Structure Determination of Outer Membrane Protein G in Native Lipids by Solid-State NMR Spectroscopy

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)
submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Joren Sebastian Retel

from Drimmelen (The Netherlands)

2016

Doctorate studies where conducted from 2010 untill 2016 under the supervision of Hartmut Oschkinat at the Department of NMR-Supported Structural Biology of the Leibniz-Intitut für Molekulare Pharmakologie.

1st reviewer: Prof. Dr. Hartmut Oschkinat
2nd reviewer: Prof. Dr. Bernd Reif

Date of defence: 24.11.2016

Acknowledgements

I would like to thank all the people in some way involved in this project or in the lab. First of all I want to thank Matthias Hiller. In many ways we have a completely complementary approach to things, which was very much needed in this project. I feel that if there would not have been such a strong personal connection between us, these years would have been a whole lot less enjoyable. At least for me, and most likely for you as well. You always listened to me and always gave a balanced answer. I appreciate that incredibly. Next up is Andy Nieuwkoop. Your dedication and perfectionism really make a difference. This project would simply not have worked without someone like you, who really freaking cares. Besides that, I had a blast with you during those summer weeks in Lyon. We should definitely go grab some Belgium beer soon. Probably we already did this by the time you are actually reading this. I would like to thank Vicky Higman for introducing me to the project those first months in 2010 and also for the helpful email correspondence whenever I was somehow completely stuck in the assignment process. Trent Franks, thank you for always helping me and the others out and for recording those first proton detected experiments on OmpG in Lyon. I guess you and Barth taught me more NMR during lunch and the beer hour than any text book. It is a pity we are not allowed to reference to “mentioned by Trent at the beer hour”. I already mentioned Barth van Rossum who was also always there to help and to add some humor to all situations. I want to explicitly thank Barth, Matthias and Andy as well for proof reading this thesis. I would like to thank Liselotte Handel for preparing OmpG samples and always making time to help. Thank you, Benjamin Bardiaux, for answering my questions surrounding the structure calculation process and for implementing new features in ARIA that made the calculation a lot more straight-forward. I also want to thank you for being interested in my assignment algorithm, when you were still around in our lab. The fact that you thought it was a good idea convinced me that I should make the initial concept code into a more mature program. I would like to thank Anne Diehl for her supportive criticism and discussions.

Thanks to all people I shared an office with trough the years: Daniela Lalli, Matthias Dorn, Tolga Helmbrecht, Anup Chowdhury, Anja Voreck, Wing Ying Chow and Masheed Sohrabi. I also want to thank all the other people that were either part of the group or I had a lot of interaction with: Sascha Lange, Arne Linden, Shakeel Ahmad Shahid, Stefan Marcovic, Madhu Nagaraj, Mônica Santos de Freitas, Nils Cremer, Marco Röben, Andrea Steuer, Sam Asami, Pascal Fricke, Max Zinke, Johanna Münkemer, Frank Eisenmenger, Kristina Rehbein, Martina Leidert, Jean-Philippe Demers, Akis Liokatis, François-Xavier Theillet and Simon Erlendsson. It was very kind of Sascha to show us (Trent, my wife and me) all the dirt tracks in northern Germany by taking the shortest route back home from Arne's wedding. Good times. Daniel Stöppler and Michel-Andreas Geiger, you have become really good friends and Helium filling would have been a lot more boring without you guys. I would like to thank Frédéric Muench and Everton D'Andrea for our colaborations.

Thank you very much, Hartmut Oschkinat, for letting me work on this topic. You have been incredibly supportive and I am really happy that you were stubborn enough to keep focus on this project although at some point it seemed like an almost impossible nut to crack. I guess you saw where we were heading, while I often did not. Also, you have been very understanding during this last period while I was writing my thesis, for which I am really thankful. I enjoyed that we could always openly discuss everything in a very informal way.

I would like to thank Bernd Reif for agreeing to be the second reviewer of this thesis.

I would also really like to thank all the people in Lyon that were involved in this project: Emeline Barbet-Massin, Jan Stanek, Loren Andreas, Tanguy Le Marchand and Guido Pintacuda. It was always a pleasure coming over to Lyon, because the atmosphere in the group was always great. I will definitely return to Lyon just for fun and will let you know.

