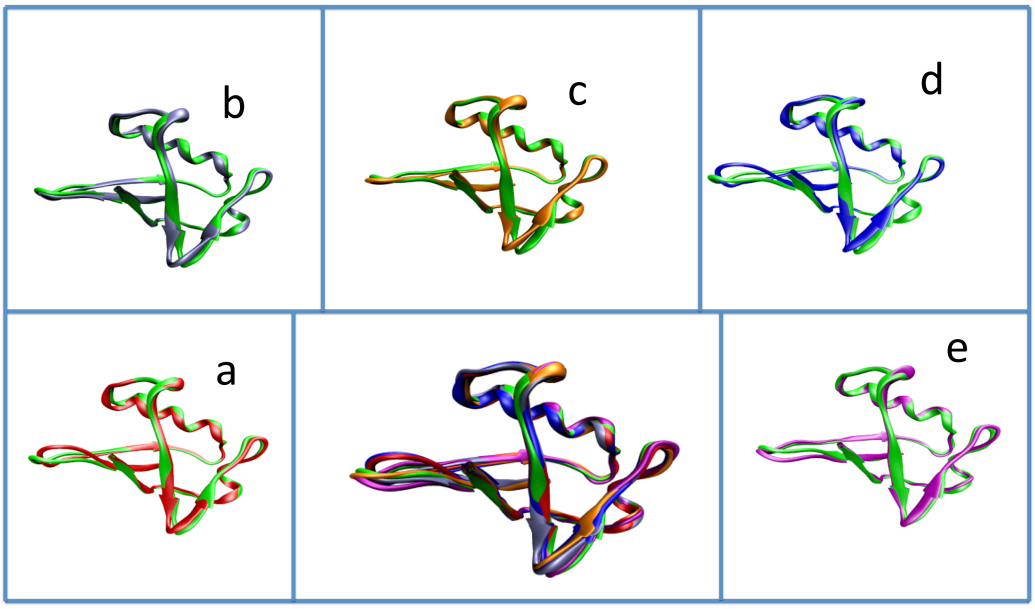

Figure 2.2: The final weighted X-ray ensemble that consists of six clusters (see Table 2.9) and representative conformations for each cluster. Center - All the structures overlaid onto one another, 1UBQ-A (cluster 1)– green, 2G45-E (cluster 2)– red, 2DX5-B (cluster 3) – ice blue, 3HIU-A (cluster 4) – orange, 1YD8-V (cluster 5) – purple and 1TBE-A (cluster 6) – blue.

Panels *a through e* compares 1UBQ-A with 2G45-E (red), 2DX5-B (ice blue), 3HIU-A (orange), 1TBE-A (blue), and 1YD8-V (purple) respectively.

Cross validation.

The individual Q-factors obtained for the different bond vectors are shown in Table 2.10 for the weighted X-ray ensemble along with other recently derived ensembles. By partitioning the ensemble into six sub-ensembles (represented by the clusters) and assigning them relative populations, the Q-factors of all the individual bond vectors are significantly lowered. Remarkably, the cross validation Q-factor, that of CAHA, is also lowered from 0.161 to 0.145 for the weighted X-ray ensemble. This significant

improvement in Q-factors further confirms the validity of clustering and relative population assignments discussed above

Table 2.10: Q-factors of the different bond vectors of the weighted X-ray ensemble as well as some other ensembles. CaHa is used for cross-validation

NH	CaC	CaHa	CN	CHN	Description
0.122	0.097	0.145	0.088	0.186	Weighted X-ray
0.184	0.108	0.161	0.099	0.228	Unweighted X-ray
0.071	0.118	0.069	0.138	0.188	EROS
0.213	0.118	0.128	0.138	0.234	EROS reprotonated
0.066	0.140	0.167	0.096	0.182	ERNST
0.180	0.141	0.177	0.095	0.207	ERNST reprotonated
0.244	0.180	0.236	0.171	0.266	1UBQ
0.114	0.105	0.084	0.120	0.163	1D3Z
0.231	0.175	0.196	0.233	0.281	MUMO(PDB id: 2RN2)

In contrast to those of the single structure representation, residue-wise Q-factors of unweighted and weighted ensembles are shown in Figure 2.3. It is seen that for most of the residues, the unweighted ensemble has lower Q-factors than the single structure, 1UBQ, while the weighted ensemble further lowers the Q-factors.

The Q-factor results of the weighted X-ray ensemble are on the par even with 1D3Z, NMR ensemble that was determined using RDC as one of the restraints and are noticeably better than MUMO, a Ubiquitin ensemble computationally determined using NOE and order parameters as constraints. When compared with EROS and ERNST (34), the weighted X-ray ensemble falls short especially in the NH and CAHA datasets. However, as was pointed out in (62), the conformations in EROS ensemble may have incorrect geometry. Indeed both, reprotonated EROS and reprotonated ERNST display much higher Q-factor values, see Table 2.10.

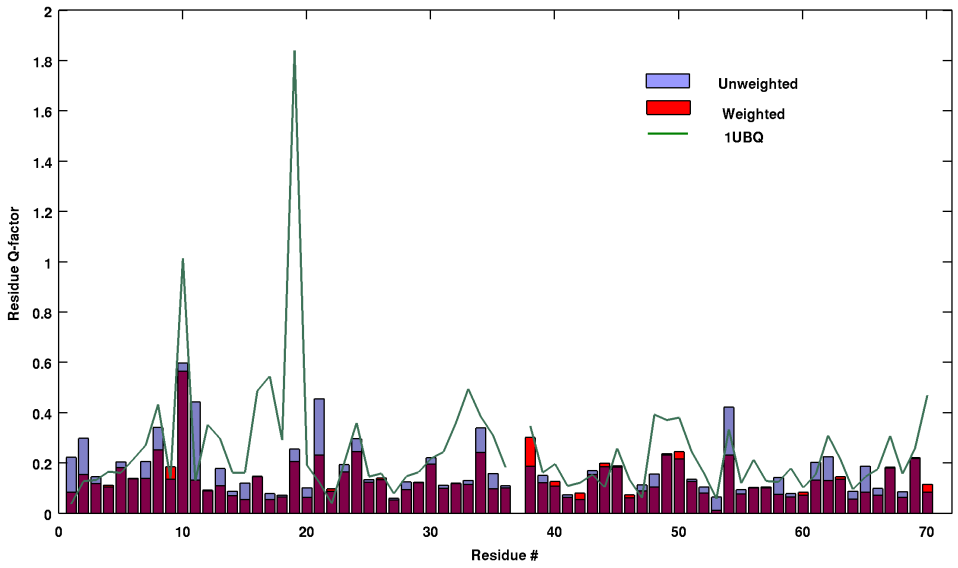

Figure 2.3: Residue-wise Q-factors from 1UBQ, the unweighted and weighted X-ray ensemble. The unweighted Q-factors are plotted in blue bars, the weighted Q-factors in red bars, and the Q-factors obtained from 1UBQ are plotted in a green line. The common region between the unweighted and weighted is colored maroon.

Uncertainty in weight assignments.

Uncertainty in weight assignments can be computed when they are multiple datasets (see Materials and Methods). In the case of ubiquitin, there are 24 NH RDC data sets along with two multi-vector RDC datasets. The whole datasets are partitioned into two subsets such that each subset contains one multi-vector dataset along with an equal proportion of NH RDC datasets. Weights obtained from each subset are compared and their standard deviations are used for representing the uncertainty in weight assignments (see Table 2.9).

Effect of weighting on conformational features of ensemble:

In addition to improving the reproduction of experimental data, weighting alters conformational properties of the ensemble. One of the interesting features of Ubiquitin structure is the presence of a "switched" conformation, which is hypothesized to have a biological function(61). The dihedral angles ϕ of residue 53 and ψ of residue 52 play an important role in facilitating the switch. While ϕ_{53} and ψ_{52} of the "switched" conformation exists in the range of ~100^{O} and ~130^{O} respectively, the same two dihedrals are in the range of ~-90^{O} and ~-50^{O} respectively for the unswitched conformation such as in 1UBQ. We look into the changes in the population distributions of these dihedral angles before and after reweighting and the results are presented in Figure 2.4.

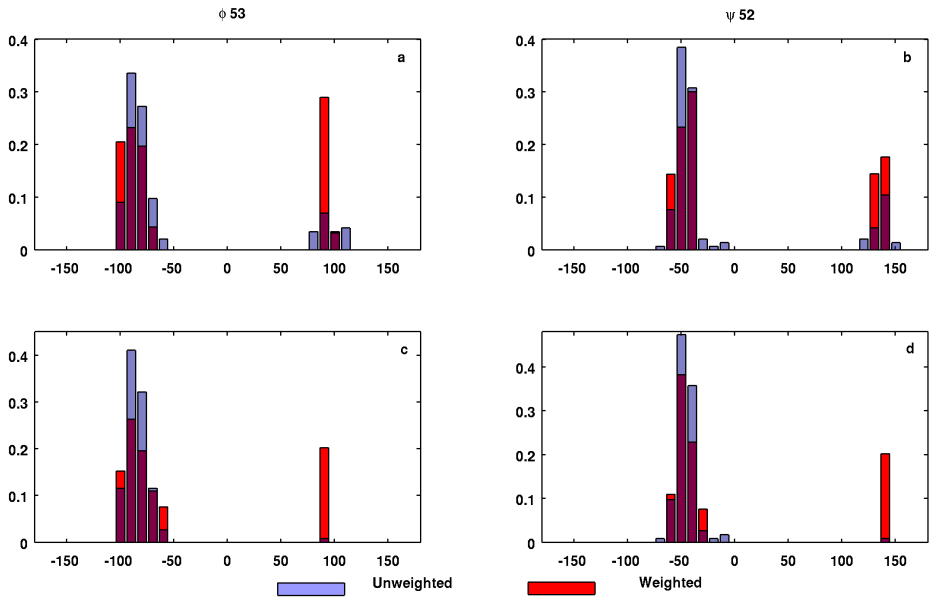

Figure 2.4: Effects of weighting on the conformational features of X-ray ensembles. Panels a and b, show the population distributions of ϕ_{53} and ψ_{52} dihedral angles before (blue bars) and after (red bars) weighting of the X-ray ensemble. Panels c and d, show the same population distributions but for a modified X-ray ensemble whose "switched" conformations except one are all taken out (see the text). The common region between the unweighted and weighted is colored maroon.

In the first row of Figure 2.4 (panels a and b) are shown the differences in the population distributions of dihedral angles ϕ_{53} and ψ_{52} between before and after reweighting the 143-conformation X-ray ensemble. The reweighting significantly alters the dihedral angle distributions of ϕ_{53} and ψ_{52} shifting much of the populations from being around the unswitched conformation to the switched conformation. The reweighting also reduces the overall ranges of the dihedral angle distributions and makes the two population peaks narrower and sharper. To further demonstrate how strong an effect weighting can have on dihedral angle distributions, all the switched conformations except 2G45 (chain E, a "switched" conformation) are removed from the 143 conformations. The population of the switched conformation in this reduced ensemble before weighting is now less than 1%. The second row of panels (c and d) of Figure 2.4 show the difference in population distributions upon reweighting this ensemble. As is seen, after reweighting the population of the "switched" conformation increases dramatically from less than 1% to as high as 20%.

Application on a computationally-determined ensemble:

In the recent years many ubiquitin ensembles have been determined computationally. ERNST, standing for ensemble refinement for native proteins using a single alignment tensor, was refined using NOEs and RDCs (34). ERNST does a very good reproduction of the experimental RDCs as seen from the low Q-factors in Table 2.11. But as with EROS, there is a significant increase in Q-factors once the ensemble is reprotonated using standard tools. Though the validity of reprotonation is debatable, such a significant increase in Q-factors could be due to the covalently incorrect placement of

hydrogen atoms (62). Therefore, we choose to apply our weighting algorithm to the reprotonated ERNST ensemble instead to avoid introducing into weights errors due to incorrect covalent geometry. The Q-factors obtained after weighting the reprotonated ERNST ensemble are shown in Table 2.11. From the table it is seen that though weighting lowers the Q-factors, the decreases are mostly quite nominal and the new Q-factors are not as good as those of the weighted X-ray ensemble.

Table 2.11: Q-factors of the different bond vectors of the ERNST ensembles. CaHa is used for cross-validation. ERNST reprotonated is the same as ERNST except the hydrogen atoms are replaced using standard geometry. In the last row, the reprotonated ERNST is first enhanced with a switch conformation 2G45-E before the population reweighting is applied.

NH	CaC	CaHa	CN	CHN	Description
0.066	0.140	0.167	0.096	0.182	ERNST
0.180	0.141	0.177	0.095	0.207	ERNST reprotonated
0.147	0.145	0.178	0.098	0.190	Weighted reprotonated ERNST
0.123	0.112	0.132	0.093	0.172	Weighted (reprotonated ERNST + 2G45-E)

A close look at dihedral angle distributions of ϕ_{53} and ψ_{52} of the ENRST ensemble as we did to the X-ray ensemble in Figure 2.4 reveals the reason. Figure 2.5 (panels a and b) shows that ENRST does not sample the "switched" conformational state at all and all the conformations have ϕ_{53} and ψ_{52} angles similar to 1UBQ. To assess the importance of "switched" conformational state, we add 2G45-E (a representative switched conformation) to ERNST ensemble and then reweighted it. Interestingly, the Q-factors now improve significantly (see Table 2.11) and reach to a level similar to the weighted X-ray ensemble. Moreover, the switched conformation (2G45-E) is assigned to a relative population of 0.30, which is highly similar to the weight of the "switched"

conformation in the weighted X-ray ensemble (which is 0.29). This is remarkable since it shows that common conformational features emerge after reweighting even though the two ensembles to which the reweighting scheme has been applied are rather different. In Figure 2.5, panels a and b show the population distributions of the ϕ_{53} and ψ_{52} before and after weighting of the reprotonated ERNST ensemble, while panels c and d show the distributions of the same dihedral angles of the same reprotonated ERNST ensemble after a switched conformation is added to it. By comparing between the dihedral angle distributions in Figures 2.4 and 2.5, it is seen that while ERNST (reprotonated) itself does not have similar properties as the weighted X-ray ensemble, the weighted ERNST + 2G45-E (see panels c and d of Figure 2.5) shows highly similar conformational properties to the weighted X-ray ensemble (see panels a and b of Figure 2.4) especially with respect to the dihedral angle distributions of ϕ_{53} and ψ_{52}.

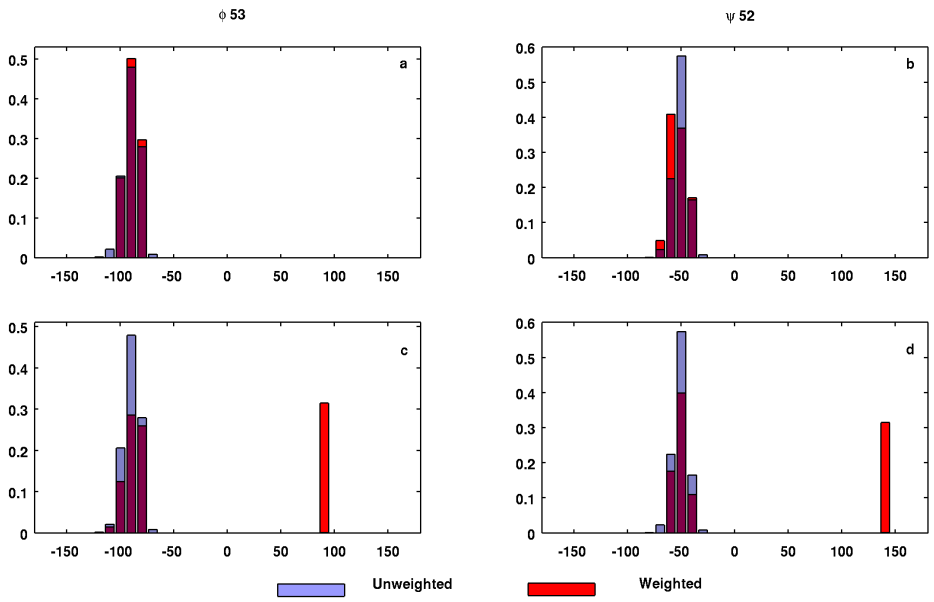

Figure 2.5: Dihedral distributions of ϕ_{53} and ψ_{52}in the ERNST ensembles. Panels a and b, show the population distributions of ϕ_{53} and ψ_{52} dihedral angles before (blue bars) and after (red bars) weighting of the ERNST ensemble. Panels c and d, show the populations of the same dihedral angles before (blue bars) and after (red bars) weighting of an enhanced ERNST ensemble (with 2G45-E, a "switched" conformation, added). The common region between the unweighted and weighted is colored maroon.

2.5 Discussion and Conclusions

Proteins are dynamic molecules and even the native state of a protein is not a single static structure but spread over a broader region of the conformation space. As a result, for many proteins, an ensemble of conformations provides a better depiction of the native states.

In this work we present a method to improve ensembles and their ability to depict the native states. The method works by identifying conformation states within an ensemble and assigning appropriate relative populations, or weights, to them. Each of

these conformation states is represented by a sub-ensemble formed by a subset of the conformations.

Our results demonstrate that such weight assignment is feasible and the weights are significant. Since the weights are computed by least squares fitting to the experimental RDC data, one may naturally question the significance of the weights. Are the weights significant and physically meaningful? Or are they merely a result of over-fitting to the noise in the experimental data? To address this concern, we design a sensitive measure to recognize the onset of over-fitting and finish the weight assignment before over-fitting starts to occur. Lastly, the significance of the weights is further examined and verified by cross validation.

The method presented in this work uses experimental RDC data as constraints to assign relative populations to conformations within an ensemble. In order for this method to succeed, what is the requirement on the ensemble and its conformations? Our results indicate the following:

- Undersampling in conformation states, where some conformation states are represented by few conformations, does not hinder weight determination. Experimental structures of the same protein obtained under different conditions or bound states have been suggested to form a native state ensemble of the protein (17). Such a native state ensemble may cover all the important conformation states of the protein, but not necessarily proportionally, and some of the states

may be severely undersampled. As seen from case I and II, undersampling does not hinder weight assignment and our algorithm can be readily applied to determine the relative populations.

- Noise conformations in an ensemble that do not represent any conformation states can be mostly filtered out. In case II, we create a situation where the ensemble contains a cluster of conformations that do not belong to any conformation state. Case III represents another situation where each conformation state is mixed with a large amount of noise conformations. The presence of noise conformations may make it difficult to identify conformation states, or to separate conformations representing a conformation state from those that do not. However, test results show that our method is able to effectively filter out most of the noise conformations.
- While cases I to III show that given an ensemble with good coverage and completeness, the weighting algorithm is able to identify the clusters and assign them with proper weights and thus lower the Q-factors, case IV indicates the converse is not necessarily true: low Q-factors do not necessarily mean that an ensemble is of good quality. Therefore, cautions must be taken in future ensemble determination and assessment. Measures other than Q-factors are needed to check the quality of computer-generated ensembles. It is not clear what these measures are, but their discovery and identification are going to be critical to the field's progress.

We apply our method to a Ubiquitin ensemble of 143 conformations and identify six conformation states. The two most populated conformation states, one of which represents the conformation state near the free state of ubiquitin while the other the ``switched'' conformer, match closely with conformation states identified by other studies. The relative populations assigned to these two states by our method, agree extremely well with the findings by Shaw's group through long MD simulations(60). The validity of such conformation state identification and weight assignments are further confirmed by significant improvement in Q-factors and cross-validation.

We apply our method also on a computationally derived ensemble, ERNST, which was refined against RDCs and NOEs. Even though the reproduction of experimental data, RDCs in this case, worsens after reprotonation, we are able to significantly improve the Q-factors by augmenting the ensemble with a switched conformation and reweighting. In doing so we observe the emergence of common dihedral angle distributions in both the augmented ERNST ensemble and X-ray ensemble.

The method presented in this work can be applied to other proteins to identify conformation states and assign relative populations, provided that sufficient RDC data exist. A good question to ask is how much RDC data is required for weight assignment? And what type of RDC data is required, NH RDCs, multi-vector RDCs, or both? We plan to study this in future work.

The number of conformation states recognized by our method can be used to guide the selection of ensemble size in ensemble determination. Most ensemble determination methods try out different sizes for replica ensembles, usually from 1, 2, 4, 8, up to 16. The method presented here provides an informed estimation of the right size for the ensemble. Since the method requires an ensemble as a starting point, it could be applied alternatively with an existing ensemble determination method until the process converges and a right ensemble size is identified. Our results strongly suggest that relative weights, instead of the default equal-weights, should be considered as parameters in ensemble determination.

2.6 Acknowledgements

Funding from National Science Foundation (CAREER award, CCF-0953517) is gratefully acknowledged. The authors would also like to thank the two anonymous reviewers for their insightful comments.

2.7 Appendix:

Calculation of RDCs from single structure

Given a 3D structure of a protein, the RDC D_{ij} can be expressed using the molecular frame. First, the elements of Saupe matrix is defined as:

$$S_{lm} = \langle \frac{3cos\beta_l cos\beta_m - k_{lm}}{2} \rangle \quad (4)$$

where β_l denotes the orientation of the l-th molecular axis with respect to the external magnetic field. The RDC D_{ij} can be reformulated in the molecular frame as:

$$D_{ij} = \frac{-\mu h r_i r_j}{(2\pi r)^3} \left(\alpha_y^2 - \alpha_x^2;\ \alpha_z^2 - \alpha_x^2;\ 2\alpha_{x\alpha_y}; 2\alpha_{x\alpha_z}; 2\alpha_{y\alpha_z} \right) \begin{pmatrix} S_{yy} \\ S_{zz} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} \quad (5)$$

where α_x, α_y, and α_z are the cosines of the angles between the bond vector of the two nuclei and the x, y, and z axes of the molecular frame. Let α_{xk}, α_{yk}, and α_{zk} represent the k-th α_x, α_y, and α_z. When all the bond vectors are considered, we have the following formula:

$$D_{\text{exp}} = \left(\frac{-\mu h r_{\text{i}} r_{\text{j}}}{(2\pi r)^3}\right) \begin{pmatrix} \alpha_{\text{y,1}}^2 - \alpha_{\text{x,1}}^2 & \cdots & 2\alpha_{\text{y,1}}\alpha_{\text{z,1}} \\ \vdots & \ddots & \vdots \\ \alpha_{\text{y,N}}^2 - \alpha_{\text{x,N}}^2 & \cdots & 2\alpha_{\text{y,N}}\alpha_{\text{z,N}} \end{pmatrix} \begin{pmatrix} S_{\text{yy}} \\ S_{\text{zz}} \\ S_{\text{xy}} \\ S_{\text{xz}} \\ S_{\text{yz}} \end{pmatrix} \quad (6)$$

where D_{exp} is the experimental RDCs and N is the total number of data points. Equation 6 can be rewritten in the following matrix form:

$$D_{\text{exp}} = cAS \quad (7)$$

where c is the constant $\frac{-\mu h r_{\text{i}} r_{\text{j}}}{(2\pi r)^3}$ and A is the Nx5 matrix in equation 6 and S is the 5x1 vector. Optimal S and thereby D_{calc} (i.e., the calculated RDCs) can be computed by singular value decomposition using Moore-Penrose pseudoinverse of matrix A:

$$S = A^{-1} D_{\text{exp}} \quad (8)$$

$$D_{\text{calc}} = AA^{-1} D_{\text{exp}} \quad (9)$$

Residual dipolar coupling (RDC) calculation from an ensemble

The RDC calculation method for a single structure can be extended to take ensemble averaging into account so that the ensemble D_{calc} can be obtained. First let us consider the assumption that all structures have equal contributions toward the

experimental RDC: D_{exp}. When an ensemble with equal weights is considered, we have the following formula:

$$\left(\frac{A_1}{n}+\frac{A_2}{n}+\cdots+\frac{A_k}{n}+\cdots\frac{A_n}{n}\right)S = D_{\text{exp}} \tag{10}$$

where A_{k} is the A matrix obtained from the k-th structure in the ensemble. S can be obtained from the following equation:

$$S = \left(\frac{A_1}{n}+\frac{A_2}{n}+\cdots+\frac{A_k}{n}+\cdots\frac{A_n}{n}\right)^{-1} D_{\text{exp}} \tag{11}$$

Strictly speaking, the Saupe matrix might vary for different conformations of the protein. In this work we assume the same Saupe matrix for all the conformations. This assumption is reasonable especially for proteins that make only small conformation changes, as is the case with Ubiquitin.

