Chemical Shift Optimization and Ensemble Averaging in Protein NMR Spectroscopy

Dissertation zur Erlangung des Doktorgrades
der Naturwissenschaften (dr. rer. nat.)
der Naturwissenschaftlichen Fakultt III
Biologie und Vorklinische Medizin
der Universität Regensburg

vorgelegt von
Kumaran Baskaran

aus
Tirupattur, India

im Jahr 2010

Das Promotionsgesuch wurde eingereicht am : 08.02.2010

Die Arbeit wurde angeleitet von : Prof. Dr. Dr. Hans Robert Kalbitzer

Prüfungsausschuss:

Vorsitzende : Prof. Dr. med. Rosemarie Baumann

Erstgutachter : Prof. Dr. Dr. Hans Robert Kalbitzer

Zweitgutachter : PD Dr. Wolfram Gronwald

Drittprüfer : Prof. Dr. Jaroslav Fabian

ABSTRACT

A problem often encountered in multidimensional NMR-spectroscopy is that an existing chemical shift list of a protein has to be used to assign an experimental spectrum but does not fit sufficiently well for a safe assignment. A similar problem occurs when temperature or pressure series of n-dimensional spectra are to be evaluated automatically. Two slightly different algorithms, AUREMOL-SHIFTOPT1 and AUREMOL-SHIFTOPT2 have developed here that fulfill this task. Their performance is analyzed employing a set of simulated and experimental two-dimensional and three-dimensional spectra obtained from three different proteins. Peak probability and atom type based weighted averaging is introduced in order to reduce the influence of the wrong assignment during the assignment process.

Chemical shift prediction programs often use a single energy minimized structure as input, but ensemble averaging of chemical shifts gives better prediction values irrespective of the prediction method. This is in agreement with the fact that proteins in solution occur in multiple conformational states in fast exchange on the chemical shift time scale. However, in contrast to the real conditions in solution at ambient temperatures, the chemical shift prediction methods seems optimal to predict the lowest energy ground state structure that is only weakly populated under these conditions. An analysis of the data shows that a chemical shift prediction can be used as measure to define the minimum size of the structural bundle required for a faithful description of the structural ensemble.

Reliable homo and heteronuclear chemical shift distributions are required for the automated assignment procedures. However, the statistics derived from the Biological Magnetic Resonance Bank (BMRB) is not clean and is not structurally unbiased. Therefore, refined chemical shift statistics was created from a structural database of non-homologous proteins (Nh3D) that comprises 806 different three-dimensional structures. The chemical shift data base was created by calculating the resulting chemical shifts with the prediction programs SHIFTS and SHIFTX. Analysis of the obtained data set shows that unbiased chemical shift statistics improves the a priori probability values for resonance assignment, removes ambiguities in assignment to certain level and helps to make stereochemical assignments.

ACKNOWLEDGEMENTS

First of all, I would like to offer my sincere thanks to **Prof. Dr. Dr. Hans Robert Kalbitzer**, for his guidance during my research and study at University of Regensburg. His perpetual energy and enthusiasm in research had motivated all his advisees, including me.

I thank **PD. Dr. Wolfram Gronwald** for his initial guidance and fruitful discussions with him. Our discussions during the coffee break were quite useful and productive. A special thanks to him, for his friendly welcome at Munich airport when I first time arrived to Germany and for the typical German dinner at his home on the very next day, which made me feel like home.

I would like to thank **Mrs Ingrid Kulbartz**, our department secretary, for her immense support in the administrative level. I was overwhelmed by her hospitality and questions like '*Is everything fine? Do you need any help?*'during my initial days in Regensburg. She welcomes everyone to her office with a smile and everyone leaves her office with a smile.

I owe my sincere thanks to the International Graduate School GK 638 **Non linearity and non equilibrium in condensed matter physics** and **Deutsche Forschungsgemeinschaft (German Research Foundation)** for the scholarship and funding. I thank the Professors and the members of the Graduate School for organising interdisciplinary seminars and annual workshops.

This thesis would not have been possible without the support of my colleagues, especially the AUREMOL Development Team (ADT). It is a pleasure to thank **Dr. Konrad Bruner**, **Dr. Carolina cano**, **Dr. Jochen Trenner**, **Dr. Andre Fischer** and **Mr. Harald Donaubaurer** for their co-operation and support during my thesis work. Thanks to **Dr. Cloudia Munte** for providing excellent experimental data for my calculations.

I would like to thank Konrad, Jochen, Alexander and Torsten for teaching me '*schafkopf*'(German card game) and table tennis. Outings and dinners with them made my stay in Regensburg a memorable one. The most memorable one was the wedding of Konarad. I was so lucky to witness a traditional German wedding ceremony. I should really thank Konrad for the invitation to his wedding.

I would like to thank my IITM classmate and a good friend **Eswar**, who is now a graduate student at TU Dortmund and my sister **Amudha**, who is a graudate student at FHI Berlin for their constant encoragement and fruitful discussions.

I thank the Indian community at the University of Regensburg for their weekend parties and site seeings in and around Germany.

I owe my deepest gratitude to my family in India, for their support and encouragement. My father, **Mr. Baskaran** often says '*I didn't get a chance to go for higher studies, now I want you to do that!*'. These words are still echoing in my mind. Even though my mother **Mrs. Lakshmi** doesn't like to see her son leaving the home, she understood the need and offered her unconditional love and support to me. My brother **Mr. Nagarajan**, who is more like a friend, encouraged me to come to Germany to pursue my higher studies. I thank all of them for their moral support.

The above mentioned is just a short list. There are countless people who helped me to achieve this milestone in my life. Due to the space restriction I could mention only few of them. I apologize for those, whose names are missing and and thank them as well.

Kumaran Baskaran

"கேடில் விழுச்செல்வம் கல்வி யொருவற்கு
மாடல்ல மற்றை யவை"

-திருக்குறள் (400)
திருவள்ளுவர் (~ கி.மு. முதலாம் நூற்றாண்டு)

"Learning is excellence of wealth that none destroy
To man nought else affords reality of joy"

- ThirukkuRaL (400)
Thiruvalluvar (~ BC 1st Century)

Dedicated to my elementary school teachers....

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

BMRB	Biological Magnetic Resonace data Bank
COSY	COrrelation SpectroscopY
CSI	Chemical Shift Index
DG	Distance Geometry
HPr	Histidine Containing Phosphocarrier Protein
HSQC	Heteronuclear Single Quantum Coherence
KDE	Kernel Density Estimation
MCD	Main Chain Directed
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhouser Effect
NOESY	Nuclear Overhouser Effect Spectroscopy
PDB	Protein Data Bank
PDF	Probability Density Function
RDC	Residual Dipolar Coupling
RMSD	Root Mean Square Deviation
rMD	restrained molecular dynamics
SA	Simulated Annealing
TOCSY	Total Crrelation SpectroscopY

CHAPTER 1

INTRODUCTION

1.1 Proteins

Proteins are biological macromolecules which are essential parts of organisms and participate virtually in every process within cells. They are made up of linear chain of amino acids joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the nucleotide sequences of the corresponding gene. In general, the genetic code specifies 20 standard amino acids. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism.

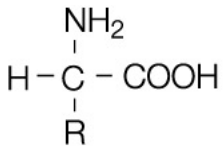


Figure 1.1: Basic structure of amino acid

What is remarkable is that more than 30, 000 proteins in our bodies are produced from a set of only 20 building blocks, known as amino acids. All amino acids have the same basic structure, an amino group, a carboxyl group and a hydrogen atom, but differ due to the presence of a side-chain known as R(Figure:1.1); This side-chain varies dramatically between amino acids, from a simple hydrogen atom in the amino acid glycine to a complex structure found in tryptophan. Depending on the nature of the side-chain, an amino acid can be hydrophilic (water-attracting) or hydrophobic (water-repelling), acidic or basic; and it is this diversity in side-chain properties that gives each protein its specific character.

1.1.1 Native structure

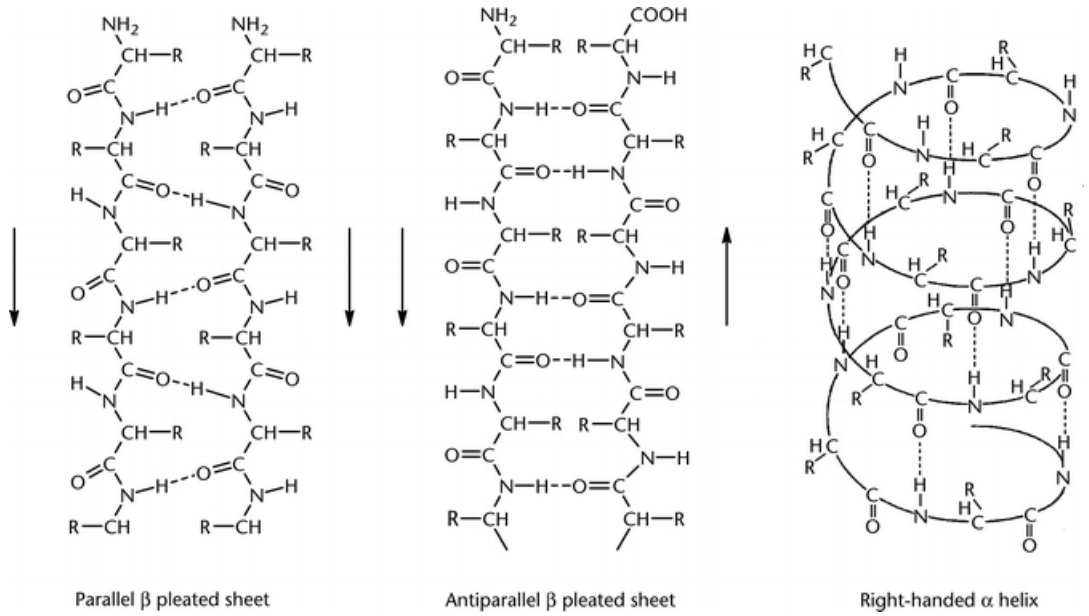


Figure 1.2: Secondary structure elements

The sequence of amino acids in a protein defines its primary structure. The blueprint for each amino acid is laid down by sets of three letters, known as base triplets, that are found in the coding regions of genes. These base triplets are recognized by ribosomes, the protein building sites of the cell, which create and successively join the amino acids together. This is a remarkably quick process: a protein of 300 amino acids will be made in little more than a minute. The result is a linear chain of amino acids, but this only becomes a functional protein when it folds into its three-dimensional native structure. This occurs through an intermediate form, known as secondary structure, the most common of which are the α-helix and the β-sheet (Figure:1.2). These secondary structures are formed by a small number of amino acids that are close together, which then, in turn, interact, fold and coil to produce the tertiary structure that contains its functional regions (called domains). Figure 1.3 shows the native 3D structure of Histidine Containing Phosphocarrier Protein from Staphylococcus aureus [Maurer et al., 2004]. Although it is possible to deduce the primary structure of a protein from a gene's sequence, its tertiary structure cannot be

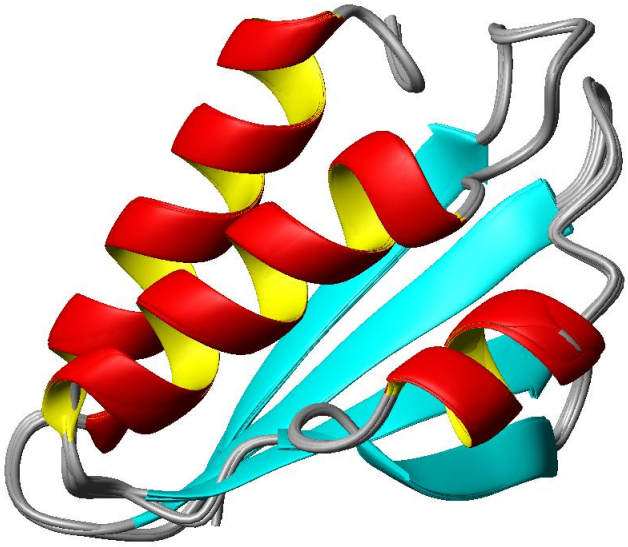


Figure 1.3: HPr from Staphylococcus aureus

determined straight away, but could be possible to make predictions when more tertiary structures are submitted to databases. It can only be determined by complex experimental analyses and, at present, this information is only known for about 10% of proteins. It is therefore not yet known how an amino-acid chain folds into its tertiary structure in the short time scale (fractions of a second) that occurs in the cell. So, there is a huge gap in our knowledge of how we move from protein sequence to function in living organisms: the line of sight from the genetic blueprint for a protein to its biological function is blocked by the impenetrable jungle of protein folding, and some researchers believe that clearing this jungle is the most important task in biochemistry at present.

1.2 Structure determination

There are many techniques to study different aspects of structures of cellular components, but two techniques allow a resolution at the level of distinguishing individual atoms: X-ray crystallography and Nuclear Magnetic Resonance or NMR technique.

1.2.1 X-Ray crystallography

X-ray crystallography has been used to determine the structure of inorganic and organic crystals since the early years of the last century. The technique was first used for the elucidation of salt crystal structure, which for example gave Linus Pauling the instrumentation to study atomic distances from which he developed his theory of the chemical bond (combining structural information with quantum mechanical calculations). From the knowledge obtained from salt crystals Pauling, who focused his attention on protein structures, proposed the alpha helical and beta strand secondary structures, both of which have been confirmed by X-ray crystallographic analysis for the first time using crystals of myoglobin and hemoglobin in the early sixties by Kendrew and co-workers [Kendrew et al., 1960]

X-ray structures are high resolution structures enabling the distinction of two points in space as close as $2A^o$ apart. Yet they depict a static structure, the result of a technique which requires large, stable protein crystals, within which each protein unit is lined up in a regular lattice. It was soon recognized that these static structures didn't really help explaining function because the structures are mostly the average of millions of identical units. 'Loose' structural parts like surface loops often failed to be resolved leaving some protein structures incomplete. The development of nuclear magnetic resonance techniques, NMR, could be used to overcome this problem. In contrast to protein crystals needed for X-ray diffraction, NMR made use of protein solutions allowing for the determination of structures at very short time ranges. Consequently those flexible loop and domain structures could be solved successfully.

1.2.2 Nuclear Magnetic Resonance(NMR)

NMR spectroscopy plays a major role in the determination of the structures and dynamics of proteins and other biological macromolecules. Chemical shifts are the most readily and accurately measurable NMR parameters, and they reflect with great specificity the conformations of native and non-native states of proteins.The chemical shift of a nucleus is the difference between the resonance frequency of the nucleus and a standard atom, normalized to the spectrometer frequency. This quantity is reported in ppm and given the

symbol δ,

$$\delta = \left(\frac{\omega - \omega_{ref}}{\omega_{ref}} \right) \times 10^6 ppm \tag{1.1}$$

In biological NMR spectroscopy, this standard is often 4, 4-dimethyl-4-silapentane-1-sulfonic acid, $C_6H_{16}O_3SSi$, abbreviated DSS[Nowick et al., 2003]. The chemical shift

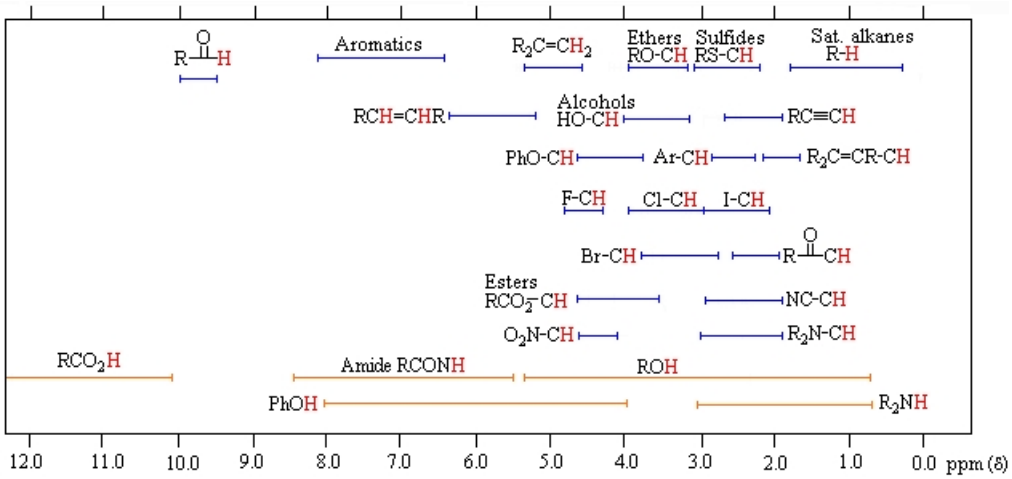


Figure 1.4: ^1H chemical shift ranges

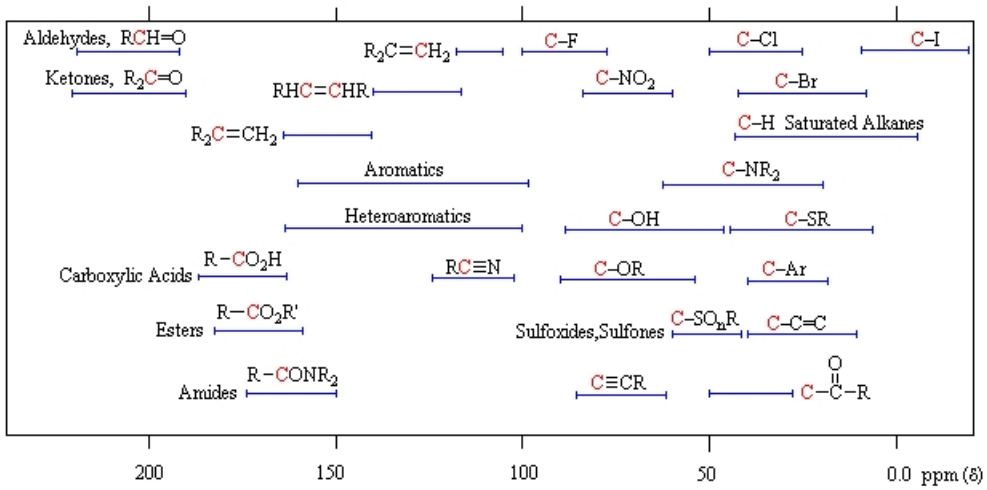


Figure 1.5: ^{13}C Chemical shift ranges

is a very precise metric of the chemical environment around a nucleus. For example, the hydrogen chemical shift of a CH_2 hydrogen next to a Cl will be different than that of a

CH_3 next to the same Cl. The expected range for ^1H and ^{13}C chemical shifts in various chemical shift environment is shown in Figures 1.4, 1.5

Unlike X-ray crystallography, NMR spectroscopy method is a not straight forward method for structure determination. In addition to experimental data, reasonable amount of data base analysis and statistics are also required. Even though there are great improvements in the instrumentation and the experimental aspects, statistical informations are essential to make some intelligent guesses in the structure determination process. The complexity involved in structure determination process and the size of experimental data require computer assistance for this process. Researchers are now mainly focused on automation techniques for structure determination process. In such cases, statistical information plays predominant roll. It is also necessary that the data sets used to calculate the statistical parameters should be fairly unbiased.

In general, statistics combined with conditional probability method are essential part of computational biology and biophysics. The deterministic nature of the classical world has been over ruled by the probabilistic nature of the quantum world. This is true for the proteins also. The ensemble representation from NMR methods against the crystal structure from X-ray methods, demonstrates the true nature of the proteins.

1.3 Structure determination by NMR

The various steps involved in protein structure determination by NMR spectroscopy is shown in Figure 1.6.The NMR data analysis consists of three steps.

- Resonance assignment
- Restraint calculation
- Molecular dynamics.

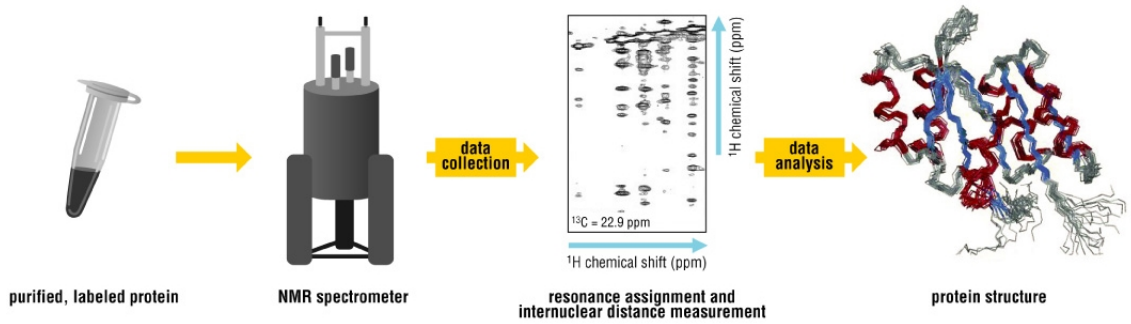


Figure 1.6: Protein Structure determination by NMR

1.3.1 Resonance assignment

In the initial stage of any investigation by NMR spectroscopy, each resonance in the NMR spectrum must be associated with a specific nucleus in the molecule under investigation. Resonance assignments must be sequence specific: each resonance must be assigned to a spin in a particular amino acid residue in the protein sequence. NMR spectroscopy provides three types of information useful for spectral assignments: through-bond interactions (via scalar couplings), through- space interactions (via dipolar couplings), and chemical environment (via isotropic chemical shifts). The strategies employed for resonance assignments depend upon whether only homonuclear ^{1}H NMR spectra are available (unlabelled proteins) or whether ^{13}C and ^{15}N heteronuclear correlation spectra are available (isotopically labelled proteins).

The procedures for obtaining ^{1}H *sequential resonance assignments* are based upon the following critical observation: with few exceptions, correlations resulting from $^{1}H^{1}H$ scalar couplings normally are only observed between ^{1}H nuclei separated by two or three bonds in proteins. Cross- peaks in ^{1}H homonuclear correlation NMR spectra occur between ^{1}H spins within the same amino acid residue or spin system. Cross-peaks do not occur between ^{1}H spins in different residues, because the inter- residue $^{4}J_{H^{N}_{i+1}H^{\alpha}_{i}}$ coupling constant is negligible. Therefore, scalar correlation experiments, such as COSY, MQF-COSY, MQ spectroscopy, and TOCSY, [Cavanagh et al., 1995] are used to identify resonance positions within each amino acid spin system, and the NOESY experiment

[Cavanagh et al., 1995] is used to sequentially connect the amino acid spin systems. Two-dimensional NOESY and TOCSY experiments also can be combined to yield homonuclear 3D experiments, which would reduce the difficulty of sequential assignments.

An alternative strategy, known as the *main chain directed* (MCD) approach, has been developed by Englander and Wand [Feng et al., 1991, Nelson et al., 1991]. In the MCD approach, scalar coupling connectivities are used initially to identify ${}^1H^N - {}^1H^\alpha - {}^1H^\beta$ units only. Assignment of the spin systems by amino acid type is not attempted. Next, the ${}^1H^N - {}^1H^\alpha - {}^1H^\beta$ units are aligned sequentially by systematically searching the NOESY spectrum for patterns of sequential NOEs. Different elements of secondary structure give rise to specific patterns of NOEs [Wüthrich, 1986], and a search is made for these motifs in the following order: helix, anti parallel sheet, parallel sheet, turns, and loops. Once all of the backbone coupling units have been aligned sequentially and categorized by secondary structural element, determination of the amino acid type of several side chains permits the defined elements of secondary structure to be aligned with the primary sequence.

1.3.2 NMR-derived structure restraints

Essentially all parameters that can be measured by NMR spectroscopy are sensitive in some, more-or-less complex, manner to molecular conformation; therefore, quantification of these parameters permits structural analysis by NMR spectroscopy. At present, dipolar cross- relaxation (NOE) rate constants, scalar coupling constants, isotropic chemical shifts, and residual dipoledipole coupling constants (RDCs) are the most commonly utilized parameters for protein structure determination.

NOE distance restraints

The most important NMR-observable parameter used in determining protein structure is the NOE. The dipolar cross-relaxation rate constant is proportional to the inverse sixth power of the distance between two interacting ${}^1H$ spins [Cavanagh et al., 1995]. In the initial rate approximation, NOE cross-peak intensities are proportional to the cross-relaxation rate constants. Thus, if one inter proton distance, r_{ref}, is known (e.g., from covalent ge-

ometry), then another, unknown inter proton distance, r_i, is determined by the relationship (ignoring internal mobility)

$$r_i = r_{ref}\left(\frac{S_{ref}}{S_i}\right)^{\frac{1}{6}} \tag{1.2}$$

in which S^{ref} and S_i are the cross-peak intensities.

Dihedral angle restraints from scalar coupling constants

As was first described by Karplus [Karplus, 1959], the magnitude of a ^3J scalar coupling constant is a function of the dihedral angle formed by the three covalent bonds:

$$^3J = Acos^2\theta + Bcos\theta + C \tag{1.3}$$

The constants A, B, and C depend upon the particular nuclei involved, and θ is the dihedral angle. Historically, dihedral angle restraints for ϕ and $\chi 1$ dihedral angles have been derived only from $^3J_{H^NH^\alpha}$ and $^3J_{H^\alpha H^\beta}$ coupling constants, respectively [Karplus, 1960, Pardi et al., 1984, Wagner et al., 1987]. Recently, numerous experiments have been developed that allow measurement of $^{13}C^{13}C$, $^{13}C^{15}N$, $^1H^{15}N$, and $^1H^{13}C$ three-bond coupling constants in isotopically enriched proteins [Cavanagh et al., 1995, Krishna and Berliner, 1999]

Dihedral angle restraints from isotropic chemical shifts

Isotropic chemical shifts are exquisitely sensitive to local molecular conformation, but this extreme sensitivity also complicates the interpretation of chemical shifts in atomic detail. Fortunately, the dependence of chemical shifts of backbone nuclei, particularly $^1H^\alpha$, ^{13}CO, $^{13}C^\alpha$, and $^{13}C^\beta$, on secondary structure is well-established [Wishart and Case, 2002, Sitkoff, 1998]. Thus, the secondary chemical shift, defined as the observed value of the shift minus the value expected for the same residue in a random coil peptide, exhibits characteristic patterns for regular elements of secondary structure [Schwarzinger et al., 2000]. This correlation forms the basis for the chemical shift index (CSI) method for identifying elements of secondary structure in proteins [Wishart et al., 1992, Wishart and Sykes, 1994]. The TALOS program compares observed chemical shifts to a database of proteins with

$^{1}H^{\alpha}$, ^{13}CO, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and ^{15}N resonance assignments and high-resolution structures to obtain dihedral angle restraints for incorporation into structure calculations [Cornilescu et al., 1999].

Restraints from residual dipolar coupling constants

Recently, a new class of structural restraint has been introduced that is not strictly local in nature and which represents a major advance in NMR structural studies. These restraints are based upon the measurement of residual dipolar couplings (RDCs) between pairs of NMR active nuclei in partially aligned molecules.

The most straightforward method for incorporating RDC data into structure calculation protocols is by direct refinement of the orientation of individual bond vectors against the measured values of the RDCs [Clore, 1998]. In this case, the orientation of each individual bond vector is changed to satisfy dipolar couplings as structures are calculated. Direct refinement has been shown to improve the accuracy and precision of structures when used in conjunction with nearly complete sets of NOE, coupling constant, and chemical shift data [Clore, 1998]. This approach can be used with limited sets of RDCs; for example, many applications use only the backbone NH RDCs. However, direct refinement requires that relatively high-quality initial structures have been determined from NOE, scalar coupling, and other restraints, because many local minima are encountered when refining RDCs [Fischer et al., 1999, Mueller et al., 2000]. Bax and Grishaev discuss the difficulties that can result from refining against too limited a set of RDCs, such as the $^{1}H^{15}N$ RDCs alone [Bax and Grishaev, 2005].

Hydrogen bond restraints from amide protonsolvent Exchange

Slow rates of amide exchange are associated with shielding of amide ^{1}HN atoms from solvent, and most commonly result from hydrogen bonding interactions [Wagner, 1983]. Amide exchange rates are usually measured in one of two ways, depending on the rate of exchange. When the rate is comparable to or faster than the spinlattice relaxation rate ($k_{ex} > 0.1s^{-1}$), the rate constant is most easily determined from a saturation trans-

fer experiment [Forsén and Hoffman, 1963]. For slower rates ($k_{ex} < 0.1s^{-1}$), exchange usually is measured by rapidly transferring the protein from H_2O into D_2O solution, and repeatedly acquiring homonuclear TOCSY [Cavanagh et al., 1995] or $^1H^{15}N$ HSQC [Cavanagh et al., 1995] spectra to observe the decrease in amide proton resonance intensities with time. Observation of a slow amide proton exchange rate implies that the ^1HN atom may be involved in a hydrogen bond, but does not identify the atoms acting as hydrogen bond acceptors (and cannot exclude the possibility that the reduced exchange rate results from steric effects rather than hydrogen bonding). Hydrogen bond restraints have a large impact on the nature and precision of the resulting structures and are usually only enforced in well-defined regions of regular secondary structure, in which only one possible hydrogen bond acceptor is consistent with the NOE data.

1.4 Molecular dynamics

Details of the local backbone geometry can be obtained by an extension of the sequential assignment process; the relative intensities of $d_{NN}, d_{\alpha N}$, and $_{\beta N}$ NOE cross-peaks and the measurement of the backbone $^3J_{H^N H^\alpha}$ are required. The observation of intense d_{NN} NOEs and small $^3J_{H^N H^\alpha}$ coupling constants ($< 6.0Hz$) are indicative of helical or turn sections of polypeptide; observation of intense $d_{\alpha N}$, weak d_{NN}, and $d_{\beta N}$ NOEs and large $^3J_{H^N H^\alpha}$ coupling constants ($> 8.0Hz$) are indicative of extended β-strands of polypeptide [Wüthrich, 1986]. The combination of sequential NOE and $^3J_{H^N H^\alpha}$ coupling constant data with medium-range and a few long-range NOEs is capable of providing details of the regions of regular secondary structure within the protein. The elements of secondary structures can be connected together to give a crude view of the global fold by the identification of a few key long-range NOEs. Thus without recourse to extensive calculations, important structural results (albeit of low absolute resolution) can be obtained in a straightforward manner.

A variety of methods have been developed to calculate atomic resolution protein structures using restraints derived from experimental NMR data [Güntert, 2003, Güntert, 1998, Grishaev and Llinás, 2005]. Importantly, NMR data do not uniquely define the three- di-

mensional structure of a protein or other biological macromolecule, because the restraints are included as ranges of allowed values, the data contain experimental uncertainties, and only a sparse subset of all possible restraints are observable. To increase the efficiency and accuracy of structure calculations, the experimentally derived restraints normally are supplemented by restraints specifically imposed to enforce proper covalent structure of the protein, including bond lengths, bond angles, and other elements of standard covalent geometry (chirality and the planarity of aromatic rings and peptide units). Protocols for structure determination aim to find coordinates for the protein atoms that will satisfy the input restraints in an unbiased fashion while exploring all of the regions of conformational space compatible with these restraints. Because of these considerations, structure calculations are repeated many times to determine an ensemble of (low energy) structures consistent with the input NMR data. Thus, a *good* ensemble of structures minimizes violations of the input restraints and maximizes the RMSD between members of the ensemble [Renugopalakrishnan et al., 1991, Hyberts et al., 1992]

The two most common approaches to generation of structures are distance geometry (DG) and restrained molecular dynamics (rMD). Historically, DG was the first approach utilized for structure determination; at an intermediate stage of development, DG frequently was used to generate initial structures for subsequent refinement by rMD methods. In modern approaches for structure determination, rMD has become the predominant technique. However, other approaches to structure determination continue to be pursued and future developments can be expected [Rieping et al., 2005].

Popular implementations of DG use either the metric matrix algorithm [Crippen and Havel, 1988, Havel, 1991] or the variable target function approach [Braun and Go, 1985, Güntert et al., 1991]. Distance geometry determines ensembles of three-dimensional structures consistent with an incomplete set of distance restraints. The restraints are incomplete because not all distances can be characterized (the NOE is limited to distances less than approximately $5A^0$) and because the distance restraints are not known precisely. The metric matrix algorithms in particular tend to be computationally expensive as the size of the protein increases.

Restrained molecular dynamics algorithms use either Cartesian or torsion-angle coor-

dinate systems [Güntert, 1998]. Torsion-angle rMD has become the preferred method due to advances in computational algorithms. In either approach, molecular dynamics force fields are supplemented by pseudo-energy terms based on the NMR-derived restraints [Clore et al., 1986, Brünger et al., 1986] These potentials drive the structure toward a conformation that will reduce the violation of the restraints during a forced heat-up and cool-down annealing cycle. The most computationally efficient implementations of the rMD method use a simplified force field in which bond length, bond angle, and repulsive van der Waals terms are retained (electrostatic and attractive van der Waals terms are ignored), and are referred to as dynamical simulated annealing (SA) [Nilges et al., 1988]. Due to advances in computational power, structures determined using simplified force fields now frequently are refined using complete force fields and including explicit or implicit solvent models [Xia et al., 2002, Linge et al., 2003b]

1.5 Automation

Considerable efforts have been made to partially or fully automate the process of resonance assignment [Zimmerman, 1995, Hiller et al., 2008, Shimotakahara et al., 1997]
[Linge et al., 2003a, Moseley and Montelione, 1999, Li and Sanctuary, 1997b]
[Li and Sanctuary, 1997a, Koradi et al., 1998, Croft et al., 1997, Bailey-Kellogg et al., 2000].
Extensive efforts also are being made to automate the process of structure determination. Most automated structure calculation programs take as input a (sufficiently complete) list of resonance assignments and one or more lists of cross peak positions and volumes from nD NOESY spectra. The programs then automatically assign the NOESY cross-peaks and calculate the three-dimensional structure of the protein. Current state-of-the-art methods for automated structure determination have been reviewed in literature by many people [Gronwald and Kalbitzer, 2004, Baskaran et al., 2009, Baran et al., 2004, Altieri and Byrd, 2004, Güntert, 2003, Grishaev and Llinás, 2005]. A comparison of conventional structure determination protocols with an optimized pipeline consisting of fast data acquisition [Cavanagh et al., 1995], automated resonance assignments, and automated structure calculation has been reported by Szyperski and co-workers [Liu et al., 2005].

The automation process has two limitations. First is the limitations in experimental procedures namely the noise, artefacts and impurities in the sample. The solvent suppression techniques are not so efficient, that they may produce additional artefact in the spectra. Second limitation comes from the limited statistical information about the structure based chemical shift information. The complete statistical correlation between structure and chemical shift is not yet known.

CHAPTER 2

CHEMICAL SHIFT OPTIMIZATION IN MULTIDIMENSIONAL NMR SPECTRA

2.1 Introduction

Nuclear Magnetic Resonance is an important tool for structure elucidation of biological macro molecules, and quite useful to study the dynamical behaviour of molecules. Structural dynamics can be studied by the careful measurement of chemical shifts of each atom. Measuring the *true* chemical shifts accurately in experimental spectra is not straightforward in NMR because of severe overlap of resonance peaks and the presence of noise and artefacts. Here, a number of optimized peak picking routines were developed [Neidig et al., 1984, Glaser, 1987]. The inverse problem, the projection of known chemical shifts (assignments) to an experimental spectrum is also not trivial because of the same reasons: the peak maximum may be shifted by noise or by superposition with other peaks or artefacts. In addition, the digital resolution provides a general limit of accuracy. However, by far the most important problem is caused by chemical shift variations due to temperature shifts or small changes of the sample composition and the buffer conditions (e. g. pH and ionic strength). Here, the already existing chemical shift table (usually created from a large set of multidimensional spectra) does not correspond exactly to the spectrum under investigation. Another application would be TROSY-spectroscopy where the cross peaks are shifted by $J/2$. Since for structural determination information from whole set of nD-spectra has to be combined, the variation in chemical shifts between the different spectra has to be taken into account. In principle, chemical shift recognition is part of automated procedures for assigning peaks in multidimensional spectra. Sev-

eral automated peak assign procedures are reported in the literature [Catasti et al., 1990, Zimmerman, 1997, Xu et al., 2001, Herrmann et al., 2002, Gronwald et al., 2002] using neural networks and other optimization techniques. But all these methods are aimed at structure elucidation and not giving much importance to optimization of chemical shift of every atom. For the chemical shift optimization of an individual spectrum that is recorded at different conditions they are not useful. In the present chapter we propose two different algorithms to adapt a given chemical shift table optimally. They are compared with each other and their accuracy for different spectral types is assessed. The proposed algorithms can also be used in other fields such metabolomics where an alignment of peaks in spectra of different mixtures improves a multivariate analysis.

2.2 Materials and methods

2.2.1 Simulation of NOESY data sets

Test data sets were created by spectral simulation of HPr (histidine containing phosphocarrier protein) from Staphylococcus aureus, a 88 residue phosphocarrier protein. The ^{1}H and ^{15}N NMR spectra are completely assigned [Maurer et al., 2004] and are deposited in the BioMag data base. The NMR structure is deposited on the PDB data base (PDB ID:1ka5). A 2D NOESY spectrum of HPr was simulated using RELAX [Görler and Kalbitzer, 1997, Gronwald et al., 2000] module in AUREMOL [Gronwald and Kalbitzer, 2004]. Test data set was created using 466 proton chemical shifts of HPr protein with mixing time of $250\,ms$, relaxation delay of $1.75\,s$ and having 2048 data points in both dimensions. The spectrometer frequency was set to 800.2 MHz. The simulation used an overall rotational correlation time of $3.9\,ns$, included internal mobility on the basis of the model-free approach with standard main chain and side chain order parameters, fast methyl rotations and slow ring flip motions. J-coupling and chemical shift anisotropy was not included in the simulation. The detection limit was set to $0.5\,nm$leading to 9035 resonance peaks in the simulated NOESY spectrum.

3D carbon and nitrogen edited NOESY-HSQC spectra were simulated analogously.

The chemical shift table of carbons contained 368 entries that of nitrogen 95 entries. The digital resolutions in δ1, δ2, and δ3 were 0.0987, 2.198, and 0.016 ppm/point for the ^{13}C edited spectra. The proton resonance frequency was 800.2 MHz. For the ^{15}N edited spectra, the digital resolutions in δ1, δ2, and δ3 were 0.0987, 2.055, and 0.011 ppm/point.

Gaussian noise was added to the simulated spectrum with a standard deviation scaled to the mean cross peak intensity of the spectrum $< I >$. 10 % noise would correspond to a standard deviation σ of $0.1 < I >$. The noise is created by randomly picking probability densities $p(z)$ at the normalized intensity z from a Gaussian distribution function with a mean of zero. A second random-number generator was used to decide if z is accepted or not. The random numbers x are projected to the interval $[0, pmax]$. For $x < p(z)$ z is accepted, otherwise rejected.[George and Muller, 1958].

2.2.2 Experimental test spectra

Experimental ^{1}H, ^{15}N-HSQC spectra of HPr from S. carnosus were recorded at different pressures and temperatures as described earlier [Kalbitzer et al., 2000]. The ^{1}H frequency was $750.MHz$. The ^{1}H and ^{15}N spectral width were 13.9474 ppm and 50.012 ppm, respectively. The corresponding data size of the time domain data were 2048 x 256 complex data points. The spectrum recorded at temperature 298 K and pressure 100 MPa was manually reassigned with the data published earlier [Görler et al., 1999] and was used as a reference spectrum to assign the rest of the spectra.

In addition a three-dimensional HNCA spectrum of Saratin [Gronwald et al., 2008] was taken. The spectral widths were 8.76 ppm, 41.11 ppm, and 41.41 ppm, in the ^{1}H, ^{15}N and ^{13}C dimensions, respectively. The corresponding number of data points 1280, 128, and 256, respectively. The proton resonance frequency was 600 MHz.

2.3 Theoretical considerations

2.3.1 Chemical shift optimization

Our aim is to find chemical shift values that optimally explain a given spectrum starting from an initial chemical shift table S_0 for atoms (spins) or group of atoms (e. g. methyl protons)(j) $\{\delta(j)|j = 1,2,...J\}$. The chemical shifts $\delta(j)$ can be degenerated that is more than one atom in the protein may have the same chemical shift $(\delta(j) = \delta(k))$. In principle, an atom can have more than one chemical shift $\delta(j)$ value when it occurs in different states n (e. g. local conformations of the proteins), a fact that can be described by introduction of the corresponding superscript to $\delta^n(j)$. In addition, in experimental spectra where a spin is represented in more than one dimension (as in typical homonuclear spectra) the experimental chemical shifts of the same atom (spin) may be different because (1) errors with referencing did occur or because (2) differences in digital resolution dominate the peak positions. Error (1) can be reduced to ± 1 data point by carefully referencing the spectra, error (2) cannot be avoided but again it should be smaller/equal to one data point. The list S_0 is usually incomplete in protein spectra since often for some atoms of a protein the resonance frequencies cannot be identified. At the end a final chemical shift table Sf is generated that optimally fulfils some optimization criteria often in the form of a target function.

