uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Steve Chao-Chung Shih
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Chemistry)
GRADE / DEGREE

Department of Chemistry
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Exploring Structure Calculation Strategies for Polytopic α- Helical Membrane Protein Structure Determination by Solution NMR

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. N. Goto
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Dr. Y. Aubin

Dr. D. Bryce

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Exploring structure calculation strategies for polytopic α-helical membrane protein structure determination by solution NMR

Steve Chao-Chung Shih

Thesis submitted to the
Faculty of Graduate & Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for the M.Sc. degree in the
Ottawa-Carleton Chemistry Institute

Thèse soumise à
L'école des études supérieures et de la recherche
Université d'Ottawa
En vue de l'obtention de la maîtrise science à
L'Institut de chimie d'Ottawa-Carleton

uOttawa

L'Université canadienn
Canada's university

Candidate	Supervisor
	
Steve Chao-Chung Shih	Natalie Goto

© Steve Chao-Chung Shih, Ottawa, Canada, 2007

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-34110-0
Our file Notre référence
ISBN: 978-0-494-34110-0

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Structure determination of membrane proteins by solution nuclear magnetic resonance spectroscopy require that they can be incorporated into detergent micelles in order to ensure that they are maintained in a folded, water-soluble state. However, detergent molecules produce micelle-protein complexes that pose unique challenges for solution NMR studies. To improve the spectroscopic properties of these complexes, it is possible to use specific isotope labeling strategies developed for the study of large water-soluble proteins by solution NMR. Specifically, the use of a highly deuterated sample that retains protons only in the Val, Leu and Ile ($\delta 1$) methyl groups has the potential to be used to determine structures of polytopic helical membrane proteins. However, while the quality of structure that can be obtained from this approach has been examined for water-soluble proteins, its utility for this class of membrane proteins has not been systematically investigated. This thesis evaluates the utility of this method for structure determination of membrane proteins by exploring structure calculation strategies from simulated NMR data sets from membrane proteins of known structure. Here I present results for a set of membrane proteins that consist of pairs or bundles of hydrophobic α-helices. I analyzed the impact of intra- and inter-helical NOEs on structure quality for samples specifically labeled with protons only at the methyl and amide positions. In addition, I also explored the role of dihedral angles and residual dipolar couplings in an effort to find general trends to improve the accuracy of structures for polytopic α-helical membrane proteins. Based on the results of these calculations it appears that obtaining a uniform distribution of inter-helical NOEs and accurately identifying transmembrane helices for dihedral angle restraints are the most

important factors determining the accuracy of these simulated membrane protein structures. These results should help to guide future structure determinations for polytopic α-helical membrane proteins of unknown structure.

Table of Contents

List of Figures

List of Tables

Abbreviations

AMC	ADP/ATP mitchondria carrier
DPC	Dodecylphosphocholine
E.coli	Escherichia coli
GlpF	Glycerol Facilatator conducting channel
GlyR	Glycine Receptor
GpA	Glycophorin A
Halor	Halorhodopsin
kDa	kiloDalton
LPPC	1-palmitoyl-2-hydroxy-sn-glycero-3-phosphocholine
MD	Molecular Dynamics
NOE	Nuclear Overhauser Effect
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
ppm	Parts Per Million
RDC	Residual Dipolar Couplings
rf	Radiofrequency
rmsd	Root mean square deviation
SA	Simulated annealing
SDS	Sodium Dodecyl Sulfate
T_1	Longitudinal relaxation time
T_2	Transverse relaxation time
τ_c	Overall rotational correlation time
TROSY	Transverse Relaxation Optimized Spectroscopy

Acknowledgements

First and foremost I would like thank my supervisor, Dr. Natalie Goto, who has given me a chance to explore the limitless possibilities of research. She is truly an excellent mentor and her patience has allowed me to learn new facets of science that I was not aware of before. I am really happy that I had the opportunity to learn from her, not just in research, not just in science, but overall as a person. You have truly 'opened' the door for my future, and I will forever be grateful for that.

I would also like to thank Dr. Thierry Ducat who has 'bossed' me around and put me into place around the laboratory. He is truly an excellent scientist, but most importantly a great friend. I will not forget our trip to Germany and I wish you and your family the best for the future.

I would also like to thank all the people in the lab that has made my experience more enjoyable and unforgettable. I will never forget Dennis' (D-rod) loud laughter and our outings to pubs, my sports talk with Sudeep, Jenny.C's spontaneous insults, Lisette's (Dr. Heydaddy) social events and our outings to 'Genji' and all the other sushi places, Allison's 'slow Mac' and how we both love subway (but you hate sushi) and Jenny Wu who has helped me so much in learning the ways of NMR. Thanks also to the summer students who has made my experience a great one: Fred, Tatjana (my other lover), Yuan (the 'funny' dance), Tabs (kazzam – when are you going to show me a magic trick?), Asma (my LB buddy), Benson who is attached to Ifrah, Nina (the quiet one) and Sara (who is not part of the Goto lab, but is my 'road trip' buddy – see you in T.O). All of you have made this part of life very special and I'm so grateful and lucky to have met and to continue to be friends with all of you.

Finally, thank you to Heidi who has also helped and matured me and brought me back down to earth when I needed it. I really appreciate the love and patience you have given me during this time. And especially I would like to thank my parents and my brother who has helped me and pushed me during these years, without you I would not be where I am right now.

Chapter 1

INTRODUCTION

1.1 Membrane protein structures

Integral membrane proteins reside in an asymmetric environment that is created by the phospholipid bilayer. The regions of these membrane proteins that traverse the lipid bilayer are exposed to a highly hydrophobic environment that would present a high energetic barrier to the incorporation of polar groups such as a peptide amide group. However, participation in hydrogen bonding interactions can decrease the energetic penalty of insertion into the hydrophobic lipid core. Therefore membrane protein structures tend to maximize the number of hydrogen bonded amide groups that would be inserted into the hydrophobic interior of the lipid bilayer (1). This can be achieved by adopting two types of secondary structure: an α-helix or β-barrel conformation (Figure 1.1). In contrast with β-strands, α-helices can exist as an independent stable monotopic helix in the lipid bilayer. It can also pack or hydrogen bond with other α-helices to traverse the membrane more than once, forming a polytopic helical membrane protein (Figure 1.2).

Both monotopic and polytopic membrane proteins fulfill vital functions in many biological processes such as cellular transport of salts and toxins, inter-cellular signaling, growth and regulation. They are also of interest in structural genomics and are targets for many drugs (2); for example, members of the 7-helix bundle G-protein coupled receptor

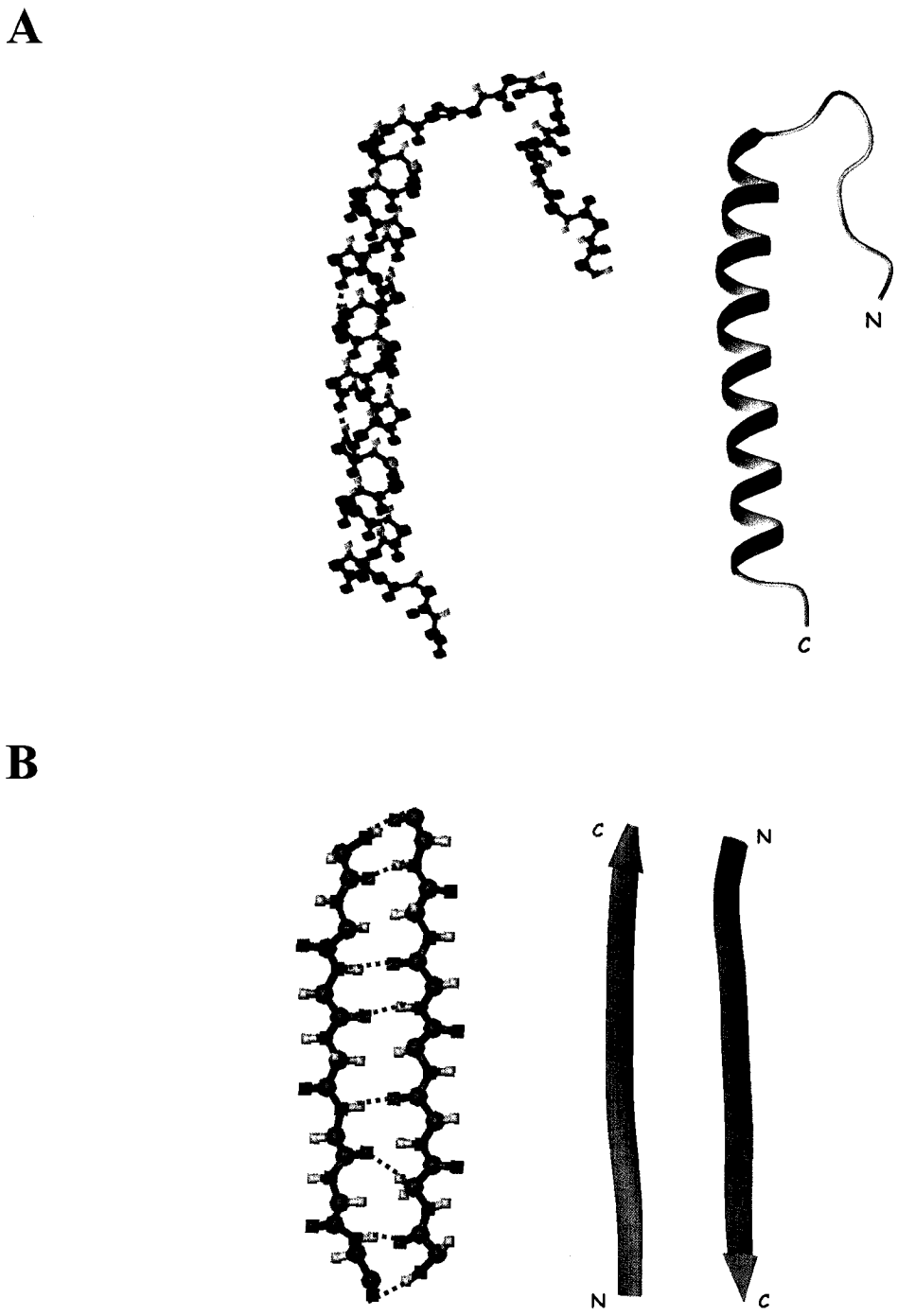

Figure 1.1 – Principal protein secondary structure elements. A ball and stick representation of the backbone atoms and their respective ribbon diagrams are represented with the carbon atoms shown in black, nitrogen atoms in blue, oxygen atoms in red and hydrogen atoms in grey. Hydrogen bonds are represented by the green dashed lines. A) A canonical α-helix. Intra-helical hydrogen bonds between the carbonyl oxygen at position n and the amide hydrogen at n+4 are highlighted. B) An anti-parallel β-sheet held together by inter-strand hydrogen bonds.

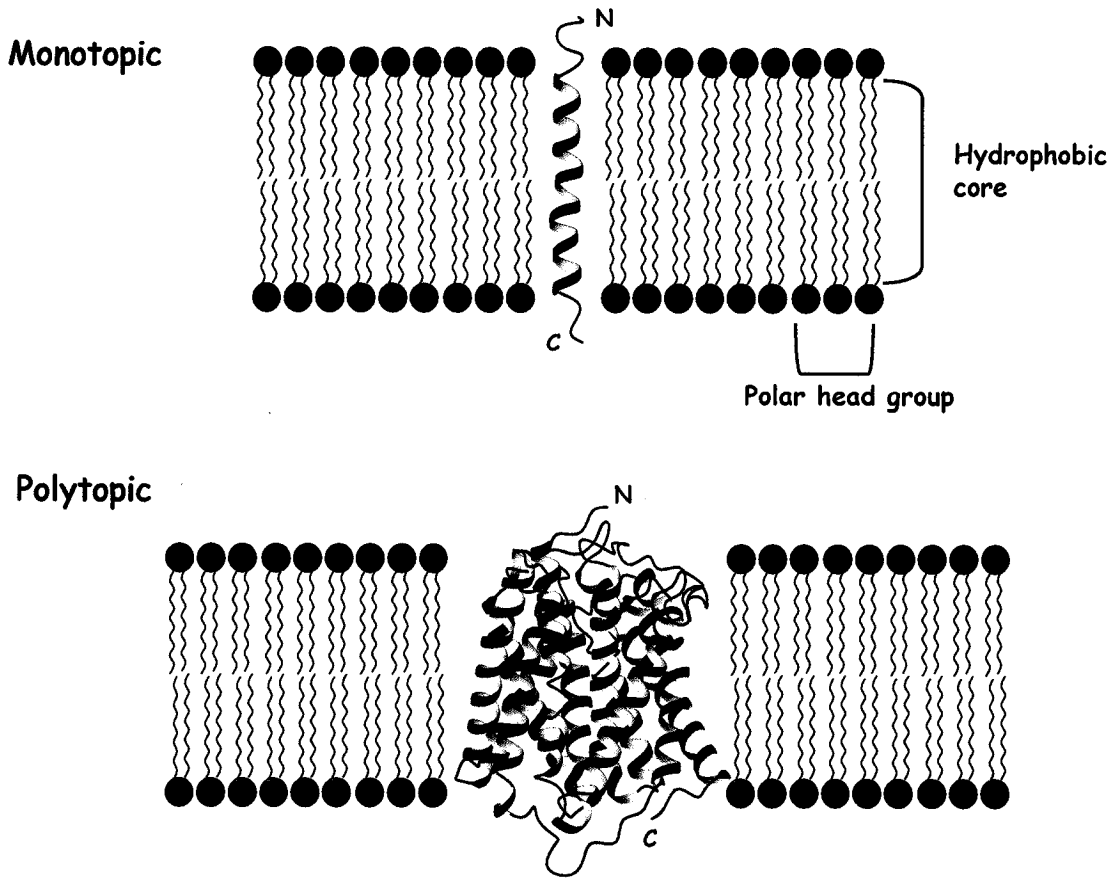

Figure 1.2 – Schematic diagram showing the arrangement of a monotopic and polytopic membrane protein in the lipid bilayer. A monotopic membrane proteins consists of a hydrophobic helix (shown in blue) crossing a lipid bilayer comprised of polar headgroups (red circles) and hydrophobic alkyl chains (green). A polytopic membrane protein is often made up of a bundle of hydrophobic helices spanning the lipid bilayer, as shown for the ammonia channel (PDB code: 1U7G).

family of proteins are associated with a wide range of human disease conditions, including cardiovascular, CNS and inflammatory diseases (3-5). Given their functional and pharmacological significance, there is great interest in understanding the function and mechanism of membrane proteins. Towards this goal, atomic resolution structures can be a useful source of information since there is a close relationship between protein structure and function. High-resolution structures have the potential to provide insights into the

different molecular mechanisms that can be used by these proteins at the membrane and also have the potential to be used to aid drug design.

Of the more than 37,000 high resolution protein structures known (6) only 120 membrane proteins are currently represented in the protein structure databank (PDB) (7), most of which were determined by x-ray crystallography. One major bottleneck in membrane protein structure determination arises from difficulties in sample production and handling. In particular, the significant hydrophobicity of membrane-embedded segments drives them to unfold and precipitate when extracted from the native lipid environment. Therefore, solubility considerations introduce a need for detergent to be present to form a protein-detergent micelle complex (Figure 1.3). This complicates the crystallographic analysis of membrane proteins since the detergents introduce heterogeneity that can complicate crystal formation.

An alternative to x-ray crystallography that avoids the requirement for crystal formation is nuclear magnetic resonance (NMR), an approach that has been widely applied to determine structures of smaller (~<20 kDa) water-soluble proteins (8). Since then, major improvements in sample preparation, spectrometer hardware, and pulse sequence methodology have dramatically increased the range and utility of NMR for the elucidation of structures of membrane proteins that are solubilized in detergents (9-11). For example, the availability of > 13 Tesla NMR magnets and transverse relaxation-optimized spectroscopy (TROSY) (12) has allowed the acquisition of high-resolution NMR spectra for proteins with molecular weights exceeding 50-60 kDa (13-19). These improvements provide much better spectra, which has allowed additional NMR structural data to be obtained to help define the tertiary fold of the protein. However, in spite of

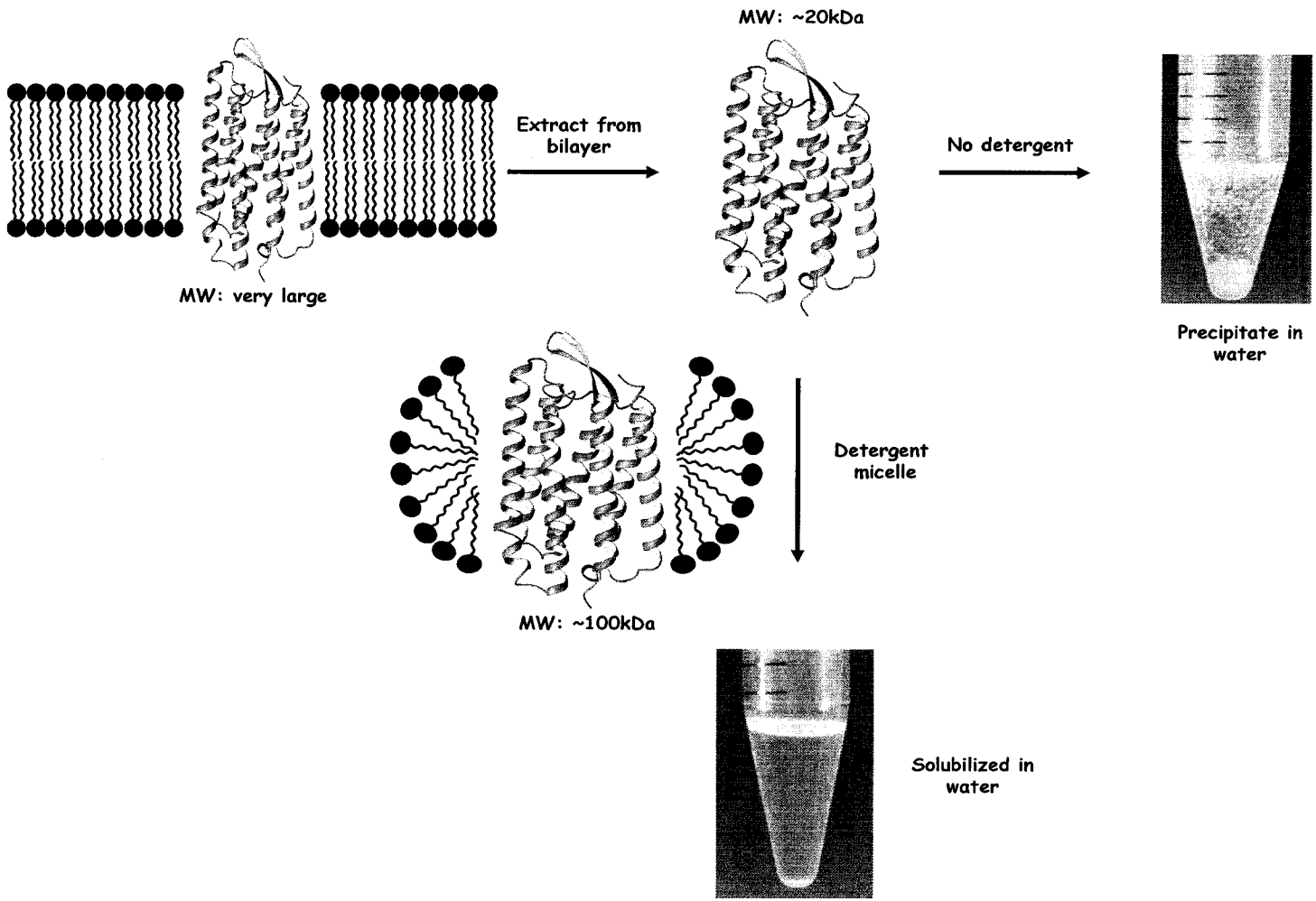

Figure 1.3 – Schematic representation of the role of detergents in the extraction of a membrane protein from the lipid bilayer. Extraction of membrane proteins from the lipid bilayer requires the addition of a detergent. In the absence of detergent this membrane protein will aggregate, leading to the formation of a precipitate. The addition of a detergent maintains the membrane protein in solution, but forms a high molecular weight (~100kDa) micelle-detergent complex which can be problematic for solution NMR studies. The multi-spanning protein halorhodopsin (PDB code: 1E12) is shown by a ribbon representation and has a molecular weight of 20 kDa in the absence of detergent.

these impressive advances, peak broadening and overlap problems still arise for large membrane protein-detergent micelle complexes, limiting progress in this field.

1.2 Protein structure determination by solution NMR

NMR is a technique that utilizes the magnetic properties of certain nuclei that possess intrinsic angular momentum, or "spin" (I). Solution NMR spectroscopy of biomolecules such as proteins mainly utilize nuclei with a spin quantum number of I = ½ (e.g. ^{1}H, ^{15}N, ^{13}C). Briefly, this technique relies on an induced dipole moment in these spins that interacts with an external magnetic field (supplied by the spectrometer) to orient in either a high energy or low energy state. As shown in Figure 1.4, NMR relies on the ability to use radiofrequency energy pulses to induce transitions between these two states. The frequency at which absorption occurs is known as the Larmor frequency. However, it depends, in part, on the local chemical environment of the nucleus, giving rise to a spectrum where each magnetically distinct nucleus can absorb at a unique resonance frequency (Larmor frequency with a shielding term). (A more complete description of the NMR phenomenon can be found in (20, 21)).

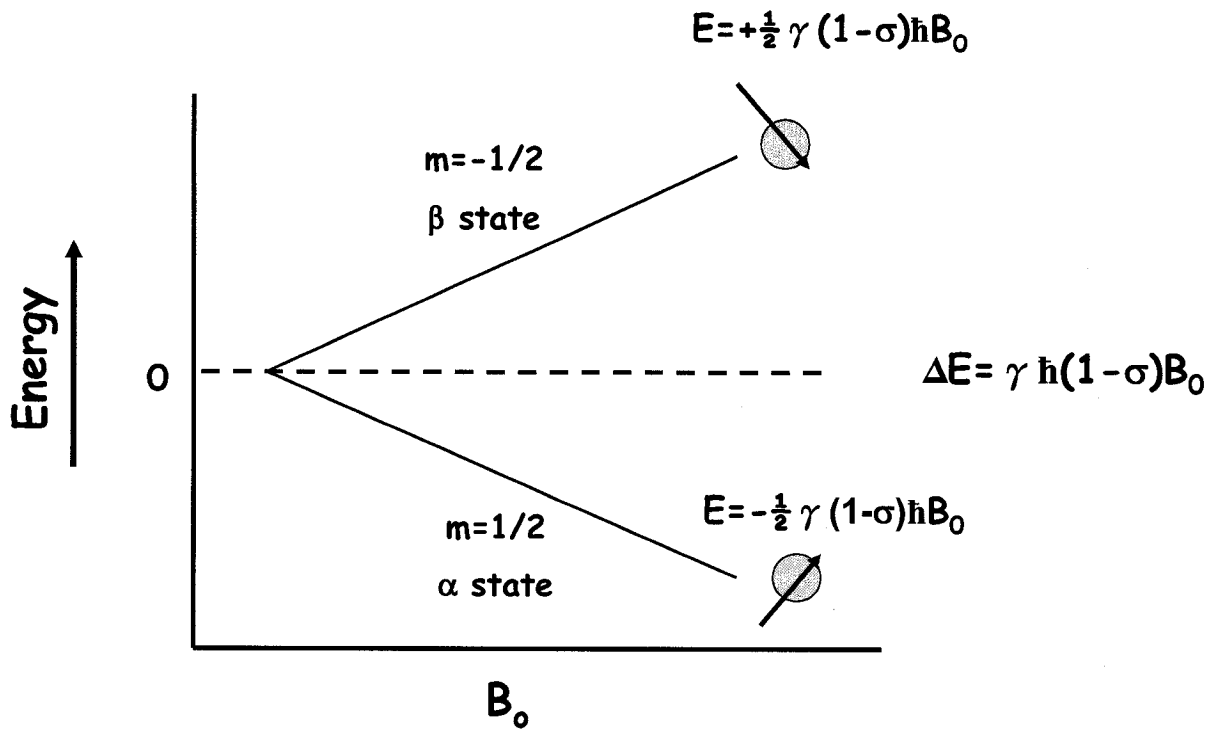

Figure 1.4 – Energy level diagram for an isolated proton with spin quantum number ½ in the absence (left side) and presence (right side) of an external magnetic field B_0. In an external magnetic field, the proton spin can adopt one of two possible orientations each with a slightly different energy. The α-state (m=1/2) has the lowest energy and it is denoted by the spin-up arrow while the higher energy β-state (m=-1/2) is shown by the spin-down arrow. The energy of these states is given by $E = -m\gamma\hbar(1-\sigma)B_o$ where γ is the gyromagnetic ratio, h is Planck's constant and σ is nuclear shielding. If the system is irradiated with a radiofrequency energy that matches the energy difference between these two states, (*i.e.* $\Delta E = \gamma\hbar(1-\sigma)B_o$) transitions between the two states can occur and a NMR signal is detected. The larger the strength of the external magnetic field, the larger this energy difference. Different local chemical environments give rise to differences in σ, and therefore different absorption frequencies for the same type of spin can result.