Now let us consider the case that structures in an ensemble have different populations and thus different amounts of contributions toward the experimental observations D_{exp}. Therefore, weights (representing the relative populations) are given to different structures and the following formula is used to represent the combination:

$$(w_1A_1 + w_2A_2 + \cdots + w_kA_k + \cdots w_nA_n)S = D_{\text{exp}} \tag{12}$$

where n is the total number of structures and w_k and A_k are respectively the relative population (or weight) and A matrix of the k-th structure. Thus, S can be obtained from the following formula:

$$S = (w_1A_1 + w_2A_2 + \cdots + w_kA_k + \cdots w_nA_n)^{-1}D_{\text{exp}} \qquad (13)$$

Our problem is thus to find the optimal relative populations for the structures in the ensemble so that the experimental RDCs are best reproduced.

Least Squares Fitting Algorithm

The iterative least squares fitting algorithm to a single RDC data set

Iterative Least Squares Fitting ($[A_1\ A_2\ \ldots\ A_n]$, D_{exp})

for i = 1 to n do

    new_weights(i) = 1/n

end for

repeat

    old weights = new_weights

    A = old_weights(1)*A_1 + ... + old_weights(n)*A_n

    S = pseudo_inverse(A) * D_{exp}

    AS = [A_1S A_2S . . . . A_nS]

    new_weights = non_negative_least_squares(AS, D_{exp})

Until old_weights and new_weights converge.

return new_weights

The iterative least squares fitting algorithm to multiple RDC data sets

Iterative Least Squares Fitting Multiple RDCs ([A_1 A_2 . . . A_n], [D_1, D_2 . . . D_m])

for i = 1 to n do

new_weights(i) = 1/n

end for

repeat

old weights = new_weights

A = old_weights(1)*A_1 + ... + old_weights(n)*A_n

for i = 1 to m do

S(i) = pseudo_inverse(A) * D_i

AS(i) = [A_1S(i) A_2S(i) . . . . A_nS(i)]

end for

AS_all = $\begin{pmatrix} AS(1) \\ AS(2) \\ \vdots \\ AS(m) \end{pmatrix}$

D_all = $\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_m \end{pmatrix}$

new_weights = non_negative_least_squares(AS_all, D_all)

Until old_weights and new_weights converge.

return new_weights

CHAPTER 3. ENSEMBLES OF A SMALL NUMBER OF CONFORMATIONS WITH RELATIVE POPULATIONS

A paper under review by Journal of Biomolecular NMR

Vijay Vammi and Guang Song

3.1 Abstract

In this work, we propose a new way to represent protein native states, using Ensembles of a Small number of conformations with relative Populations, or ESP in short. Using Ubiquitin as an example, we show that using a small number of conformations can greatly reduce the potential of overfitting and assigning relative populations to protein ensembles can significantly improve their quality. To demonstrate that ESP is an excellent alternative to represent protein native states, we compare the quality of two ESP ensembles of Ubiquitin with several well-known regular ensembles or average structure representations. Extensive amount of significant experimental data are employed to achieve a thorough assessment. Our results demonstrate that ESP ensembles, though much smaller in size comparing to regular ensembles, perform equally or even better sometimes in all four different types of experimental data used in the assessment, namely, the residual dipolar couplings (RDCs), residual chemical shift anisotropy, hydrogen exchange rates, and solution scattering profiles. This work underlines the significance of having relative populations in describing the native states.

3.2 Introduction

Proteins are dynamic molecules and often occupy multiple conformational states in their native states. The functional behavior of a protein is thus best understood from the distribution and dynamic transition among these conformational states that form the native state ensemble (1, 5, 6, 15, 42).

Nuclear Magnetic Resonance (NMR) experiments have played a pivotal role in capturing the dynamics of proteins in their native states. Data obtained from NMR experiments have been used as constraints in recovering the underlying structures or ensembles. In that process, two different refinement schemes are routinely followed:

i). *Average structure representation*: In this scheme, a single structure is used to explain all the observed experimental data. For Ubiquitin, one of the most studied proteins, a single structure has been shown to be sufficient in reproducing most experimental data (29, 30). But it was also pointed out that average structure representations, due to the lack of structural variance, cannot fully capture the underlying dynamics (31, 32). This representation becomes less complete when the studied protein occupies multiple distinct sub-states, since the refinement protocol would be over-restrained (under-fitting)(33).

ii). *Ensemble representation*: In this representation, an ensemble of conformations is used to explain the experimental data. In the case of Ubiquitin, there has been a number of recent work aimed at determining an ensemble of conformations for

the protein, such as MUMO(33), EROS (23) and ERNST (34). All of these ensembles are shown to represent the dynamics well but there is little confidence that any given conformation within the ensemble truly belongs to the native state ensemble, since the ensemble might be under-constrained or over-fitted (9, 35).

In this work, we propose a third representation,

iii). *Ensembles of a Small number of conformations with relative Populations or ESP in short:* In our recent work (63), we showed that the conformation space could be represented by far fewer conformations than the aforementioned ensemble representations and the conformations could be clustered into conformation states and these conformation states could be assigned relative populations, corresponding to their Boltzmann weights. The advantage of using ESP over an average structure is that it overcomes underfitting. The advantage of using ESP over using an ensemble with hundreds of conformations is that it minimizes overfitting. ESP uses a much smaller number of conformations than regular ensembles.

The objective of this work is to establish ESP as a better representation for describing the native states of a protein. To demonstrate that ESP ensembles are indeed of high quality and minimize overfitting, we resort to a series to significant experimental data that are not used in the determination of these ensembles, and show that ESP ensembles, though having a much smaller number of conformations, are able to

reproduce these experimental data equally well or even better sometimes and with less overfitting. Weighted ensembles had been successfully used in modeling unfolded protein conformational ensembles (39, 64) and was considered also in loop modeling (65), but they are usually not used in determining native state protein ensembles.

Though cross-validation using a subset of the data points that were left out during the ensemble determination stage has been commonly used, unused experimental data of different types present an even better resource for assessing the quality of the ensembles since they are even more unbiased. Since all of the aforementioned ensembles, namely, MUMO, EROS, and ERNST, use NOEs or RDCs as constraints in their construction, experimental data on Residual Chemical Shift Anisotropies (RCSA), amide exchange reactivities, and solution scattering profiles are employed in this study for cross-validation.

Our ensemble representation with relative populations could be thought of as an intermediate scheme between the two refinement schemes aforementioned: *average structure representation* or *ensemble representation*. Both representations have strengths and weaknesses. Average structure representation is the simplest in form but lacks structure variance, while ensemble representation captures the dynamics of the conformation space well but may suffer the problem of over-fitting and there is little confidence that any given conformation within the ensemble truly belongs to the native stateensemble. The advantage of ESP representation is that it has a very limited number of conformation states whose relative populations are rigorously determined (63) without

over-fitting. Consequently, there is high confidence on the validity of these conformation states.

3.3 Materials and Methods

3.3.1 Ensembles of a Small number of conformations with Relative Populations (ESP):

Two ESP ensembles were reported in our previous work (63) and will be used in this work as example ESPs.

a). *Weighted X-ray ensemble*: X-ray conformations resolved in different conditions have been shown to form a native state ensemble (17). In our previous work (63), 143 such structures of Ubiquitin were collected from PDB (10) to form an unweighted X-ray ensemble. After applying the weighting protocol, 16 of these structures were selected to form the weighted X-ray ensemble and six conformational states were identified (63). The weights assigned to the conformational states are in agreement with what was found in the 1 μs equilibrium simulation conducted by Shaw's group (66). The conformational state adopted by Ubiquitin when bound to de-ubiquitinating proteins, also called the "switched" conformation (67, 68), was given a weight of ~0.30.

b). *Enhanced ERNST ensemble*: Besides the X-ray ensemble, our conformation weighting algorithm was applied to another computationally derived ensemble, ERNST (34) to produce an enhanced ERNST ensemble. After introducing a "switched" conformation to the ensemble and then assigning relative populations

to the conformations in the ensemble, it was found that the enhanced ERNST ensemble was able to reproduce experimental data in a comparable accuracy to the weighed X-ray ensemble. This enhanced ERNST ensemble contains one X-ray switched conformation and 35 conformations selected from the original ENRST ensemble that has 640 conformations.

In this work, these two ESP ensembles are compared with three regular ensembles determined for Ubiquitin: MUMO (33) (pdb-id: 2NR2), EROS (23) (pdb-id: 2K39), and ERNST (34) (pdb-id: 2KOX), as well as two NMR structures with pdb-ids 1D3Z (29) and 2MJB (30) and one crystal structure 1UBQ (69).

3.3.2 Residual Dipolar Couplings (RDC):

Residual dipolar coupling comes from the interaction of two nuclear spins (dipole-dipole) in the presence of the external magnetic field and is defined (20-22, 29) as:

$$D_{\{AB\}} = \sum_{i=x,y,z} -\frac{\mu h r_A r_B}{(2\pi r)^3} \cos^2 \emptyset_i \; A_{ii} \qquad (1)$$

where r_A and r_B are the nuclear magnetogyric ratios of nuclei A and B respectively, h is Plank's constant, μ is permittivity of space, r is the internuclear distance between the two nuclei, A_{ii} the principal moment of the alignment tensor and $\emptyset_i$ is the angle between the internuclear vector and i^{th} principal axis of the alignment tensor. The alignment tensor could be determined by fitting a single structure or ensemble to the experimental data.

Normally, the residual dipolar coupling reduces to zero because of isotropic tumbling. The anisotropic measurement can be obtained by the aid of various types of liquid crystalline media. Details regarding back-calculation of RDC's were given in the appendix of our previous work (63).

Experimental RDCs used in this work: The RDCs used to determine the weights for the X-ray ensemble and enhanced ERNST ensemble are given in (63), along with the codes assigned to them according to (55). The Q-factors reported in this work use the newly determined RDC dataset in Squalamine and Pf1 media (30), in addition to the Ottiger dataset used in the determination of 1D3Z (56).

3.3.3 Q-factor:

Q-factor is a commonly used measure of the agreement between the experimental and calculated RDCs and is defined as:

$$Q\text{-}factor = \frac{\sqrt{\sum\left(D_{calc} - D_{\exp}\right)^2}}{\sqrt{\sum\left(D_{\exp}\right)^2}} \qquad (2)$$

where D_{calc} is the calculated RDC and D_{exp} is the experimental RDC.

3.3.4 Residual Chemical Shift Anisotropy (RCSA):

Along with RDC's, chemical shifts also change upon shifting from an isotropic medium to an anisotropic medium (29, 70-72). The change is defined by:

$$\Delta\delta = \sum_{i=x,y,z} \sum_{j=x,y,z} A_{jj} \cos^2 \theta_{ij} \, \delta_{ii} \qquad (3)$$

where δ_{ii} is the principal moment of the chemical shift tensor, A_{jj} the principal moment of the alignment tensor and θ_{ij} is the angle between i^{th} principal axis of the chemical shift tensor and j^{th} principal axis of the alignment tensor. The alignment tensor used in RCSA back-calculations is generally the same as the one computed from RDCs using either a single conformation or an ensemble (63). More information regarding the relation between RDC and RCSA back-calculation of a conformation can be found in (70).

The experimental dataset of RCSA used in this work were reported in (29) along with the RDC dataset used for obtaining the alignment tensor. Magnitudes and orientations of the chemical shift tensors reported in (72) are used in this work.

3.3.5 Amide Hydrogen Reactivity:

Hydroxide catalyzed amide hydrogen rates were used as a measure to assess conformational distribution of various ensembles (73, 74). The experimental rate constants of amide hydrogen exchange depend not only on the solvent accessibility but

also on the chemical environment surrounding the amide hydrogen. Even rarely exposed amide hydrogen could therefore exhibit a high exchange rate if the chemical environment is conducive for such an exchange. This property makes amide hydrogen reactivity a very sensitive measure of the conformational distribution of the native states.

Poisson Boltzmann electrostatic calculations: The experimental exchange rate constants for all the backbone amide hydrogens of Ubiquitin were reported in the work by LeMaster et al (73). In this work, electrostatics calculations needed to predict the exchange rates of conformational ensembles are performed in a similar way to what was described in a previous work (74). Briefly, surface exposure of amide hydrogens in all the conformations belonging to the ensemble is computed using Naccess (75), using default values for the atomic radii and 1 Å for the radius of the probe sphere. For all the amide hydrogens that are not involved in any hydrogen bonding (computed using HBplus (76)) and have a surface exposure greater than 0.5 Å, Poisson-Boltzmann continuum electrostatic computations are done using Delphi (77). The CHARMM22 atomic charge and radius values (78) are used in the electrostatic computations. To make the comparisons feasible between different conformations of the ensemble, N-methylacetamide is added to the grid in such a way that the molecule is at least 16 Å away from any atom of the protein. The charge distribution of N-methylacetamide (or its anionic form) is taken from (73). Serines or threonines are mutated to alanine or α-aminobutyrate respectively before the electrostatic potential is computed.

Gauche side chain χ_1 conformers have remarkably low solvent exposure than their trans counterparts. To account for this, for every conformation, in addition to computing electrostatic potential in the original side chain configuration, a gauche χ_1 rotated side chain configuration also is used (whenever such a rotation was possible) (73). The side chain position with the higher exchange rate is used for further processing.

3.3.6 Solution Scattering Profile:

Small Angle X-ray scattering (SAXS) and wide angle X-ray scattering (WAXS) data encode the information about the shape and size of the bio-molecules in solution (79, 80). The observed intensities from X-ray scattering are sensitive to the overall conformational distribution of the protein and are being regularly used as complementary data to those obtained from NMR or X-ray crystallographic studies (81, 82). Predicting the scattering profiles from either single structure or an ensemble are most routinely done using the Crysol software package (83). Along with significantly improving the predictions, AXES (Analysis of X-ray scattering data for Ensemble of structures)(84) webserver, provides an easy method to predict such intensities from ensembles. The predicted intensities of all the ensembles or single structures reported in this paper are computed using a local version of AXES webserver, generously provided by Bax's group. The experimental SAXS/WAXS data used in this work are reported in (84). The agreement between the predicted and experimental scattering intensities is most commonly denoted by the chi value that is defined as:

$$chi = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{I_{exp}(q_i) - I_{calc}(q_i)}{\sigma(q_i)}\right)^2} \quad (4)$$

where I_{exp} and I_{calc} are the experimental and predicted scattering intensities at q_i with a error of σ_i and M is the number of observed scattering intensities.

3.4 Results and Discussion

3.4.1 Agreement with Experimental RDCs:

Table 3.1: Q-factors obtained for different bond types by different representations of Ubiquitin. The experimental RDCs used for computing these Q-factors consist of the newly obtained Squalamine and pf1 dataset (30) and the Ottiger's dataset(56), the latter of which was used in the refinement of 1D3Z and in the fitting of weights for the weighted X-ray and ERNST ensembles.

NH	CaC	CaHa	CN	CHN	Description
0.12	0.10	0.15	0.12	0.23	Weighted X-ray
0.18	0.11	0.16	0.12	0.26	Unweighted X-ray
0.10	0.14	0.18	0.11	0.23	ERNST (34)
0.12	0.12	0.15	0.12	0.22	Enhanced ERNST
0.07	0.12	0.07	0.14	0.19	EROS (23)
0.23	0.17	0.20	0.23	0.28	MUMO (33)
0.20	0.18	0.22	0.18	0.30	1UBQ (69)
0.11	0.10	0.08	0.12	0.16	1D3Z (29)
0.069	0.097	0.083	0.096	0.2	2MJB (30)

Table 3.1 lists the Q-factors obtained for different bond types using different representations of Ubiquitin. The RDC datasets used for computing these Q-factors consist of Ottiger's multi-vector dataset (56) and the newly obtained RDC dataset in Squalamine and Pf1 media (30). Weighted X-ray, ERNST, enhanced ERNST, EROS and 1D3Z used only the Ottiger's dataset (56) in their structure/ensemble refinement or

weighting, while the 2MJB ensemble used both datasets. Therefore the Q-factors reported here in Table 3.1 serve not as a complete cross-validation but as a comparison of these Ubiquitin representations regarding their ability to reproduce existing or new RDC data. It is worth pointing out that the two ESP ensembles, the weighted X-ray ensemble and the enhanced ERNST ensemble, are able to well reproduce the new RDC dataset (in Squalamine and Pf1 media) even though the dataset was not used in determining these two ensembles (63).

The RDC Q-factors obtained for bonds with hydrogen atoms (NH, CaHa, CHN) are highly sensitive to the positions of the hydrogen atoms. Allowing a certain degree of deviation from the ideal covalent geometry can lower the Q-factors significantly. It should be noted that no such optimization of hydrogen atom positions was applied to our weighted X-ray or enhanced ERNST ensemble, while it was to the other representations, whose refinement protocols allowed such deviations from the ideal covalent geometry to better fit experimental RDC data. Nevertheless, the two ESP ensembles have a comparable performance in RDC Q-factors to the other ensembles or average structures. Structure 2MJB gives the best RDC Q-factors, which is not surprising since it utilizes all the RDC data in its refinement process.

3.4.2 ESP ensembles give Better Agreements with Residual Chemical Shift Anisotropies (RCSA):

Table 3.2 compares the RMSDs between experimental and computed residual chemical shift anisotropies (RCSAs) for carbonyl carbons, nitrogens, and amide

hydrogens, using different Ubiquitin ensembles. Since chemical shift anisotropies were not used in determining any of the above structures or ensembles, they can serve as an unbiased dataset for assessing the accuracy of different structures or ensembles. From the table it is seen that weighted X-ray (an ESP ensemble) outperforms its unweighted counterpart in predicting RCSAs: the RMS values of all three atom types are significantly reduced (see Table 3.2, row 1 and 2). Except for a nominal increase in RMSD for amide hydrogens, enhanced ERNST (another ESP ensemble) also performs better than ERNST itself.

Table 3.2: RMSDs of residual chemical shift anisotropy (RCSA) as predicted by different representations of Ubiquitin. None of the adjustable parameters in the RCSA was modified while predicting the chemical shifts. Q_{NH} is the RDC Q-factor of the NH dataset that was used in obtaining the alignment tensor. The same alignment tensor was used in the RCSA computations.

Carbonyl	Nitrogen	Amide	Q_{NH}	Description
6.37	16.2	1.57	0.11	Weighted X-ray
6.87	17.3	1.61	0.17	Unweighted X-ray
10.7	16.0	1.53	0.06	ERNST
7.84	15.7	1.61	0.11	Enhanced ERNST
8.63	16.6	1.51	0.07	EROS
13.2	19.63	1.67	0.22	MUMO
13.1	18.6	1.68	0.18	1UBQ
8.59	14.17	1.47	0.10	1D3Z
7.71	15.39	1.50	0.07	2MJB

Similar to the sensitivity to hydrogen atom positions in RDC calculations, calculations of the chemical shift tensors of nitrogens and amide hydrogens , and thus their RCSA predictions, depend on the orientations of the amide bond vectors. Comparisons of RCSAs regarding these two atom types should thus be done cautiously and with this in mind. From Table 3.2, it is seen that both ESP ensembles outperform other representations in carbonyl carbon RCSA. While for nitrogens and amide

hydrogens, the performance of ESP ensembles is slightly worse than average structure representations but comparable to other ensemble representations.

In ideal situations, a refinement/weighting using RDC data would implicitly improve the RSCA predictions of the structure/ensemble as an optimization of the bond vector orientation by the RDC data also improves the chemical shift tensor orientation of the involved atoms (chemical shift tensor orientations of N and HN atoms depend upon NH bond vector orientation encoded in NH RDC data while those of Carbonyl atoms depend upon CN bond vector orientation provided by CN RDC data). However, noises in experimental RDC data along with errors in structure/ensemble models preclude such ideal situations. Consequently, RCSAs are considered mostly independent from RDC data and were commonly used in cross-validation for observables determined by RDCs (29).

Importance of the "switched" conformation: The "switched" conformation, represented by 2G45-E, was given a population weight of ~0.30 by our weighting protocol (63) in the enhanced ERNST ensemble. This weight was higher than the expected weight given in one previous work (85). Comparing ERNST without the "switched" conformation and that with (rows 3 & 4), the latter performs better, confirming the importance of the "switched" conformation.

3.4.3 ESP ensembles Reproduce Amide Exchange Rates Well:

Ensembles naturally incorporate backbone flexibility, potentially increasing the number of surface exposed amide hydrogens than an average structure representation. Therefore, in this section only ensemble representations of Ubiquitin are used for comparison. Figure 3.1 plots the orders of differences between experimental and predicted pKa values (both in log scale) by various ensemble representations of Ubiquitin. A single index, the squared sum of the deviations, is given to every ensemble in the figure to give an overall sense of the quality of the predictions. Only residues exposed significantly in the X-ray, MUMO, EROS and ERNST ensembles and having an experimental pKa value of ~ 5.0 or higher are shown. (Since a different program was used to compute surface accessibility, our pKa predictions differ from LeMaster and colleagues' computations for some of the residues (74)).