The experimental N-dimensional spectrum contains cross peaks at positions δ_i

$$\delta_i = \begin{pmatrix} \delta_1^i \\ ... \\ \delta_N^i \end{pmatrix} \tag{2.1}$$

where the possible combinations of the components of vector δ_i depend on the actual type experiment and the sample composition. In a classical multidimensional spectrum the allowed frequencies are a subset of all resonance frequencies $\delta(j)$. Assignment of a cross peak at δ_i would mean the assignment of different $\delta(j) \in S_0$ to all components $\delta_k^i(k = 1..N)$ of δ_i .The chemical shift optimization can be then formulated as the search

for a diagonal matrix A with

$$S_f = (1 + A)\, S_0 \tag{2.2}$$

and the initial and the final chemical shift tables S_0 and S_f written as vectors. The simulation of the spectra with S_f should optimally explain the experimental spectrum under consideration. In an ideal, noise-free spectrum all experimental cross peaks at positions δ_i^e should be explained by simulated cross peaks at positions δ_m^s.

2.3.2 Algorithms and general definitions

Starting with a chemical shift table S_0 as input two different procedures were developed here that are combined in AUREMOL-SHIFTOPT. They are tested on different types of spectra for showing their advantages and disadvantages. Both methods include a set of common procedures that are already existing in AUREMOL, namely a peak picking procedure [Neidig et al., 1984], the calculation of peak probabilities to discriminate true resonances from artefacts and noise [Antz et al., 1995, Schulte, 1997], the assignment of peaks from the input shift table, the calculation of the assignment probability from the chemical shift deviation (and optional additional information), and the calculation of a corrected chemical shift table from the probability weighted cross peak coordinates Figure2.1. The last three steps may be iterated until the target function is optimized. In NOESY-type spectra the peak volume is an important additional source of information. Experimental peak volumes are obtained by an iterative segmentation procedure [Geyer, 1995], the simulated peak volumes are calculated on the basis of the full relaxation matrix formalism [Görler and Kalbitzer, 1997, Görler et al., 1999, Ried et al., 2004]. When an experimental structure is not yet available, the relaxation matrix is built from the most likely pair wise distances obtained from an unbiased structural data base. In other cases the spectrum is simulated by a procedure established in AUREMOL that allows to predict any n-dimensional spectrum from a chemical shift list and an internal description of the expected cross peak patterns on a semi quantitative basis.

Since it is the general assumption that a spectrum back calculated with S_0 should be similar to the experimental spectra, the corresponding experimental cross peaks should

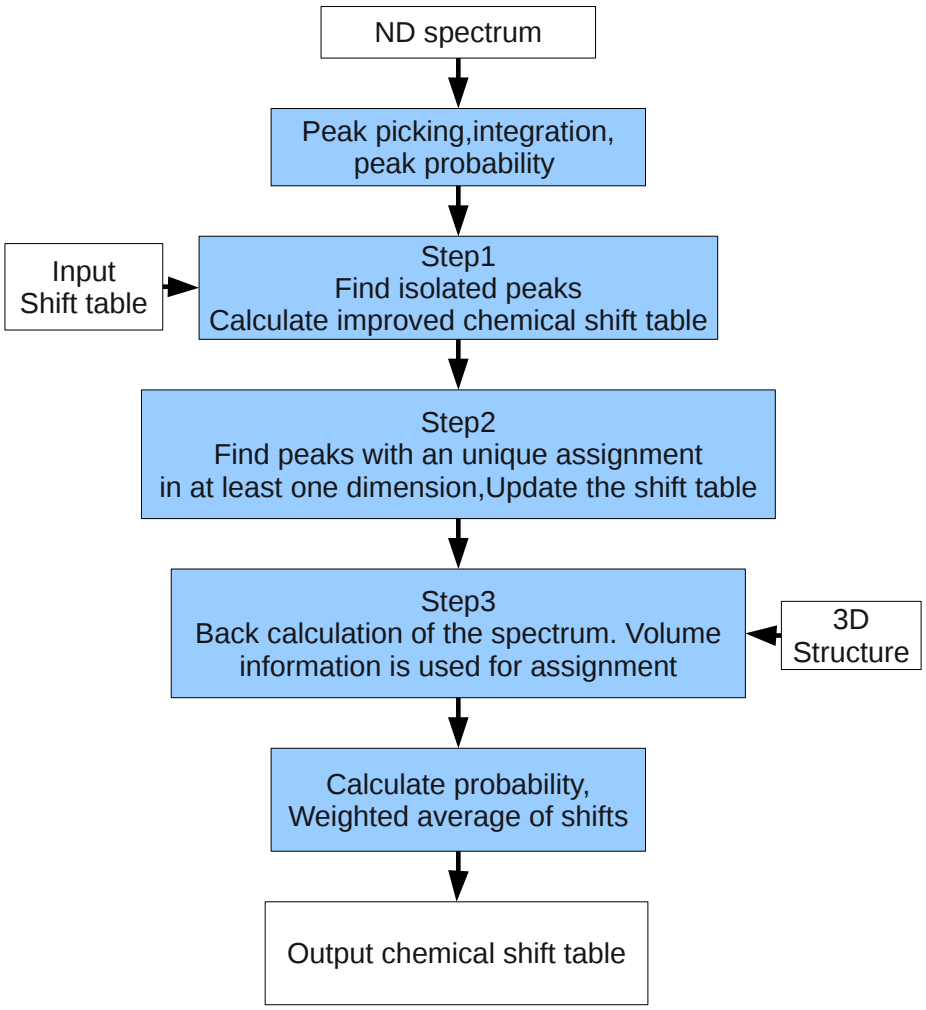


Figure 2.1: The three-step chemical shift optimization of SHIFTOPT1

not be too far away from the simulated cross peaks in terms of a distance metric in the chemical shift space.

In a multidimensional NMR spectrum with the dimension N the cross peak position is defined by the N-dimensional vector δ_i. Since the digital resolution R_k in all dimensions k is usually not identical, the precision of the peak position is also different in the different dimensions. A lower error bound T_k^l is given by

$$T_k^l = \pm \left(R_k + \frac{L_k}{2} \right) \tag{2.3}$$

L_k is the expected line width in dimension k. The upper error bound T_k^u depends on the experimental conditions and defines the initial search range for a peak. The vectors

$$T^l = \begin{pmatrix} T_1^l \\ ... \\ T_N^l \end{pmatrix} \tag{2.4}$$

$$T^u = \begin{pmatrix} T_1^u \\ ... \\ T_N^u \end{pmatrix} \tag{2.5}$$

define the upper and lower search range in all N dimensions.

2.3.3 SHIFTOPT1

This first approach aims mainly on cases where experimental NOESY type spectra plus additional structural information are available for shift optimization. Employing the initial shift list and the structural information NOESY spectra are simulated. However, it is not limited to NOESY-spectra but can be applied to other types of spectra. The simulated spectra predict the frequency combinations δ_m where cross peaks s_m^s are to be expected and the approximate cross peak volume for a given structure. Peak positions and volumes in the simulated spectra can be directly compared with those in the experimental spectra obtained by peak picking and peak integration of the experimental cross peaks s_j^e. Since the experimental spectrum contains true signals as well as noise and artefact peaks the

signals are classified by a Bayesian analysis [Antz et al., 1995, Schulte, 1997] and probabilities P_j are calculated that an experimental cross peak is a true signal. The experimental peaks inside the search areas are sorted according to their probability and all peaks with a low signal probability $P_j < P_t$ are removed from the peak list, the other peaks are included in the analysis but with consideration of P_j. P_t is calculated in such a way that in the search area 10% more peaks are left as expected from the spectrum simulation. Another criterion for the general acceptance of cross peaks is the cross peak volume that should be in the correct range. In the NOESY-spectra cross peaks with intensities that are significantly larger than that of the smallest possible $H-H$ distance of $0.18\,nm$ can be omitted from the assignment procedure. When a reliable structure exists, a more detailed volume comparison can be done on the basis of an assignment hypothesis. Here, the full relaxation matrix formalism provides good estimates of the cross peak intensities to be expected.

The actual procedure used is a three step procedure (Figure 2.1). Step 1 selects those cross peaks s_j^e that are isolated and uniquely assignable to a simulated peak s_m^s inside the error limits of$\delta_m \pm T^u$. The corresponding chemical shift values of S_1 are set to new values. In step 2 those cross peaks are selected where at least in one dimension a unique chemical shift assignment exists. Here, as well as in the step 3 (no unique chemical shifts) volume information is used to solve ambiguities. Simultaneously with the assignment procedure the chemical shift tolerance is reduced for individual shifts to T^l.

The peak assignment and chemical shift refinements are performed in an iterative way. Initially, all experimental peaks are unassigned and are part of list of unassigned experimental peaks (U-list) that consists of the experimental peak positions δ_i^e, the volumesV_i^e and the probability values P_i. In step 1 first the uniquely assignable peaks are identified and written to the A-list. Starting with an arbitrary peak s_i^e from the A-list with chemical shifts δ_i^e the chemical shift table S is slowly refined by updating iteratively their chemical shifts $\delta(j)$

$$\delta(j) = \delta(j) + \xi_i \left(\delta_{i,k}^e - \delta(j) \right) \tag{2.6}$$

with

$$\xi_i = \frac{P_i}{a + P_i} \tag{2.7}$$

and $\delta^e_{i,k}$ the chemical shift coordinate assigned to $\delta(j)$. Using the updated chemical shift table S a new A-list is created that may now contain different peaks and the procedure is repeated. ξ_i controls the influence of an individual experimental peak on the updated chemical shift table S. Especially it is ensured that the influence of artefact and noise peaks is limited by the inclusion of the peak probability values P_i in ξ_i, while the parameter a controls the general influence of an individual experimental peak. Initially the parameter a is set to 10 and is now increased by 1. After 5 cycles the tolerance is reduced to max $(\frac{T^u}{2}, T^l)$ and the procedure is repeated N-times (typically N = 15). The final A-list now contains the uniquely assigned cross peaks. Simultaneously the chemical shift table S has been updated and consists of two subsets S_{nr} and S_r that contain the non-refined and the refined chemical shift values, respectively. In case of NOESY spectra the peaks of the final A-list are used to normalize a back calculated NOESY-spectrum for the next steps.

In step 2 again all cross peaks are scanned iteratively (usually 20 times) with the new chemical shift table S and the corresponding error bounds. Only cross peaks are accepted that fulfill the condition that at least in one dimension an unambiguous assignment is possible. In addition, for NOESY spectra the peak volume V must be inside the allowed range corresponding to the distance range between 0.18 and 5 nm. If more than one assignment is possible, the assignment of a cross peak s^e_i to a simulated peak sms where the deviation of chemical shifts is smallest and the volume is closest to the expectation is taken. For making that decision a $z-score$ Q is defined as

$$Q(S^s_m) = \frac{1}{b}\sqrt{\sum_{k=1}^{n} z^2_k(S^s_m) + z^2(S^s_m)} \tag{2.8}$$

With n the dimension in the N-dimensional spectrum considered. The parameters z_k are defined by

$$z_k = \frac{\delta_k(S^s_m) - \delta_k(S^e_i)}{\sigma_k} \tag{2.9}$$

with $\delta_k(S^s_m)$ and $\delta_k(S^e_i)$ the chemical shifts in dimension j of the experimental and the simulated peaks. The expected standard deviations σ_k are defined by

$$\sigma_k = \frac{T^l_k}{2} \tag{2.10}$$

The parameter z is defined as a function of the normalized cross peak volumes with

$$\sigma = \frac{|V_{max}^{-\frac{1}{6}} - V_{min}^{-\frac{1}{6}}|}{2} \tag{2.11}$$

and V_{max} and V_{min} are the volumes corresponding to the smallest possible distance (0.18 nm) and the maximum detectable distance (0.5 nm), respectively. The constant b in Eq.2.8 is given by

$$b = \sqrt{\frac{V_{min}^{-\frac{1}{6}}}{\sigma} + \sum_{k=1}^{n} \frac{T_k^2}{\sigma_k^2}} \tag{2.12}$$

with T_k the actual error range of the chemical shift. Only solutions with $Q \leq 0.5$ are considered. If such a solution exists the assignment with the lowest Q value is taken and the shift list is updated as in step 2.

In the last step also peaks with ambiguous assignments are taken and the solution with the lowest Q value is selected. After typically 20 cycles the final shift list is calculated from all assigned cross peaks as the average weighted with the peak probabilities P_i, i.e., when the component k of the chemical shift vectors $\delta_i (i = 1, .., N)$ is assigned to a specific atom j then

$$\delta_{final}(j) = \frac{1}{\sum_{i=1}^{N} P_i} \sum_{i=1}^{N} P_i \delta_i(k) \tag{2.13}$$

with P_i the Bayesian peak probability. When an atom is represented in more than one dimension in a spectrum, e.g., in dimension k and p, than the P_i is replaced by P_i^* to

$$P_i^* = \frac{D_R(k) P_i}{D_R(k) + D_R(p)} \tag{2.14}$$

where $D_R(k)$ is the digital resolution of dimension k.

2.3.4 SHIFTOPT2

SHIFTOPT2 represents an alternative way to optimize a chemical shift table and is useful for spectra with a not too large number of cross peaks as HSQC or HNCA spectra. Again a model spectrum is generated from the given input chemical shift table. This model

spectrum is compared with corresponding experimental spectrum. The general preparation of the data corresponds to SHIFTOPT1 where after peak picking and Bayesian analysis the most probable experimental peaks s_i^e in the search areas are selected as defined above. The number of simulated cross peaks is reduced to n_s cross peaks by removing all simulated cross peaks s_m^s where an experimental cross peak s_i^e does not exist with δ_i^e ie inside the error limits of $\delta_m^s \pm T^u$. Two probability matrices Q^e and Q^s are constructed for the experimental and simulated cross peaks with the elements Q_{im}^e and Q_{im}^s. Q_{im}^e represents a measure for the probability of an experimental peak s_i^e at position δ_i^e to be assigned to a simulated peak s_m^s at position δ_m^s. $Q^s im$ represents a measure for the probability of a simulated peak s_m^s at position δ_m^s to be assigned to the experimental peaks s_i^e at position δ_i^e. Q_{im}^e and Q_{jm}^s are given analogously to Eq.2.8 as a function of a generalized variable z^{im}. The components z_k^{im} im of the vector z^{im} are defined by

$$z_k^{im} = \frac{\delta_i^e - \delta_m^s}{P_i \sigma_j} \tag{2.15}$$

The values $\frac{1}{\sigma_j}$ are atom and amino acid specific weighting factors with respect to the assignment of the simulated peak as defined earlier [Schumann et al., 2007]. For atoms or molecules not contained in the data base the averages of the data base are taken with $\sigma(^1H)$ 1.55 ppm, $\sigma(^{15}N)$ 0.236 ppm, and $\sigma(^{13}C)$ 0.447 ppm [Schumann et al., 2007]. The signal probability p_i is obtained from the Bayesian analysis of the experimental spectra. The elements Q_{im}^e are given as

$$Q_{im}^e = \frac{exp(\frac{-(z^{im})^2}{2})}{\sum_r exp(\frac{-(z^{ir})^2}{2})} \tag{2.16}$$

with summation over all peaks s_r^s in the search range. Correspondingly, Q_{im}^s is defined by

$$Q_{im}^s = \frac{exp(\frac{-(z^{im})^2}{2})}{\sum_s exp(\frac{-(z^{sm})^2}{2})} \tag{2.17}$$

with summation over all peaks s_s^e in the search range.

From the two matrices an averaged probability matrix Q is calculated by

$$Q = \frac{Q^e + Q^s}{2} \tag{2.18}$$

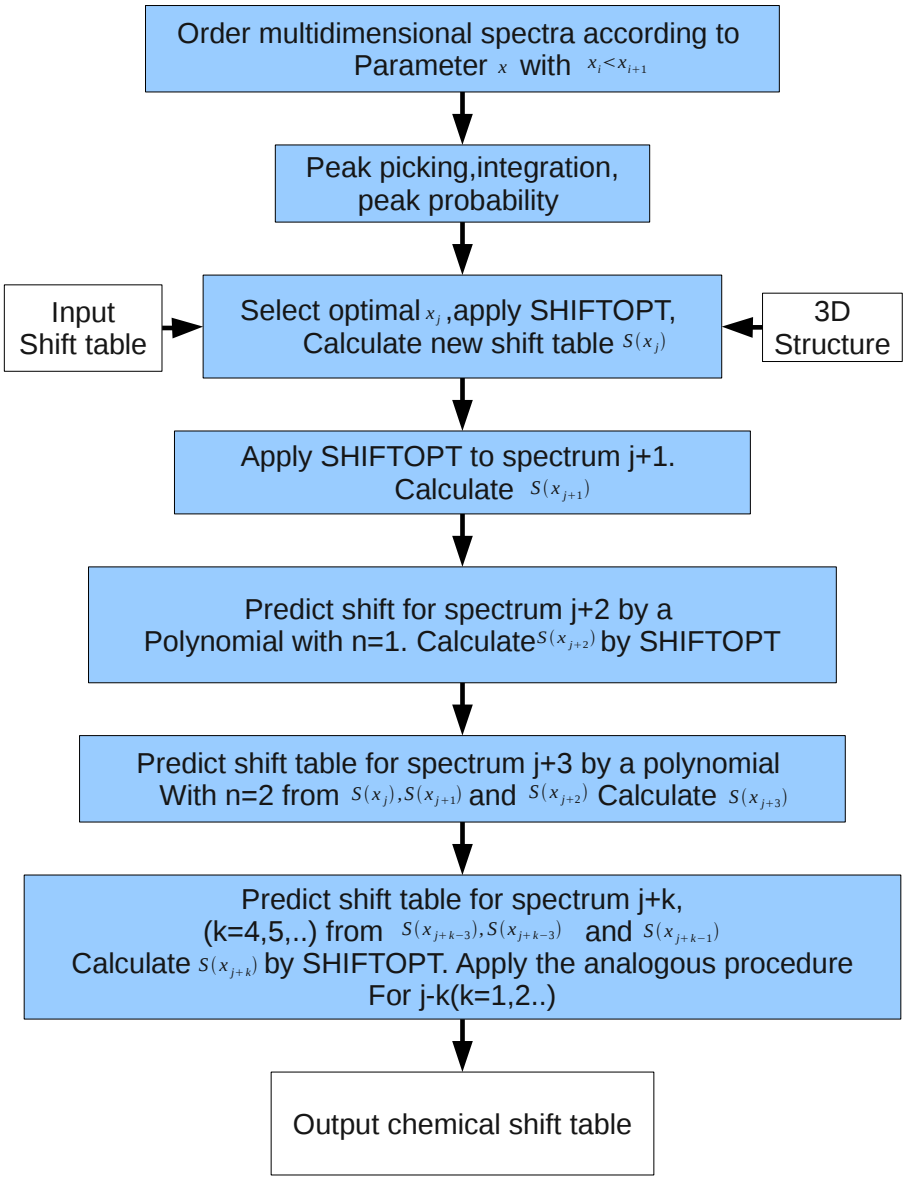


Figure 2.2: Schematical view of the chemical shift optimization in a series of spectra. Note that a polynomial function of the order of 2 is used. In general n + 1 chemical shift lists are required for the prediction

In step 1 pairs of peaks s_i^e and s_m^s with $Q_{im} = 1$ are uniquely assigned and all matrix elements $Q_{ik}(k = 1..n_s; k \neq m)$ and $Q_{lm}(m = 1..n_e; l \neq i)$ are set to 0. For the remaining peaks new probabilities are calculated from the reduced set of peaks. The procedure is repeated until no further element with $Q_{im} = 1$ are found. In step 2 the maximum element Q_{im}^{max} of Q is identified and Q is renormalized by $Q = \frac{Q}{Q_{im}^{max}}$. Step 1 is repeated but only peaks are taken as assigned where $Q_{im} = 1$ and $Q_{ik} < 1(k = 1..n_s; k \neq m)$ and $Q_{im} < 1(m = 1..n_e; l \neq i)$ holds. This procedure is repeated until no new assignments are found.

2.3.5 Adaptation of assignments to a series of spectra

In many cases the chemical shifts in a series of n-dimensional NMR spectra can be represented as continuous functions of a parameter x (such as temperature, pressure, pH or ligand concentration). When more than two spectra are available the approximate positions of cross peaks in spectrum $i+1$ can be predicted from the already assigned spectra by a polynomial of the order n set by the user.

In SHIFTOPT the following strategy is used (Fig.2.2): First the spectra are ordered according to the parameter x in such a way that for all spectra i and $i+1$ $x_i < x_{i+1}$ holds. In a next step that spectrum j is selected where the $x(j)$ is closest to the conditions where the chemical shift table S_0 fits to. The spectrum is assigned with SHIFTOPT and the obtained chemical shift table S_j is used to assign spectrum $j+1$ or $j-1$. If $x_{j+1} - x_j \leq x_j - x_{j-1}$ holds spectrum $k = j+1$ is selected, otherwise spectrum $j-1$. For the following we describe the algorithm for increasing values of k but it can (and generally has to) be applied also for decreasing values of $k, k < j$. The chemical shifts $\delta_m(s_i^s)$ of the simulated peaks in dimension m of spectrum $k = j+2$ are then predicted by a polynomial of order 1 from the optimized $\delta_m(s_i^e)$ from spectrum j and $j+1$. After optimization of the chemical shift list for x_{j+2}, new coefficients a_{ml}, are calculated by

$$\delta_m(s_i^s, x_k) = \sum_{l=0}^{n} a_{ml} x_k^l \tag{2.19}$$

The coefficients *aml* can be calculated by rewriting Eq.2.19 in matrix notation, using the

Vandermonde matrix

$$\begin{bmatrix} 1 & x_0 & x_0^2 & . & . & . & x_0^n \\ 1 & x_1 & x_1^2 & . & . & . & x_1^n \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_n & x_n^2 & . & . & . & x_n^n \end{bmatrix} \begin{bmatrix} a_{m0} \\ a_{m1} \\ . \\ . \\ . \\ a_{mn} \end{bmatrix} = \begin{bmatrix} \delta_m(s_i^s, x_0) \\ \delta_m(s_i^s, x_1) \\ . \\ . \\ . \\ \delta_m(s_i^s, x_n) \end{bmatrix} \tag{2.20}$$

The coefficients a_{ml} can be obtained from the linear Eq.2.20 using standard techniques for solving simultaneous equations [Press et al., 1992].

2.4 Results

2.4.0.1 Stability of the search algorithms

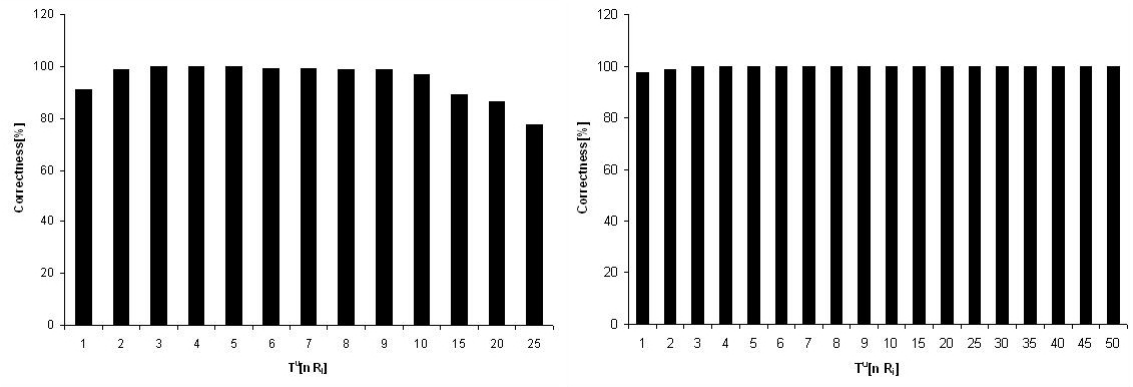


Figure 2.3: Stability of the algorithms as a function of the search range T^u. Application of SHIFTOPT1 to a simulated 2D NOESY-spectrum HPr from S. aureus with a ^1H digital resolution R_i in both dimensions of 0.0062 ppm and b application of SHIFTOPT2 to an experimental $2D-^1H,^{15}N$ HSQC-spectrum from HPr from S. carnosus with a ^1H and a ^{15}N digital resolution R_i of 0.0068 and 0.195 ppm, respectively. The percentage of correct solutions with ($|\delta_i^{opt} - \delta_i^e| \leq R_i$) is plotted as a function of n with n a multiple of the search range $T^u = nR_2.\delta_i^{opt}$ and δ_i are the chemical shift values after optimization and the correct values before optimization, respectively

The size of the search area T^u determines the number of possibilities to solve the optimization problem. A too small size will lead to incorrect solutions for resonances

outside the search interval. A too large size increases the computational time and may also increase the ambiguities and thus the error probability. Therefore, we have systematically increased the size of the search range from a value corresponding to the digital resolutions R_i to very large values (Fig. 2.3). Two different spectra were used, a simulated 2D-NOESY spectrum containing 9, 035 cross peaks and an experimental $^1H,^{15}N-HSQC$ spectrum containing 79 cross peaks. The spectra were subjected to peak picking, integration and the Bayesian peak recognition routine of AUREMOL was applied. When we start with the correct chemical shift table, false chemical shift values are only obtained when the search range T^u is either too small or at very large values of T^u. At very small search ranges the peak maximum may be shifted out of the search range since overlapping of peaks may shift the peak maxima. Very large search ranges increase the ambiguities and thus may lead to errors. This is especially important in the crowded NOESY-type spectra. Here, SHIFTOPT1 gets some incorrect result after the search range is increased to values larger than 0.03 ppm and becomes significantly less efficient at values larger than 0.06 ppm (48 Hz), whereas for the much less crowded HSQC spectrum the size of the search range virtually does not play any role.

However, when an ideal spectrum without noise and artefacts and infinitely small line widths was created by calculating a peak list directly from the chemical shifts both methods are very stable for all search ranges tested.

2.4.1 Performance in the absence of noise

In a next test the chemical shift table was perturbed by adding or subtracting a value $\Delta\delta_i$ to the chemical shifts δ_i. The values $\Delta\delta_i$ were randomly selected from a Gaussian distribution with a standard deviation $\sigma$and a mean of δi. For nuclei X others than ^1H the standard deviation was modified by multiplying $\sigma(H)$ with $\frac{\gamma_H}{\gamma_X}$ Only values less than 3σ were taken. In this way, a chemical shift table S_0 was produced that does not fit optimally to the experimentally observed chemical shifts δ_i^e. For small values of σ and thus for small chemical shift variations between the simulated shifts δ_m^s and the experimental shifts δ_i^e excellent results are obtained with SHIFTOPT1 for the 2D-NOESY (Fig. 4a) spectrum as well as for the 3D ^{15}N edited NOESY spectrum (Fig. 4c and d). Up to a σ of 0.01

ppm all chemical shifts values are correct after the application of SHIFTOPT1 in the 2D-NOESY spectrum, up to σ of 0.02 ppm all values are either improved or not modified. Only after σ has increased to 0.03 ppm, a few chemical shift values (2 out of 466) are worse than before. A similar picture is obtained for the 3D-NOESY spectrum. Up to a σ of 0.01 and 0.1 ppm for ^{1}H and ^{15}N all chemical shifts values are correct, up to a σ of 0.02 and 0.3 ppm all ^{1}H and ^{15}N chemical shifts values are either improved or unchanged. Only after σ has increased to 0.03 and 0.4 ppm, a few chemical shift values are slightly worse than before. In its last step SHIFTOPT1 relies strongly on the calculation of cross peak volumes, whereas SHIFTOPT2 does not use this information. If this information is available and the cross peak volumes vary much as it is typical for NOESY-type spectra the performance of SHIFTOPT2 should be inferior to that of SHIFTOPT1. Indeed, at values of the standard deviation where SHIFTOPT1 works perfectly, SHIFTOPT2 already makes some errors (Fig. 2.4b).

A different type of data sets are represented by 2D $^{1}H,^{15}N-HSQC$ spectra or 3D HNCA spectra that contain only a small number of cross peaks with a relatively small dynamic range of cross peak intensities. Here, the fast, direct algorithm of SHIFTOPT2 should perform well. Figure 2.5c and d show that this is indeed true. Up to a σ of 0.01 ppm all chemical shifts values are correct after the application of SHIFTOPT2 in the 2D-HSQC spectrum, up to a σ of 0.02 ppm all ^{1}H chemical shift values are either improved or not modified. Only after σ has increased to 0.05 ppm, a few chemical shift values are worse than before. In the ^{15}N domain standard deviations of up to 0.4 ppm (that is up to a maximum error of 3σ = 1.2 ppm) are accepted without error. A similar picture is obtained for the 3D-HNCA spectrum (Fig. 2.6). Up to a σ of 0.01, 0.1, and 0.1 ppm all ^{1}H, ^{15}N and ^{13}C chemical shifts values are correct after the application of SHIFTOPT2 in the 3D-HNCA spectrum, up to a σ of 0.03, 0.3, and 0.3 ppm all ^{1}H, ^{15}N and ^{13}C all values are either improved or not modified. Only after σ has increased to 0.04, 0.4, and 0.4 ppm for the ^{1}H, ^{15}N and ^{13}C, respectively, a few chemical shift values are slightly worse than before. SHIFTOPT1 (Fig. 5a, b) does not perform as well as SHIFTOPT2 in

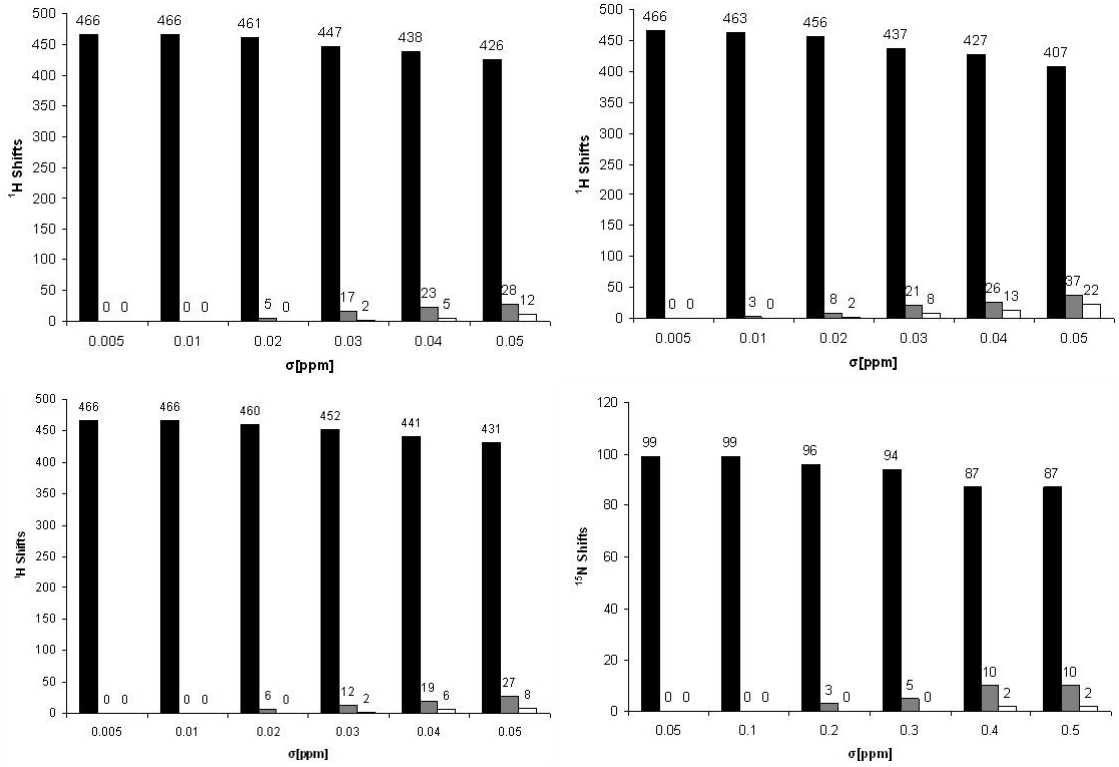


Figure 2.4: Reliability of SHIFTOPT1 and SHIFTOPT2 for NOESY-type spectra in the absence of noise. The number of completely correctly predicted chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq R_i$) (black bars), of improved or unchanged chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq |\delta_m^s - \delta_i^e|$) (grey bars), and inadequately optimized chemical shifts ($|\delta_i^{opt} - \delta_i^e| > |\delta_m^s - \delta_i^e|$) (white bars) are plotted as a function of σ in the dimension k under consideration. δ_i^{opt} is the chemical shift after optimization. Note that for nuclei X others than ^1H the standard deviation was modified by multiplying $\sigma(H)$ with $\frac{\gamma_H}{\gamma_X}$. *a* Simulated 2D NOESY-spectrum with a ^1H digital resolution of 0.0062 ppm, application of SHIFTOPT1. The spectrum contains 9, 035 cross peaks from the protein. *b* as (a) but after application of SHIFTOPT2. c, d Simulated 3D ^{15}N edited NOESY spectrum with a ^1H digital resolution of 0.005 and of 0.098 ppm in the direct and indirect dimension, a ^{15}N digital resolution of 0.764 ppm. Only data for SHIFTOPT1 are shown

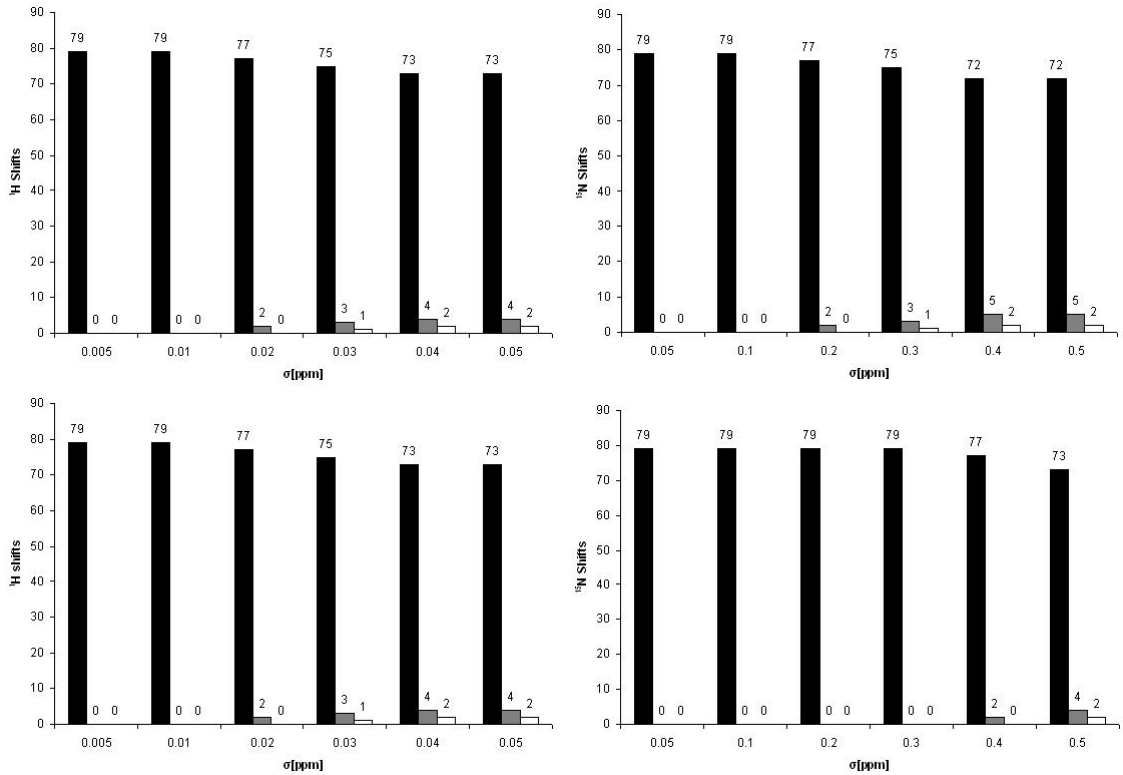


Figure 2.5: Reliability of SHIFTOPT1 and SHIFTOPT2 for HSQC-type spectra in the absence of noise. The number of completely correctly predicted chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq R_i$) (black bars), of improved or unchanged chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq |\delta_m^s - \delta_i^e|$) (grey bars), and inadequately optimized chemical shifts ($|\delta_i^{opt} - \delta_i^e| > |\delta_m^s - \delta_i^e|$) (white bars) are plotted as a function σ in the dimension k under consideration. δ_i^{opt} is the chemical shift after optimization. Note that or nuclei X others than $1H$ the standard deviation was modified by multiplying $\sigma(H)$ with $\frac{\gamma_H}{\gamma_X}$. *a*, *b* Application of SHIFTOPT1 to a 2D $^1H,^{15}NHSQC$ spectrum with a ^1H digital resolution of 0.0068 ppm and a ^{15}N digital resolution of 0.19 ppm, where all noise peaks were removed after peak picking. *c*, *d* As (a,b) but using SHIFTOPT2 on a 2D $^1H,^{15}N-HSQC$-spectrum

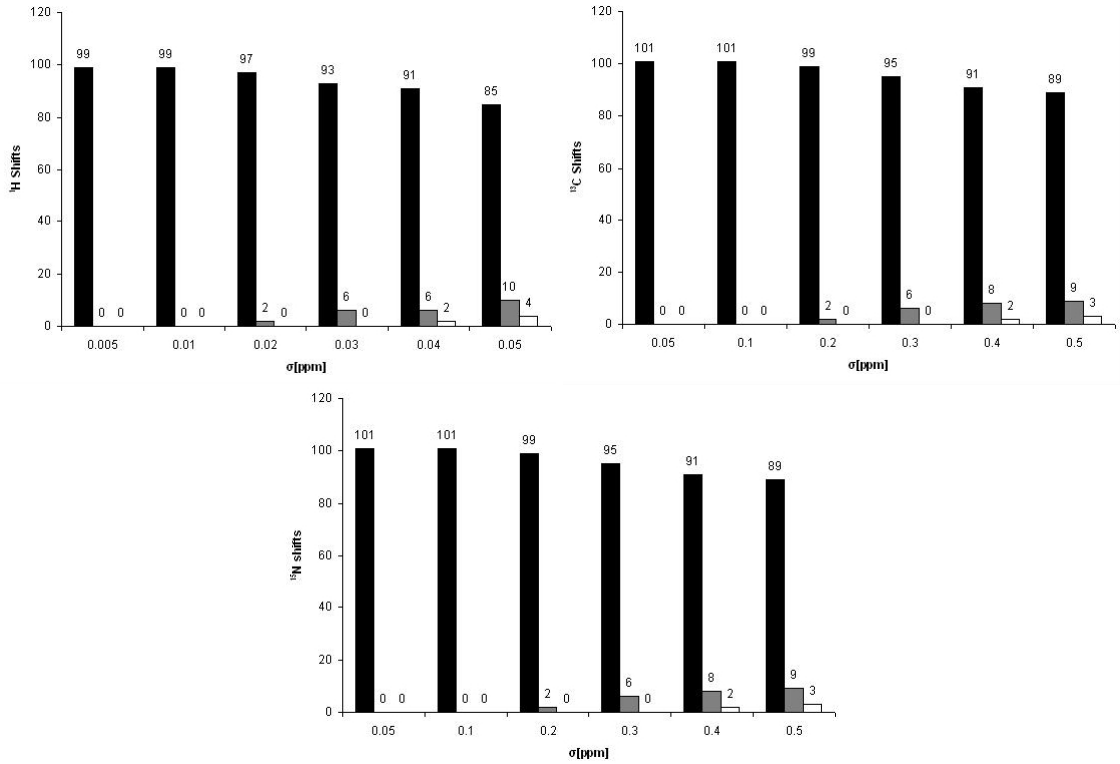


Figure 2.6: Reliability of SHIFTOPT2 for a 3D HNCA spectrum in the absence of noise. The number of completely correctly predicted chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq R_i$) (black bars), of improved or unchanged chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq |\delta_m^s - \delta_i^e|$) (grey bars), and inadequately optimized chemical shifts ($|\delta_i^{opt} - \delta_i^e| > |\delta_m^s - \delta_i^e|$) (white bars) are plotted as a function σ in the dimension k under consideration. δ_i^{opt} is the chemical shift after optimization. Note that or nuclei X others than $1H$ the standard deviation was modified by multiplying $\sigma(H)$ with $\frac{\gamma_H}{\gamma_X}$. 3D HNCA with a ^1H digital resolution of 0.0068 ppm (a), ^{13}C digital resolution of 0.1617 ppm (b), and c a ^{15}N digital resolution of 0.321 ppm

a HSQC spectrum (Fig. 2.5c, d) giving wrong results for ^1H and ^{15}N shift deviations of $\sigma > 0.02ppm$ and $> 0.2ppm$, respectively.