NMR spectroscopy has been used for protein structure determination using the general strategy outlined in Figure 1.5 (22, 23) . In this approach a homogenous, concentrated (~1 mM) protein sample must be prepared, usually with the NMR-active isotopes ^{15}N and/or ^{13}C incorporated (although structures of smaller proteins ~10 kDa and under can be solved without the need for isotope-enrichment). This is usually achieved

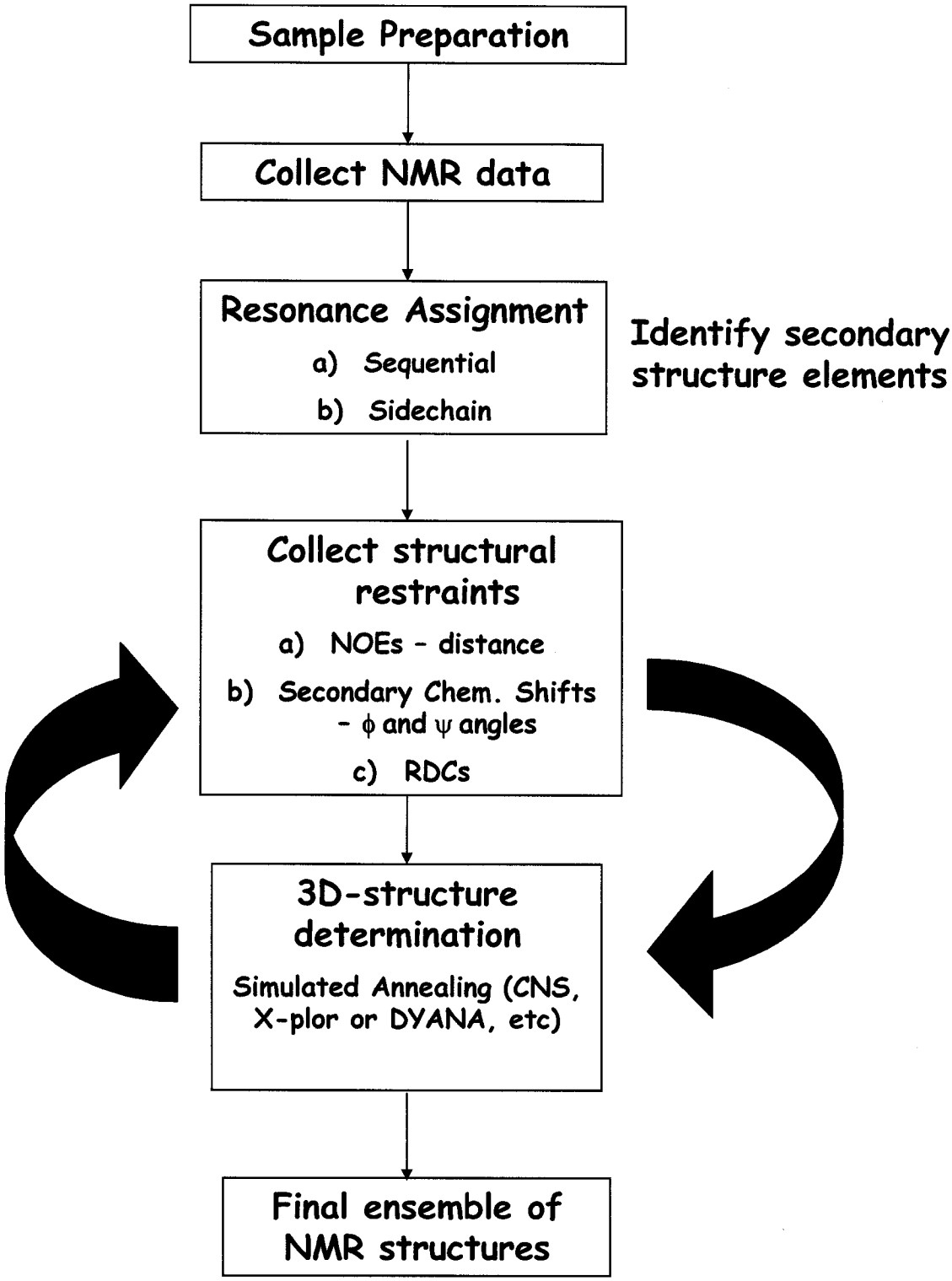

Figure 1.5 – Outline of general strategy to determine a protein structure by NMR spectroscopy.

by over-expressing the protein in a suitable host, such as *E. coli* that is grown in a medium containing isotope-labeled metabolic precursors ($^{15}NH_4Cl$ and ^{13}C-glucose in the case of *E. coli*). After purification of this sample, a standard series of triple resonance NMR experiments can be run to determine the chemical shifts of backbone atoms such as $^{13}C\alpha$, $^{13}C\beta$, ^{13}CO, ^{15}N, and $^{1}H_N$ (21, 22). The basis for all these experiments is the two-dimensional ^{15}N-^{1}H HSQC (heteronuclear single quantum coherence) experiment that correlates the amide proton chemical shift with that of its directly attached nitrogen atom. This ^{15}N-$^{1}H_N$ HSQC spectrum is sometimes referred to as a protein "fingerprint spectrum" since each protein gives rise to a unique pattern of peaks. Backbone assignment experiments are 3D versions of this experiment that provide correlations between this amide proton-nitrogen pair and backbone carbon atoms. For example, the HNCO experiment connects the amide proton and nitrogen chemical shifts with the shift of the carbonyl carbon from the preceding residue. A typical suite of backbone assignment spectra usually include this experiment, along with the HNCACB, and CBCA(CO)NH experiments (22, 24). Together these spectra can be linked together to identify series of peaks arising from sequential residues which ultimately should allow for identification of >95% of the backbone chemical shifts. A similar type of process is then followed to assign sidechain chemical shifts, often using experiments that give through-bond correlations (e.g. COSY and TOCSY spectra). Given the larger number of atoms and the higher degree of overlap for sidechain versus backbone atoms, the assignment process is much more time-consuming for the side-chain atoms.

Once chemical shift assignments have been completed then it is necessary to record experiments that can provide longer-range structural information, since this is the primary

data used to define the protein fold. These are nuclear Overhauser enhancements (NOEs), secondary shifts, and residual dipolar couplings (RDCs), as described in the following sections.

1.2.1 Nuclear Overhauser Effect (NOE) – Distance restraints

The main source of geometric information used in NMR protein-structure determination lies in the nuclear Overhauser effect (25), which is due to the local field experienced by a spin with neighboring magnetic nuclei. NOEs are related to the "through-space" distance between a pair of atoms that are either linked by covalent bonds (intra-molecular NOE) or not at all (inter-molecular NOE). Depending on the relative orientation of the two nuclei to one another, the field generated by a neighboring spin will either reinforce or diminish the effect of the external magnetic field, with the time-dependent fluctuation produced by the rotational motion of the molecule in which these nuclei are situated. The magnitude of this effect depends on the inter-nuclear separation distance, r, which is proportional to the inverse sixth power as given by Equation 1.

$$NOE = \frac{\mu_o^{\ 2}}{16\pi^2}\hbar^2\gamma_I^4 r^{-6}\{-J(0)+6J(2\omega_I)\} \qquad [1]$$

where ω_I is the resonance frequency of spin I, $J(\omega)$ is the spectral density function at frequency ω, and μ_o is the permittivity constant. The NOE represents the change in intensity of one resonance when the transitions of another spin are perturbed from their equilibrium populations. The dependence of the NOE on specific spectral density function terms means that the sign and magnitude will depend on the size of the molecule being probed. For large molecules that tumble slowly in solution, the spectral density function is

large at small frequencies (i.e. $J(0)$ is large) and therefore efficient magnetization transfer can occur via this effect. Proteins can be considered large molecules in this respect, and as a result, it is possible to observe NOE correlations between protons that are approximately <5 Å apart in proteins. In protein NMR spectroscopy, NOEs between protons are the primary source of structural information.

For protein structure determination, NOE correlations are obtained from NOESY spectra that must be assigned to identify those atoms that are close together in space. Assignments are made by matching the peak positions in the NOESY spectrum with chemical shifts from the list of assigned atoms obtained from the backbone and sidechain experiments. In addition, peak intensities in the NOESY spectrum can be related to the distance separating the two atoms since protons that are closer together will give rise to a more intense NOE cross-peak. However, these peak intensities cannot be used to determine exact distances since contributions can also arise from spin diffusion. Specifically, magnetization transfer between two spins through multiple short steps may be more efficient than a one-step transfer over the longer, direct distance (Figure 1.6). Since the density of protons in the interior of proteins is relatively high, the distance estimate is often about 20% smaller than what is expected experimentally (20) and as a result, NOE intensities are converted to approximate inter-proton distances for the purpose of structure calculations. In this approach NOE intensities are classified as weak, medium or strong and then assumed to correspond to internuclear separation distances that are long, intermediate or short, respectively. Although the specific values that are used for these distance bins tend to vary, typical values used as conservative estimates are 1.8-3.0 Å, 1.8-4.0 Å and 1.8-5.0 Å for strong, medium and weak NOEs respectively.

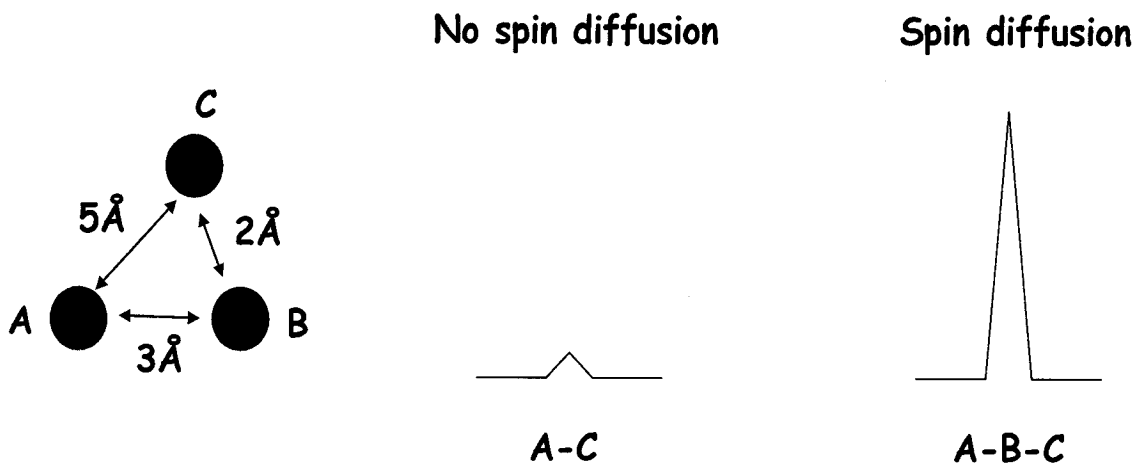

Figure 1.6 – Schematic representation of the effect of spin diffusion on NOE intensity. Due to the r^{-6} dependency, the NOE between spin A and C would be more intense if it occurred through spin diffusion (i.e. A to B to C) compared to the intensity of the NOE without spin diffusion (i.e. A to C directly). Consequently in the presence of spin diffusion the internuclear distance estimate would be much smaller than the actual distance in the molecule.

1.2.2 Dihedral Angle restraints

Backbone dihedral angle ϕ and ψ values for a protein (Figure 1.7) can be indirectly estimated from the chemical shifts of backbone nuclei that are assigned during the initial stages of the structure determination process. In particular, a strong correlation exists between $^{1}H\alpha$ chemical shifts and secondary structure (26, 27). When ^{13}C isotopic enrichment is also available a wider range of secondary shifts can be determined, improving the predictive power of this measurement. This is generally done by calculating the random coil chemical shift value (determined by residue type) and subtracting this from the observed chemical shift to obtain the secondary chemical shift as originally described by Wishart *et al.* (28-30). Since this secondary chemical shift depends heavily upon the type of secondary structure it can provide an accurate prediction of the secondary structure.

For example, both $^{13}C\alpha$ and ^{13}CO atoms tend to experience an upfield shift in α-helices and a downfield shift in β-strands. For these atoms observation of positive secondary chemical shift for both atoms would provide a strong indication of α-helical structure while negative shifts would suggest β-stand structure (29).

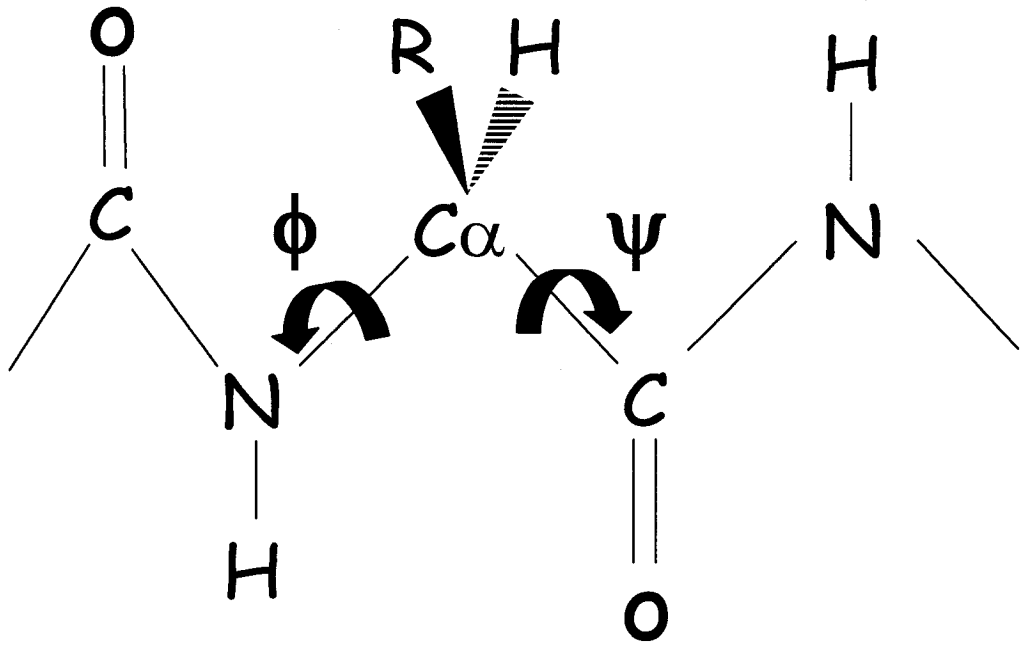

Figure 1.7 – Polypeptide backbone ϕ and ψ dihedral angles. Rotations about the H^N-N – Cα - Hα and Hα - Cα – C - O bonds can be described by phi (ϕ) and psi (ψ), respectively. By convention, both ϕ and ψ are defined as 180° when the polypeptide is in its fully extended conformation and all peptide groups are in the same plane.

Once the secondary chemical shifts have been used to predict the secondary structure, the dihedral angles ϕ and ψ can be defined. For an α-helix, the values are typically in the range of ϕ = -57° and ψ = -47° while for β-sheets they are ϕ = 120° and ψ = -120° (31). However, a range of dihedral angle values can be accommodated in these types of secondary structures. The range of dihedral angle values that can be adopted by the

protein backbone also depends on steric effects which reduce the allowable degrees of phi and psi to favored regions of the Ramachandran plot (Figure 1.8). For these reasons, a range of dihedral angle values are usually considered to be equally valid for a given secondary structure prediction. Typically for α-helices ϕ is allowed to vary by 20° and ψ by 30°, although wider bounds can also be used. It is also possible to more precisely predict these dihedral angle values through an empirical approach that uses a combination of five kinds (Hα, Cα, Cβ, CO, N) of chemical shift assignments for comparison with database values collected for a suite of proteins of known structure (31). This strategy has been implemented in a freely available software package called TALOS that has been widely used to generate dihedral angle restraints for protein structure determination.

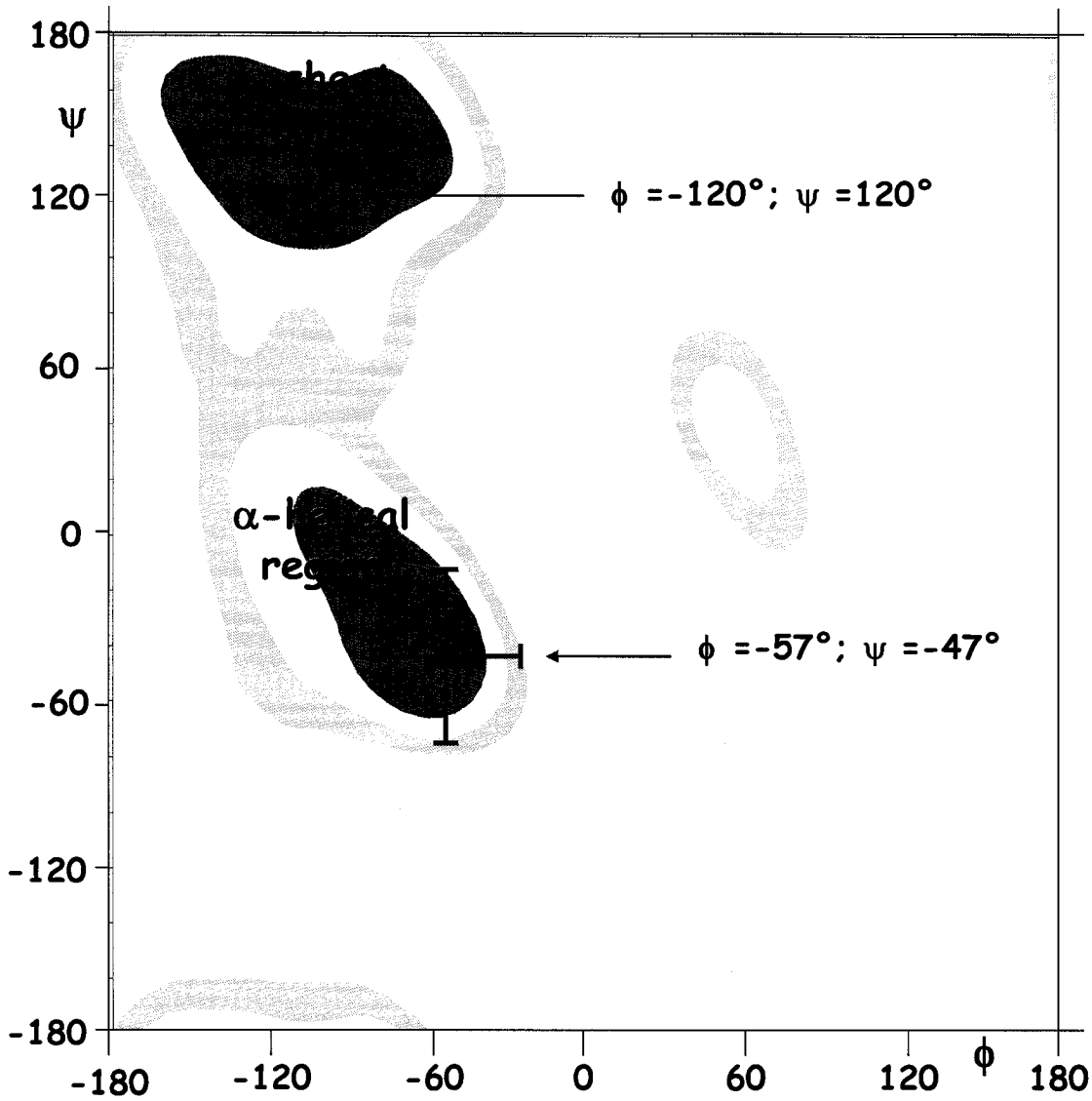

Figure 1.8 – The Ramachandran plot for proteins showing accessible backbone torsion angles. Most energetically favored regions are highlighted in green and with less favorable but permissible areas in yellow and pink. The canonical α-helix has dihedral angles corresponding to ϕ = -57° and ψ = -47° with +/-20° and +/-30° boundaries respectively (shown as error bars) which are mainly used for structure calculations as shown by a red dot. Similarly, an ideal β-sheets has ϕ = -120° and ψ = 120°.

1.2.3 *Dipolar Couplings*

Dipolar couplings are a through-space interaction that arises between any two magnetically active nuclei (i.e. ^{1}H, ^{15}N, ^{13}C) where the magnetic flux from one-spin influences the effective magnetic field and thus the resonance frequency of another spin. They provide orientational information, both short range and long range, which is dependent on the inter-nuclear vector orientation relative to the external magnetic field. They are complementary to the NOE, since the NOE only provides distance information for two nuclei and does not provide any information on how they might be related to another pair of nuclei that are distant in the molecule. In contrast, dipolar couplings are defined in terms of a three-axis reference alignment frame that provides orientational information on each internuclear vector (Figure 1.9) and consequently this information does not depend on the translational distance between two dipole vectors (32, 33).

The magnitude of a static dipolar coupling interaction between two atoms A and B is given by:

$$D_{AB} = -\frac{\mu_o \gamma_A \gamma_B h}{(2\pi r)^3}\left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle \qquad [2]$$

where μ_o is the permittivity of free space, γ is the gyromagnetic ratio, h is Planck's constant, θ is the angle between the magnetic field and the internuclear vector and r is the distance between interacting spins. If the dipolar coupling is measured between directly bonded nuclei (e.g. ^{1}H-^{15}N) the distance between the two interacting spins is set to a fixed value (34), and θ becomes the primary variable. However, the angle brackets indicate that the angular component of the measurement is averaged over the ensemble, leading to an

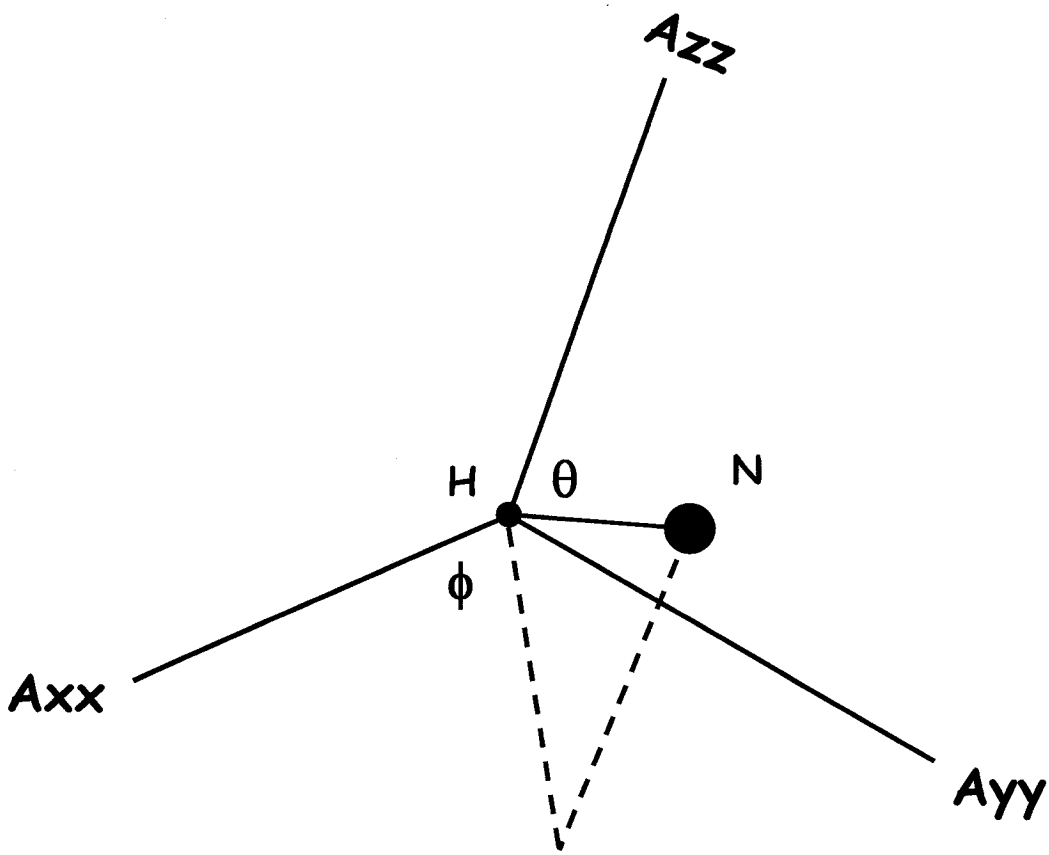

Figure 1.9 – Schematic diagram of the parameters used to calculate the dipolar coupling from equation 3. The ^{1}H-^{15}N internuclear vector orientation can be defined with polar angles θ and ϕ with respect to an alignment tensor with unit axes Axx, Ayy and Azz. The orientation of the alignment tensor with respect to the molecular frame is given by Euler angles (α, β, γ) and the magnitude of the tensor by Aa and the asymmetry about the long axis by R.

average dipolar coupling of zero for an isotropic distribution of vectors (32, 35, 36). For this reason in the solution phase it is necessary to include anisotropic elements (e.g. stretched polyacrylamide gels, liquid crystalline media, *etc.*) in the sample (37-39) to induce a small degree of alignment in the solute molecules such that this term will not average to zero. These residual dipolar contributions add to through-bond (J) couplings and hence D_{AB} can be measured as the difference in couplings obtained under isotropic conditions (J) and anisotropic conditions (J+D) (Figure 1.10).

A set of residual dipolar couplings (RDCs) can be obtained for each residue in a protein and can be used as NMR restraints to improve the quality of structures determined by NMR (40-43). During this process it is more practical to cast the dipolar coupling interaction in the frame of an alignment tensor according to:

$$D_{AB} = A_a^{AB}[(3\cos^2\theta - 1) + \frac{3}{2}R(\sin^2\theta\cos 2\phi)] \qquad [3]$$

where $A_a^{AB} = -\left(\frac{\mu_o h}{16\pi^3}\right) S\gamma_A\gamma_B\left\langle r_{AB}^{-3}\right\rangle A_a$ [4],

A_a^{AB} and R are the axial and rhombic components respectively, of the alignment tensor, S is an order parameter, γ is the gyromagnetic ratio of the nuclei, r_{AB} is the time averaged internuclear distance, and θ and ϕ are the polar angles describing the orientation of the internuclear vector with respect to the alignment tensor axes (Figure 1.9). When the alignment tensor parameters are known this equation can be used to calculate a theoretical set of dipolar couplings for a given structure which can then be compared to experimental values to help assess the accuracy of the structure.

Although RDCs have helped reduce the reliance on NOEs in structure determination of proteins (40, 44-46) there is degeneracy in the solutions that can be obtained with RDCs that can reduce their utility in protein structure determination. Specifically, there are eight possible orientations for the alignment tensor that are consistent with the measured values since the angular terms in Equation 2 allow an inversion of θ,ϕ with respect to any of the axes of the molecular alignment tensor to result in the same RDC value. Measuring RDCs in a second alignment medium that imposes a different alignment tensor from the first can reduce this degeneracy (47). However, this is not always

necessary because, in some cases, only one of the possible orientations is consistent with NOE or other structural data.