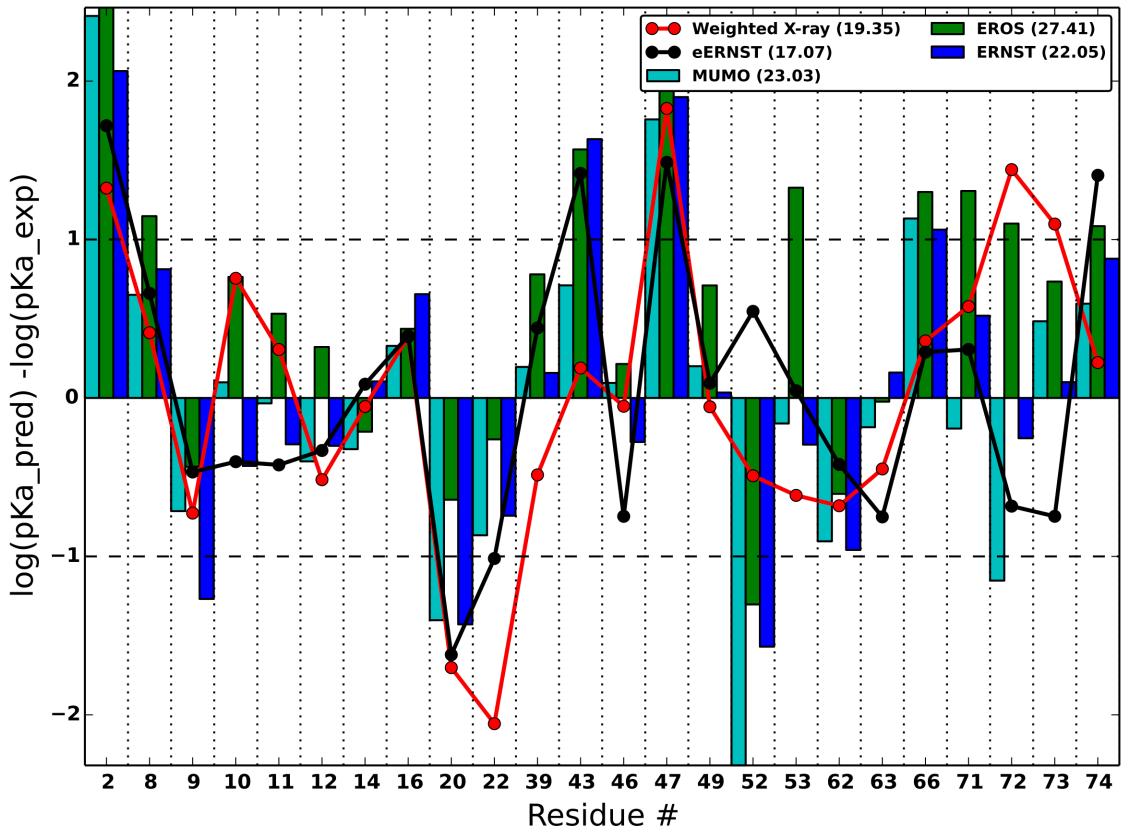

Figure 3.1: Residue-wise differences between experimental amide hydrogen reactivity data (in log scale) and those predicted by different representations of Ubiquitin. Only the hydrogens that are significantly exposed in all the ensembles (X-ray, EROS, ERNST, and MUMO) are shown here. A single index, the squared sum of the deviations, is given to every ensemble to give an overall sense of the quality of the predictions.

pKa predictions are not possible for residues 24, 31-36, 40-42, 48, 51 and 57-60, even though these residues exhibit high experimental exchange rates. This is because none of the ensembles has any surface exposed amides for these residues, which is needed to reproduce pKa values properly.

The weighted X-ray ensemble predicts the experimental pKa values quite well, having an overall performance better than all the unweighted ensembles. Likewise, the

enhanced ERNST ensemble also predicts the experimental pKa's better than the unweighted ensembles. Comparing with ERNST itself (blue bars), enhanced ERNST (weighted) performs significantly better on many residues.

In summary, both ESP ensembles (i.e., weighted X-ray and enhanced ERNST) perform well in predicting the experimental pKa values. This further validates that ESP ensembles are of high quality.

3.4.4 Solution Scattering Profile:

Solution scattering profiles are observed scattered intensities of X-rays that are collected as a function of the scattering vector q. Typically a q value of 0 to ~0.3 Å falls into the Small Angle X-ray scattering (SAXS) regime while the range for the Wide Angle X-ray scattering (WAXS) regime is ~0.1 to 2.5 Å. The information encoded in these two regimes along with the results obtained for different structure or ensemble representations of Ubiquitin are presented in the following two sections.

3.4.4.1 Small Angle X-ray Scattering (SAXS):

Scattering intensities observed at SAXS encode information about the overall size and shape of the molecule, radius of gyration (R_g) and other low-resolution information (86). Table 3.3 lists the chi value obtained by different representations of Ubiquitin.

Table 3.3: SAXS or WAXS Chi values obtained for different representations of Ubiquitin.

SAXS chi	WAXS chi	Ensemble
1.24	3.45	Weighted X-ray
1.28	3.65	Unweighted X-ray
1.49	3.75	ERNST
1.37	3.00	Enhanced ERNST
1.27	4.53	EROS
1.36	3.99	MUMO
1.04	4.87	1UBQ
1.17	3.40	1D3Z
0.84	4.98	2MJB

From Table 3.3 it is seen that, both weighted X-ray and ERNST ensembles have better chi values than their unweighted counterparts. The decreases in chi value confirm that conformations selected to form these two ESP ensembles and the weights assigned to them are meaningful. However, since SAXS data are of low resolution and are not the best data for validating ensembles, this should be taken only as a weak confirmation. Indeed, average structure representations (1D3Z, 1UBQ, or 2MJB) produce an excellent agreement with the experimental data, implying that at low resolution the native states of Ubiquitin appear to be mostly a single conformation.

Figure 3.2 plots the relative intensities (I_{exp}/I_{calc}) for different representations. While all the representations perform highly similarly at smaller values of q, at higher values of q (> 0.14 Å) single structure representations perform the best, followed by the weighted X-ray ensemble.

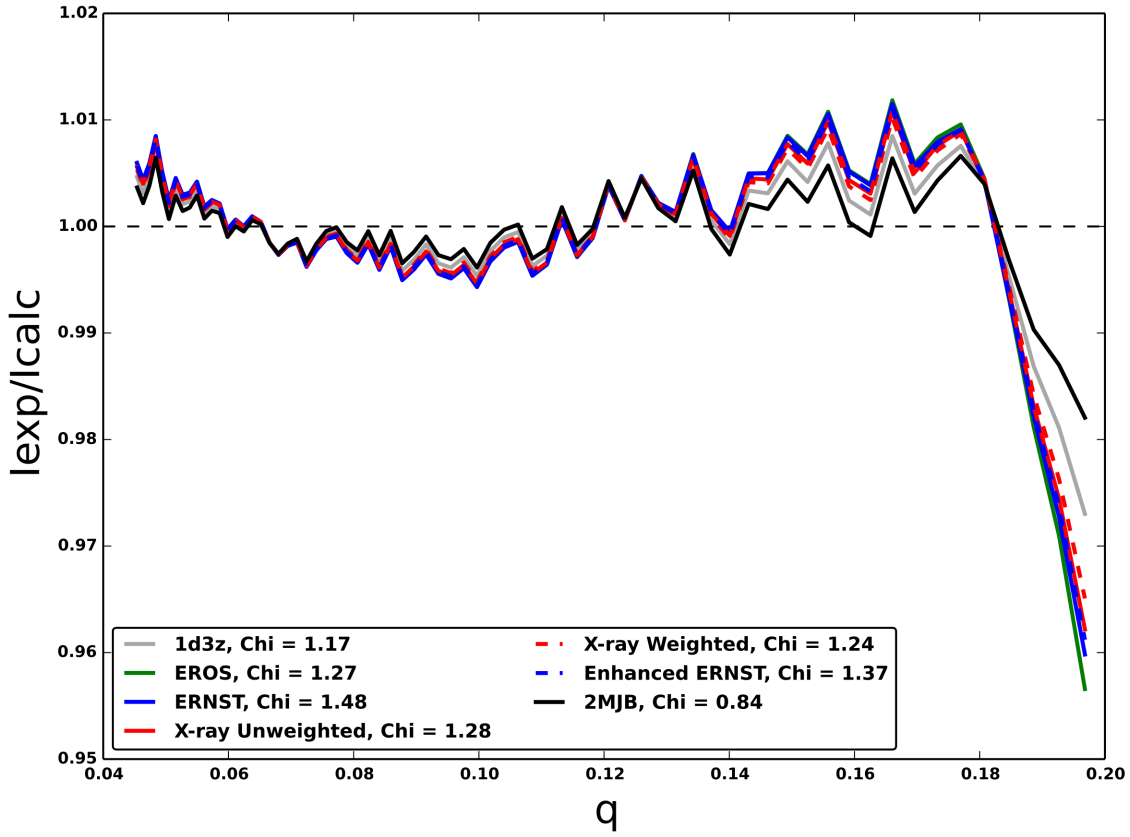

Figure 3.2: Relative intensities (I_{exp}/I_{calc}) as a function of q in the SAXS regime for different representations of Ubiquitin. The chi values obtained for different representations also are given.

3.4.4.2 Wide Angle X-ray Scattering (WAXS):

Scattering intensities observed at wider angles (higher *q*) encode information of higher resolution than SAXS but at the cost of potentially bringing in a higher noise level since the intensity of solution scattering also increases. Since data used in this analysis are limited to the range of *q* values that are less than 1.0 Å, the extent of this noise is limited. WAXS data are often used to validate structural models and to identify structural changes (86). Table 3.3 lists the WAXS chi values obtained for different representations of Ubiquitin.

Because of its much higher resolution, WAXS data is able to detect conformation state heterogeneity within the native state ensemble. Our first observation based on the WAXS results in Table 3.3 is that ensemble representations generally do better than the average structure representations (1UBQ and 2MJB). 1D3Z is an exception. Secondly, weighted X-ray and enhanced ERNST (the two ESP ensembles) are better than unweighted X-ray ensemble and ERNST ensemble respectively. Thirdly, though weighted X-ray (16 conformations) and weighted ERNST (36 conformations) have significantly fewer conformations than the unweighted X-ray (143 conformations) and ERNST (640 conformations), and the other ensembles such as EROS (116 conformations) and MUMO (144 conformations), these two ESP ensembles clearly outperform the other ensembles in WAXS chi values. This implies that ensemble sizes ought to be fairly limited to avoid overfitting, and that conformations in an ensemble should not be too spread out, and that having too many conformations makes an ensemble highly susceptible to overfitting. Put these together, it seems that the optimal way to represent the native states of a protein is to use i) an *ensemble,* of ii) a *small number* of conformations, and iii) with *relative populations*, as in ESP ensembles.

Figure 3.3 plots the detailed, relative intensities (I_{exp}/I_{calc}) computed from different Ubiquitin representations in the WAXS regime.

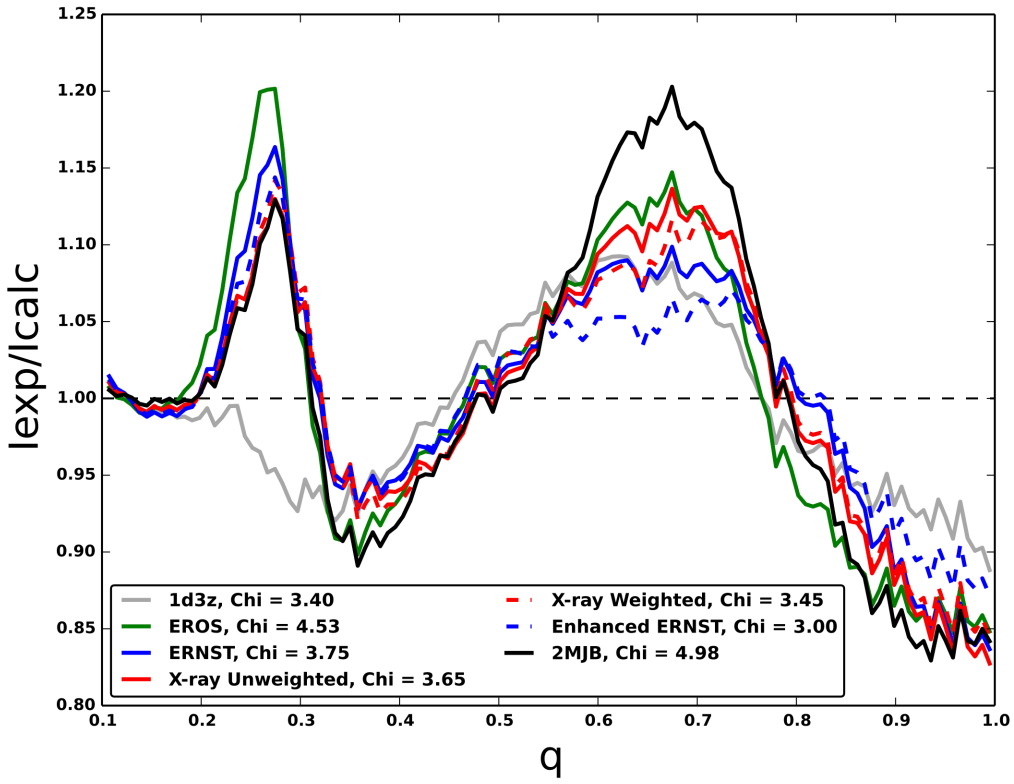

Figure 3.3: Relative intensities (I_{exp}/I_{calc}) as a function of q in the WAXS regime for different representations of Ubiquitin. The chi values obtained for different representations also are given.

3.5 Conclusions

In this work, by using Ubiquitin as example and extensive experimental data validations, we demonstrate that it is significant to assign relative populations to conformation ensembles and that ESP ensembles, though having a much smaller number of conformations, are of better quality than regular unweighted ensembles. Specifically, we carry out a thorough cross-validation of two ESP ensembles of Ubiquitin that were determined in an earlier work (63), namely, the weighted X-ray ensemble and the enhanced ERNST ensemble, and show that these two ensembles perform extremely well in all four different types of experimental data: the residual dipolar couplings (RDCs),

residual chemical shift anisotropy, hydrogen exchange rates, and solution scattering profile. This is not the case with other ensembles. For example, the MUMO ensemble, which performs well in predicting hydrogen exchange rates, does rather poorly in predicting RDCs. The ERNST or EROS ensemble does well in predicting RDCs but does not perform well in predicting hydrogen exchange rates or the residual chemical shift anisotropies. All these three ensembles (namely MUMO, EROS, and ERNST) do rather poorly in reproducing WAXS chi values. As a result, it is reasonable to conclude that the two ESP ensembles portray the Ubiquitin native states more accurately. Both ensembles reveal that there are six conformation states in Ubiquitin native states, two of which has dominating populations over the others. The conformation state with the largest population contains the unbounded conformation of ubiquitin, 1UBQ, while the one with the second largest population corresponds to the "switched" conformation, consisting exclusively of ubiquitin structures in complex with deubiquitinating enzymes (63).

Qualitatively speaking, the idea of having an ensemble with a small number of conformation states is advantageous. It both captures the dynamical nature of the native state (for which a single average structure is often insufficient to account for) and maintains a strong confidence on the validity of the conformation states. It is the most natural extension of the average structure representation. In contrast, confidence on any individual conformation that it truly belongs to the ensemble is elusive in regular Ubiquitin ensembles since they contain so many conformations and the removal of any single conformation hardly affects the ensemble. Consequently, these ensembles are highly susceptible to over-fitting.

3.6 Acknowledgement

Funding from National Science Foundation (CAREER award, CCF-0953517) is gratefully acknowledged. The authors would also like to thank Dr. LeMaster for his invaluable help in the initial phase of hydrogen exchange calculations. The authors would also like to extend thanks to Dr. Grishaev for his invaluable advice and help in computing solution scattering profiles.

CHAPTER 4. DETERMINE THE MINIMAL REQUIREMENT FOR EXPERIMENTAL DATA IN ASSIGNING RELATIVE POPULATIONS TO ENSEMBLE

A manuscript in its final stages of preparation.

4.1 Abstract

The function and dynamics of many proteins are best understood not from a single structure but from an ensemble. In our previous work, using Ubiquitin as an example we have shown that the native state ensemble could be represented by a few appropriately weighted conformations using Residual Dipolar Couplings as constraints. Ubiquitin, being a model protein for dynamics studies using NMR experiments, has abundant experimental data to construct such ensembles but this is not true for many other proteins. To make the method generally applicable to other proteins, it is important to identify the minimal experimental data necessary to construct such ensembles. In this work, we show that such weighted ensembles can be derived using only a few NH RDCs and the ensemble thus obtained is of similar quality to the previous ensembles constructed using both NH RDCs and multi-vector RDCs. We extend the method to Hen Egg White Lysozyme (HEWL) and show that a weighted HEWL ensemble consisting of 3 conformational states reproduces the cross-validation experimental data, RCSA, and solution scattering profiles as accurately as, or even better than other solutions of HEWL reported in the literature.

4.2 Introduction

The functional behaviors of the proteins are often realized by complex conformational changes they undergo (1, 5, 6, 9, 15, 42, 43). Significant advances in experimental techniques, especially in Nuclear Magnetic Resonance (NMR) and Solution scattering profile, have presented opportunities to observe these conformational changes in biologically relevant time scales (21, 87). Exciting computational techniques have also been developed to interpret these experimental data for a better understanding of the underlying energy landscape (23, 31, 32, 34, 88).

Experimental data obtained from NMR have been routinely interpreted using either a single average conformation, or an ensemble that may contains hundreds of conformations. The average conformation, being the least complex, could suffer from under-fitting. It may underrepresent structural variance that exists in the native states of a protein, especially when the protein occupies multiple sub-states (7). On the other hand, ensembles that contain hundreds of conformations are highly susceptible to over-fitting and there is little confidence that any given conformation within the ensemble truly belongs to the native state ensemble (9, 89).

In our previous work (63), we showed that the conformation space could be represented by far fewer conformations than the aforementioned ensemble representations and the conformations could be clustered into conformation states and these conformation states could be assigned relative populations, corresponding to their Boltzmann weights. We name such ensembles ESP ensembles, or Ensembles of Small

number of conformations with relative Populations. The advantage of using an ESP ensemble over an average structure is that it can capture the intrinsic dynamics existed in the native states of many proteins that a single structure misses due to lack of structural variance. The advantage of using ESP over using an ensemble with hundreds of conformations is that it minimizes over-fitting as ESP uses a much smaller number of conformations than regular ensembles.

The objective of this work is to determine what is the minimal amount of experimental data, especially Residual Dipolar Coupling (RDC) data that is required to reliably generate ESP ensembles. ESP ensembles were constructed for Ubiquitin using 22 sets of NH RDC data and 2 sets of multi-vector RDCs (63). Are all these data sets needed? Or only a smaller set of them are necessary? These questions are important since for most proteins we don't have the luxury of having as many sets of RDC data as there are for Ubiquitin. A definite answer to these questions can assist experimentalists to determine what and how much data need to be collected in order to use the method as prescribed in (63) to assign relative populations to a protein ensemble of interest.

In the rest of this work, we show that ESP ensembles, similar to those determined in (63), could be determined using a small amount of experimental data that is as little as a few NH RDCs. To ensure the weights assigned are significant and these ensembles are still of high quality, we carry out a series of evaluations and careful cross-validations. Lastly, as an application, we apply the proposed protocol to Hen Egg White Lysozyme (HEWL) that has 8 sets of NH RDC data. The newly determined ESP ensemble for

Lysozyme is shown to reproduce the cross-validation experimental data, RCSA, and solution scattering profiles as accurately as, or even better than other solutions of HEWL reported in the literature.

4.3 Materials and Methods

4.3.1 Residual Dipolar Couplings (RDC)

Residual dipolar coupling originates from the interaction of two nuclear spins (dipole-dipole) in the presence of the external magnetic field and is defined (20-22, 29) as:

$$D_{\{AB\}} = \sum_{i=x,y,z} -\frac{\mu h r_A r_B}{(2\pi r)^3} \cos^2 \emptyset_i \; A_{ii} \qquad (1)$$

Where r_A and r_B are the nuclearmagnetogyric ratios of nuclei A and B respectively, h is Plank's constant, μ is permittivity of space, r is the internuclear distance between the two nuclei, A_{ii} the principal moment of the alignment tensor and $\emptyset_i$ is the angle between the internuclear vector and i^{th} principal axis of the alignment tensor. The alignment tensor could be determined by fitting a single structure or ensemble to the experimental data. Normally, the residual dipolar coupling reduces to zero because of isotropic tumbling. The anisotropic measurement can be obtained by the aid of various types of liquid crystalline media. Detailed steps on back calculations of RDC's were given in our previous work (63).

The NH RDC datasets along with multi-vector RDC datasets for Ubiquitin used in this work are given in Appendix Table 4.11. The NH RDC datasets used for HEWL are given in Appendix Table 4.12.

4.3.2 Q-factor

Q-factor is a commonly used measure of the agreement between the experimental and calculated RDCs and is defined as:

$$Q\text{-}factor = \frac{\sqrt{\sum\left(D_{calc} - D_{\exp}\right)^2}}{\sqrt{\sum\left(D_{\exp}\right)^2}} \qquad (2)$$

where D_{calc} is the calculated RDC and D_{exp} is the experimental RDC.

4.3.3 Residual Chemical Shift Anisotropy (RCSA)

Along with RDC's, chemical shifts also change upon shifting from an isotropic medium to an anisotropic medium (29, 70, 71). The change is defined by:

$$\Delta\delta = \sum_{i=x,y,z} \sum_{j=x,y,z} A_{jj} \cos^2 \theta_{ij}\, \delta_{ii} \qquad (3)$$

where δ_{ii} is the principal moment of the chemical shift tensor, A_{jj} the principal moment of the alignment tensor and θ_{ij} is the angle between i^{th} principal axis of the chemical shift tensor and j^{th} principal axis of the alignment tensor. The alignment tensor used in RCSA back-calculations is generally the same as the one computed from RDCs using either a

single conformation or an ensemble (63). More information regarding the relation between RDC and RCSA back-calculation of a conformation can be found in (70).

The experimental dataset of RCSA for Ubiquitin used in this work were reported in (29) along with the RDC dataset used for obtaining the alignment tensor. Magnitudes and orientations of the chemical shift tensors are the same as those reported in (72). The experimental RCSAs for HEWL used in this work were taken from (90).

4.3.4 Creating an Artificial Conformation Ensemble and Artificial RDC Data

As in our previous work (63), we first create an artificial native state ensemble and use it as the reference, similar to what was done in (33). We then generate artificial RDC data based on the ensemble composition. The advantage of using artificial ensembles and artificial RDCs is that one has perfect control over their compositions and their noise levels. The artificial native state ensemble used in (63) is used again in this work, with slight modifications.

Briefly, five different conformations of Ubiquitin are assumed to be the centers of the five conformational states of the protein. These five conformations are chosen such that the minimum backbone RMSD between any two conformations is greater than 1.5 Å. We then locally sampled (less than 1 Å away from the center) more conformations around these centers and used them, together with the centers, to represent the conformation states. The Boltzmann weight of each conformational state is set to be proportional to the number of conformations in that energy well, except for conformation

state 1, which are given a weight of 0. Conformational states whose relative populations are less than 10% are not considered, as there is much less confidence in the weights assigned to them. To include noise in the ensemble, noise conformations are sampled at 1.5 Å distance away from the centers. All the conformations, except for the conformations belonging to state one, are then used to generate artificial RDC's. 22 sets of artificial NH RDCs are generated using alignment tensors that are taken from the 22 sets of real experimental NH RDCS of Ubiquitin. To simulate experimental noise in the RDC data, Gaussian noise is added to the artificial RDC's. The number of conformations sampled in each sub-ensemble and the associated Boltzmann weight are given in Table 4.1.

Table 4.1: Boltzmann weights of the conformational states in the artificial ensemble. Conformational state one is not used in the experimental data generation and hence has a Boltzmann weight of 0.

Conformational State	Two	Three	Four	Five	Total
# of Conformations	200	350	500	700	1750
Boltzmann weight	0.114	0.20	0.285	0.40	1

4.3.5 A Sampling of the Artificial Energy Landscape

Next we create a sampling of the energy landscape as defined by the above artificial ensemble. Since generally it is not realistic to expect conformation samplings to be proportional to the Boltzmann distribution, we purposely select a biased sampling of the ensemble. Specifically, 21, 60, 6, 7, 290 conformations are randomly selected from conformation state one, two, three, four, and five respectively. These conformations, along with an equal amount of noise conformations generated around each conformation state are mixed together to form a "sample ensemble" (see Table 4.2). It should be noted

that, the samplings around the conformational states are noisy as a result and are not proportional to the Boltzmann weights of the conformational states. The input to the weighting algorithm as described next is this "sample ensemble" and the artificial experimental data created out of the native state ensemble.