2.4.2 Performance in the presence of noise

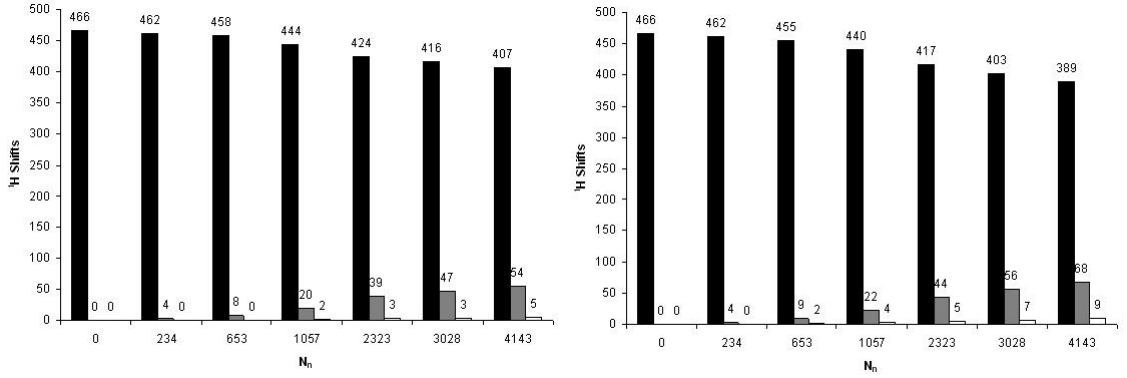


Figure 2.7: Reliability of the shift optimization procedure SHIFTOPT1 as a function of noise level. The number of completely correctly predicted chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq R_i$) (black bars), of improved or unchanged chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq |\delta_m^s - \delta_i^e|$) (grey bars), and inadequately optimized chemical shifts ($|\delta_i^{opt} - \delta_i^e| > |\delta_m^s - \delta_i^e|$) (white bars) are plotted as a function of the number of noise peaks in the dimension k under consideration. The noise level was increased gradually, so that at the peak picking threshold N additional noise peaks were identified. δ_i^{opt} is the chemical shift after optimization. *a* Simulated 800 MHz 2D NOESY-spectrum with a ^1H digital resolution of 0.0062 ppm, application of SHIFTOPT1. The spectrum contains 9, 035 valid protein cross peaks. Variations of chemical shifts with a standard deviation $\sigma = 0.01ppm$. *b* Same as (a) but with a σ of 0.02 ppm

Since noise and artefact peaks can lead to false assignments, Gaussian noise was added to the simulated 2D NOESY spectrum before peak picking. Additional cross peaks at random positions were added in the case of the experimental 2D HSQC spectrum. Since the performance of SHIFTOPT1 was superior for NOESY-type spectra to SHIFTOPT2 but inferior for HSQC-type spectra, SHIFTOPT1 was only tested for NOESY-type spectra (Fig. 2.7) and SHIFTOPT2 for HSQC-type spectra (Fig. 2.8).

Two different cases of practical importance were studied, a relatively small maximum chemical shift variation of 0.01 ppm (8 Hz at 800 MHz) and 0.02 ppm (18 Hz at 800 MHz).

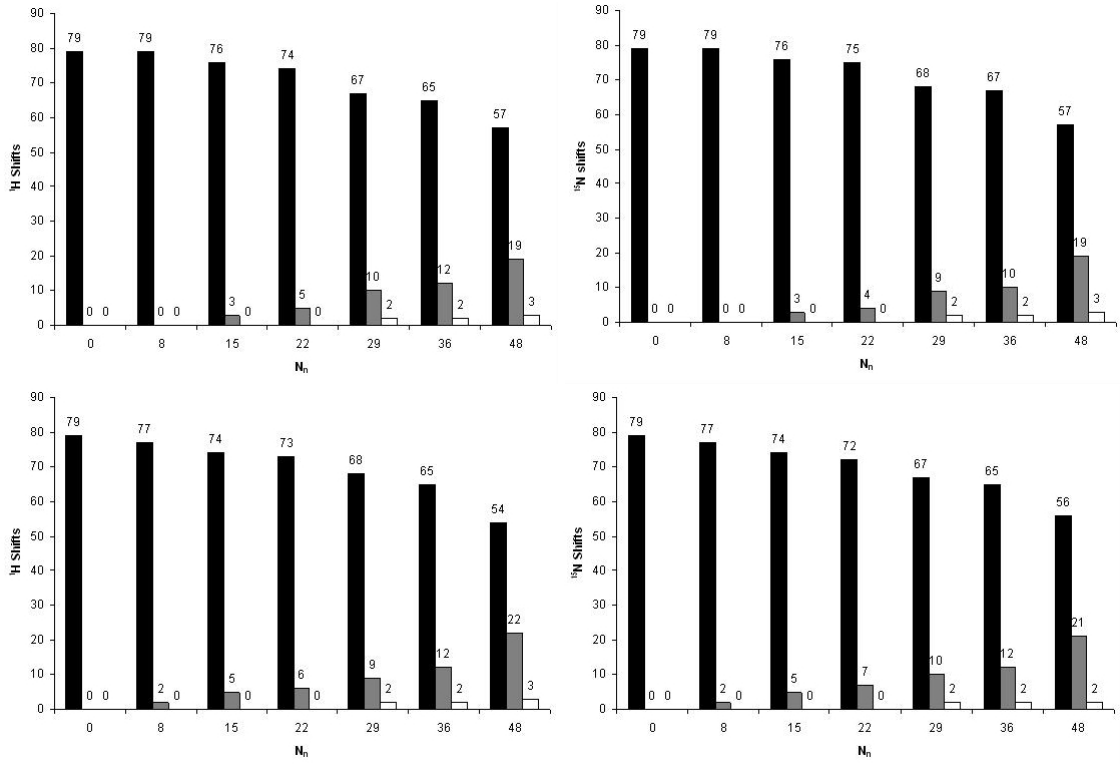


Figure 2.8: Reliability of the shift optimization procedure SHIFTOPT2 as a function of noise level. The number of completely correctly predicted chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq R_i$) (black bars), of improved or unchanged chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq |\delta_m^s - \delta_i^e|$) (grey bars), and inadequately optimized chemical shifts ($|\delta_i^{opt} - \delta_i^e| > |\delta_m^s - \delta_i^e|$) (white bars) are plotted as a function of the number of N. The spectrum contains 79 valid protein cross peaks. δ_i^{opt} is the chemical shift after optimization. Note that or nuclei X others than $1H$ the standard deviation was modified by multiplying $\sigma(H)$ with $\frac{\gamma_H}{\gamma_X}$. (a, b) Experimental 2D ^1H,$^{15}NHSQC$ spectrum, ^1H digital resolution of 0.0068 ppm and ^{15}N digital resolution of 0.019 ppm. Variations of chemical shifts with $\sigma = 0.01ppm$ in the direct dimension and 0.1 ppm in the indirect dimension. c, d Same as (a, b) but with σ-values of 0.02 ppm and of 0.2 ppm

The number of noise peaks wrongly recognized as signals were increased by increasing the level of the standard deviation of the Gaussian noise continuously. The spectrum contained 9, 035 true cross peaks. When 653 additional peaks where wrongly recognized all chemical shifts where improved or at least not changed. When 4, 143 noise peaks were recognized 407 of the 466 chemical shifts where corrected perfectly, 54 improved or unchanged, and only 5 (1.1%) corrected in the wrong way. At a σ of 0.02 ppm still most of the peaks were improved (Fig. 7b).

The application of SHIFTOPT2 to an experimental $^{1}H,^{15}NHSQC$ gives similar results: Up to 22 additional cross peaks are tolerated when 79 true signals are present and the chemical shift list contains errors up to 0.03 (^{1}H) and 0.3 ppm (^{15}N). When 48 additional cross peaks are added, the majority of the chemical shifts is still improved, only 3 chemical shifts are modified in the wrong direction (Fig. 2.8a, b). At the higher maximum chemical shift deviation of 0.06 and 0.6 ppm respectively (Fig. 2.8c, d) a similar picture is obtained, the number of wrongly corrected chemical shifts is still unchanged.

2.4.3 Automated chemical shift assignment in a set of pressure dependent HSQC spectra

Figure 2.9a shows a part of a $^{1}H,^{15}NHSQC$ spectrum of a HPr from S. carnosus measured at different pressures. The initial ^{1}H and ^{15}N chemical shift values at ambient pressure were taken from [Görler et al., 1999]. In a first step SHIFTOPT2 was applied to a data set recorded at 3 MPa [Kalbitzer et al., 2000] and could assign all shift values correctly (Fig. 2.9b). This optimized chemical shift table was then applied to a second data set recorded at 50 MPa where again all chemical shifts were correctly found. In a next step, the chemical shifts expected at 100 MPa were predicted by a linear extrapolation and than optimized by applying SHIFOPT2. The chemical shifts were used for a second order prediction and the procedure was repeated as before. The chemical shifts of the 79 amide groups could be correctly obtained for all pressures. As a comparison, all chemical shifts were obtained by applying SHIFTOPT2 without prior prediction of the chemical shift development. Here, errors occurred for higher pressures, although most of the chemical shifts obtained were

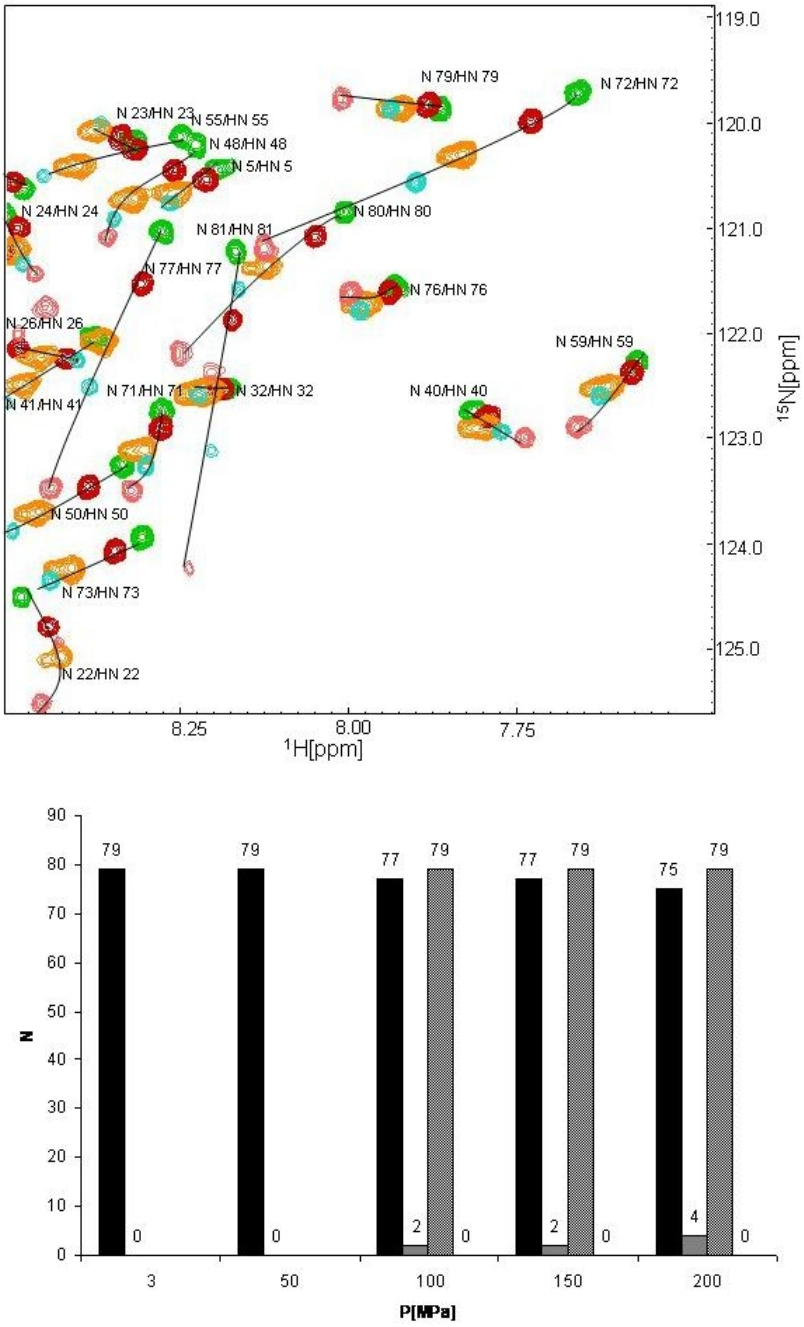


Figure 2.9: Automated chemical shift recognition in a set of pressure dependent HSQC-spectra. a A set of $^1H,^{15}N$ NMR spectra of ^{15}N enriched HPr from S. carnosus was recorded at 298 K and various pressures. (green) 3 MPa, (red) 50 MPa, (yellow) 100 MPa, (blue) 150 MPa, (pink) 200 MPa. Only part of the spectrum is shown. Solid lines connect residues automatically assigned using a polynomial of the order of 2. *b* The number of completely correctly optimized chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq R_i$) (black bars), of improved or unchanged chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq |\delta_m^s - \delta_i^e|$) and (grey bars) are plotted as a function of the pressure. Using the predicted shifts from the chemical shift polynomial as input shift, the assignment getting better. The number of completely correctly optimized chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq R_i$) (black doted bars), improved or unchanged chemical shifts ($|\delta_i^{opt} - \delta_i^e| \leq |\delta_m^s - \delta_i^e|$) (grey doted bars) using the polynomial is plotted in the figure. The spectra contain 79 valid protein cross peaks

still correct (Fig. 2.9).

2.5 Discussion

Even when using the same sample chemical shifts of cross peaks vary slightly from spectrum to spectrum due to small variations of temperature (caused e.g., by broadband decoupling) and differences in digital resolution. In practice, chemical shift lists as published in the BMRB data base are composed from different data sets measured under various experimental conditions (e.g., data recorded in D_2O and H_2O), and thus do usually not fit exactly to a given experimental spectrum. Here, a chemical shift optimization as implemented in SHIFTOPT1 and SHIFTOPT2 can help.

2.5.1 Limits of accuracy

Although a completely correct result of the chemical shift optimization is the ultimate goal, for most applications, an improvement of the chemical shift lists is still a satisfactory result. There are several factors that lead with high probability to insufficient results: (1) Incomplete spectra, where some chemical shifts are not represented at all do not contain the required information and thus cannot be used for calculating optimized shifts. When working on peak lists as SHIFTOPT1 and SHIFTOPT2 do, artefact peaks with sufficient intensity may be wrongly interpreted as true signals and may be used for the chemical shift calculation. This is especially dangerous in low redundancy spectra such as HSQC-spectra where they may be wrongly assigned because of the used distance metric. Here, a limited search range is an important mean to avoid a misinterpretation, since an artefact inside in the search range cannot be recognized as such. However, the application of a powerful artefact recognition routine prior to the application of SHIFTOPT reduces the likelihood of such a wrong interpretation and the starting value is not modified. The Bayesian routine used in AUREMOL assigns signal probabilities to all peaks. It has been proved powerful to calculate the number of cross peaks K expected in the search areas, and to retain only the a*K peaks (a = 1.1) with the highest probabilities in the peak list. In our experience about 10% more peaks should be retained, that is a is set to 1.1. (2)

Chemical shift degeneracy as it often occurs in 2D-HSQC spectra with the amide proton and nitrogen resonances identical cannot be handled by the actual routines, since there is no information available, to decide if the cross peaks are superposed or one cross peak is missing or shifted significantly. However, in NOESY-type spectra usually the redundancy is very high and normally at least one cross peak with different chemical shift combinations is available and can be used. (3) Very strong shifts in crowded spectra may lead to wrong decisions since the used metric favours small shift changes.

2.5.2 Performance of the routines

In the absence of significant noise peaks in NOESY type spectra SHIFTOPT1 works perfectly up to a sigma of 0.01 ppm (Fig. 2.4a) that corresponds to maximum chemical shift changes of $\pm 0.03 ppm$. For higher chemical shift changes a few shifts are only partly corrected. At a maximum shift change of 0.09 ppm 2 (out of 462) shift values are not improved anymore, but are wrongly corrected by a shift change in the wrong direction. This is usually a pair wise ambiguity where the used metric favours the wrong assignment for two resonances with very close chemical shifts. In the 3D-NOESY-HSQC spectrum comparable results are obtained, up to maximum shifts of the order of the line width very good results are obtained. When larger shift variations are allowed, the used distance metric leads in a few cases to wrong decisions. In practice, it means that the subsequent steps still have to allow an appropriate error of the chemical shifts.

When more than one spectrum exists with continuous shift changes, the proposed prediction procedure leads to accurate results.

CHAPTER 3

CHEMICAL SHIFT PREDICTION

3.1 Introduction

Since the advent of protein structure determination it has been a long time debate whether X-ray crystallography is clearly superior to NMR spectroscopy, because X-ray structures are very well defined when compared to NMR structures. This fact is true when the precision of the coordinates are taken in to consideration. If the available crystals diffract sufficiently well, then one could get a structure with high precision. It is often not realized that the two methods cannot calculate directly the three-dimensional structures of proteins from the experimental data, but use iterative search algorithms to find a solution that is consistent with the data. The main differences is when compared to NMR spectroscopy that in general the number and precision of the structural restraints are superior in X-ray crystallography. Apart from that the back calculation of NMR spectra from structural models is not as straight forward as the back calculation of diffraction patterns.

However, for proteins in solution the question may be ill-posed since the minimum energy state in the crystal lattice (or in better words, the weighted average of the structural ensemble in the crystal) may not correspond well enough to the average of the structural ensemble in solution. The obvious reason is they both are in different physiological conditions. There are a number of examples for that in literature, a typical example is HPr from S.faecalis where we could show that the dominant active centre in solution [Hahmann et al., 1998] clearly differs from that of crystal structure [Jia et al., 1994]. In many of these cases, probably more than one conformational state (defined by different local energy minima) exists in solution but only one of them is selected by the crystallization conditions. In addition, software packages used in crystallography tend to suppress

alternative conformations even when they are present in the single crystals. A prominent example is the Ras protein complexed with Mg2+.GppNHp that exists in solution in two conformational states [Geyer et al., 1996, Spoerner et al., 2001]. The published crystal structure [Pai et al., 1990] shows only a single, well-resolved structure but solid-state NMR on the same crystals proves that the two conformational states coexist also in the single crystals [Stumber et al., 2002, Iuga et al., 2004].

Even when only one global energy minimum exists that is identical in crystal and solution, the thermally populated states create a conformational ensemble that in general is different for solid state and solution, since the details of the local energy surface will most probably be different. A full structural description of a protein would require the knowledge of the whole ensemble of structures not only the lowest energy structure because the properties of the protein may depend on a subset of structures which are similar but not identical to the lowest energy structure.

In crystallography, the B-factor is used to describe the thermally induced conformers as the atoms move from their average positions. In solution, classically relaxation time measurements give information about atomic motions in the local conformational space. Chemical shifts could also provide information on the local conformational equilibrium, since they are strongly structure dependent. Close to a single energy minimum regime they are usually population averaged, due to the the exchange between neighbouring states should be fast on the NMR time scale.

Chemical shifts are the ”mileposts” of NMR spectroscopy. Not only are they important as spectral markers, but their dependency on multiple electronic and geometric factors means that chemical shifts can potentially provide a rich source of structural information. However, these multiple dependencies make both the interpretation and accurate prediction of chemical shifts exceedingly difficult, particularly for large molecules such as proteins. Fortunately, over the past decade, significant progress in chemical shift prediction has been made, both through computational advances [Williamson and Asakura, 1997, Case, 2000, Case, 1998, Wishart and Case, 2002] and through the rapid expansion of biomolecular chemical shift databases [Seavey et al., 1991, Zhang et al., 2003, Ulrich et al., 2007]

The main problem in chemical shift prediction is that protein chemical shifts cannot

be predicted with high accuracy by using a single structure. In spite of the many groups that have worked on this problem over the years. The first attempts to calculate protein chemical shifts started already in 1977 [Perkins et al., 1977]. In the mean time, a number of programs are available that are able to predict chemical shifts from a structural data base or calculate chemical shifts from a physical model. Popular examples are SHIFTS [Xu and Case, 2001] and SHIFTX [Neal et al., 2003]. SHIFTS is a program for predicting ^{1}H, ^{15}N, $^{13}C^{\alpha}{}^{13}C^{\beta}$, and $^{13}C'$ chemical shifts from protein structures. SHIFTX can be used to predict all backbone and some of side chain ^{1}H, ^{13}C and ^{15}N protein chemical shifts using only its PDB file as input. SHIFTX uses a unique semi-empirical approach to calculate protein chemical shifts.

Currently there are three main approaches for calculating protein chemical shifts from atomic coordinates: (1) Quantum mechanical, (2) classical, and (3) empirical. Quantum mechanical (QM) approaches employing density functional theory (DFT) have been used to very accurately calculate ^{1}H,^{13}C and ^{15}N shifts for selected classes of residues in proteins [de Dios et al., 1993, Le et al., 1995, Xu and Case, 2001]. Classical approaches, which employ simplified or empirical equations derived from classical physics and experimental data, have been used to accurately calculate ^{1}H shifts for quite some time [Wagner et al., 1983, Dalgarno et al., 1983, Osapay and Case, 1991, Ösapay and Case, 1994, Wishart, 1991, Herranz et al., 1992, Williamson et al., 1992]. Empirical approaches, which rely on chemical shift hypersurfaces calculated from databases of observed chemical shifts, are capable of rapid, but only modestly accurate calculation of ^{1}H,^{13}C and ^{15}N shifts [Spera and Bax, 1991, Le and Oldfield, 1994, Beger and Bolton, 1997, Wishart and Nip, 1998, Iwadate et al., 1999]. These hypersurfaces relate chemical shifts to various empirical parameters (backbone angles, nearest neighbors, sidechain angles, secondary structure, etc.). Pre-calculated chemical shift hyper-surfaces are also used in QM approaches to greatly accelerate the speed of their calculations [Xu and Case, 2001, Xu and Case, 2002, Le et al., 1995].

3.1.1 Prediction programs

There are many prediction programs available in the internet for chemical shift prediction. Some of them predict the chemical shifts based on sequence, but they are not so accurate. SHIFTS and SHIFTX are well known structure based prediction programs. They get the coordinates of atoms(PDB) as am input and will produce the complete backbone and some of side chain chemical shifts list as an output.

3.1.1.1 SHIFTS

SHIFTS [Osapay and Case, 1991, Ösapay and Case, 1994, Sitkoff and Case, 1997, Sitkoff, 1998, Xu and Case, 2002] is a program for predicting ^{1}H, ^{15}N, ^{13}Ca, ^{13}Cb, and ^{13}C chemical shifts from protein structures. It was developed based on an additive model of chemical shift contributions, corresponding to various conformational effects found in a database of density functional theory (DFT) calculations on more than 2000 peptides. Some empirical extensions were used for covering additional conformation regions and residue types. When experimental shifts are available, an optional refinement process for side-chain orientation can also be carried out, which may help identify problems in either the structure or the shift assignments themselves.

3.1.1.2 SHIFTX

SHIFTX[Neal et al., 2003] is a computer program, which rapidly and accurately calculates the diamagnetic ^{1}H, ^{13}C and ^{15}N chemical shifts of both backbone and side chain atoms in proteins. The program uses a hybrid predictive approach that employs pre-calculated, empirically derived chemical shift hyper surfaces in combination with classical or semi-classical equations (for ring current, electric field, hydrogen bond and solvent effects) to calculate ^{1}H, ^{13}C and ^{15}N chemical shifts from atomic coordinates. The chemical shift hyper surfaces capture dihedral angle, side chain orientation, secondary structure and nearest neighbor effects that cannot easily be translated to analytical formula or predicted via classical means. The chemical shift hyper surfaces were generated using a database of IUPAC-referenced protein chemical shifts RefDB [Zhang et al., 2003], and a correspond-

ing set of high resolution ($< 2.1A^o$) X-ray structures. Data mining techniques were used to extract the largest pairwise contributors (from a list of $\sim$ 20 derived geometric, sequential and structural parameters) to generate the necessary hyper surfaces. SHIFTX is rapid ($<$ 1 CPU second for a complete shift calculation of 100 residues) and accurate.

Overall, the program was able to attain a correlation coefficient (r) between observed and calculated shifts of 0.911($^1H_\alpha$), 0.980($^{13}C_\alpha$), 0.996($^{13}C_\beta$), 0.863(^{13}CO), 0.909(^{15}N), 0.741(^1HN), and 0.907 (side chain ^1H) with RMS errors of 0.23, 0.98, 1.10, 1.16, 2.43, 0.49, and 0.30 ppm, respectively on test data sets. It is further shown that the agreement between observed and SHIFTX calculated chemical shifts can be an extremely sensitive measure of the quality of protein structures. They argue that if NMR-derived structures could be refined using heteronuclear chemical shifts calculated by SHIFTX, their precision could approach that of the highest resolution X-ray structures.

3.2 Materials and methods

3.2.1 NMR spectroscopy and structures

The sequential assignments of the NMR signals of the set of proteins (Table 3.1) were taken from the BMRB data base, the corresponding NMR structures from the PDB-data base. Sequential assignments for wild type HPr(wt) from S.aureus were taken from [Maurer et al., 2004], for the mutant HPr(H15A) from Munte et al. (manuscript in preparation).

3.2.2 Molecular dynamics calculations

Structure calculations were performed using the molecular dynamics program CNS v.1.2. (Crystallography and NMR System for crystallographic and NMR structure determination) [Brunger, 1992, Brunger, 2007] employing the restraints (Table 3.4) in a simulated annealing protocol using extended-strand as starting structures. High-temperature torsional angle dynamics were run at 50,000 K for 3000 steps with a time step of 5 fs. The high number of restraints required a threefold reduction of the time step for the integration

of the equation of motion to 5 fs and a reduction of the ceiling value to 15 for around 30 restraints per residue for the NOE-energies (the default value is 30 for 16 restraints per residue). In the first cooling stage, torsional angle dynamics were used for 3000 steps with a starting temperature of 50,000 K and a time step $5fs$. The second cooling stage was performed with 3000 steps of Cartesian dynamics with a time step of $5fs$ and a starting temperature of 3000 K. In the final stage, 2000 steps of energy minimization were performed. The structures were accepted based on the NOE violations. Those structures having more than 5% NOE violations are rejected during simulated annealing process. Once 2000 structures were calculated using simulated annealing, they were refined in explicit water (Linge et al., 2004). After the water refinement the population distribution is fitted with a Gaussian distribution and those structures whose energy is $> 5\sigma$ is removed and refined again with different initial seeds until their energies were $< 5\sigma$.

3.2.3 Programs and structure validation

The program PROCHECK_NMR[Laskowski et al., 1996] was employed to check the stereochemical quality by calculating Ramachandran plots. The program MOLMOL was used to display the structures and to calculate the RMSD-values (Koradi et al., 1996). The combined chemical shift based error ε were calculated with the chemical shift and atom specific weighting factors published by Schumann et al[Schumann et al., 2007].

3.2.4 Theory

In solution, a protein is described by a multistate energetic profile. At a given time t it is described by a space ensemble $S^V : \{s_1, s_2, s_3, ...s_{N_V}\}$, where N_V the number of molecules in the solution. In a typical NMR experiment (0.5 mL of a 1 mM solution) the value of N is approximately equals to 3.01×10^{20}. In addition, for each individual molecule in solution a time ensemble $S^T : \{s_1, s_2, s_3, ...s_{N_T}\}$ is defined as all structural states visited in a time interval Δt, where N_T is the maximum number of possible states in time interval Δt. The actual size of the ensemble is given by the direct product space $S^V \times S^T$. The NMR spectrum obtained in a typical repetitive nD-NMR experiment represents a non-uniform

spatial and temporal average of these states. However, the averaging mechanism depends on different NMR properties (e.g.chemical shift and J-coupling), vary from atom to atom in the same molecule, and may also depend on the path on the energy landscape.

In a time interval Δt the ensemble can be divided in subsets where the exchange between different states is fast on the NMR time-scale for a given atom i. For this subset S_f with N_f molecule, the chemical shift δ_i of an atom i corresponds to the population average $< \delta_i(s_j) >$ of the shifts $\{\delta_i(s_1), \delta_i(s_2), \delta_i(s_3), ...\delta_i(s_{N_f})\}$ is given by,

$$< \delta_i >= \frac{1}{N_f} \sum_{j=1}^{N_f} \delta_i(s_j) \tag{3.1}$$

The fast exchange condition can be defined by,

$$\frac{1}{\tau(s_j, s_k)} >> |\omega_i(s_j) - \omega_i(s_k)| \tag{3.2}$$

where, $\tau(s_j, s_k)$ is the exchange correlation time for the transition between states s_j and s_k and $\omega_i(s_j)$ and $\omega_i(s_k)$ are the resonance frequencies of nucleus i in states s_j and s_k respectively. In its simplest form the fast exchange condition must apply for all pairs of states s_j and s_k. Equation 3.1 can also be written as follows using free enthalpy.

$$< \delta_i >= \sum_{j=1}^{N_f} p(s_j)\delta_i(s_j) = \frac{1}{Z} \sum_{j=1}^{N_f} \delta_i(s_j) e^{\frac{G(s_j)}{RT}} \tag{3.3}$$

Where, $p(s_j)$ is the probability to find state s_j, Z is the state sum over all possible states and $G(s_j)$ is the corresponding free enthalpy. Here only the atoms which satisfy the fast exchange condition are taken into account.

For the sake of simplicity we will restrict ourselves to an ensemble where for all (or essentially all) structures fast exchange conditions is satisfied. Let us denote the experimentally measured chemical shift of atom i as δ_i^e, the predicted average chemical shift of the same atom in a structure s_j in the ensemble as δ_i^p. Then the mean predicted and experimental chemical shifts over the ensemble of N structures is given by,

$$< \delta_i^p >= \frac{1}{N} \sum_{j=1}^{N} \delta_i^p(s_j) \tag{3.4}$$

Using the Hamming distance the error ε_i in the prediction of a single atom i can be defined as,

$$\varepsilon_i = | < \delta_i^p - \delta_i^e > | \tag{3.5}$$

Alternatively, the pairwise RMSD can be defined as,

$$\varepsilon_i^{rmsd} = \frac{1}{N}\sqrt{\sum_{j=1}^{N}(\delta_i^p(s_j) - \delta_i^e)^2} \tag{3.6}$$

The mean error ε for a subset of n atoms (e. g.all backbone atoms HN, N, C^{α}, C in the protein or all atoms of a given amino acid) of structural ensemble is defined as,

$$\varepsilon = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i\varepsilon_i \tag{3.7}$$

where, w_i is the atom specific weighting factor [Schumann et al., 2007] calculated from standard deviation and n is the total number of atoms taken into account. Different magnitudes of the the error values arise from ^1H, ^{13}C and ^{15}N are normalized by the weighting factor. The second moment σ^2 of the errors ε_i for n atoms is given by,

$$\sigma^2 = < w_i^2\varepsilon_i^2 > - < w_i\varepsilon_i >^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i\varepsilon_i - \varepsilon)^2 \tag{3.8}$$

The expectation value of ε should go to zero if (1) the experimental data are error free, (2) the experimental ensemble and the ensemble used for the prediction of the chemical shifts are identical and if (3) the chemical shift calculation is perfect. In practice, all three conditions are not fulfilled. The experimental data have errors that are caused by assignment errors and the limited precision of chemical shift measurements. The experimental ensemble is not known but has to be replaced by an ensemble obtained from the structure calculation, usually from a restrained molecular dynamics simulation. In general, the number of structures used for prediction is also much smaller than the experimental ensemble for which the size is of the order of 10^{20}. Up to now the classical chemical shift calculations are far from perfect although they get better with time. A main point is the parametrization that is usually derived from X-ray structures that are only approximate representatives of the solution ensemble.

The error ε can then be written as a function of the experimental error in the determination of experimental chemical shifts $\Delta\delta^e$, the differences between the experimental ensemble and the calculated ensemble ΔS and the error of the chemical shift prediction methods $\Delta\delta^s$. ε can be written as a Tyalor expanstion to the first order as,

$$\varepsilon(\Delta\delta^e, \Delta S, \Delta\delta^s) = \varepsilon(0,0,0) + \frac{\partial\varepsilon(0,0,0)}{\partial\Delta\delta^e}\Delta\delta^e + \frac{\partial\varepsilon(0,0,0)}{\partial\Delta S}\Delta S + \frac{\partial\varepsilon(0,0,0)}{\partial\Delta\delta^s}\Delta\delta^s \tag{3.9}$$

Although the prediction error $\Delta\delta^s$ not only depends on the simulation method C used and the atom types T included in calculation but also oThe sequential assignments of the NMR signals of the set of proteins (Table 1) were taken from the BMRB data base, the corresponding NMR structures from the PDB-data base. Sequential assignments for wild type HPr(wt) from S.aureus were taken from Maurer et al. (2004), for the mutant HPr(H15A) from Munte et al. (manuscript in preparation). n specific structural properties of the protein under consideration. For a given method C it can be approximated by a constant $\Delta\delta^s(C,T)$ and the corresponding derivative by 1 when enough atoms are involved in the calculation. This means that equation 3.9 can be written as,

$$\varepsilon = \frac{\partial\varepsilon(0,0,0)}{\partial\Delta S} + \Delta\delta^s(C,T) \tag{3.10}$$

At first glance, the error ΔS of the structural ensemble depends on two factors the correctness of the structures and the minimum number of structures and their selection for describing the ensemble from the point of averaged chemical shifts. Assuming a harmonic potential for the energy, (Gronwald and Kalbitzer, 2004) Molecular dynamic simulations will produce a Gaussian probability distribution for the population of states. Therefore, ΔS is a function of the arbitrarily chosen number of structures N that are either ordered according to their energies with the lowest energy assigned to structure s_i or according to the probability.

3.3 Results

3.3.1 Prediction of chemical shifts in a test data set

Wang and Jardetzky [Wang and Jardetzky, 2002] prepared a data set of proteins where high resolution structures and heteronuclear chemical shift data were available for the development of a new method of secondary structure prediction. Here, we use a subset of 16 NMR structures for an analysis of the chemical shift predictions (Table 3.1) where structural bundles are deposited in PDB database. For obtaining an estimate of $\Delta\delta^s$ the chemical shifts were calculated with the program SHIFTX and SHIFTS for these structures and compared with the experimental data. The mean ε and the second moment σ^2

were calculated using equations 3.7 and 3.8.

Table 3.1: Test Data Set

Protein (no.of residues)	BMRB ID (pH, T)	PDB ID (ensemble size)
Bet v 1-L (159)	4417 (pH 7.0, 298 K)	1B6F (23)
P14a (135)	4301 (pH 5.5, 303 K)	1CFE (20)
CA RSV (262)	4384 (pH 6.0, 303 K)	1D1D (20)
Pathogenesis-protein (159)	4671 (pH 7.0, 298K)	1E09 (22)
Dynein light chain 8 (89)	4911 (pH 7.0, 298K)	1F96 (20)
Phosphoglycerate mutase (211)	4648 (pH 6.4, 310K)	1FZT (21)
HTL V-1 capsid protein(134)	4649 (pH 6.0, 302K)	1GO3 (20)
β2-GP1 domain V (86)	4981 (pH 6.0, 298K)	1G4F (20)
ERp29 C-domain(120)	4920 (pH 4.9, 308K)	1G7D (20)
ERp29 N-domain(137)	4919 (pH 4.9, 308K)	1G7E (20)
CDC4P (141)	4851 (pH 6.5, 303K)	1GGW (26)
Vam3p N-terminal (123)	4945 (pH 6.0, 302K)	1HS7 (20)
Mouse doppel (132)	4938 (pH 5.2, 299K)	1I17 (20)
Core binding factor (143)	4092 (pH 6.6, 293K)	2JHB (20)
Rabphilin_3_C2B (140)	4360 (pH 6.1, 304K)	3RPB (20)
P55 (166)	4321 (pH 6.5, 300K)	5GCN(24)

The deviation of the predicted combined chemical shifts $\Delta\delta_{comb}$ of the backbone atoms from the experimental chemical shifts are shown in Figure 3.1. The chemical shifts were calculated from the first structure in the data base (usually the lowest energy structure) and the total ensemble. In general the performance of SHIFTX is slightly better for all structures studied than that of SHIFTS, for the single structure as well as for the ensemble. The weighted mean of error ε over all 16 proteins drops from 0.63 ppm to 0.58 ppm for SHIFTX and from 0.66 ppm to 0.61 ppm for SHIFTS. The non-weighted average for all atoms (Tables 3.2 and 3.3) drops from 0.66 ppm to 0.61 ppm for SHIFTS and

from 0.63 ppm to 0.58 oom for SHIFTX. The same trend, that SHIFTX gives a more correct prediction than SHIFTS, is observed for the majority of the predicted chemical shifts of groups of atoms (Table 3.2 and 3.3). Especially the predicted chemical shifts of the backbone resonances are more precise for SHIFTX.

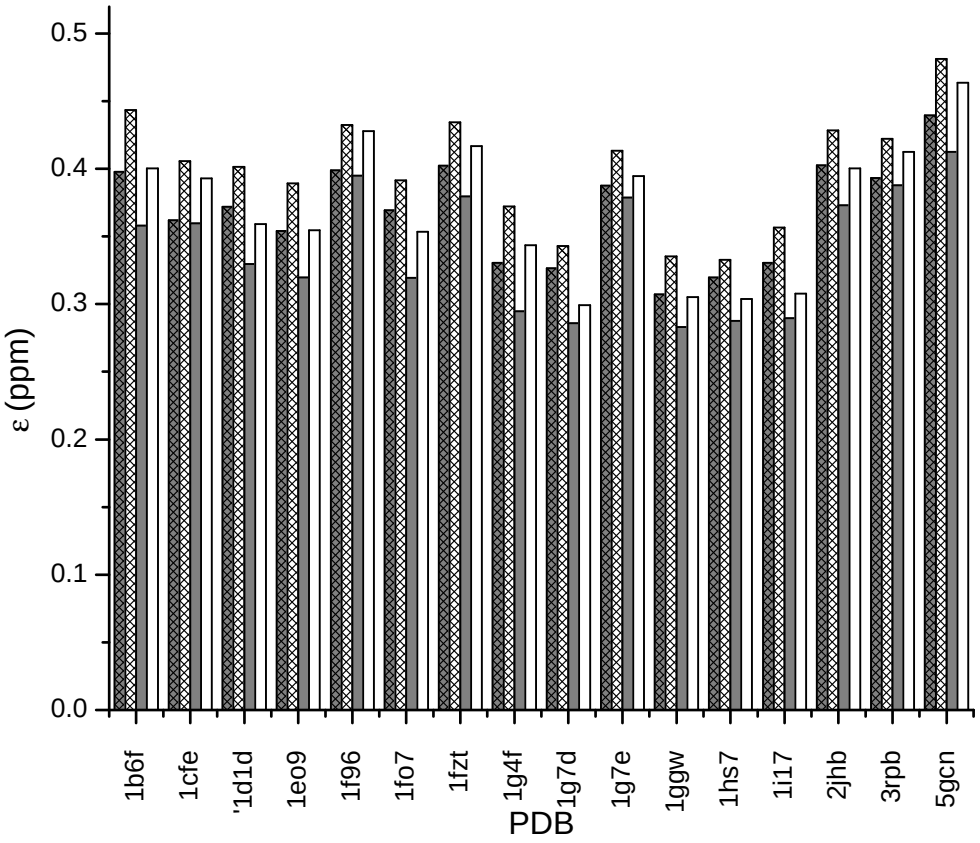


Figure 3.1: Accuracy of chemical shift predictions. For the structures listed in Table 1 chemical shifts were calculated from the lowest energy structure (SHIFTX(white bar),SHIFTS(checked white bars)) and the structural ensemble (SHIFTX(gray bars), SHIFTS(checked gray bars)).