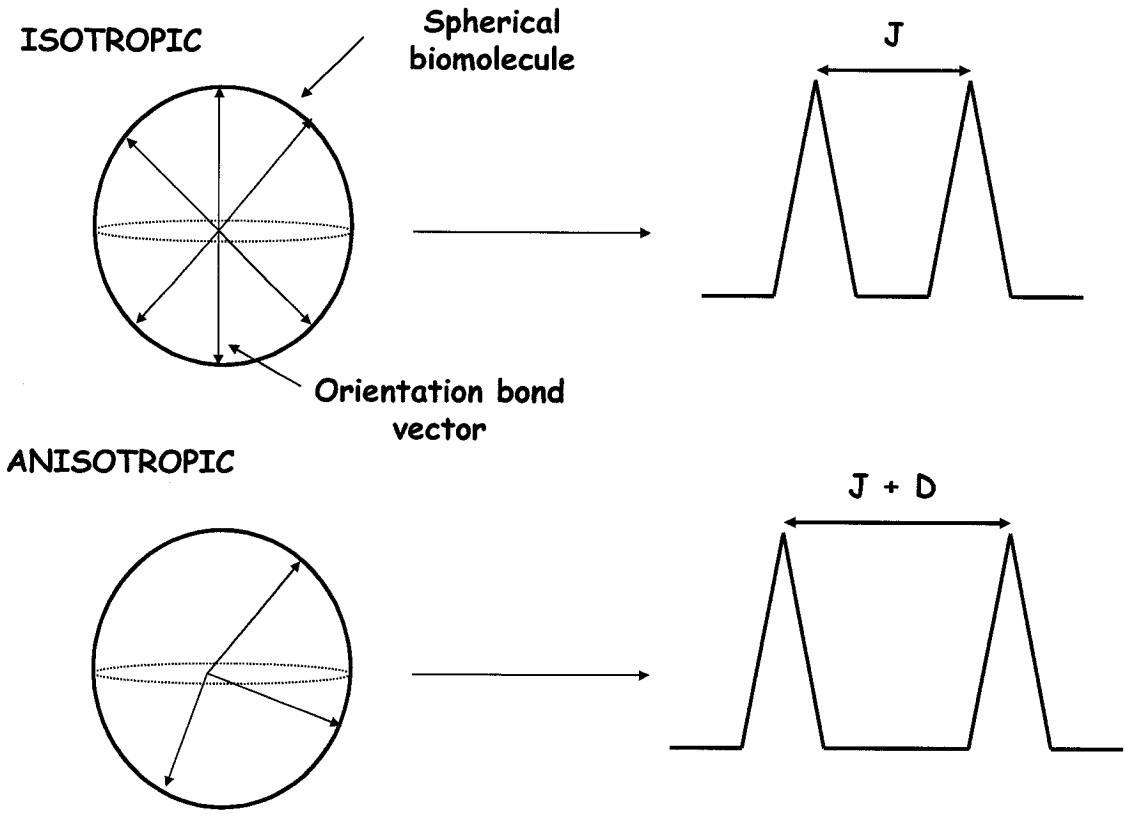

Figure 1.10 – Schematic representation of how an RDC is obtained and measured by solution NMR. In isotropic solution, all bond vector orientations are randomly oriented and $(3\cos^2\theta - 1)$ averages to zero. The coupled 1D spectrum of a directly bonded pair of spins (e.g. ^{1}H-^{15}N) would be split into two peaks separated by the J-couplings. In anisotropic conditions the bond vectors orient themselves in a preferred direction relative to the magnetic field. In this case, the angular term will not average to zero and the peak splitting will include the dipolar coupling as well as the J-coupling interaction. Subtraction of the two peak splittings will give the numerical value for the dipolar coupling.

1.3 Structure calculations via simulated annealing

Once all the experimental restraints have been acquired they can be used as input in a simulated annealing protocol for structure determination. The ultimate goal of structure calculation algorithms is to optimize agreement between an atomic model and the observed data as well as *a priori* chemical information (*e.g.* bond lengths, bond angles, and non-bonded interactions) (48). Most algorithms used for structure calculation aim to find the global minimum of an energy function E_{target}:

$$E_{target} = E_{chem} + E_{nmr} \quad [5]$$

where $E_{chem} = E_{bonds} + E_{angles} + E_{non-bonded}$ [6]

and $E_{nmr} = E_{NOE} + E_{dihedral} + E_{RDC}$ [7]

E_{chem} contains energy terms for covalent bonds, bond angles, and non-bonded repulsion, which include the electrostatic repulsion and van der Waals interactions, while E_{nmr} is comprised of energy terms describing the experimentally determined structural restraints. For experimental restraints these energy terms are usually described by a flat-bottomed parabolic function (Figure 1.11) where a range of values are accepted without penalty. Specifically, distance, angle, and orientational data are provided with a lower (L) and upper (U) limit that reflects the fact that they cannot be precisely determined from NMR experiments. If a structure is determined such that the calculated value for this restraint is outside of this range, then the energy of the system increases exponentially. The shape of these energy functions are typically set so that NOE violations over 0.5 Å, dihedral angle violations over 5°, and RDC violations over 0.5 Hz are usually not tolerated.

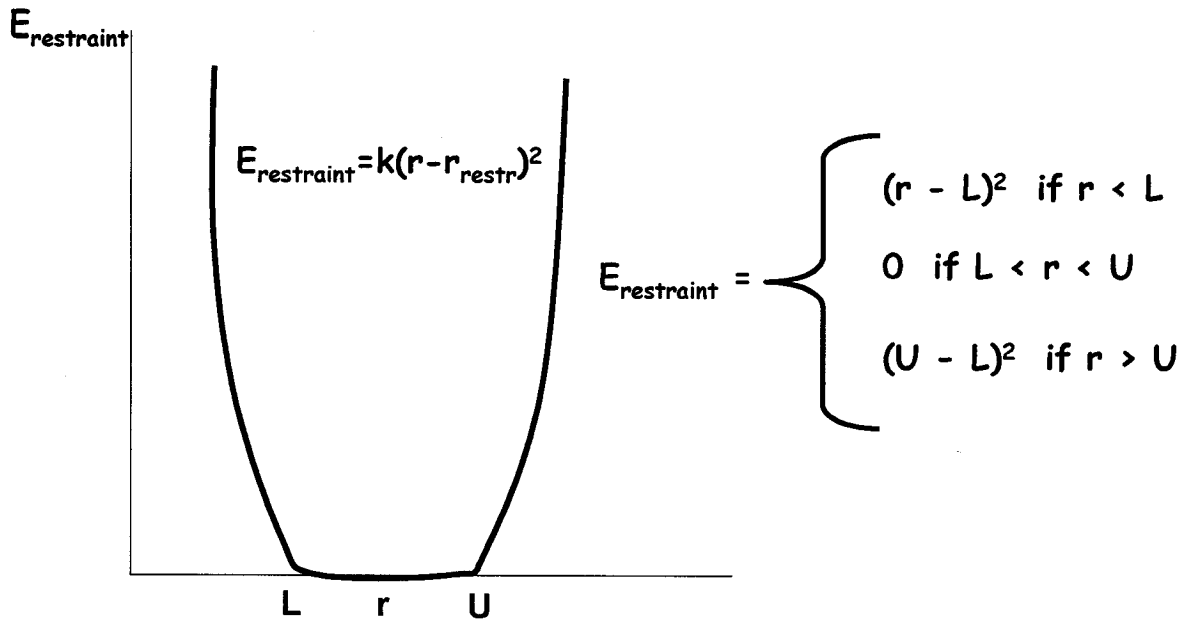

Figure 1.11 – Schematic representation of the flat-bottom potential energy well used in structure calculations. A pseudo-potential energy function is used for NMR-derived structural restraints (r_{rest} in this example) so that a penalty is incurred when the value of this parameter in the model (r) does not agree with the experimentally determined value. Usually a range of values is accepted (L<r<U in this example) where the pseudo-energy term would be zero. Any conformation that gives rise to a value of r that is less than L or greater than U would add energy to the system. The force constant (k) must not be too large or physically unreasonable structures that deviate from ideal geometry could result.

During the process of structure determination, an extended polypeptide chain is typically used as a starting structure, and the atomic positions are adjusted until the energy of the system reaches a minimum. During this process the conformation of the polypeptide chain is adjusted using molecular dynamics (MD) simulations where the acceleration of atoms is determined by Newton's equation and the velocity from Maxwell's distribution for a given temperature. However, the large number of variables that contribute to the total energy of the system gives rise to the so-called 'multiple minima problem' where the target function contains many local energy minima in addition to the global minimum (Figure 1.12). For this reason, simulated annealing (SA) approaches are typically used, allowing a

larger range of conformational space to be sampled (49-51). This method simulates the heating of the system so that there is enough energy to overcome these energy barriers and the system allowed to evolve according to the potential energy function defined by the empirical and *a priori* energy terms. The temperature of the system is then lowered in a number of steps (usually ~3000) and MD used to find the energy minimum at each temperature step. The result of this procedure should be a structure that represents the global energy minimum of the system. This overall procedure is repeated a number of times (usually ~100) and the 10 lowest energy structures are taken to represent the ensemble of NMR structures.

Once the lowest energy structures (~10 structures) are obtained from the simulation, a variety of statistical analyses are performed to evaluate the structural quality. The most common calculation is of the root mean square deviation of the ensemble (precision) or to a previously determined structure (accuracy). Typically, high quality structures have a large number of restraints (>20 NOEs per residue and two backbone dihedral angles per residue), small deviations from ideal geometry (< 1Å accuracy) and good agreement with restraints (no violations).

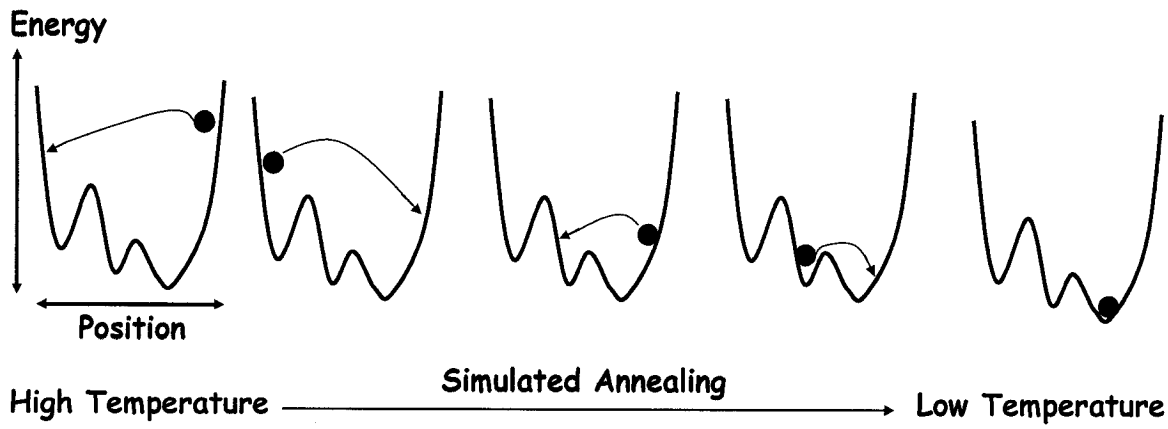

Figure 1.12 – Schematic representation of simulated annealing. A biomolecule (shown as the blue dot) starts in a high energy conformation (e.g. an unfolded extended state) and high temperature. Molecular dynamics is carried out as the temperature is reduced, allowing the conformation to access lower energy configurations. Although the biomolecule may sample shallow energy wells (local energy minima) during this process, the temperature is high enough to overcome small barriers such that it can reach the global minimum.

1.4 Relaxation and NMR Spectral Quality

Successful determination of protein structures using solution NMR hinges on the ability to acquire high quality spectra. This is a function of sample sensitivity and resolution where intense, well-separated, narrow peaks would comprise the ideal spectrum. Among all the different factors that affect spectral quality, the size of the molecule being analyzed is one of the most important since it is related to transverse (T_2) relaxation, a factor that is inversely proportional to peak width. Faster transverse relaxation leads to rapid decay of the NMR signal, decreasing sensitivity and resolution. The transverse relaxation time is determined by the spectral density function $J(\omega)$ at specific frequencies since:

$$\frac{1}{T_2} = \frac{3\gamma^4\hbar^2}{20r^6}[3J(0) + 5J(\omega_o) + J(2\omega_o)] \qquad [8]$$

where γ is gyromagnetic ratio, $\hbar$ is Planck's constant, r is the distance between the spins. The spectral density function describes the power available within a molecular system to induce transitions depending on the overall motion of the macromolecule as a whole and can be described by:

$$J(\omega) = \frac{2}{5}\frac{S^2\tau_c}{1+(\omega\tau_c)^2} \qquad [9]$$

where S is the order parameter which measures the degree of intramolecular motion, and τ_c is the rotational correlation time of the molecule. For a spherical molecule, τ_c can be approximated using the Debye equation (11).

$$\tau_c = \frac{4\pi\eta r_H^3}{3k_bT} \quad [9]$$

where V is the volume of the molecule, η is the viscosity of the medium, k_B is the Boltzmann constant, T is the temperature, and r_H is the hydrodynamic radius of the protein. According to this relationship, larger molecules will have longer rotational correlation times, which reflects the fact that they tumble more slowly in solution. According to Equation 9, a larger proportion of this population of slowly tumbling molecules will oscillate at transition-inducing frequencies, specifically those at and close to zero corresponding to large values of $J(0)$. Since T_2 is inversely proportional to $J(0)$, larger molecules will have a shorter T_2, and therefore a faster loss of a signal and broader line width. This rapid transverse relaxation will reduce both signal intensity and resolution and may lead to signal overlap in the NMR spectrum.

Rapid transverse relaxation is a particular issue for solution NMR studies of membrane proteins, since detergent must be used to maintain the hydrophobic protein in solution. The result is a membrane protein-detergent micelle complex, usually of high molecular mass with long correlation times. There have been studies that use a mixture of organic solvents to study membrane protein structures (10, 52), but these solvents are generally poor at preserving tertiary structure in membrane proteins. Any stable secondary structure obtained in this environment cannot be assumed to reflect native-like structure since these solvents are generally poor mimics of the lipid bilayer environment (53). The detergent micelle is a better mimic for the lipid bilayer since it has both a hydrophilic and hydrophobic domain that is capable of preserving the active form of a membrane protein. However, the added mass of the detergent micelle to the membrane protein presents a challenge due to extensive line broadening, making it much harder to measure and assign these resonances (Figure 1.13).

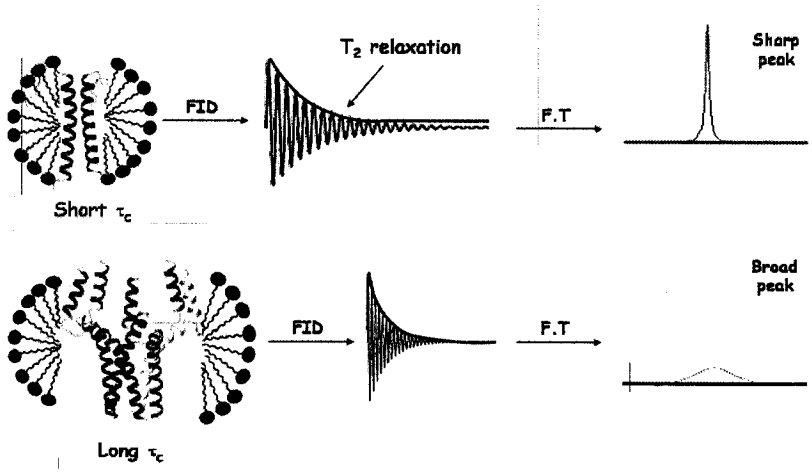

Figure 1.13 – Schematic diagram representing the impact of protein size in a solution NMR study. Smaller protein-detergent micelle complexes (<20kDa) have more rapid rotational correlation times in solution that leads to longer T_2 relaxation times which produces a longer-lived free induction decay (FID). Fourier tranformation of the FID produces a frequency-domain spectrum with relatively narrow peaks. Larger complexes tumble more slowly, giving rise to longer rotational correlation times, ultimately leading to extensive line broadening in the NMR spectrum.

1.5 Progress to date in solution NMR of membrane proteins

In spite of the difficulties associated with solution NMR of membrane proteins, tremendous progress has been made in the past few years, enabling NMR-based structure determination of several β-barrel membrane proteins as well as α-helical membrane proteins (Figure 1.14). Development of transverse relaxation optimized NMR methods (TROSY) allowed structures of several small β-barrel proteins (OmpA, OmpX, and PagP) to be determined under micellar conditions in which the aggregate molecular weights were in the range of 50-60 kDa (13-15, 17, 18). For example, the structure of OmpA in DPC micelles could be determined from only 91 NOE restraints (49 $^{1}H_{N}$-$^{1}H_{N}$ and 42 $^{1}H_{N}$-$^{1}H_{\alpha}$ obtained from a 3D HN-NOESY-TROSY spectrum) and 142 dihedral angle restraints from secondary shift calculations, giving rise to an ensemble with backbone rmsd 1.19+/-0.29 over 82 residues (14). Improvements to this structure were also made through measurement of RDCs in micelle samples aligned in strained polyacrylamide gels (19).

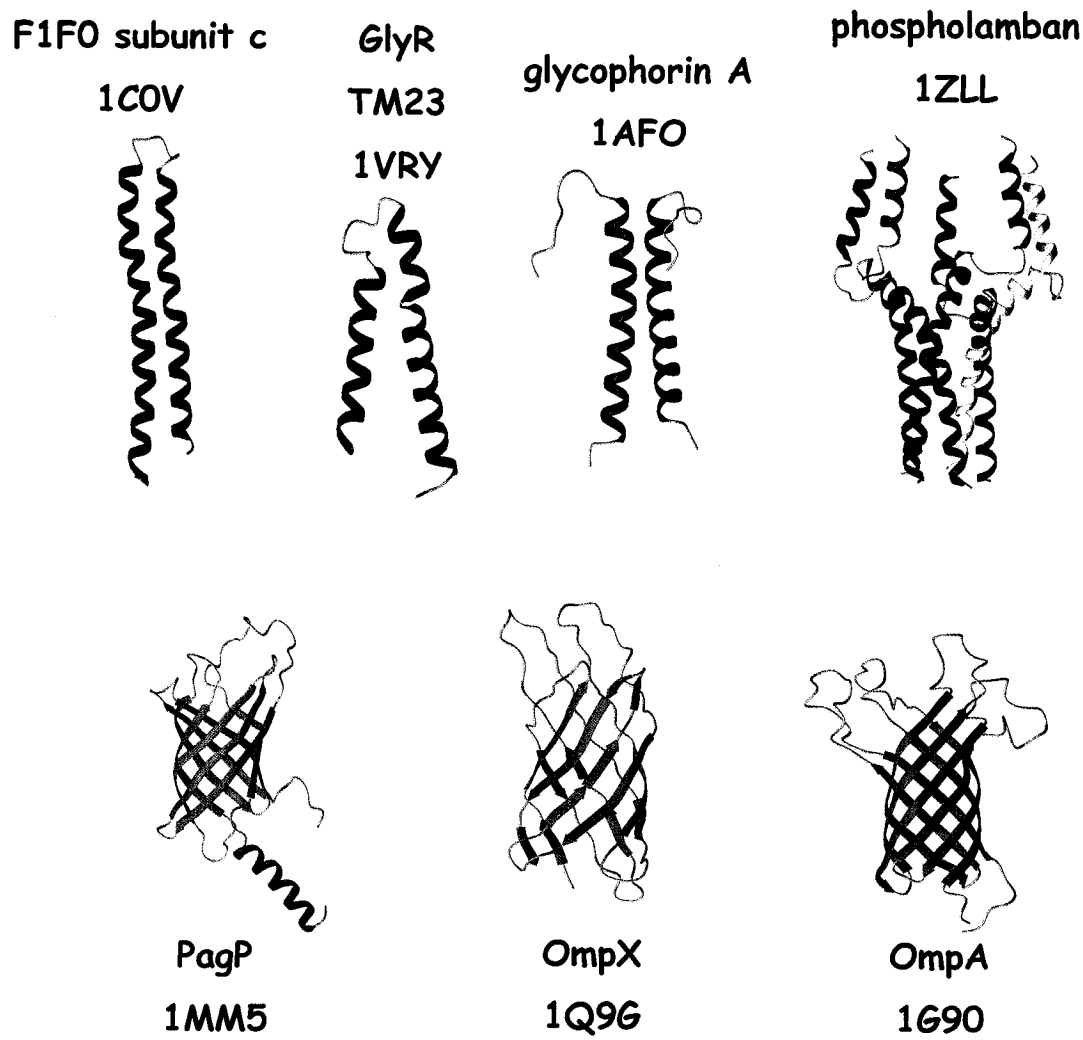

Figure 1.14 – Ribbon representation of all polytopic membrane proteins solved by solution NMR to date. Two of the α-helical membrane proteins: F1F0 subunit and the glycine receptor (GlyR) were determined in organic solvents instead of detergent micelles. All other proteins were solved under detergent micelle conditions (SDS, dodecylphosphocholine or β-octyl glucoside). PDB accession codes are indicated below each structure

Although solution NMR has made impressive gains in the study of β-barrel types of membrane proteins, it is unlikely that such sparse NOE datasets would allow global folds to be determined to a similar degree of resolution for α-helical membrane proteins. This class of membrane proteins often tends to exhibit narrower chemical shift dispersions since the

transmembrane sections of these proteins consist of predominantly of hydrophobic amino acids and very few charged amino acids. In addition, amide-amide NOEs are a rich source of long range distance information for β-barrel proteins, but occur very rarely between helices in polytopic helical membrane proteins. Hence, unlike the case for β-barrel proteins, side chain protons from polytopic α-helical proteins must also be assigned in order to obtain long-range information required to define the tertiary fold. This is a particularly challenging task for larger proteins due to the poor NMR spectral quality and large number of resonances to assign.

To date, most of the solved structures of α-helical membrane proteins are monotopic and hence their biological relevance has been limited. To truly understand the function and mechanism of biological process that occur in the membrane at a molecular level of detail, more polytopic α-helical membrane protein structures must be determined. There are only four α-helical membrane proteins solved by solution NMR that provide insight into inter-helical interactions (Figure 1.14), two of which were determined in the more native-like micelle environment (54, 55). However, larger single-chain polytopic α-helical membrane protein structures have yet to be determined by solution NMR. Given the large anticipated size of the detergent-protein complex, it is clear that additional strategies that can address the problems associated with the large size must be employed to make progress in this field. One of the approaches that have been used for the study of large water soluble proteins that may be useful in this application is selective protonation of deuterated proteins.

1.6 Protein deuteration and methyl protonation

The added mass of detergent (e.g. SDS 18kDa, DPC 25-28kDa, LPPC 92kDa) (116) in protein-detergent complexes means that smaller membrane proteins behave like large molecules, such that rapid relaxation becomes a major obstacle. Fully protonated proteins suffer from efficient ^{1}H-^{1}H and heteronuclear ^{1}H -X (X – ^{15}N and ^{13}C) dipolar spin relaxation pathways. This leads to decreased magnetization transfer efficiency resulting in a loss of both sensitivity and resolution that can be quite serious for proteins with high molecular weights (56). However, replacing these protons with deuterons can eliminate many of the relaxation pathways due to the 6.7-fold lower gyromagnetic ratio of ^{2}H relative to ^{1}H. Therefore, complete, or even partial proton replacement leads to significant reduction of transverse relaxation rates. For ^{13}C atoms that have directly attached protons, deuteration removes the strong dipolar interaction of the attached proton. For example, estimates of dipolar-dipolar and CSA relaxation rates for amide protons in proteins with deuteration of the Cα carbon results a tenfold decrease in transverse relaxation rates can be acheived for proteins with long correlation times (57). Deuteration therefore dramatically enhances signal-to-noise ratios in protein NMR spectra by increasing T_2 for ^{13}C and ^{1}H, thereby narrowing the peaks in the NMR spectrum.

The application of deuteration to protein NMR was first demonstrated by Crespi and Jardetsky who used selective deuteration to reduce the complexity of one-dimensional (1D) ^{1}H spectra (58, 59). Since these experiments, random incorporation of ^{2}H into non-exchangeable sites has been used to obtain well-defined NMR structures for proteins as large as 30 kDa (57). However, there is a trade-off for NMR experiments that involve aliphatic protons, since an increasing percentage of deuteration will improve relaxation

rates, but will also reduce the number of protons being used for signal observation (60, 61). These competing effects can hinder assignment of side-chain resonances as well as the ability to detect NOEs. In the extreme case, where all non-exchangeable sites are deuterated, NOE observations are limited to $^{1}H_{N}$-$^{1}H_{N}$ NOEs. The sparse proton density in these samples leads to relatively few long-range NOEs and, as a consequence, to poorly resolved structures (62).

With this problem in mind, Rosen *et al.* (63) developed a protocol for the production of ^{15}N, ^{13}C, ^{2}H labeled proteins with protonation of the methyl groups of alanine, valine, leucine, and isoleucine ($\gamma 2$ only). These are attractive targets for selective protonation as methyl groups since they are enriched in protein hydrophobic cores (64). In addition, protonation of the methyl groups give reasonably well resolved ^{13}C-^{1}H correlations due to the rapid rotation about the methyl symmetry axis (65). A more cost-effective method was subsequently developed that allowed for the protonation of leucine, valine and isoleucine (only $\delta 1$) groups (66). This method reintroduces protons in a perdeuterated protein by growing *E. coli* that is expressing the protein of interest in D_2O minimal media with selectively labeled α-ketoisovalerate and α-ketobutyrate provided as a supplement to ^{13}C, ^{2}H-glucose, and $^{15}NH_4Cl$ (66) (Figure 1.15). High levels of protonation are obtained for these methyl groups such that the isotopomers $^{13}CH_2D$ and $^{13}CHD_2$ are not produced, preventing the appearance of extraneous peaks in the NMR spectrum (63).

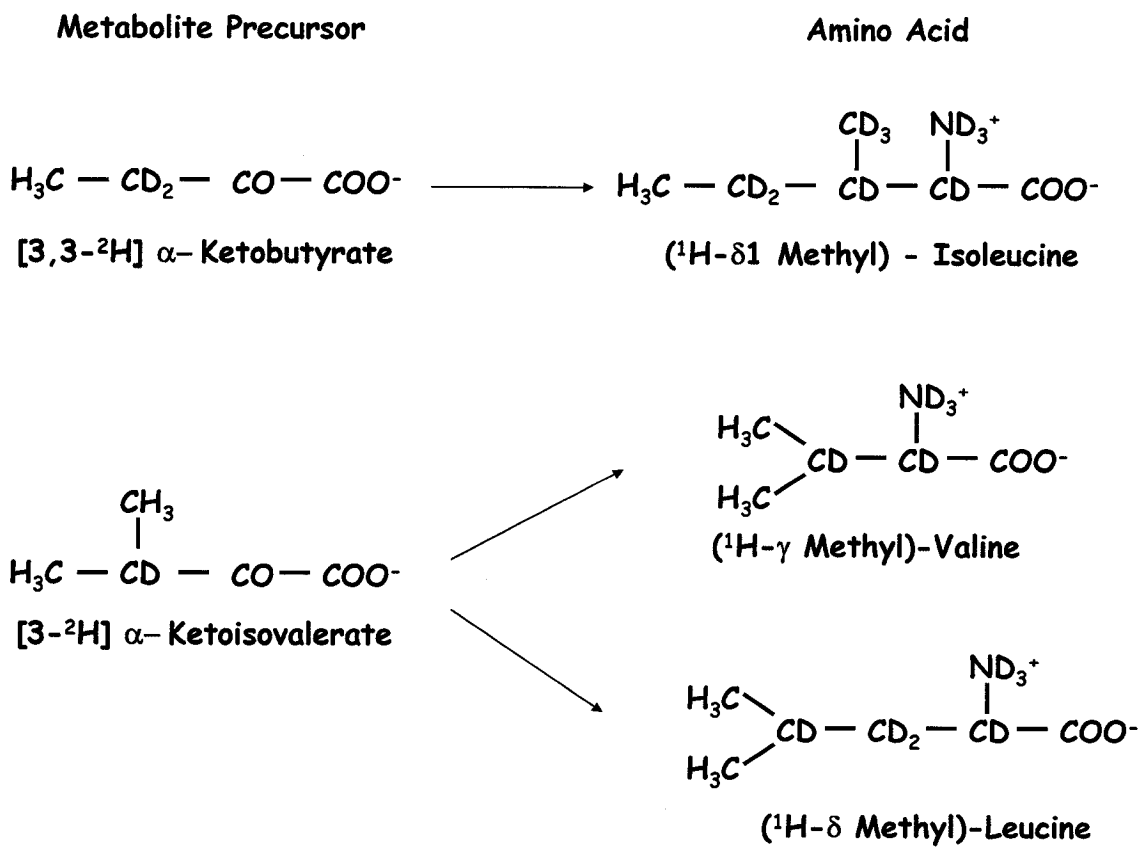

Figure 1.15 – Chemical structures of metabolites involved in selective methyl isotope labeling strategies. Selectively methyl-protonated metabolites on the left are used to supplement *E.coli* minimal media made with D_2O and ^{13}C, ^{2}H-glucose to give rise to proteins containing the selectively protonated amino acids on the right.