Table 4.2: The composition of the sample ensemble.

Conformational State	One	Two	Three	Four	Five	Total
# of conf.	21	60	6	7	290	384
# of noise conf.	21	60	6	7	290	384

4.3.6 Ensemble of Small Number of Conformations with Relative Populations (ESP):

In our previous work (63), ESP ensembles of Ubiquitin were determined by using experimental data that consisted of 22 NH RDC datasets and two multi-vector RDC datasets. In this work, we aim to demonstrate that NH RDCs alone are sufficient for the weight assignment and to determine the minimum number of NH RDCs that is required. This objective is significant because it will put much less a burden on experimental data collection and will make it easier to extend the method to assign relative populations to other protein ensembles. Using only NH RDCs does present some new challenges: since multi-vector RDC datasets are not used, the numbers of experimental data points, or RDC constraints used to determine the ensemble, become significantly fewer. In the following sections we review our previous method and present the modifications needed to still apply it when only NH RDC datasets are available.

Given as input an ensemble of conformations and experimental RDC data as constraints, the original method, as detailed in (63), had two key steps in constructing an

ESP ensemble: 1) identify and select representative conformations out of the given ensemble of conformations; 2) using the RDC data as constraints and a guide, merge the representative conformations into conformational states and assign to them relative populations while minimizing over-fitting.

Step 1: Identifying representative conformations for the conformation states

Assuming that the starting ensemble has all the conformational states represented with reasonable accuracy, there are two alternative ways to identify representative conformations for the conformation states to be determined.

a) Identify representative conformations by least square fitting: In situations where the number of conformations or conformation clusters (i.e., a small number of conformations that are tightly close to one another in terms of RMSD distance) is equal or less than the number of RDC constraints, least square fitting can be applied to the conformations or clusters to find the weight assignments that best fit the RDC data. Conformations or conformation clusters that has non-zero weights are then selected as representative conformations. Though being a over-fitting, least square fitting at this step was found to be effective in eliminating the majority of noise conformations from the ensemble (63). However, it should be noted that the weights obtained at this point are over-fitted. This is not a problem since these over-fitted weights are used only to separate noise conformations from representative conformations.

b). Identify representative conformations through many randomized runs: this option is used when the number of initial conformations is greater than the number of RDC constraints. This is more likely to be the case when only NH RDCs are used. In such a case, only a subset of initial conformations or conformation clusters, instead of all of them as is in the case of (a), are selected randomly and used as input for the weighting procedure. The weights assigned to these conformations or clusters are then recorded. This procedure is repeated many times and the weights assigned to each conformation are tallied and averaged. Conformations or conformation clusters that consistently receive significant weights are then identified as representative conformations. Specifically, we consider a conformation to have a significant average population if the mean of its populations tallied over the randomization runs plus the standard deviation is greater than 0.07, a population around and below which becomes hardly detectable experimentally.

Step 2: Form conformation states while avoiding Over-fitting:

Once the representative conformations are identified, the algorithm proceeds to form conformation states. Careful consideration is taken to identify and avoid over-fitting in the process. An example that illustrates the steps involved is given in Figure 4.1. Briefly, the procedure is:

i). The representative conformations obtained after pruning the noise conformations are clustered together to form a hierarchical tree in which all the representative conformations form the leaf nodes and closest pairs of nodes are

iteratively merged together to form internal nodes (or sub-ensembles). (see Figure 4.1, panel a). The experimental data, RDC's, are replicated into M sets. These M sets are identical to one another except for a small amount of random Gaussian noise (replica noise) added to them. The purpose for this RDC data replication is to guard against over-fitting (63).

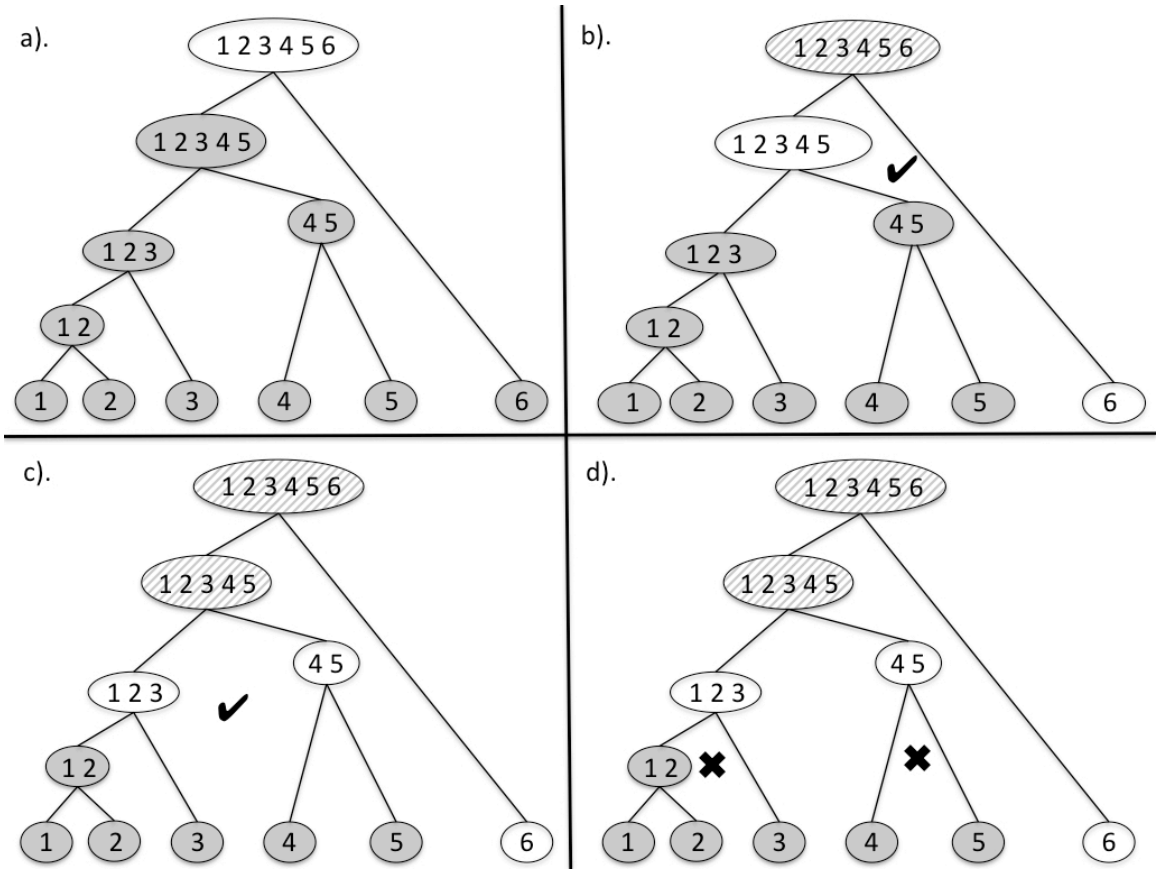

Figure 4.1: In this illustration, all the solid shaded ovals are conformational states not yet reached. Line shaded ovals are conformational states reached and split into new ones. The unshaded ovals are the current conformational states. In the panel a, the hierarchical tree is formed by merging conformational states closest to each other. The only state visible at the start of protocol is root of the tree. In panel b, two new states are discovered and the split is approved by the RDCs. In panel c, one more cluster is discarded into 2 new states and again the split is approved by RDC's. In panel d, splitting exposes a few more conformational states but are found to be over-fitted by RDC's. The final states approved by RDC's are conformational clusters 123, 45, 6.

ii). A traversal from the root node towards the leaf node introduces a new conformational state at every step. In our study, we found that one can confidently

move beyond representing the entire ensemble as one conformation state (the root node) to reach at a level where there are several sub-ensembles and assign relative populations to these sub-ensembles but not to the point where every conformation (leaf node) is assigned a weight. These exists a limit beyond which one cannot further divide a sub-ensemble without incurring over-fitting. This limit represents the extent to which relative populations can be assigned and it depends on the quality of the ensemble and the quality and quantity of the experimental data.

iii). At every step of the traversal, the conformational states are weighed using the M replicas of experimental data. If the newly obtained conformational states are valid and are not subjected to over-fitting, the weights assigned by the M replicas of experimental data should highly correlate between one another (see Figure 4.1, panels b and c). The onset of over-fitting is when such correlations start to greatly degrade, signifying that the data is now being fitted to the random noise added to the replicas instead. A traversal is cancelled if it causes over-fitting. The process stops when no more traversal towards to the leaf nodes can be made (Figure 4.1, panel d).

4.3.7 Picking the Right Level of Replica Noise $\sigma_{replica}$ to Promptly Detect Over-fitting

Each of the RDC data-point in the replicas used in the above fitting procedure can be expressed as:

$$RDC_{replica} = RDC_{dipolar} + \sigma_{exp} + \sigma_{replica} \quad (4)$$

where $RDC_{dipolar}$ represents actual dipolar couplings, σ_{exp} experimental measurement noise in RDC that is the same across all the M sets of experimental data replicas, while $\sigma_{replica}$ is the random Gaussian noise added and is different for different replica.

In our study, we find that the experimental noise σ_{exp} is large enough that when only one single NH RDC dataset (one data-point for all available residues) is used, it results in under-fitting, i.e., the conformation states cannot be fully separated and identified. Using multiple RDC datasets decreases the experimental noise by a factor of square root of n, where n is the number of datasets. A similar equation to (4) but for *n* RDC datasets could then be written as:

$$RDC_{replica} = RDC_{dipolar} + \sigma_{exp}/\sqrt{n} + \sigma_{replica}/\sqrt{n} \qquad (5)$$

Along with decreasing the experimental noise, using multiple NH RDC datasets also helps in capturing the dynamics present in the native state ensemble well. In our study, we find that for a given number of NH RDC datasets some combination of NH RDC datasets can resolve the four states accurately while others fail to do so resulting in either under-fitting or over-fitting. Increasing the number of NH RDC to 4 or more datasets alleviates this problem significantly. This observation is in accordance to the degeneracy problem present in RDCs and studies in the past have suggested using multiple RDCs obtained in independent media to fully capture the dynamics (91).

Picking the right level of replica noise $\sigma_{replica}$ is critical in identifying over-fitting.

a). If $\sigma_{replica}$ is too high, it will introduce too much uncertainty to the data and make it impossible to distinguish the conformation states.

b). If $\sigma_{replica}$ is too low, it will not cause enough perturbation to the system that is needed to identify over-fitting.

Intuitively speaking, the replica noise represents a "shaking" to the solutions found by least square fitting. If there is an over-fitting or it is fitting to noises at a given step, the solution is unstable and some shaking in the noise level will produce a different solution. On the other hand, if it is fitting to the data, then the solution should be stable and some shaking will not disturb it.

As with experimental noise, the net effect of replica noise $\sigma_{replica}$ also gets reduced at a rate of the square root of n when n NH datasets are used (see Eq. (5)). Therefore, when n NH datasets are used, $\sigma_{replica}$ shouldbe increased proportionally (by $\sqrt{n}$ times) to maintain the same optimal level so that it can produce enough perturbation to identify over-fitting. Therefore the equation in (5) for *n* NH RDC datasets could be expressed as:

$$RDC_{replica} = RDC_{dipolar} + \sigma_{exp}/\sqrt{n} + (\sqrt{n}\,\sigma_{optimal})/\sqrt{n}$$

$$= RDC_{dipolar} + \sigma_{exp}/\sqrt{n} + \sigma_{optimal}. \qquad (6)$$

where $\sigma_{optimal}$ is the optimal noise level required for single NH RDC dataset. In doing this, the level of experimental noise gets effectively reduced while the effective replica noise level remains the same. Details on obtaining $\sigma_{optimal}$ are given in the Results section.

4.3.8 Solution Scattering Profile

Small Angle X-ray scattering (SAXS) and wide angle X-ray scattering (WAXS) data encode the information about the shape and size of the bio-molecules in solution (79, 80). The observed intensities from SAXS/WAXS are sensitive to the overall conformational distribution of the protein and are being regularly used as complementary data to those obtained from NMR or X-ray crystallographic studies (82, 87). Predicting the scattering profiles from either single structure or an ensemble are most routinely done using the Crysol software package (83). Along with significantly improving the predictions, AXES (Analysis of X-ray scattering data for Ensemble of structures) (84) webserver, provides an easy method to predict such intensities from ensembles. The predicted intensities of all the ensembles or single structures used in this work are computed using a local version of AXES webserver, generously provided by Bax's group. The experimental SAXS/WAXS data used in this work are reported in (84). The agreement between the predicted and experimental scattering intensities is most commonly denoted by the chi value that is defined as:

$$chi = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{I_{exp}(q_i) - I_{calc}(q_i)}{\sigma(q_i)}\right)^2} \quad (7)$$

where I_{exp} and I_{calc} are the experimental and predicted scattering intensities at q_i with a error of σ_i and M is the number of observed scattering intensities.

4.4 Results and Discussion

In the following sections, we first obtain the optimal replica noise level (see Methods) required to identify over-fitting and then evaluate the success rate in resolving the four conformational states of the artificial native energy landscape when an increasing number of NH RDCs are used, up to 16 NH RDCs. We then apply the method to Ubiquitin and Hen Egg White Lysozyme (HEWL) ensembles to assign relative populations and examine the quality of weighted ensembles thus determined.

4.4.1 Obtaining Optimal Replica Noise ($\sigma_{optimal}$)

Recall that the role played by the replica noise is to cause just enough perturbation to identify over-fitting but such a perturbation may result also in under-fitting. Because of this behavior of optimization, the replica noise is chosen such that the four conformational states of the artificial energy landscape are fully resolved for the majority (>60%) of the time. This value was found to be 0.036 Hz per one experimental data point

and should be increased, by $\sqrt{n}$ times, when n such data points are present when multiple experimental data sets are used.

Accuracy of resolving the conformational states using different number of NH RDC datasets:

Using multiple NH RDC datasets helps reduce the degeneracy problem associated with RDC data sets. To estimate the errors introduced by fewer NH datasets, we run the protocol 100 times for a given number of NH datasets, using a random combination of NH datasets at every iteration. Each combination of NH datasets is then run another 100 times and is marked as success if all four conformational states are identified for at least 60% of the time. The frequencies of success, as a function of the number of NH datasets used, are plotted in Figure 4.2.

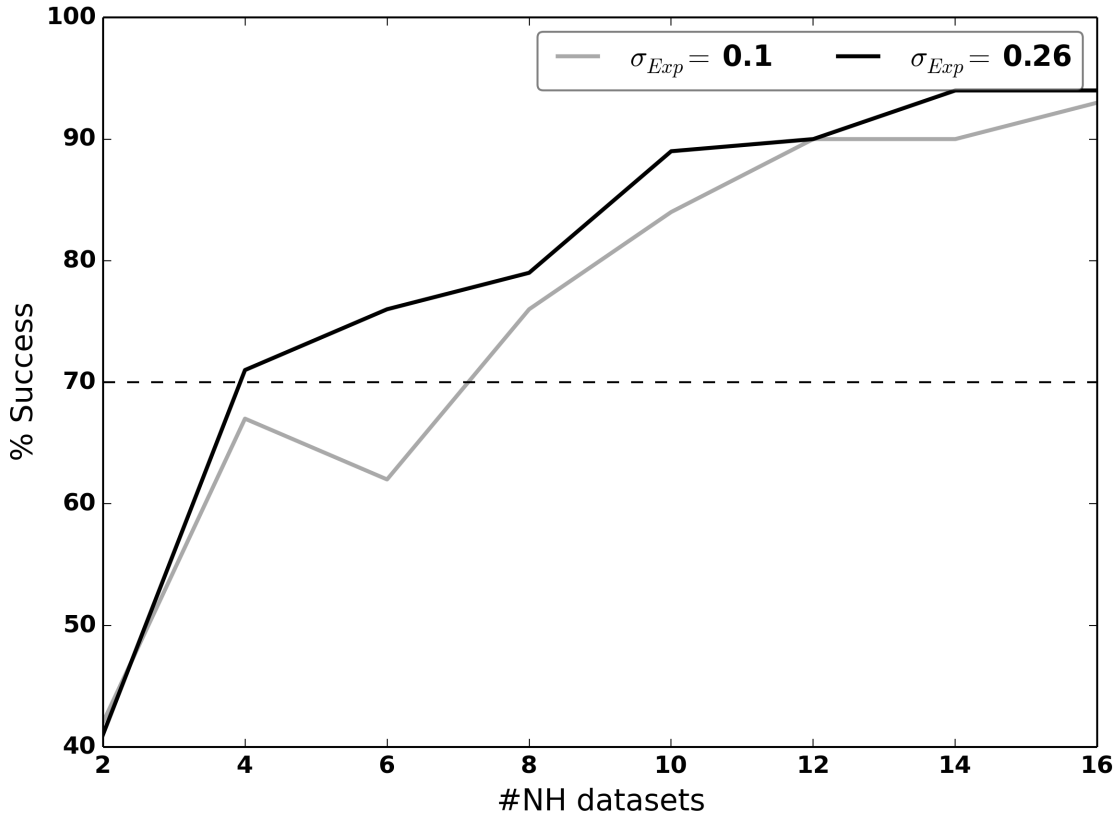

Figure 4.2: The frequency of success, defined as identifying the four conformational states accurately more than 60% of time, against the number of NH RDC datasets.

As expected, the success rates while using fewer (<4) NH datasets is low. It should be noted that, this behavior is not due to the presence of experimental noise or replica noise but RDC data itself as a few combinations of NH RDC datasets, even when using only 2 NH datasets, perform well. Increasing the number of datasets beyond 6 increases the success rates beyond ~70% for all the experimental noises tested in this work.

Experimental data requirements and recommended procedure

Based on our results on artificial energy landscape, the confidence on the number of conformational states in the energy landscape depends on the number of experimental datasets and their capability to capture the native state dynamics. Beyond 6 NH datasets, the experimental noise does not seem to play a role and the conformational states could be resolved with ~70% confidence level. Highly erroneous data sets ($\sigma_{Exp} > 0.50\ Hz$) are not tested in this work but are shown to unreliable by (92).

Recommendation:

To extending this work to actual protein ensembles, we recommend to use as many NH datasets as available with a minimum of 6 NH datasets. Confidence on the correctness of the results (in terms of success rate) could be estimated based on Figure 4.2.

4.4.2 Weighted Ubiquitin Ensembles

In our previous work (63), we determined a weighted X-ray ensemble using all the available data for Ubiquitin which consisted of 22 NH RDC datasets and 2 multi-vector RDC datasets. 16 conformations out of 143 conformations were selected to form a weighted ensemble. The PDB ids of the 143 X-ray structures are given in 4.13. The weighted X-ray ensemble is shown in Table 4.3 along with the compositions of the conformational states and the weights assigned to each conformational state. In

comparison, the weighted X-ray ensemble determined using only NH RDC datasets is shown in Table 4.4.

Table 4.3: The six conformational clusters and their weights of the weighted X-ray ensemble using all the possible data including multi-vector datasets. The conformations included in each cluster are listed by their PDB ids as well as chain identifiers.

Cluster	Final weight ± std	Composition
Cluster1	0.55 ± 0.03	1AAR-B, 1UBQ-A, 2C7M-B, 2C7N-H, 2QHO-A, 3EHV-C, 3M3J-A, 3M3J-E
Cluster2	0.29 ± 0.03	2G45-B, 2G45-E, 2HD5-B
Cluster3	0.064 ± 0.001	2DX5-B, 3KW5-B
Cluster4	0.043 ± 0.002	1YD8-V
Cluster5	0.027 ± 0.004	3HIU-A
Cluster6	0.026 + 0.001	1TBE-A

Table 4.4: The three conformational clusters and their weights of weighted X-ray ensemble using 22 NH RDC datasets. The conformations included in each cluster are listed by their PDB ids as well as chain identifiers.

Cluster	Final weight	Composition
Cluster1	0.57	1CMX-B, 1UBI-A, 1UBQ-A, 1WR6-H, 1XD3-D, 2C7M-B, 2WWZ-A, 3EHV-C, 3M3J-E
Cluster2	0.28	2G45-B, 2G45-E, 2HD5-B, 2IBI-B, 3A9J-B, 3A9K-B, 3EHV-B
Cluster3	0.15	1AAR-B, 2QHO-A, 3M3J-A, 3M3J-C

Cluster 1 of the weighted X-ray ensemble determined using only NH RDC datasets (see Table 4.4) contains conformations very similar to 1UBQ and the composition matches to Cluster 1 determined using both NH RDC and multi-vector datasets (see Table 4.3). The weights assigned in both cases also are highly similar.

The second largest weighted cluster, Cluster 2 in Table 4.3, consisted solely of "switched" conformations. The corresponding cluster in Table 4.4, also cluster 2, consists

mostly of the conformations belonging to the "switched" state such as 2G45-E, 2G45-B, 2HD5-B, 2IBI-B, 3A9J-B, and 3A9K-B. In addition, it contains a conformation (3EHV-B) that is not bound to deubiquitinating enzymes and is not in the "switch" state. The weight assigned to this conformational state also is similar between Tables 4.3 and 4.4.

Cluster3 in Table 4.4 also consists of conformations from Cluster1 in Table 4.3, conformations close to 1UBQ but distant in comparison to the conformations in Cluster1 of Table 4.4. The clusters in Table III that have less significant weights (< 0.10) are not selected at all when only NH RDC datasets are used for assigning weights.

Table 4.5: Q-factors of the different bond vectors of the weighted X-ray ensemble as well as some other ensembles. For the weighted ensemble using only NH RDC, except NH all the remaining bond vectors act as cross-validation while CAHA serves as a cross-validation for ensembles using NH RDCs along with multi-vector datasets.

NH	CaC	CaHa	CN	CHN	Description
0.18	0.11	0.16	0.10	0.228	Unweighted X-ray
0.12	0.10	0.14	0.09	0.19	Weighted X-ray using multi-vector datasets
0.13	0.10	0.14	0.09	0.19	Weighted X-ray using only NH RDC

Table 4.5 lists the RDC Q-factors of different bond vectors. From the table, we can see that the weighted ensemble obtained using only NH RDCs have similar performance to the one that is obtained using 22 NH RDCs and 2 multi-vector RDC datasets. Both perform significantly better than the unweighted ensemble. For the weighted ensemble obtained using only NH RDC's, the remaining bond vectors serve as

cross-validations and it should be noted that there is a consistent improvement in all their Q-factors, indicating no or minimal over-fitting.

Table 4.6: RMSDs of residual chemical shift anisotropy (RCSA) as predicted by representations of Ubiquitin. None of the adjustable parameters in the RCSA was modified while predicting the chemical shifts. Q_{NH} is the Q-factor of the NH dataset that was used in obtaining the alignment tensor. The same alignment tensor was used in the RCSA computations.

Carbonyl	Nitrogen	Amide	Q_{NH}	Description
6.89	17.3	1.61	0.17	Unweighted X-ray
6.37	16.2	1.57	0.11	Weighted X-ray using multi-vector datasets
6.85	16.9	1.56	0.11	Weighted X-ray using only NH RDC

Table 4.6 lists the RMSDs between the experimental Residual Chemical Shift Anisotropy (RCSA) and back-calculated ones from the ensembles. Both weighted ensemble representations show an improvement over the unweighted ensemble. Since RCSA was not used in the weight-fitting process, this further corroborates that there is none or minimal over-fitting in the weighting process. The RMSDs of the weighted ensemble obtained using multi-vector RDCs is very similar to the RMSDs of the ensemble using only NH RDCs, indicating that the two representations have similar quality.