Table 3.2: Average performance of chemical shift prediction for specific atoms using SHIFTS. The average chemical shift differences ε_i were calculated using the Hamming distance (equation 3.6,3.7) and a weighting factor $w_i = 1$, that is the ε_i were calculated as average of all proteins listed in Table 3.1. The second moments $< \sigma_i >$(values in brackets) were calculated by applying equation3.8. For stereospecifically not assigned atoms such as methylene protons the chemical shifts of the corresponding protons were averaged before calculating the difference

Atom type	Lowest Energy Structure $\varepsilon^S(\sigma)$ppm	Ensemble $\varepsilon^E(\sigma)$ppm	$\frac{\varepsilon^E-\varepsilon^S}{\varepsilon^S}$
HN	0.53 (0.43)	0.50 (0.41)	-0.06
N	3.50 (2.54)	3.45(2.55)	-0.01
H^{α}	0.31 (0.28)	0.28 (0.26)	-0.10
C^{α}	1.22 (1.03)	1.06 (0.92)	-0.13
C	1.61 (1.30)	1.48 (1.18)	-0.08
H^{β}(methylen)	0.25 (0.25)	0.23 (0.23)	-0.08
H^{β}(methyl)	0.18 (0.18)	0.16 (0.17)	-0.11
C^{β}(methylen)	1.29 (1.13)	1.10 (0.97)	-0.15
C^{β}(methyl)	1.19 (1.02)	1.03 (0.85)	-0.13
H^{γ}(methylen)	0.25 (0.26)	0.21 (0.21)	-0.16
H^{γ}(methyl)	0.23 (0.23)	0.21 (0.20)	-0.09
H^{δ}(methylen)	0.21 (0.22)	0.20 (0.19)	-0.05
H^{δ}(methyl)	0.19 (0.20)	0.17 (0.17)	-0.11
H^{δ}(aromatic)	0.23 (0.20)	0.21 (0.17)	-0.09
H^{δ}(amide)	0.59 (0.32)	0.56 (0.29)	-0.05
H^{ε}(methylen)	0.17 (0.20)	0.15 (0.16)	-0.18
H^{ε}(methyl)	0.28 (0.16)	0.21 (0.13)	-0.25
H^{ε}(aromatic)	0.31 (0.25)	0.30 (0.23)	-0.03
H^{η}(aromatic)	0.42 (0.20)	0.39 (0.20)	-0.07
H^{ζ}(aromatic)	0.24 (0.23)	0.23 (0.21)	-0.04
Average error for all atoms	0.66 (0.53)	0.61 (0.49)	-0.08
Weighted average for all atoms	0.40 (0.38)	0.37 (0.35)	-0.08
Weighted average for backbone atoms	0.77 (0.72)	0.70 (0.69)	-0.09
Weighted average for all sidechain atoms	0.38(0.36)	0.36 (0.33)	-0.05

Table 3.3: Average performance of chemical shift prediction for specific atoms using SHIFTX. The average chemical shift differences ε_i were calculated using the Hamming distance (equation 3.6,3.7) and a weighting factor $w_i = 1$, that is the ε_i were calculated as average of all proteins listed in Table 3.1. The second moments $< \sigma_i >$(values in brackets) were calculated by applying equation3.8. For stereospecifically not assigned atoms such as methylene protons the chemical shifts of the corresponding protons were averaged before calculating the difference

Atom type	Lowest Energy Structure $\varepsilon^S(\sigma)$ppm	Ensemble $\varepsilon^E(\sigma)$ppm	$\frac{\varepsilon^E-\varepsilon^S}{\varepsilon^S}$
HN	0.52 (0.43)	0.48 (0.40)	-0.08
N	2.53 (1.98)	2.38(1.88)	-0.02
H^{α}	0.28 (0.25)	0.24 (0.22)	-0.14
C^{α}	1.04 (0.93)	0.95 (0.85)	-0.08
C	1.28 (1.07)	1.22 (1.01)	-0.05
H^{β}(methylen)	0.23 (0.22)	0.21 (0.21)	-0.09
H^{β}(methyl)	0.20 (0.18)	0.18 (0.16)	-0.10
C^{β}(methylen)	1.13 (1.01)	1.01 (0.91)	-0.12
C^{β}(methyl)	0.99 (0.94)	0.84 (0.78)	-0.15
H^{γ}(methylen)	0.23 (0.23)	0.21 (0.21)	-0.09
H^{γ}(methyl)	0.22 (0.23)	0.20 (0.20)	-0.09
H^{δ}(methylen)	0.21 (0.20)	0.19 (0.17)	-0.10
H^{δ}(methyl)	0.26 (0.24)	0.22 (0.20)	-0.15
H^{δ}(amide)	0.63 (0.41)	0.55 (0.35)	-0.13
H^{ε}(methylen)	0.15 (0.15)	0.14 (0.14)	-0.07
H^{ε}(methyl)	0.38 (0.24)	0.39 (0.23)	-0.03
H^{ε}(amide)	0.42 (0.30)	0.40 (0.27)	-0.05
Average error for all atoms	0.63 (0.53)	0.58 (0.48)	-0.08
Weighted average for all atoms	0.37 (0.34)	0.34 (0.30)	-0.08
Weighted average for backbone atoms	0.70 (0.68)	0.64 (0.63)	-0.09
Weighted average for all sidechain atoms	0.35(0.33)	0.33 (0.32)	-0.06

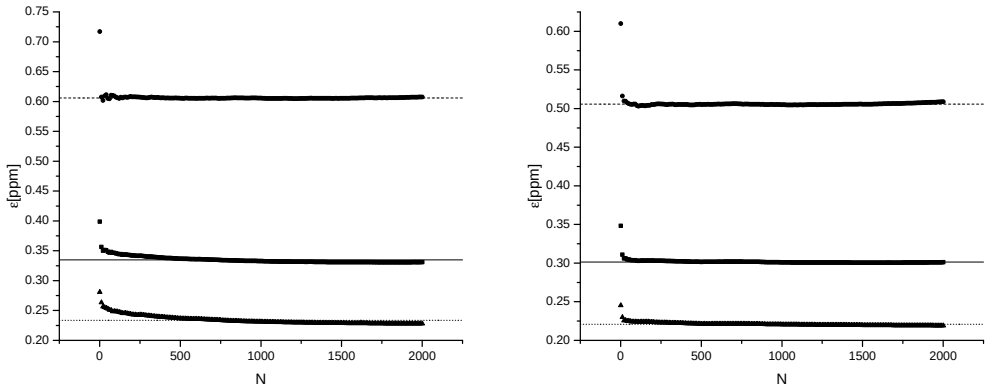


Figure 3.2: Dependence of the chemical shift error ε to the size of the structural ensemble before water refinement for HPr(WT) using SHIFTS & SHIFTX. The mean error of the back bone atoms $HN, H^{\alpha}, N, C^{\alpha}, C$ (circle), side chain atoms(triangle) and all atoms(square) were plotted as a function of the size of the ensemble. The data points were fitted with the function $\varepsilon = \frac{1}{N\sqrt{2\pi}} e^{-\frac{(lnN)^2}{2\sigma^2}} + C$ and the value of C shown as dashed line for backbone atoms,dotted line for side chian atoms and solid line for all atoms.

3.3.1.1 Effect of ensemble size and quality

In line with the fact that the experimentally observed chemical shifts are ensemble averages, the chemical shift predictions calculated as averages over the structural ensemble, the prediction of the chemical shifts from the ensembles is always more correct. This is clearly seen for the weighted shifts depicted in Figure 3.1 where the chemical shift prediction is more accurate for all proteins when ensembles are used. Also for the ensembles the predictions by SHIFTX are again always better than those of SHIFTS. Also for the individual groups the ensemble prediction is always more precise than that obtained from the lowest energy structure independent of the prediction method used (Table 3.2 and Table 3.3).

The data base used contains only relatively small structural ensembles, usually of the order of 20 structures. A complete description of a real structural ensemble would require a much larger number of structures. Therefore, we calculated large ensembles of 2000 structures each for wildtype HPr [Maurer et al., 2004] and a mutant of HPr, HPr(H15A)

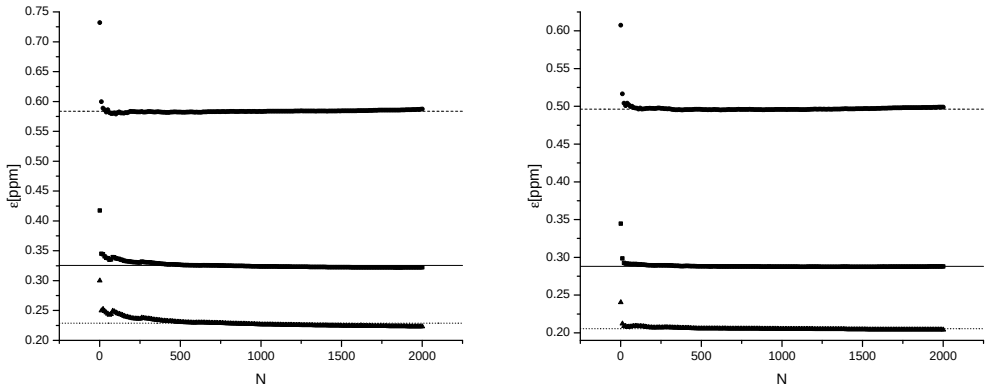


Figure 3.3: Dependence of the chemical shift error ε to the size of the structural ensemble before water refinement for HPr(H15A) using SHIFTS & SHIFTX. The mean error of the back bone atoms $HN, H^{\alpha}, N, C^{\alpha}, C$ (circle), side chain atoms(triangle) and all atoms(square) were plotted as a function of the size of the ensemble. The data points were fitted with the function $\varepsilon = \frac{1}{N\sqrt{2\pi}} e^{-\frac{(lnN)^2}{2\sigma^2}} + C$ and the value of C shown as dashed line for backbone atoms,dotted line for side chian atoms and solid line for all atoms.

from S. aureus (Munte et al., to be published). The number of experimental restraints used for the simulated annealing (SA) is given in Table 3.4. Structures that obviously did not converge and therefore showed large violations of the experimental restraints (see Materials and Methods) were removed before analysis. The remaining structures were ordered according to their total energies (not including the pseudo energies from experimental restraints) and the weighted cumulative shift difference $\varepsilon(N)$ (with s_1 the lowest energy structure) was plotted for the backbone atoms as well as for all atoms (more precisely for all atoms with assigned chemical shifts). Figure 3.2 and 3.3 shows clearly that $\varepsilon(N)$ first decreases substantially for the two proteins and shows an asymptotic stable behaviour. The shape of the function is virtually independent on the prediction method used; however, the magnitude of the effect strongly depends on the atoms selected: the lowest values are obtained for the side chain atoms, the highest values for the main chain atoms, and intermediate values for the weighted average of all atoms (constant C in Table 3.5). The data can rather well be fitted by a *lognormal* distribution with an additional offset. Refinement of the obtained structures in explicit water in general leads to an improved quality of the structures and possibly also to a change of $\varepsilon(N)$. Therefore, all the 2000 structures were

subjected to a water refinement and ε was recalculated. When the refined structures are given as input, the prediction error decreases significantly at the same ensemble size (Figure 3.4 and 3.5). However, for large ensembles the asymptotic value differs only slightly by a few percent (Table 3.5). The largest differences are observed for small ensemble sizes (Figure. 3.6). Here, water refinement leads to much smaller initial values for the prediction errors. The data can be sufficiently well fitted by the lognormal distribution, however at very small sample sizes a substructure is clearly observable. Without water refinement about 18 structures are necessary for coming close to the asymptotic value, after water refinement only 10 structures are required.

For the structures contained in the experimental data base (Table 3.1) the same analysis leads to analogous results when we assume that the structures are ordered according to their energies (a fact not known). In general, the ensemble gives better performance of the chemical shift prediction by SHIFTX and SHIFTS (Figure 3.7. In most cases a minimum value seems to be reached when about 18 structures are used for the calculations.

Table 3.4: Experimental NMR restraints

	HPr(WT)	HPr(H15A)
NOE Restraints	1219	1248
Dihedral	130	130
J Coupling	78	69
H bond	51	53

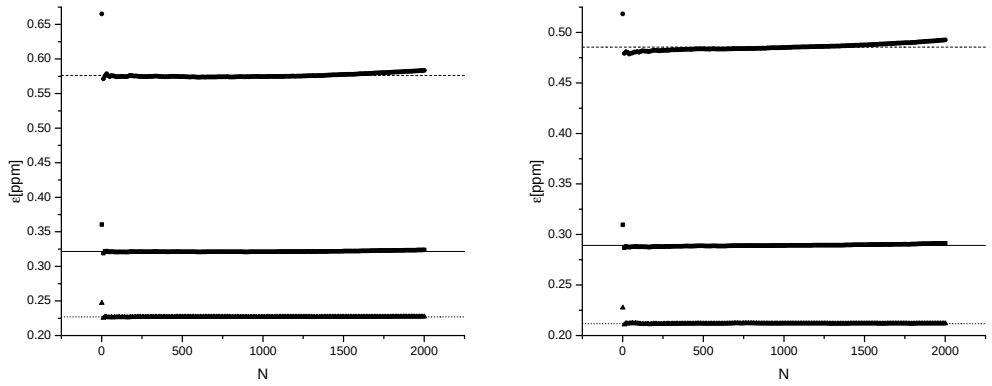


Figure 3.4: Dependence of the chemical shift error ε to the size of the structural ensemble after water refinement for HPr(WT) using SHIFTS & SHIFTX. The mean error of the back bone atoms $HN, H^{\alpha}, N, C^{\alpha}, C$ (circle), side chain atoms(triangle) and all atoms(square) were plotted as a function of the size of the ensemble. The data points were fitted with the function $\varepsilon = \frac{1}{N\sqrt{2\pi}} e^{-\frac{(lnN)^2}{2\sigma^2}} + C$ and the value of C shown as dashed line for backbone atoms,dotted line for side chian atoms and solid line for all atoms.

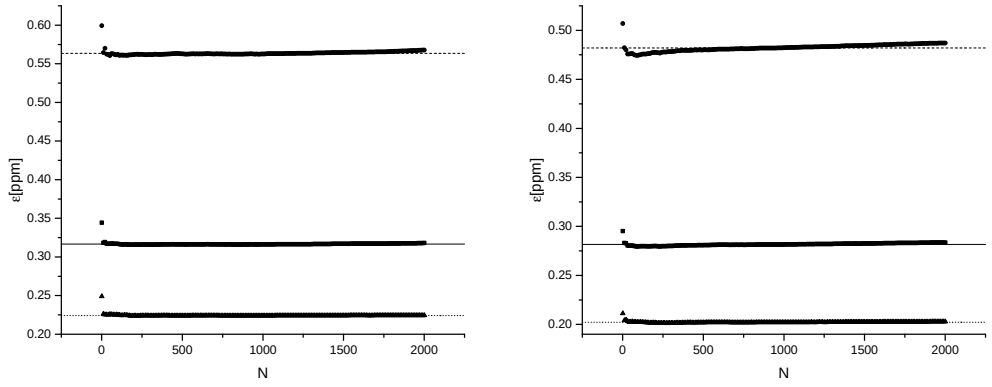


Figure 3.5: Dependence of the chemical shift error ε to the size of the structural ensemble after water refinement for HPr(H15A) using SHIFTS & SHIFTX. The mean error of the back bone atoms $HN, H^{\alpha}, N, C^{\alpha}, C$ (circle), side chain atoms(triangle) and all atoms(square) were plotted as a function of the size of the ensemble. The data points were fitted with the function $\varepsilon = \frac{1}{N\sqrt{2\pi}} e^{-\frac{(lnN)^2}{2\sigma^2}} + C$ and the value of C shown as dashed line for backbone atoms,dotted line for side chian atoms and solid line for all atoms.

Table 3.5: Minimum ensemble size and error offset. Fit parameters for the function $\varepsilon = \frac{1}{N\sqrt{2\pi}} e^{-\frac{(lnN)^2}{2\sigma^2}} + C$

Ensemble	Atoms	σ(SHIFTS) (ppm)	C(SHIFTS) (ppm)	σ(SHIFTX) (ppm)	C(SHIFTX) (ppm)
HPr(WT)	All	0.026	0.33	0.019	0.30
SA	Backbone	0.044	0.61	0.041	0.51
	Sidechain	0.019	0.23	0.010	0.22
HPr(WT)	All	0.015	0.32	0.008	0.29
WREF	Backbone	0.036	0.58	0.013	0.49
	Sidechain	0.008	0.23	0.006	0.21
HPr(H15A)	All	0.037	0.32	0.022	0.29
SA	Backbone	0.059	0.58	0.044	0.50
	Sidechain	0.028	0.23	0.014	0.21
HPr(H15A)	All	0.011	0.32	0.005	0.28
WREF	Backbone	0.014	0.56	0.010	0.48
	Sidechain	0.009	0.22	0.004	0.20

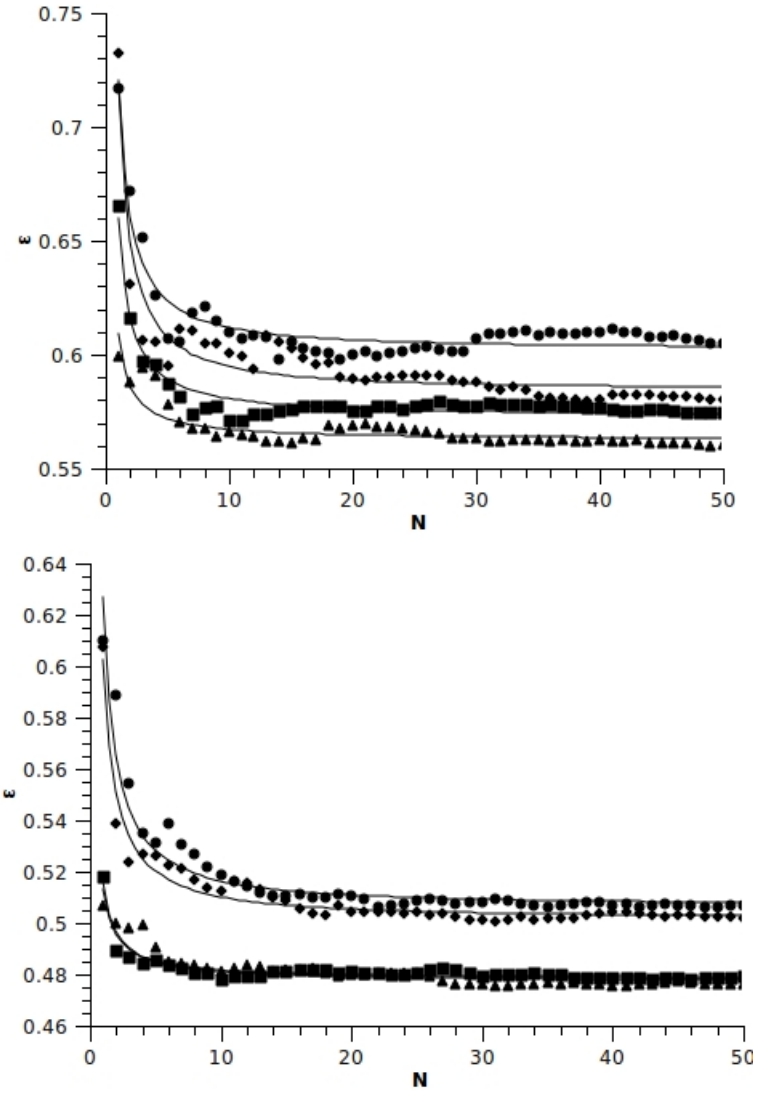


Figure 3.6: chemical shift error as a function of the ensemble size. The mean error of the backbone atoms using (A)SHIFTS and (B)SHIFTX are plotted as function of the size of the ensemble . Only the first 50 structures are shown. HPr-wildtype before (circle) and after (square) water refinement, HPr(H15A) before (diamond) and after (triangle) water refinement. Solid line shows the lognormal($\varepsilon = \frac{1}{N\sqrt{2\pi}} e^{-\frac{(lnN)^2}{2\sigma^2}} + C$) for the corresponding data points.

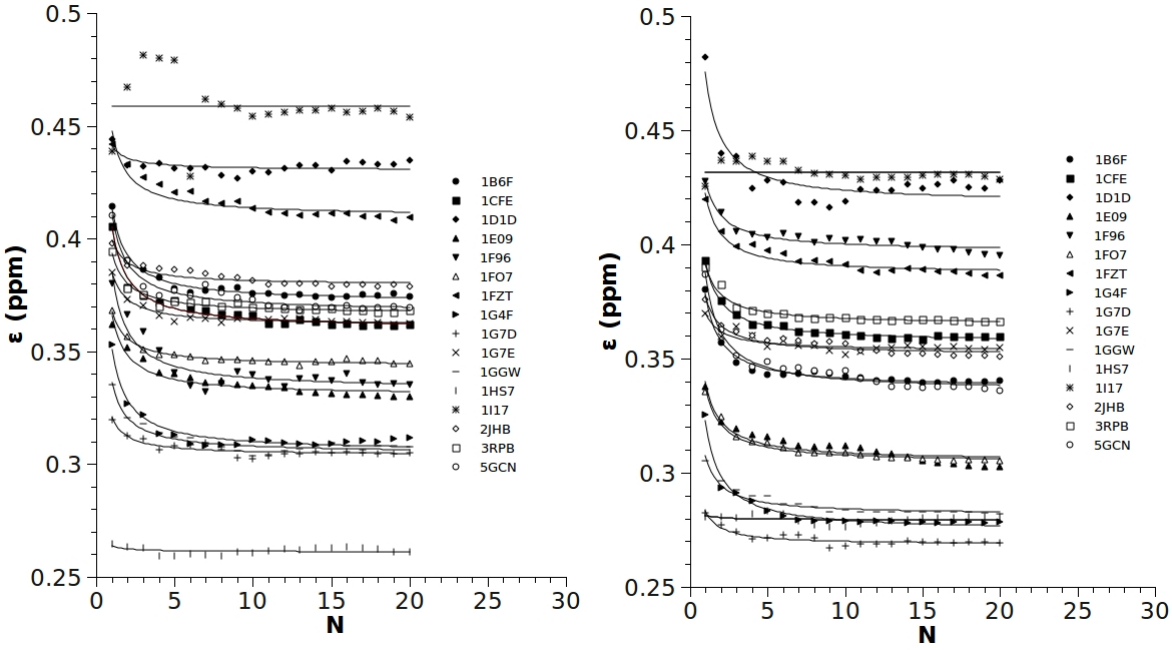


Figure 3.7: Chemical shift error as a function of the ensemble size for the structural data base. The curves shown are fit curves as defined in Figure 3.2 of the structures contained in the experimental data base (Table 3.1). The chemical shift predictions were performed with SHIFTS and SHIFTX.

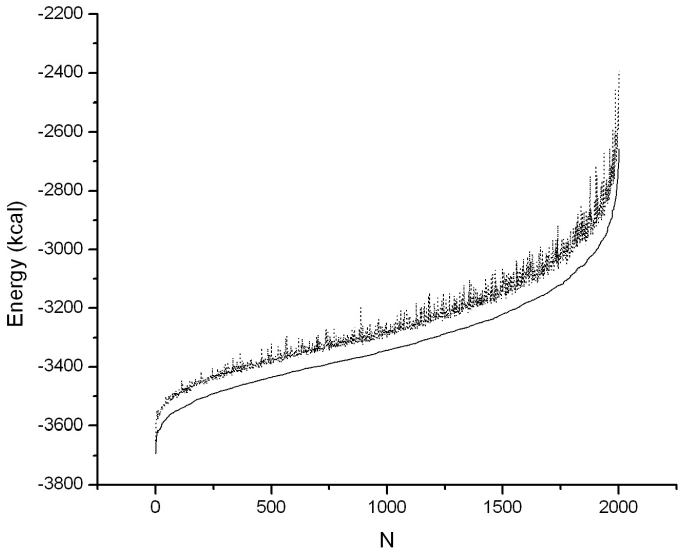


Figure 3.8: The structures of HPr(WT) were ordered according to their total energy (N= 1,...,2000). The total energy(___) and the sum of the total energy and the violation energy(...) were plotted as a function of N

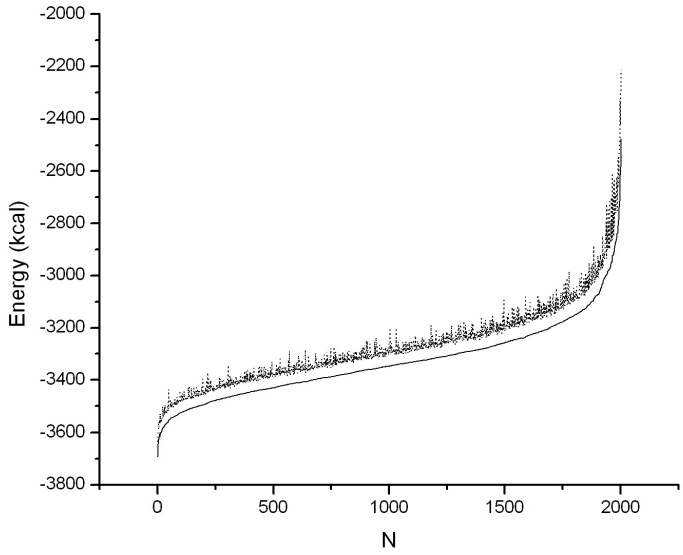


Figure 3.9: The structures of HPr(H15A) were ordered according to their total energy (N= 1,...,2000). The total energy(___) and the sum of the total energy and the violation energy(...) were plotted as a function of N

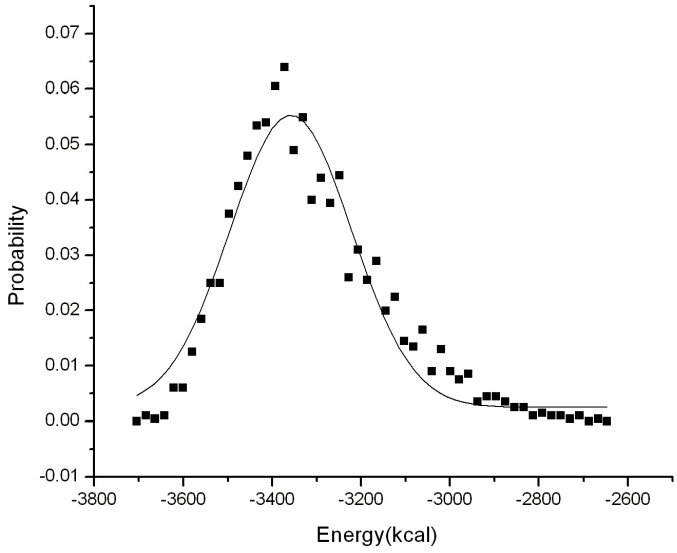


Figure 3.10: The probability of each energy state was plotted as function of total energy (all energies excluding the experimental pseudo energies) and fitted with Gaussian function for HPr(WT): $< E >= -3358.5\ kcal/mol$, $\sigma = 274.0\ kcal/mol$

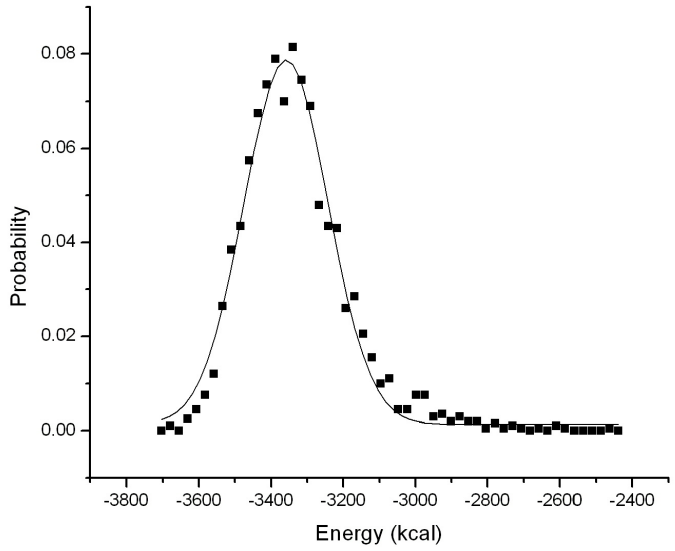


Figure 3.11: The probability of each energy state was plotted as function of total energy (all energies excluding the experimental pseudo energies) and fitted with Gaussian function for HPr(H15A): $< E >= -3358.5\ kcal/mol$, $\sigma = 233.5\ kcal/mol$

3.3.1.2 Energy distributions and their impart on the chemical shift prediction

The ensembles obtained for wild type HPr and its mutant by simulated annealing followed by refinement in explicit water should not determined mainly by the experimental pseudo energies but by the physical model itself. Figure 3.8,3.9 shows the energies with and without inclusion of the pseudo energies resulting from the restraint violation. The energies were ordered according to their magnitude for the 2000 structures. It is obvious that the restraint violations only contribute little to the energy. The probability distributions of the energies are represented in Figure 3.10 and 3.11 can be fitted in a good approximation by a Gaussian. The quality of the chemical shift prediction of smaller sets of structures may depend on the total energy of the structures under consideration. This was tested for wildtype and mutant HPr in two different ensemble types:

- Ensembl of structures having same energies(in terms of standard deviation σ) or
- Enselble of structures haveing fixed(20) number of structures.

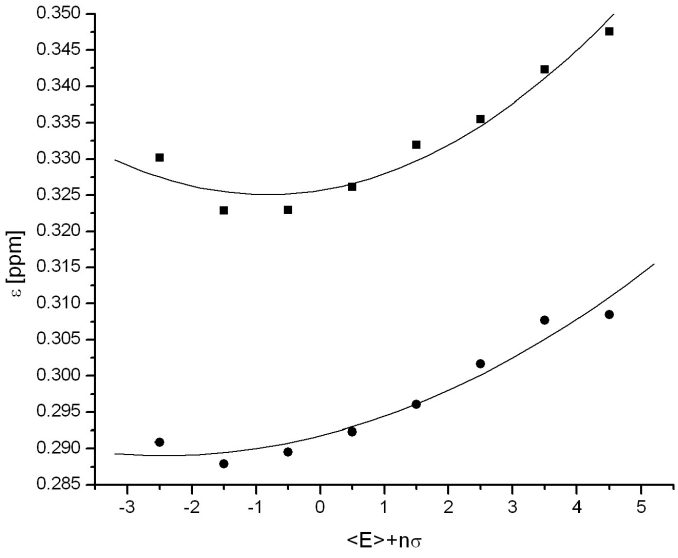


Figure 3.12: The prediction error for the backbone atoms is plotted as a function of the total energies for HPr(WT). The ensembles were created based on energy. The data were fitted with a polynomial of the second order. Squares:prediction with SHIFTS, circles:prediction with SHIFTX

An ensemble can be characterised by parameters like the number of structures N and the mean energy $E_{average}$. Structures from a large ensemble can be subdivided into smaller ensembles in many ways. One would define a $E_{average}$ for a small ensemble and select those structures whose energy lies between $E_{average} + \Delta E$ and $E_{average} - \Delta E$. If such a selection criteria is used, than the size of the ensembles may differ. Figure 3.12 and 3.13 indicates the error dependence on ensemble selected from different energy regime. Having $< E_{total} >$ as the most probable energy and σ as standard deviation of the large ensemble, the upper and lower bound of smaller ensembles can be defined as $< E_{total} > + n\sigma$ and $< E_{total} > + (n+1)\sigma$, where n runs from -5 to +5. Hence average energy of each small ensemble is equal to $< E_{total} > + \frac{n}{2}\sigma$.

On the other hand, smaller ensembles can created by subdividing the large ensemble with equal size ensemble. Figure 3.14 and 3.15 shows the error dependence on equal

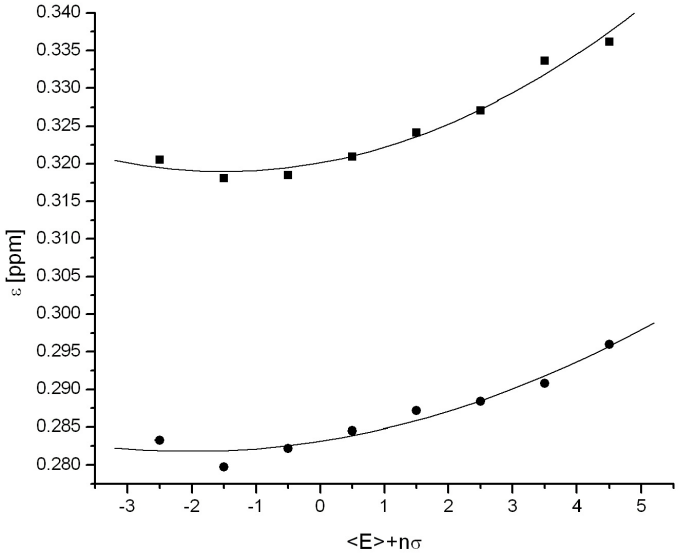


Figure 3.13: The prediction error for the backbone atoms is plotted as a function of the total energies for HPr(H15A). The ensembles were created based on energy. The data were fitted with a polynomial of the second order. Squares:prediction with SHIFTS, circles:prediction with SHIFTX

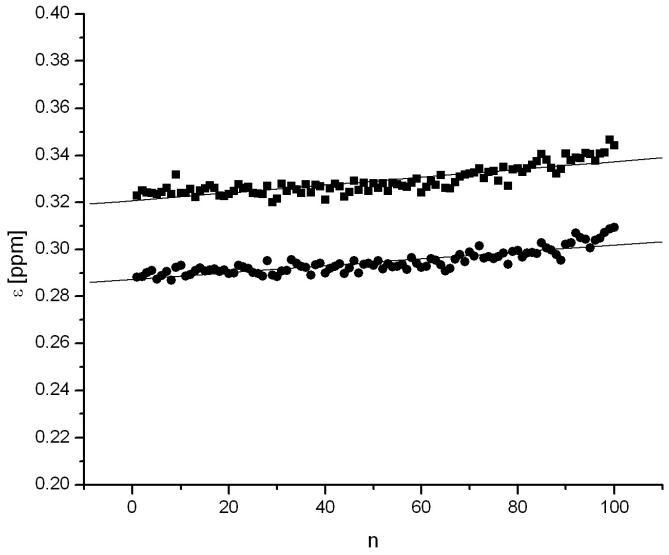


Figure 3.14: The prediction error for the backbone atoms is plotted as a function of the total energies for HPr(WT). The ensembles were taken as equal size(20 structures). The data were fitted with a polynomial of the first order. Squares:prediction with SHIFTS, circles:prediction with SHIFTX

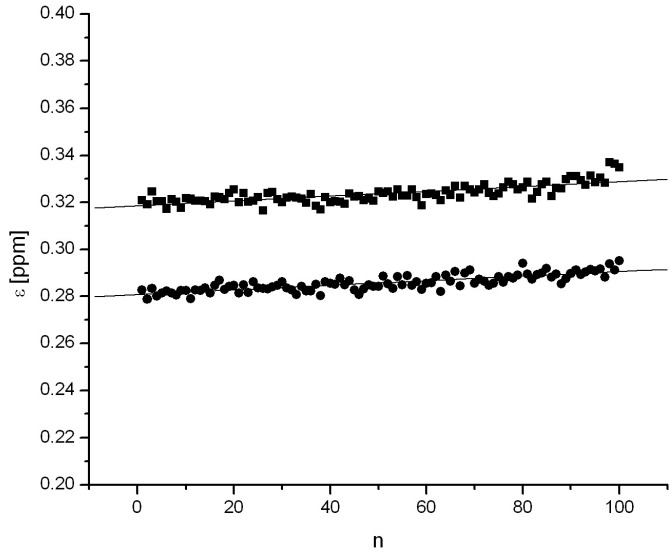


Figure 3.15: The prediction error for the backbone atoms is plotted as a function of the total energies for HPr(H15A). The ensembles were taken as equal size(20 structures). The data were fitted with a polynomial of the first order. Squares:prediction with SHIFTS, circles:prediction with SHIFTX

size of ensemble. The 2000 structures are subdivided in to 100 small ensemble having 20 structures each. The data points can be fitted well with straight line having slopes of the order of 0.5×10^{-06} ppm.

As shown in Figure 3.12,3.13,3.14 and 3.15 the first case, a second order polynomial was required to fit the data, in the second case a first order polynomial was necessary.

3.3.2 Discussion

Multi-conformational ensembles and conformational averaging of chemical shifts

From first principles of thermodynamics it is clear that protein structures in solution forms a large ensemble of multiple conformational states. A complete description of a protein would require the knowledge of all coexisting structures; even the knowledge of energetically unfavourable states that are only weakly populated may be important since functional excited states may be contained in the higher energy part of the energy landscape[Kalbitzer et al., 2009]. A complete representation of all states is practically not

possible because of the extremely large number of states. Even when one restricts to the ground state, only the representation of a limited number of structures is feasible. However, it is not clear, which structures should be selected and how many structures must be included for a faithful representation of the ensemble.

Indeed, the definition of a faithful representation of the ensemble depends on the properties of the ensemble that should be represented. In biochemistry, it would often focus on the explanation of functional properties. In the present context, it is the general question (1) if in agreement with theory the quality of the chemical shift prediction can be increased by using ensembles of structures, (2) if the improvement of the prediction by using ensembles is independent of the method used, (3) what is the minimum size of the ensemble required for optimum chemical shift prediction, and (4) can the chemical shift prediction be used to define the representative ensemble.

Since in the ground state (assumed as a single minimum in the energy landscape) the transition between the conformations should be fast, the ensemble averaged chemical shifts may represent a suitable measure to determine the size of a representative structural set. In fact, when chemical shifts could be calculated perfectly, the difference between the experimental shifts and the population averaged chemical shift calculated from the complete structural ensemble should be zero.

Quality of the chemical shift prediction.

For testing and quantifying the quality of the chemical shift calculation we used a set of NMR structures (Table 3.1) previously designed by [Wang and Jardetzky, 2002]. In general, from the two programs tested here SHIFTX performs somewhat better than SHIFTS, the weighted average error of all atoms of 0.340 ppm calculated with SHIFTX for the lowest energy structure is about 8.2% lower than 0.368 ppm calculated with SHIFTS (Tables 3.2 and 3.3). The standard deviations of the errors ε are for all atom groups rather high (almost as large as the mean itself), indicating that either the structural quality varies much or that the parametrization is not optimal for all conditions found in the structures. From the data itself this cannot be decided but probably variations in the structural quality may be the dominant factor for these variations. The prediction error varies for the different

atom types. For SHIFTX we found mean errors of 0.52, 0.28, 2.53, 1.04, and 1.28 ppm for the *HN*, H^{α}, *N*, C^{α}, *and C* atoms respectively. Similar results were published most recently by [Lehtivarjo et al., 2009] for a smaller data base of protein structures with 0.55 and 0.37 ppm for the *HN and* H^{α} resonances. Similar results were also published earlier by [Arun and Langmead, 2004]. For the side chain atoms the prediction error is usually much smaller (Tables 3.2 and 3.3), one factor is the smaller chemical shift variations found here experimentally. However, this cannot be the only reason since the backbone prediction error is also larger than the side chain prediction error when it is calculated with the amino acid type and atom type specific weighting factors [Schumann et al., 2007] that correct for the chemical shift distribution of the atoms under consideration. The chemical shift prediction by SHIFTS and SHIFTX (an by all methods published so far) is still more than one order of magnitude too inaccurate when it should be used for a direct assignment of resonances: here a precision of the order of the typical line width would be required that is about 0.01 ppm for proton and about 0.1 ppm for nitrogen resonances.

Prediction of chemical shifts from the ensemble vs the lowest energy structure

In accordance with the fact that chemical shifts represent ensemble averages the use of ensembles generally improves the chemical shift prediction for all structures of the data base (Fig. 3.1 and for most of the atoms taken into account (Tables 3.2 and 3.3). The weighted mean error of all atoms decreases by 8.82% when SHIFTX is used and 8.30% when SHIFTS is used. It also is to be expected that an improvement by averaging over ensembles is independent on the prediction method used as it is shown here for SHIFTS and SHIFTX. However,for 4D-shift predictions[Lehtivarjo et al., 2009] a similar result has been reported recently.

Minimum size of the ensemble required for shift prediction

Usually 10 to 20 NMR structures are stored in the data base. Our result indicates that this is sufficient as far as the chemical shift prediction is concerned. The extensive simulation of the HPr structures shows that an asymptotic value is reached before water refinement when about 18 structures are averaged (Fig. 3.6 and 3.7), after water refinement when

about 10 structures are averaged (Fig. 3.6). Under this aspect the traditional way to deposit NMR structures can be considered as sufficient. When during the calculation of the structures those structures are removed that show large violations of the experimental restraints and thus have not converged properly, the error dependence of the chemical shift prediction on the size of the ensemble can be sufficiently well described by a lognormal distribution with a constant offset. Whereas the description with a lognormal distribution is purely empirical, the asymptotic behaviour to a constant value can be expected from the chemical shift averaging. However, when the structures with larger pseudo energies are included, a continuous increase of the prediction error with the number N can be observed.

Dependence of the prediction error on the energy distribution

From a general point of view it is surprising that a very small number of low energy structures can lead to a virtual optimum ensemble energy structures, although they clearly are not representative for the experimental ensemble but only represent weakly populated states. The obtained energy distributions are shown in Fig. 3.10 and 3.11 for HPr that can be approximated well by a Gaussian. Structures having energy values less than -2 relative to the mean energy are considered as lowest energy structures, energy values between $-\sigma$ to $+\sigma$ are considered as most probable structures and structures above +2 are considered as high energy structures. When the prediction error is plotted as a function of the deviation from the mean a minimum prediction error is detected close to the most probable ensemble at the mean energy (Fig. 3.12,3.13) as to be expected from theory. However, the effect is rather small. In the intervals between $[< E > -\sigma, < E >]$ and $[< E >, < E > +\sigma]$ the number of calculated structures is much higher than in the other intervals. This may cause a bias on the data evaluation favouring the chemical shift prediction from the larger ensemble of structures. Therefore, the structures were sorted according to their energies and sets of identical size (20 structures) were taken for the chemical shift prediction. Here, a minimum of the error cannot be detected any more but the prediction error is almost constant and can be well approximated by a straight line with a very small positive slope (Fig. 3.14,3.15).