I am grateful to my parents, brother, grandparents and the rest of the family for the support and always showing interest from the sideline. It is a blessing to have such a warm family. As my mom says: it is better to have children you only see a few times a year, but have a very solid relationship with, than to have children living next door that you can not stand. The same thing is true for parents.

I would like to thank my friends back in the Netherlands for keeping up to date on what is happening. Whenever I see you guys, it is like I never left the country. These are Roel Bolsius, Rohola Hosseini, Bouke de Jong, Robert de Jonge, Marinus Tiehatten, Eric Alberts, Bas van Altvorst, Arnoud Baalhuis, Hans Broos, Jorn Bode, Jules Witte and Wijnand Smulders. I would also like to thank some of the friends Beatriz and I made here in Berlin: Frederico Reichel, Luís Furtado, Carise Fernandes, Bartho Valk, Rianne Valk-Baerselman, Martijn Kers, Elske Baerselman, Oriol Sallés and Christian El Khoury.

Most of everyone I have to thank Beatriz. You came here with me and together we started everything from scratch, which was not always easy. You have made big concessions, changed your profession and put off many plans because I was "almost finishing my Ph.D." for a very long time. You are the one I can always rely on. You went through this process together with me and I am eternally grateful for that. Love you very much.

Contents

Appendices **137**

Chapter 1

Introduction: Structural Biology of Membrane Proteins

This thesis revolves around the determination of the three-dimensional structure of the outer membrane protein G (OmpG) from the *E. coli* bacterium using solid-state nuclear magnetic resonance (NMR) spectroscopy. Knowledge about the three-dimensional structure of biomolecules such as nucleic acids and proteins is of great interest for understanding the function and mechanism of these molecules. Therefore the field of structural biology is a central part of biochemistry and molecular biology, and for years almost every single edition of the leading scientific journals features at least one article concerning the structure and function of a biomolecular system.

Detailed knowledge of the function of biomolecules is interesting from a purely scientific point of view. However, there are many examples in which knowledge of specific biomolecular mechanisms can be exploited. For example, understanding the biomolecular mechanisms connected to diseases can lead to the development of therapies. Knowledge of how specific enzymes catalyze reactions can help engineering them to make them more suitable for applications such as the development of renewable energy sources and wastewater treatment.

Here we are interested in the important subclass of proteins that are found in the membranes of cells. Compared to proteins located in the watery environment inside the cell, membrane protein structures are relatively poorly understood even though they are of great importance in many biological mechanisms, play a role in many diseases and are targets of the majority of medicines. The major difficulty is, that two of the most commonly used techniques in structural biology (x-ray crystalography and solution NMR) are of limited use for membrane proteins. One of the big advantages of solid-state NMR is that it can be used to study membrane proteins inside their native environment, the membrane. In this study, OmpG is used to develop new solid-state NMR methods that can be used to study membrane proteins.

Biological Membranes

Some form of compartmentalization is essential for the existence of life. In a semi-closed off system, energy and organic matter can reach sufficiently high concentrations to support the rise of complex structures. Therefore, every theory on the origin of life in one way or another involves the development of spatial compartments [1]. These original compartments are often believed to be of non-biological nature such as mineral surfaces [2][3]. At some point during the early evolution of life, first biological cells were formed that possessed some kind of lipid membrane so that life could break free from these pre-existing compartments. Whether this happened before or after the last universal common ancestor and what the composition of this early membrane exactly was, is still under some debate [4][5]. Fact is that all modern cells have membranes, though.

Membranes form the barrier between the inside and the outside of cells. In eukaryotes, membranes are also present within the cell and divide it in different compartments, such as the nucleus, mitochondria, Golgi apparatus and, in the case of plants and algae, the chloroplasts. Membranes do not only function purely as separators but play an active role determining the cell's shape, locomotion, interaction with other organisms or neighboring cells and the extracellular matrix in the case of multicellular organisms. For example, proton gradients over the membranes of mitochondria and chloroplasts drive the synthesis of ATP and ion gradients over the membranes of our neurons allow them to conduct electric signals.

The lipids in membranes are arranged to form a bilayer. The hydrophobic tails are pointing towards the center of the bilayer while the hydrophilic head groups point towards the aqueous solution. Because of the various different roles that membranes can play, lipid composition in biological membranes is very diverse and varies widely between organisms and cell types [6].