This selective methyl protonation labeling strategy was successfully employed in the structure determination of the β-barrel membrane protein OmpX (18). A large number of additional NOEs relative to the perdeuterated sample (526 vs. 107) were derived and assigned from their NMR spectra. The additional methyl NOE-derived distance restraints improved the backbone rmsd from 3.11 to 1.42 Å. In addition, this labeling strategy has been incorporated to determine the structures of large α-helical water-soluble proteins by solution NMR (63, 67-69). However, it was found that this class of water-soluble protein

was the most challenging to accurately define using this labeling approach, with backbone rmsds of >5 Å sometimes being obtained (63, 69) . Since this type of fold is most similar to that of the polytopic helical membrane protein, these results suggest that a similar quality of structure would result if this strategy was applied to helical membrane proteins. However, these membrane proteins have a higher proportion of these hydrophobic amino acids relative to water soluble proteins (70) raising the possibility that a higher degree of accuracy may be accessible for these proteins. In addition, the anisotropy of the lipid bilayer environment makes it less likely for helices to end in the interior of the lipid bilayer, potentially leading to simpler folds that can be more accurately determined. However, no study has yet been undertaken to assess the utility of this approach for helical membrane protein structures. Considering the expense associated with the production of these types of methyl protonated samples it would be particularly useful to have some theoretical measure of the range of accuracies that can be obtained in this case.

1.7 Thesis Objectives

The primary objective of this thesis is to evaluate the utility of the selective methyl labeling strategy for structure determination of α-helical membrane proteins by solution NMR. For this purpose I explored a range of structure calculation strategies using simulated NMR data from membrane proteins of known structure. In Chapter 3, I present results for a set of membrane proteins that mainly consist of pairs or bundles of hydrophobic α-helices. I analyzed the impact of intra- and inter-helical NOEs on structure quality for samples specifically labeled with protons only at the methyl and amide positions

as described above. In addition, I also explored the role of dihedral angles and residual dipolar couplings in an effort to find general trends to improve the accuracy of polytopic α-helical membrane proteins. Based on the results of these calculations I have developed a set of recommendations to consider for future structure determinations of polytopic α-helical membrane of unknown structures.

Chapter 2

MATERIALS AND METHODS

2.1 Membrane protein dataset

Five different known membrane protein structures were used to generate simulated NMR structural data sets: the dimeric transmembrane domain of human glycophorin A (GpA) (55); the second and third transmembrane helices of human glycine receptor (GlyR) (71); the glycerol-conducting facilitator channel from *E.coli* (GlpF) (72); the light-driven chloride pump halorhodopsin (Halor) (73) and the mitochondrial ADP/ATP carrier (AMC) (74). For the NMR structures GpA and GlyR, an ensemble of 20 structures were obtained from the Protein Data Bank. The first model from each ensemble was chosen as the representative structural model for this study. All X-ray structures chosen for this study had all protons added into the structure using Molmol (75).

2.2 Generation of restraints: NOEs, Dihedral Angles, RDCs

NOE restraints were simulated for H^N-H^N, H^N-CH_3, and CH_3-CH_3 pairs in α-helical regions for each membrane protein structure. Following typical observable distances from NOESY experiments (76) only H^N-H^N distances less than 5.0Å and H^N-CH_3, and CH_3-CH_3 distances less than 6.0Å were taken as NOE restraints. In order to simulate the selective methyl labeling strategy as described in Chapter 1 (66), distances involving Val, Leu, and

Ile (δ1 only) methyl protons were measured to their pseudo-atoms (for nomenclature see (77)). Distance restraints were categorized as strong (1.8-3.0 Å), medium (3.0-4.0 Å), and weak (4.0-5.0 Å or 6.0 Å for methyl protons) and then each bin was set to a target value with upper and lower limits, following a similar study for soluble proteins (62). Specifically distance restraints of 1.8 - 4.0 Å, 2.8 - 5.2 Å, and 3.2 - 6.7 Å for strong, medium and weak bins were used. Only inter-residue NOE restraints were included and the list was generated following the format required by the structure calculation protein CNS with default energy constants (78).

Backbone dihedral restraints for α-helices were used in all the simulations with canonical helix values of $\phi = -57^\circ$ and $\psi = -47^\circ$, similar to values used in previous studies (62, 79-81). (ϕ, ψ) dihedral angle ranges of either (+/-2.5°, +/-5.0°) for strong restraints or (+/-20°, +/-30°) for medium strength restraints were used, along with the default energy constant for CNS (78).

The set of ^{1}H-^{15}N RDCs were generated only for α-helical regions by using Equation 3 as described previously (33, 82) with alignment tensor parameters obtained for each structure from PALES (83) using a simulated bicelle concentration of 0.05%. All program scripts used to create NMR restraint lists are given in the Appendix.

Table 2.1 - Summary of membrane protein structures used in simulations

Protein	Type of structure	# of TM segments	# of helical residues	Max no. of NOEs	
				Intra-helical	Inter-helical
GpA	NMR	1 (dimer)	54	140	24
GlyR	NMR	2	47	199	9
AMC	X-ray (2.2Å)	6	168	641	99
Halor	X-ray (1.8Å)	7	188	758	195
GlpF	X-ray (2.2Å)	8	181	638	145

2.3 Structure calculations

Using the simulated data for each membrane protein, structures were calculated using a torsion-angle molecular dynamics simulated annealing protocol implemented in the program CNS v1.1 (51) on a Dell workstation with an Intel Pentium IV processor containing 1 GB of memory and a 1.2 GHz clock speed. All calculations were performed starting from extended structures and using slow cooling temperatures, variable molecular time steps, and variable cooling steps as summarized in Table 2.2. In most cases default CNS parameters were employed, but were adjusted for larger proteins to ensure that convergence was obtained. An example of a typical input script used for this series of structure calculations is given in the Appendix.

Table 2.2 Simulated Annealing parameters

Parameters	Stage 1	Stage 2	Stage 3	Stage 4
	High temperature torsion-angle MD	Slow cooling torsion molecular dynamics	Slow cooling Cartesian MD	Final energy minimization
Temperature	50000	50000 ---> 2000K	2000K ---> 300K	-----
Cooling steps	-----	10000 (55000)	3000 (40000)	-----
Timestep	0.015 ps	0.015 ps (0.005 ps)	0.005 ps	-----
kNOE	150	150	150	75
kDA	100	200	200	400
KRDC	-----	0.003 ---> 0.6	0.6	0.6

Note: Values in brackets signify the values used for GlpF, Halor and AMC membrane protein simulations in order to ensure convergence

2.4 Evaluation of structure quality

From the torsion angle SA protocol, 100 structures were generated and the 10 lowest energy structures were used to form an ensemble (as chosen by the accept.inp script from the CNS website). To evaluate the accuracy of structures that were generated, the ensemble was superimposed onto the target structure that was used to generate the set of structural restraints in the program MOLMOL (75). The root mean squared difference (rmsd) between each member of the ensemble and the target structure was calculated for backbone atoms in the helical parts of the protein. The average rmsd was taken as a measure of the accuracy of the ensemble, while the standard deviation in the rmsd is a measure of the variability in the accuracy. The precision of the ensemble was also determined for backbone helical atoms as the average rmsd from the 10 lowest energy structures. In general, higher quality structures have small deviations from ideal geometry, good agreement with their restraints (no violations), and a low overall ensemble energy.

The one-way analysis of variance (ANOVA) test was used to evaluate whether the differences in rmsds obtained from different simulations were significant (84). The ANOVA test evaluates the aggregate differences and the random variability between the groups of data. In order for the data sets to be significant, the aggregate differences must be larger than the random variability between the groups resulting in a high f-value. F-Values of <0.05 (<0.01) indicate that there is a >95% (>99%) probability that the difference is significant.

Chapter 3

RESULTS

3.1 Choice of membrane protein dataset

In order to examine the quality of α-helical structures that can be obtained from the selective methyl labeling strategy, we first must choose representative structures of polytopic α-helical membrane proteins that can be used to generate simulated NMR restraints. In the PDB, there are approximately ~120 3D polytopic membrane proteins structures available at the time our study was initiated (9). However, only 66 of these structures were comprised of α-helical transmembrane segments. Also, many of these are homologues, and hence the total number of non-homologous α-helical membrane proteins was 23. From this set, any protein that had more than 300 residues was also removed since structure determination of larger membrane proteins may be too challenging to attempt using this strategy. Twenty proteins were removed according to this criteria, leaving 3 non-homologous polytopic α-helical membrane proteins; namely the 1.8 Å structure of the 7 TM segment halorhodopsin (PDB code: 1E12) (73), the 2.2 Å structure of the 6 TM helix glycerol facilitator (PDB code: 1FX8) (72), and the 2.2 Å structure of the 6 TM helix adenine nucleotide carrier (PDB code: 1OKC) (74).

Given the limited number of unique polytopic α-helical membrane protein structures that was available for this analysis, we extended the range to include two simpler

systems that had been determined by solution NMR. These included the structure of a well characterized single-spanning homodimer called glycophorin A (PDB code: 1AFO) determined in dodecyl phosphocholine (DPC) micelles, (55) and a two TM segment fragment of the human glycine receptor (PDB code: 1VRY) determined in trifluoroethanol (TFE) since it provided a single polypeptide similar in size to the GpA dimer. Overall, the database used for our study included four polytopic α-helical membrane proteins with one monotopic α-helical membrane protein that forms a dimer under micellar conditions (Figure 3.1).

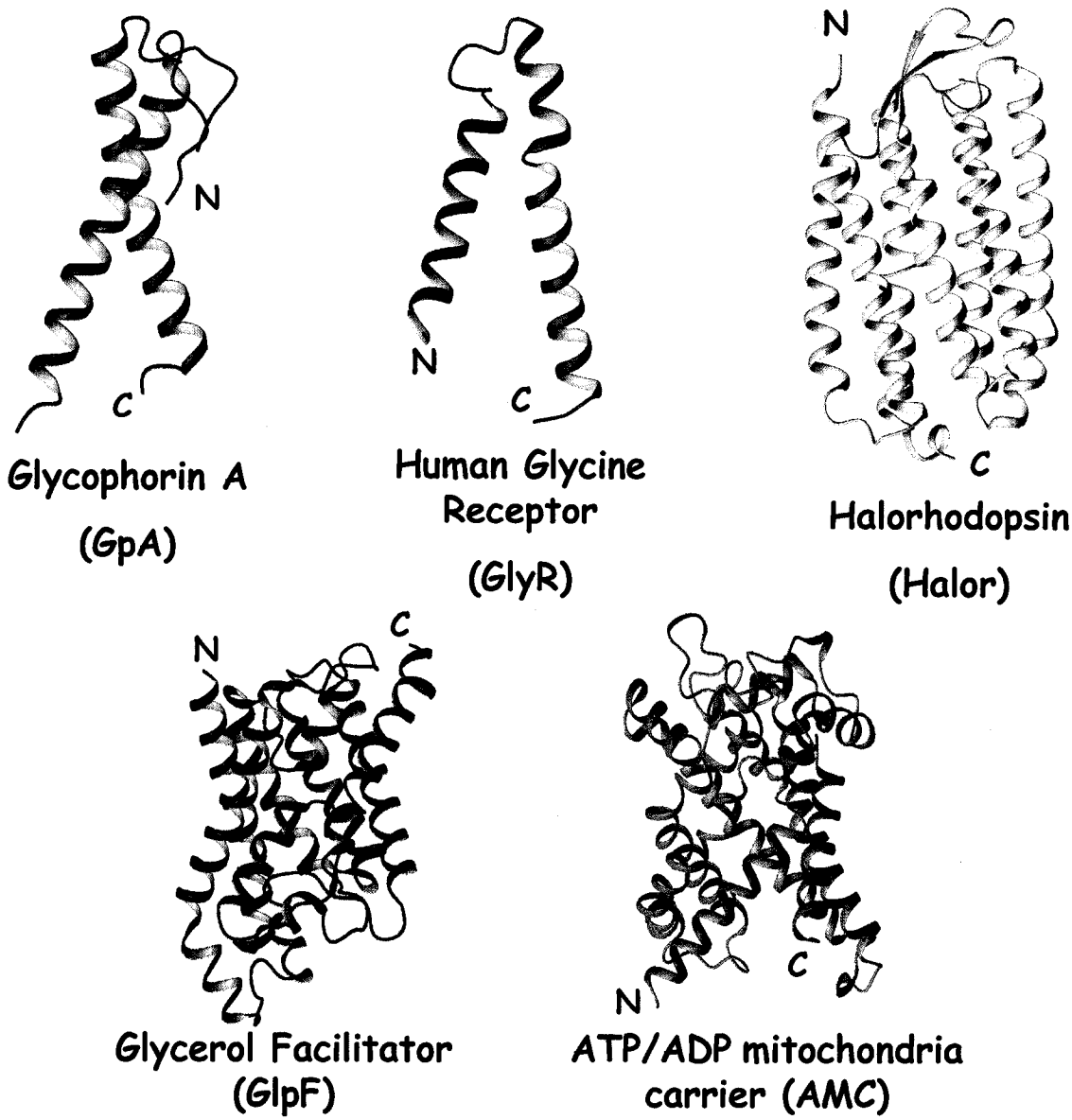

Figure 3.1 – Ribbon diagrams showing the five α-helical membrane proteins used in this study. Three of the membrane proteins (Halor, GlpF, and AMC) were solved by x-ray crystallography with resolution 2.2 Å or higher. The other two membrane proteins, GpA and GlyR, were solved by solution NMR in DPC micelles and TFE-d2 solvent respectively.

3.2 Investigation of the selective methyl protonation strategy for α-helical membrane protein structure determination

One factor we found to be important for these structure calculations was the number of cooling steps used during the MD simulations. Previous studies have shown that the protein size influences the ability of the system to reach convergence, with larger proteins requiring more cooling steps (85). This was shown to be especially important for situations when an extensive set of NOEs are unavailable, as is the case when selectively methyl protonated samples are used. Slower cooling allows more time for thermal equilibration and minimizes the risk that the structure will become trapped in local energy minima (85). In this case the quality of structures can be improved by simply increasing the number of cooling steps, a strategy that provides more efficient improvements than previous suggestions that more structures be calculated instead (48). For this reason, for proteins GlpF, Halor, and AMC the number of cooling steps was increased to ensure that the system was reaching convergence. As shown in Figure 3.2, the accuracy of GlpF increased with the number of cooling steps, with a plateau occurring at 55000 steps. Therefore for all subsequent calculations for these proteins the 55000 cooling steps was used. A similar protocol was also performed to optimize the molecular time steps during the simulation. It was found that GlpF, Halor, and AMC structure calculations required a shorter time step, and hence a 0.005ps time step was used for these proteins.

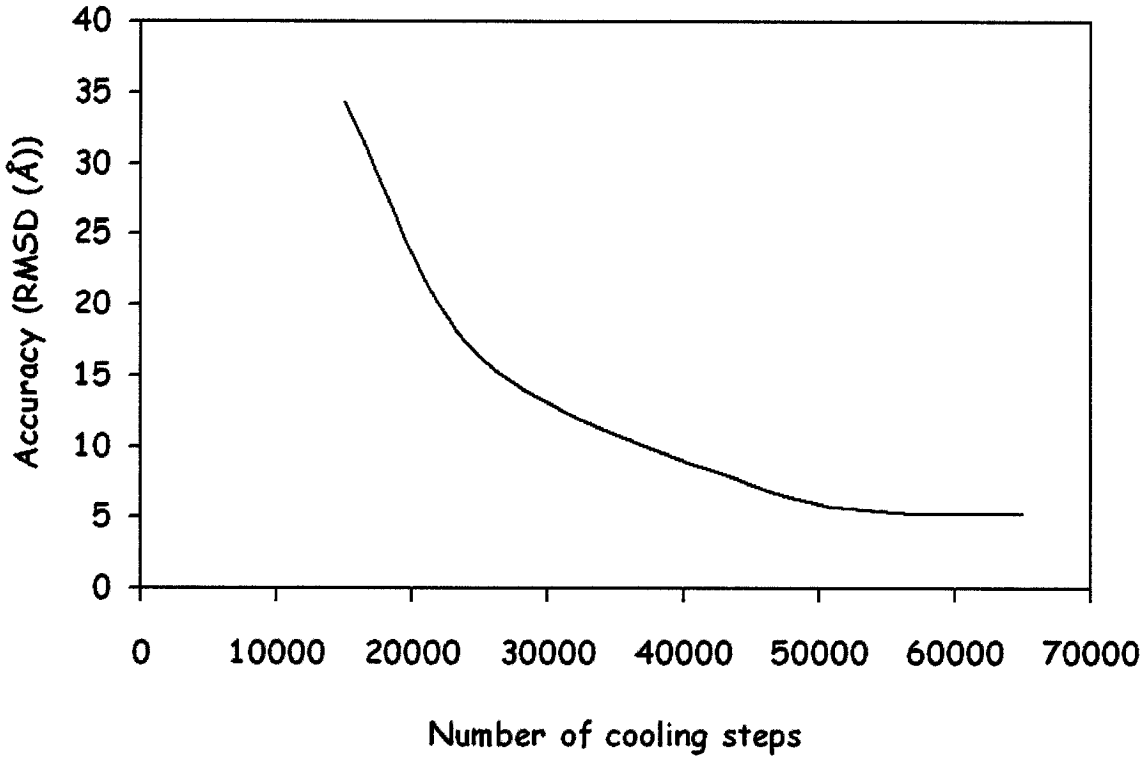

Figure 3.2 – Chart showing the accuracy of GlpF structure calculated with different number of cooling steps. Each simulation was calculated with 3.4 intra-helical NOEs, 145 inter-helical NOEs and a complete set of backbone dihedral angle restraints.

In order to evaluate the utility of this strategy for polytopic α-helical membrane protein structure determination, a simulated set of H^N-H^N, methyl-H^N, and methyl-methyl distance restraints were generated for each protein in our data set. In each case a 'full' NOE restraint list was created for the best-case scenario that would arise if it were possible to observe and assign NOEs for all proton pairs that are within 5Å of each other. As shown in Table 2.1, the majority of these simulated NOEs were short-range intra-helical restraints, as would be expected due to the relatively high density of amide protons in the α-helical backbone. Approximately 4 NOEs/helical residue were present, except for the structure of GlyR, which only had 2.5 NOEs/helical residue. Conversely, relatively few simulated

NOEs were obtained between helices with 0.2 (GpA) - 1 (Halor) interhelical NOE per residue being available. Since the number of interhelical NOEs that can be measured for these selectively methyl protonated samples is not constant, it is already apparent that significant differences in structure quality can be anticipated for these different membrane proteins.

3.2.1 Evaluation of intra-helical NOE influence on structure quality

Since intrahelical NOEs only provide local information that mainly affects the quality of secondary structure elements (86, 87) we wanted to evaluate the impact of these restraints on the global structure in the best case scenario where all potential interhelical NOEs are available (88). For this purpose we performed a series of structure calculations using decreasing numbers of intra-helical NOEs while keeping the number of inter-helical NOE restraints constant. Since backbone secondary shifts can be used to identify secondary structure elements with a high degree of accuracy, dihedral angle restraints for the TM helices were also included in the same way as is routinely done for structure calculations of water-soluble proteins (62). For each simulated dataset, 100 structures were calculated in CNS using standard simulated annealing protocols and the 10 lowest energy structures were compared to the target structure that had originally been used to generate the dataset. The root mean squared difference (rmsd) between each member of the ensemble and the target structure was averaged over the ensemble and used as a measure of accuracy for each simulated structure determination. According to this measure the lower the average rmsd, the higher the accuracy of the structure determination.

As shown in Figure 3.3 a wide range of accuracies was obtained from these simulations, depending on the protein under investigation. The best accuracies were obtained for GpA and Halor with rmsds to the target structure of 2 and 3 Å respectively. Meanwhile, lower accuracies were achieved for GlyR, GlpF and AMC with rmsds of 4 to 6 Å. Given the different sizes of proteins that yield these lower accuracy structures (2 - 7 TM segment proteins), it appears that the quality of structures obtained using this labeling method is not correlated with the size of the protein, in line with previous observations with water-soluble proteins (62).

In order to evaluate the impact that intra-helical NOEs have on the accuracy of these structures these calculations were repeated with the same set of restraints with a decreasing number of intra-helical NOEs. As shown in Figure 3.3 the accuracy values did not show a significant change, indicating that the ensemble of structures that are generated even with an extremely limited set of intra-helical NOEs per residue are just as accurate as structures calculated with the full set of intra-helical NOEs. In addition, the variability in the accuracy values over the ensemble (shown as the error bars) did not show any significant changes when calculated using a full or limited set of intra-helical NOEs.

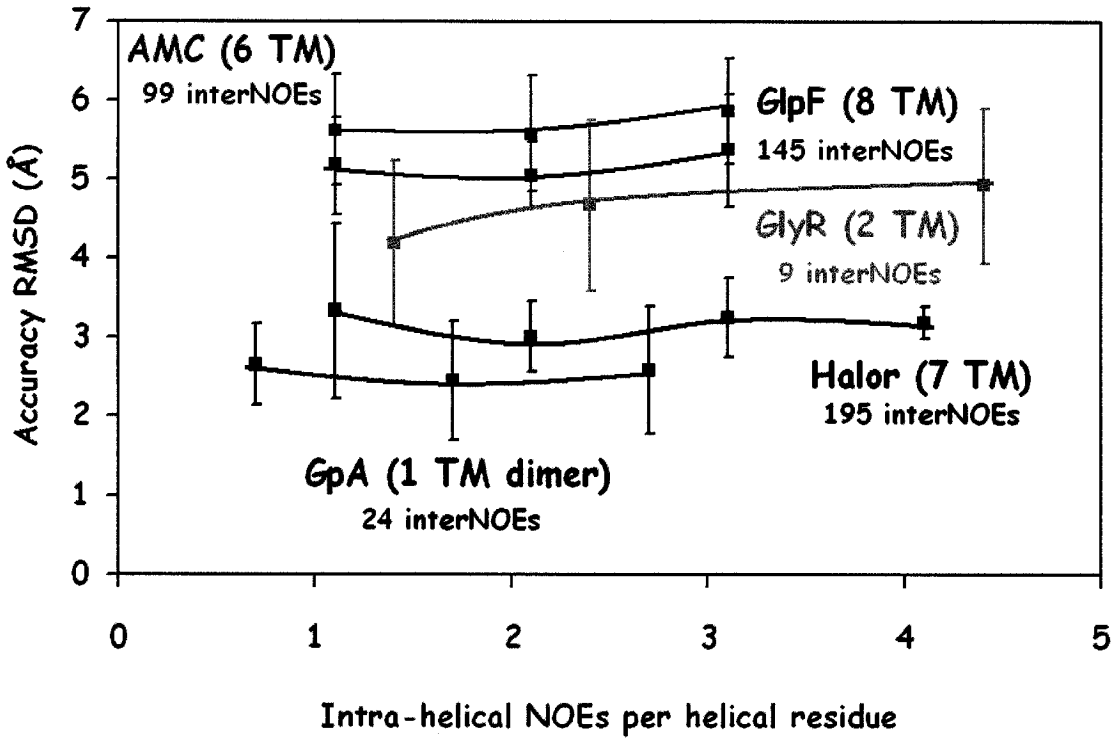

Figure 3.3 – Impact of intra-helical NOEs on accuracy of structures determination of selectively methyl protonated α-helical membrane proteins. The average accuracy is shown as the rmsd difference from the target structure that was originally used to generate the simulated set of NOEs. For each membrane protein all structure calculations were done using the same number of inter-helical NOEs and a decreasing number of intra-helical NOEs. Lines connect points in each series of calculations to guide the eye. Error bars determined by the standard deviation in the rmsd values represent the variability of the accuracy values for each calculation.

3.2.2 Evaluation of the influence of inter-helical NOEs on structure quality

In contrast with intra-helical NOEs, inter-helical NOEs are exclusively long-range restraints, and hence they should have a significant influence on structure quality. Therefore, for each protein a series of simulated datasets was constructed with a decreasing number of inter-helical NOEs and a constant number of intra-helical NOEs. Simulated structure calculations were then performed exactly as described for the previous set of

experiments and the average accuracy determined for each ensemble. As shown in Figure 3.4, for all proteins except GlyR, decreasing the number of inter-helical NOEs decreased the accuracy of the resulting structures. In the case of GlyR there were very few inter-helical NOEs to begin with and so this may have already been too small a number of long-range restraints to allow much difference in structure quality to be observed. However, for all other samples a 50% or greater reduction in the number of inter-helical NOEs led to an approximate 1-2 Å increase in rmsd from the target structure. In addition, this trend was conserved regardless of the number of intra-helical NOEs used, indicating that the number of long-range restraints was a dominant factor in determining structure quality, as expected.