Prediction Error

The experimentally observed prediction error $\Delta\delta^s(C,T)$ (eq. 3.10) is still rather large for all prediction methods C and for all atoms T considered and is especially large for the backbone atoms. It is much larger than the effects resulting from the ensemble averaging itself, according to the analysis of our structural data basis (Table 3.1) the ensemble effect ΔS is of the order of 10% to 20% of the $\Delta\delta^s(C,T)$. Therefore, it is not surprising that we can only observe small effects from the selection of the structural ensemble (Fig. 3.12,3.13). In fact, using the same ensemble size, the quality of the ensemble prediction slightly decreases with the mean energy of the structural set. This bias may be caused by the parametrization procedure of the calculation methods itself that are optimized to the lowest energy NMR structures and/or the crystal structures that clearly do not represent the solution ensemble measured experimentally. In a most recent paper, that appeared during the preparation of this manuscript, [Lehtivarjo et al., 2009] show that indeed better results can be obtained when MD-ensembles are used for parametrization.

There are a number of good reasons to use an ensemble of structures instead of a single energy minimized structure. The minimum energy structure calculated by restraint molecular dynamics, simulated annealing, and water refinement are actually calculated from subset of structures in the conformation space. There is no guaranty that this subset contains the true lowest energy structure. The other good reason to use ensemble representations is to account for the presence of different conformers (local energy minima), that is the lowest energy structure may not be unique. Experimentally, chemical shift prediction can be improved significantly when a structural ensemble is used. An ensemble size of about 20 structures is sufficient, further increase of the size seems not to lead to better results. The conclusion primarily holds for the two tested, most popular prediction programs SHIFTX and SHIFTS but for theoretical reason most probably also applies for other prediction programs.

CHAPTER 4

REFINED CHEMICAL SHIFT STATISTICS

4.1 Introduction

Statistical informations are essential to calculate certain probability values to make decisions in the automation process. An unbiased set of experimental observation and structural information is necessary to obtain useful statistical parameters for structure determination. In general structure determination starts with primary structure (sequence) followed by experimental observation and finally ends in structure calculation. In X-ray crystallography, experimental observations can be directly correlated to 3D-structure by analytical functions (Fourier transform). But in case of protein structure determination by NMR spectroscopy, experimental observations has to be interpreted based on the knowledge about previous experimental observation. However previous experimental observation can not be readily taken as a standard, since the protein under observation may exhibit completely different structural features. The best way to overcome this problem is to have a statistical correlation between primary structure (sequence),experimental observation and final 3D structure. It is important to construct statistically unbiased data bases, before studding statistical correlation between them.

Two different data bases exist today to assist the structure determination as well as to study the structural dynamics of proteins and other biological macromolecules. They are Protein Data Bank(PDB) and Biological Magnetic Resonance Data Bank (BMRB).

4.1.1 Protein data bank

The Protein Data Bank (PDB)[Kouranov et al., 2006, Berman et al., 2000, Berman et al., 2003, Deshpande et al., 2005] is a repository for the 3-D structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography or NMR spectroscopy, submitted by biologists and biochemists from around the world, can be accessed at no charge on the internet. The PDB is overseen by an organization called the Worldwide Protein Data Bank (wwPDB).

4.1.1.1 Data entries

The number of structures deposited in PDB, determined by various experimental techniques are listed in Table4.1

In the total number of protein structures deposited in the data base, only 6864(12.82%)

Table 4.1: PDB contents as on June 2 2009

Experiment Methods	Proteins	Nucleic Acid	Protein & Nucleic Acid Complexes	Other	Total
X-ray	46383	1147	2141	17	49688
NMR	6864	856	146	6	7872
Electron Microscopy	168	16	59	0	243
Hybrid	13	1	1	1	16
Other	108	4	4	9	125
Total	53536	2024	2351	33	57944

structures were determined by NMR spectroscopy. Among this 6864 structures, 5154 structures are provided with NMR restraint information.

4.1.1.2 Usefulness of protein data bank

Protein data base is quite useful for homology modelling, based on sequential analysis. Approximate structures could be estimated, by matching segments of target protein, with

the protein sequences in the database.

The number of structures in the Protein Data Bank has grown to over 50,000. Many of the proteins in the PDB are homologous, i.e. have descended from a common ancestor, conserving significant aspects of their structure, function, and sequence. This would increase the statistical weight for a specific topology, which may result a biased statistics. For purposes such as a statistical analysis of protein structure features, a subset of the PDB is required in which structural features can be presumed to be independently distributed, i.e. unbiased with respect to evolutionary decent. Apart from that, the real NMR observable (chemical shift) information is missing in this database. This makes it difficult to make correlated statistics between experimental observation and the 3D structure.

4.1.2 Biological magnetic resonance data bank

Biological Magnetic Resonance Data Bank (BMRB)[Seavey et al., 1991, Ulrich et al., 2007] is the publicly-accessible depository for NMR results from peptides, proteins, and nucleic acids recognized by the International Society of Magnetic Resonance and by the IUPAC-IUBMB-IUPAB Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy. BMRB's mission is to collect, archive, and disseminate (worldwide in the public domain) the important quantitative data derived from NMR spectroscopic investigations of biological macromolecules.

4.1.2.1 Data entries

Chemical shift statistical information can be obtained directly from the BMRB web page (*http://www.bmrb.wisc.edu/ref_info/*). They have separate statistics for Proteins, DNA and RNA. For Proteins they filtered data set which contains only diamagnetic atoms.

Full set

The statistics presented in this table were calculated from the full BMRB database. This includes paramagnetic proteins, proteins with aromatic prosthetic groups, and entries where chemical shifts are reported relative to uncommon chemical shift references. The calcu-

lated statistics are derived from a total of 3145570 chemical shifts.

Diamagnetic only

BMRB Entries not included in the calculations for this table contained chemical shifts outside eight standard deviations from the mean calculated for the full BMRB database or a chemical shift for at least one carbon bound proton that was greater than 10ppm or was less than -2.5ppm. These criteria were used to eliminate from the calculations chemical shifts from paramagnetic proteins, from proteins with aromatic prosthetic groups, and from entries where unusual chemical shift referencing was used. Of the 3145570 possible chemical shifts in the BMRB database, 2483054 were included in calculating this table.

4.1.2.2 Usefulness of BMRB

The BMRB statistics are useful in the initial stages of resonance assignment. The distribution of chemical shift of a specific atom calculated from BMRB statistics, allow us to create a search window to locate the resonance peak in the spectra. Even though it is useful for resonance assignment, structure dependencies on chemical shifts can not be studied using this data base. Moreover the entries in the BMRB may have the following errors.

Referencing error: The chemical shift of an atom is not measured absolutely, instead it has to be measured with respect to a reference molecule like TMS. Direct referencing is straight forward, but when they go for indirect referencing chances are more to commit numerical errors. Sometimes people use non-standard references, which may totally give different shift values

Experimental error: Extracting chemical shift table from NMR spectrum is not so easy. Artefacts and noises are often interpreted as peaks. Recording a clean spectra without artifacts is hardly possible. These limitations leads to wrong chemical shift assignment and produce considerable error in the shift value.

Assignment error: Though the nomenclature of amino acids are well defined [Markley et al., 1998] most of the users do not pay much attention to it or the programs they use to handle the data written with wrong nomenclature. This leads to ambiguity in the stereo specific assignments. This can be seen for the atoms HE21,HE22 in glutamine

(GLN) and HD21,HD22 in asparagine (ASN). In these two atoms one of them must be down field shifted with respect to other, but the bmrb statistics , they both have approximately same mean.

4.1.3 Statistics

4.1.4 Chemical shift statistics

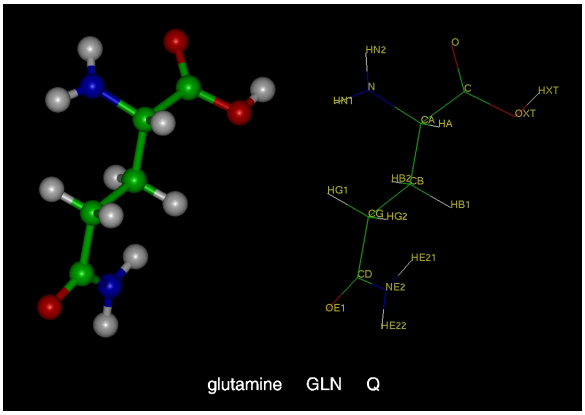


Figure 4.1: 3D structure of glutamine. Stereo specific atoms HE21 and HE22 are shown

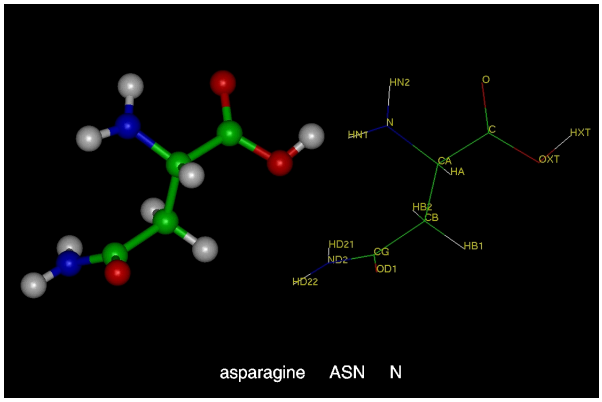


Figure 4.2: 3D structure of asparagine. Stereo specific atoms HD21 and HD22 are shown

Chemical shift distribution for every atom in the standard 20 amino acids can be obtained from BMRB data base. These statics can be directly obtained from BMRB webpage

in two categories, namely full set and restricted set. Paramagnetic atoms and non-standard referencing are excluded in the restricted data set.

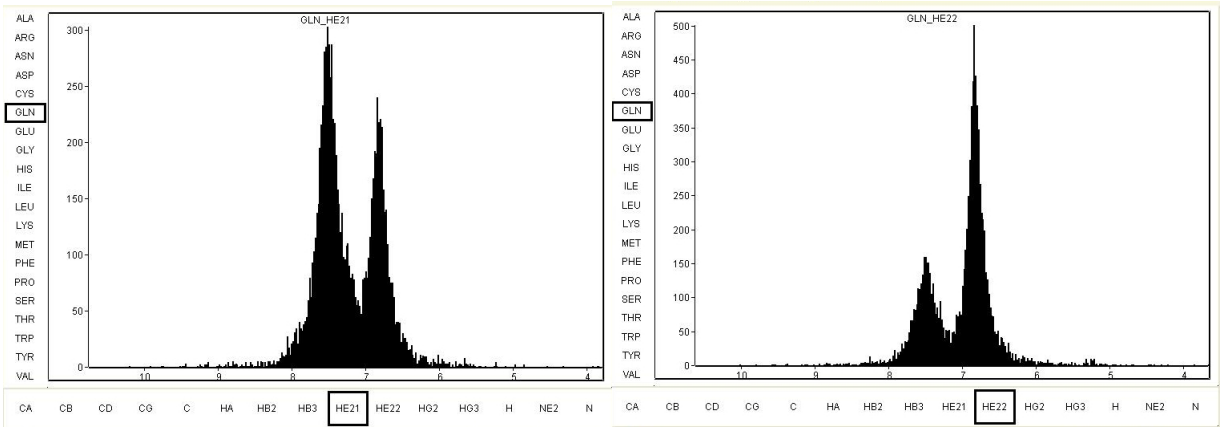


Figure 4.3: Chemical shift statistics of GLN HE21 (left), HE22 (right) from BMRB

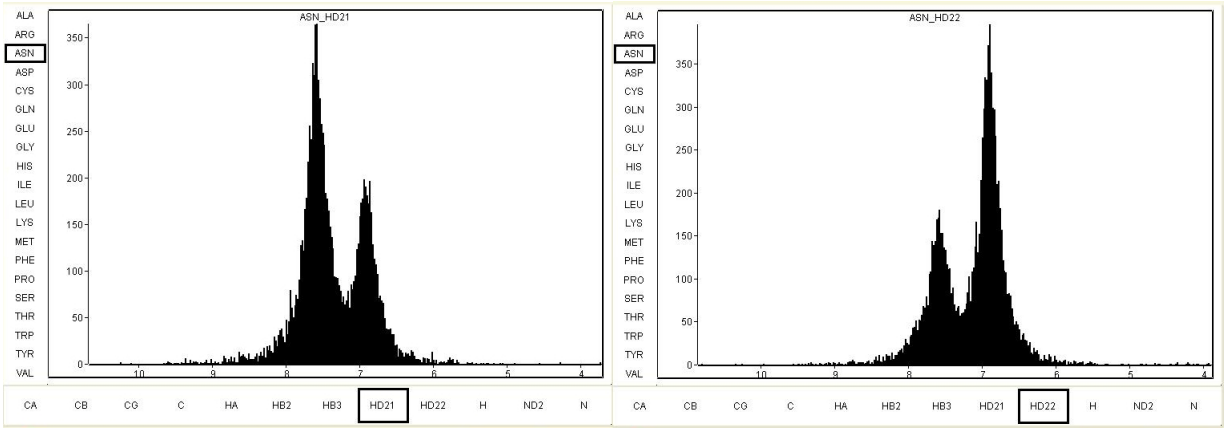


Figure 4.4: Chemical shift statistics of ASN HD21 (left), HD22 (right) from BMRB

But still experimental errors and assignment errors can not be eliminated completely. This can been seen in the case of stereo specific atoms HE21,HE22 in glutamine (GLU) and HD21,HD22 in asparagine (ASN). Figure 4.1, 4.2 these two stereo specific atoms are in slightly different environment, should give two distinct chemical shift values. Often people misunderstand the nomenclature and report wrongly to BMRB data base. This can be seen in Figure 4.3,4.4, where those atoms has two maximums instead of one. As a

result they both have almost the same mean. This might increase the ambiguity in stereo specific assignment and those stereo specific atoms are indistinguishable while resonance assignment.

Another subtlety in BMRB statistics is that it contains no structural information. The chemical shifts reported in the data base might come from proteins which are not in its native folded state. This may produce errors, when apply this statics for a folded protein.

4.1.5 Unbiased Structural Database

The statistical analysis of protein structures requires datasets in which structural features can be considered independently distributed, i.e. not related through common ancestry, and also have to fulfil minimal requirements regarding the experimental quality of the structures it contains. However, non-redundant datasets based on sequence similarity invariably contain distantly related homologies. Nh3D [Thiruv et al., 2005] is structural database, created mainly for the purpose of statical data analysis. It provides a reference dataset of non-homologous protein domains, assuming that structural dissimilarity at the topology level is incompatible with recognizable common ancestry. The dataset is based on domains at the topology level of the CATH [Orengo et al., 1997] database which hierarchically classifies all protein structures. It contains the best refined representatives of each topology level, validates structural dissimilarity and removes internally duplicated fragments. Nh3D database can be downloaded from *http://www.schematikon.org/Nh3D.php*

The CATH database is a hierarchical domain classification of protein structures in the Protein Data Bank [Kouranov et al., 2006]. Only crystal structures solved to resolution better than 4.0 A^0 are considered, together with NMR structures. All non-proteins, models, and structures with greater than 30% *C-alpha only* are excluded from CATH. This filtering of the PDB is performed using the SIFT protocol [Michie, 1996]. Protein structures are classified using a combination of automated and manual procedures. There are four major levels [Orengo et al., 1997] in this hierarchy:

- Class
- Architecture

- Topology (fold family)
- Homologous superfamily

Each level is described below, together with the methods used for defining domain boundaries and assigning structures to a specific family

4.1.6 CATH hierarchy and classification

All the classification is performed on individual protein domains. To divide multi domain protein structures into their constituent domains, a combination of automatic and manual techniques are used. If a given protein chain has sufficiently high sequence identity and structural similarity (ie. 80% sequence identity, SSAP (Single Amino Acid Polymorphism) score $>= 80$) with a chain that has previously been chopped, the domain boundary assignment is performed automatically by inheriting the boundaries from the other chain (ChopClose). Otherwise, the domain boundaries are assigned manually, based on an analysis of results derived from a range of algorithms which include structure based methods CATHEDRAL, SSAP, DETECTIVE [Swindells, 1995], PUU [Holm and Sander, 1994], DOMAK [Galzitskaya and Melnik, 2003] and sequence based methods.

4.1.6.1 Class

Class is determined according to the secondary structure composition and packing within the structure. Three major classes are recognized; mainly-alpha, mainly-beta and alpha-beta. This last class (alpha-beta) includes both alternating alpha/beta structures and alpha+beta structures, as originally defined by [Levitt and Chothia, 1976]. A fourth class is also identified which contains protein domains which have low secondary structure content.

4.1.6.2 Architecture

This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. It is currently assigned manually using a simple description of the secondary structure

arrangement e.g. barrel or 3-layer sandwich. Reference is made to the literature for well-known architectures (e.g the beta-propeller or alpha four helix bundle).

4.1.6.3 Topology

Structures are grouped according to whether they share the same topology or fold in the core of the domain, that is, if they share the same overall shape and connectivity of the secondary structures in the domain core. Domains in the same fold group may have different structural decorations to the common core. Some fold groups are very highly populated [Orengo et al., 1994, Orengo and Thornton, 2005] particularly within the mainly-beta 2-layer sandwich architectures and the alpha-beta 3-layer sandwich architectures.

4.1.6.4 Homologous superfamily

This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are identified either by high sequence identity or structure comparison using SSAP. Structures are clustered into the same homologous superfamily if they satisfy one of the following criteria:

- Sequence identity $\geq$ 35%, overlap$\geq$ 60% of larger structure equivalent to smaller.
- SSAP score $\geq$ 80.0, sequence identity $\geq$ 20%, 60% of larger structure equivalent to smaller.
- SSAP score $\geq$ 70.0, 60% of larger structure equivalent to smaller, and domains which have related functions, which is informed by the literature and Pfam protein family database,
- Significant similarity from HMM-sequence searches and HMM-HMM comparisons using SAM [Sjölander et al., 1996], HMMER (http://hmmer.wustl.edu) and PRC (http://supfam.org/PRC).

4.1.7 Nh3D

Nh3D is a reference dataset [Thiruv et al., 2005] of structures of non-homologous proteins. It contains a dataset of structurally dissimilar proteins. This dataset has been compiled by selecting well resolved representatives from the Topology level of the CATH database. These have been been pruned to remove

- domains that may contain homologous elements (by pairwise sequence comparison and structural superposition of aligned residues)
- internal duplications (by repeat detection)
- regions with high B-Factor (average B-Factor greater than 60A) .

The current Nh3D list contains 570 domains with a total of 90780 residues. It covers more than 70% of folds at the topology level of the CATH database and represents more than 90% of the structures in the PDB that have been classified by CATH. It is observed that even though all protein pairs are structurally dissimilar, some pairwise sequence identities after global alignment are greater than 30%. The current version Nh3D-v3.0 contains 806 structures. Those structures are classified into 4 major classes

1. Mainly Alpha
2. Mainly Beta
3. Alpha Beta
4. Few Secondary Structures.

In Nh3D structures are named according the numbers mentioned above. Structures starting with 1 are mainly alpha, starting with 2 are mainly beta, starting with are alpha and beta and starting with 4 are less structured.

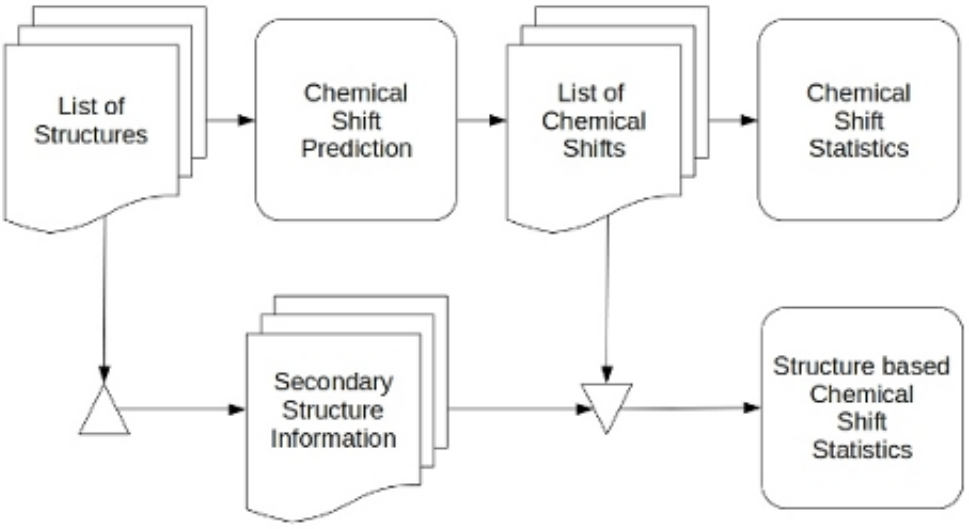


Figure 4.5: Flowchart for creating refined chemical shift statistics

4.2 Materials and methods

4.2.1 Non homologous chemical shift statistics

The general idea to create an unbiased chemical shift statistics is explained in Figure 4.5. The predicted chemical shifts of structures from the non-homologous structural database(Nh3D) can be used to create a non-homologous chemical shift statistics. In addition to the chemical shift information, structural information can also be extracted from the data base in order to create a more sophisticated correlated statistics. The procedure to create the distribution involved various steps.

4.2.1.1 Unbiased chemical shift database

The chemical shift prediction program SHIFTS [Xu and Case, 2001] and SHIFTX [Neal et al., 2003] were analysed in chapter three. First step to create chemical shift statistics is to create complete list of chemical shifts for each structure in Nh3D data base which are structurally unbiased. SHIFTS has no option to predict all the chemical shifts in single

run. There are separate options to predict ^{1}H chemical shifts and ^{13}C,^{15}N. Those two outputs were merged to get a complete list. SHIFTX can produce complete list in single run. They both produce output in different file formats. The outputs are formatted to standard ASCII formate using python scripts. We can use a single prediction program, or both to create a chemical shift database. It is better to use both prediction methods, since we take advantage from both prediction algorithm.

4.2.1.2 Grouping of chemical shifts

Once the unbiased chemical shift data base is created, chemical shifts are grouped by atom types. For every atom $a : \{H, HA, HB, C, CA, CB, ..\}$ in amino acid $A : \{ALA, GLY, ..\}$ a chemical shift list $\delta_{A,a} : \left\{\delta^1_{A,a}, \delta^2_{A,a}, \delta^3_{A,a}, .., \delta^{n_{A,a}}_{A,a}\right\}$ is created. $n_{A,a}$ is the maximum number of chemical shifts of atom a from amino acid A found in the database.

Chemical shifts can be further grouped with respect to secondary structure elements $S : \{H, E, C\}$ as shown in the Figure 4.5, where H,E and C stands for Helix, Sheet and Coil respectively. This group of chemical shifts can be represented as $\delta_{A,a,S} : \left\{\delta^1_{A,a,S}, \delta^2_{A,a,S}, \delta^3_{A,a,S}, .., \delta^{n_{A,a,S}}_{A,a,S}\right\}$, where $n_{A,a,S}$ is the maximum number of chemical shifts of atom a from amino acid A found in a particular secondary structure S

4.2.2 Probability density function

Probability Density Function (PDF) is a distribution function of any random variable in a given data set. Here it means chemical shift probability density function of a given atom (or) atom type in the unbiased chemical shift database. There are two possible ways to construct PDFs based on the following tow assumptions,

1. chemical shifts distributed around single value (single maximum)

2. chemical shifts distributed around more than one value (multiple maximum)

As a simple case, single maximum distribution is treated with Gaussian model and for multiple maximum distribution case, Kernel Density Estimation(KDE) is used.

4.2.2.1 Gaussian model

In general the distribution of a random variable such as chemical shifts could be assumed to follow a Gaussian distribution, characterized by a mean and a standard deviation. For a set to chemical shift values $\delta_{A,a} : \left\{\delta^1_{A,a}, \delta^2_{A,a}, \delta^3_{A,a}, .., \delta^{n_{A,a}}_{A,a}\right\}$ the mean $\overline{\delta_{A,a}}$ and $\sigma_{A,a}$ is given by the following relation,

$$\overline{\delta}_{A,a} = \frac{1}{n_{A,a}} \sum_{i=1}^{n_{A,a}} \delta^i_{A,a} \tag{4.1}$$

$$\sigma_{A,a} = \sqrt{\frac{1}{n_{A,a}} \sum_{i=1}^{n_{A,a}} (\delta^i_{A,a} - \overline{\delta}_{A,a})^2} \tag{4.2}$$

Probability density function(PDF) for Gaussian model Φ_G can be constructed using the mean $\overline{\delta}_{A,a}$ and sigma $\sigma_{A,a}$ from the data base.

$$\Phi_G(\delta) = \frac{1}{\sigma_{A,a}\sqrt{2\pi}} e^{\frac{(\delta - \overline{\delta}_{A,a})^2}{2\sigma^2_{A,a}}} \tag{4.3}$$

For a given chemical shift δ, the probability $P^G(\delta|A,a)$ that it belongs to atom a from amino acid A can be calculated by integrating the PDF between the limits $\delta + \frac{\Delta\delta}{2}$ and $\delta - \frac{\Delta\delta}{2}$

$$P^G(\delta|A,a) = \frac{1}{\sigma_{A,a}\sqrt{2\pi}} \int_{\delta - \frac{\Delta\delta}{2}}^{\delta + \frac{\Delta\delta}{2}} e^{\frac{(\delta - \overline{\delta}_{A,a})^2}{2\sigma^2_{A,a}}} .d\delta \tag{4.4}$$

4.2.2.2 Multiple-Gaussian model

If the distribution showing non Gaussian nature, it is better to use Kernel Density estimation [Brodsky and Darkhovsky, 2000] to calculated PDF. In this case it is also called Kernel Density Function. For a given set of chemical shifts $\delta_{A,a} : \left\{\delta^1_{A,a}, \delta^2_{A,a}, \delta^3_{A,a}, .., \delta^{n_{A,a}}_{A,a}\right\}$, the kernel density function $\Phi_K$is given as,

$$\Phi_K(\delta, h)(\delta_{A,a}) = \frac{1}{n_{A,a}h} \sum_{i=1}^{n_{A,a}} G\left(\frac{\overline{\delta}_{A,a} - \delta^i}{h}\right) \tag{4.5}$$

where G is the kernel usually a standard Gaussian function with mean zero and variance 1 and h is the bandwidth

$$G\left(\frac{\overline{\delta}_{A,a} - \delta^i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\overline{\delta}_{A,a} - \delta^i)^2}{2h^2}} \tag{4.6}$$

The value of h determines the smoothness of function. If h is too small than the resulting function is under smoothed, and if h is too large than the resulting function is over smoothed. There are several ways to optimize the bandwidth. A common way to calculate the bandwidth is usng Asymptotic Mean Integrated Squared Error (AMISE).[Turlach, 1993] For a given chemical shift δ, the probability $P^K(\delta|A,a)$ that it belongs to atom a from amino acid A can be calculated by integrating the kernel density function between the limits $\delta + \frac{\Delta\delta}{2}$ and $\delta - \frac{\Delta\delta}{2}$

$$P^K(\delta|A,aa) = \int_{\delta-\frac{\Delta\delta}{2}}^{\delta+\frac{\Delta\delta}{2}} \Phi_K(\delta_{A,a}).d\delta \tag{4.7}$$

The value of $\Delta\delta$ is calculated from over all σ of specific atom type.

$$\Delta\delta(a) = \frac{\sigma_a}{C_1} \tag{4.8}$$

where, C_1 is a constant which decides the integration with $\Delta\delta$. The width of $\Delta\delta$ should be infinitely small to estimate the probability of given chemical shift. But due to computational limitations, it can not be kept very low. $C_1 = 50$ will be a moderate value, which could be used for all type of atoms.

4.2.3 Assignment probabilities

The assignment probability P that a given chemical shift δ can be assigned to a specific atom a can be calculated using equations 4.4 and 4.7 with different probability density functions created out of different level of information. One can directly take the mean and standard deviation from BMRB statistics and calculate the assignment probability P_{bmrb} using equation 4.4, In addition to that, one can use Gaussian model and Kernel density estimation method from the unbiased chemical shift database, to calculate the assignment probability P_{gauss} and P_{kde} respectively.

These assignment probabilities are useful to remove some ambiguity during resonance assignment. For a known assigned chemical shift list, probability for *correct assignment* can be calculated using three different method as mentioned in the previous paragraph. Let us call those probabilities as P^+_{bmrb},P^+_{gauss} and P^+_{kde}. The relative improvement (only for

correct assignment) in the probability of P^{+}_{gauss} and P^{+}_{kde} when compared to P^{+}_{bmrb} is given by,

$$G_{imp} = \frac{\sum_{i=1}^{N} P^{+}_{gauss}(A_i, a_i, \delta_i) - \sum_{i=1}^{N} P^{+}_{bmrb}(A_i, a_i, \delta_i)}{\sum_{i=1}^{N} P^{+}_{bmrb}(A_i, a_i, \delta_i)} \times 100 \tag{4.9}$$

$$K_{imp} = \frac{\sum_{i=1}^{N} P^{+}_{kde}(A_i, a_i, \delta_i) - \sum_{i=1}^{N} P^{+}_{bmrb}(A_i, a_i, \delta_i)}{\sum_{i=1}^{N} P^{+}_{bmrb}(A_i, a_i, \delta_i)} \times 100 \tag{4.10}$$

where G_{imp} and K_{imp} are the improvements in "%" produced by Gaussian and KDE models respectively for a given protein with N atoms. The sign of G_{imp} and K_{imp} tells us which method is better. If these quantities are positive then we can conclude that unbiased chemical shift statistics is better than BMRB statistics.

4.2.4 Resonance assignment

The procedure to assign the experimentally observed chemical shift values to its corresponding atom is called resonance assignment. This is the first step in protein structure determination. The improvement in these a priory probabilities P^{+}_{gauss} and P^{+}_{kde} are useful in resonance assignment, only when the probability values uniquely improved for the correct assignment. One of the main problem in resonance assignment is the degeneracy in chemical shift values. Many atoms may have almost same chemical shift values. The success rate in the assignment process depends on the fact how well these chemical shifts are distinguished by the use of statistical information in our hand. Many observerd chemical shifts will be inside the chemical shift distribution of single atom if the large standard deviation of chemical shift distribution is too large. This will increase the ambiguity in many cases and leaving most of the chemical shifts wrongly assigned. The multiple Gaussian model from refined statistics may reduce the ambiguity by giving low probability values for some chemical shifts depending on the nature(shape) of the distribution. One has to check that P^{+}_{gauss} and P^{+}_{kde} are improving only for correct assignments. This can be tested by randomizing a known assigned list and re doing resonance assignment using standard assignment procedure like Hungarian method[Schmid, 1978, Wright, 1990, Jonker and Volgenant, 1986] with the probabilities P_{bmrb}, P_{gauss} and P_{kde}.

4.2.4.1 Hungarian algorithm

The resonance assignment is complex process in NMR spectroscopy. The presence of noise and artefacts makes it harder to identify real signal and label them with the corresponding atoms. In ideal case where there is no artefacts and no missing peaks, the problem reduced to a *combinatorial optimization* problem with equal number of chemical shifts and atoms. One of the well known methods to solve a combinatorial problem is the Hungarian algorithm [Schmid, 1978, Wright, 1990, Jonker and Volgenant, 1986].

Let us consider a list of chemical shifts $\delta \equiv \{\delta_1, \delta_2, \delta_3 .... \delta_N\}$ and list of atoms $A \equiv \{a_1, a_2, a_3, ... a_N\}$. Assigning a chemical shift δ_i to an atom a_j will cost some energy denoted by $\xi_{i,j}$. The cost matrix C is given by,

$$\xi_{i,j} = \xi\left(\delta_i, a_j\right), \delta_i \in \delta, a_j \in A \tag{4.11}$$

The objective of the assignment problem is to find a particular mapping between chemical shift δ_i and atom A_j

$$\delta_i \mapsto \prod(a_j), 1 \leq i \leq N, 1 \leq j \leq N \tag{4.12}$$

such that the total assignment energy Ξ for a set of such matching,

$$\Xi = \sum_{i}^{N} \xi_{\delta_i, \prod(a_i)} \tag{4.13}$$

is minimized over all permutations of all atoms $\prod$. This is a one to one mapping, hence no two chemical shift indices cannot be assigned to same atom or no two atom indices can not be assigned to same chemical shift. If one wants to do it in a straight forward way, he has to calculate all possible assignments in order to chose the one which has lowest energy. The cost matrix C consists of all possible assignments is given by,

$$C = \begin{bmatrix} \xi_{1,1} & \xi_{1,2} & \xi_{1,3} & \cdots & \xi_{1,N} \\ \xi_{2,1} & \xi_{2,2} & \xi_{2,3} & \cdots & \xi_{2,N} \\ \xi_{3,1} & \xi_{3,2} & \xi_{3,3} & \cdots & \xi_{3,N} \\ \cdots & \cdots & \xi_{i,j} & \cdots & \cdots \\ \xi_{N,1} & \xi_{N,2} & \xi_{N,3} & \cdots & \xi_{N,N} \end{bmatrix} \tag{4.14}$$

A brute-force algorithm for solving the assignment problem involves generating all independent sets of the matrix C, computing the total costs of each assignment and a search of all assignment to find a minimal-sum independent set. The complexity of this method is driven by the number of independent assignments possible in an $N \times N$ matrix. There are N choices for the first assignment, $N-1$ choices for the second assignment and so on, giving $N!$ possible assignment sets. Therefore, this approach has, at least, an exponential runtime complexity.

Hungarian algorithm basically tries to minimize the total energy with a set of optimum assignments. It can solve the assignment problem in polynomial time. This is done in the following steps.

1. For each row of the matrix, find the smallest element and subtract it from every element in its row. Go to Step 2.

2. Find a zero (Z) in the resulting matrix. If there is no starred zero in its row or column, star Z. Repeat for each element in the matrix. Go to Step 3.

3. Cover each column containing a starred zero. If N columns are covered, the starred zeros describe a complete set of unique assignments. In this case, Go to DONE, otherwise, Go to Step 4.

4. Find a non-covered zero and prime it. If there is no starred zero in the row containing this primed zero, Go to Step 5. Otherwise, cover this row and uncover the column containing the starred zero. Continue in this manner until there are no uncovered zeros left. Save the smallest uncovered value and Go to Step 6.

5. Construct a series of alternating primed and starred zeros as follows. Let Z0 represent the uncovered primed zero found in Step 4. Let Z1 denote the starred zero in the column of Z0 (if any). Let Z2 denote the primed zero in the row of Z1 (there will always be one). Continue until the series terminates at a primed zero that has no starred zero in its column. Unstar each starred zero of the series, star each primed zero of the series, erase all primes and uncover every line in the matrix. Return to Step 3.

6. Add the value found in Step 4 to every element of each covered row, and subtract it from every element of each uncovered column. Return to Step 4 without altering any stars, primes, or covered lines.

DONE: Assignment pairs are indicated by the positions of the starred zeros in the cost matrix. If $C(i,j)$ is a starred zero, then the element associated with chemical shift δ_i is assigned to the element associated with atom a_j.

Some of these descriptions require careful interpretation. In Step 4, for example, the possible situations are, that there is a non-covered zero which get primed and if there is no starred zero in its row the program goes onto Step 5. The other possible way out of Step 4 is that there are no non-covered zeros at all, in which case the program goes to Step 6. Though the algorithm looks complicated, now a days it is readily available in many software tools like SCILAB and Python Numerics. Here python scripts are used to calculate the optimum assignments from the cost matrix C.

The important question here is *how to define the energy term?*. It is possible to use directly the probability value with the negative sign as a pseudo energy term. But in a better way $\xi_{i,j}$ can be defined as follows,

$$\xi_{i,j} = -log\left(\frac{P^m(\delta_i, a_j)}{P^z(\delta_i, a)}\right) \tag{4.15}$$

where, $P^m(\delta_i, a_j)$ is the probability calculated using a specific statistical model(BMRB / SHIFTS / SHIFTX / Both) and $P^z(\delta_i, a)$ is the probability calculated from zero model created using kernel density estimation. The reason for introducing zero model is avoid the cases, in which no statistics will work. Zero model probabilities are derived from statistics created by using minimal information or no (zero) information. For example one can think of a zero model, which includes all ^1H chemical shifts (independent of whether it comes from specific atom group or specific amino acid). This zero model serves as a reference in order to handle the cases where no statistics will work, which means if our statistical models giving same probabilities as compared to zero model statistics, then the atom has random distribution of chemical shift. The over all distribution (independent of amino acid) of ^1H,^{13}C and ^{15}N can also be taken as zero model. It will give a low reference probability value for any atom. The value of $\xi_{i,j}$ can be positive or negative or

even zero depending on the following condition.

- $\xi_{i,j}$ is negative *if* $P^m(\delta_i, a_j) > P^z(\delta_i, a_j)$, which means our model is better
- $\xi_{i,j}$ is positive *if* $P^m(\delta_i, a_j) < P^z(\delta_i, a_j)$, which means zero model is better
- $\xi_{i,j}$ is zero *if* $P^m(\delta_i, a_j) = P^z(\delta_i, a_j)$, which means both models are same

Its not a good idea to construct a zero model including all type of atoms, since ^1H,^{13}C and ^{15}N chemical shifts are found in different range. Hence it is better to have three separate zero model for the three different type of atoms. For a given protein a pseudo energy matrix can be constructed for specific atom type in order to find optimum assignment. For example if we have a list which contains only ^1H chemical shifts and the sequence of the protein is known, one could construct the matrix for ^1H as given below

$$\begin{bmatrix} \xi_{1,1}(^1H) & \xi_{1,2}(^1H) & \xi_{1,3}(^1H) & ... & \xi_{1,N}(^1H) \\ \xi_{2,1}(^1H) & \xi_{2,2}(^1H) & \xi_{2,3}(^1H) & ... & \xi_{2,N}(^1H) \\ \xi_{3,1}(^1H) & \xi_{3,2}(^1H) & \xi_{3,3}(^1H) & ... & \xi_{3,N}(^1H) \\ ... & ... & \xi_{i,j}(^1H) & ... & ... \\ \xi_{N,1}(^1H) & \xi_{N,2}(^1H) & \xi_{N,3}(^1H) & ... & \xi_{N,N}(^1H) \end{bmatrix} \tag{4.16}$$

$$\xi_{i,j}(^1H) = -log\left(\frac{P^m(\delta_i, a_j)}{P^z(\delta_i, ^1H)}\right) \tag{4.17}$$

similarly matrices can be constructed for ^{13}C and ^{15}N also

$$\begin{bmatrix} \xi_{1,1}(^{13}C) & \xi_{1,2}(^{13}C) & \xi_{1,3}(^{13}C) & ... & \xi_{1,N}(^{13}C) \\ \xi_{2,1}(^{13}C) & \xi_{2,2}(^{13}C) & \xi_{2,3}(^{13}C) & ... & \xi_{2,N}(^{13}C) \\ \xi_{3,1}(^{13}C) & \xi_{3,2}(^{13}C) & \xi_{3,3}(^{13}C) & ... & \xi_{3,N}(^{13}C) \\ ... & ... & \xi_{i,j}(^{13}C) & ... & ... \\ \xi_{N,1}(^{13}C) & \xi_{N,2}(^{13}C) & \xi_{N,3}(^{13}C) & ... & \xi_{N,N}(^{13}C) \end{bmatrix} \tag{4.18}$$

$$\xi_{i,j}(^{13}C) = -log\left(\frac{P^m(\delta_i, a_j)}{P^z(\delta_i, ^{13}C)}\right) \tag{4.19}$$

$$\begin{bmatrix} \xi_{1,1}(^{15}N) & \xi_{1,2}(^{15}N) & \xi_{1,3}(^{15}N) & ... & \xi_{1,N}(^{15}N) \\ \xi_{2,1}(^{15}N) & \xi_{2,2}(^{15}N) & \xi_{2,3}(^{15}N) & ... & \xi_{2,N}(^{15}N) \\ \xi_{3,1}(^{15}N) & \xi_{3,2}(^{15}N) & \xi_{3,3}(^{15}N) & ... & \xi_{3,N}(^{15}N) \\ ... & ... & \xi_{i,j}(^{15}N) & ... & ... \\ \xi_{N,1}(^{15}N) & \xi_{N,2}(^{15}N) & \xi_{N,3}(^{15}N) & ... & \xi_{N,N}(^{15}N) \end{bmatrix} \tag{4.20}$$

$$\xi_{i,j}(^{15}N) = -log\left(\frac{P^m(\delta_i, a_j)}{P^z(\delta_i, ^{15}N)}\right) \tag{4.21}$$

The best possible assignments can be found on each matrix by applying hungarian algorithm.