In general, the majority of lipids in membranes are phospholipids. In addition, eukaryotic membranes also contain sterols influencing rigidity and permeability [7]. Furthermore, membranes of plant cells contain large amounts of glycolipids. Knowledge about the exact chemical composition of different lipids in cells is mostly obtained by a combination of mass spectroscopy and liquid chromatography and generated its own "omics" field, logically called lipidomics [8][9]. This is a very complex field since there is no simple basic paradigm like in the study of proteins and nucleic acids, in which case there is a more or less direct transcription/translation between DNA sequences and RNA/proteins. In eukaryotes, thousands of different lipids can be present based on the combination of different head groups and chain lengths [10][11]. An intriguing difference exists between the phospholipid composition in the membranes of bacteria and eukaryotes on the one hand, and archea on the other [12][13]. The most important distinction here is that in archaea the opposite glycerol stereoisomer is used to synthesize the phospholipid backbone as compared to the other two branches of life. The fact that there is such a large number of lipids and that organisms spend expensive resources to maintain this variety, indicates its functional relevance. The mix of lipids making up the membrane directly influences properties such as its flexibility, curvature, permeability and interaction with membrane proteins [14].

The combination of lipids does not only vary between different membranes but also between the two leaflets that compose the lipid bilayer. For instance, for the outer membrane of *E. coli.* the outer leaflet is composed of lipopolysaccharides (LPS), and the inner leaflet of the more usual phospholipids of which the majority is phosphatidylethanolamine (PE) (75%), phosphatidylglycerol (PG) (20%) and cardiolipin [15][16][17][18].

Membrane Proteins

Besides lipids, the membrane consists for a large part of membrane proteins. In mouse liver the fraction of protein by weight is about 45% and in *E. coli* this is 75% [19]. Some membrane proteins are outside of membrane and are anchored to the membrane with a covalently bound lipid tag or hydrophobic α-helix. These proteins are known as peripheral or monotopic membrane proteins. Others span the entire bilayer and have parts of the proteins stick out on both sides of the membrane. These proteins are called integral membrane proteins. In turn, there are two large classes of integral membrane proteins: proteins consistent of multiple membrane spanning α-helices and proteins that form β-barrels. The reason that all integral membrane proteins have one of these two topologies is that it minimizes the number of unfulfilled hydrogen bonds. This is very important, because for membrane proteins there is a high energy penalty for unfulfilled hydrogen bonds within the protein, since there are virtually no hydrogen bond partners present in the non-polar part of the membrane [20]. Therefore all CO and NH pairs in the protein backbone should hydrogen bond within the protein itself, leaving only these two basic topologies. β-sheets naturally roll up into closed barrels, otherwise non-hydrogen bonded residues would be present on both extremes of the sheet.

In most organisms, 20-30% of the genes code for membrane proteins [21]. Membrane proteins play a role in numerous important biological events. Receptors transmit information from the outside of the cell to the inside. Transporters enable the flux of molecules and ions. Membrane proteins catalyze reaction such as the before mentioned synthesis of ATP. Almost the entire photosynthesis machinery consists of membrane proteins. The flagellar motor that lets some bacteria swim is membrane embedded. Of course membrane proteins are also active in processes during which the shape of the membrane has to be transformed, such as endocytosis and cell division. Furthermore, over 60% of all approved drugs target a membrane protein [22][23].

For these reasons it is important to gather a detailed understanding of how these proteins work. However, the number of unique proteins in the database of membrane proteins of known 3D structure is at the moment (4.5.2016) 612, of which 22 are porins, like OmpG [24]. This is only 1.2% of the total number of unique structures in the protein data bank (PDB) [25]. The reason for this underrepresentation is that the two major workhorses for the structure elucidation at atomic length scales, x-ray crystallography and solution NMR spectroscopy, work very well with soluble protein but less so with large insoluble membrane proteins [26]. The most used technique to grow crystals