Based on these results, we can confidently state that the native energy landscape of Ubiquitin can be described by 2 conformational states, conformations similar to 1UBQ with relative population weight of ~0.7 and the "switched" conformation with relative population weight of ~0.3. These results agree well with what was found in the 1 μs simulation conducted by Shaw's group (66).

4.4.3 Weighted Hen Egg White Lysozyme (HEWL) Ensembles

While Ubiquitin has been the model protein for NMR studies, Hen Egg White Lysozyme (HEWL) is studied extensively using X-ray crystallography with more than 400 different X-ray structures available in the PDB (10). The PDB ids along with the chain identifiers of all the X-ray conformations used in this work are listed in the Table 4.14. The X-ray conformation, 193L (93), is used a reference conformation for the unbound HEWL.

A single co-ordinate called the pincer angle can conveniently track the dynamics of HEWL. This angle is defined as the angle between the alpha helix (residues: 111, 112, 113, 114), the hinge region (residues: 80, 81, 82, 83, 84, 90, 91, 92, 93) and the beta sheet (residues: 44, 45, 51, 52) (88). While the unbound form of HEWL has a narrow range of motion (55^o to 56^o), the conformations bound to anti-bodies or substrate exhibit a slightly broader range (55^o to 59^o).

HEWL has been studied using NMR techniques also (88, 90, 91, 94) and up to 8 NH RDC's in multiple media and backbone nitrogen RCSA in two media are available in the literature. This makes HEWL an ideal case for our method to be applied to identify conformational states within its native state ensemble. Following the same procedure outlined for Ubiquitin, we next extend our method to the HEWL X-ray ensemble.

Table 4.7 lists the different conformational states identified for HEWL native state ensemble along with their PDB ids and the corresponding pincer angle distribution. All the conformations belonging to Cluster1 exhibit a narrow range of pincer angle distribution commonly exhibited by unbound HEWL. All the conformations except 1HEW and 4GN5 are unbound HEWL conformations. The conformations identified in Cluster2 exhibit a distinct pincer angle distribution from the conformations obtained in Cluster1 and all these conformations are bound to anti-body. 1SQ2, the lone member of Cluster3 is bound to anti-body but with a pincer angle less than ones observed in Cluster2.

Table 4.7: The three conformational clusters and their weights of weighted X-ray ensemble of HEWL along with the pincer angle distribution. The conformations included in each cluster are listed by their PDB ids as well as chain identifiers.

Cluster	Final weight	Pincer angle distribution	Composition
Cluster1	0.57	55.04^{o} to 57.06^{o}	1AKI, 1B0D, 1F0W, 1HEW, 1HF4_B, 1JPO, 1LJ3_B, 1LJ4_B, 1LJE_B, 1LJF_B, 1LJG_B, 1LJH_B, 1LJI_B, 1LJJ_B, 1LJK_B, 1LZB, 1LZC, 1T6V_M, 1UC0, 1UCO_B, 2LYZ, 3LYO, 3LYZ, 4GN5_C, 5LYM_B, 6LYZ, 7LYZ, 8LYZ, 9LYZ
Cluster2	0.30	57.23^{o} to 59.54^{o}	1JTO-L, 1MEL-L, 1MEL-M
Cluster3	0.13	56.24^{o}	1SQ2-L

The Q-factors for NH RDCs obtained in different alignment media are given in Table 4.8 for different representations of HEWL. 1E8L, the NMR solution structure of HEWL, performs well only on NH1and NH3, RDCs that were used in the refinement process. It performs very poorly on the rest of the RDCs that were not used in its refinement. This observation raises doubts about the quality of this solution structure. It

also indicates that the native state of HEWL may consist of distinct substates that a single average structure cannot fully capture.

Table 4.8: Q-factors for NH RDCs obtained for different representations of HEWL.

NH1	NH2	NH3	NH4	NH5	NH6	NH7	NH8	Ensemble/ Structure
0.26	0.34	0.22	0.25	0.25	0.23	0.23	0.28	Unweighted X-ray
0.21	0.30	0.18	0.21	0.20	0.20	0.23	0.28	Weighted X-ray
0.08	0.36	0.08	0.27	0.31	0.28	0.40	0.35	1E8L (94)
0.26	0.34	0.23	0.26	0.27	0.24	0.26	0.28	193L (93)
0.21	0.31	0.16	0.20	0.29	0.20	0.27	0.25	193L with optimized H positions (91)
0.17	0.14	0.14	0.18	0.22	0.16	0.22	0.20	RDC restrained Ensemble (88)

193L, the reference conformation of unbound HEWL and a crystal structure, performs better than the solution structure for all the RDCs that were not used in the refinement of 1E8L. In (91), Redfield and co-workers used the same reference X-ray conformation 193L but with hydrogen positions optimized to reproduce NH RDCs and found it performed better than 193L itself. The RDC restrained ensemble (88), consisting of hundreds of conformations, refined using NH1 and NH3 datasets seems to perform the best among all the representations. But it should be noted that the refinement protocol used to obtain the ensemble implicitly optimizes hydrogen atom positions to best reproduce NH RDCs.

The weighted X-ray ensemble as determined by our method does not have any such optimizations performed on it but uses the RDCs only to select and assign weights

to conformational states in the ensemble. Comparing to the unweighted X-ray ensemble, the weighted X-ray ensemble shows a consistent decrease in Q-factors indicating minimal or no over-fitting in the process. Except for Q-factor obtained for NH2, weighted X-ray, 193L with optimized H position, and RDC restrained ensemble perform similarly in reproducing the RDCs.

Interestingly, these three solutions represent three distinct options of representing the native state of a protein. 193L represents an average/single structure solution. The RDC restrained ensemble is a regular ensemble representation that has been commonly used and consists of hundreds of diverse conformations. Lastly, the weighted X-ray ensemble is an ESP ensemble that we propose. It is an ensemble that consists of only a small number of conformations but has relative populations assigned to them. As aforementioned, these three representations of HEWL give similar performance in reproducing RDCs (see Table 4.8). It thus would be highly interesting if there are some other experimental data that can be used to further distinguish the quality of these three representations. In the following, we look into RCSA and SAXS/WAXS data. The results indicate that these three representations are not of the same quality. The RDC restrained ensemble does poorly in reproducing the WAXS data, suggesting that it might have over-fitted the RDCs data using its hundreds of conformations. 193L with modified hydrogen positions performs similarly to the ESP representation. However, it might be difficult to fully justify the validity of using modified hydrogen positions.

Cross-validation using Residual Chemical Shift Anisotropy (RCSA)

Table 4.9: RMSDs of residual chemical shift anisotropy (RCSA) as predicted by representations of HEWL. None of the adjustable parameters in the RCSA was modified while predicting the chemical shifts. Q_{NH} is the Q-factor of the NH dataset that was used in obtaining the alignment tensor. The same alignment tensor was used in the RCSA computations. RCSA RMSD for RDC restrained ensemble (88) was not reported in the literature.

NH1 medium		NH3 Medium		
RDC Q-factor	RCSA RMSD	RDC Q-factor	RCSA RMSD	Ensemble/Structure
0.26	22.7	0.22	24.02	Unweighted X-ray
0.21	21.41	0.18	20.61	Weighted X-ray
0.08	27.39	0.08	27.86	1E8L (94)
0.26	23.12	0.23	24.87	193L (93)
0.21	20.93	0.18	19.22	193L with optimized H positions (91)

As with Ubiquitin, RCSA serve as an unbiased cross-validation since they are not used in the refinement or the weight fitting protocol itself. In Table 4.9, we list the RMSD between the experimental 15N RCSA to the back-calculated ones from the ensembles. From the table it is seen that the weighted X-ray ensemble is significantly better than the unweighted, confirming the significance of weighting. 1E8L, the solution structure that was determined using NH1 and NH3 RDCs as constraints, performs much more poorly also in RCSA RMSD. This, together with its unusually high RDC Q-factors in media whose data were not used in its refinement, indicates this solution structure probably was overly fitted to the NH1 and NH3 RDC data. The weighted ensemble performs better also than a single structure such as 193L. The 193L structure with optimized H position, however, performs similarly to the weighted X-ray ensemble. The RCSA RMSD for the RDC restrained ensemble (88) was not reported in the literature.

Cross-validation using Solution Scattering profile

Solution scattering profiles are observed scattered intensities of X-rays that are collected as a function of the scattering vector q. Typically a q value of 0 to ~0.3 Å falls into the Small Angle X-ray scattering (SAXS) regime while the range for the Wide Angle X-ray scattering (WAXS) regime is ~0.1 to 2.5 Å. The information encoded in these two regimes along with the results obtained for different structure or ensemble representations of HEWL are presented in the following two sections.

Small Angle X-ray Scattering (SAXS)

Scattering intensities observed at SAXS encode information about the overall size and shape of the molecule, radius of gyration (R_g) and other low-resolution information (86). Table 4.10 lists the chi value obtained by different representations of HEWL.

Table 4.10: SAXS or WAXS Chi values obtained for different representations of HEWL. A representative ensemble of 188 conformations of the RDC restrained ensemble is used for computing SAXS/WAXS profiles.

SAXS chi	WAXS chi	Merged Chi	Ensemble/Structure
0.12	1.52	1.04	Unweighted X-ray
0.12	1.46	0.98	Weighted X-ray
0.15	3.51	2.47	1E8L (94)
0.12	1.62	1.01	193L (93)
0.13	1.66	0.96	193L with optimized H positions (91)
0.12	2.21	2.21	RDC restrained Ensemble (88)

From the Table 4.10 we can see that at a low resolution of SAXS regime, all the representations of HEWL perform equally well. The weighted X-ray ensemble and unweighted X-ray ensemble chi values are the same, even though the number of conformations in the weighted X-ray ensemble (28 conformations) is only a small

fraction of the unweighted X-ray ensemble (432 conformations). This adds confidence on the conformations selected by the protocol to form the weighted X-ray ensemble.

Figure 4.3 plots the relative intensities (I_{exp}/I_{calc}) for different representations. As observed in the chi values, all the representations perform equally well with 1E8L deviating the most amongst the possible representations.

Wide Angle X-ray Scattering (WAXS)

Scattering intensities observed at wider angles (higher q) encode information of higher resolution than SAXS but at the cost of potentially bringing in a higher noise level since the intensity of solution scattering also increases. Since data used in this analysis are limited to the range of q values that are less than 1.0 Å, the extent of this noise is limited. WAXS data are often used to validate structural models and to identify structural changes (86). Table 4.10 lists the WAXS chi values obtained for different representations of HEWL.

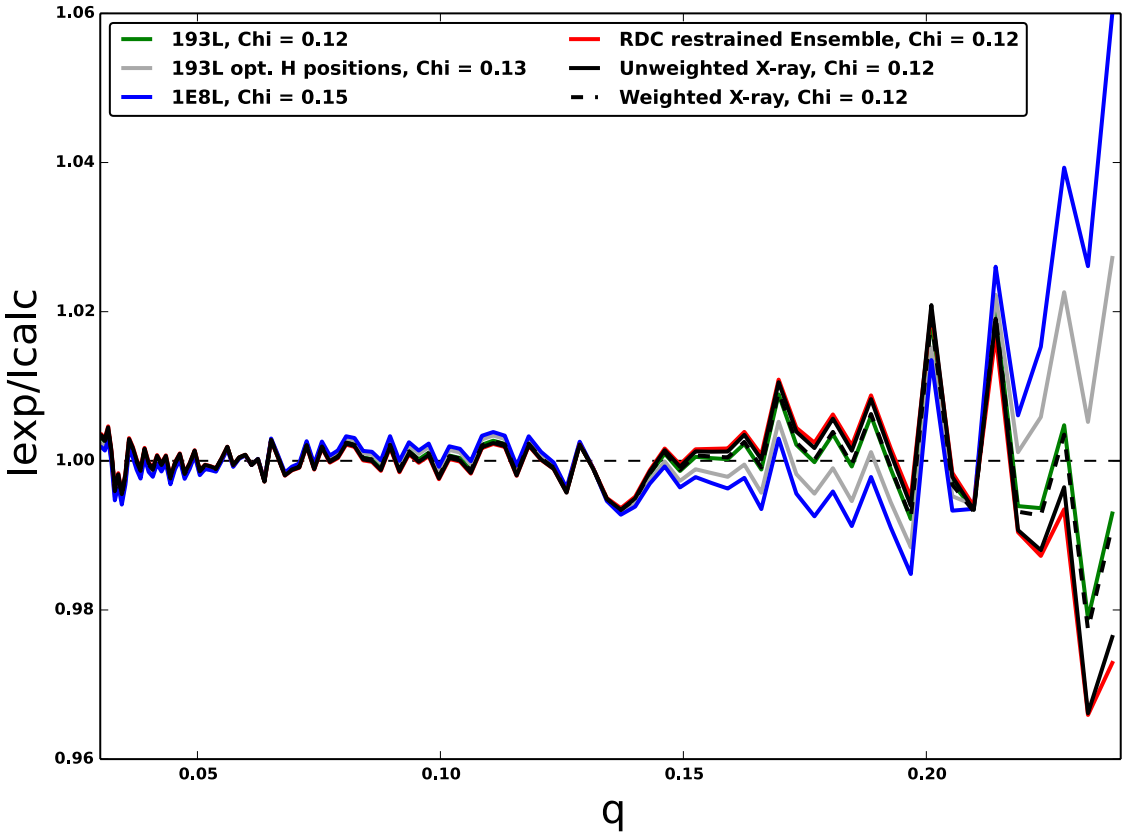

Figure 4.3: Relative intensities (I_{exp}/I_{calc}) as a function of q in the SAXS regime for different representations of HEWL. The chi values obtained for different representations also are given.

Table 4.10 shows that the single conformation 193L reproduces WAXS data extremely well. Weighted X-ray ensemble performs better than the unweighted, again confirming the significance of the weights. The reweighting process also excludes a large number of structures from the ensemble and keeps only a small subset of structures, probably those of higher quality, to represent the native state. As a result, weighted X-ray ensemble does better than both the unweighted X-ray ensemble and 193L itself. RDC restrained ensemble (De Simone et al., 2013) has many conformations very different from 193L and does poorly in reproducing WAXS data. 1E8L performs the worst amongst all the representations.

The chi values obtained by merging SAXS and WAXS data are also reported in Table 4.10. Figure 4.4 plots the relative intensities (I_{exp}/I_{calc}) for different representations.

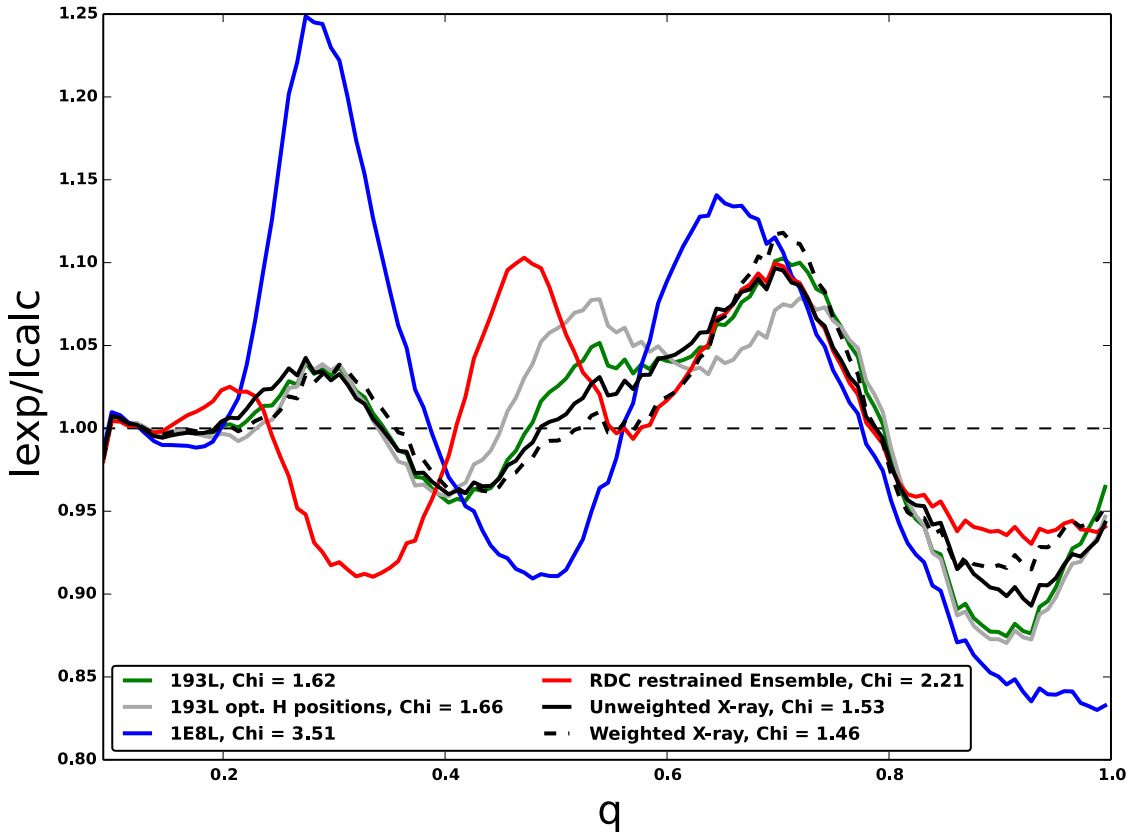

Figure 4.4: Relative intensities (I_{exp}/I_{calc}) as a function of q in the WAXS regime for different representations of HEWL. The chi values obtained for different representations also are given.

4.5 Conclusions

As a continuation of our previous work (63) that demonstrated the significance of assigning relative populations to protein ensembles and how it can greatly enhance an ensemble's quality, the focus of this work is to determine what is the minimal amount of

experimental data required to do such weight assignments. Though our previous study used both a large number of NH RDCs and several multi-vector RDCs, we demonstrate in this work that a much lesser amount of RDCs are needed. Specifically, we show that a minimum of 6 NH RDCs is sufficient to confidently assign weights with a high confidence. This is highly significant since it puts much less a burden on experimental data collection and thus increases the potential for the method to be applied to many more proteins to improve ensemble quality.

To establish that NH RDCs are sufficient for weight assignment, we perform a series of tests both on an artificial ensemble and on Ubiquitin. Our results show that the weighted Ubiquitin ensemble determined using only NH RDCs has a similar quality to the one determined using 22 NH RDCs and 2 multi-vector RDCs.

We then extend and apply the method to hen egg white Lysozyme (HEWL), and determine a weighted, 3-conformation state ensemble. This newly constructed, weighted ensemble performs significantly better than the unweighted ensemble, NMR solution structure, or crystal structure without optimized H positions, in RDCs, RCSA, or solution scattering profiles. Moreover, the two dominating conformation states match closely to, i) the unbounded state with a smaller pincer angle, and ii) the antibody-bound state that has a larger pincer angle, respectively. Our results indict that the unbound state has a population of nearly 60%, while the antibody-bound state a population of 30%.

Average structure representation of the native state of HEWL, 193L with optimized H positions, reproduces RDCs, RCSA and the solution scattering profile accurately. But this representation could possibly be uninformative, if the native states of HEWL indeed occupy multiple substates. Additionally, using optimized H position is not fully justified. Ensembles naturally incorporate multiple substates and we show that weighted X-ray ensemble provides a natural extension without compromising the quality of experimental data reproduction. Weighted X-ray ensemble does better also in capturing the solution scattering profiles significantly better than the RDC restrained ensemble (88), implying the presence of over-fitting in the latter.

Along with our previous work, results from this work corroborate that for some proteins, a single structure is not sufficient to represent its native state fully. An ensemble of conformation is better suited for that purpose and introducing relative populations to the ensembles can significantly improve the quality of the ensembles and greatly reduce the potential risk of over-fitting. The improvement is reflected not only in greatly reduced RDC factors, but also in greatly reduced RCSA RMSDs.

We believe that our method can help significantly improve the quality of ensembles for many proteins. And we thus recommend to the NMR community that RDC data be collected for more proteins, especially for those proteins for which there already exist a large number of experimental structures in the PDB (10), such as HIV protease, Adelnyte Kinase, etc. Once sufficient experimental RDC data are collected and becomes available for these proteins that have high biological significance, our method can be

applied to, i) identify the conformation states of these proteins; ii) determine the relative population of each conformation state. Such in-depth knowledge of the conformation states such as their compositions and their relative populations should provide new understanding of the native states of these proteins and insights into their functional mechanisms. RCSA and SAXS/WAXS data also are recommended to be collected for these proteins, as they can serve as a good cross-validation.

4.6 Appendix

Table 4.11: RDC datasets used for Ubiquitin, coded according to (55)

Experimental data type	RDC data
NH	A1, A2, A4, A6, A7, A8, A9, A10, A11, A12, A13, A16, A21, A22, A23, A24, A25, A26, A27, A28, A29, A34, A36
NH, CN, CHN, CaC and CaHa	(56) (2 sets)

Table 4.12: RDC datasets used for HEWL along with the table code assigned in (91)

NH RDC code	Table code in (91)
NH1	S2
NH2	S3
NH3	S4
NH4	S5
NH5	S6
NH6	S7
NH7	S8
NH8	S9

Table 4.13: PDB ids as well as chain identifiers of the 143 Ubiquitin X-ray conformations used in this work to form the Ubiquitin X-ray ensemble.

1AAR-A, 1AAR-B, 1CMX-B, 1F9J-A, 1F9J-B, 1NBF-C, 1NBF-D, 1OGW- A, 1P3Q-U, 1P3Q-V, 1S1Q-B, 1S1Q-D, 1TBE-A, 1TBE-B, 1UBI-A, 1UBQ- A, 1UZX-B, 1WR6-E, 1WR6-F, 1WR6-G, 1WR6-H, 1WRD-B, 1XD3-B, 1XD3-D, 1YD8-U, 1YD8-V, 2AYO-B, 2C7M-B, 2C7N-B, 2C7N-D, 2C7N- F, 2C7N-H, 2C7N-J, 2C7N-L, 2D3G-A, 2D3G-B, 2DX5-B, 2FID-A, 2FIF- A, 2FIF-C, 2FIF-E, 2G45-B, 2G45-E, 2GMI-C, 2HD5-B, 2HTH-A, 2IBI- B, 2J7Q-B, 2J7Q-D, 2JF5-A, 2JF5-B, 2O6V-A, 2O6V-C, 2O6V-E, 2O6V- G, 2OOB-B, 2QHO-A, 2QHO-C, 2QHO-E, 2QHO-G, 2WDT-B, 2WDT- D, 2WWZ-A, 2WWZ-B, 2WX0-A, 2WX0-B, 2WX0-E, 2WX0-F, 2WX1- A, 2XEW-A, 2XEW-B, 2XEW-C, 2XEW-D, 2XEW-E, 2XEW-F, 2XEW- G, 2XEW-H, 2XEW-I, 2XEW-J, 2XEW-K, 2XEW-L, 2XK5-A, 2ZCC-C, 2ZNV-C, 3A1Q-A, 3A1Q-D, 3A33-B, 3A9J-B, 3A9K-B, 3ALB-A, 3ALB- B, 3ALB-C, 3ALB-D, 3BY4-B, 3C0R-B, 3C0R-D, 3EEC-A, 3EEC-B, 3EFU-A, 3EHV-B, 3EHV-C, 3H1U-A, 3H1U-B, 3H7P-B, 3H7S-A, 3H7S-B, 3HM3-A, 3HM3-B, 3HM3-C, 3HM3-D, 3I3T-B, 3I3T-D, 3I3T-F, 3I3T-H, 3IFW-B, 3IHP-C, 3IHP-D, 3JSV-B, 3JVZ-X, 3JVZ-Y, 3JW0-X, 3JW0-Y, 3K9P-B, 3KVF-B, 3KW5-B, 3LDZ-E, 3LDZ-F, 3LDZ-G, 3M3J-A, 3M3J-B, 3M3J-C, 3M3J-D, 3M3J-E, 3M3J-F, 3MHS-D, 3NHE-B, 3NOB-B, 3NOB- C, 3NOB-D, 3NOB-E, 3NOB-F, 3NOB-G, 3NOB-H

Table 4.14: PDB ids of Hen Egg White Lysozyme (HEWL) X-ray conformations used in this work to form the X-ray ensemble.