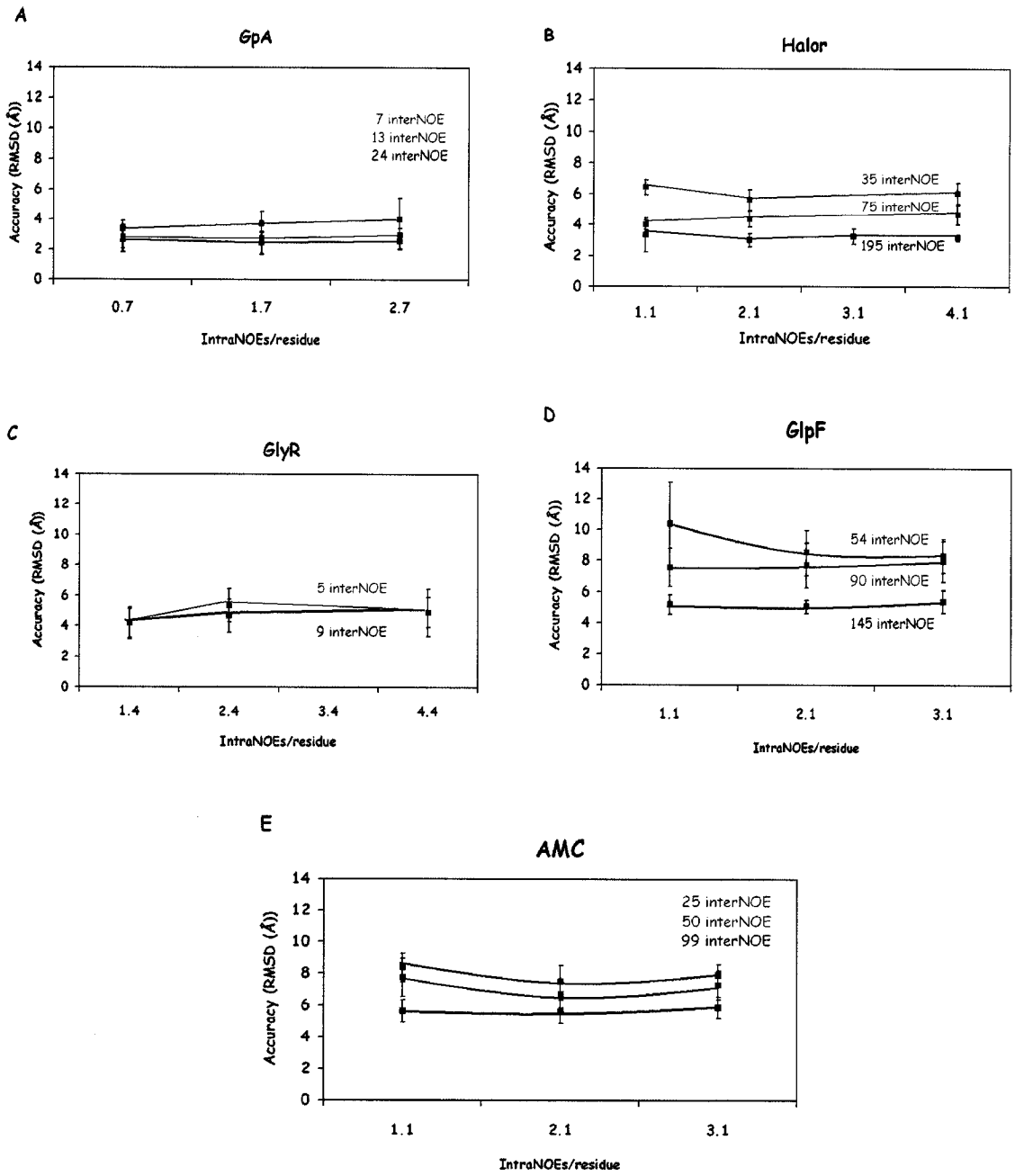

Figure 3.4 – Impact of the inter-helical NOEs on structure determination of α-helical membrane proteins. Each graph shows the accuracy as the rmsd from the target structure as a function of the number of intra-helical NOEs per helical residue. Each line connects a series of simulations performed with the same number of interhelical NOEs and a different number of intrahelical NOEs. Note that the blue series is the series run with the 'complete' set of inter-helical NOEs that is summarized in Figure 3.2. Error bars show the standard deviation in accuracy values for each ensemble.

3.2.3 Structure determination without intra-helical NOEs

Given the apparent insensitivity of the structure quality to the number of intra-helical NOEs, we were interested to determine whether the removal of these restraints could produce structures with comparable accuracy values. For this purpose, a series of structure calculations was performed with the full set of inter-helical NOEs as done in the previous section except that intra-helical NOEs were no longer included in the simulated dataset. As shown in Figure 3.5, the accuracies of structures determined with no intra-helical NOEs (red bars) were indistinguishable from those determined with the maximum number of intra-helical NOEs (blue bars) (Figure 3.5). These results demonstrate that selectively methyl labeled membrane protein structures can be determined with equal accuracy using inter-helical and dihedral angle restraints alone, and that the additional information provided by intra-helical NOEs does not improve the structure.

To investigate how the accuracy would be affected with a smaller number of inter-helical restraints in the absence of intra-helical NOEs a series of calculations was done with a decreasing number of inter-helical NOEs (Figure 3.6). As anticipated, the trend for each protein showed a decrease in structure quality with decreasing numbers of inter-helical restraints. Interestingly, these differences only became significant for Halor, AMC and GlpF at levels of 0.4 - 0.6 interhelical restraints per residue, while GpA did not show significant differences in structure quality until 0.1 inter-helical restraints/residue were used. GlyR did not show a significant difference when a smaller number of NOEs was used, indicating that other factors, in addition to number of inter-helical restraints, could affect the quality of the structures being obtained.

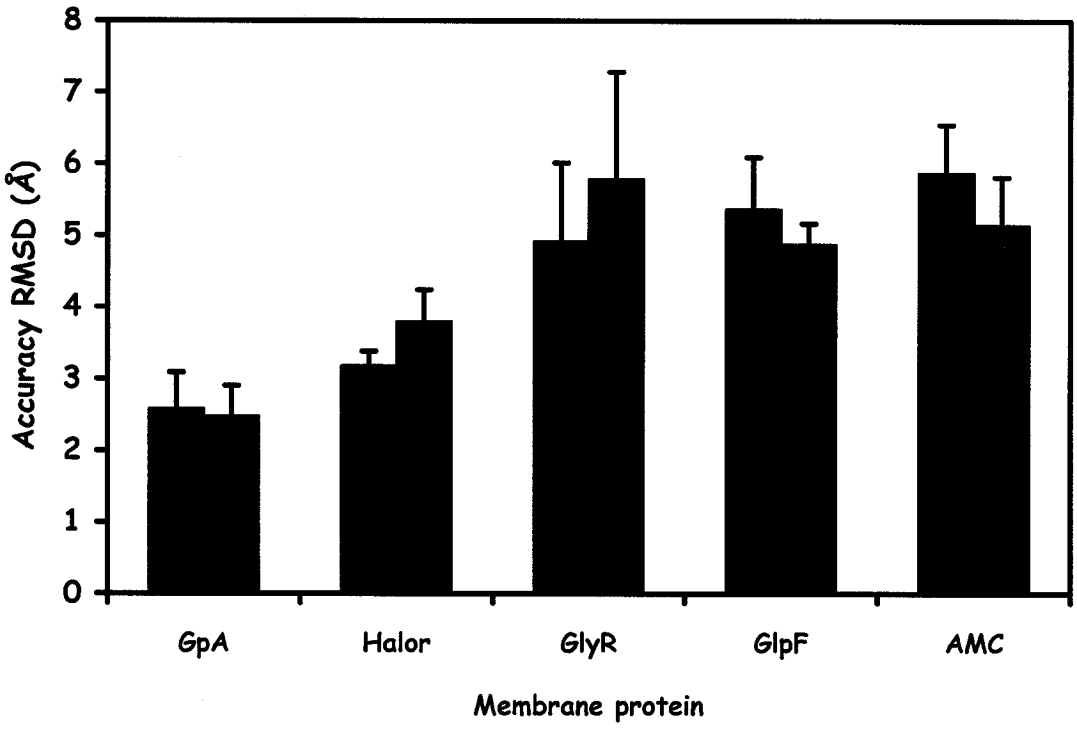

Figure 3.5 – Average rmsd from target structure for ensembles determined either with (blue bars) or without (red bars) intra-helical NOEs (red bar) for the indicated membrane protein. The maximum possible number of inter-helical NOEs was used for all calculations.

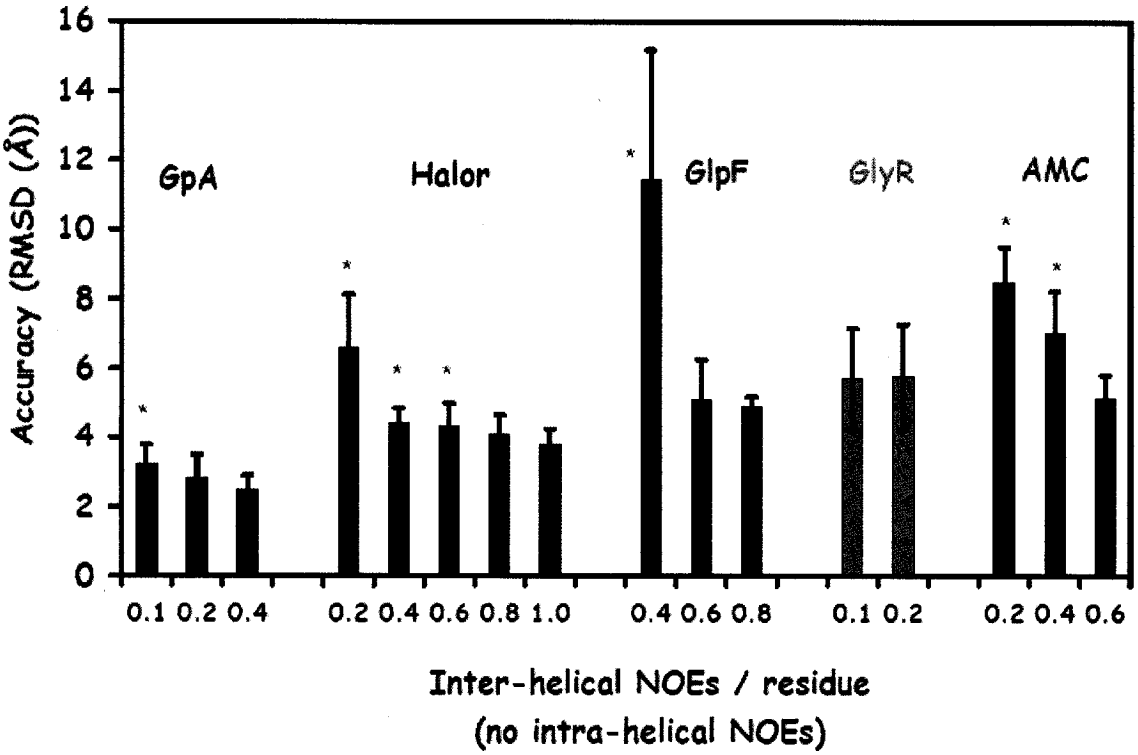

Figure 3.6 – The impact of inter-helical NOEs on structure quality in the absence of intra-helical NOEs. The average rmsd from the target structure is shown as a function of inter-helical NOEs for the indicated membrane protein. Each series was done with the maximum number of inter-helical NOEs (defined by the 'complete' dataset) and with randomly reduced NOE datasets. The asterisk (*) indicates accuracy values that are significantly worse than that obtained for the same protein with the complete set of inter-helical NOEs according to a one-way ANOVA test ($p>0.01$).

3.2.4 Inter-helical NOE distributions

Since the quality of structures that could be obtained was not straightforwardly related to the number of inter-helical NOEs, I was interested in evaluating the possibility that the distribution of inter-helical NOE restraints could explain the observed differences. As shown in Figure 3.7, except for GlyR, qualitative examination of all the proteins indicates that the distribution of inter-helical NOEs is relatively uniform. In the case of GlyR, the 9 inter-helical NOEs were concentrated in the central region of the interface between the two TM helices. This poor distribution of inter-helical restraints for GlyR would help account for the relatively low accuracy obtained for this small protein in our simulated calculations. The high redundancy of the structural information provided by these restraints also explains why the quality does not change when fewer inter-helical NOEs are used for the GlyR calculations. Hence these results indicate that the uniformity of the NOE distribution is more important than the numbers of NOEs for determining accurate structures.

The inter-helical NOE distribution may also explain the relatively low accuracies obtained for GlpF and AMC structures since some parts of the inter-helical interfaces did not have restraints to define them. For example, as shown in Figure 3.8 there are a number of regions in these two proteins that would not give rise to inter-helical NOE restraints in a selectively methyl protonated sample, corresponding to 13% and 15% of the helical regions of GlpF and AMC, respectively. In contrast, Halor has fewer of these restraint-poor regions, which only make up 9% of its transmembrane helices. GpA also has uniformly distributed inter-helical NOEs, which would account for the relatively high quality of structures that can be obtained for this protein.

Figure 3.7 – Ribbon representation of the target membrane protein structures showing the distribution of inter-helical NOEs (red or yellow tubes) for the selectively methyl protonated condition used to create the simulated NOE datasets. From top left to the bottom right for the identity of the proteins are: GpA, GlyR, Halor, GlpF and AMC.

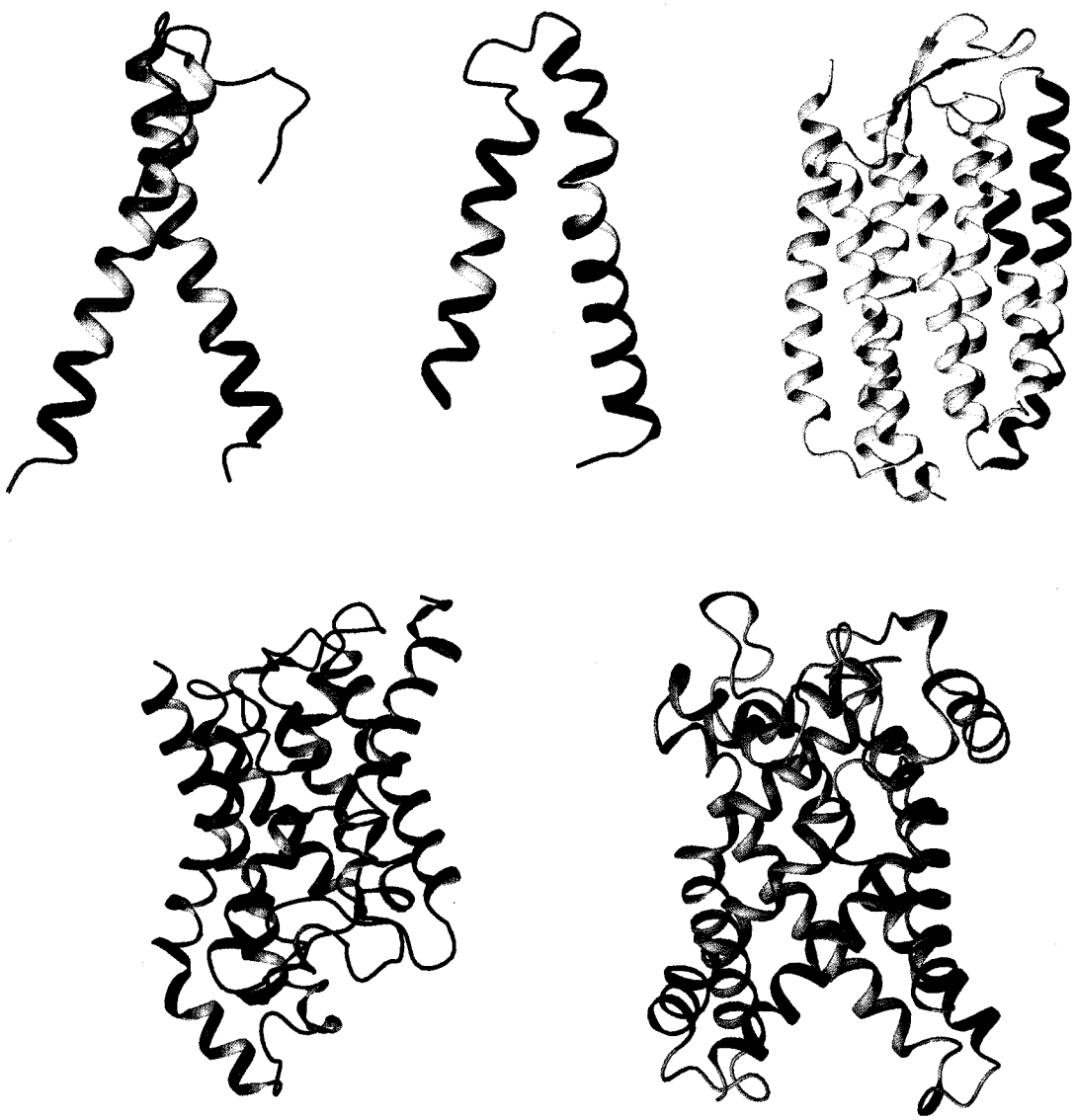

Figure 3.8 – Transmembrane regions lacking inter-helical NOEs. The blue regions represent residues that do not have inter-helical NOE restraints within a 5 Å radius from its backbone amide proton.

In order to investigate the impact of non-uniform inter-helical NOE distributions on structure quality, a reduced inter-helical NOE restraint list was generated to simulate the effect of non-uniform NOE distributions for GpA and Halor. These new NOE lists were created by removing all inter-helical NOEs located within a 5.5 Å radius from a backbone amide in either an N-terminal, central, or C-terminal residue of a transmembrane helix (Figure 3.9). For the multi-spanning Halor this was done for all helices together so that NOEs were missing from either the cytoplasmic, central or extracellular side of the protein, Structure calculations identical to those described in the previous simulations were then performed and the quality of the resulting ensembles calculated as an rmsd from the target structure. As shown in Figure 3.10, removal of NOEs from the central region of the protein were well tolerated while removal NOEs from one side of the protein in most cases led to statistically significant differences in structure accuracy relative to an ensemble generated with a similar number of restraints with a uniform distribution. Although the low accuracy (rmsd: 5.60+/-0.39 Å) for the Halor structures obtained without the C-terminal NOEs might be attributed to the smaller number of inter-helical NOEs in that dataset, it is also significantly worse than the ensemble calculated with 75 uniformly distributed NOEs (rmsd: 4.44+/-0.41 Å). Similarly, GpA had only one inter-helical NOE in the C-terminal region which, when removed, gave rise to an ensemble with an rmsd of 5.48+/-2.72 Å from the target structure, a $>$ 2 Å increase over that of the ensemble calculated with a uniform distribution of fewer NOEs. These results suggest that it is particularly important to obtain long-range restraints between the ends of transmembrane helices. In addition, the fact that the NOE-poor regions of AMC and GlpF are largely found at, or close to, the ends of transmembrane helices (Figure 3.8) suggests that this was likely an important factor that reduced the quality of the ensembles for these proteins. Although this was also true for

Halor, the fact the NOE-poor regions were concentrated in one area of the protein allowed the main part of the structure to be determined with reasonable accuracy.

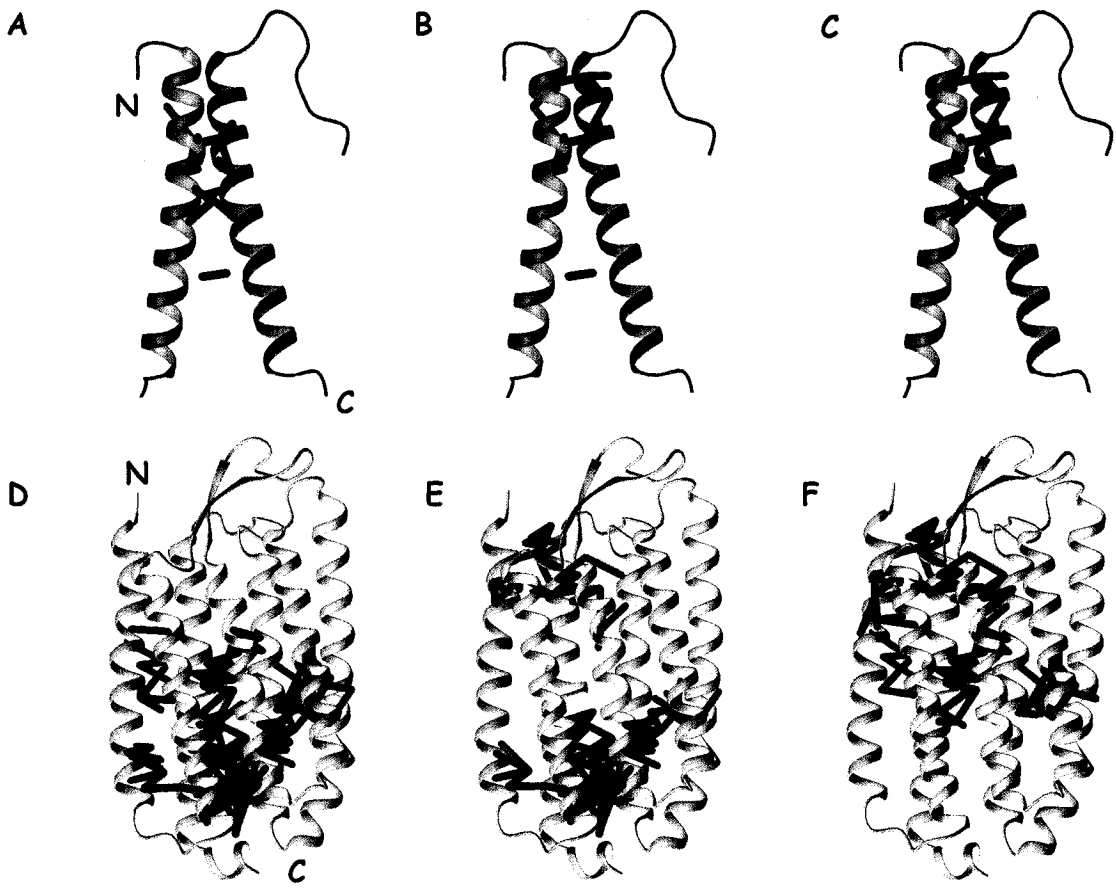

Figure 3.9 – Non-uniform inter-helical NOE datasets used for in GpA (top) and Halor (bottom). Inter-helical NOEs were removed from either a) and d) the N-terminal, b) and e) the middle, or c) and f) the C-terminal regions. These non-uniform distributions of NOEs were made by removing all inter-helical NOEs that were within a 5.5 Å radius of a backbone amide proton of an N-terminal, C-terminal or central residue from the transmembrane helices.

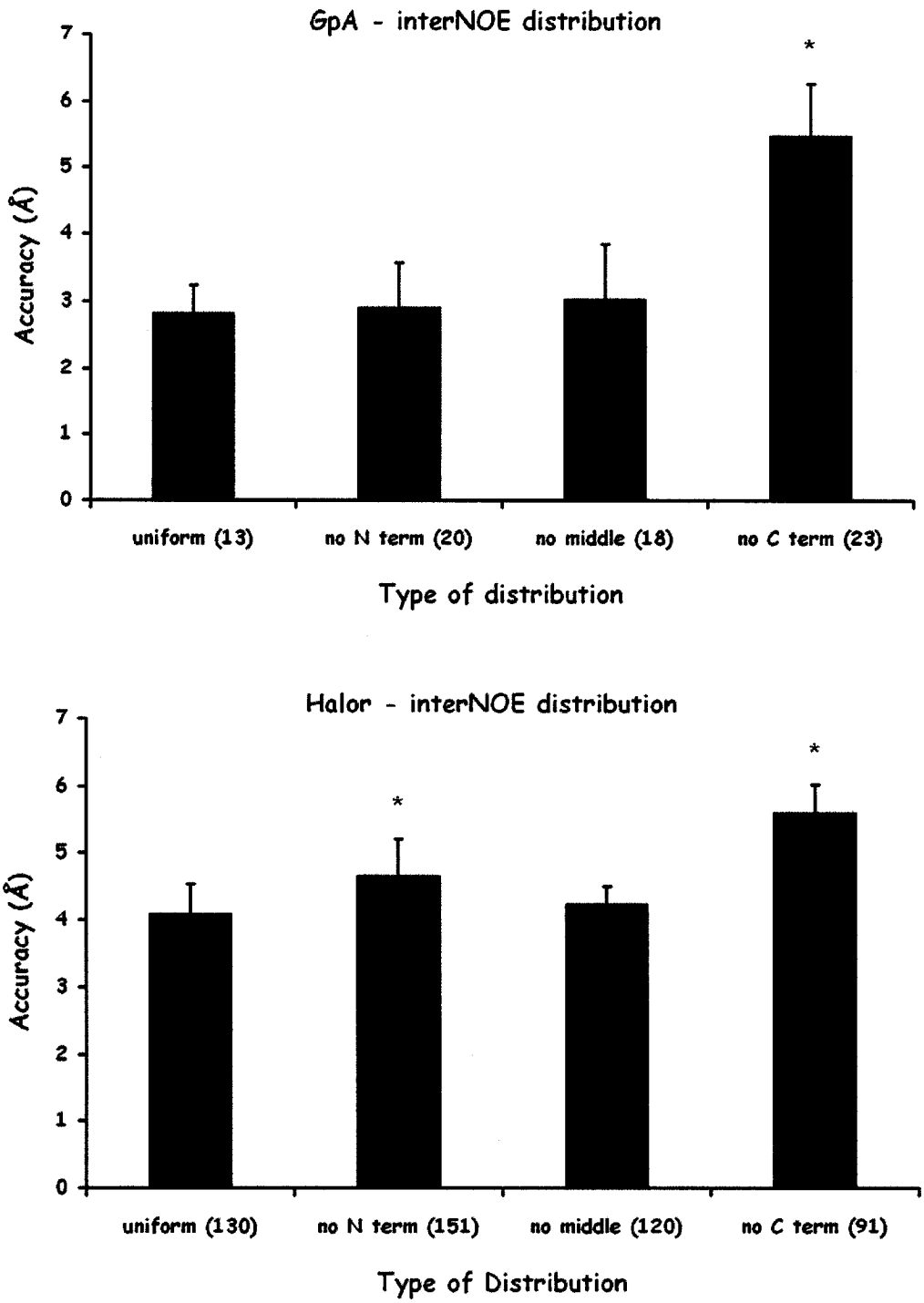

Figure 3.10 – The impact of inter-helical NOE distribution for GpA and Halor structures. Each bar represents the average backbone accuracy of 10 lowest energy structures. The asterisk (*) indicates accuracy values that are significantly worse than that obtained for the same protein with the uniform set of inter-helical NOEs according to a one-way ANOVA test ($p>0.01$).

3.3 Exploring other possibilities to improve the quality of structures from methyl-protonated α-helical membrane proteins

Although obtaining a uniform distribution and a large number of inter-helical NOEs seemed to improve the quality of membrane proteins, low accuracy ensembles (rmsds of ~5-6 Å) were obtained for GlpF and AMC even in the ideal case when the 'full' set of NOEs from the selectively methyl protonated sample was available. This limits the utility of this labeling strategy since a higher quality structure (~2-3 Å accuracy) would provide more insight into its molecular structure. In addition, while the labeling strategy gives rise to more accurate structures when uniformly distributed inter-helical NOEs are available, in many cases only a non-uniform distribution of inter-helical NOEs can be obtained. For these reasons, we decided to investigate additional strategies that might improve the quality of these large α-helical membrane protein structures and reduce the dependency on inter-helical NOEs uniformity.

3.3.1 Effect of dihedral angle restraints on structure quality

In all of the structure calculations performed so far in this work dihedral angle restraints were employed using a relatively conservative range of values as is typically done in routine structure calculations, with ϕ and ψ being allowed to vary by 20 and 30 degrees about the canonical helical values respectively. This is done to account for the uncertainty associated with the dihedral angle assignment that comes from the secondary shift data used to identify secondary structure elements. However, in the case of membrane proteins

the strength of the secondary structure prediction can be augmented using hydropathy profiles (89) which are also used to identify transmembrane helices. Therefore it should be possible to introduce tighter dihedral angle boundaries once hydropathy profiles and secondary shift data have identified the parts of the membrane protein that form transmembrane segments. This should promote higher accuracy structures as has been demonstrated for the water soluble protein G IgG-binding domain III (PDB code: 2IGD) (80).