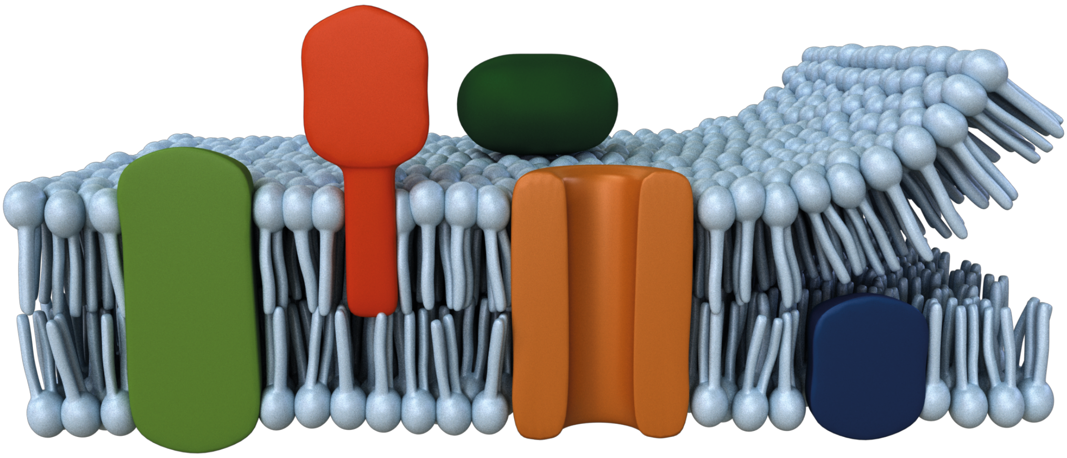


Figure 1.1: Biological membrane with embedded proteins. Figure provided by Dr. Barth-Jan van Rossum.

for x-ray crystallography is the hanging drop method. This method relies on proteins being free in solution. To accomplish this for membrane proteins, detergents have to be added. As the volume of the drop shrinks and the concentration of both protein and detergent rises, often a phase separation takes place that negatively influences the formation of crystals [26]. An alternative method to produce crystals specifically designed for the crystallization of membrane proteins is to employ a lipid cubic phase which is a complex but ordered matrix of lipid bilayers [27][28]. For solution NMR studies, membrane proteins are often introduced in micelles or nanodisks [29][30]. Although these techniques in crystallography and solution NMR have enabled the structure determination of the majority of membrane proteins in the protein data bank (PDB), see table 1.1, the growth in the number of deposited membrane protein structures still dramatically lags behind.

Table 1.1: Statistics of membrane proteins deposited in the protein data bank (PDB), for four structure determination methods [25]. To prevent over-representation of the number of protein entries for each method, they have been filtered for 95% sequence similarity. Note that the numbers of structures determined by individual methods do not add up to the total number of unique structures because the structure of several proteins has been determined by more than one method. Also the total number of structures is slightly higher than the number reported in the database of membrane proteins of known 3D structure.

	α-helical	β-barrel	monotopic
total unique	463	133	46
x-ray crystallography	398	124	46
electron microscopy	34	5	1
solution NMR	55	16	1
solid-state NMR	9	1	-

Recently, the development of direct electron detectors has allowed impressive progress in single molecule cryo-electron microscopy (cyo-EM) [31][32][33]. It has been shown to be applicable to membrane proteins as well and will likely allow the structure determination of many membrane proteins in the future [34][35]. However, cry-EM also has limitations. Because the method is based on aligning thousands of individual noisy images, the protein (complexes) under investigation should be relatively large. Also, the images are taken of a flash frozen solution, which means that, like for solution NMR, membrane proteins have to be reconstituted in some sort of vehicle. Several structures of membrane proteins reconstituted in liposomes have been published [36][37]. Liposomes are small vesicles closed off by a lipid bilayer. Therefore, depending on the mixture of lipids they are composed of, they should be a reasonable approximation of a real biological bilayer. However, all high-resolution structures (<10 Å) that have been determined thus far, employ detergents, such as DDM, amphipols and digitonin, or nanodisks. The review of Vinothkumar provides a good listing of the different detergents used in recent cryo-EM microscopy studies [35].

Protein-lipid interactions

In order for membrane proteins to be incorporated into the membrane, the residues facing the hydrophobic lipids almost exclusively have hydrophobic side-chains [38]. The hydrophobic core of a membrane is roughly 30 Å thick, meaning that there must be around 19 hydrophobic residues to span this distance with an α-helix. For a β-barrel only every second residue points into the lipid bilayer and the length of a stretch spanning the hydrophobic core depends on its shear number but is normally around 10 residues. Another common feature of membrane proteins is that at the interface between the hydrophobic core of the membrane and the head groups of the lipids, often tyrosine and tryptophan residues are found [39][40]. On the basis of these kind of features, algorithms are written to predict from the primary sequence whether a protein is a membrane protein and what its topology is. Because of the larger amount of residues needed to span the membrane and because all residues should be hydrophobic, it is easier to detect α-helical membrane proteins than β-barrels in genome databases [41].