193L, 194L, 1AKI, 1AZF, 1B0D, 1B2K, 1BGI, 1BVX, 1BWH, 1BWI, 1BWJ, 1C08, 1C10, 1DPW, 1DPX, 1DQJ, 1F0W, 1F10, 1FDL, 1G7H, 1G7I, 1G7J, 1G7L, 1G7M, 1GPQ, 1GWD, 1H87, 1HC0, 1HEL, 1HEW, 1HF4, 1HSW, 1HSX, 1IC4, 1IC5, 1IC7, 1IEE, 1J1O, 1J1P, 1J1X, 1JIS, 1JIT, 1JIY, 1JJ0, 1JJ1, 1JJ3, 1JPO, 1JTO, 1JTT, 1KIP, 1KIQ, 1KIR, 1LCN, 1LJ3, 1LJ4, 1LJE, 1LJF, 1LJG, 1LJH, 1LJI, 1LJJ, 1LJK, 1LKR, 1LKS, 1LMA, 1LPI, 1LSA, 1LSB, 1LSC, 1LSD, 1LSE, 1LSF, 1LYO, 1LYS, 1LZ8, 1LZ9, 1LZA, 1LZB, 1LZC, 1LZT, 1MEL, 1MLC, 1N4F, 1NDM, 1P2C, 1PS5, 1QIO, 1QTK, 1RCM, 1RFP, 1RI8, 1RJC, 1SQ2, 1T3P, 1T6V, 1UA6, 1UC0, 1UCO, 1UIG, 1UIH, 1UUZ, 1V7S, 1V7T, 1VAT, 1VAU, 1VDP, 1VDQ, 1VDS, 1VDT, 1VED, 1VFB, 1W6Z, 1WTM, 1WTN, 1XEI, 1XEJ, 1XEK, 1XFP, 1XGP, 1XGQ, 1XGR, 1XGT, 1XGU, 1YIK, 1YIL, 1YKX, 1YKY, 1YKZ, 1YL0, 1YL1, 1YQV, 1Z55, 1ZV5, 1ZVY, 2A7D, 2A7F, 2AUB, 2BLX, 2BLY, 2BPU, 2C8O, 2C8P, 2CDS, 2CGI, 2D4I, 2D4J, 2D4K, 2D6B, 2D91, 2DQC, 2DQD, 2DQE, 2DQF, 2DQG, 2DQH, 2DQI, 2DQJ, 2EIZ, 2EKS, 2EPE, 2F2N, 2F30, 2F4A, 2F4G, 2FBB, 2G4P, 2G4Q, 2H9J, 2H9K, 2HTX, 2HU1, 2HU3, 2HUB, 2I25, 2I26, 2I6Z, 2LYM, 2LYO, 2LYZ, 2LZT, 2PC2, 2Q0M, 2VB1, 2W1L, 2W1M, 2W1X, 2W1Y, 2X0A, 2XBR, 2XBS, 2XJW, 2XTH, 2YBH, 2YBI, 2YBJ, 2YBL, 2YBM, 2YBN, 2YDG, 2YSS, 2YVB, 2Z12, 2Z18, 2Z19, 2ZNX, 2ZQ3, 2ZQ4, 2ZYP, 3A34, 3A67, 3A6B, 3A6C, 3A8Z, 3A90, 3A91, 3A92, 3A93, 3A94, 3A95, 3A96, 3AGG, 3AGH, 3AGI, 3AJN, 3ATN, 3ATO, 3AW6, 3AW7, 3AZ4, 3AZ6, 3AZ7, 3B6L, 3B72, 3D9A, 3E3D, 3EMS, 3EXD, 3F6Z, 3IJU, 3IJV, 3KAM, 3LYM, 3LYO, 3LYT, 3LYZ, 3LZT, 3M18, 3M3U, 3N9A, 3N9C, 3N9E, 3P4Z, 3P64, 3P65, 3P66, 3P68, 3QE8, 3QNG, 3RNX, 3RT5, 3RU5, 3RW8, 3RZ4, 3SP3, 3T6U, 3TMU, 3TMV, 3TMW, 3TMX, 3TXB, 3TXD, 3TXE, 3TXF, 3TXG, 3TXH, 3TXI, 3TXJ, 3ULR, 3VFX, 3W6A, 3ZEK, 4A7D, 4AGA, 4AXT, 4B0D, 4B1A, 4B49, 4B4E, 4B4I, 4B4J, 4BAD, 4BAF, 4BAP, 4BS7, 4C3W, 4D9Z, 4DD0, 4DD1, 4DD2, 4DD3, 4DD4, 4DD6, 4DD7, 4DD9, 4DDA, 4DDC, 4DT3, 4E3U, 4EOF, 4ET8, 4ET9, 4ETA, 4ETB, 4ETC, 4ETD, 4ETE, 4FJR, 4G49, 4G4A, 4G4B, 4G4C, 4G4H, 4GCB, 4GCC, 4GN3, 4GN4, 4GN5, 4H1P, 4H8X, 4H8Y, 4H8Z, 4H90, 4H91, 4H92, 4H93, 4H94, 4H9A, 4H9B, 4H9C, 4H9E, 4H9F, 4H9H, 4H9I, 4HP0, 4HPI, 4HSF, 4HTK, 4HTN, 4HTQ, 4HV1, 4I8S, 4IAT, 4II8, 4J1A, 4J1B, 4J7V, 4KXI, 4LFP, 4LFX, 4LGK, 4LT0, 4LT1, 4LT2, 4LT3, 4LYM, 4LYO, 4LYT, 4LYZ, 4LZT, 4M4O, 4MR1, 4N5R, 4NEB, 4NFV, 4NG1, 4NG8, 4NGI, 4NGJ, 4NGK, 4NGL, 4NGO, 4NGV, 4NGW, 4NGY, 4NGZ, 4NY5, 4O34, 4OOO, 4P2E, 4QEQ, 4TUN, 5LYM, 5LYT, 5LYZ, 6LYT, 6LYZ, 7LYZ, 8LYZ, 9LYZ

CHAPTER 5. DO ENSEMBLE REFINEMENTS USING RESIDUAL DIPOLAR COUPLING IMPROVE THE STRUCTURAL QUALITY?

The following contains preliminary results of an ongoing work.

5.1 Abstract

Nuclear Magnetic Resonance (NMR) has played a pivotal role in capturing the structure and dynamics of proteins in native state. Traditionally such dynamic data has been structurally modeled using single or average structure. But for more and more proteins, it is becoming increasingly evident that an ensemble of conformations rather than a single structure might capture the dynamics better. Indeed, a number of recent works on ensemble refinement saw a significant increase in the quality of reproduction of experimental data. However, it is unclear whether the increase is due to a better description of protein native states or due to over-fitting. In this work, using synthetic experimental data on Residual dipolar Couplings (RDCs) and Nuclear Over-hauser effects (NOEs) , we show that ensemble refinements of arbitrary number of conformations do not increase the structural quality of the solution and the cross-validation data typically used, CaHa RDC, can be well reproduced even where there is over-fitting. Such overfitting can be avoided if good initial conformations are provided and appropriate relative populations are assigned to them.

5.2 Introduction

Proteins in solution can occupy multiple conformational states and the stability of a conformational state depends upon the relative free energy of the state(3, 5, 42). Nuclear Magnetic resonance (NMR) experiments probe the dynamics of bio-molecules in their native states and the resultant data from such experiments are time and ensemble averages. Even though the data obtained from NMR experiments have been traditionally used to model the underlying native state ensemble, most of these uses are found in loop modeling (65), rigid body docking (95) and force-field optimization (96), to name a few. In this work, we focus exclusively on protocols used in structural modeling of protein native states.

Single structure refinements or single-copy refinements enforce that all NMR data constraints along with the covalent geometry regulations or empirical constraints should be satisfied in one single conformation. Generally such a refinement is carried many times and the best 10-20 structures are reported. It is unclear if any structure in such an ensemble represents a "true" conformational state of the protein, as single structure refinements would result in average conformations. Extracting dynamics from such an ensemble may also be difficult. For Ubiquitin, one of the most studied proteins, a single structure has been shown to be sufficient in reproducing most experimental data (29, 30).

On the other hand, ensemble refinements are attempted in the recent past where instead of forcing all the NMR constraints on one conformation, an ensemble of conformations is used. In all of these cases, all the members of the ensemble are given

equal weight. The R-factors (quality of fit indicators) of such a refinement have been considerably improved compared to single structure refinements but it is unclear whether the fit is better because of a better structural representation of the underlying native states or the large increase in parameters when multiple conformations are used. In an elegant work by Clore et al. (32), it was shown that an ensemble refinement of 2 structures significantly improves the cross-validation, the CaHa RDC R-factors, compared to single structure refinement and increasing the ensemble size beyond that does not show much improvement in the R-factors. While this work is important in identifying the minimal number of conformations required to satisfy the experimental data without over-fitting, there was no analysis on the quality of the solution itself.

There has been a lot of recent work aimed at determining an ensemble of conformations for Ubiquitin, such as MUMO (33), EROS (23) and ERNST (34). All of these refinement protocols used an ensemble of size 8 in their refinement protocols to satisfy experimental data. Typically the refinement protocol was run in many cycles and the resulting solutions from every cycle were pooled to form an ensemble. Consequently, these ensembles contain 100s of conformations. All of these ensembles are shown to represent the dynamics well but there is little confidence that any given conformation within the ensemble truly belongs to the native state ensemble, since the ensemble might be under-constrained or over-fitted(9, 35).

In our extensive cross-validation studies using varied experimental data (see chapter 3), we have shown that ensembles of 100s of conformations do not perform significantly better than a weighted ensemble that has two significantly populated

conformational states (63) or even an average structure representation. These observations raise some fundamental questions on the quality of the solution provided by ensemble refinements.

In this work, we aim to assess the quality of ensemble refinements using synthetic data. The simplest representation of an ensemble is a two membered ensemble with equal or unequal weights. Two distinct conformations of Ubiquitin are used to represent the native state ensemble, from which synthetic experimental data, RDCs and NOEs, is generated. These synthetic data are then used to guide refinements in a similar manner to conventional ensemble refinements and the resulting solution is verified with respect to the reproduction of experimental data (RDCs and NOEs) as well as structural similarity to the reference ensemble.

5.3 Materials and Methods

5.3.1 Reference Structures and Dynamics

To thoroughly test the quality of solution obtained by average structure or ensemble refinements we used an artificial native state ensemble whose structural and dynamic properties are known. The benefit of using such reference ensembles is that they can be used as a standard to assess the quality of any obtained solution from the refinement protocols (33). To keep the synthetic set-up as close as possible to native state, 2 distinct states of Ubiquitin 1AAR-B and 2HD5-B, representing unbound state and "switched" state of Ubiquitin respectively, are chosen to form the reference structures.

These two conformations have a backbone RMSD of ~1 Å away from each other and are shown to represent the majorly populated conformation states of the protein (63).

Assuming that these two states are the only possible conformational states, the dynamics in the native state are obtained by local sampling around these states using CONCOORD (59). A damp factor of 0.5 is used, which effectively produces sampling conformations that are about 0.5 Å away from the reference structures. The first 3 principal components (PCs) in the principal component analysis (PCA) of the resultant ensemble capture more than 75% of the dynamics.

5.3.2 Synthetic Experimental Data

Under the assumption that the native state only contains 2 conformations, synthetic NOE constraints and RDCs are generated from the reference structures mimicking the experimental data constraints obtained from NMR experiments as closely as possible. These synthetic experimental data constraints are used in the structure/ensemble refinement protocols that follow.

Nuclear Over-hauser Effect constraints:

The distance constraints used in the refinement of 1D3Z (29) was used to generate synthetic experimental data using the 2 reference structures. A total of 2727 NOEs are available for 1D3Z, out of which 1320 distance restraints are used after removing ambiguous restraints. Any NOE distance constraint less than 5 Å was given a lower bound of 1.8 Å. To simulate experimental errors observed in NOEs, Gaussian noise of 10% of the magnitudes of the observed values was added to the distance constraints (41).

Residual Dipolar Couplings:

Synthetic RDC datasets matching the composition of the real experimental RDC data of Ubiquitin are generated using the two reference structures. The RDC datasets along with the codes assigned by Lakomek (55, 57) are given in (63) along with details on computing RDCs from ensembles. Briefly, the average directional cosine matrix of the ensemble is first calculated from the ensemble. Then for each of the experimental datasets the best-fit Saupe matrix is determined using 1D3Z NMR ensemble. Multiplying the average directional cosine matrix with this Saupe matrix produces synthetic RDC datasets. At this point, these RDC datasets are noise-free. In reality, experimental data contains noise of about 0.5 to 1.0 Hz (24, 31), we added Guassian noise to the artificially generated RDC data that are originally noise-free. The standard deviations of the noise are 0.26 Hz, 0.1 Hz, 0.5 Hz, 0.1 Hz and 0.1 Hz for NH, CaC, CaHa, CN and CHN datasets respectively, as was done in (Clore and Schwieters, 2004a). Note that because of the way in which the synthetic RDC data are generated, the given conformation ensemble can perfectly reproduce these RDC data prior to the adding of the noise, but not so after.

Synthetic experimental data are obtained using only the 2 reference structures under equal weighted and un-equal weighted conditions by applying the weights appropriately during the back-calculations of the data.

Cross-Validation:

R-factor is a commonly used measure of the agreement between the experimental and calculated RDCs and is defined as:

$$R\text{-}factor = \frac{\sqrt{\sum\left(D_{calc} - D_{\exp}\right)^2}}{\sqrt{2\sum\left(D_{\exp}\right)^2}} \qquad (3)$$

where D_{calc} is the calculated RDC and D_{exp} is the experimental RDC.

Unlike Lange et al (23)who used CN vector for cross-validation, the CaHa dataset was used for cross-validation in this work. Given that the data used in refinement includes CaC, CHN, NH vector orientations, CN RDC might not be the best choice. CaHa vector, on the other hand, is not in the peptide plane and is thus independent of other bond vector orientations, making it a better cross-validation dataset. Along with CaHa RDC, 20% of NOE distance restraints are randomly chosen to be left out as additional cross-validation.

Structural properties of the refinement solution:

RMSD has been used as a measure to assess the structural similarity between conformations. Every conformation belonging to the refinement solution is assigned to one of the 2 reference structures based on its RMSDs to the reference structures.

5.3.3 Refinement Protocol

Refinement protocol similar to the one detailed in (32) is used in the average structure or ensemble refinements. Briefly, torsional angle dynamics and minimization followed by Cartesian minimization are performed. The final step of Cartesian minimization is important to allow acceptable degree of deviation from ideal covalent geometry. The force constants for bond and angular terms are optimized so as to minimize the deviation from ideal covalent geometry. The maximum allowed deviation is set to be 5^{o} and the average deviation is less than 2.5^{o}. The force constants for RDCs are scaled with respect to NH RDC (CN: 25, CAC: 15, CHN: 5) and are ramped up geometrically from 0.4 to 4 kcal mol^{-1} Hz^{-2} throughout the 2500 steps of simulated annealing protocol with the temperature cooling from 400 K to 300 K. Each cycle is performed 16 times and the final solution is reported. 24 such solutions are generated for all the ensemble/structures used in this work. XPLOR-NIH (28) software suite is used to perform the simulated annealing and minimization.

5.4 Results and Discussion

In this section we test the quality of solutions achieved by various refinement protocols using cross validation R-factors and structural similarities to the reference ensemble. We present the analysis of a simple equal weighted reference ensemble of two conformations and then increase the complexity by allowing unequal weights in the reference ensemble.

5.4.1 Equal weighted Reference Ensemble

In this scenario, the two conformations in the reference ensemble are given equal weights and are used in the generation of the artificial RDC data. 1UBQ-A was used as the starting conformation(s). Different ensemble sizes of 1, 2, 3, 4, 6 and 8 were tested. All conformations in the ensemble are given the same weight.

Table 5.1: Quality of solutions for assessed by reproduction of experimental data and structural similarity to the reference ensemble. The data shown here are the R-factors for RDCs, RMSD for NOE distance constraints and RMSD for structural similarity. The percentage of solutions close to reference structure 1 or 2 is computed by finding the fraction of conformations closer to reference structure 1 or 2. The Ne number represents the ensemble size. The solution denoted by Ne value of 2u is generated by starting the refinement with initial conformations close to the reference structures.

R-factors for bond vectors	Ne=1	Ne=2	Ne=3	Ne=4	Ne=6	Ne=8	Ne=2u
NH	9.9	4.9	5.8	3.8	3.3	3.0	4.5
CAC	9.7	5.5	6.7	4.4	3.5	2.7	5.3
CHN	9.8	4.0	4.9	3.6	3.4	3.0	4.0
CN	12.6	6.7	6.7	5.3	4.5	4.0	6.1
NOE	0.14	0.09	0.08	0.07	0.07	0.07	0.09
Cross-Validation R-factors and NOE RMSD							
CAHA	13.2	12.3	11.4	10.7	10.06	11.1	8.3
NOE (free)	0.29	0.28	0.28	0.31	0.32	0.34	0.25
Structural Properties: percentages of conformations closer to reference structure 1 than 2 (or vice versa) and their mean RMSD to the closer reference structure							
% conf. closer to Ref1	0	8	24	32	41	41	50
Mean RMSD to Ref1	NA	1.21	1.53	1.78	2.15	2.44	0.79

Table 5.1 continued

% conf. closer to Ref2	100	92	76	68	59	59	50
Mean RMSD to Ref2	0.6	1.1	1.51	1.70	2.1	2.40	0.69

Figure 5.1 plots the distributions of the solutions of different ensemble sizes in the principal component space defined by the first two PCs of the reference ensemble. To this end, two hundred conformations are first generated by following CONCORD procedure near both conformation 1 and conformation 2 in the reference ensemble (the gray dots in Figure 5.1). The PCs are obtained by applying PCA to these 400 conformations (gray dots). The cross-validation R-facs for different ensemble sizes also are given in the figure. To examine the importance of having 'good' initial conformations in refinement, a refinement starting with two conformations close to the two reference structures was also performed and the resulting distribution is shown as *2u* in the figure. Figure 5.2 plots the same distribution against the first and third PC of the reference ensemble. The statistics on the reproduction quality of both experimental data (both working and cross-validation) and structural properties are given in Table 5.1.

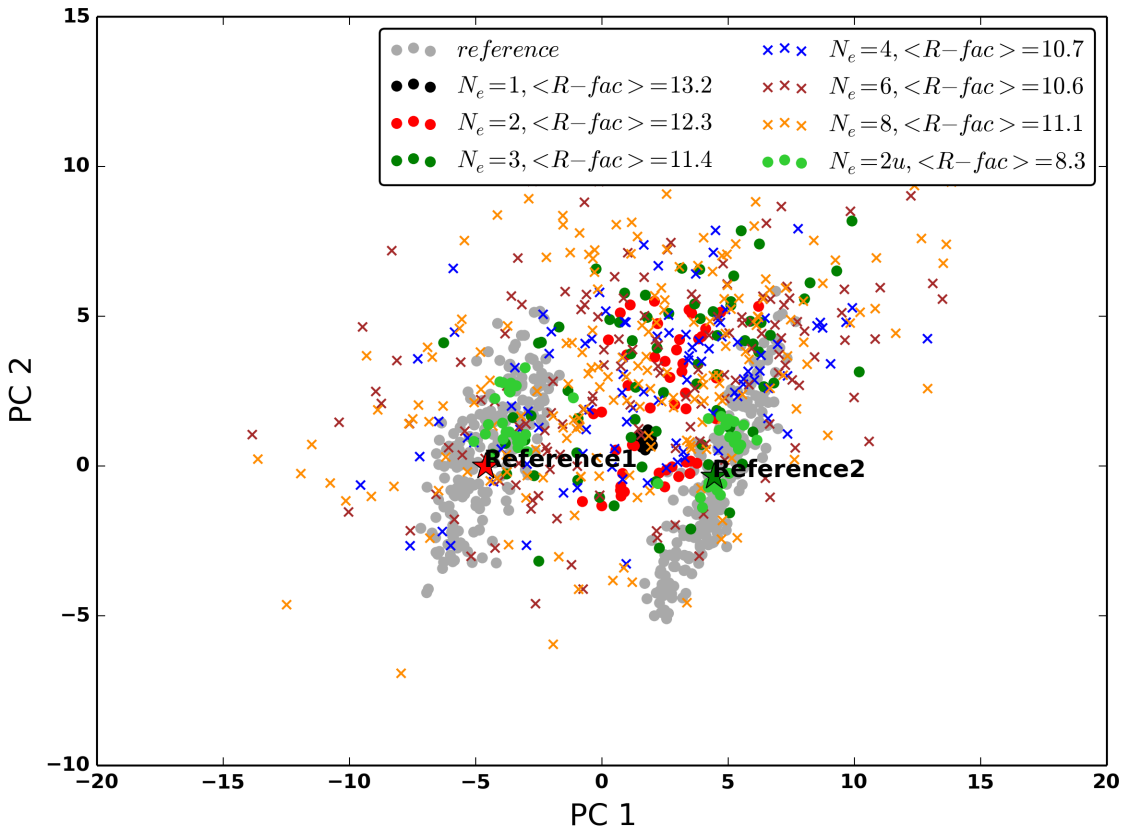

Figure 5.1: Distribution of equal weighted refinement solutions of different ensemble sizes on the principal component space defined by the first two PCs of the reference ensemble. The experimental data used to guide the refinement is generated by assigning equal weights to the reference structures. The gray dots are the distribution of the structures generated by local sampling around the reference structures by CONCOORD. The solution marked *2u* is obtained by a refinement starting with conformations close to the reference structures (one of the gray dots).