4.2.4.2 Residue assignment

The invention of multidimensional NMR spectroscopy has improved the resolution, by spreading the peaks in more than one dimensions. The same techniques could be used for assignment process also. Instead of creating one big matrix for Hungarian algorithm, one can create matrices for specific atom types (matrices for HA , CA and N). Triple resonance experiments like HNCA will give the connectivity between those atoms which belong the same reside. Chemical shifts are arranged using this information and the same order is used in all the matrices. If the protein has n residue, we will get 3 $n \times n$ matrices. All these matrices are weighted averaged to create a new matrix. Now Hungarian algorithm is applied to this final matrix. Each energy term in the final matrix is created from the different atom from the same residue. Since all three matrices chemical shifts and atoms are arranged in the same order, once the assignment is done, same assignment could be copied to all three matrices. By doing this we can assign all the three atom types (HA,CA and N) simultaneously.

The advantage we have here is similar to the advantage we have in multidimensional NMR spectra. Suppose a protein has two residues which may have similar chemical shift distribution functions for HA, but dissimilar chemical shift distribution functions for CA. By taking weighted average [Schumann et al., 2007] of the pseudo energy matrices, we take the advantage of CA probability to reduce HA ambiguity for a given residue. The pseudo energy of each residue in the weighted average matrix is given by

$$\xi_{i,j}(A) = \frac{1}{\sum_{k=1}^{m} w_k} \sum_{k=1}^{m} w_k \xi_{i,j}(aa_k) \tag{4.22}$$

where, A is the residue, aa is the atom in the residue,w_k is the atom specific weighting factor [Schumann et al., 2007] and m is the maximum number of atoms used for the assignment process.

4.3 Test cases and results

The main use of this statistical method is to assist the automation of structure determination. At various levels these statistical informations are used to estimate certain probabilities and to also to optimize the search ranges at the very first step. The difficult part here is that it is hardly possible to identify the atom unambiguously from a given chemical shift value δ. But still it is possible to estimate the probability that the given chemical shift should come from a specific atom. This is where these statistics plays important role.

4.3.1 Probability test

The assignment probability is defined as the probability for a given chemical shift δ, should come from a specific atom a in amino acid A. For a given atom a from amino acid A with given chemical shift δ this probability can be calculated using three different statistics, namely

1. BMRB statistics
2. refined statistics (Gaussian model)
3. refined statistics (KDE)

using equations 4.4,4.7. A Gaussian PDF is constructed from BMRB statistics similar to equation 4.4 using the mean and standard deviation obtained from BMRB website. Three types of predicted data base used for refined statistics. They are based on SHIFTS, SHIFTX and the combined data set of both predictions method.

4.3.1.1 Test data set

The predicted shifts can not be used to test these statistics, because it will over fit the data and give very good results. The best test data set would be the real experimental chemical shifts.When a statistics tested on a same data set in which it is created, will give better test results. This is known as over fitting. One can directly take BMRB entries as a test data set, even though that will over fit with BMRB statistics. The aim of this test is to show improvements in probability values for real data set using refined statistics. As discussed in chapter three, the BMRB entries may contain experimental errors or might have come from an experiment with unusual physiological conditions. In such cases, any statistics would give worst probability values. What we are interested here is the relative improvement in the performance when compared to BMBR statistics. The question here is *'If BMRB statistics could give a definite probability for a correct assignment, then how much it is improved when refined statistics is used?'*. For every entry in the BMRB the relative improvement in the probability values were calculated using equation 4.9 and 4.10 and averaged over all entries.

Randomly selected 1000 BMRB entries were used for assignment test. The chemical shift list of each entry is randomized and re assigned using Hungarian method with pseudo energy terms calculated over probabilities from BMRB and refined statistics. The output of the Hungarian method were compared with the input assignments and the correctness is calculated in percentage.

4.3.1.2 Results

Figure 4.6 shows the relative improvement in over all probability values, when refined statics is used. These mean percentages are calculated from 2820 BMRB entries with varying number of residues. Figure 4.7 and 4.8 are the similar plots, but only back bone and side chin atoms are treated separately. The multiple Gaussian statistics (KDE) gives much better probability value, when compared to BMRB and single Gaussian models. It increases the a priory probability by 20%, which is very useful to reduce the ambiguity in

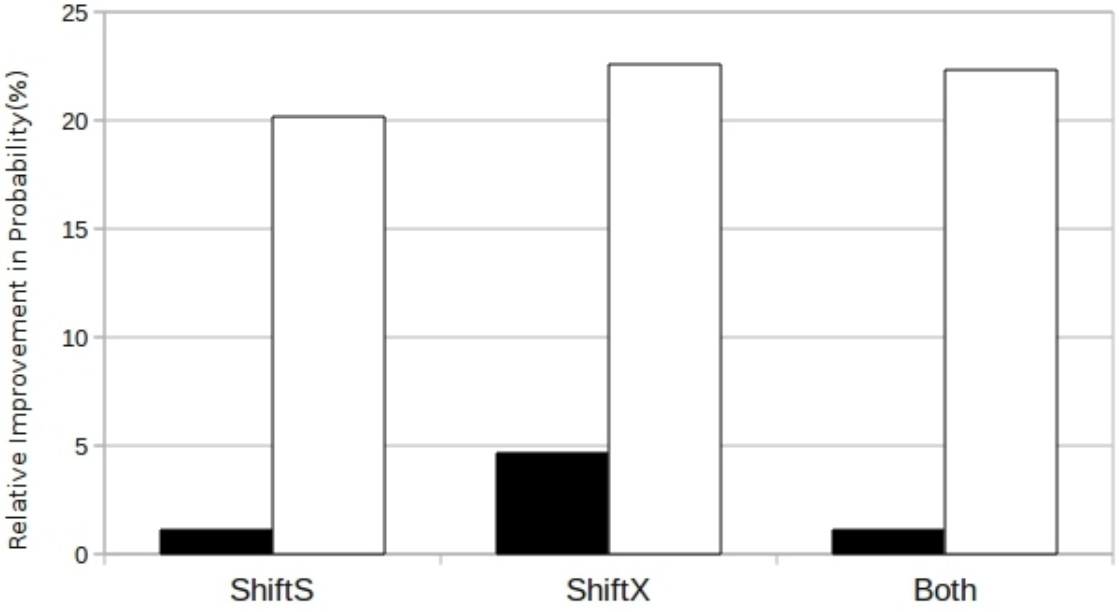


Figure 4.6: Relative improvement in probability of all atoms averaged over 2820 BMRB entries. Black bars indicates the mean of G_{imp} calculated using equation 4.9, and white bars indicates the mean of K_{imp} using equation 4.10

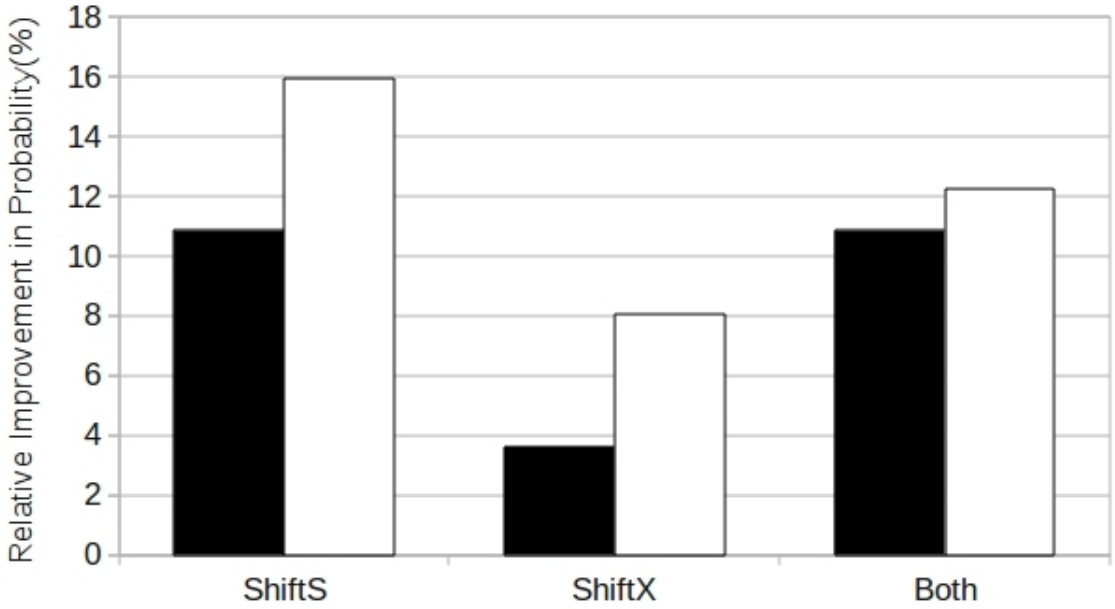


Figure 4.7: Relative improvement in probability of back bone atoms averaged over 2820 BMRB entries. Black bars indicates the mean of G_{imp} calculated using equation 4.9, and white bars indicates the mean of K_{imp} using equation 4.10

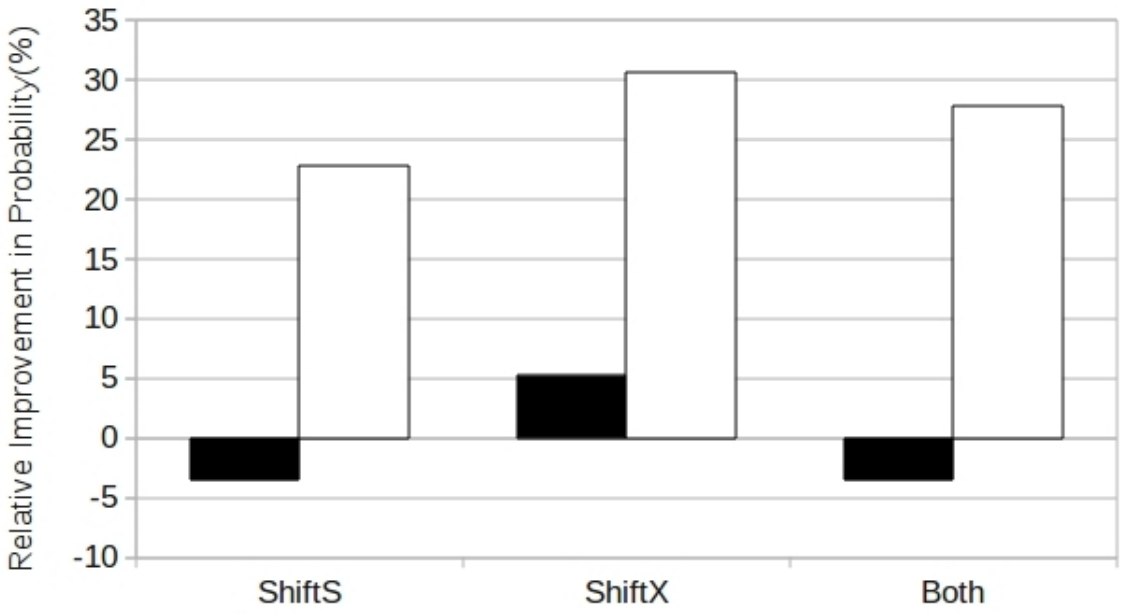


Figure 4.8: Relative improvement in probability of side chain atoms calculated over 2820. Black bars indicates the mean of G_{imp} calculated using equation 4.9, and white bars indicates the mean of K_{imp} using equation 4.10

the very first stage of the assignment process. The side chain showed negative values for Gaussian models in two cases (Figure 4.8). This may be due to the strong overlap in side chain resonances and the chemical shift degeneracy of the side chain chemical shifts.

Figure 4.9 shows the results of resonance assignment of randomly selected 1000 BMRB entries using Hungarian algorithm. Resonance assignment using only proton chemical shift is very difficult, since it has many degenerate chemical shift values. However ^{13}C resonance assignment works better than protons. The unique distribution of ^{13}C chemical shifts are really helpful in the assignment process.

Figure 4.10 shows the results obtained using weighted average matrix using HA, CA and CB atoms. By using the weighted average of chemical shift pseudo energies, the ambiguity is reduced in a considerable amount, and we get more than 60% correct assignment. This method has no limitation in the number of matrices used for averaging. The weighted average method shows 10% improvement when compared to correct assignments using BMRB statistics.

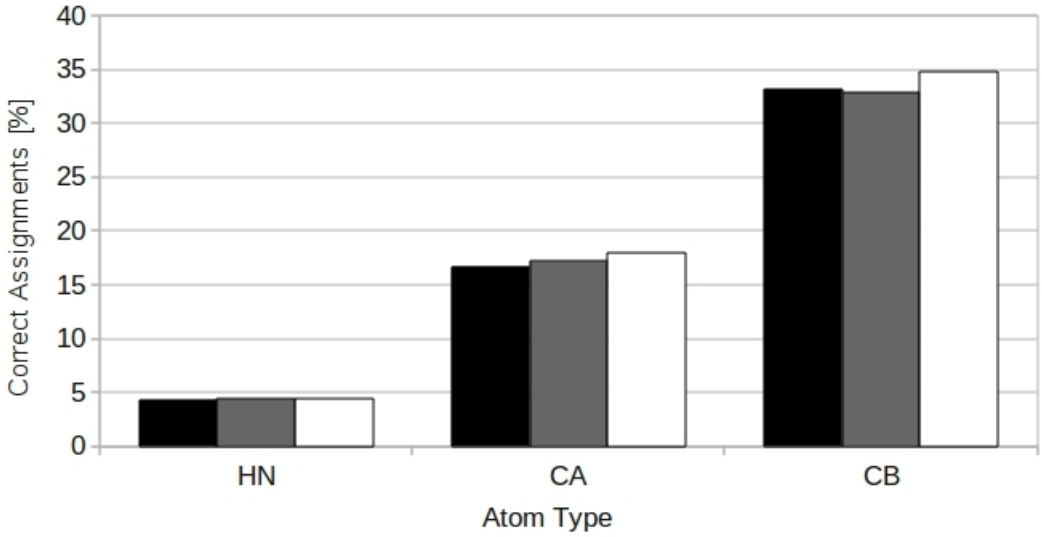


Figure 4.9: Average percentage of correct resonance assignment of randomly selected 1000 BMRB entries using different atom types. Black bar indicates BMRB statistics, gray bar indicates Gaussian model from refined statistics and white bar indicates KDE using refined statistics.The pseudo energies for assignment are calculated using equations 4.17 and 4.19.

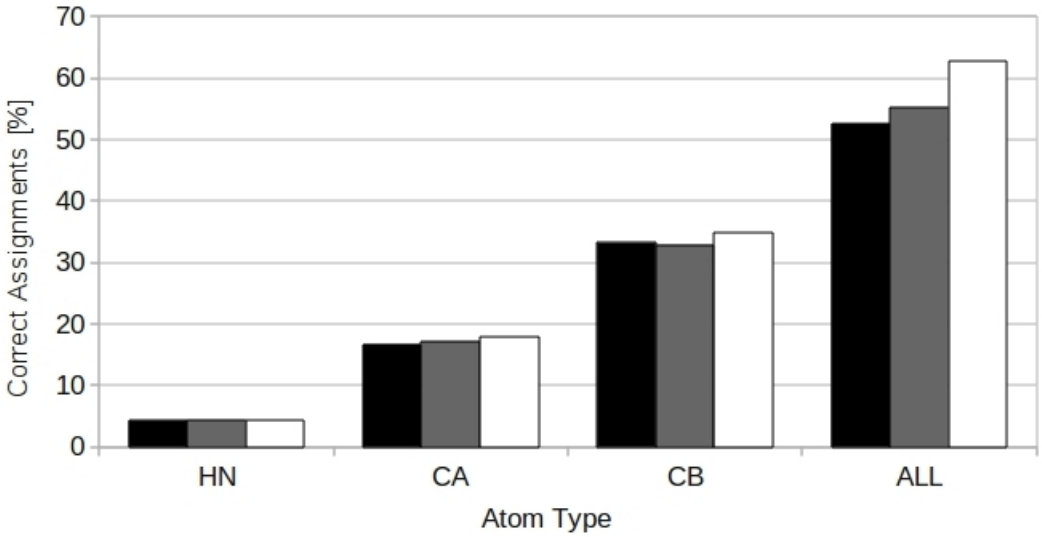


Figure 4.10: Average percentage of correct residue assignment of randomly selected 1000 BMRB entries using different atom types and using weighted pseudo energies. Black bar indicates BMRB statistics, gray bar indicates Gaussian model from refined statistics and white bar indicates KDE using refined statistics. The pseudo energies for assignment are calculated using equations 4.17,4.19 and 4.22.

4.4 Discussion

4.4.1 Advantages and limitations

4.4.1.1 Advantages

The various problems in the BMRB statistics discussed in chapter 2 has been partially or fully eliminated in the new refined non homologous chemical shift statistics. For example, structural independence is guarantied by Nh3D, hence the same is true for chemical shift data base.

Table 4.2: Comparison of $\overline{\delta}$ and σ_δ for stereo specific assignments.

Stereo specific atom	BMRB $\overline{\delta}(\sigma_\delta)$	SHIFTS $\overline{\delta}(\sigma_\delta)$	SHIFTX $\overline{\delta}(\sigma_\delta)$
GLN-HE21	7.23(0.47)	7.64(0.28)	7.35(0.24)
GLN-HE22	7.02(0.46)	6.90(0.24)	6.75(0.29)
ASN-HD21	7.34(0.48)	7.69(0.32)	7.53(0.32)
ASN-HD22	7.14(0.5)	6.96(0.30)	6.88(0.34)

The misunderstanding in the nomenclature of atoms has been partially removed in the refined chemical shift statistics. Even though some of the atom names predicted using SHIFTS and SHIFTX are not matching with the BMRB IUPAC format, they can be corrected(based on our experimental knowledge) using a routine. Unlike BMRB, the error in the naming occurred in the same way in all the predicted chemical shifts. The standard deviation calculated from BMRB, is almost twice as big as when compared to SHIFTS and SHIFTX (Table 4.2). Hence the ambiguity caused by the BMRB statistics while assigning the stereo specific atoms has been reduced by the use of refined statistics. The distribution of stereo specific atoms like GLN-HE21,HE22 and ASN-HD21,HD22 shows clear distinction in refined chemical shift statistics. Table 4.2 shows that SHIFTS and SHIFTX

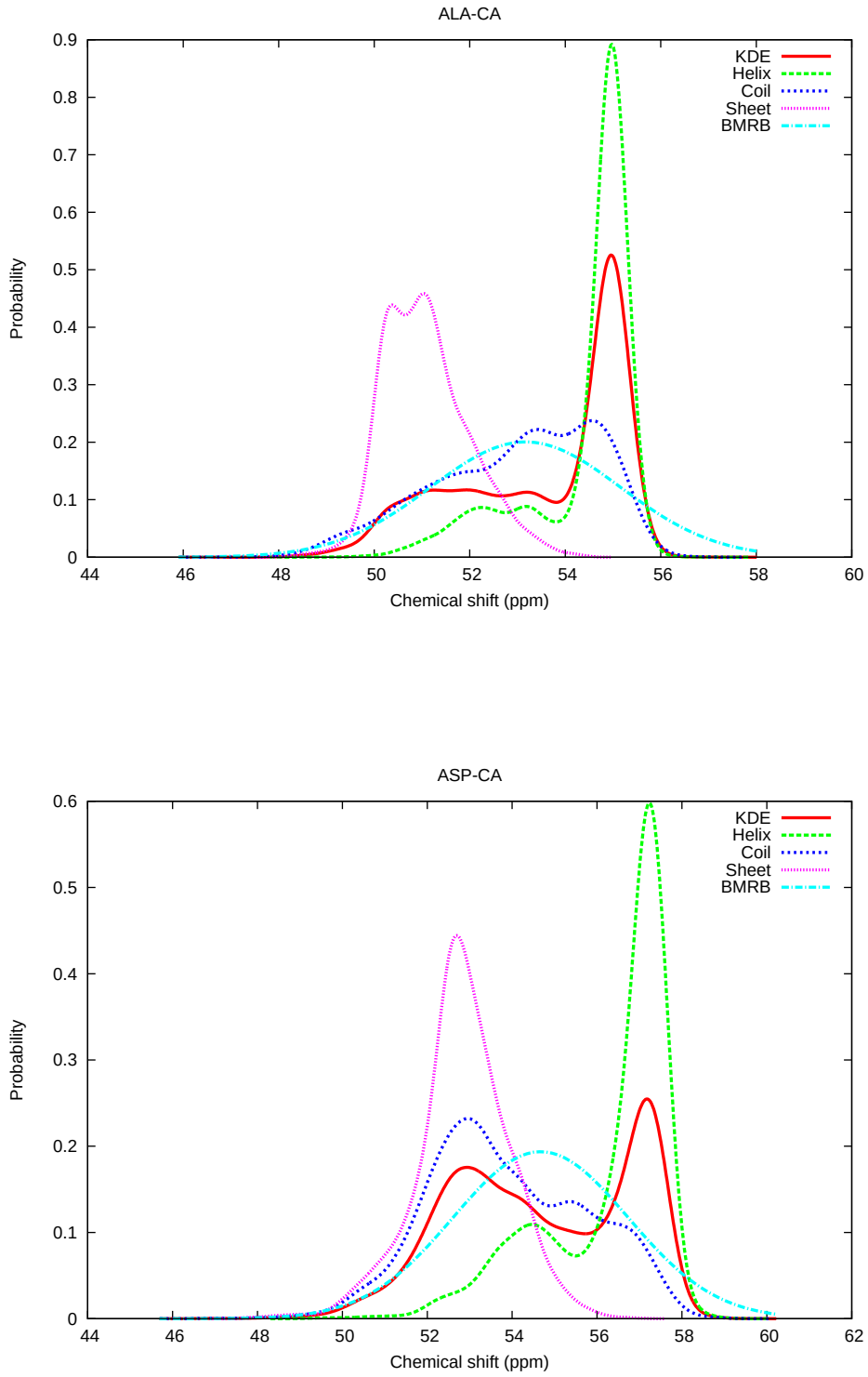


Figure 4.11: Comparison of CA-chemical shift distribution of (ALA & ASP) created from unbiased statistics and BMRB. The KDE,helix,sheet and coil functions were created using equation 4.5 and BMRB is created using 4.3

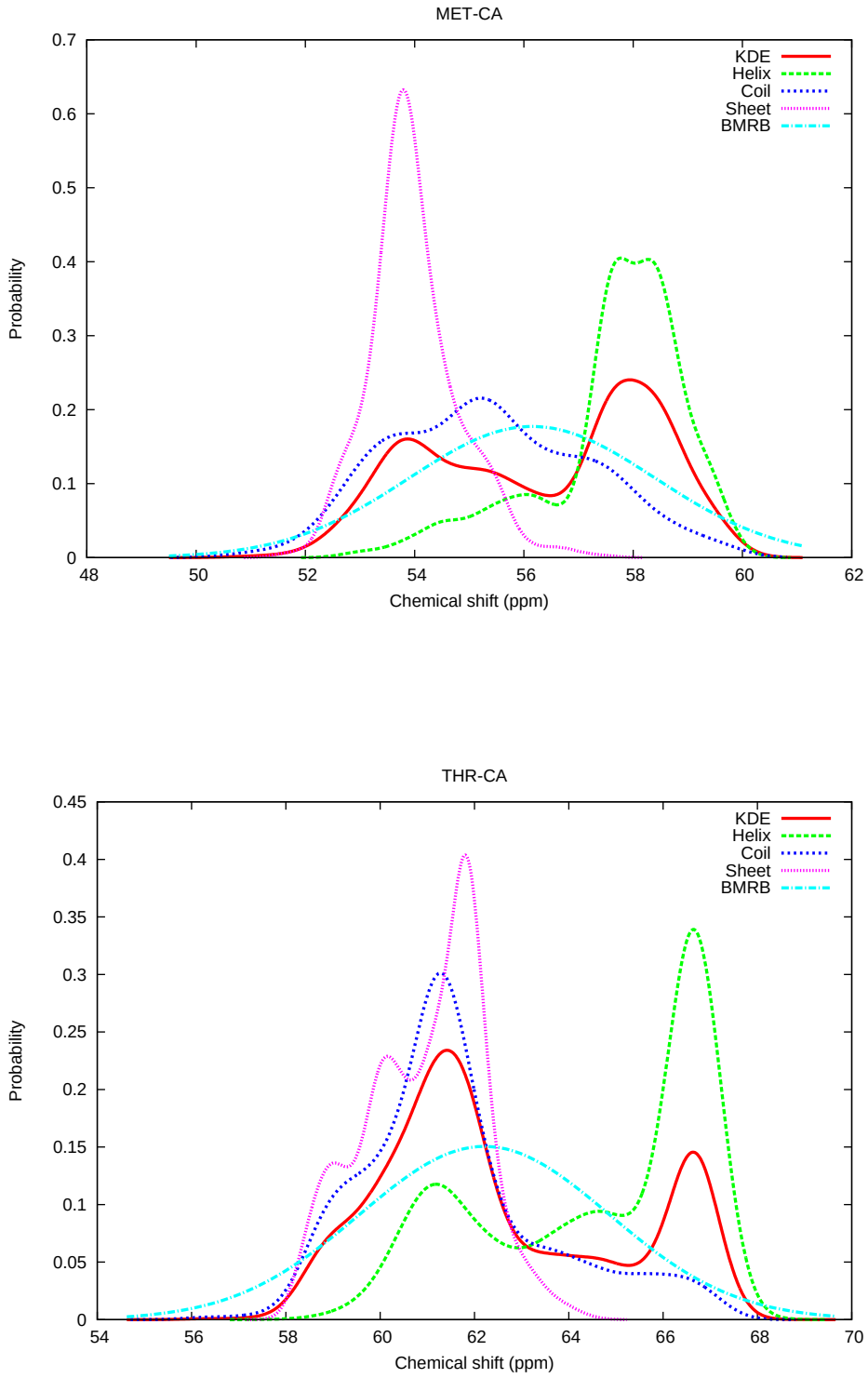


Figure 4.12: Comparison of CA-chemical shift distribution(MET & THR) created from unbiased statistics and BMRB. The KDE,helix,sheet and coil functions were created using equation 4.5 and BMRB is created using 4.3

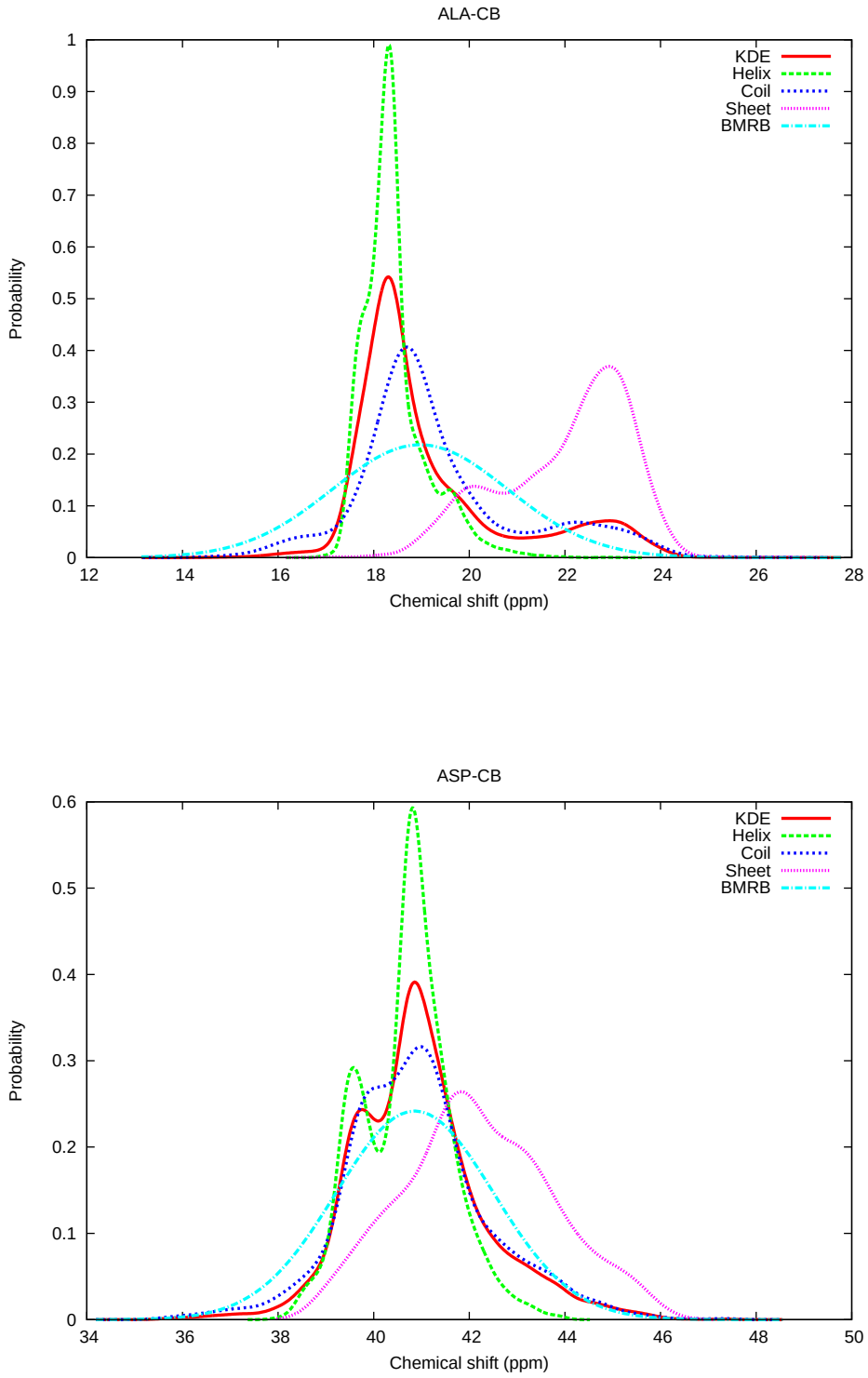


Figure 4.13: Comparison CB-chemical shift distribution(ALA & ASP) created from unbiased statistics and BMRB. The KDE,helix,sheet and coil functions were created using equation 4.5 and BMRB is created using 4.3

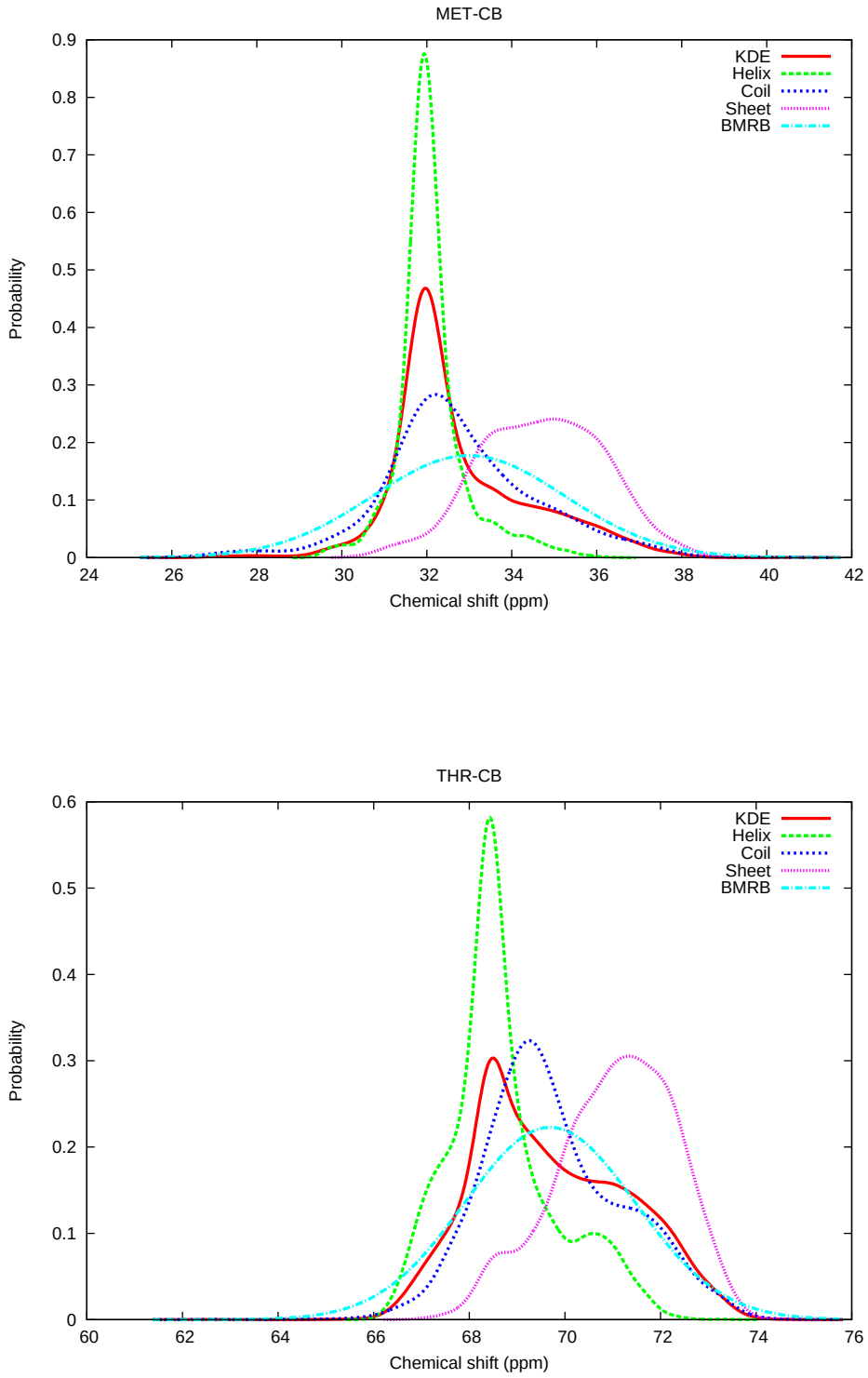


Figure 4.14: Comparison CB-chemical shift distribution(MET & THR) created from unbiased statistics and BMRB. The KDE,helix,sheet and coil functions were created using equation 4.5 and BMRB is created using 4.3

statistics showing distinct mean and standard deviation values for stereo specific atoms, whereas BMRB statistics showing almost close values and its standard deviations are almost twice as big as compared to SHIFTS and SHIFTX statistics. The standard deviation of almost all atoms are smaller (Appendix Table 7.1) compared to BMRB statistics.

Several studies [Wishart, 1991, Sharma and Rajarathnam, 2000, Pastore, 1990] [Oldfield, 1995, Asakura et al., 1995, Cornilescu et al., 1999, Laws et al., 1993] [Asakura, 1999, Morris et al., 2004, Wishart and Case, 2002, Case, 2000, Case, 1998] [Ando, 2001, Szilagyi, 1995] have been done on the secondary structure dependence on the chemical shifts. In the case of known structures, secondary structure based based statistics could be used to get better results. This is not possible from BMRB statistics, since no structural information is available there. Even in the case of unknown structure, one could use the KDE statistics to include secondary structure information indirectly. The multiple Gaussian feature in the KDE statistics indirectly includes the characteristic features of secondary structures. Figures 4.11,4.12,4.13 and 4.14 are the examples for multiple maximum in the chemical shift distribution. Similar kind of distributions can be obtained for any atom or atom type. Figure shows only CA and CB chemical shifts which show strong dependence on secondary structures when compared to any other atom types.

4.4.1.2 Limitations

The refined chemical shift statistics has certain limitations. First of all it is artificial, the accuracy of which mostly depends on the accuracy of the prediction programs. Even though backbone atoms are predicted with reasonably good accuracy, snide chain atoms, especially aromatic side chains are having problems due to ring current effects.

The chemical shift prediction is also incomplete. Some of the side chain atoms like $C\xi$ and $C\zeta$ are not predicted by these programs. The reason could be the lac of training data sets in the prediction program or due to the computational difficulties like ring current effects in the density functional formulations.

The distribution function created from this statistics are mainly from the proteins in standard physiological environment. Hence this can not be used to study a protein in a non-standard environment like in acidic medium or at different pressures and temperatures.

CHAPTER 5

CONFORMERS IN PROTEINS

5.1 Introduction

The fact that back calculation(prediction) of chemical shifts using SHIFTS[Xu and Case, 2001] and SHIFTX[Neal et al., 2003] could be improved, when an ensemble of structures is used instead of a single energy minimized structure has been been discussed in chapter three. This demonstrates the fact the protein itself found as an ensemble of structures in solution. This kind of ensemble representation of protein structures is more realistic and physically meaningful, when compared to a representation of protein as a rigid single structure determined by X-ray crystallography. As an additional outcome, prediction methods can also be used to identify the conformers in the given protein ensemble.

Protein conformation plays important role in the biological function of the protein. In a protein sample every individual protein molecule may not have exactly same three dimensional structure. Slight fluctuations in the structure are quite normal due the tumbling motion of the molecules and the temperature. But all of them can be successfully averaged to a single structure in NMR time scale. In some cases, these structures may be averaged to more than one structures corresponding to different energy minima in the conformational space. These average structures (conformers) may have different population in a given sample. Often these conformations arise due to the changes in the active site of the protein, where it can be found at two different functional states.

The presence of the conformers can be identified in many ways. If the conformers undergo exchange that is slow on NMR time scale, the spectrum may contain either multiple peaks or broad peak for atoms associated with the structural changes. If the atoms undergo fast exchange then it is hard to observe in the spectrum. In such situations one

could identify the conformers by plotting the ϕ and ψ angles of the given residue from a large ensemble. The clustering of points at different regions in the Ramachandran plot [Ramachandran et al., 1963] is a clear signature of the presence of two separate conformers. Another indirect way to detect the presence of conformers is to plot the chemical shift distribution of atoms which are supposed to be responsible for structural changes. If the ensemble contains structures whose average is just only one mean structure, then the chemical shift distribution will have only one maximum. On the other hand, if the ensemble can be grouped into two classes whose mean structures are slightly different from each other, then the chemical shift distribution will have more than one maximum. The relative heights of the normalized chemical shift distribution, will give information about the relative abundance of the conformation in the given sample.

5.2 Materials and Methods

5.2.1 NOE-Chemical shift correlation

Structural changes are often reflected in chemical shift changes. In NMR spectroscopy chemical shifts are very sensitive to the changes in the three dimensional structure. This could be seen in the correlation between NOE changes and chemical shift changes. Here experimentally measured NOEs and chemical shifts of HPr(WT) and HPr(H15A) were used to study the correlation between NOE changes and chemical shift changes. For each atom in a residue, the weighted average of change in chemical shift and change is NOE between HPr(WT) and HPr(H15A) is calculated as follows: Let us suppose $\{a_1, a_2, ..., a_m\}$ be the list of atoms present in both HPr(WT) and HPr(H15A), which have measurable NOE with atom A_i of residue R_i. The change in NOE is given by,

$$NOE_{diff}(A_i) = \frac{1}{m}\sum_{j=1}^{m} |NOE(A_i^{WT}, a_j^{WT}) - NOE(A_i^{H15A}, a_j^{H15A})| \tag{5.1}$$

and the chemical shift difference is given by,

$$CS_{diff}(A_i) = \frac{1}{\sum_{j=1}^{m} w_j}\sum_{j=1}^{m} w_j |\delta(a_j^{WT}) - \delta(a_j^{H15A})| \tag{5.2}$$

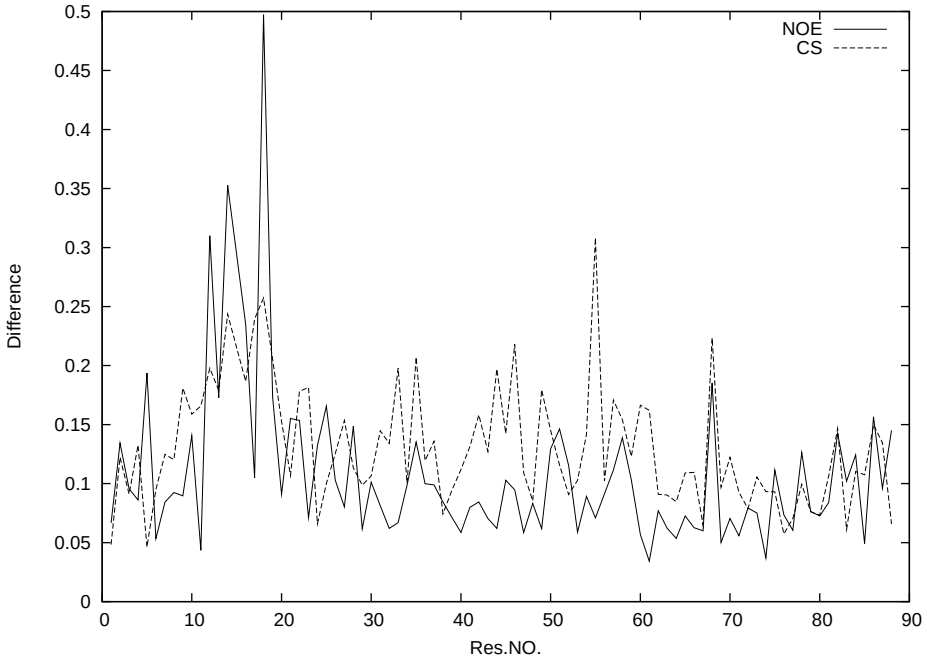


Figure 5.1: NOE-chemical shift correlation. The solid line and dotted line indicates the NOE difference and chemical shift difference between HPr(WT) and HPr(H15A) calculated using Equation 5.1 and 5.2

where w_j atom specific chemical shift weighting factor [Schumann et al., 2007]. These quantities are summed over all atoms in the residue and plotted versus residue number as show in Figure 5.1. From the figure, it is clear that in the vicinity of the mutation (H15A), the NOE changes matches with the chemical shift changes. The figure shows only the qualitative behaviour. The quantitative estimation needs much more information about the structure and symmetries. In order to model such correlations, one needs to know the exact dependence of chemical shift with distances between atoms. As a fundamental fact in NMR spectroscopy, the chemical shifts should reflect the structural changes and that is shown in Figure 5.1

Let us suppose that a given protein ensemble has n_c conformational states in given temperature pressure and pH. Often proteins have two or three conformational states. The structural changes in most cases come from one or two residues. When the structure of a protein and its mutant is compared, one would expect a structural changes in the region close to the location of mutant residue. The back bone atoms are sensitive to structural changes, and hence by plotting the distribution of back bone CA chemical shifts one can easily identify the structural changes.