In recent years, it has become clear though that the interaction between proteins and surrounding lipids is a lot more complex than just aspecific hydrophobic interactions between the protein and the membrane. If the size of the hydrophobic part of the protein is different than the length of the fatty acid chain, there is a hydrophobic mismatch which might cause an incorrect geometry of the protein [42]. Not only membrane thickness plays a role. Because the relative width of the head groups and fatty acid tails varies between lipids, some of them are cylindrical whether others are conical, which can cause a specific pressure profile within the membrane [43]. Indeed, membrane proteins can be regulated by physical properties such as thickness and the intrinsic curvature of the membrane [44]. The thickness of the membrane can even vary locally around the circumference of a protein [45]. Furthermore, in some very interesting examples, detecting changes in the physical properties of the membrane is the sole purpose of a membrane protein, as is for instance the case for mechanosensing channels [46].

Besides from bulk mechanical properties induced by the lipids composition of the membrane specific lipids might be needed in close proximity to the protein. It might be useful to divide these protein-lipid interactions into three classes of interaction modes [47]. The first class consists of lipids that form an annular shell that directly surrounds the membrane protein but are not tightly bound to the protein. A second class of lipids is bound more specifically to structural features of the protein. Lipids that act as substrates of membrane proteins form the third class. The identity of the first class of lipids is the most problematic to retrieve since they are often removed by the use of harsh detergents in extraction protocols. However, by extracting membrane proteins from membranes in a more controlled manner, the identity of these lipids can be analyzed by mass spectrometry to various extends [48][47]. More information is available about the more tightly bound lipids in the last two groups. To obtain crystals for x-ray crystallography, these specific lipids often have to be present. In a number of crystal structures, these lipids could be observed tightly bound to the protein. Multiple reviews have been written summarizing all observed lipids in crystal structures [49][50][51].

A recent review by Yeagle lists more than a hundred crystal structures with bound lipids such as cholesterol, cardiolipin, phosphatidylethanolamine (PE) and phosphatidylglycerol (PG) [52]. Many of these lipids were not specifically added during crystallization but were so tightly bound to the protein that they were not removed during the purification. There does not seem to be one general principle as to how these lipids interact with membrane proteins. In some structures, binding is mediated by the head groups, in others by the tails or by both head group and tails. Furthermore, in many structures detergents are found, most likely occupying a place of a removed lipid. Although some information became available in recent years, our real understanding of protein-lipid interactions is still very limited.