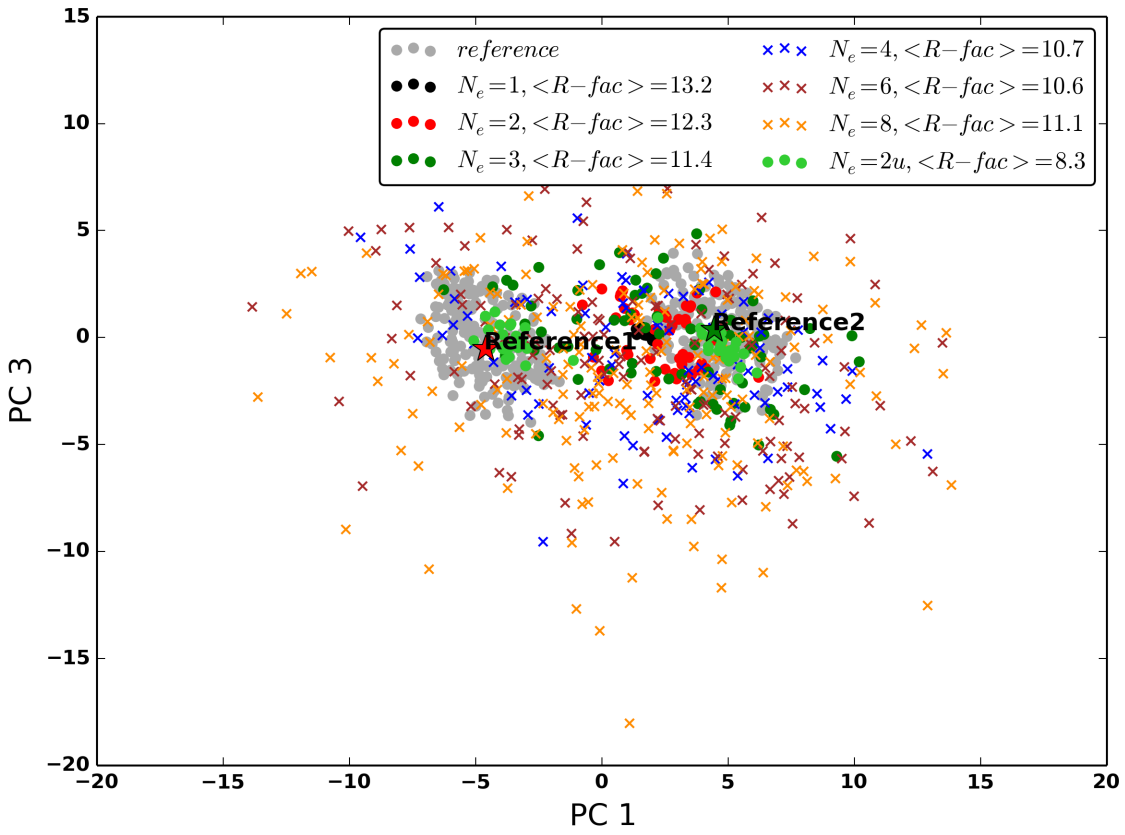

Figure 5.2: Distribution of equal weighted refinement solutions of different ensemble sizes on the principal component space defined by the first and third PCs of the reference ensemble. The experimental data used to guide the refinement is generated by assigning equal weights to the reference structures. The gray dots are the distribution of the structures generated by local sampling around the reference structures by CONCOORD. The solution marked *2u* is obtained by a refinement starting with conformations close to the reference structures (one of the gray dots).

From Figures 5.1 and 5.2, we see that single structure refinement (i.e., Ne=1), represented by black dots, resolves into an average structure that lies nearly at the center of the two reference structures and all the structures are slightly closer to second reference structure than the first reference structure (for no particular reason). Increasing the ensemble size to 2 increases the spread away from the average structure but does not necessarily sample conformations closer to the reference structures. As a matter of fact, we see that the conformations move further away from the reference structures than the average structure, as indicated by the increasing RMSD distances to the reference

structures in Table 5.1. Interestingly, we see also a decrease in cross-validation R-factor as ensemble sizes increase even though the structural quality is not getting better. Though increasing the ensemble size increases the sampling away from the average structure and has the potential of sampling conformations closer to the native states, in our testing it ends up sampling conformations that actually further away from the native state. Lange et al. (23) in their seminal work found that nearly all the X-ray conformations of Ubiquitin in the Protein Data bank (10) are close to at least one of conformations in the ensemble that they determined using RDCs and NOEs as constraints and thereby concluded that conformational selection should be favored over induced-fit model in explaining Ubiquitin binding modes. Based on our results given above, it becomes doubtful if such inferences can be drawn confidently, since conformations obtained by ensemble refinements may be further away from the native state even though they appear to reduce cross-validation R-factors.

The importance of having good initial conformations is clearly seen in the last column of Table 5.1 and the results denoted by 2u in Figure 5.1. The ensemble refinement using two structures close to the reference ensemble as starting point (denoted by *2u* in Figure 5.1 and 5.2 and in Table 5.1) achieves the best sampling around the reference states along with the lowest R-factors amongst all the refinements. A plausible explanation is that there are many local energy minima on the energy landscape defined by the RDCs (97) and a starting point(s) far away from the global minima can be stuck in these local minima, as a result of which the resulting ensemble may be far way from the reference ensemble structure wise though it reduces cross-validation R-factors.

5.4.2 Un-equal weighted Refinements

In this scenario, the reference structures in the ensemble are given unequal weights. Specifically reference 1 is given a weight of 70% and reference2 a weight of 30%. This is a close representation of the Ubiquitin native states where structures close to 1UBQ (represented by reference1) are shown to have a weight of about70% and the "switched" conformation (represented by reference2) a weight of 30% (63). Similar to the first scenario, 1UBQ-A is used as the starting conformation(s) and different ensemble sizes of 1, 2, 3, 4, 6 and 8 are tested. All conformations are given equal weights.

Explicit Weighting vs Equal weighted Refinements:

In addition to the equal weighted ensembles of different sizes, a refinement starting with 2 structures close to reference structures and with appropriate weights assigned to each conformation (70% to the conformation close to reference 1 and 30% to the conformation close to reference 2) is also performed. The results obtained using this refinement scheme, called the explicit-weighting scheme, are denoted by *2w*.

Figure 5.3 plots the distributions of the refinement solutions for different ensemble sizes in the principal component space defined by the first two PCs of the reference ensemble, along with the cross-validation R-factors for different refinements. Figure 5.4 plots the distributions in the space of the first and third PCs of the reference ensemble. The statistics on the reproduction quality of both experimental data (both working and cross-validation) and structural properties are shown in Table 5.2.

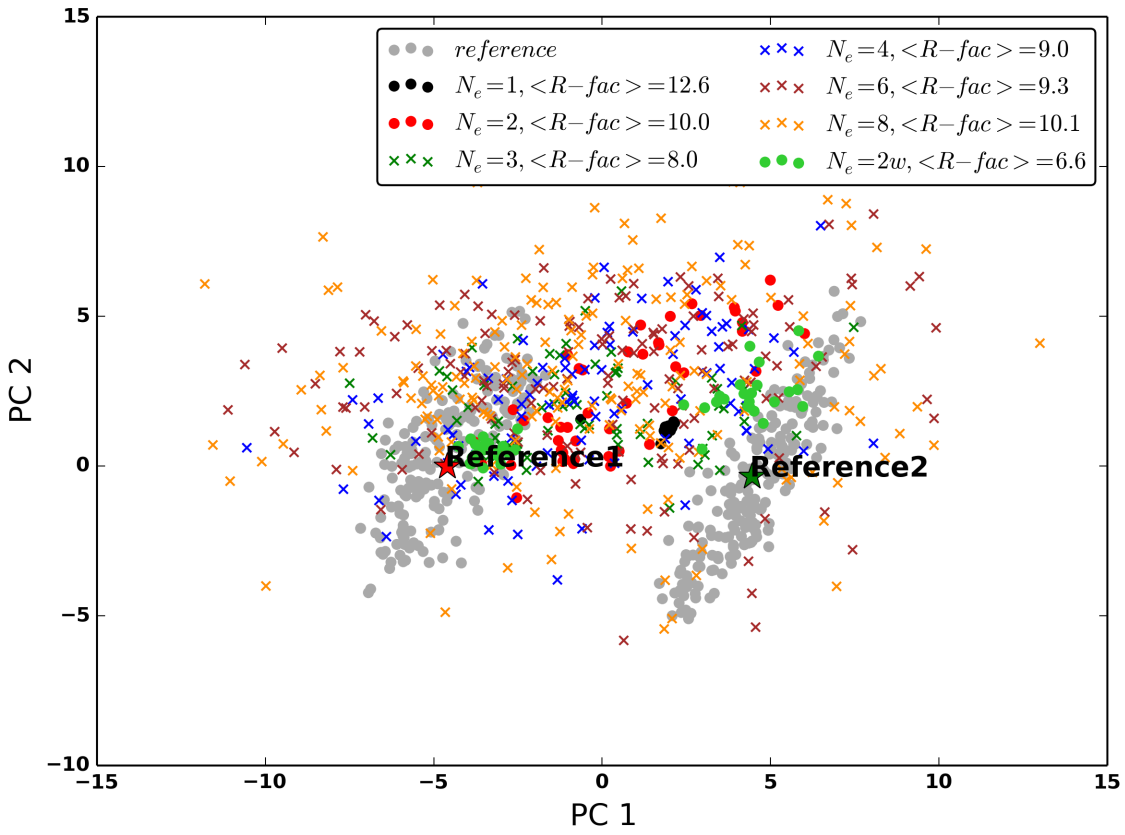

Figure 5.3: Distribution of equal weighted refinement solutions of different ensemble sizes on the principal component space defined by the first two PCs of the reference ensemble. The experimental data used to guide the refinement is generated by assigning un-equal weights to the reference structures. The gray dots are the distribution of the structures generated by local sampling around the reference structures by CONCOORD. The solution marked *2w* is obtained by a refinement starting with conformations, appropriately weighted, close to the reference structures (one of the gray dots).

As observed with the equal-weighted reference ensemble, single or average structure refinement produces solutions that lie closer to one reference conformation than other and cannot capture the underlying dynamics of the native state. Increasing the ensemble size to 2 spreads the sampling around the mean conformation but does not capture the 2 states any better than single structure refinement, as indicated by the increase in RMSD in Table II. Further increase in the ensemble size results in an even wider distribution that moves farther away from the reference structures. The importance

of having ‘good’ initial conformations can be seen again, as explicit weighted refinement (last column in Table 5.2) results in both a good sampling around the reference states and lowest cross-validation statistics. It should be noted that the quality of sampling around conformational state with lower relative population (30%, reference2) is poorer in comparison to the other conformation with a higher relative population (reference1). This is attributed to the procedure used to compute alignment tensors in XPLOR-NIH suite. XPLOR-NIH estimates the tensors based on pseudo atom approach and restrains the rhombicity and magnitude of the alignment instead of using singular value decomposition of the direction cosine matrix computed from molecular co-ordinates, which is more accurate. Though the pseudo atom approach works fairly well for many cases including single and equal weighted refinements, the estimation introduces errors when the experimental data is obtained from a weighted ensemble with one dominant conformation. The error is large enough that even starting with the reference structures themselves and with the exact weights can still result in a solution that deviates from the reference structures, with RMSDs similar to those observed in case 2w.

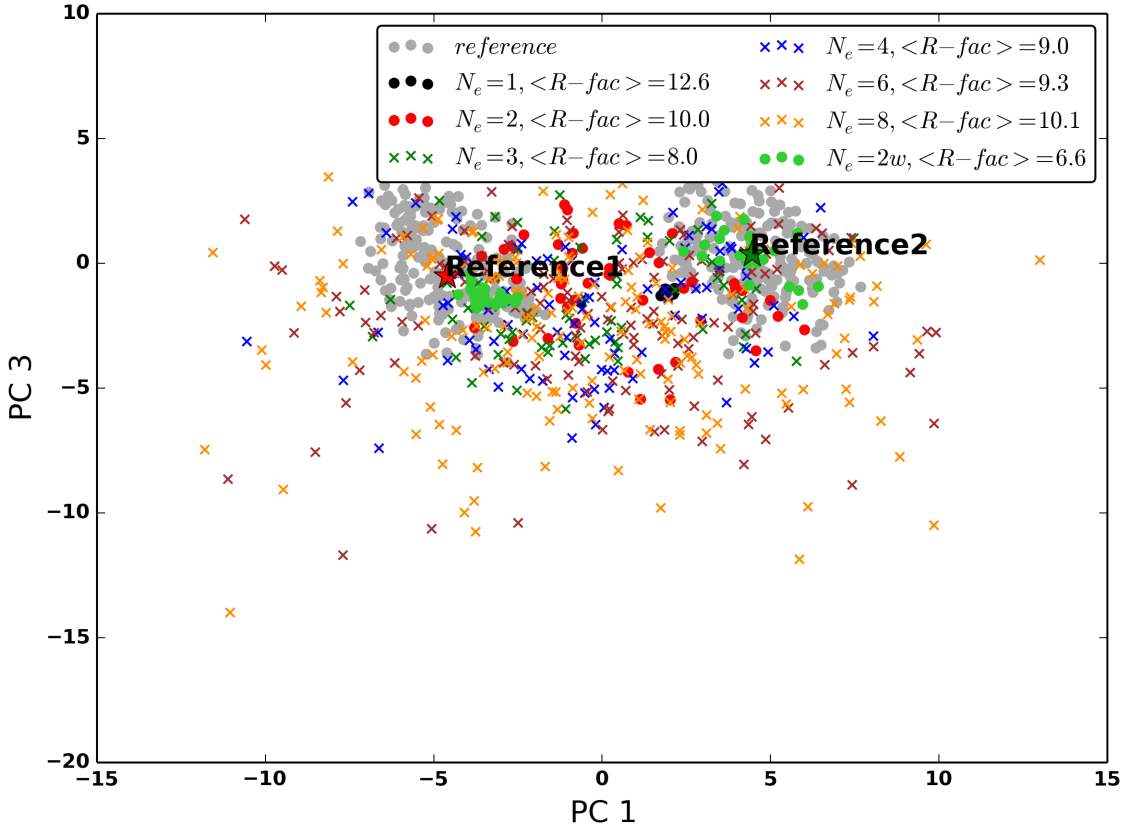

Figure 5.4: Distribution of equal weighted refinement solutions of different ensemble sizes on the principal component space defined by the first and third PCs of the reference ensemble. The experimental data used to guide the refinement is generated by assigning un-equal weights to the reference structures. The gray dots are the distribution of the structures generated by local sampling around the reference structures by CONCOORD. The solution marked *2w* is obtained by a refinement starting with conformations, appropriately weighted, close to the reference structures (one of the gray dots).

Table 5.2: Quality of solutions for an un-equal weighted reference ensemble assessed by reproduction of experimental data and structural similarity to the reference ensemble. The data shown here are the R-factors for RDCs, RMSD for NOE distance constraints and RMSD for structural similarity. The percentage of solutions closer to one reference structure than the other is computed by finding the fraction of conformations that are closer to that reference structure. The Ne number represents the ensemble size. The solution denoted by Ne value of 2w is generated by starting the refinement with initial conformations close to the reference structure and with appropriate weights.

R-factors for bond vectors	Ne=1	Ne=2	Ne=3	Ne=4	Ne=6	Ne=8	Ne=2w
NH	13.1	5.3	4.5	3.8	3.3	3.2	4.7
CAC	11.0	5.4	4.5	3.7	2.9	2.5	5.0
CHN	16.8	4.9	4.0	3.7	3.3	3.1	4.2
CN	17.0	7.4	5.6	4.9	4.0	3.7	6.3
NOE	0.13	0.08	0.07	0.06	0.07	0.06	0.07
Cross-Validation R-factors and NOE RMSD							
CAHA	12.6	10.0	8.0	9.0	9.3	10.1	6.6
NOE (free)	0.21	0.21	0.23	0.26	0.32	0.29	0.20
Structural Properties: percentages of conformations closer to reference structure 1 than 2 (or vice versa) and their mean RMSD to the closer reference structure							
% conf. closer to Ref1	4	47	61	61	51	55	50
Mean RMSD to Ref1	0.73	0.93	1.24	1.55	1.90	2.14	0.64
% conf. closer to Ref2	96	53	39	39	49	45	50
Mean RMSD to Ref2	0.66	1.22	1.27	1.56	2.06	2.53	1.1

Implicit weighting refinement scheme:

In the explicit weighting scheme, the weights are explicitly assigned to the initial conformations in the refinement protocol. Alternatively, such weights can be reflected implicitly by choosing an equal weighted ensemble whose numbers of structures that are close to each of reference structures are in proportion to the relative populations of the reference structures. To test if implicit refinement scheme would result in similar solution to the explicit weighting scheme, we test equal weighted implicit refinements of different ensemble sizes starting from 3. The ensemble composition is set to approximately reflect the population weights of the reference structures. The resulting structural quality of such refinements is given in Table 5.3 and the distributions of the conformations along the principal components of the reference ensemble are shown in Figures 5.5 and 5.6.

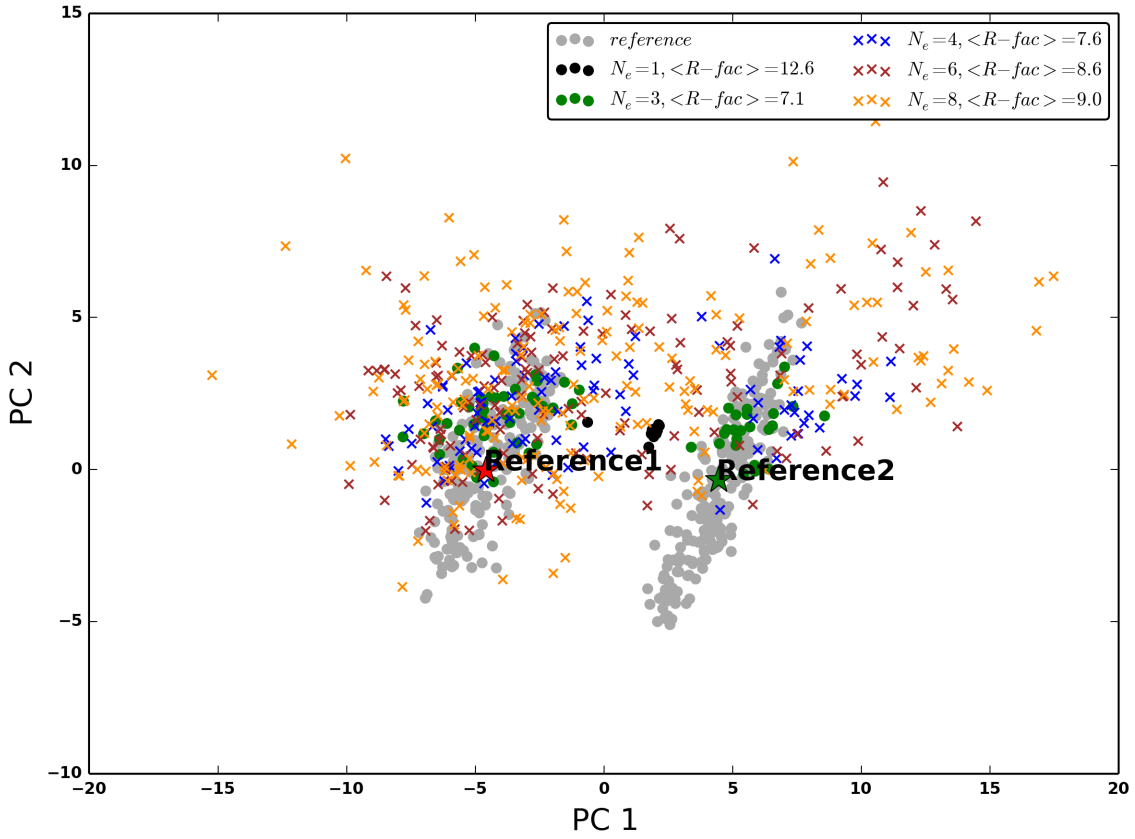

Figure 5.5: Distribution of implicit weighted refinement solutions of different ensemble sizes on the principal component space defined by the first two PCs of the reference ensemble. The experimental data used to guide the refinement is generated by assigning un-equal weights to the reference structures. The gray dots are the distribution of the structures generated by local sampling around the reference structures by CONCOORD. The solution marked *2w* is obtained by a refinement starting with conformations, appropriately weighted, close to the reference structures (one of the gray dots).

From the Figures 5.5 and 5.6, we see that the results of the minimal possible representation that encodes implicit weights, Ne=3, stay closer to the reference ensemble than larger ensembles. The quality of its solution conformations, in terms of its closeness to the reference states, is slightly worse than explicit weighting refinement in Table 5.3. Increasing the number of conformations in the implicit weighting scheme results in wider conformational distribution and larger deviation from the reference structures. This result

strongly indicates that ensembles that use sheer numbers of conformations to represent relative populations of conformation states are much more vulnerable than those that use a minimum number of conformations but with proper relative populations assigned to them.

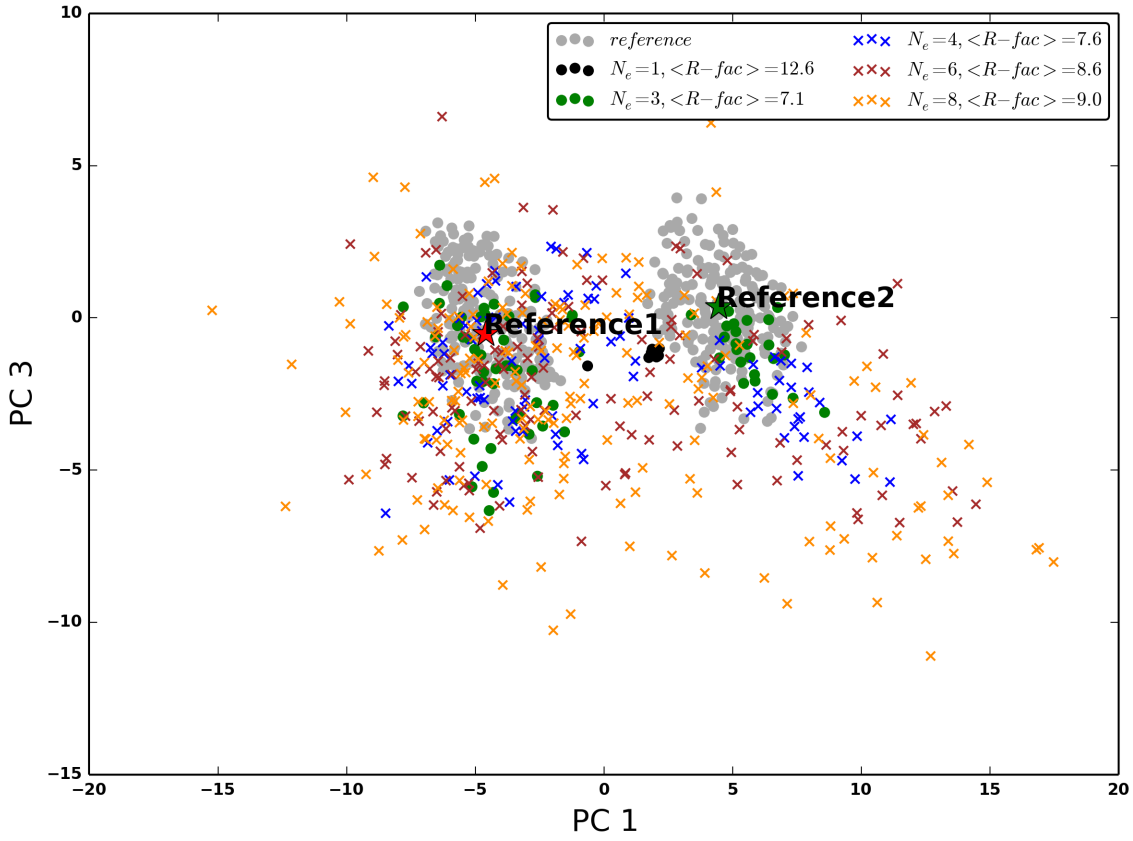

Figure 5.6: Distribution of implicit weighted refinement solutions of different ensemble sizes on the principal component space defined by the first and third PCs of the reference ensemble. The experimental data used to guide the refinement is generated by assigning un-equal weights to the reference structures. The gray dots are the distribution of the structures generated by local sampling around the reference structures by CONCOORD. The solution marked *2w* is obtained by a refinement starting with conformations close to the reference structures (one of the gray dots).