5.2.2 Test data set

As a test case, ensembles of HPr(WT) and HPr(H15A) were taken. Structure calculations were performed using the molecular dynamics program CNS v.1.2. (Crystallography and NMR System for crystallographic and NMR structure determination)[Brunger, 1992, Brunger, 2007] employing the restraints (Table 3.4) in a simulated annealing protocol using extended-strand as starting structures. High-temperature torsional angle dynamics were run at 50,000 K for 3000 steps with a time step of 5 fs. The high number of restraints required a threefold reduction of the time step for the integration of the equation of motion to 5 fs and a reduction of the ceiling value to 15 for around 30 restraints per residue for the NOE-energies (the default value is 30 for 16 restraints per residue). In the first cooling stage, torsional angle dynamics were used for 3000 steps with a starting temperature of 50,000 K and a time step $5fs$. The second cooling stage was performed with 3000 steps of Cartesian dynamics with a time step of $5fs$ and a starting temperature of 3000 K. In

the final stage, 2000 steps of energy minimization were performed. The structures were accepted based on the NOE violations. Those structures having more than 5% NOE violations are rejected during simulated annealing process. Once 2000 structures were calculated using simulated annealing, they were refined in explicit water [Linge et al., 2003b]. After the water refinement the population distribution is fitted with a Gaussian distribution and those structures whose energy is $> 5\sigma$ is removed and refined again with different initial seeds until their energies were $< 5\sigma$.

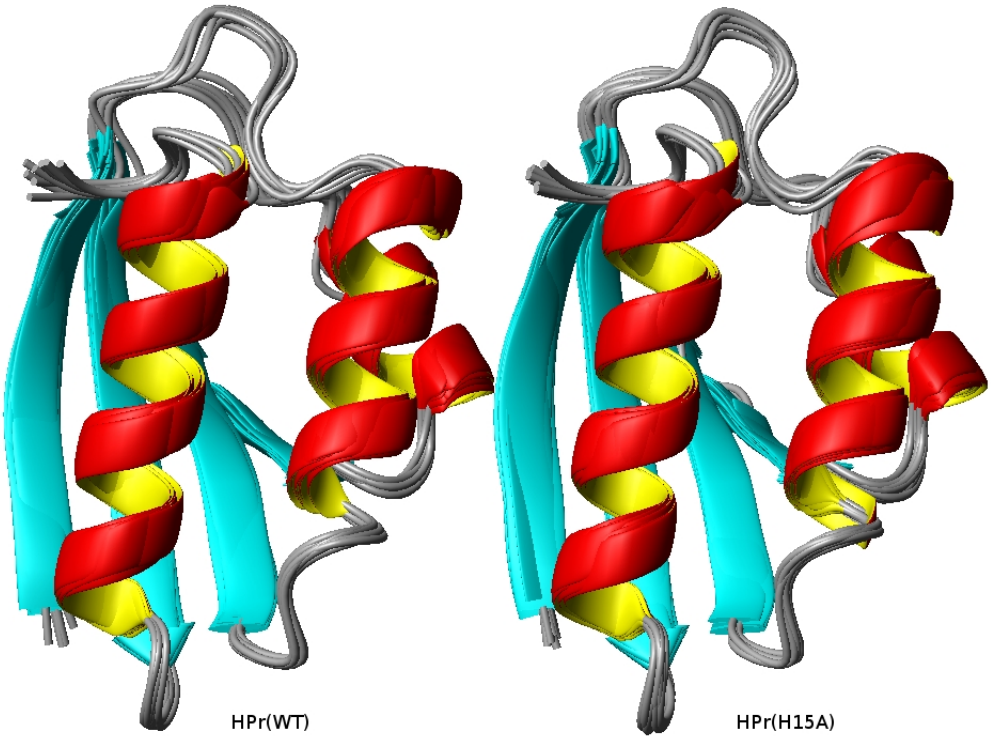


Figure 5.2: 10 lowest energy structures of HPr(WT) and HPr(H15A)

Figure 5.2 shows the 10 lowest energy structures of HPr(WT) and HPr(H15A). The structure of HPr(H15A) is expected to show slight structural changes in the nearby by mutant region. The structural changes can be detected either by Ramachandran plot or by plotting the chemical shift distribution of the near by residue. In this case ALA 16 has taken as a probe residue to detect the changes. The ϕ and ψ angle distribution of ALA 16 from 100 lowest energy structure is shown in Figure 5.3. In our test case the protein HPr has active site at residue number 15. The X-ray structure of Hpr shows an dihedral angle

strain close to active center of the protein[Jia et al., 1993], for which the ψ,ϕ angle plot of ALA 16 lies in the forbidden region of the Ramachandran [Ramachandran et al., 1963] plot . Similar conformation are observed in the solution case also. In Figure 5.3 only 3 out of 100 points for HPr(WT) lies in the forbidden region of Ramachandran plot, but for HPr(H15A) 36 out of 100 points lies in that forbidden region. This shows the presence of two conformer in HPr(H15A).

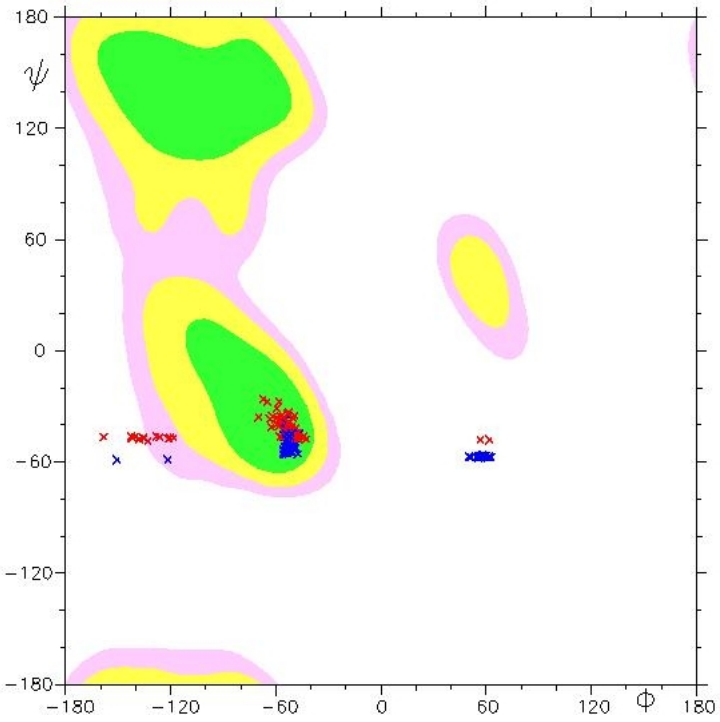


Figure 5.3: Ramachandran plot of 100 lowest energy structures :ALA 16 Red:HPr(WT) Blue:HPr(H15A)

These structural changes at residue 16 can be further confirmed by plotting the chemical shift distribution using kernel density mention as described in chatter four. The probability density function for *CA* chemical shifts of ALA 16 from the 2000 structures can be calculated equation 4.5. This distribution will have information about multiple conformers in the ensemble. If the distribution has more than one maxima , then there may be a chance for more than one conformers present in the ensemble.

Though the chemical shift distribution gives information about the presence of conformers, it does not give the information about how probable they are in a given energy.

Suppose we consider only the 10 or 20 lowest energy structures, those conformational states may not be present in that itself. If those conformers are present in the lowest energy structures also, then we could conclude that its genuine, otherwise it may be just an artefact of molecular dynamics program used to calculate the structure. This is verified by plotting the probability distribution of conformational sates.

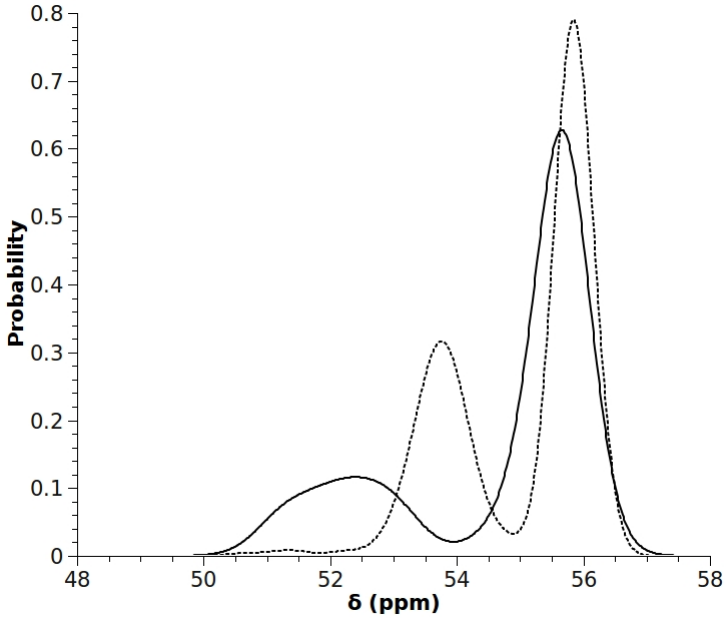


Figure 5.4: Chemical shift distribution of ALA 16 CA calculated using 4.5 from 2000 structures. Solid line indicates HPr(WT) and dotted line indicated HPr(H15A)

5.3 Results and discussions

The two maxima in Figure 5.4 and the two groups in Figure 5.3 indicate that the protein has possibly two conformer. The two maxima in Figure 5.4 has different heights in HPr(WT) and HPr(H15A). The chemical shift ranges of conformers C1 and C2 are shown in Table 5.1. This however does not tell you how probable they are in a given energy. This can be found out easily by calculating the probability of each conformer states. This is done as follows. Structures are grouped according to the chemical shift criteria given in table 5.1 and counted according to that. Using those numbers we can calculate the abun-

dance of the conformers in the given ensemble. In the 2000 structures of HPr(WT) only 5.75% are in conformational state C1 and 94.25% are in conformational state C2. In the 2000 structures of HPr(H15A) 35.9% are in conformational state C1 and 64.1% are in conformational state C2. Figure 5.5,5.6 shows the probability distribution of conformers

Table 5.1: Identifying conformers using chemical shifts distribution of ALA 16 CA. N_{C1} and N_{C2} are the number of structures found in the corresponding chemical shift range

Ensemble	C1	C2	N_{C1}	N_{C2}
HPr(WT)	50 ppm to 54 ppm	54 ppm to 58ppm	115	1885
HPr(H15A)	52 ppm to 55ppm	55 ppm to 58 ppm	718	1282

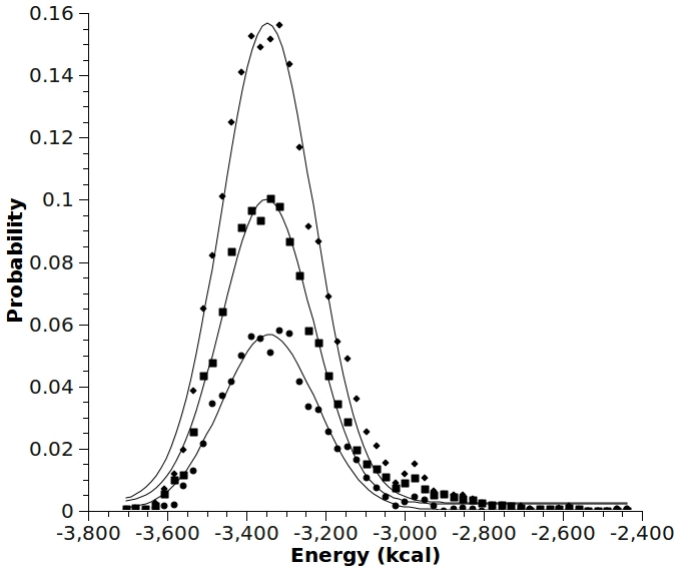


Figure 5.5: Probability of different conformers in HPr(H15A) for given energy calculated form 2000 structures. circle indicates conformer C1, square indicates conformer C2 and diamond indicates the overall. Solid line indicates the Gaussian fit for the data points.

for a given energy. Figure 5.6 shows the the probability of the conformer C1 is extremely low in HPr(WT). Hence only HPr(H15A) has the conformers and they are likely to be present even in lowest energy structures. The two maxima for HPr(H15A) in the Figure

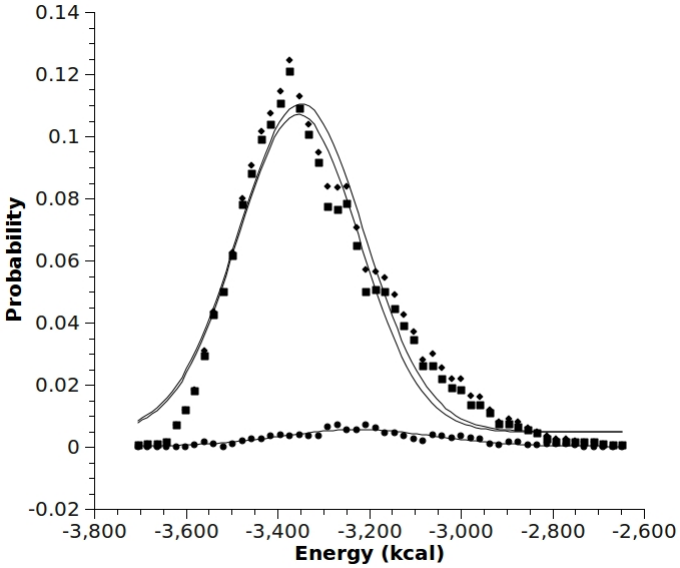


Figure 5.6: Probability of different conformers in HPr(WT) for given energy calculated form 2000 structures. circle indicates conformer C1, square indicates conformer C2 and diamond indicates the overall. Solid line indicates the Gaussian fit for the data points.

5.4 comes from the two conformers. This is verified in another way. As shown in Figure 5.3, the structures are grouped based on the ϕ, ψ angles. Those lie in the forbidden region of Ramachandran plot were considered as conformer C1 and the rest as conformer C2. Figure 5.8 shows the chemical shift distribution of conformers separately. Here the conformers are identified using ϕ, ψ angles of ALA 16. The method described here enables us to identify conformers using chemical shift distributions. In case of HPr(H15A) the conformers are hardly noticeable to human eye. Still it could be detected by Ramachandran plot as well as the chemical shift distribution method.

Figure 5.2 shows the top 10 energy minimized structures of HPr(WT) and HPr(H15A). Even though the ϕ, ψ plot and the chemical shift distribution shows the presence of conformers, it is hard to detect them visually. Figure 5.7 shows the HPr(H15A) in two different conformation. Only one structure in each case is taken and shown in different colour. The slight difference can be seen near to ALA 16, which is shown in different colour in Figure 5.7.

The correctness of this method depends on the molecular dynamics program and the

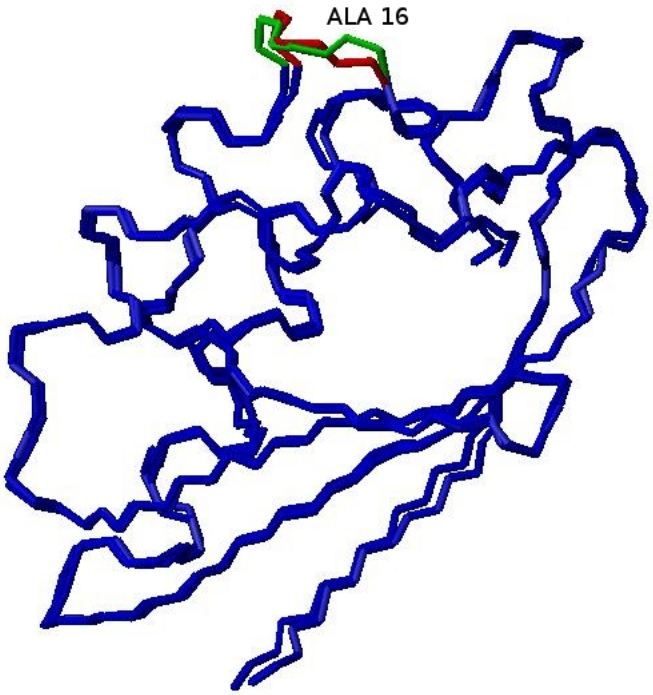
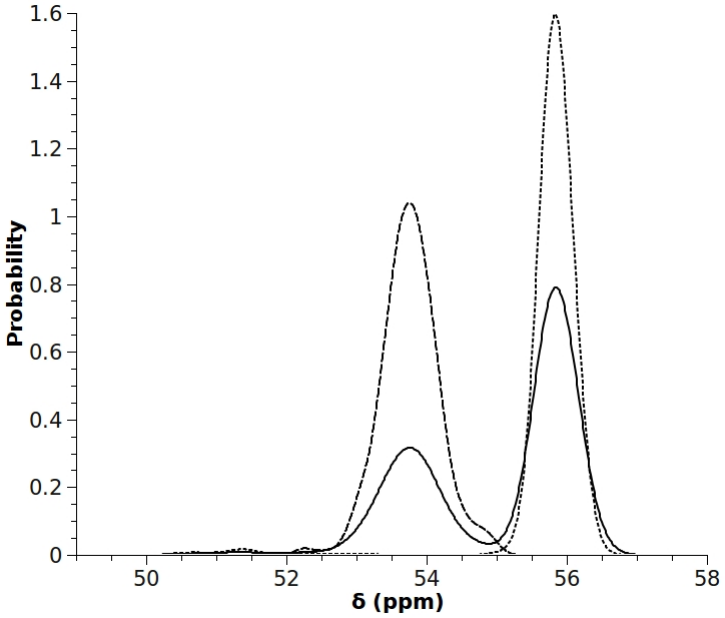


Figure 5.7: HPr(H15A) Conformations. Only one lowest energy structure from each conformer has shown in the figure. Residue 16 is showin in different color. Red belongs to the conformer whose ϕ, ψ angles are in the forbidden region of Ramachandran plot and Green belongs to the conformer whose ϕ, ψ angles are in the allowed region of Ramachandran plot



of

Figure 5.8: Chemical shift distribution of ALA 16 CA from HPr(H15A). Solid line indicates overall, dash lines indicates the distribution of chemical shifts which lies in the forbidden region of Ramachandran plot and dotted line indicates the distribution of chemical shifts of which lies in the allowed and favourable region in the Ramachandran plot

chemical shift prediction methods. The structural changes certainly reflects in chemical shift distributions. If the molecular dynamics program for a given experimental restraints finds two possible minima, then there is a possibility of presence of conformers. In order to check this one has to calculate large number of structures. The chemical shift distributions are mainly created from predictions program. Hence its not a good idea to use this method to identify conformers. Instead we can use this as an additional validation tool for the presence of conformers. It is also possible that some cases the conformational states of a given atom are in chemically equal environment. In such cases the chemical shift distributions may not have multiple maxima.

Figure 5.8 is a proof that chemical shift distribution also yields the same results when compared to Ramachandran plot. The maxima in the overall chemical shift distribution matches with the chemical shift distribution of conformers(selected based on ϕ, ψ angles). So this method could serve as additional tool to confirm the presence of conformers.

CHAPTER 6

CONCLUSIONS AND DISCUSSIONS

6.1 Conclusions

Experimental techniques and data processing are the two sides of the coin in NMR spectroscopy. The recent methodological development in the high field NMR spectroscopy is testing the limits of accuracy one could achieve. These developments are useful only when the statistical methods used for data processing has also been improved. Since the final product depends on both, it is important to concentrate on both experimental methods as well as statistical techniques. The necessary part here is the computational tools to incorporate the results and to convert the findings into a useful three dimensional structure. People have achieved to get a high resolution NMR spectra with the help of high field magnets, cryoprobes and advanced pulse sequences. The dimensionality of the multidimensional NMR spectra already gone beyond six. But still the assignment problem and the structure determination needs considerable amount of human effort and computational time.

The problem here is, NMR spectroscopy needs some statistical methods and computational techniques to explain the experimental findings. People pay less attention to these statical methods. The reason may be that the outcome of these statical methods are probabilistic in nature, and people may think that *'anyway it won't produce accurate results '*. In a sense it is true, the outcome is probabilistic, but the methods to calculate those probabilities should be accurate.

6.1.1 Chemical shift optimization

In Chapter 2, we deal with the question *'how one can extract optimum information from a given spectrum or a series of spectra?'*. The accurate measurement of the chemical shift is very important, since it caries valuable information about the structure and dynamics. Most researchers are interested only in resonance assignment and structure determination. In chapter 2 our main focus is to get a complete and correct chemical shift list with their error bounds. Chemical shift optimization of multi dimensional NMR spectra is really useful for variable temperature and pressure experiments and also to study the protein-protein and protein-ligand interactions. The use of more than one n-dimensional spectra could help us to get a more accurate chemical shift value and its error bound. It is also shown that this improvement could reduce the ambiguity in the assignment process and improve the number of assignment. It is also shown that with a help of model structure, optimization of NOESY type spectra could be performed in a better way. The weighed average performed at the end could reduce the influence of wrong assignments and artifices.

6.1.2 Chemical shift prediction

The chemical shift back calculation(prediction) programs like SHIFTS [Xu and Case, 2001] and SHIFTX [Neal et al., 2003] uses only single structure, though a bundle of structures were given as input. In chapter 3 its shown that the prediction could be improved, when an ensemble of structures is used instead of a single energy minimized structure. This demonstrates the fact that the protein itself found as an ensemble of structures in solution. This kind of ensemble representation of protein structures is more realistic and physically meaningful, when compared to a representation of protein as a rigid single structure determined by X-ray crystallography.

Structurally unbiased chemical shift distributions of atoms in all 20 standard amino acids are useful to start resonance assignment for a unknown protein. This missing link between structural database (PDB) and chemical shift database (BMRB) is an obstacle on our way to create a unbiased chemical shift distribution. This problem is solved by the use of non-homologous structural database (Nh3D) and the chemical shift prediction

program. Though the chemical shift distributions are artificial, they are structurally unbiased which is essential for statistical data processing. Relative improvements in a priory probabilities and assignments using the new refined chemical shift data base is discussed in chapter 4. The general procedure described in chapter 4 could give perfect chemical shift distributions, once we succeed in creating perfect structural database and prediction program.

When a large ensemble is present, several additional information could be extracted from it. The correlation between experimental NOEs and chemical shifts indicates that chemical shifts are sensitive to structural fluctuations. Hence chemical shifts carry the information about conformers in a protein ensemble. The chemical shift distribution of predicted chemical shifts from an ensemble shows the signs of conformers. This is shown using a large ensemble of HPr in chapter 5. This type of chemical shift distribution is useful to identify the conformers and to calculate their population in a given ensemble. This is not possible, when we have single energy minimized structure.

6.2 Applications

The automation of protein structure determination using NMR spectroscopy
[Gronwald et al., 2002, Gorler, 1999, Zimmerman, 1995, Shimotakahara et al., 1997]
[Moseley and Montelione, 1999, Li and Sanctuary, 1997b, Li and Sanctuary, 1997a]
[Koradi et al., 1998, Croft et al., 1997] is still an unsolved problem. The challenging task here is the resonance assignment. The success in assignment process mainly depends on the following facts.

- quality of the spectrum
- search range
- assignment probability

6.2.1 Search Range

For a given protein sequence, every atom is searched for its resonance peak in a specific location in the spectrum. The search range is usually centred at the expected value, and spreads over as a function of the standard deviation of that chemical shift distribution calculated from given statistics. These expected value and standard deviation could be derived from the refined chemical shift statistics. If the protein structure is already known, a secondary structure based search ranges could be defined using the refined statistics. The optimized search ranges could reduce the ambiguity in the first step itself and will produce better results.

6.2.2 A priory probability

Better assignment probabilities could be obtained using the refined statistics during the assignment process. It is not only improving the assignment probabilities of the most probables, but also the reduce the probabilities for the wrong assignments. This is shown by the improvement in assignments using Hungarian algorithm. This a priory probabilities can be coupled with other experimental information, to achieve maximum number of assignments.

6.2.3 Structure Refinement

The usual way to chose a structural ensemble from the output of molecular dynamics calculation is to sort them according to their energy and taking only lowest energy structures assuming that this subset is a better representation of a given protein. The true measurable quantity in NMR spectroscopy is the chemical shifts. Hence it is better to chose the ensemble which gives minimum prediction error when compared to experimental chemical shifts. This will guaranty that the experimental measurements have agreement with simulation.

At present situation, this may not be a good idea, since the prediction programs are not perfect. The prediction methods are not error free and the prediction method which we analysed will not consider the pressure and temperature effects on structures. May be

in future, when an ideal structure based prediction method is available, this kind sorting technique will become the gold standard.

In an another aspect, the difference between observed chemical shift and predicted chemical shift could be used as pseudo energy term in the refinement process. This structure refinement using chemical shift prediction is already existing, but only a single structure is used for prediction. Instead of a single structure, if we use ensemble we will get better results.

The idea of creating refined chemical shift statistics could be easily generalized and not specific to the data set Nh3D or prediction programs like SHIFTS and SHIFTX. The errors in this method might come from either from the non-homologous data set or from the prediction methods. In future, an ideal non-homologous data set and and an ideal structure based prediction program is found, this refined chemical shifts statistics will give much better results.

CHAPTER 7

APPENDIX

7.1 Refined Chemical Shift Statistics

Comparison of Mean and Standard deviation between BMRB and predicted chemical shifts from Nh3D. Missing chemical shifts are not predicted by prediction programs

Res	Name	BMRB Mean	BMRB Std.Dev	SHIFTS Mean	SHIFTS Std.Dev	SHIFTX Mean	SHIFTX Std.Dev	BOTH Mean	BOTH Std.Dev
ALA	HN	8.190	0.600	8.133	0.396	8.190	0.503	8.163	0.456
ALA	HA	4.260	0.440	4.138	0.371	4.236	0.452	4.190	0.419
ALA	HB	1.360	0.250	1.371	0.258	1.265	0.258	1.315	0.263
ALA	CO	177.760	2.180	178.378	2.085	178.040	1.725	178.200	1.911
ALA	CA	53.170	1.990	53.676	1.476	53.435	1.947	53.549	1.745
ALA	CB	18.970	1.830	19.150	1.793	19.220	1.588	19.187	1.688
ALA	N	123.200	3.570	121.036	4.227	122.401	3.024	121.756	3.705
ARG	HN	8.240	0.610	8.192	0.406	8.185	0.535	8.188	0.478
ARG	HA	4.300	0.460	4.208	0.387	4.222	0.469	4.215	0.432
ARG	HB2	1.800	0.270	1.891	0.302	1.826	0.258	1.857	0.282
ARG	HB3	1.770	0.280	1.745	0.302	1.873	0.265	1.812	0.290
ARG	HG2	1.570	0.270	1.467	0.350	1.550	0.265	1.510	0.311
ARG	HG3	1.550	0.280	1.505	0.367	1.592	0.276	1.551	0.326
ARG	HD2	3.120	0.240	3.062	0.324	3.122	0.261	3.094	0.291
ARG	HD3	3.110	0.250	3.059	0.349	3.150	0.257	3.107	0.305
ARG	HE	7.390	0.650	—	—	7.390	0.170	7.390	0.170
ARG	HH11	6.890	0.450	—	—	—	—	—	—
ARG	HH12	6.820	0.440	—	—	—	—	—	—
ARG	HH21	6.790	0.390	—	—	—	—	—	—
ARG	HH22	6.780	0.460	—	—	—	—	—	—
ARG	CO	176.430	2.060	177.719	2.732	176.849	1.819	177.260	2.336
ARG	CA	56.790	2.330	57.345	2.269	57.152	2.060	57.243	2.163
ARG	CB	30.670	1.860	30.979	1.768	30.922	1.531	30.949	1.648
ARG	CG	27.210	1.230	—	—	—	—	—	—
ARG	CD	43.150	0.930	—	—	—	—	—	—
ARG	CZ	160.130	3.450	—	—	—	—	—	—
ARG	N	120.810	3.730	117.923	4.975	119.686	3.364	118.853	4.294
ARG	NE	91.730	14.050	—	—	—	—	—	—
ARG	NH1	74.650	13.190	—	—	—	—	—	—
ARG	NH2	76.230	9.860	—	—	—	—	—	—
ASP	HN	8.320	0.580	8.086	0.394	8.319	0.507	8.209	0.472
ASP	HA	4.600	0.310	4.548	0.313	4.596	0.300	4.574	0.306
ASP	HB2	2.720	0.270	2.784	0.220	2.752	0.190	2.767	0.204
ASP	HB3	2.670	0.280	2.732	0.228	2.673	0.205	2.701	0.217
ASP	CO	176.430	1.790	176.836	1.472	176.661	1.596	176.743	1.541
ASP	CA	54.670	2.060	54.655	2.189	54.776	2.010	54.720	2.096
ASP	CB	40.860	1.650	40.595	1.391	41.257	1.318	40.947	1.392

ASP	CG	179.200	1.840	—	—	—	—	—	—
ASP	N	120.730	3.940	118.619	5.043	120.293	3.503	119.509	4.374
ASN	HN	8.350	0.630	8.216	0.412	8.363	0.547	8.294	0.493
ASN	HA	4.670	0.370	4.666	0.357	4.627	0.400	4.645	0.381
ASN	HB2	2.810	0.310	2.894	0.324	2.762	0.270	2.824	0.304
ASN	HB3	2.770	0.330	2.773	0.312	2.785	0.273	2.780	0.291
ASN	HD21	7.340	0.480	7.692	0.317	7.530	0.551	7.606	0.466
ASN	HD22	7.140	0.500	6.959	0.299	6.815	0.360	6.882	0.341
ASN	CO	175.310	1.830	175.772	1.598	175.101	1.462	175.414	1.563
ASN	CA	53.540	1.910	53.854	1.960	53.644	1.863	53.742	1.912
ASN	CB	38.670	1.710	38.716	1.588	39.021	1.496	38.879	1.547
ASN	CG	176.790	1.370	—	—	—	—	—	—
ASN	N	118.990	4.050	115.521	5.571	118.196	3.669	116.948	4.841
ASN	ND2	112.810	2.320	—	—	—	—	—	—
CYS	HN	8.400	0.670	8.287	0.438	8.339	0.549	8.314	0.499
CYS	HA	4.670	0.560	4.511	0.433	4.574	0.521	4.543	0.482
CYS	HB2	2.950	0.460	2.925	0.395	2.948	0.349	2.937	0.372
CYS	HB3	2.890	0.490	2.869	0.388	2.915	0.350	2.893	0.370
CYS	HG	1.880	1.440	—	—	—	—	—	—
CYS	CO	174.770	2.070	—	—	174.116	1.416	174.116	1.416
CYS	CA	58.060	3.400	—	—	58.845	2.579	58.845	2.579
CYS	CB	33.060	6.360	—	—	34.138	5.741	34.138	5.741
CYS	N	120.180	4.610	—	—	118.638	3.787	118.638	3.787
GLU	HN	8.340	0.600	8.287	0.394	8.362	0.492	8.326	0.449
GLU	HA	4.250	0.410	4.155	0.348	4.214	0.424	4.186	0.390
GLU	HB2	2.030	0.210	2.137	0.222	2.030	0.187	2.081	0.212
GLU	HB3	2.010	0.210	1.996	0.199	1.950	0.182	1.972	0.190
GLU	HG2	2.280	0.210	2.240	0.260	2.298	0.185	2.271	0.224
GLU	HG3	2.260	0.220	2.258	0.282	2.331	0.200	2.297	0.243
GLU	CO	176.940	2.000	177.775	1.900	177.345	1.837	177.550	1.880
GLU	CA	57.360	2.110	57.374	1.912	57.611	1.981	57.498	1.952
GLU	CB	29.990	1.750	30.170	1.628	30.217	1.353	30.194	1.490
GLU	CG	36.050	1.270	—	—	—	—	—	—
GLU	CD	182.460	2.130	—	—	—	—	—	—
GLU	N	120.700	3.560	118.866	4.311	119.969	3.288	119.443	3.850
GLN	HN	8.220	0.590	8.174	0.407	8.240	0.509	8.209	0.465
GLN	HA	4.270	0.430	4.221	0.371	4.219	0.431	4.220	0.404
GLN	HB2	2.050	0.260	2.169	0.288	2.092	0.240	2.129	0.265
GLN	HB3	2.020	0.270	1.989	0.266	2.066	0.242	2.029	0.256
GLN	HG2	2.320	0.280	2.261	0.336	2.312	0.265	2.288	0.298
GLN	HG3	2.300	0.290	2.258	0.365	2.353	0.279	2.309	0.324
GLN	HE21	7.230	0.470	7.641	0.289	7.348	0.321	7.495	0.312
GLN	HE22	7.020	0.460	6.909	0.240	6.757	0.262	6.888	0.259
GLN	CO	176.350	1.990	176.917	2.122	176.599	1.885	176.749	2.007
GLN	CA	56.590	2.150	56.774	2.145	57.022	2.037	56.905	2.093
GLN	CB	29.160	1.850	29.358	1.876	29.563	1.478	29.467	1.680
GLN	CG	33.760	1.150	—	—	—	—	—	—
GLN	CD	179.630	1.430	—	—	—	—	—	—
GLN	N	119.910	3.680	118.530	4.468	119.874	3.277	119.241	3.941
GLN	NE2	111.860	1.890	—	—	—	—	—	—
GLY	HN	8.330	0.660	8.095	0.453	8.311	0.679	8.209	0.593
GLY	HA2	3.970	0.380	3.926	0.395	3.906	0.357	3.926	0.395
GLY	HA3	3.900	0.380	3.970	0.381	3.906	0.357	3.970	0.381
GLY	CO	173.950	1.900	173.960	1.848	173.601	1.469	173.769	1.667
GLY	CA	45.350	1.300	45.448	1.077	45.807	1.254	45.639	1.189
GLY	N	109.690	3.960	106.279	4.669	108.734	3.582	107.586	4.304
HIS	HN	8.250	0.680	8.170	0.459	8.380	0.554	8.282	0.522
HIS	HA	4.620	0.450	4.559	0.380	4.575	0.468	4.567	0.429
HIS	HB2	3.110	0.370	3.235	0.350	3.149	0.323	3.189	0.339
HIS	HB3	3.050	0.390	3.141	0.325	3.243	0.307	3.196	0.317
HIS	HD1	9.020	2.810	—	—	—	—	—	—
HIS	HD2	7.040	0.470	6.784	0.389	5.713	0.166	6.215	0.611

HIS	HE1	7.980	0.530	7.637	0.308	6.899	0.187	7.245	0.446
HIS	HE2	10.100	2.720	—	—	—	—	—	—
HIS	CO	175.230	2.020	—	—	175.082	1.893	175.082	1.893
HIS	CA	56.460	2.390	—	—	56.308	2.049	56.308	2.049
HIS	CB	30.210	2.100	—	—	30.099	1.650	30.099	1.650
HIS	CG	131.370	3.440	—	—	—	—	—	—
HIS	CD2	119.910	2.900	—	—	—	—	—	—
HIS	CE1	137.240	2.490	—	—	—	—	—	—
HIS	N	119.550	4.100	—	—	118.360	3.788	118.360	3.788
HIS	ND1	195.990	32.980	—	—	—	—	—	—
HIS	NE2	180.100	19.240	—	—	—	—	—	—
ILE	HN	8.280	0.690	8.155	0.428	8.071	0.569	8.111	0.508
ILE	HA	4.180	0.560	4.186	0.425	4.194	0.557	4.190	0.498
ILE	HB	1.790	0.290	1.907	0.266	1.857	0.234	1.881	0.251
ILE	HG12	1.270	0.410	1.308	0.350	1.300	0.324	1.304	0.334
ILE	HG13	1.210	0.410	1.000	0.356	0.782	0.367	0.886	0.378
ILE	HG2	0.780	0.270	0.672	0.279	0.842	0.274	0.761	0.289
ILE	HD1	0.680	0.290	0.648	0.288	0.775	0.279	0.775	0.279
ILE	CO	175.880	1.960	176.275	2.056	176.256	1.731	176.265	1.892
ILE	CA	61.610	2.710	62.016	2.649	61.702	2.244	61.851	2.449
ILE	CB	38.600	2.050	38.645	1.856	38.516	1.572	38.577	1.714
ILE	CG1	27.700	1.820	—	—	—	—	—	—
ILE	CG2	17.520	1.460	—	—	—	—	—	—
ILE	CD1	13.450	1.720	—	—	—	—	—	—
ILE	N	121.510	4.340	120.056	5.238	121.093	4.219	120.602	4.758
LEU	HN	8.230	0.650	8.204	0.420	8.170	0.537	8.186	0.485
LEU	HA	4.320	0.470	4.232	0.383	4.297	0.477	4.266	0.436
LEU	HB2	1.620	0.340	1.631	0.310	1.664	0.276	1.648	0.293
LEU	HB3	1.540	0.360	1.569	0.286	1.612	0.265	1.592	0.276
LEU	HG	1.510	0.330	1.452	0.320	1.520	0.294	1.487	0.309
LEU	HD1	0.760	0.280	0.718	0.289	0.840	0.279	0.782	0.290
LEU	HD2	0.740	0.280	0.682	0.282	0.761	0.348	0.723	0.321
LEU	CO	177.010	2.030	177.620	1.883	177.480	1.735	177.546	1.808
LEU	CA	55.650	2.150	56.148	2.203	55.835	2.077	55.983	2.143
LEU	CB	42.290	1.890	42.322	1.735	42.582	1.335	42.459	1.543
LEU	CG	26.770	1.190	—	—	—	—	—	—
LEU	CD1	24.650	1.650	—	—	—	—	—	—
LEU	CD2	24.090	1.720	—	—	—	—	—	—
LEU	N	121.860	3.950	119.390	4.667	121.378	3.468	120.437	4.199
LYS	HN	8.190	0.610	8.116	0.402	8.221	0.530	8.171	0.476
LYS	HA	4.270	0.440	4.165	0.390	4.200	0.424	4.183	0.408
LYS	HB2	1.780	0.250	1.862	0.299	1.764	0.241	1.811	0.275
LYS	HB3	1.750	0.270	1.743	0.276	1.759	0.242	1.751	0.258
LYS	HG2	1.370	0.270	1.274	0.333	1.364	0.255	1.321	0.296
LYS	HG3	1.360	0.290	1.310	0.348	1.399	0.274	1.357	0.311
LYS	HD2	1.610	0.240	1.560	0.258	1.649	0.221	1.607	0.241
LYS	HD3	1.600	0.230	1.548	0.252	1.643	0.221	1.598	0.239
LYS	HE2	2.920	0.190	2.869	0.257	2.910	0.189	2.892	0.218
LYS	HE3	2.910	0.200	2.873	0.239	2.926	0.204	2.902	0.216
LYS	HZ	7.380	0.770	—	—	—	—	—	—
LYS	CO	176.710	2.000	177.324	2.023	177.234	1.677	177.277	1.850
LYS	CA	56.980	2.220	57.250	1.962	57.417	2.014	57.338	1.991
LYS	CB	32.760	1.820	33.139	1.554	32.936	1.474	33.032	1.516
LYS	CG	24.890	1.190	—	—	—	—	—	—
LYS	CD	28.920	1.230	—	—	—	—	—	—
LYS	CE	41.870	0.880	—	—	—	—	—	—
LYS	N	121.050	3.830	119.124	4.461	120.514	3.395	119.854	3.997
LYS	NZ	67.690	40.980	—	—	—	—	—	—
MET	HN	8.250	0.600	8.255	0.412	8.241	0.523	8.248	0.474
MET	HA	4.400	0.480	4.368	0.410	4.326	0.492	4.346	0.456
MET	HB2	2.030	0.350	2.148	0.287	1.984	0.296	2.061	0.303
MET	HB3	2.000	0.350	1.945	0.314	2.026	0.284	1.988	0.301