Outer membrane protein G

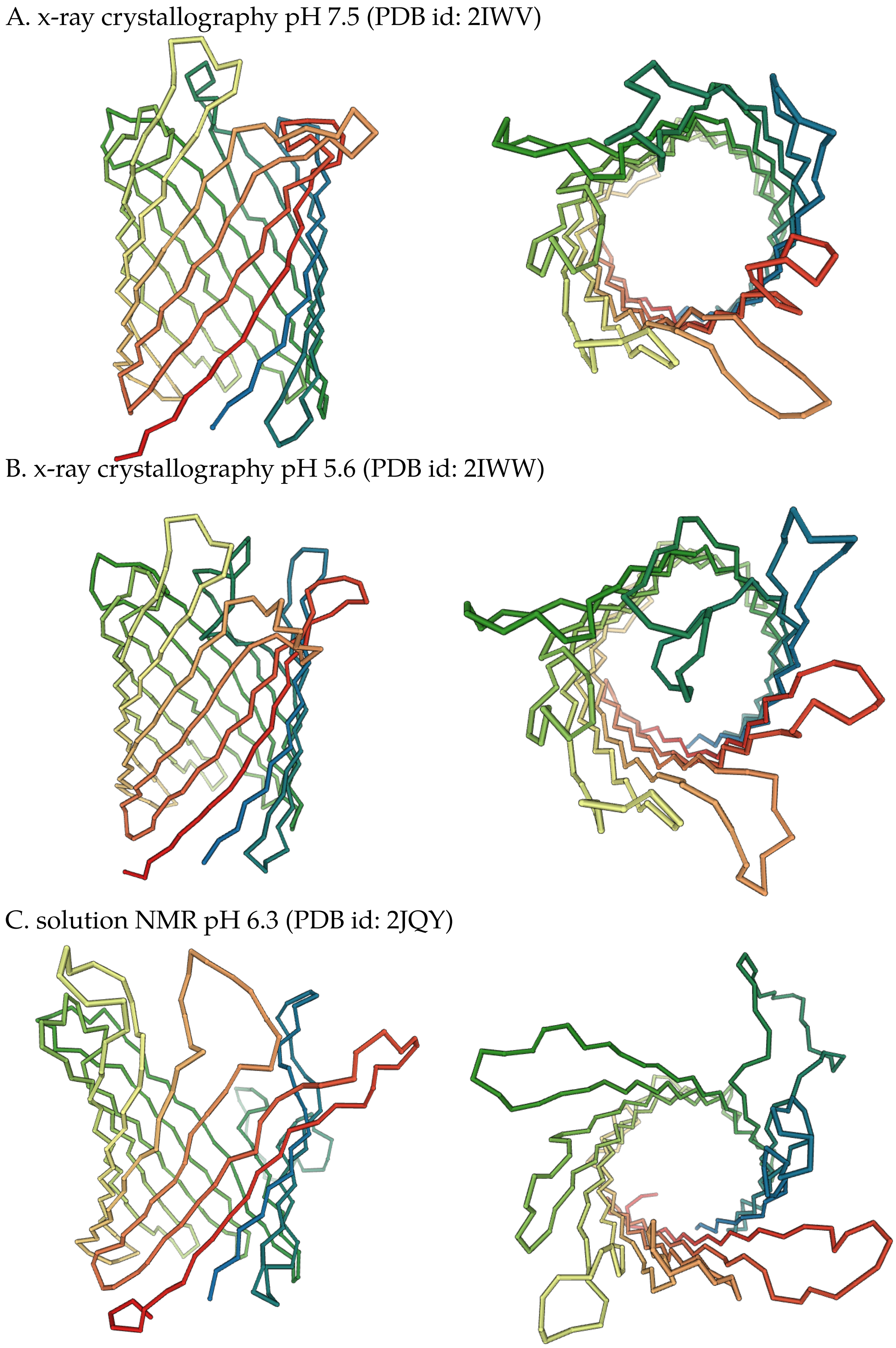
Outer membrane protein G (OmpG) is a 34 kDa β-barrel protein found in the outer membrane of *E. coli.* Besides Gram-negative bacteria like *E. coli* also mitochondria and chloroplasts have outer membranes. The outer membranes of Gram-negative bacteria are exclusively populated by proteins with a β-barrel topology while in mitochondria and chloroplasts also some α-helical outer membrane proteins (Omps) are found [53]. In the case of Gram-negative bacteria, Omps are produced in the cytoplasm and moved into the periplasm where they are inserted into the outer membrane by the β-barrel assembly machinery (BAM) complex, of which one of the proteins, BamA, itself contains a membrane embedded β-barrel subunit [54]. Outer membrane proteins perform a host of different functions that are needed on the interface between the inside and outside of the cell/organelle [55]. They can act as enzymes, transporters and/or receptors. Many are autotransporters that translocate one of their domains to the extracellular space, often acting as adhesins helping with the invasion of other cells and therefor linked to infectious decease [56]. OmpG belongs to a class of outer membrane proteins known as porins. These proteins act as pores allowing the passive but selective uptake and secretion of nutrients, ions and proteins. In general, porins in Gram-negative bacteria have short turns on the periplasmic side and long loops on the extracellular side [53].

The main porins for the uptake of sugars through the outer membrane of Gram-negative bacteria are LamB and OmpF. Following deletion of genes in *E. coli* coding for LamB and OmpF and a selection procedure to generate phenotypes able to grow on a maltodextrin medium, a number of different mutations were found. One of those mutations allowed the otherwise not expressed OmpG to come to expression [57]. Interestingly, low levels of OmpG expression were found in Salmonella and Shigella bacteria [58]. Further biochemical analysis showed that OmpG is able to import mono-, di- and trisaccharides [58]. The OmpG gene codes for 301 amino acids of which the first 20 are a signal sequence that gets cleaved off upon arrival in the periplasm [58]. It was discovered that OmpG exists as a monomer, which is exceptional since most porins are composed of trimers. No evidence of an oligomeric form could be found in native/denaturing PAGE analysis and cross-linking experiments [58]. Further evidence from electrophysiology studies confirmed the monomeric nature of OmpG

[59].