To test this strategy, the dihedral angle boundaries were reduced to 2.5° and 5.0° for ϕ and ψ respectively. Structures were then calculated using the complete set of inter-helical NOEs and the accuracy of those ensembles were compared to those previously determined with the more loosely restrained dihedral angles. As shown in Figure 3.11, all the ensembles that were calculated with strong dihedral angle restraints showed a statistically significant improvement in the quality of structure, with the best improvement being for AMC and GlyR, which brought the accuracy rmsd down to 4.80 Å and 3.38 Å respectively. In the case of Halor and GlpF much smaller improvements were obtained, although the differences were still statistically significant. Overall these results indicate that more narrowly defined dihedral angle restraints could be employed as a general strategy to improve the accuracy of selective methyl protonated polytopic helical membrane protein structures.

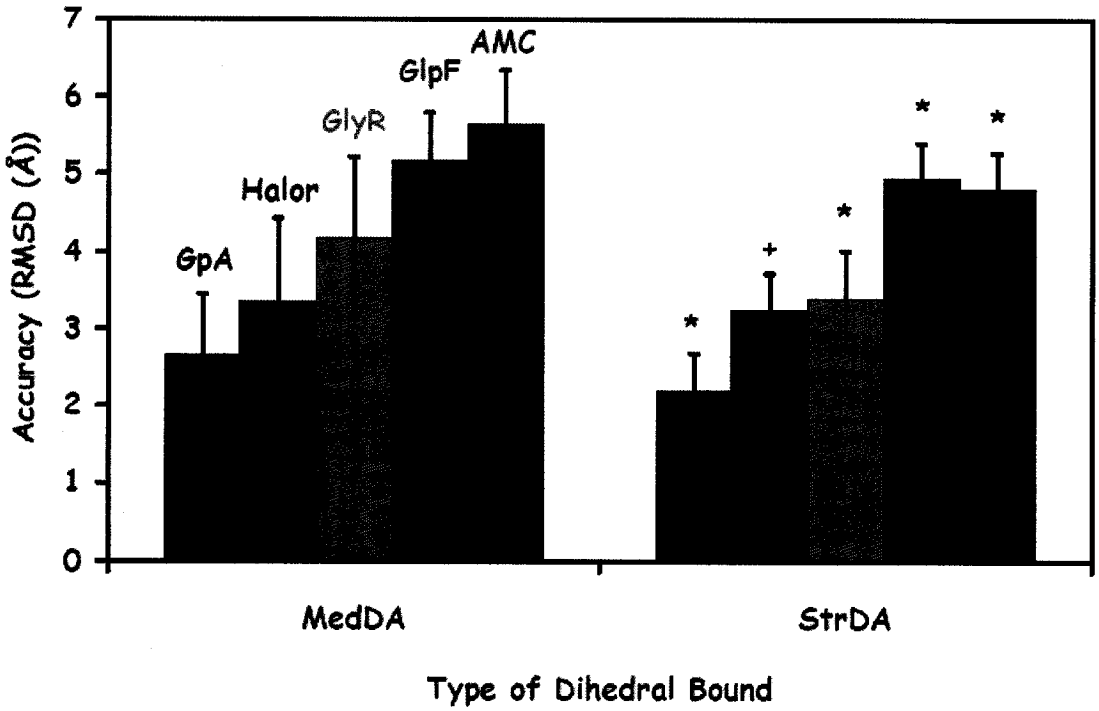

Figure 3.11 – The impact of medium (left) and strong (right) dihedral angle restraints on structure determination of α-helical membrane proteins. Each bar represents the average accuracy of the ensemble calculated with the full list of inter-helical NOEs and a reduced list of intra-helical NOEs (~1.0 intraNOE/residue) as a function of dihedral angle restraint boundary where medium corresponds to a +/- 20° and +/-30° range of accepted dihedral angles and strong to a +/- 2.5° and +/- 5° range for phi and psi respectively. Statistically significant differences between ensembles calculated with strong and medium strength dihedral angle restraints are with $p>0.01$ (*) or $p>0.05$ (+) are indicated.

3.3.2 ^{1}H-^{15}N RDCs

Although strengthening the dihedral angle restraints improved the quality of the structures, we were interested to determine whether the quality can be improved further by other means. Specifically, residual dipolar couplings (RDCs) have the potential to add new sources of long-range information that can improve the quality of structures. For example,

structure refinement using ^{1}H-^{15}N RDCs has been successfully demonstrated for detergent-solubilized membrane proteins such as the single-spanning α-helical proteins Vpu (79) and Pf1 coat protein (90). The addition of ^{15}N – ^{13}CO and ^{13}CO-$^{13}C\alpha$ RDCs has also been done to refine structures of the β-barrel OmpA (19), and the pentameric form of phospholamban (54). In addition, as they only require assignment of chemical shifts for backbone atoms, this data can be simpler to acquire relative to the long-range NOEs (19, 42, 91). These observations have led us to investigate the possibility of improving the quality of α-helical membrane protein structures using ^{1}H-^{15}N RDCs. Although other types of RDC can be measured in these samples, the ^{1}H-^{15}N RDC is the most commonly used for structure refinement.

For this purpose a series of structure calculations was carried out exactly as described in the previous sections using the complete set of inter-helical NOEs except that ^{1}H-^{15}N RDCs were added for the transmembrane helices. As shown in Figure 3.12, the backbone rmsd to target structures for ensembles determined with the simulated RDC data was not significantly different from those determined without RDCs. These results, in agreement with other results from previous studies (42, 92), suggest that the ^{1}H-^{15}N RDCs might need to be complemented by the measurement of different types of RDCs (e.g. $^{1}H\alpha$-$^{13}C\alpha$) in order to exhibit a more significant effect on structure accuracy.

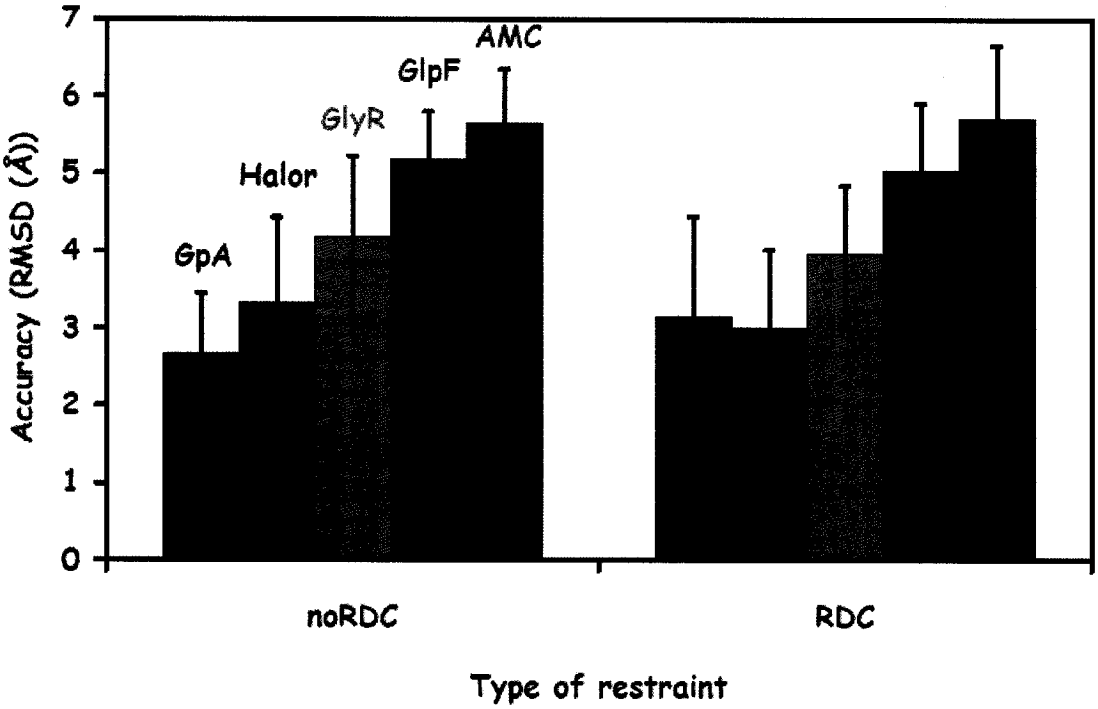

Figure 3.12 – The impact of $^{1}H^{N}$-^{15}N RDCs on structure accuracies. Each bar represents the average accuracy of the ensemble calculated with the full list of inter-helical NOEs and a reduced list of intra-helical NOEs (~1.0 intraNOE/residue) and medium strength (+/-20°,+/-30°) dihedral angle restraints. No statistically significant difference was obtained with the addition of RDCs.

3.3.3 Deviations from ideal helicity as a contribution to structure inaccuracy

Although it appeared that structure accuracy was improved through the use of stronger dihedral angle restraints we were interested to determine whether the presence of non-ideal helix structures was a significant source of error in our structure determination. For this purpose we examined the amount of error arising from deviations from canonical helical values for two types of helices: ideal and kinked (Figure 3.13). A series of calculations was performed with either normal or strong dihedral angle restraints for a

sample of single helices taken from GpA, Halor, GlpF or GlyR and the accuracy of these secondary structure

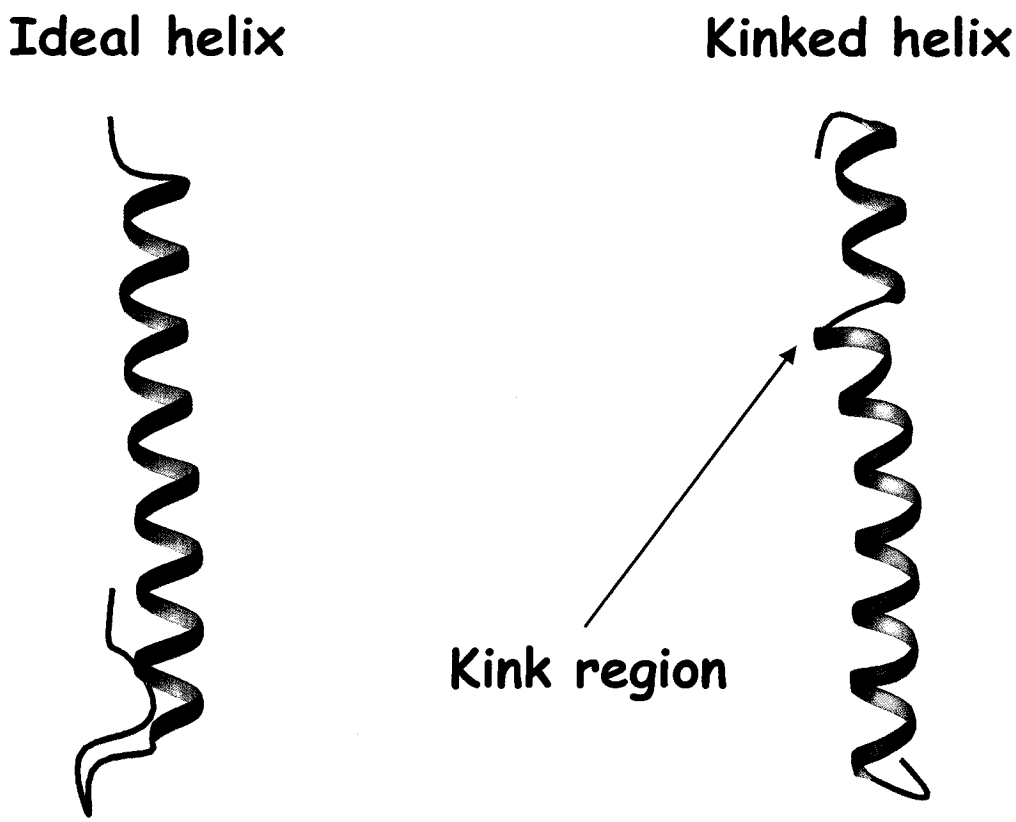

Figure 3.13 – Schematic representation of the two types of helices that were studied in the single-helix calculations.

elements evaluated. As shown from Table 3.1, the accuracy of the helix was generally better for ideal structures with rmsds of ~1 Å being obtained. However, when a kink was present the rmsd was significantly higher, with a range of 1.6 - 2.2 Å being observed. Nonetheless, the significant rmsd obtained for single helices in both cases indicates that a large part of the inaccuracies obtained for the structure

Table 3.1 The backbone rmsd for single helix calculations

Protein	Residues	Length of helix	Restraints			
			w/o RDCs		with RDCs	
			Med D.A.	Strong D.A	Med D.A.	Strong D.A
GpA	72-98	27	1.15+/-0.20	0.75+/-0.10	1.12+/-0.30	0.83+/-0.10
GlyR	0-16	17	0.85+/-0.09	0.76+/-0.08	0.75+/-0.09	0.59+/-0.07
	25-54	**30**	**2.01+/-0.30**	**1.60+/-0.20**	**1.98+/-0.10**	**1.62+/-0.20**
Halor	26-51	26	1.12+/-0.40	0.95+/-0.30	1.03+/-0.50	0.88+/-0.20
	106-127	22	0.94+/-0.20	0.90+/-0.10	0.90+/-0.20	0.83+/-0.10
	158-190	**33**	**2.28+/-0.80**	**2.10+/-0.80**	**2.22+/-0.50**	**2.08+/-0.80**
	193-216	24	1.13+/-0.50	0.81+/-0.30	0.74+/-0.20	0.62+/-0.10
	227-255	29	1.02+/-0.40	0.87+/-0.30	0.95+/-0.20	0.72+/-0.10
GlpF	7-34	28	1.58+/-0.50	1.25+/-0.40	1.42+/-0.30	0.72+/-0.10
	41-63	**23**	**1.65+/-0.40**	**1.23+/-0.10**	**1.64+/-0.30**	**1.21+/-0.50**
	69-78	10	0.55+/-0.10	0.36+/-0.03	0.50+/-0.10	0.31+/-0.10
	83-108	26	1.00+/-0.30	0.70+/-0.10	0.93+/-0.50	0.64+/-0.10
	145-167	23	0.82+/-0.30	0.48+/-0.06	0.77+/-0.20	0.42+/-0.10

Note: The bold rows are the kinked helices

calculations of the full-length proteins can be attributed to inaccuracies in defining the secondary structure elements themselves. As shown in Table 3.2, the lower accuracy structures obtained for GlpF and AMC may have been caused by the lower percentage of canonical helical structure relative to that in more accurately determined structures such as Halor.

Table 3.2 Percentage of non-ideal helical residues in transmembrane helices

Protein	% Non-ideal*
GpA	0
Halor	8.5
GlpF	16
GlyR	17
AMC	17

* Non-ideal helical residues correspond to residues with φ and ψ angles that are outside -57° +\- 20° and -47° +\- 30° respectively.

Somewhat unexpectedly, when strong dihedral angles restraints were imposed the accuracy slightly improved for both ideal and kinked helices, with about a ~0.2 Å decrease in rmsd values typically being obtained. This provides additional support for the strategy of imposing strong dihedral angle restraints for helical transmembrane segments, and suggests that this strategy would not be detrimental even when the non-canonical helical structures are present.

We also tested the ability of ^{1}H-^{15}N RDCs to improve the quality of the structures of the single helices, since it may be possible for kinked regions to be defined by this data. Although very small improvements were observed, they were generally smaller than those obtained with the application of strong dihedral angle restraints in the absence of RDCs. This is in line with the very small but statistically insignificant improvement in structure accuracies that were obtained for the full-length proteins.

Chapter 4

DISCUSSION

Given the large size of a detergent-protein complex that can be anticipated for polytopic helical membrane proteins, it will be necessary to use methods geared towards the spectroscopic properties of these types of samples. In this work we have focused on a selective labeling method widely applied to the study of large water-soluble proteins whereby otherwise perdeuterated proteins are protonated only at amide positions and at the methyl groups of Val, Leu, Ile residues (57, 63). Although this labeling method has allowed global folds to be determined for water-soluble proteins more than 50 kDa in size (43, 62, 67, 93), it has yet to be applied towards polytopic α-helical membrane proteins. In addition, prior to this work the quality of membrane protein structures that can be obtained from this labeling strategy was not known and hence gaining insight into the utility of this approach was the purpose of our study.

4.1 Intra-helical NOEs versus dihedral angle restraints

Although it is generally accepted that a larger number of NOEs give better quality structures (67, 93, 94), this did not seem to generally apply for our sparse datasets since only inter-helical NOEs affected the accuracy of our structures, while intra-helical NOEs did not seem to be important. Based on our results it appears that helices can be well defined without any intra-helical NOEs, so long as helical regions can be identified by

other methods (i.e. secondary chemical shifts) such that dihedral angle restraints can be accurately imposed. This reflects the fact that the intra-helical NOEs were all short-range, with ~90% involving backbone amide protons. Of these, 30% were sequential (i, i+1), 30% were (i, i+2), and 15% were (i, i+3) NOEs. Although the inclusion of these NOEs in the structure calculation may reduce the range of potential secondary structures that can be adopted, they do not exclusively define an α-helical conformation. In fact, a variety of helix (i.e. 3_{10}) and turn structures are equally consistent with the same pattern of NOEs (95), making it impossible for an α-helix to be uniquely defined by these restraints.

In contrast with the selectively protonated samples, fully protonated helices give rise to a diverse sample of NOEs that help to define the α-helix, including $^1H^{\alpha}(i)$-$^1H^{\beta}(i+3)$, $^1H^{\alpha}(i)$-$^1H^{N}(i+4)$, $^1H^{\alpha}(i)$-$^1H^{N}(i+3)$, and sequential $^1H^{\alpha}-^1H^{N}$ NOEs. The $^1H^{\alpha}(i)$-$^1H^{N}(i+4)$ correlation is particularly important as it most uniquely defines an α-helical secondary structure (95). Since the selective methyl protonation strategy does not allow the observation of these NOEs, the unambiguous definition of an α-helical secondary structure cannot be achieved. This can partially explain why the inclusion of intra-helical NOEs from selectively methyl protonated samples in structure calculations did not improve the accuracy of the structure.

The structural ambiguity associated with the short-range amide-amide NOEs has the potential to be resolved by the NOEs involving protonated methyl groups, although these comprised a small part of the dataset for our test set of proteins. Intra-helical methyl-methyl NOEs were particularly sparse (<5% of the dataset), reflecting the fact that they can only occur when Ile, Val or Leu are on the same side of the helix, separated by no more than 4 residues in the primary sequence. This can give rise to an additional problem since, unlike

the amide-amide NOEs, the methyl-methyl NOEs are not uniformly distributed along the length of the helix. Consequently, in the sparsely defined case of the methyl-protonated protein, an asymmetric distribution of methyl-methyl NOEs about the helical axis could actually deform the helix during the structure calculation. For example, the helical structure could partially unwind on the side of the helix lacking methyl-methyl correlations, giving rise to a bent structure that could still be consistent with the available data. This would help to explain why inclusion of NOEs involving methyl protons did not appear to improve the accuracy of the helical secondary structure.

In light of these factors it is not surprising that dihedral angle restraints were particularly important in the simulations, to the point where dihedral angles alone were sufficient to define the helical structure. In addition, a statistically significant improvement in the accuracy was obtained when more narrowly defined dihedral angle boundaries were used, even though the actual structures used to generate the simulated restraint list had a significant proportion of dihedral angles outside this narrowly defined range. The difference between the actual and imposed dihedral angles does enforce a lower limit on the accuracy that can be obtained by this approach. As shown in Table 4.1 this limit (defined by the rmsd between a canonical α-helix and the target helix) is realized in almost all cases when strong dihedral angle restraints are imposed for the structure calculations with the full-length proteins. Interestingly, the dihedral angle restraints alone were not capable of enforcing this structure in the absence of inter-helical NOEs from the other parts of the protein. In none of the samples examined did the accuracy become worse when more narrowly defined dihedral angle restraints were used, even when the helix was kinked or

non-α-helical (e.g. 3_{10} helices). This is a surprising result since the dihedral angles for kinks or non-alpha helical regions lie outside this narrow boundary.

Table 4.1 Single TM helix accuracies

Proteins	Residues	Rmsd (Å) between the target helix structure and the helix structure from:		
		Single helix structure calculations*	Full length structure calculations+	Canonical α-helix
GpA	72-98	0.75+/-0.10	0.53+/-0.09	0.21
GlyR	0-16	0.76+/-0.08	0.68+/-0.14	0.63
	25-54	1.60+/-0.20	1.54+/-0.22	1.46
Halor	26-51	0.95+/-0.30	0.92+/-0.22	0.93
	106-127	0.90+/-0.10	0.91+/-0.25	0.81
	158-190	2.10+/-0.80	1.89+/-0.31	1.83
	193-216	0.81+/-0.30	0.73+/-0.17	0.70
	227-255	0.87+/-0.30	0.79+/-0.11	0.74
GlpF	7-34	1.25+/-0.40	1.19+/-0.08	1.16
	41-63	1.23+/-0.10	1.24+/-0.13	1.20
	69-78	0.36+/-0.03	0.33+/-0.06	0.30
	83-108	0.70+/-0.10	0.65+/-0.30	0.60
	145-167	0.48+/-0.06	0.40+/-0.07	0.38

* Structure calculations performed with strong dihedral angle restraints and no NOEs for isolated TM helices.
+ Structure calculations were performed with strong dihedral angle restraints and the maximum number of possible inter-helical NOEs.

Therefore the results of our simulations indicate that strong dihedral angle restraints should be imposed for helical membrane protein structures determined with selectively methyl protonated samples. Intra-helical NOEs are not required, although they could be used to substantiate secondary structure predictions provided by the secondary chemical shifts.

4.2 Inter-helical NOE distributions

One of the findings that arose from our study was that, in addition to the number of inter-helical NOEs, the distribution of these restraints is an important factor determining structure accuracy. For example GpA had a higher accuracy structure compared to GlyR even when a similar number of inter-helical NOEs was employed due to differences in the distribution of inter-helical NOEs (Figure 3.7). However, significant gaps in long-range structural information could be tolerated so long as they occurred in central parts of the structure that were flanked by NOE-defined regions on both sides. This was not the case if NOEs were missing from the terminal regions of transmembrane helices, indicating that it is more important to obtain long-range structural information in these regions. Unfortunately this may pose a particular problem for polytopic α-helical proteins since the ends of the transmembrane helix are known to have a higher preference for less hydrophobic residues (i.e. Lys, Arg, Phe, and Trp) while Val, Leu and Ile are preferentially found in the hydrophobic core of TM helices (70, 96). This distribution problem was also apparent for the membrane proteins from our simulations, since most of the gaps in inter-helical NOE data involved the terminal regions of the helices. This helps account for the differences in accuracies obtained for the different membrane proteins simulated, with AMC and GlpF having the poorest accuracies and the largest number of gaps in inter-helical data at helix termini.

4.3 Comparison with water-soluble α-helical proteins

While our calculations with the virtually methyl-protonated membrane proteins in this series gave rise to a range of low- to medium-range accuracy structures, this is in line with observations for similar types of simulations run for water-soluble proteins. As shown in Table 4.2, similar studies done for water-soluble α-helical proteins (structures shown in Figure 4.1) showed that a comparable level of NOEs would be obtained for selectively methyl protonated samples relative to those in our membrane protein dataset. Therefore, in spite of the more frequent occurrence of Ile, Leu and Val in membrane proteins relative to water soluble proteins, additional structural information does not seem to be available for these membrane proteins. This reflects the fact that the hydrophobic cores of water soluble proteins are actually very similar to those of membrane proteins in terms of packing and hydrophobicity (97, 98). Consequently, the backbone accuracy rmsd values for the water-soluble proteins were similar to the ones determined from this study.

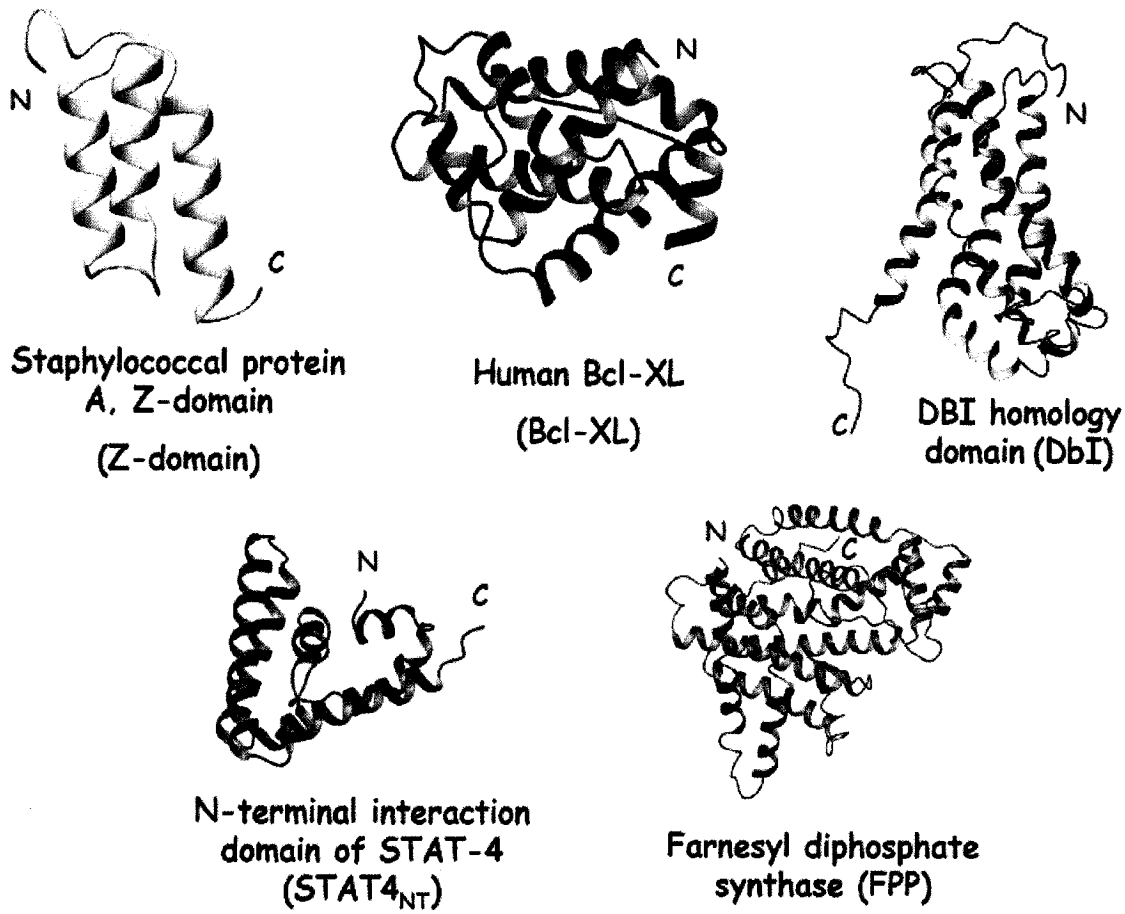

Figure 4.1 – Ribbon representations of the water-soluble proteins used to evaluate the ability of specific methyl protonation to accurately determine their structures (as summarized in Table 4.2).