Table 5.3: The structural quality of solution generated by implicit weighting in comparison to explicit weighting in the refinement scheme. For a given ensemble size, conformations close to the reference ensemble were chosen proportional to the weights assigned in the experimental data for implicitly weighted refinement scheme. The explicitly weighted refinement scheme, denoted by Ne=2w, used only two conformations close to the reference structures along with explicit assignment of weights in the refinement protocol.

	Ne=3	Ne=4	Ne=6	Ne=8	Ne=2w
Ensemble composition. (# conf close to ref. 1, # conf close to ref. 2)	(2, 1)	(3,1)	(4,2)	(6,2)	(1,1)
% conf. close to Ref. 1	66	66	63	64	50
RMSD to Ref1	1.06	1.33	1.68	1.93	0.64
% conf. close to Ref. 2	34	34	37	36	50
RMSD to Ref2	0.92	1.35	1.78	2.14	1.1

5.5 Future Work

Ensemble refinements, instead of single structure refinements, have been proposed to better capture the structure and dynamics of the native states of biomolecules. Ensemble refinements of 8 or more replicas have been used routinely in refinements of Ubiquitin (23, 34) and Lysozyme (88) and are shown to reproduce the experimental data better than single structure refinements. Though work by Clore et al (32) identified the minimum number of conformations required to satisfy the

experimental data for Ubiquitin, there has been no detailed study on the *structural quality* of the solutions generated by ensemble refinements.

In this work by using synthetic data, we show that ensemble refinements do not necessarily improve the structure quality of the solutions but can result in conformations further away from the native states, though the ensembles may appear to be able to reproduce experimental RDC/NOE data better and even pass cross-validations. Our results show that a decrease in cross-validation R-factors does not necessarily mean the solution is moving closer structurally.

Our results show that having good initial conformations in refinement can alleviate the problem. Good initial conformations are those structurally similar to the target structures. Practically speaking, these could be existing X-ray structures of the protein being studied. Previous work in our lab has extensively focused on identifying representative conformational states among existing structures and assigning appropriate weights to them based on RDC data. We envision that ensemble refinements based on these starting conformations should further improve the quality of solutions and help better quantify the under-lying conformational states of the protein.

CHAPTER 6. CONCLUSIONS AND FUTURE WORK

Proteins are dynamic molecules and even the native states of a protein are not a single static structure but spread over a broader region of the conformation space. As a result, for many proteins, an ensemble of conformations provide a more accurate depiction of the native states (7, 9). When using ensembles to represent the dynamic nature of the native states and to gain insights into protein functional mechanisms, care must be taken in their construction, making sure that they capture the underlying native states reliably. But constructing such reliable ensembles based on experimental data constraints proves to be challenging as it is a heavily under-constrained problem and can be easily over-fitted (35).

The two key requirements for deriving high quality ensembles are:

1). Experimental data capturing the dynamics of the native states in the biologically relevant time scales.

2). Conformational sampling capturing the representative conformations of the native states.

In our work described in Chapter 2, we have shown that, given a reasonable conformational sampling, RDCs could be used as a guide to construct a high quality ensemble. Specifically, we have shown that the native state of Ubiquitin could be well represented by a few conformational states whose relative populations are determined using RDCs as constraints. In addition to the abundant experimental data that are

available, Ubiquitin has also over a hundred X-ray structures. Such collections of structures of one same protein were hypothesized to represent different conformational states of the native states (17), Putting them all together, it makes Ubiquitin an ideal model system for constructing such weighted ensembles. The conformational states identified by our method (Vammi et al., 2014) (see Chapter 2) were shown to be biologically relevant to the functions of Ubiquitin and the relative populations assigned to them matched closely to observations from a long molecular dynamics simulation (66).

Traditionally the native state of proteins is structurally represented by a single or average structure representation, or an equal weighted ensemble consisting of hundreds of conformations. Single structure representations, owing to their lack of structural variance, may suffer under-fitting while ensembles of hundreds of conformations are highly susceptible to over-fitting due their large number of model parameters. The weighted ensemble representation can be thought of as an intermediate scheme between these two representations. The weighted ensemble representation uses minimal conformational states, whose relative populations are determined using experimental data, thus minimizing the problem of over-fitting while still capturing the dynamics that a single structure misses. To assess the quality of such a representation, we have performed extensive cross validation studies using varied experimental data and compared it against the traditional structural representations in Chapter 3. The cross-validation results clearly show that weighted ensembles represent the native state equally well or in some cases better, than traditional representations.

To make the method developed in Chapter 2 more generally applicable to other proteins, two significant bottlenecks have to be solved:

1). Ubiquitin, being the model protein for NMR studies, has abundant experimental data, which is not the case for many other proteins.

2). Reasonable conformational sampling to allow reliable construction of weighted ensembles.

In Chapter 4, we resolve the first bottleneck by identifying the minimal experimental data that are required to construct weighted ensembles. We show that weighted ensembles could be constructed using as few as 6 NH RDC datasets and theses ensembles are of similar quality to the ones constructed using both NH RDCs and multi-vector RDCs. To test if this observation can be extended to other proteins, we choose Hen Egg White Lysozyme (HEWL), the model protein in X-ray crystallographic studies, as an example since there are hundreds of HEWL structures deposited in PDB and eight NH RDCs available in the literature. In our extensive cross-validation studies, the weighted ensemble representation of HEWL is shown to perform better than any other ensemble representation in literature and competes well with the average structure representation.

One may wonder if the RDC requirement for assigning relative populations could be further lowered. Though we recommend using 6 or more NH datasets to construct weighted ensembles in Chapter 4, in our studies we did observe that some combinations of even two NH RDC datasets are good enough to construct weighted ensembles.

However, currently it is not known how to identify and select such combinations. A careful study that takes into account both the dynamics present in the native states and the dynamics encoded in the RDCs is needed and could potentially minimize the NH RDC requirements for ensemble weight assignment even further.

Good quality conformational sampling remains to be the biggest bottleneck in applying the method in Chapter 4 to other proteins that have satisfied the minimal experimental data requirement. We have attempted to resolve this by resorting to structural or ensemble refinements using experimental data as constraints in Chapter 5. Our preliminary results show that ensemble refinements with arbitrary starting points, though increasing the chances of sampling conformations away from the average structure and having the potential of sampling conformations close to the native states, often ends up sampling conformations farther away from the native states.

In Chapter 5, we find also that ensemble refinements with a reasonably good starting point (i.e., good initial conformations) are able to solve the aforementioned problem of sampling conformations farther away from the native states and sample conformations close to the native states. The work presented in Chapter 2 and Chapter 4 focuses exclusively on identifying such good starting points with minimal over-fitting. An immediate future work is to use solutions obtained by the methods described in Chapters 2 and 4 as initial conformations (or starting points) for further refinements (Chapter 5).

The methods described in Chapter 2 and Chapter 4 are tolerant of structural sampling noise but they require the majorly populated conformational states of the native state represented sufficiently well in the initial pool of conformations, an essential requirement for any sample and select strategy. As long as this criterion is satisfied, the weighting protocol would identify the representative conformations and rightly assign weights. Molecular dynamics simulations have been often used to generate such ensembles (62) for proteins and nucleic acids (36) and it would be interesting to see if such ensembles result in similar solutions.

Experimental data capturing the dynamics at different resolutions and time scales are integrated to further enhance our understanding of dynamics of bio-molecules (81, 98, 99). Typically solution scattering profiles are used along with residual dipolar couplings to provide orientation restraints (100). Solution scattering profiles themselves contain low-resolution information on the dynamics of protein and can be used for large proteins that are beyond the regime of NMR experiments. Similar to our work using RDCs, a lot of work on obtaining the minimal weighted ensemble using solution scatter profiles has been done by different groups (101-103). Another potential direction for future research is to extend our RDCs-based method and use solution scattering profiles instead to construct weighted ensembles.

BIBLIOGRAPHY

1. Frauenfelder H, Sligar SG, & Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254(5038):1598-1603.
2. Miyashita O, Onuchic JN, & Wolynes PG (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences* 100(22):12570-12575.
3. Bryngelson JD, Onuchic JN, Socci ND, & Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics* 21(3):167-195.
4. Dill KA & Chan HS (1997) From Levinthal to pathways to funnels. *Nat. Struct. Biol.* 4:10-19.
5. Austin RH, Beeson KW, Eisenstein L, Frauenfelder H, & Gunsalus IC (1975) Dynamics of ligand binding to myoglobin. *Biochemistry* 14(24):5355-5373.
6. Boehr DD, Nussinov R, & Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* 5(11):789-796.
7. Furnham N, Blundell TL, DePristo MA, & Terwilliger TC (2006) Is one solution good enough? *Nat Struct Mol Biol* 13(3):184-185.
8. Karplus M & McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nature Struct Biol* 9(9):646-652.
9. Phillips GN (2009) Describing protein conformational ensembles: beyond static snapshots. *F1000 Biol Rep* 1.
10. Berman HM, *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235-242.
11. Burling FT & Brunger AT (1994) Thermal motion and conformational disorder in protein crystal-structures - comparison of multi-conformer and time-averaging models. *Isr J Chem* 34:165-175.
12. Fraser JS, *et al.* (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proceedings of the National Academy of Sciences* 108(39):16247-16252.
13. van den Bedem H, Dhanik A, Latombe J-C, & Deacon AM (2009) Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. *Acta Crystallogr D* 65(10):1107-1117.
14. Lang PT, Holton JM, Fraser JS, & Alber T (2014) Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proceedings of the National Academy of Sciences* 111(1):237-242.
15. DePristo MA, de Bakker PI, & Blundell TL (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure* 12(5):831-838.

16. Knight JL*, et al.* (2008) Exploring structural variability in X-ray crystallographic models using protein local optimization by torsion-angle sampling. *Acta Crystallogr D* 64(4):383-396.
17. Best RB, Lindorff-Larsen K, DePristo MA, & Vendruscolo M (2006) Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci U S A* 103(29):10901-10906.
18. Wüthrich K (1995) *NMR in structural biology: a collection of papers by Kurt Wüthrich* (World Scientific Pub Co Inc).
19. Crippen GM & Havel TF (1988) *Distance geometry and molecular conformation* (Research Studies Press Taunton, England).
20. Kontaxis G & Bax A (2001) Multiplet component separation for measurement of methyl 13C-1H dipolar couplings in weakly aligned proteins. *J Biomol NMR* 20(1):77-82.
21. Prestegard J (1998) New techniques in structural NMR—anisotropic interactions. *Nature Structural & Molecular Biology* 5:517-522.
22. Tolman JR, Flanagan JM, Kennedy MA, & Prestegard JH (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proceedings of the National Academy of Sciences* 92(20):9279-9283.
23. Lange OF*, et al.* (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* 320(5882):1471-1475.
24. Clore GM & Schwieters CD (2006) Concordance of Residual Dipolar Couplings, Backbone Order Parameters and Crystallographic B-factors for a Small α/β Protein: A Unified Picture of High Probability, Fast Atomic Motions in Proteins. *Journal of Molecular Biology* 355(5):879-886.
25. Chen K & Tjandra N (2012) The use of residual dipolar coupling in studying proteins by NMR. *NMR of Proteins and Small Biomolecules*, (Springer), pp 47-67.
26. Torda AE & van Gunsteren WF (1991) The refinement of NMR structures by molecular dynamics simulation. *Computer Physics Communications* 62(2):289-296.
27. Scheek RM, Torda AE, Kemmink J, & van Gunsteren WF (1991) Structure Determination by NMR - the Modeling of NMR Parameters as Ensemble Averages. *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy*:209-217.
28. Schwieters CD, Kuszewski JJ, Tjandra N, & Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance* 160(1):65-73.
29. Cornilescu G, Marquardt JL, Ottiger M, & Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc* 120(27):6836-6837.
30. Maltsev AS, Grishaev A, Roche J, Zasloff M, & Bax A (2014) Improved Cross Validation of a Static Ubiquitin Structure Derived from High Precision Residual Dipolar Couplings Measured in a Drug-Based Liquid Crystalline Phase. *Journal of the American Chemical Society* 136(10):3752-3755.

31. Clore GM & Schwieters CD (2004) Amplitudes of protein backbone dynamics and correlated motions in a small α/β protein: correspondence of dipolar coupling and heteronuclear relaxation measurements. *Biochemistry* 43(33):10678-10691.
32. Clore GM & Schwieters CD (2004) How Much Backbone Motion in Ubiquitin Is Required To Account for Dipolar Coupling Data Measured in Multiple Alignment Media as Assessed by Independent Cross-Validation? *Journal of the American Chemical Society* 126(9):2923-2938.
33. Richter B, Gsponer J, Varnai P, Salvatella X, & Vendruscolo M (2007) The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J Biomol NMR* 37(2):117-135.
34. Fenwick RB*, et al.* (2011) Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. *Journal of the American Chemical Society* 133(27):10336-10339.
35. Ángyán AF & Gáspári Z (2013) Ensemble-based Interpretations of NMR structural data to describe protein internal dynamics. *Molecules* 18(9):10548-10567.
36. Salmon L, Yang S, & Al-Hashimi HM (2014) Advances in the Determination of Nucleic Acid Conformational Ensembles. *Annual review of physical chemistry* 65:293-316.
37. Chen Y, Campbell SL, & Dokholyan NV (2007) Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophysical journal* 93(7):2300-2306.
38. Fisher CK, Huang A, & Stultz CM (2010) Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *Journal of the American Chemical Society* 132(42):14919-14927.
39. Choy WY & Forman-Kay J (2001) Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J Mol Biol* 308:1011-1032.
40. Brüschweiler R, Blackledge M, & Ernst R (1991) Multi-conformational peptide dynamics derived from NMR data: a new search algorithm and its application to antamanide. *Journal of biomolecular NMR* 1(1):3-11.
41. Bonvin AMJJ & Brunger AT (1996) Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? *J Biomol NMR* 7:72-76.
42. Frauenfelder H, McMahon BH, Austin RH, Chu K, & Groves JT (2001) The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proc Natl Acad Sci U S A* 98(5):2370-2374.
43. Levin EJaK, D. A. and Wesenberg, G. E. and Phillips, G. N., Jr. (2007) Ensemble refinement of protein crystal structures: validation and application. *Structure* 15(9):1040-1052.
44. Frauenfelder HaSSGaWPG (1991) The energy landscapes and motions of proteins. *Science* 254(5038):1598-1603.
45. Miyashita OaOJNaWPG (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. 100(22):12570-12575.

46. Clore GMaSCD (2004) Amplitudes of Protein Backbone Dynamics and Correlated Motions in a Small Î±/Î² Protein:â Correspondence of Dipolar Coupling and Heteronuclear Relaxation Measurementsâ. *Biochemistry* 43(33):10678-10691.
47. Clore GMaSCD (2004) How Much Backbone Motion in Ubiquitin Is Required To Account for Dipolar Coupling Data Measured in Multiple Alignment Media as Assessed by Independent Cross-Validation? *Journal of the American Chemical Society* 126(9):2923-2938.
48. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, & Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. *Nature* 433(7022):128-132.
49. Shao J, Tanner SW, Thompson N, & Cheatham TE (2007) Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation* 3(6):2312-2334.
50. Daura X*, et al.* (1999) Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition* 38(1-2):236-240.
51. Eastwood MP, Hardin C, Luthey-Schulten Z, & Wolynes PG (2001) Evaluating protein structure-prediction schemes using energy landscape theory. *IBM Journal of Research and Development* 45(3.4):475-497.
52. Tolman JRaFJMaKMAaPJH (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proceedings of the National Academy of Sciences* 92(20):9279-9283.
53. Lawson CL & Hanson RJ (1995) *Solving least squares problems* (SIAM, Philadelphia).
54. Word JM, Lovell SC, Richardson JS, & Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology* 285(4):1735-1747.
55. Lakomek NA*, et al.* (2008) Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics. *J Biomol NMR* 41(3):139-155.
56. Ottiger M & Bax A (1998) Determination of relative N-H-N N-C', C-alpha-C', andC(alpha)-H-alpha effective bond lengths in a protein by NMR in a dilute liquid crystalline phase. *J Am Chem Soc* 120(47):12334-12341.
57. Lakomek NA, Carlomagno T, Becker S, Griesinger C, & Meiler J (2006) A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. *J Biomol NMR* 34(2):101-115.
58. Donald Hamelberg and John Mongan and JAM (2004) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics* 120(24):11919-11929.
59. de Groot BL*, et al.* (1997) Prediction of protein conformational freedom from distance constraints. *Proteins* 29(2):240-251.
60. Piana SaL-LKaSDE (2013) Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences* 110(15):5915-5920.

61. Huang KYaAGAaTLaMA (2011) The structure of human ubiquitin in 2-methyl-2,4-pentanediol: A new conformational switch. *Protein Science* 20(3):630--639.
62. Markwick PRLaBGaSLaMJAaNMaBM (2009) Toward a Unified Representation of Protein Structural Dynamics in Solution. *Journal of the American Chemical Society* 131(46):16968-16975.
63. Vammi V, Lin T-L, & Song G (2014) Enhancing the quality of protein conformation ensembles with relative populations. *Journal of Biomolecular NMR* 58(3):209-225.
64. Fisher CK, Huang A, & Stultz CM (2010) Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *Journal of the American Chemical Society* 132(42):14919-14927.
65. Tripathy C, Zeng J, Zhou P, & Donald BR (2012) Protein loop closure using orientational restraints from NMR data. *Proteins: Structure, Function, and Bioinformatics* 80(2):433-453.
66. Piana S, Lindorff-Larsen K, & Shaw DE (2013) Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences* 110(15):5915-5920.
67. Sidhu A, Surolia A, Robertson AD, & Sundd M (2011) A hydrogen bond regulates slow motions in ubiquitin by modulating a β-turn flip. *Journal of molecular biology* 411(5):1037-1048.
68. Huang KY, Amodeo GA, Tong L, & McDermott A (2011) The structure of human ubiquitin in 2 ‐ methyl ‐ 2, 4 ‐ pentanediol: A new conformational switch. *Protein Science* 20(3):630-639.
69. Vijay-Kumar S, Bugg CE, & Cook WJ (1987) Structure of ubiquitin refined at 1.8 Åresolution. *Journal of molecular biology* 194(3):531-544.
70. Liu Y & Prestegard J (2010) A device for the measurement of residual chemical shift anisotropy and residual dipolar coupling in soluble and membrane-associated proteins. *Journal of Biomolecular NMR* 47(4):249-258.
71. Saitô H, Ando I, & Ramamoorthy A (2010) Chemical shift tensor–The heart of NMR: Insights into biological aspects of proteins. *Progress in nuclear magnetic resonance spectroscopy* 57(2):181.
72. Cornilescu G & Bax A (2000) Measurement of proton, nitrogen, and carbonyl chemical shielding anisotropies in a protein dissolved in a dilute liquid crystalline phase. *Journal of the American Chemical Society* 122(41):10143-10154.
73. LeMaster DM, Anderson JS, & Hernández G (2009) Peptide conformer acidity analysis of protein flexibility monitored by hydrogen exchange. *Biochemistry* 48(39):9256-9265.
74. Hernández G, Anderson JS, & LeMaster DM (2010) Assessing the native state conformational distribution of ubiquitin by peptide acidity. *Biophysical chemistry* 153(1):70-82.
75. Hubbard SJ & Thornton JM (1993) Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London* 2(1).

76. McDonald IK & Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology* 238(5):777-793.
77. Li L*, et al.* (2012) DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC biophysics* 5(1):9.
78. MacKerell AD*, et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B* 102(18):3586-3616.
79. Svergun DI & Koch MHJ (2003) Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics* 66(10):1735.
80. Putnam CD, Hammel M, Hura GL, & Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly reviews of biophysics* 40(03):191-285.
81. Schwieters CD & Clore GM (2007) A physical picture of atomic motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data. *Biochemistry* 46(5):1152-1166.
82. Grishaev A, Wu J, Trewhella J, & Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *Journal of the American Chemical Society* 127(47):16621-16628.
83. Svergun D, Barberato C, & Koch MHJ (1995) CRYSOL-a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography* 28(6):768-773.
84. Grishaev A, Guo L, Irving T, & Bax A (2010) Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. *Journal of the American Chemical Society* 132(44):15484-15486.
85. Massi F, Grey MJ, & Palmer AG (2005) Microsecond timescale backbone conformational dynamics in ubiquitin studied with NMR R1ρ relaxation experiments. *Protein science* 14(3):735-742.
86. Makowski L (2010) Characterization of proteins with wide-angle X-ray solution scattering (WAXS). *Journal of structural and functional genomics* 11(1):9-19.
87. Schwieters CDaCGM (2007) A Physical Picture of Atomic Motions within the Dickerson DNA Dodecamer in Solution Derived from Joint Ensemble Refinement against NMR and Large-Angle X-ray Scattering Data. *Biochemistry* 46(5):1152-1166.
88. De Simone A, Montalvao RW, Dobson CM, & Vendruscolo M (2013) Characterization of the interdomain motions in hen lysozyme using residual dipolar couplings as replica-averaged structural restraints in molecular dynamics simulations. *Biochemistry* 52(37):6480-6486.
89. F Ángyán A & Gáspári Z (2013) Ensemble-Based Interpretations of NMR Structural Data to Describe Protein Internal Dynamics. *Molecules* 18(9):10548-10567.

90. Boyd J & Redfield C (1999) Characterization of 15N chemical shift anisotropy from orientation-dependent changes to 15N chemical shifts in dilute bicelle solutions. *Journal of the American Chemical Society* 121(32):7441-7442.
91. Higman VA, Boyd J, Smith LJ, & Redfield C (2011) Residual dipolar couplings: are multiple independent alignments always possible? *Journal of biomolecular NMR* 49(1):53-60.
92. Yao L, Vögeli B, Torchia DA, & Bax A (2008) Simultaneous NMR study of protein structure and dynamics using conservative mutagenesis. *The Journal of Physical Chemistry B* 112(19):6045-6056.
93. Vaney MC, Maignan S, Ries-Kautt M, & Ducruix A (1996) High-resolution structure (1.33 A) of a HEW lysozyme tetragonal crystal grown in the APCF apparatus. Data and structural comparison with a crystal grown under microgravity from SpaceHab-01 mission. *Acta Crystallographica Section D: Biological Crystallography* 52(3):505-517.
94. Schwalbe H*, et al.* (2001) A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Science* 10(4):677-688.
95. Dominguez C, Boelens R, & Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* 125(7):1731-1737.
96. Li D-W & Brüschweiler R (2011) Iterative optimization of molecular mechanics force fields from NMR data of full-length proteins. *Journal of Chemical Theory and Computation* 7(6):1773-1782.
97. Bax A (2003) Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Science* 12(1):1-16.
98. Shen Y, Vernon R, Baker D, & Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. *Journal of biomolecular NMR* 43(2):63-78.
99. Ihms EC & Foster MP (2015) MESMER: Minimal Ensemble Solutions to Multiple Experimental Restraints. *Bioinformatics*:btv079.
100. Grishaev A, Ying J, Canny MD, Pardi A, & Bax A (2008) Solution structure of tRNAVal from refinement of homology model against residual dipolar coupling and SAXS data. *Journal of biomolecular NMR* 42(2):99-109.
101. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, & Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *Journal of the American Chemical Society* 129(17):5656-5664.
102. Pelikan M, Hura GL, & Hammel M (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. *General physiology and biophysics* 28(2):174.
103. Rozycki B, Kim YC, & Hummer G (2011) SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* 19(1):109 - 116.