A cryo-electron microscopy projection map at 6 Å confirmed the β-barrel structure of OmpG and its observed diameter of 2.5 nm agreed with earlier made predictions that OmpG is composed of 14 strands [60][59]. In 2006 and 2007, crystal structures and a solution NMR structure were published [61][62][63]. The studies of Yildiz et al. hint at a pH-dependent opening and closing mechanism [62]. For these studies, OmpG was crystallized at pH 7.5 and pH 5.6. The structure at pH 7.5 showed an open conformation while in the structure at pH 5.6 the longest extra-cellular loop (loop 6) was folded into the pore, closing it off as a sort of lid, see figure 1.2. The crystal structure of Subbarao and van den Berg was crystallized at pH 5.5 and misses part of the residues in loop 6 (220-231) but seems to resemble the pH 7.5 structure of Yildiz et al., which is surprising [61]. The solution NMR studies were performed at pH 6.3 which is between the crystallization conditions of the structures of Yildiz et al. [63]. The entire loop 6 and parts of loop 7 could not be assigned, and almost no long-range restraints could be found for most of the extra-cellular loops, indicating motional inhomogeneity. Therefore the β-barrel in the solution NMR structure is a lot shorter on the extra-cellular side in comparison to the crystal structures. This smaller β-sheet size fits the probable thickness of the outer membrane of *E. coli* which is around 27 Å corresponding to around 10 residues to cross the membrane [41]. Also, the barrel in the crystal structures is extended very far beyond the ring of outward facing tryptophans and tyrosines that are likely at the membrane interface, see figure 1.3A. The conformation in the crystal structure can be explained by crystal contacts. In figure 1.3B the crystal packing for one of the crystal structures, 2IWV,is shown [62]. The molecules in the crystal are stacked in such a way that they form a continuing barrel. For the solution NMR structure, the motion of the extra-cellular loops was confirmed by heteronuclear NOESY experiments [63].

Yildiz et al. proposed that the protonation state of two histidines (231 and 261) determines whether the protein is in an open or closed configuration, see figure 1.4. At a lower pH, the histidine side-chains get protonated which causes them to repel each other. This disrupts the hydrogen bond pattern and thereby allows loop 6 to fold into the pore. Although this seems a simple and attractive explanation, later studies showed that the exact mechanism for the pH-dependent opening and closing of OmpG is more complex and is still not completely clear. Because OmpG is a monomer, it is a good candidate to form the basis for a stochastic biosensor for the detection of analytes. The detection method of this type of biosensor relies on the binding of an analyte to cause a disruption in the ion flow when an electrical potential is applied over the membrane [65][66]. Within this context, the opening and closing mechanism of OmpG is of interest since it is necessary to remove all spontaneous opening and closing events of the pore which manifests itself as noise. To reach this goal glycine 231 and aspartic acid 262, which directly neighbor the before mentioned histidines, were mutated to cysteines to form disulfide bond keeping strands 12 and 13 together (dark green in figure 1.4). Also aspartic acid 215 was deleted to remove a beta bulge and thereby reinforcing the hydrogen bonding pattern (light green in figure 1.4). This combination of mutations removed 95% of the spontaneous gating [67].



A. x-ray crystallography pH 7.5 (PDB id: 2IWV)

B. x-ray crystallography pH 5.6 (PDB id: 2IWW)

C. solution NMR pH 6.3 (PDB id: 2JQY)

Figure 1.2: Crystal structures by Yildiz et al. and solution NMR structure by Liang and Tamm.