MET	HG2	2.430	0.360	2.437	0.375	2.529	0.343	2.486	0.362
MET	HG3	2.400	0.390	2.373	0.411	2.496	0.339	2.438	0.380
MET	HE	1.880	0.490	1.876	0.320	1.943	0.536	1.911	0.449
MET	CO	176.230	2.100	177.116	2.041	176.719	1.807	176.901	1.928
MET	CA	56.150	2.250	56.351	2.074	56.308	2.035	56.328	2.053
MET	CB	32.990	2.250	32.637	1.738	32.911	1.501	32.785	1.620
MET	CG	32.030	1.310	—	—	—	—	—	—
MET	CE	17.160	2.010	—	—	—	—	—	—
MET	N	120.050	3.600	117.489	4.275	119.980	3.078	118.834	3.881
PHE	HN	8.360	0.720	8.221	0.505	8.380	0.631	8.305	0.581
PHE	HA	4.620	0.570	4.555	0.431	4.595	0.552	4.577	0.499
PHE	HB2	2.990	0.370	3.108	0.356	3.030	0.318	3.067	0.339
PHE	HB3	2.950	0.390	2.992	0.329	2.987	0.300	2.989	0.314
PHE	HD1	7.060	0.310	—	—	—	—	—	—
PHE	HD2	7.060	0.310	7.053	0.329	—	—	7.053	0.329
PHE	HE1	7.090	0.320	—	—	—	—	—	—
PHE	HE2	7.090	0.320	7.144	0.298	—	—	7.144	0.298
PHE	HZ	7.020	0.410	7.083	0.388	—	—	7.083	0.388
PHE	CO	175.510	2.020	176.397	1.790	175.602	1.577	175.971	1.725
PHE	CA	58.150	2.620	58.308	2.411	58.166	2.285	58.232	2.345
PHE	CB	39.900	2.080	39.900	1.617	40.168	1.503	40.043	1.563
PHE	CG	138.010	1.870	—	—	—	—	—	—
PHE	CD1	131.490	1.240	—	—	—	—	—	—
PHE	CD2	131.530	1.150	—	—	—	—	—	—
PHE	CE1	130.650	1.470	—	—	—	—	—	—
PHE	CE2	130.750	1.160	—	—	—	—	—	—
PHE	CZ	129.230	1.610	—	—	—	—	—	—
PHE	N	120.560	4.180	118.125	4.962	120.253	3.999	119.265	4.596
PRO	H2	8.510	0.000	—	—	—	—	—	—
PRO	HA	4.400	0.340	4.325	0.329	4.396	0.339	4.362	0.337
PRO	HB2	2.070	0.350	2.167	0.303	2.062	0.292	2.111	0.302
PRO	HB3	2.020	0.360	1.987	0.347	2.014	0.315	2.001	0.330
PRO	HG2	1.930	0.320	1.958	0.336	1.961	0.296	1.960	0.314
PRO	HG3	1.910	0.340	1.988	0.361	1.963	0.317	1.975	0.339
PRO	HD2	3.650	0.360	3.759	0.404	3.623	0.319	3.687	0.367
PRO	HD3	3.620	0.390	3.823	0.438	3.638	0.376	3.726	0.417
PRO	CO	176.720	1.570	177.420	1.534	176.792	1.248	177.086	1.424
PRO	CA	63.320	1.570	63.509	1.598	63.619	1.480	63.567	1.537
PRO	CB	31.830	1.220	32.026	0.643	31.884	0.702	31.951	0.679
PRO	CG	27.200	1.140	—	—	—	—	—	—
PRO	CD	50.340	1.020	—	—	—	—	—	—
PRO	N	132.500	10.420	134.004	4.815	—	—	134.004	4.815
SER	HN	8.290	0.590	8.164	0.411	8.268	0.562	8.219	0.499
SER	HA	4.490	0.410	4.399	0.386	4.454	0.418	4.428	0.401
SER	HB2	3.880	0.260	3.966	0.274	3.920	0.225	3.942	0.249
SER	HB3	3.850	0.280	3.924	0.287	3.890	0.255	3.907	0.271
SER	HG	5.330	1.070	—	—	—	—	—	—
SER	CO	174.670	1.780	174.964	1.738	174.589	1.477	174.766	1.616
SER	CA	58.720	2.120	58.737	1.971	58.920	1.913	58.834	1.942
SER	CB	63.800	1.500	63.735	1.507	63.939	1.536	63.843	1.526
SER	N	116.290	3.610	114.440	4.939	115.736	3.299	115.125	4.203
THR	HN	8.240	0.620	8.044	0.421	8.171	0.554	8.111	0.500
THR	HA	4.460	0.480	4.431	0.389	4.390	0.481	4.410	0.441
THR	HB	4.170	0.330	4.342	0.242	4.248	0.207	4.292	0.229
THR	HG1	5.150	1.380	—	—	—	—	—	—
THR	HG2	1.140	0.230	1.042	0.255	1.182	0.235	1.116	0.254
THR	CO	174.560	1.790	174.865	1.416	174.589	1.496	174.718	1.465
THR	CA	62.210	2.650	62.355	2.662	62.666	2.446	62.520	2.554
THR	CB	69.690	1.790	69.336	1.585	70.022	1.513	69.701	1.584
THR	CG2	21.530	1.170	—	—	—	—	—	—
THR	N	115.480	4.840	112.557	5.875	114.397	4.673	113.535	5.350
TRP	HN	8.310	0.790	7.891	0.514	8.392	0.630	8.164	0.632

TRP	HA	4.690	0.530	4.523	0.434	4.612	0.541	4.572	0.497
TRP	HB2	3.190	0.360	3.190	0.365	3.342	0.330	3.273	0.354
TRP	HB3	3.130	0.370	3.173	0.325	3.128	0.298	3.148	0.311
TRP	HD1	7.150	0.370	7.045	0.375	—	—	7.045	0.375
TRP	HE1	10.090	0.580	10.145	0.362	—	—	10.145	0.362
TRP	HE3	7.320	0.410	7.328	0.448	—	—	7.328	0.448
TRP	HZ2	7.290	0.320	7.286	0.307	—	—	7.286	0.307
TRP	HZ3	6.860	0.410	6.899	0.421	—	—	6.899	0.421
TRP	HH2	6.980	0.400	7.006	0.361	—	—	7.006	0.361
TRP	CO	176.120	2.010	176.518	2.118	176.326	1.749	176.413	1.927
TRP	CA	57.670	2.580	58.001	2.459	57.744	2.152	57.860	2.299
TRP	CB	29.980	2.020	29.399	1.605	30.568	1.605	30.040	1.707
TRP	CG	110.320	1.660	—	—	—	—	—	—
TRP	CD1	126.440	1.980	—	—	—	—	—	—
TRP	CD2	127.390	1.360	—	—	—	—	—	—
TRP	CE2	138.250	7.530	—	—	—	—	—	—
TRP	CE3	120.420	1.810	—	—	—	—	—	—
TRP	CZ2	114.170	1.590	—	—	—	—	—	—
TRP	CZ3	121.400	1.620	—	—	—	—	—	—
TRP	CH2	123.720	1.920	—	—	—	—	—	—
TRP	N	121.740	4.270	119.576	4.667	123.646	5.728	121.806	5.650
TRP	NE1	129.350	2.330	—	—	—	—	—	—
TYR	HN	8.320	0.740	8.166	0.492	8.193	0.634	8.180	0.572
TYR	HA	4.630	0.560	4.535	0.410	4.535	0.518	4.535	0.470
TYR	HB2	2.900	0.380	3.021	0.367	2.976	0.324	2.997	0.345
TYR	HB3	2.850	0.400	2.925	0.322	2.903	0.289	2.913	0.305
TYR	HD1	6.940	0.300	—	—	—	—	—	—
TYR	HD2	6.930	0.300	6.942	0.307	—	—	6.942	0.307
TYR	HE1	6.710	0.230	—	—	—	—	—	—
TYR	HE2	6.710	0.240	6.699	0.219	—	—	6.699	0.219
TYR	HH	9.240	1.510	—	—	—	—	—	—
TYR	CO	175.430	2.020	175.907	2.029	175.814	1.925	175.857	1.974
TYR	CA	58.120	2.570	58.000	2.501	58.110	2.084	58.059	2.287
TYR	CB	39.300	2.170	39.424	1.416	39.580	1.627	39.507	1.534
TYR	CG	129.240	2.570	—	—	—	—	—	—
TYR	CD1	132.720	1.330	—	—	—	—	—	—
TYR	CD2	132.650	1.530	—	—	—	—	—	—
TYR	CE1	117.910	1.350	—	—	—	—	—	—
TYR	CE2	117.920	1.470	—	—	—	—	—	—
TYR	CZ	156.490	2.130	—	—	—	—	—	—
TYR	N	120.680	4.300	118.407	5.055	119.519	3.661	119.003	4.399
VAL	HN	8.290	0.690	8.146	0.441	8.123	0.571	8.134	0.513
VAL	HA	4.190	0.580	4.165	0.411	4.132	0.557	4.147	0.494
VAL	HB	1.980	0.320	2.116	0.286	2.063	0.254	2.089	0.269
VAL	HG1	0.830	0.260	0.724	0.267	0.835	0.298	0.782	0.289
VAL	HG2	0.810	0.280	0.762	0.275	0.888	0.263	0.828	0.277
VAL	CO	175.640	1.930	176.010	1.918	175.883	1.685	175.943	1.800
VAL	CA	62.460	2.900	63.081	2.888	62.751	2.405	62.907	2.650
VAL	CB	32.720	1.830	32.845	1.685	32.934	1.583	32.892	1.633
VAL	CG1	21.460	1.420	—	—	—	—	—	—
VAL	CG2	21.300	1.600	—	—	—	—	—	—
VAL	N	121.150	4.620	118.788	5.394	119.969	4.406	119.409	4.934

BIBLIOGRAPHY

[Altieri and Byrd, 2004] Altieri, A. and Byrd, R. (2004). Automation of NMR structure determination of proteins. *Current Opinion in Structural Biology*, 14(5):547–553.

[Ando, 2001] Ando, I. (2001). NMR chemical shift calculations and structural characterizations of polymers. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 39(2):79–133.

[Antz et al., 1995] Antz, C., Neidig, K.-P., and Kalbitzer, H. R. (1995). A general bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. *Journal of Biomolecular NMR*, 5(3):287–296.

[Arun and Langmead, 2004] Arun, K. and Langmead, C. J. (2004). Large scale testing of chemical shift prediction algorithms and improved machine learning based approaches to shift prediction. In *Computational Systems Bioinformatics Conference, 2004*, pages 712–713. IEEE.

[Asakura, 1999] Asakura, T. (1999). Structural analysis of silk with ^{13}C NMR chemical shift contour plots. *International Journal of Biological Macromolecules*, 24(2-3):167–171.

[Asakura et al., 1995] Asakura, T., Taoka, K., Demura, M., and Williamson, M. P. (1995). The relationship between amide proton chemical shifts and secondary structure in proteins. *Journal of Biomolecular NMR*, 6(3):227–236.

[Bailey-Kellogg et al., 2000] Bailey-Kellogg, C., Widge, A., Kelley, J. J., Berardi, M. J., Bushweller, J. H., and Donald, B. R. (2000). The NOESY jigsaw: automated protein

secondary structure and main-chain assignment from sparse, unassigned NMR data. *Journal of Computational Biology*, 7(3-4):537–558.

[Baran et al., 2004] Baran, M. C., Huang, Y. J., Moseley, H. N., and Montelione, G. T. (2004). Automated analysis of protein NMR assignments and structures. *Chemical reviews*, 104(8):3541–3556.

[Baskaran et al., 2009] Baskaran, K., Kirchhöfer, R., Huber, F., Trenner, J., Brunner, K., Gronwald, W., Neidig, K.-P., and Kalbitzer, H. (2009). Chemical shift optimization in multidimensional NMR spectra by AUREMOL-SHIFTOPT. *Journal of Biomolecular NMR*, 43(4):197–210.

[Bax and Grishaev, 2005] Bax, A. and Grishaev, A. (2005). Weak alignment NMR: a hawk-eyed view of biomolecular structure. *Current opinion in structural biology*, 15(5):563–570.

[Beger and Bolton, 1997] Beger, R. D. and Bolton, P. H. (1997). Protein ϕ and ψ dihedral restraints determined from multidimensional hypersurface correlations of backbone chemical shifts and their use in the determination of protein tertiary structures. *Journal of Biomolecular NMR*, 10(2):129–142.

[Berman et al., 2003] Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature Structural and Molecular Biology*, 10(12):980.

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28(1):235–242.

[Braun and Go, 1985] Braun, W. and Go, N. (1985). Calculation of protein conformations by proton-proton distance constraints: A new efficient algorithm. *Journal of Molecular Biology*, 186(3):611–626.

[Brodsky and Darkhovsky, 2000] Brodsky, B. E. and Darkhovsky, B. S. (2000). *Non-Parametric Statistical Diagnosis: Problems and Methods (Mathematics and Its Applications)*. Springer, 1 edition.

[Brunger, 1992] Brunger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475.

[Brunger, 2007] Brunger, A. T. (2007). Version 1.2 of the crystallography and NMR system. *Nature Protocols*, 2(11):2728–2733.

[Brünger et al., 1986] Brünger, A. T., Clore, G. M., Gronenborn, A. M., and Karplus, M. (1986). Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proceedings of the National Academy of Sciences of the United States of America*, 83(11):3801–3805.

[Case, 1998] Case, D. (1998). The use of chemical shifts and their anisotropies in biomolecular structure determination. *Current Opinion in Structural Biology*, 8(5):624–630.

[Case, 2000] Case, D. (2000). Interpretation of chemical shifts and coupling constants in macromolecules. *Current Opinion in Structural Biology*, 10(2):197–203.

[Catasti et al., 1990] Catasti, P., Carrara, E., and Nicolini, C. (1990). Pepto: An expert system for automatic peak assignment of two-dimensional nuclear magnetic resonance spectra of proteins. *Journal of Computational Chemistry*, 11(7):805–818.

[Cavanagh et al., 1995] Cavanagh, J., Fairbrother, W. J., Arthur, and Skelton, N. J. (1995). *Protein NMR Spectroscopy: Principles and Practice*. Academic Press.

[Clore, 1998] Clore, G. (1998). Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. *Journal of Magnetic Resonance*, 131(1):159–162.

[Clore et al., 1986] Clore, G. M., Brünger, A. T., Karplus, M., and Gronenborn, A. M. (1986). Application of molecular dynamics with interproton distance restraints to three-

dimensional protein structure determination. a model study of crambin. *Journal of molecular biology*, 191(3):523–551.

[Cornilescu et al., 1999] Cornilescu, G., Delaglio, F., and Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR*, 13(3):289–302.

[Crippen and Havel, 1988] Crippen, G. M. and Havel, T. F. (1988). *Distance Geometry and Molecular Conformation (Chemometrics Series)*. Research Studies Pr.

[Croft et al., 1997] Croft, D., Kemmink, J., Neidig, K.-P., and Oschkinat, H. (1997). Tools for the automated assignment of high-resolution three-dimensional protein NMR spectra based on pattern recognition techniques. *Journal of Biomolecular NMR*, 10(3):207–219.

[Dalgarno et al., 1983] Dalgarno, D. C., Levine, B. A., and Williams, R. J. P. (1983). Structural information from NMR secondary chemical shifts of peptide α CH protons in proteins. *Bioscience Reports*, 3(5):443–452.

[de Dios et al., 1993] de Dios, A. C., Pearson, J. G., and Oldfield, E. (1993). Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. *Science*, 260(5113):1491–1496.

[Deshpande et al., 2005] Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M., and Bourne, P. E. (2005). The RCSB protein data bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Research*, 33(suppl_1):D233–237.

[Feng et al., 1991] Feng, Y. Q., Wand, A. J., Roder, H., and Englander, S. W. (1991). Chemical exchange in two dimensions in the ^1H NMR assignment of cytochrome c. *Biophysical journal*, 59(2):323–328.

[Fischer et al., 1999] Fischer, M. W., Losonczi, J. A., Weaver, J. L., and Prestegard, J. H. (1999). Domain orientation and dynamics in multidomain proteins from residual dipolar couplings. *Biochemistry*, 38(28):9013–9022.

[Forsén and Hoffman, 1963] Forsén, S. and Hoffman, R. A. (1963). Study of moderately rapid chemical exchange reactions by means of nuclear magnetic double resonance. *The Journal of Chemical Physics*, 39(11):2892–2901.

[Galzitskaya and Melnik, 2003] Galzitskaya, O. V. and Melnik, B. S. (2003). Prediction of protein domain boundaries from sequence alone. *Protein Science*, 12(4):696–701.

[George and Muller, 1958] George and Muller, M. E. (1958). A note on the generation of random normal deviates. *Ann.˜Math.˜Stat.*, 29(2):610–611.

[Geyer, 1995] Geyer, M. (1995). Automated peak integration in multidimensional NMR spectra by an optimized iterative segmentation procedure. *Journal of Magnetic Resonance, Series B*, 109(1):31–38.

[Geyer et al., 1996] Geyer, M., Schweins, T., Herrmann, C., Prisner, T., Wittinghofer, A., and Kalbitzer, H. R. (1996). Conformational transitions in p21ras and in its complexes with the effector protein Raf-RBD and the GTPase activating protein GAP. *Biochemistry*, 35(32):10308–10320.

[Glaser, 1987] Glaser, S. (1987). Automated recognition and assessment of cross peaks in two-dimensional NMR spectra of macromolecules. *Journal of Magnetic Resonance (1969)*, 74(3):450–463.

[Gorler, 1999] Gorler, A. (1999). Computer assisted assignment of^{13}C or^{15}N edited 3D-NOESY-HSQC spectra using back calculated and experimental spectra. *Journal of Magnetic Resonance*, 137(1):39–45.

[Görler et al., 1999] Görler, A., Hengstenberg, W., Kravanja, M., Beneicke, W., Maurer, T., and Kalbitzer, H. R. (1999). Solution structure of the histidine-containing phosphocarrier protein fromstaphylococcus carnosus. *Applied Magnetic Resonance*, 17(2):465–480.

[Görler and Kalbitzer, 1997] Görler, A. and Kalbitzer, H. R. (1997). Relax, a flexible program for the back calculation of NOESY spectra based on complete-relaxation-matrix formalism. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 124(1):177–188.

[Grishaev and Llinás, 2005] Grishaev, A. and Llinás, M. (2005). Protein structure elucidation from minimal NMR data: the CLOUDS approach. *Methods in enzymology*, 394:261–295.

[Gronwald et al., 2008] Gronwald, W., Bomke, J., Maurer, T., Domogalla, B., Huber, F., Schumann, F., Kremer, W., Fink, F., Rysiok, T., and Frech, M. (2008). Structure of the leech protein saratin and characterization of its binding to collagen. *Journal of Molecular Biology*, 381(4):913–927.

[Gronwald and Kalbitzer, 2004] Gronwald, W. and Kalbitzer, H. R. (2004). Automated structure determination of proteins by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 44(1-2):33–96.

[Gronwald et al., 2000] Gronwald, W., Kirchhöfer, R., Görler, A., Kremer, W., Ganslmeier, B., Neidig, K. P., and Kalbitzer, H. R. (2000). RFAC, a program for automated NMR r-factor estimation. *Journal of biomolecular NMR*, 17(2):137–151.

[Gronwald et al., 2002] Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J., Kremer, W., Neidig, K.-P., and Kalbitzer, H. R. (2002). Automated assignment of noesy NMR spectra using a knowledge based method (KNOWNOE). *Journal of Biomolecular NMR*, 23(4):271–287.

[Güntert, 1998] Güntert, P. (1998). Structure calculation of biological macromolecules from NMR data. *Quarterly Reviews of Biophysics*, 31(02):145–237.

[Güntert, 2003] Güntert, P. (2003). Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43(3-4):105–125.

[Güntert et al., 1991] Güntert, P., Braun, W., and Wüthrich, K. (1991). Efficient computation of three-dimensional protein structures in solution from nuclear magnetic reso-

nance data using the program diana and the supporting programs CALIBA, HABAS and GLOMSA. *Journal of Molecular Biology*, 217(3):517–530.

[Hahmann et al., 1998] Hahmann, M., Maurer, T., Lorenz, M., Hengstenberg, W., Glaser, S., and Kalbitzer, H. R. (1998). Structural studies of histidine-containing phosphocarrier protein from enterococcus faecalis. *European Journal of Biochemistry*, 252(1):51–58.

[Havel, 1991] Havel, T. F. (1991). An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Progress in biophysics and molecular biology*, 56(1):43–78.

[Herranz et al., 1992] Herranz, J., González, C., Rico, M., Nieto, J. L., Santoro, J., Jiménez, M. A., Bruix, M., Neira, J. L., and Blanco, F. J. (1992). Peptide group chemical shift computation. *Magnetic Resonance in Chemistry*, 30(10):1012–1018.

[Herrmann et al., 2002] Herrmann, T., Güntert, P., and Wüthrich, K. (2002). Protein NMR structure determination with automated noe assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology*, 319(1):209–227.

[Hiller et al., 2008] Hiller, S., Joss, R., and Wider, G. (2008). Automated NMR assignment of protein side chain resonances using automated projection spectroscopy (APSY). *Journal of the American Chemical Society*, 130(36):12073–12079.

[Holm and Sander, 1994] Holm, L. and Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research*, 22(17):3600–3609.

[Hyberts et al., 1992] Hyberts, S. G., Goldberg, M. S., Havel, T. F., and Wagner, G. (1992). The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. *Protein Science*, 1(6):736–751.

[Iuga et al., 2004] Iuga, A., Spoerner, M., Kalbitzer, H. R., and Brunner, E. (2004). Solid-state 31P NMR spectroscopy of microcrystals of the ras protein and its effector loop

mutants: Comparison between crystalline and solution state. *Journal of Molecular Biology*, 342(3):1033–1040.

[Iwadate et al., 1999] Iwadate, M., Asakura, T., and Williamson, M. P. (1999). C alpha and C beta carbon-13 chemical shifts in proteins from an empirical database. *Journal of biomolecular NMR*, 13(3):199–211.

[Jia et al., 1994] Jia, Z., Vandonselaar, M., Hengstenberg, W., Quail, J. W., and Delbaere, L. T. (1994). The 1.6 A structure of histidine-containing phosphotransfer protein HPr from streptococcus faecalis. *Journal of Molecular Biology*, 236(5):1341–1355.

[Jia et al., 1993] Jia, Z., Vandonselaar, M., Quail, J. W., and Delbaere, L. T. (1993). Active-centre torsion-angle strain revealed in 1.6 A-resolution structure of histidine-containing phosphocarrier protein. *Nature*, 361(6407):94–97.

[Jonker and Volgenant, 1986] Jonker, R. and Volgenant, T. (1986). Improving the hungarian assignment algorithm. *Operations Research Letters*, 5(4):171–175.

[Kalbitzer et al., 2000] Kalbitzer, H. R, A. G. R., Hua, L. I., Dubovsk, P. V., Wolfgang Hengstenber, G., Claudia Kowoli, K., Hiroaki Yamad, A., and Kazuyuki Akasak, A. (2000). ^{15}N and ^{1}H NMR study of histidine containing protein (HPr) from staphylococcus carnosus at high pressure. *Protein Science*, 9(04):693–703.

[Kalbitzer et al., 2009] Kalbitzer, H. R., Spoerner, M., Ganser, P., Hozsa, C., and Kremer, W. (2009). Fundamental link between folding states and functional states of proteins. *Journal of the American Chemical Society*, 131(46):16714–16719.

[Karplus, 1959] Karplus, M. (1959). Contact electron-spin coupling of nuclear magnetic moments. *The Journal of Chemical Physics*, 30(1):11–15.

[Karplus, 1960] Karplus, M. (1960). Weak interactions in molecular quantum mechanics. *Reviews of Modern Physics*, 32(2):455–460.

[Kendrew et al., 1960] Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., and Shore, V. C. (1960). Structure of myoglobin: A three-dimensional fourier synthesis at 2 A. resolution. *Nature*, 185(4711):422–427.

[Koradi et al., 1998] Koradi, R., Billeter, M., Engeli, M., Guntert, P., and Wuthrich, K. (1998). Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of Magnetic Resonance*, 135(2):288–297.

[Kouranov et al., 2006] Kouranov, A., Xie, L., de La, Chen, L., Westbrook, J., Bourne, P. E., and Berman, H. M. (2006). The RCSB PDB information portal for structural genomics. *Nucleic acids research*, 34(Database issue).

[Krishna and Berliner, 1999] Krishna, R. N. and Berliner, L. J. (1999). *Biological Magnetic Resonance - Volume 16: Modern Techniques in Protein NMR*. Springer, 1 edition.

[Laskowski et al., 1996] Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *Journal of biomolecular NMR*, 8(4):477–486.

[Laws et al., 1993] Laws, D. D., de Dios, A. C., and Oldfield, E. (1993). NMR chemical shifts and structure refinement in proteins. *Journal of Biomolecular NMR*, 3(5):607–612.

[Le and Oldfield, 1994] Le, H. and Oldfield, E. (1994). Correlation between ^{15}N NMR chemical shifts in proteins and secondary structure. *Journal of biomolecular NMR*, 4(3):341–348.

[Le et al., 1995] Le, H.-b., Pearson, J. G., de Dios, A. C., and Oldfield, E. (1995). Protein structure refinement and prediction via NMR chemical shifts and quantum chemistry. *Journal of the American Chemical Society*, 117(13):3800–3807.

[Lehtivarjo et al., 2009] Lehtivarjo, J., Hassinen, T., Korhonen, S.-P., Peräkylä, M., and Laatikainen, R. (2009). 4D prediction of protein 1h chemical shifts. *Journal of Biomolecular NMR*, 45(4):413–426.

[Levitt and Chothia, 1976] Levitt, M. and Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, 261(5561):552–558.

[Li and Sanctuary, 1997a] Li, K.-B. and Sanctuary, B. C. (1997a). Automated resonance assignment of proteins using heteronuclear 3D NMR. 1. backbone spin systems extraction and creation of polypeptides. *Journal of Chemical Information and Computer Sciences*, 37(2):359–366.

[Li and Sanctuary, 1997b] Li, K.-B. and Sanctuary, B. C. (1997b). Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. side chain and sequence-specific assignment. *Journal of Chemical Information and Computer Sciences*, 37(3):467–477.

[Linge et al., 2003a] Linge, J. P., Habeck, M., Rieping, W., and Nilges, M. (2003a). ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics (Oxford, England)*, 19(2):315–316.

[Linge et al., 2003b] Linge, J. P., Williams, M. A., Spronk, C. A. E. M., Bonvin, A. M. J. J., and Nilges, M. (2003b). Refinement of protein structures in explicit solvent. *Proteins: Structure, Function, and Genetics*, 50(3):496–506.

[Liu et al., 2005] Liu, G., Shen, Y., Atreya, H. S., Parish, D., Shao, Y., Sukumaran, D. K., Xiao, R., Yee, A., Lemak, A., Bhattacharya, A., Acton, T. A., Arrowsmith, C. H., Montelione, G. T., and Szyperski, T. (2005). NMR data collection and analysis protocol for high-throughput protein structure determination. *Proceedings of the National Academy of Sciences U S A*, 102(30):10487–10492.

[Markley et al., 1998] Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E., and Wüthrich, K. (1998). Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB inter-union task group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. *Journal of biomolecular NMR*, 12(1):1–23.

[Maurer et al., 2004] Maurer, T., Meier, S., Kachel, N., Munte, C. E., Hasenbein, S., Koch, B., Hengstenberg, W., and Kalbitzer, H. R. (2004). High-resolution structure

of the histidine-containing phosphocarrier protein (HPr) from staphylococcus aureus and characterization of its interaction with the bifunctional HPr kinase/phosphorylase. *Journal of bacteriology*, 186(17):5906–5918.

[Michie, 1996] Michie, A. (1996). Analysis of domain structural class using an automated class assignment protocol. *Journal of Molecular Biology*, 262(2):168–185.

[Morris et al., 2004] Morris, L. C., Valafar, H., and Prestegard, J. H. (2004). Assignment of protein backbone resonances using connectivity, torsion angles and $^{13}C\alpha$ chemical shifts. *Journal of Biomolecular NMR*, 29(1):1–9.

[Moseley and Montelione, 1999] Moseley, H. N. B. and Montelione, G. T. (1999). Automated analysis of NMR assignments and structures for proteins. *Current Opinion in Structural Biology*, 9(5):635–642.

[Mueller et al., 2000] Mueller, G. A., Choy, W. Y., Yang, D., Forman-Kay, J. D., Venters, R. A., and Kay, L. E. (2000). Global folds of proteins with low densities of noes using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. *Journal of Molecular Biology*, 300(1):197–212.

[Neal et al., 2003] Neal, S., Nip, A. M., Zhang, H., and Wishart, D. S. (2003). Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *Journal of Biomolecular NMR*, 26(3):215–240.

[Neidig et al., 1984] Neidig, K. P., Bodenmueller, H., and Kalbitzer, H. R. (1984). Computer aided evaluation of two-dimensional NMR spectra of proteins. *Biochemical and biophysical research communications*, 125(3):1143–1150.

[Nelson et al., 1991] Nelson, S. J., Schneider, D. M., and Wand, A. J. (1991). Implementation of the main chain directed assignment strategy. computer assisted approach. *Biophysical journal*, 59(5):1113–1122.

[Nilges et al., 1988] Nilges, M., Clore, G. M., and Gronenborn, A. M. (1988). Determination of three-dimensional structures of proteins from interproton distance data by

dynamical simulated annealing from a random array of atoms. circumventing problems associated with folding. *FEBS letters*, 239(1):129–136.

[Nowick et al., 2003] Nowick, J. S., Khakshoor, O., Hashemzadeh, M., and Brower, J. O. (2003). DSA: A new internal standard for NMR studies in aqueous solution. *Organic Letters*, 5(19):3511–3513.

[Oldfield, 1995] Oldfield, E. (1995). Chemical shifts and three-dimensional protein structures. *Journal of Biomolecular NMR*, 5(3):217–225.

[Orengo et al., 1994] Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, 372(6507):631–634.

[Orengo et al., 1997] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH-a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, 5(8):1093–1108.

[Orengo and Thornton, 2005] Orengo, C. A. and Thornton, J. M. (2005). Protein families and their evolution-a structural perspective. *Annual Review of Biochemistry*, 74:867–900.

[Osapay and Case, 1991] Osapay, K. and Case, D. A. (1991). A new analysis of proton chemical shifts in proteins. *Journal of the American Chemical Society*, 113(25):9436–9444.

[Ösapay and Case, 1994] Ösapay, K. and Case, D. A. (1994). Analysis of proton chemical shifts in regular secondary structure of proteins. *Journal of Biomolecular NMR*, 4(2):215–230.

[Pai et al., 1990] Pai, E. F., Krengel, U., Petsko, G. A., Goody, R. S., Kabsch, W., and Wittinghofer, A. (1990). Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: implications for the mechanism of GTP hydrolysis. *The EMBO journal*, 9(8):2351–2359.

[Pardi et al., 1984] Pardi, A., Billeter, M., and Wüthrich, K. (1984). Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3jhn alpha, in a globular protein. use of 3JHN alpha for identification of helical secondary structure. *Journal of molecular biology*, 180(3):741–751.

[Pastore, 1990] Pastore, A. (1990). The relationship between chemical shift and secondary structure in proteins. *Journal of Magnetic Resonance (1969)*, 90(1):165–176.

[Perkins et al., 1977] Perkins, S. J., Johnson, L. N., Phillips, D. C., and Dwek, R. A. (1977). Conformational changes, dynamics and assignments in ^1H NMR studies of proteins using ring current calculations. hen egg white lysozyme. *FEBS letters*, 82(1):17–22.

[Press et al., 1992] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2 edition.

[Ramachandran et al., 1963] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7:95–99.

[Renugopalakrishnan et al., 1991] Renugopalakrishnan, V., Carey, P. R., Smith, I. C. P., Huans, S. G., and Storer (1991). *Proteins: Structure, Dynamics and Design*. Springer, 1 edition.

[Ried et al., 2004] Ried, A., Gronwald, W., Trenner, J. M., Brunner, K., Neidig, K.-P., and Kalbitzer, H. R. (2004). Improved simulation of NOESY spectra by RELAX-JT2 including effects of J-coupling, transverse relaxation and chemical shift anisotropy. *Journal of Biomolecular NMR*, 30(2):121–131.

[Rieping et al., 2005] Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science*, 309(5732):303–306.

[Schmid, 1978] Schmid, H. J. (1978). A geometrical interpretation of the hungarian method. *Discrete Mathematics*, 21(3):297–308.

[Schulte, 1997] Schulte, A. (1997). Use of global symmetries in automated signal class recognition by a bayesian method. *Journal of Magnetic Resonance*, 129(2):165–172.

[Schumann et al., 2007] Schumann, F. H., Riepl, H., Maurer, T., Gronwald, W., Neidig, K.-P. P., and Kalbitzer, H. R. R. (2007). Combined chemical shift changes and amino acid specific chemical shift mapping of protein-protein interactions. *Journal of biomolecular NMR*, 39(4):275–289.

[Schwarzinger et al., 2000] Schwarzinger, S., Kroon, G. J., Foss, T. R., Wright, P. E., and Dyson, J. H. (2000). Random coil chemical shifts in acidic 8M urea: Implementation of random coil shift data in NMRView. *Journal of Biomolecular NMR*, 18(1):43–48.

[Seavey et al., 1991] Seavey, B. R., Farr, E. A., Westler, W. M., and Markley, J. L. (1991). A relational database for sequence-specific protein NMR data. *Journal of Biomolecular NMR*, 1(3):217–236.

[Sharma and Rajarathnam, 2000] Sharma, D. and Rajarathnam, K. (2000). ^{13}C NMR chemical shifts can predict disulfide bond formation. *Journal of Biomolecular NMR*, 18(2):165–171.

[Shimotakahara et al., 1997] Shimotakahara, S., Rios, C. B., Laity, J. H., Zimmerman, D. E., Scheraga, H. A., and Montelione, G. T. (1997). NMR structural analysis of an analog of an intermediate formed in the rate-determining step of one pathway in the oxidative folding of bovine pancreatic ribonuclease a: Automated analysis of ^{1}H, ^{13}C, and ^{15}N resonance assignments for wild-type and [c65s, c72s] mutant forms†. *Biochemistry*, 36(23):6915–6929.

[Sitkoff, 1998] Sitkoff, D. (1998). Theories of chemical shift anisotropies in proteins and nucleic acids. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 32(2):165–190.

[Sitkoff and Case, 1997] Sitkoff, D. and Case, D. A. (1997). Density functional calculations of proton chemical shifts in model peptides. *Journal of the American Chemical Society*, 119(50):12262–12273.

[Sjölander et al., 1996] Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., and Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Computer Applications in Biosciences*, 12(4):327–345.

[Spera and Bax, 1991] Spera, S. and Bax, A. (1991). Empirical correlation between protein backbone conformation and C.alpha. and C.beta. ^{13}C nuclear magnetic resonance chemical shifts. *Journal of the American Chemical Society*, 113(14):5490–5492.

[Spoerner et al., 2001] Spoerner, M., Herrmann, C., Vetter, I. R., Kalbitzer, H. R., and Wittinghofer, A. (2001). Dynamic properties of the ras switch i region and its importance for binding to effectors. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):4944–4949.

[Stumber et al., 2002] Stumber, M., Geyer, M., Graf, R., Robert Kalbitzer, H., Scheffzek, K., and Haeberlen, U. (2002). Observation of slow dynamic exchange processes in ras protein crystals by 31P solid state NMR spectroscopy. *Journal of Molecular Biology*, 323(5):899–907.

[Swindells, 1995] Swindells, M. B. (1995). A procedure for detecting structural domains in proteins. *Protein Science*, 4(1):103–112.

[Szilagyi, 1995] Szilagyi, L. (1995). Chemical shifts in proteins come of age. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 27(4):325–430.

[Thiruv et al., 2005] Thiruv, B., Quon, G., Saldanha, S. A., and Steipe, B. (2005). Nh3D: a reference dataset of non-homologous protein structures. *BMC Structural Biology*, 5.

[Turlach, 1993] Turlach, B. A. (1993). Bandwidth selection in kernel density estimation: A review. In *CORE and Institut de Statistique*, pages 23–493.

[Ulrich et al., 2007] Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent, Yao, H., and Markley, J. L. (2007). Biomagresbank. *Nucleic Acids Research*, pages 402–408.

[Wagner, 1983] Wagner, G. (1983). Characterization of the distribution of internal motions in the basic pancreatic trypsin inhibitor using a large number of internal NMR probes. *Quarterly reviews of biophysics*, 16(1):1–57.

[Wagner et al., 1987] Wagner, G., Braun, W., Havel, T. F., Schaumann, T., Go, N., and Wüthrich, K. (1987). Protein structures in solution by nuclear magnetic resonance and distance geometry. the polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN. *Journal of Molecular Biology*, 196(3):611–639.

[Wagner et al., 1983] Wagner, G., Pardi, A., and Wuethrich, K. (1983). Hydrogen bond length and proton NMR chemical shifts in proteins. *Journal of the American Chemical Society*, 105(18):5948–5949.

[Wang and Jardetzky, 2002] Wang, Y. and Jardetzky, O. (2002). Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein science : a publication of the Protein Society*, 11(4):852–861.

[Williamson and Asakura, 1997] Williamson, M. P. and Asakura, T. (1997). Protein chemical shifts. pages 53–69.

[Williamson et al., 1992] Williamson, M. P., Asakura, T., Nakamura, E., and Demura, M. (1992). A method for the calculation of protein alpha-ch chemical shifts. *Journal of biomolecular NMR*, 2(1):83–98.

[Wishart, 1991] Wishart, D. (1991). Simple techniques for the quantification of protein secondary structure by ^1H NMR spectroscopy. *FEBS Letters*, 293(1-2):72–80.

[Wishart and Case, 2002] Wishart, D. and Case, D. (2002). Use of chemical shifts in macromolecular structure determination. *Methods in Enzymology*, 338:3–34.

[Wishart and Nip, 1998] Wishart, D. S. and Nip, A. M. (1998). Protein chemical shift analysis: a practical guide. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 76(2-3):153–163.

[Wishart and Sykes, 1994] Wishart, D. S. and Sykes, B. D. (1994). The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *Journal of biomolecular NMR*, 4(2):171–180.

[Wishart et al., 1992] Wishart, D. S., Sykes, B. D., and Richards, F. M. (1992). The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6):1647–1651.

[Wright, 1990] Wright, M. B. (1990). Speeding up the hungarian algorithm. *Computers and Operations Research*, 17(1):95–96.

[Wüthrich, 1986] Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids (Baker Lecture Series)*. Wiley-Interscience, 1 edition.

[Xia et al., 2002] Xia, B., Tsui, V., Case, D. A., Dyson, H. J., and Wright, P. E. (2002). Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized born model, and with explicit water. *Journal of biomolecular NMR*, 22(4):317–331.

[Xu and Case, 2001] Xu, X.-P. and Case, D. (2001). Automated prediction of ^{15}N, $^{13}C\alpha$, $^{13}C\beta$ and ^{13}C chemical shifts in proteins using a density functional database. *Journal of Biomolecular NMR*, 21(4):321–333.

[Xu and Case, 2002] Xu, X.-P. and Case, D. A. (2002). Probing multiple effects on ^{15}N, $^{13}C\alpha$, $^{13}C\beta$ and ^{13}C chemical shifts in peptides using density functional theory. *Biopolymers*, 65(6):408–423.

[Xu et al., 2001] Xu, Y., Jablonsky, M. J., Jackson, P. L., Braun, W., and Krishna, N. R. (2001). Automatic assignment of NOESY cross peaks and determination of the protein structure of a new world scorpion neurotoxin using NOAH/DIAMOD. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 148(1):35–46.

[Zhang et al., 2003] Zhang, H., Neal, S., and Wishart, D. S. (2003). RefDB: a database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR*, 25(3):173–195.

[Zimmerman, 1995] Zimmerman, D. (1995). Automated analysis of nuclear magnetic resonance assignments for proteins. *Current Opinion in Structural Biology*, 5(5):664–673.

[Zimmerman, 1997] Zimmerman, D. (1997). Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology*, 269(4):592–610.

Erklärung

Hiermit erkläre ich, das ich die vorliegende Arbeit selbständig angefertigt, und keine Hilfsmittel, außer den angegebenen, benutzt habe.

Regensburg, 08-02-2010

Kumaran Baskaran