Table 4.2 Accuracy rmsd of proteins using the ILV - specific labeling strategy

Protein*	No. of helical residues	NOEs per helical residue	IntraNOEs	InterNOEs	Value of D.A restraints used in simulations (ϕ,ψ)	Accuracy RMSD†	Reference
GpA	54	3.04	2.59	0.44	-57°+/-20°; -47°+/-30°	2.58+/-0.80	this study
GlyR	47	4.43	4.23	0.19	-57°+/-20°; -47°+/-30°	4.92+/-0.99	this study
GlpF	181	4.33	3.52	0.80	-57°+/-20°; -47°+/-30°	5.37+/-0.72	this study
Halor	188	5.07	4.03	1.04	-57°+/-20°; -47°+/-30°	3.18+/-0.20	this study
AMC	168	4.40	3.82	0.59	-57°+/-20°; -47°+/-30°	5.87+/-0.67	this study
Z-domain	45	5.44	4.23	1.18	-60°+/-30°; -40°+/-40°	1.80+/-0.40[@]	(69)
$STAT4_{NT}$	100	3.46	2.87	0.59	TALOS output	2.90+/-n/a	(68)
Bcl-XL	96	5.80	n/a[#]	n/a	TALOS output	3.87+/-0.42[&]	(67)
Dbl	142	2.40	n/a	n/a	TALOS output	5.83+/-0.75	(67)
FPP	228	n/a	n/a	n/a	-70°+/-50°; -50°+/-50°	7.54+/-2.81[%]	(63)

* Dbl, Dbl homology domain of Trio; Z-domain, Z-domain from *Staphylococcal* protein A; Bcl-XL, antiapoptotic protein; FPP, farnesyl diphosphate synthase; $STAT4_{NT}$, amino-terminal protein interaction domain of STAT-4

n/a, not available

† Backbone accuracy rmsd of the 10 lowest energy structures for the secondary structure region (i.e. the α-helices); all protein simulations except the ones in this study and FPP used hydrogen bond distance restraints

@ Rmsd for core residues: 20-31, 38-49, 54-68; Rmsd for residues 14-71 – 2.9+/-0.40 Å

& This study also included some NOEs from the β-protons of Val and γ-protons of Ile and Leu

% Rmsd of the global fold; precision of the two subdomains independently is 3.18+/-0.41 Å and 5.37+/-1.68 Å

4.4 Future sources of structural data to be investigated

Given the importance of the dihedral angle restraint when only a sparse set of NOEs is available, future efforts to determine these structures should focus on the accurate identification of these helical regions since the NOEs themselves do not provide enough information to fold the helices. Although secondary chemical shifts can predict secondary structure with ~90-95% accuracy (28), other sources of information can also be used to minimize the chance that a strong dihedral angle restraint is erroneously imposed. For example, hydropathy analysis of the primary sequence can identify regions that are likely to fold into transmembrane helices (89). When combined with sequence alignment information these helices can be identified with 90 % accuracy (89). Agreement between hydropathy profiles and secondary shift information should help to identify potential problems with a helical assignment. In addition, $^{1}H^{N}$-^{15}N residual dipolar couplings exhibit specific patterns for α-helical regions (99) and can even be combined with ^{15}N chemical shifts to identify relative helix orientations (100). Therefore there are many potential sources of data to help identify the helical regions of membrane proteins which can be used together to ensure the accuracy of secondary structure assignment.

In this work dihedral angle values corresponding to a geometrically ideal α-helix were used to define the structures, as has been done in previous studies (79-81). However, the actual (ϕ ,ψ) values in these membrane proteins may be closer to ideal coiled-coil values of (-65°, -40°), which also happens to represent the average dihedral angle values for α-helices in α-helical water-soluble proteins. These values are also similar to those found in structures of membrane proteins comprised of bundles of coiled helices, including that of

halorhodopsin. In contrast, transporters and channels exhibit an average value of (-60°, -45°) which falls within the narrowly defined range of acceptable dihedral angles used in our simulations. This would suggest that ideal values were already being used for AMC and GlpF, but may have been inappropriate for halorhodopsin. Nonetheless, the most ideal scenario would be one where the dihedral angle values could be accurately defined while maintaining a narrow window of acceptable values during the structure calculation. While some approaches have been developed to extract more exact predictions for dihedral angles based on secondary chemical shifts (e.g. TALOS (31)), or dipolar couplings (e.g. PISEMA (101)), the uncertainty associated with these methods is still larger than the boundaries used in the structure calculations. Therefore it remains to be determined as to whether the utilization of these approaches will help to improve the accuracies of the overall structures.

While the addition of $^{1}H^{N}$-^{15}N RDCs did not give rise to a significant improvement in the accuracy of structures determined in our investigation, it has been established that $^{1}H^{N}$-^{15}N RDCs alone can improve structure quality (42, 91). However, many successful examples where $^{1}H^{N}$-^{15}N RDCs helped to refine the structure were with proteins that contained both α-helical and β-sheet secondary structures. It has been noted that β-sheets usually show more significant improvement with RDCs since the long-range information can help to define the twist of the sheet, a characteristic that is not well described by the local information provided by NOEs (42). In contrast, amide NH dipoles in a helix tend to be almost collinear, giving rise to a narrow distribution of dipolar couplings that does not help to provide new structural information with respect to the secondary structure itself. However, it is well established that the measurement of multiple dipolar couplings per peptide bond can give rise to a more significant improvement in structure quality,

particularly for of α-helical structures (19, 54). An example of this was provided for the Z-domain, which is small α-helical protein structure that was determined for a selectively methyl protonated sample (Table 4.2). The addition of $^{1}H^{N}$-^{15}N and $^{13}C\alpha$-^{13}CO RDCs improved the accuracy by 0.4 Å relative to the same structure determination performed without these RDCs. Therefore, although other types of RDCs are not often acquired, they should be more routinely included in future structure determinations, particularly when sparse datasets for α-helical proteins are being used.

One potential application for RDCs could be to help identify kinks in the α-helix which could then be accommodated in the structure calculation with looser dihedral angle boundaries. Kinks arises from a break in the usual hydrogen bonding network and are common in membrane proteins (102). Many of these helix deformations occur at proline residues since there is a loss of a backbone hydrogen bond with backbone carbonyl oxygens 3 and 4 residues downstream in the sequence. In addition there are steric interactions of the proline ring with the carbonyl oxygen in the preceding turn of the helix that favor deviations from ideal helix geometries (81). Helix distortions also occur at non-proline residues, with certain sequence motifs in membrane proteins being associated with these disruptions in helical structure (81, 103). Therefore, it may be possible to use sequence information combined with secondary shifts to identify these kinked regions. In addition, $^{1}H^{N}$-^{15}N RDC correlations with ^{15}N chemical shifts (100, 104), not only help to identify discontinuities in the helix structure but also have the potential to help define the dihedral angles more accurately in this region. Given the significant contributions to inaccuracies arising from kinked helices any approach that can help identify these regions should help prevent incorrect restraints from being applied.

4.5 Outlook on structure determination of α-helical membrane proteins

Results from our studies as well as those reviewed above suggest that the success of the specific methyl protonation labeling strategy for structure determination of α-helical membrane proteins will hinge on the ability to acquire a uniform distribution of inter-helical NOEs. However, if the protein fold does not permit this to be obtained from a methyl protonated sample other labeling strategies should be utilized. For example, it has been shown that the addition of protonated Phe and Tyr to selectively methyl protonated samples produces a useful source of supplemental NOEs that can significantly improve the quality of structures obtained (40, 67, 105). In the case of the helical protein Bcl-xl, it was shown that the rmsd of structures calculated with selectively protonated samples from the crystal structure dropped from 3.87+/-0.42 down to 2.61+/-0.12 Å when the protonated Phe and Tyr were included (67). Nonetheless, it can be difficult to obtain assignments for aromatic protons, particularly when a large number of aromatic residues are present. Therefore this labeling strategy may only be of limited utility for polytopic α-helical membrane proteins since these residues are often found close to the ends of TM helices (70).

A particularly promising form of selective protonation has recently been demonstrated (67, 70) that stereo-selectively incorporates just a single proton into methyl and methylene groups, and at alternating sites in aromatic groups, with the remaining non-exchangeable protons being replaced with deuterium. In addition, ^{12}C is selectively incorporated into the perdeuterated sites of Leu and Val methyl groups and the deuterated sites of aromatic groups (106, 107). This labeling pattern preserves through-bond

connectivity information needed for backbone and side chain assignment and removes the sources of spin diffusion to improve the accuracy of proton distance measurements. This method has been successfully demonstrated for the 17 kDa calmodulin (46) and the 41 kDa maltose binding protein (43) and has shown to give rise to well-defined ensembles, although interestingly, the accuracies of the ensembles were not reported. However, at the present time the selectively labeled amino acids required for this labeling strategy are not commercially available and can only be obtained by an extensive series of chemical syntheses (108, 109). In addition, incorporation of these selectively labeled amino acids into proteins requires the use of cell-free in vitro translation, a technique that has yet to be generally adopted for the generation of NMR samples (109, 110). Although the application of cell-free translation to membrane proteins is currently being investigated (111) there are a still a significant number of developments that need to occur before this approach can be generally applied to solution NMR of membrane proteins.

One type of labeling that has better potential for more general application to structure determination of α-helical membrane proteins by solution NMR is paramagnetic relaxation enhancement (PRE). The method is used to obtain more distance restraints by introducing spin-labels into specific sites and measuring the extent and location of broadened resonances. Depending on the spin label being employed PRE is able to cause substantial broadening of resonances for protons that are within 15-25Å of the spin label (112, 113). The extent of broadening can be translated into distances that are included as input for structure determination. This strategy has already been applied to the structure determination of a few membrane proteins: the α-helical membrane-integrating protein (Mistic) and the β-barrel membrane protein OmpA (114, 115) and has improved the

precision of the resulting structures. Based on the results of our study, it is possible that these long-range distances will be sufficient to accurately define the structure of a polytopic α-helical membrane protein so long as the helical regions are accurately defined. Ultimately, this strategy may open the door to NMR structure determination of polytopic α-helical membrane proteins which can help provide new insights into membrane protein structures and their helical interactions.

CLAIMS TO ORIGINAL RESEARCH

Evaluated the accuracy of solution NMR structures that can be obtained for polytopic α-helical membrane proteins with the selective methyl protonation strategy.

The main findings from this work are:

1) Intra-helical NOE restraints do not improve structure accuracy.

2) If helices can be accurately identified, imposition of narrowly defined dihedral restraints significantly improves the accuracy of structures.

3) Distribution of inter-helical distance restraints is one of the most important factors determining structure accuracy.

REFERENCES

1. Popot, J.L., and D.M. Engelman. 1990. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry* 29:4031-4037.

2. Chasman, D. 2003. Protein Structure - Determination, Analysis, and Applications for Drug Discovery. Marcel Dekker Inc., New York. 606 pp.

3. Flohr, S., M. Kurz, E. Kostenis, A. Brkovich, A. Fournier, and T. Klabunde. 2002. Identification of nonpeptidic urotensin II receptor antagonists by virtual screening based on a pharmacophore model derived from structure-activity relationships and nuclear magnetic resonance studies on urotensin II. *J Med Chem* 45:1799-1805.

4. Palczewski, K., T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B.A. Fox, I. Le Trong, D.C. Teller, T. Okada, R.E. Stenkamp, M. Yamamoto, and M. Miyano. 2000. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* 289:739-745.

5. Ames, R.S., H.M. Sarau, J.K. Chambers, R.N. Willette, N.V. Aiyar, A.M. Romanic, C.S. Louden, J.J. Foley, C.F. Sauermelch, R.W. Coatney, Z. Ao, J. Disa, S.D. Holmes, J.M. Stadel, J.D. Martin, W.S. Liu, G.I. Glover, S. Wilson, D.E. McNulty, C.E. Ellis, N.A. Elshourbagy, U. Shabon, J.J. Trill, D.W. Hay, E.H. Ohlstein, D.J. Bergsma, and S.A. Douglas. 1999. Human urotensin-II is a potent vasoconstrictor and agonist for the orphan receptor GPR14. *Nature* 401:282-286.

6. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235-242.

7. White, S.H., and J.E. Gouaux. 2005. Membrane proteins--pumping along. *Curr. Opin. Struct. Biol.* 15:375-377.

8. Montelione, G.T., D. Zheng, Y.J. Huang, K.C. Gunsalus, and T. Szyperski. 2000. Protein NMR spectroscopy in structural genomics. *Nat. Struct. Biol.* 7 Suppl:982-985.

9. White, S.H. 2004. The progress of membrane protein structure determination. *Protein Sci.* 13:1948-1949.

10. Girvin, M.E., V.K. Rastogi, F. Abildgaard, J.L. Markley, and R.H. Fillingame. 1998. Solution structure of the transmembrane H+-transporting subunit c of the F1F0 ATP synthase. *Biochemistry* 37:8817-8824.

11. Debye, P. 1945. Polar Molecules. Dover publications, New York.

12. Pervushin, K., R. Riek, G. Wider, and K. Wuthrich. 1997. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci. U S A* 94:12366-12371.

13. Hwang, P.M., W.Y. Choy, E.I. Lo, L. Chen, J.D. Forman-Kay, C.R. Raetz, G.G. Prive, R.E. Bishop, and L.E. Kay. 2002. Solution structure and dynamics of the outer membrane enzyme PagP by NMR. *Proc. Natl. Acad. Sci. U S A* 99:13560-13565.

14. Arora, A., F. Abildgaard, J.H. Bushweller, and L.K. Tamm. 2001. Structure of outer membrane protein A transmembrane domain by NMR spectroscopy. *Nat. Struct. Biol.* 8:334-338.

15. Tamm, L.K., F. Abildgaard, A. Arora, H. Blad, and J.H. Bushweller. 2003. Structure, dynamics and function of the outer membrane protein A (OmpA) and influenza hemagglutinin fusion domain in detergent micelles by solution NMR. *FEBS Lett.* 555:139-143.

16. Fernandez, C., and K. Wuthrich. 2003. NMR solution structure determination of membrane proteins reconstituted in detergent micelles. *FEBS Lett.* 555:144-150.

17. Hwang, P.M., R.E. Bishop, and L.E. Kay. 2004. The integral membrane enzyme PagP alternates between two dynamically distinct states. *Proc. Natl. Acad. Sci. U S A* 101:9618-9623.

18. Fernandez, C., K. Adeishvili, and K. Wuthrich. 2001. Transverse relaxation-optimized NMR spectroscopy with the outer membrane protein OmpX in dihexanoyl phosphatidylcholine micelles. *Proc. Natl. Acad. Sci. U S A* 98:2358-2363.

19. Cierpicki, T., B. Liang, L.K. Tamm, and J.H. Bushweller. 2006. Increasing the accuracy of solution NMR structures of membrane proteins by application of residual dipolar couplings. High-resolution structure of outer membrane protein A. *J. Am. Chem. Soc.* 128:6947-6951.

20. Claridge, T. 1999. High-Resolution NMR Techniques in Organic Chemistry. Elsevier Science Ltd, London.

21. Cavanagh, J., Fairbrother W.J., Palmer A.G., and N.J. Skeleton 1996. Protein NMR Spectroscopy. Principles and Practice. Academic Press, California.

22. Kanelis, V., J.D. Forman-Kay, and L.E. Kay. 2001. Multidimensional NMR methods for protein structure determination. *IUBMB Life* 52:291-302.

23. Clore, G.M., and A.M. Gronenborn. 1998. Determining the structures of large proteins and protein complexes by NMR. *Trends Biotechnol.* 16:22-34.

24. Sattler, M., J. Schleucher, and C. Griesinger. 1999. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Res. Spectros.* 34:93 - 158.

25. Solomon, I. 1955. Relaxation processes in a system of two spins. *Phys. Rev* 99:559-565.

26. Kuntz, I.D., Kosen, P.A. and Craig, E.C. 1991. Amide chemical shifts in many helices in peptides and proteins are periodic. *J. Am. Chem. Soc.* 113:1406-1408.

27. Ösapay, K., Case, D.A. 1994. Analysis of proton chemical shifts in regular secondary structure of proteins *J. of Biomol. NMR* 4:215-230.

28. Wishart, D.S., B.D. Sykes, and F.M. Richards. 1992. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* 31:1647-1651.

29. Wishart, D.S., and B.D. Sykes. 1994. The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. *J. Biomol. NMR.* 4:171-180.

30. Wishart, D.S., and B.D. Sykes. 1994. Chemical shifts as a tool for structure determination. *Methods Enzymol* 239:363-392.

31. Cornilescu, G., F. Delaglio, and A. Bax. 1999. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR.* 13:289-302.

32. Bax, A., G. Kontaxis, and N. Tjandra. 2001. Dipolar couplings in macromolecular structure determination. *Methods Enzymol.* 339:127-174.

33. Prestegard, J.H., H.M. al-Hashimi, and J.R. Tolman. 2000. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Q. Rev. Biophys.* 33:371-424.

34. Ottiger, M., Bax, A. 1998. Determination of relative N-HN, N-C', CA-C', and CA-HA effective bond lengths in a protein by NMR in dilute liquid crystalline phase. *J. Am. Chem. Soc.* 120:12334-12341.

35. Prestegard, J.H., C.M. Bougault, and A.I. Kishore. 2004. Residual dipolar couplings in structure determination of biomolecules. *Chem Rev* 104:3519-3540.

36. Lipsitz, R.S., and N. Tjandra. 2004. Residual dipolar couplings in NMR structure analysis. *Annu Rev Biophys Biomol Struct* 33:387-413.

37. Ruckert, M., Otting, G. 2000. Alignment of biological macromolecules in novel nonionic liquid crystalline medica for NMR experiments. *J. Am. Chem. Soc.* 122:7793-7797.

38. Bax, A. 2003. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.* 12:1-16.

39. Gaemers, S., and A. Bax. 2001. Morphology of three lyotropic liquid crystalline biological NMR media studied by translational diffusion anisotropy. *J. Am. Chem. Soc.* 123:12343-12352.

40. Clore, G.M., Starich M.R., Bewley, C.A., Cai, M.L., Kuszewski, J. 1999. Impact of residual dipolar couplings on the accuracy of NMR structures determined from a minimal number of NOE restraints. *J. Am. Chem. Soc.* 121:6513-6514.

41. Choy, W.Y., M. Tollinger, G.A. Mueller, and L.E. Kay. 2001. Direct structure refinement of high molecular weight proteins against residual dipolar couplings and carbonyl chemical shift changes upon alignment: an application to maltose binding protein. *J. Biomol. NMR.* 21:31-40.

42. Garcia-Mayoral, M.F., D. Pantoja-Uceda, J. Santoro, A. Martinez del Pozo, J.G. Gavilanes, M. Rico, and M. Bruix. 2005. Refined NMR structure of alpha-sarcin by 15N-1H residual dipolar couplings. *Eur Biophys. J.* 34:1057-1065.

43. Mueller, G.A., W.Y. Choy, D. Yang, J.D. Forman-Kay, R.A. Venters, and L.E. Kay. 2000. Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. *J. Mol. Biol.* 300:197-212.

44. Ramirez, B.E., O.N. Voloshin, R.D. Camerini-Otero, and A. Bax. 2000. Solution structure of DinI provides insight into its mode of RecA inactivation. *Protein Sci* 9:2161-2169.

45. Howell, S.C., M.F. Mesleh, and S.J. Opella. 2005. NMR structure determination of a membrane protein with two transmembrane helices in micelles: MerF of the bacterial mercury detoxification system. *Biochemistry* 44:5196-5206.

46. Chou, J.J., S. Li, C.B. Klee, and A. Bax. 2001. Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains. *Nat Struct Biol* 8:990-997.

47. Ramirez, B.E., Bax, A. 1998. Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. *J. Am. Chem. Soc.* 120:9106-9107.

48. Brunger, A.T., P.D. Adams, and L.M. Rice. 1997. New applications of simulated annealing in X-ray crystallography and solution NMR. *Structure* 5:325-336.

49. Kirkpatrick, S., Gelatt,C.D., Vecchi,M.P. 1983. Optimization by Simulated Annealing. *Science* 220:671-680.

50. Brunger, A.T., and M. Nilges. 1993. Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q Rev Biophys* 26:49-125.

51. Stein, E.G., L.M. Rice, and A.T. Brunger. 1997. Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *J. Magn. Reson.* 124:154-164.

52. Schwaiger, M., M. Lebendiker, H. Yerushalmi, M. Coles, A. Groger, C. Schwarz, S. Schuldiner, and H. Kessler. 1998. NMR investigation of the multidrug transporter EmrE, an integral membrane protein. *Eur J Biochem* 254:610-619.

53. Sonnichsen, F.D., J.E. Van Eyk, R.S. Hodges, and B.D. Sykes. 1992. Effect of trifluoroethanol on protein secondary structure: an NMR and CD study using a synthetic actin peptide. *Biochemistry* 31:8790-8798.

54. Oxenoid, K., and J.J. Chou. 2005. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc. Natl. Acad. Sci. U S A* 102:10870-10875.

55. MacKenzie, K.R., J.H. Prestegard, and D.M. Engelman. 1997. A transmembrane helix dimer: structure and implications. *Science* 276:131-133.

56. Goto, N.K., and L.E. Kay. 2000. New developments in isotope labeling strategies for protein solution NMR spectroscopy. *Curr. Opin. Struct. Biol.* 10:585-592.

57. Gardner, K.H., and L.E. Kay. 1998. The use of 2H, 13C, 15N multidimensional NMR to study the structure and dynamics of proteins. *Annu Rev Biophys Biomol Struct* 27:357-406.

58. Markley, J.L., Putter, I, Jardetzky, O. 1968. High-resolution nuclear magentic resonance studies of selectively deuterated staphylococcal nuclease. *Science* 161:1249-1251.

59. Crespi, H.L., Katz J.J. 1969. High resolution proton magnetic resonance studies of fully deuterated and isotope hybrid proteins. *Nature* 560-562.

60. Torchia, D.A., S.W. Sparks, and A. Bax. 1988. NMR signal assignments of amide protons in the alpha-helical domains of staphylococcal nuclease. *Biochemistry* 27:5135-5141.

61. Pervushin, K., Riek, R., Wider, G., Wuthrich, K. 1998. Transverse relaxation-optimized spectroscopy (TROSY) for NMR studies of aromatic spin systems in 13C-labeled proteins. *J. Am. Chem. Soc.* 94:12366-12371.

62. Gardner, K.H., M.K. Rosen, and L.E. Kay. 1997. Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry* 36:1389-1401.

63. Rosen, M.K., K.H. Gardner, R.C. Willis, W.E. Parris, T. Pawson, and L.E. Kay. 1996. Selective methyl group protonation of perdeuterated proteins. *J. Mol. Biol.* 263:627-636.

64. Janin, J., S. Miller, and C. Chothia. 1988. Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* 204:155-164.

65. Kay, L.E., Bull, T.E., Nicholson, L.K., Griesinger, C., Schwalbe, H., Bax, A., Torchia, D.A. 1992. The measurement of heteronuclear transverse relaxation-times in AX3 spin systems via polarization transfer techniques. *J. Magn. Reson.* 100:538-558.

66. Goto, N.K., K.H. Gardner, G.A. Mueller, R.C. Willis, and L.E. Kay. 1999. A robust and cost-effective method for the production of Val, Leu, Ile (delta 1) methyl-protonated 15N-, 13C-, 2H-labeled proteins. *J. Biomol. NMR.* 13:369-374.

67. Medek, A., E.T. Olejniczak, R.P. Meadows, and S.W. Fesik. 2000. An approach for high-throughput structure determination of proteins by NMR spectroscopy. *J. Biomol. NMR.* 18:229-238.

68. Gaponenko, V., S.P. Sarma, A.S. Altieri, D.A. Horita, J. Li, and R.A. Byrd. 2004. Improving the accuracy of NMR structures of large proteins using pseudocontact shifts as long-range restraints. *J. Biomol. NMR.* 28:205-212.

69. Zheng, D., Y.J. Huang, H.N. Moseley, R. Xiao, J. Aramini, G.V. Swapna, and G.T. Montelione. 2003. Automated protein fold determination using a minimal NMR constraint strategy. *Protein Sci.* 12:1232-1246.

70. Adamian, L., V. Nanda, W.F. DeGrado, and J. Liang. 2005. Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins. *Proteins* 59:496-509.

71. Ma, D., Z. Liu, L. Li, P. Tang, and Y. Xu. 2005. Structure and dynamics of the second and third transmembrane domains of human glycine receptor. *Biochemistry* 44:8790-8800.

72. Fu, D., A. Libson, L.J. Miercke, C. Weitzman, P. Nollert, J. Krucinski, and R.M. Stroud. 2000. Structure of a glycerol-conducting channel and the basis for its selectivity. *Science* 290:481-486.

73. Kolbe, M., H. Besir, L.O. Essen, and D. Oesterhelt. 2000. Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution. *Science* 288:1390-1396.

74. Pebay-Peyroula, E., C. Dahout-Gonzalez, R. Kahn, V. Trezeguet, G.J. Lauquin, and G. Brandolin. 2003. Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. *Nature* 426:39-44.

75. Koradi, R., M. Billeter, and K. Wuthrich. 1996. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* 14:51-55, 29-32.

76. Venters, R.A., W.J. Metzler, L.D. Spicer, L. Mueller, and B.T. Farmer, 2nd. 1995. Use of 1HN - 1HN NOEs to determine protein global folds in perdeuterated proteins. *J. Am. Chem. Soc.* 117:9592-9593.

77. Markley, J.L., A. Bax, Y. Arata, C.W. Hilbers, R. Kaptein, B.D. Sykes, P.E. Wright, and K. Wuthrich. 1998. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J. Mol. Biol.* 280:933-952.

78. Brunger, A.T., P.D. Adams, G.M. Clore, W.L. DeLano, P. Gros, R.W. Grosse-Kunstleve, J.S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson, and G.L. Warren. 1998. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta. Crystallogr. D. Biol. Crystallogr.* 54:905-921.

79. Park, S.H., A.A. Mrse, A.A. Nevzorov, M.F. Mesleh, M. Oblatt-Montal, M. Montal, and S.J. Opella. 2003. Three-dimensional structure of the channel-forming trans-membrane domain of virus protein "u" (Vpu) from HIV-1. *J. Mol. Biol.* 333:409-424.

80. Alexandrescu, A.T. 2004. Strategy for supplementing structure calculations using limited data with hydrophobic distance restraints. *Proteins* 56:117-129.

81. Bright, J.N., Sansom, M.S.P. 2003. The Flexing/Twirling Helix: Exploring the Flexibility about Molecular Hinges formed by proline and glycine motifs in transmembrane helices. *J. Phys. Chem. B* 107:627-636.

82. Blackledge, M. 2005. Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings *Prog. Nucl. Magn. Res. Spec.* 46:26-61.

83. Zweckstetter, M., and A. Bax. 2000. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR *J. Am. Chem. Soc.* 122 3791-3792.

84. Yates, D.S., Moore, D.S., Starnes, D.S. 2006. The Practice of Statistics. W H Freeman & Co, New York 728 pp.

85. Fossi, M., H. Oschkinat, M. Nilges, and L.J. Ball. 2005. Quantitative study of the effects of chemical shift tolerances and rates of SA cooling on structure calculation from automatically assigned NOE data. *J. Magn. Reson.* 175:92-102.

86. Smith, B.O., Ito, Y., Raine, A., Teichmann, S., Ben-Tovim, L. 1996. An approach to structure determination using limited NMR data from larger proteins selectively protonated at specific residue types. *J. Biomol. NMR.* 8:360-368.

87. Venters, R.A., Huang, C-C., Spicer, L.D., Mueller, L., Farmer B.T. II. 1995. Use of 1HN-1HN NOEs to determine protein global folds in perdeuterated proteins. *J. Am. Chem. Soc.* 117:9592-9593.

88. Li, W., Y. Zhang, D. Kihara, Y.J. Huang, D. Zheng, G.T. Montelione, A. Kolinski, and J. Skolnick. 2003. TOUCHSTONEX: protein structure prediction with sparse NMR data. *Proteins* 53:290-306.

89. Kyte, J., Doolittle, RF. 1982. A simple method for displaying the hydropathic character of a protien. *J. Mol. Biol.* 157:105.

90. Lee, S., M.F. Mesleh, and S.J. Opella. 2003. Structure and dynamics of a membrane protein in micelles from three solution NMR experiments. *J. Biomol. NMR.* 26:327-334.

91. Markus, M.A., R.B. Gerstner, D.E. Draper, and D.A. Torchia. 1999. Refining the overall structure and subdomain orientation of ribosomal protein S4 delta41 with dipolar couplings measured by NMR in uniaxial liquid crystalline phases. *J. Mol. Biol.* 292:375-387.

92. Jensen, P.R., J.B. Axelsen, M.H. Lerche, and F.M. Poulsen. 2004. Improvement of hydrogen bond geometry in protein NMR structures by residual dipolar couplings--an assessment of the interrelation of NMR restraints. *J. Biomol. NMR.* 28:31-41.

93. Fuentes, G., A.J. Nederveen, R. Kaptein, R. Boelens, and A.M. Bonvin. 2005. Describing partially unfolded states of proteins from sparse NMR data. *J. Biomol. NMR.* 33:175-186.

94. Li, W., Y. Zhang, and J. Skolnick. 2004. Application of sparse NMR restraints to large-scale protein structure prediction. *Biophys. J.* 87:1241-1248.

95. Wuthrich, K. 1986. NMR of Proteins and Nucleic Acids, U.S.A. 320 pages

96. Eyre, T.A., L. Partridge, and J.M. Thornton. 2004. Computational analysis of alpha-helical membrane protein structure: implications for the prediction of 3D structural models. *Protein Eng. Des. Sel.* 17:613-624.

97. Gimpelev, M., L.R. Forrest, D. Murray, and B. Honig. 2004. Helical packing patterns in membrane and soluble proteins. *Biophys. J.* 87:4075-4086.

98. Eilers, M., A.B. Patel, W. Liu, and S.O. Smith. 2002. Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys. J.* 82:2720-2736.

99. Mesleh, M.F., and S.J. Opella. 2003. Dipolar Waves as NMR maps of helices in proteins. *J. Magn. Reson.* 163:288-299.

100. Nevzorov, A.A., M.F. Mesleh, and S.J. Opella. 2004. Structure determination of aligned samples of membrane proteins by NMR spectroscopy. *Magn Reson Chem* 42:162-171.

101. Nevzorov, A.A., and S.J. Opella. 2003. Structural fitting of PISEMA spectra of aligned proteins. *J. Magn. Reson.* 160:33-39.

102. Ubarretxena-Belandia, I., and D.M. Engelman. 2001. Helical membrane proteins: diversity of functions in the context of simple architecture. *Curr. Opin. Struct. Biol.* 11:370-376.

103. Riek, R.P., I. Rigoutsos, J. Novotny, and R.M. Graham. 2001. Non-alpha-helical elements modulate polytopic membrane protein architecture. *J. Mol. Biol.* 306:349-362.

104. Wu, C.H.R., A.; Opella, S.J. 1994. High-Resolution Heteronuclear Dipolar Solid-State NMR Spectroscopy. *J. Magn. Reson. A* 109:270.

105. Yu, L.S., CH.; Song DY.; Shen JW.; Xu N.; Gunasekera A.; Hajduk PJ.; Olejniczak ET 2005 Nuclear magnetic resonance structural studies of a potassium channel-charybdotoxin complex *Biochemistry* 44:15834-15841

106. Kainosho, M., T. Torizawa, Y. Iwashita, T. Terauchi, A. Mei Ono, and P. Guntert. 2006. Optimal isotope labelling for NMR protein structure determinations. *Nature* 440:52-57.

107. Ikeya, T., T. Terauchi, P. Guntert, and M. Kainosho. 2006. Evaluation of stereo-array isotope labeling (SAIL) patterns for automated structural analysis of proteins with CYANA. *Magn Reson Chem* 44 Spec No:S152-157.

108. Torizawa, T., M. Shimizu, M. Taoka, H. Miyano, and M. Kainosho. 2004. Efficient production of isotopically labeled proteins by cell-free synthesis: a practical protocol. *J. Biomol. NMR.* 30:311-325.

109. Zubay, G. 1973. In vitro synthesis of protein in microbial systems. *Annu. Rev. Genet.* 7:267-287.

110. Kigawa, T., Y. Muto, and S. Yokoyama. 1995. Cell-free synthesis and amino acid-selective stable isotope labeling of proteins for NMR analysis. *J. Biomol. NMR.* 6:129-134.

111. Koglin, A., C. Klammt, N. Trbovic, D. Schwarz, B. Schneider, B. Schafer, F. Lohr, F. Bernhard, and V. Dotsch. 2006. Combination of cell-free expression and NMR

spectroscopy as a new approach for structural investigation of membrane proteins. *Magn Reson Chem* 44 Spec No:S17-23.

112. Kosen, P.A. 1989. Spin labeling of proteins. *Methods Enzymol* 177:86-121.

113. Battiste, J.L., and G. Wagner. 2000. Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. *Biochemistry* 39:5355-5365.

114. Liang, B., Bushweller, JH., Tamm, LK. 2006. Site-directed parallel spin-labelling and paramagnetic relaxation enhancement in structure determination of membrane proteins by solution NMR spectroscopy. *J. Am. Chem. Soc.* 128:4389-4397.

115. Roosild, T.P., J. Greenwald, M. Vega, S. Castronovo, R. Riek, and S. Choe. 2005. NMR structure of Mistic, a membrane-integrating protein for membrane protein expression. *Science* 307:1317-1321.

116. Koplin, R.D.K., Sorgen, P.L., Koplin, S.T.K., Torres, I.V.O., Cahill, S.M., Hicks, D.B., Grinus, L., Krulwich, T.A., and M.E. Girvin. 2003. An evaluation of detergent for NMR structural studies of membrane proteins. *J. Biomol. NMR.* 17:43-57.

APPENDIX

A.1 Script for generation of simulated ILV-NOEs in CNS format

```
      PROGRAM NOE_all

c       for do loops, residue number, total residues
        integer k,m,p,l,q,num_res,a

        integer n_atoms, n_res(15000)
          integer countCH3_CH3(4),countNH_NH(4),countNH_CH3(4)
          integer countNH_NH2(4)

          double precision x(11000), y(11000), z(11000)

          double precision d_NOE
          character(LEN=4) res(4000)
          character(LEN=4) prtc_ext(11000, 4000)

        stdi = 5
        stdo = 6
        pi = 4*datan(1.d0)

c        initialize each bin to zero
          countCH3_CH3(1) = 0
          countCH3_CH3(2) = 0
          countCH3_CH3(3) = 0
          countCH3_CH3(4) = 0

c        initialize each bin to zero
          countNH_CH3(1) = 0
          countNH_CH3(2) = 0
          countNH_CH3(3) = 0
          countNH_CH3(4) = 0

c        initialize each bin to zero
          countNH_NH(1) = 0
          countNH_NH(2) = 0
          countNH_NH(3) = 0
          countNH_NH(4) = 0

c        initialize each bin to zero
          countNH_NH2(1) = 0
          countNH_NH2(2) = 0
          countNH_NH2(3) = 0
          countNH_NH2(4) = 0

c        take the total number of atoms

          write(stdo,*)'Number of atoms??'
          read(stdi,*)n_atoms
```

```
c  -- open the pdb file and read in the residue name, number, and
identity

       open(unit=9, file='1FX8_all.list', status='old')

       do k=1, n_atoms

         read(9, FMT=120)a, res(a), prtc_ext(k,a),x(k), y(k), z(k)

120    FORMAT(i3,2x,a4,3x,a3,3x,3(f7.3,2x))

         n_res(k) = a

       enddo

       write(stdo,*)res(72)

c      now open output file to write

       open(unit=14, file='NOE_new1FX8.tbl', status='unknown')

c      write all distances smaller than 5A, between all pairs of H-s
c---A
        do p=1,188

c---B
          do k=1,n_atoms

c      this is the residue number of the reference atom
            l = n_res(k)

c      write to the file for cns
           if((l.eq.p).AND.(l.ne.n_res(k-1)))then
              write(14, FMT=300)res(l),l
300           FORMAT(/,'! assigning residue ',a4,1x, i3)
           endif

c---C
            do m=1,n_atoms


c      this is the residue number of the secondary atom
              q = n_res(m)

c-----------------------------------------------------------------------

c      this will look for amide-amide groups

c---1a
         if ((prtc_ext(k,l).eq."HN ").AND.
     $       (prtc_ext(m,q).eq."HN ").AND.((q.eq.l+1)))then


c----1b check to make sure l does not equal q - not the same  residue and
ensure that you only create the restraint once - non-redundantly
```

```
            if ((l.ne.q).AND.(l.eq.p)) then
               d_NOE = sqrt((x(k) - x(m))*(x(k) - x(m)) +(y(k) -
     $              y(m))*(y(k) - y(m)) + (z(k) - z(m))*
     $              (z(k) - z(m)))

c     now set the upper bounds for the RST file: will be 1.5 times the
pdb-distance

c     if however H-s are too close (r3 > r2 = 1.80) a supplemental
restriction:

          if(d_NOE.le.1.80) then
             d_NOE = 1.80
          endif

c     place the NOE's into their seperate bins
               if (d_NOE.le.3.0) then
                  countNH_NH(2) = countNH_NH(2) + 1
                  write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $                 3.0, 1.0, 1.2
c     debug statement
                write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $                 prtc_ext(m,q)
               endif

               if ((d_NOE.gt.3.0).AND.(d_NOE.le.4.0)) then
                 countNH_NH(3) = countNH_NH(3) + 1
                 write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $                 4.0, 1.2, 1.2
c     debug statement
                 write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $                 prtc_ext(m,q)
               endif

               if ((d_NOE.gt.4.0).AND.(d_NOE.le.5.0)) then
                 countNH_NH(4) = countNH_NH(4) + 1
                 write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $                 5.2, 1.5, 2.0
c     debug statement
                 write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $                 prtc_ext(m,q)
               endif

c     total number of restraints for amide -
               countNH_NH(1) = countNH_NH(1) + 1

250         FORMAT('assign (resid ',i3,' and name ',a4,' ) ',
     %           '(resid ',i3,' and name ',a4,' )', 3(2x,f3.1))
c----1b
            endif
c---1a
            endif

c-------------------------------------------------------------------------
-
c-------------------------------------------------------------------------
-
```

```
c     this will look for amide-amide groups

c---1a
          if ((prtc_ext(k,l).eq."HN ").AND.
     $       (prtc_ext(m,q).eq."HN ").AND.((q.eq.l+2)))then


c----1b     check to make sure l does not equal q - not the same residue
and ensure
c           that you only create the restraint once - non-redundantly
             if ((l.ne.q).AND.(l.eq.p)) then
                 d_NOE = sqrt((x(k) - x(m))*(x(k) - x(m)) +(y(k) -
     $               y(m))*(y(k) - y(m)) + (z(k) - z(m))*
     $               (z(k) - z(m)))

c     now set the upper bounds for the RST file: will be 1.5 times the
pdb-distance

c     if however H-s are too close (r3 > r2 = 1.80) a supplemental
restriction:

           if(d_NOE.le.1.80) then
              d_NOE = 1.80
           endif

c     place the NOE's into their seperate bins
                 if (d_NOE.le.3.0) then
                    countNH_NH2(2) = countNH_NH2(2) + 1
                    write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $                   3.0, 1.0, 1.2
c     debug statement
                  write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $                   prtc_ext(m,q)
                  endif

                 if ((d_NOE.gt.3.0).AND.(d_NOE.le.4.0)) then
                   countNH_NH2(3) = countNH_NH2(3) + 1
                   write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $                   4.0, 1.2, 1.2
c     debug statement
                   write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $                   prtc_ext(m,q)
                 endif

                 if ((d_NOE.gt.4.0).AND.(d_NOE.le.5.0)) then
                   countNH_NH2(4) = countNH_NH2(4) + 1
                   write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $                   5.2, 1.5, 2.0
c     debug statement
                   write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $                   prtc_ext(m,q)
                 endif

c     total number of restraints for amide -
                 countNH_NH2(1) = countNH_NH2(1) + 1
```

```
c----1b
          endif
c----1a
          endif

c-------------------------------------------------------------------------
-

c     this will look for amide-methyl groups

c---2a
          if ((prtc_ext(k,l).eq."HN ").AND.
     $              (prtc_ext(m,q).ne."HN ").AND.(q.ne.l)) then

c---2b    for methyl groups we only look at Val,Leu,Ile(QD1only)
            if ((res(q).eq."LEU ".OR.res(q).eq."VAL ")
     $          .AND.(prtc_ext(m,q).eq."QD1".OR.prtc_ext(m,q)
     $          .eq."QD2".OR.prtc_ext(m,q).eq."QG2"
     $          .OR.prtc_ext(m,q).eq."QG1").OR.(res(q).eq."ILE ".AND.
     $          prtc_ext(m,q).eq."QD1"))then

c---2c    check to make sure l does not equal q - not the same residue
and ensure
c         that you only create the restraint once - non-redundantly
             if ((l.ne.q).AND.(l.eq.p)) then
                 d_NOE = sqrt((x(k) - x(m))*(x(k) - x(m)) +(y(k) -
     $               y(m))*(y(k) - y(m)) + (z(k) - z(m))*
     $               (z(k) - z(m)))
c     now set the upper bounds for the RST file: will be 1.5 times the
pdb-distance

c     if however H-s are too close (r3 > r2 = 1.80) a supplemental
restriction:

          if(d_NOE.le.1.80) then
             d_NOE = 1.80
          endif

c     place the NOE's into their seperate bins
               if (d_NOE.le.3.0) then
                  countNH_CH3(2) = countNH_CH3(2) + 1
                  write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $                3.0, 1.0, 1.2
c     debug statement
                 write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $               prtc_ext(m,q)
               endif
               if ((d_NOE.gt.3.0).AND.(d_NOE.le.4.0)) then
                 countNH_CH3(3) = countNH_CH3(3) + 1
                 write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $               4.0, 1.2, 1.2
c     debug statement
                 write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $               prtc_ext(m,q)
               endif
               if ((d_NOE.gt.4.0).AND.(d_NOE.le.6.0)) then
                 countNH_CH3(4) = countNH_CH3(4) + 1
```

```
                write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $               5.2, 1.5, 2.0
c     debug statement
                write(stdo,*)d_NOE," ",res(l)," ",res(q)," ",
     $               prtc_ext(k,l),prtc_ext(m,q)
              endif
c     total number of restraints for methyl - methyl
              countNH_CH3(1) = countNH_CH3(1) + 1


c----2c
           endif
c---2b
          endif
c---2a
         endif

c-------------------------------------------------------------------------
-

c---3a
         if ((prtc_ext(k,l).ne."HN ").AND.
     $             (prtc_ext(m,q).ne."HN ").AND.(l.lt.q)) then

c---3b     for methyl groups we only look at Val,Leu,Ile(QD1only)
             if ((res(l).eq."LEU ".OR.res(l).eq."VAL ")
     $            .AND.(prtc_ext(k,l).eq."QD1".OR.prtc_ext(k,l)
     $           .eq."QD2".OR.prtc_ext(k,l).eq."QG2"
     $           .OR.prtc_ext(k,l).eq."QG1").OR.(res(l).eq."ILE ".AND.
     $           prtc_ext(k,l).eq."QD1"))then

c---3c     for methyl groups we only look at Val,Leu,Ile(QD1only)
             if ((res(q).eq."LEU ".OR.res(q).eq."VAL ")
     $            .AND.(prtc_ext(m,q).eq."QD1".OR.prtc_ext(m,q)
     $           .eq."QD2".OR.prtc_ext(m,q).eq."QG2"
     $           .OR.prtc_ext(m,q).eq."QG1").OR.(res(q).eq."ILE ".AND.
     $           prtc_ext(m,q).eq."QD1"))then

c----3d     check to make sure l does not equal q - not the same residue
and ensure
c          that you only create the restraint once - non-redundantly
            if ((l.ne.q).AND.(l.eq.p)) then
               d_NOE = sqrt((x(k) - x(m))*(x(k) - x(m)) +(y(k) -
     $              y(m))*(y(k) - y(m)) + (z(k) - z(m))*
     $              (z(k) - z(m)))
c     now set the upper bounds for the RST file: will be 1.5 times the
pdb-distance

c     if however H-s are too close (r3 > r2 = 1.80) a supplemental
restriction:

          if(d_NOE.le.1.80) then
             d_NOE = 1.80
          endif

c     place the NOE's into their seperate bins
              if (d_NOE.le.3.0) then
```

```
                countCH3_CH3(2) = countCH3_CH3(2) + 1
                write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $               3.0, 1.0, 1.2
c     debug statement
                write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $               prtc_ext(m,q)
              endif
              if ((d_NOE.gt.3.0).AND.(d_NOE.le.4.0)) then
                countCH3_CH3(3) = countCH3_CH3(3) + 1
                write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $               4.0, 1.2, 1.2
c     debug statement
                write(stdo,*)d_NOE,res(l)," ",res(q)," ",prtc_ext(k,l),
     $               prtc_ext(m,q)
              endif
              if ((d_NOE.gt.4.0).AND.(d_NOE.le.6.0)) then
                countCH3_CH3(4) = countCH3_CH3(4) + 1
                write(14,FMT=250)l, prtc_ext(k,l), q, prtc_ext(m,q),
     $               5.2, 1.5, 2.0
c     debug statement
                write(stdo,*)d_NOE," ", res(l)," ",res(q)," ",
     $               prtc_ext(k,l),prtc_ext(m,q)
              endif
c     total number of restraints for methyl - methyl
              countCH3_CH3(1) = countCH3_CH3(1) + 1

c---3d
           endif
c---3c
          endif
c---3b
         endif
c---3a
        endif
c-----------------------------------------------------------------

c---D
c         endif

c---C
        enddo
c---B
       enddo
c---A
      enddo


c     write the total number of NOE's for methyl-methyl proton
         write(stdo,*)'CH3-CH3 NOEs: total # is: ',countCH3_CH3(1)
         write(stdo,*)'Here are the bins: ',countCH3_CH3(2),
     $       countCH3_CH3(3),countCH3_CH3(4)
c     write the total number of NOE's for amide-amide proton
         write(stdo,*)'NH-NH(i-i+1) NOEs: total # is: ',countNH_NH(1)
         write(stdo,*)'Here are the bins: ',countNH_NH(2),
     $       countNH_NH(3),countNH_NH(4)
c     write the total number of NOE's for amide-amide proton
         write(stdo,*)'NH-NH(i-i+2) NOEs: total # is: ',countNH_NH2(1)
```

```
         write(stdo,*)'Here are the bins: ',countNH_NH2(2),
     $       countNH_NH2(3),countNH_NH2(4)
c     write the total number of NOE's for amide-methyl proton
         write(stdo,*)'NH-CH3 NOEs: total # is: ',countNH_CH3(1)
         write(stdo,*)'Here are the bins: ',countNH_CH3(2),
     $       countNH_CH3(3),countNH_CH3(4)

        END PROGRAM NOE_all
```

A.2 Script for generation of simulated dihedral angles in CNS format

```
      PROGRAM dih_cns

c
      integer k,a,p,i
        character(LEN=4)res(1000)
        character(LEN=3)prtc_ext(11000,1000)
        double precision x(11000),y(11000),z(11000)
        integer n_res(11000)

c     stdi = 5
c     stdo = 6
c        pi = 4*datan(1.d0)


c  -- open the pdb file and read in the residue name, number, and
identity

c       open(unit=9, file='helices.list', status='old')

c       do k=1,10162

c          read(9, FMT=120)a, res(a), prtc_ext(k,a),x(k), y(k), z(k)

c120    FORMAT(i5,1x,a4,3x,a3,2x,3(f7.3,2x))

c       n_res(k) = a

c       enddo


c     open output file - a dihedral angle tbl file

      open (unit=15, file='dihedral_angle.tbl', status='unknown')

c    start at the beginning of the helical residue to the end

     do p=5,261

c          if(p.eq.n_res(i).AND.p.ne.n_res(i-1))then

           write(15, FMT=401)p, p+1, p+1,
     %          p+1, 1.0, -57.0, 20.0, 2
```

```
            write(15, FMT=402)p+1, p+1, p+1,
     %           p+2, 1.0, -47.0,30.0, 2
c          endif
c        enddo
       enddo

401     FORMAT(' assign (resid ',i3,'  and name c )',
     %  ' (resid ',i3,'  and name n )',/,
     %  8x,'(resid ',i3,'  and name ca)',
     %  ' (resid ',i3,'  and name c )',3x,f3.1,2x,f6.1,1x,f6.1,1x,i1,/)

402    FORMAT(' assign (resid ',i3,'  and name n )',
     %  ' (resid ',i3,'  and name ca)',/,
     %  8x,'(resid ',i3,'  and name c )',
     %  ' (resid ',i3,'  and name n ) ',2x,f3.1,2x,f6.1,1x,f6.1,1x,i1,/)

            END PROGRAM dih_cns
```

A.3 Tables of precision and accuracy values

Protein	IntraNOEs/residue	Accuracy (Å)	Precision (Å)
GpA	1.4	2.65+/-0.51	3.69+/-1.20
	2.4	2.45+/-0.75	3.42+/-0.80
	3.4	2.58+/-0.80	3.02+/-0.73
GlyR	1.4	4.18+/-1.04	4.94+/-1.66
	2.4	4.67+/-1.09	4.81+/-1.50
	4.4	4.92+/-0.99	5.96+/-1.69
Halor	1.4	3.33+/-1.10	2.21+/-0.97
	2.4	3.01+/-0.44	1.75+/-0.62
	3.4	3.25+/-0.50	1.93+/-0.95
	4.4	3.18+/-0.20	1.95+/-0.66
GlpF	1.4	5.17+/-0.62	3.74+/-0.67
	2.4	5.05+/-0.43	3.62+/-0.53
	3.4	5.37+/-0.72	3.62+/-0.69
AMC	1.4	5.63+/-0.71	4.87+/-1.31
	2.4	5.58+/-0.74	5.48+/-1.13
	3.4	5.87+/-0.67	5.09+/-1.36

A.4 Tables of precision and accuracy values

Proteins	# of interNOEs	IntraNOEs/residue	Accuracy (Å)	Precision (Å)
GpA	24	0.7	2.65+/-0.51	3.69+/-1.20
		1.7	2.45+/-0.75	3.42+/-0.80
		2.7	2.58+/-0.80	3.02+/-0.73
	13	0.7	2.84+/-0.78	3.52+/-1.15
		1.7	2.77+/-0.37	3.33+/-0.78
		2.7	2.98+/-0.45	3.76+/-0.94
	7	0.7	3.33+/-0.57	3.88+/-1.06
		1.7	3.67+/-0.86	4.73+/-1.42
		2.7	3.99+/-1.44	5.45+/-2.06
GlyR	9	1.4	4.18+/-1.04	4.94+/-1.66
		2.4	4.67+/-1.09	4.81+/-1.50
		4.4	4.92+/-0.99	5.96+/-1.69
	5	1.4	4.17+/-0.96	5.14+/-1.50
		2.4	5.35+/-1.08	6.55+/-2.11
		3.4	4.90+/-1.55	6.69+/-2.05
Halor	195	1.1	3.33+/-1.10	2.21+/-0.97
		2.1	3.01+/-0.44	1.75+/-0.62
		3.1	3.25+/-0.50	1.93+/-0.95
		4.1	3.18+/-0.20	1.95+/-0.66
	75	1.1	4.00+/-0.30	3.87+/-0.80
		2.1	4.36+/-0.49	3.43+/-0.53
		4.1	4.69+/-0.65	3.36+/-0.60
	35	1.1	6.41+/-0.49	5.21+/-1.06
		2.1	5.60+/-0.69	4.87+/-1.19
		4.1	6.00+/-0.74	6.76+/-3.89
GlpF	145	1.1	5.17+/-0.62	3.74+/-0.67
		2.1	5.05+/-0.43	3.62+/-0.53
		3.1	5.37+/-0.72	3.62+/-0.69
	90	1.1	7.56+/-1.22	6.27+/-1.64
		2.1	7.70+/-1.44	6.33+/-1.58
		3.1	7.96+/-1.31	7.12+/-1.64
	54	1.1	10.35+/-2.77	9.93+/-3.65
		2.1	8.51+/-1.48	7.62+/-2.21
		3.1	8.30+/-1.11	7.18+/-1.78
AMC	99	1.1	5.63+/-0.71	4.87+/-1.31
		2.1	5.58+/-0.74	5.48+/-1.13
		3.1	5.87+/-0.67	5.09+/-1.36
	50	1.1	7.74+/-1.20	5.83+/-0.95
		2.1	6.71+/-0.88	5.96+/-1.11
		3.1	7.27+/-0.90	6.55+/-1.91
	25	1.1	8.38+/-0.90	7.38+/-1.57
		2.1	7.48+/-1.03	8.16+/-1.37
		3.1	7.88+/-0.71	6.86+/-1.21