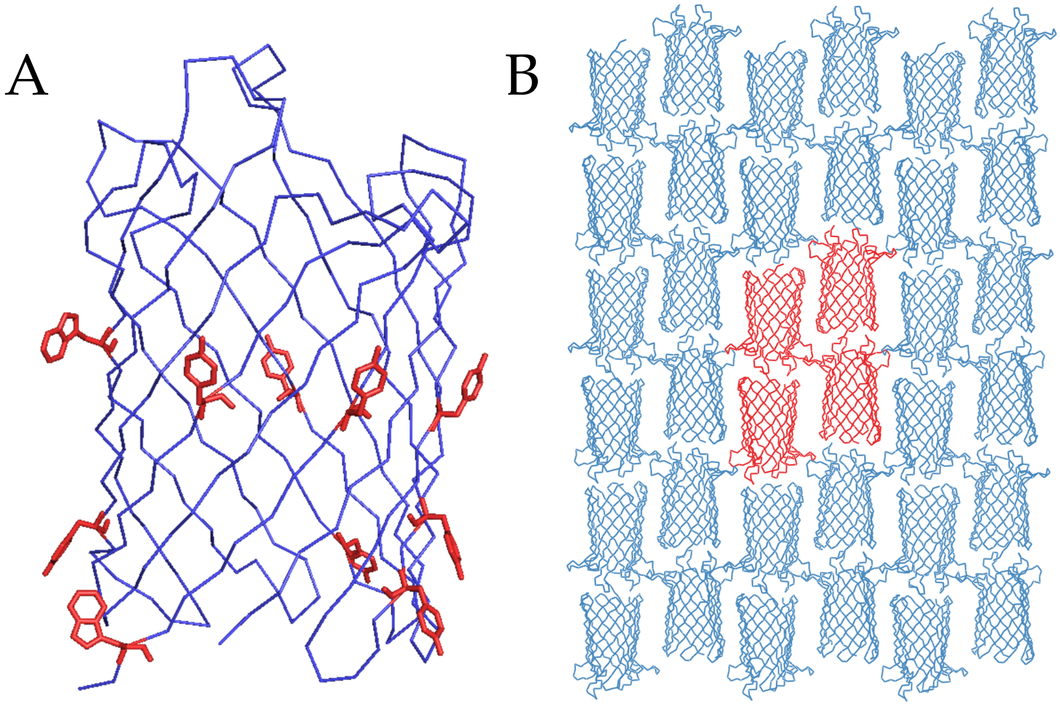


Figure 1.3: A) Crystal structure 2IWV with outward facing tryptophans and tyrosines highlighted in red indicating the approximate position of the membrane [62]. There are more tryptophans and tyrosines in the molecule but they face inside into the pore. B) the same crystal structure with surrounding unit cells. One unit cell contains four OmpG molecules. One of the unit cells is depicted in red. The individual OmpG molecules are stacked to form an quasi infinite barrel with several crystal contact stabilizing this conformation. Figure produced using pymol [64].

Another study, partially performed by the same people that published the crystal structures in two pH states, using Fourier transform infrared spectroscopy, indicated that there is an increase in β-sheet rigidity and thermostability at higher pH [68]. In this study three different mutants were created. In two mutants, the histidine pair was mutated to alanines or cysteines (dark blue in figure 1.4). In the third mutant, 9 residues in loop 6 were deleted (light blue in figure 1.4). In the first two mutants no pH-dependent alteration of secondary structure content could be observed. These mutants always showed a similar secondary structure content as in wild type OmpG at pH 7.5. The crystal structure of the alanine mutant at pH 6.5 resembled the wild-type structure at pH 7.5. However, for the third mutant a slight pH-dependent change in secondary structure was observed. From these studies it was concluded that the protonation state of the histidine pair directly determines the opening and closing of the pore. However, in a follow-up study by the same author this conclusion was softened somewhat, and it was pointed out that other factors, such as the state of the charged aspartic acid, glutamic acid and arginine residues in the lumen of the pore, might play a role as well [69]. This was further supported by a study that tried to completely block the spontaneous gating of OmpG [70]. In that study large parts of all extra-cellular loops were deleted to create a minimal pore(red circles in figure 1.4). Even in the absence of the loops and one of the histidines of the pair, pH-dependent gating was detected. Furthermore, solution NMR ensembles obtained by attaching paramagnetic relaxation enhancers to the flexible loops indicated that not only loop 6 might be involved in the opening and closing mechanism but also the other loops [71]. Therefore a combination of all these factors might govern the gating of OmpG.

Besides that there are some unresolved biological questions surrounding OmpG, this membrane protein provides a good model for the development of solid-state NMR methods for the structure elucidation of membrane proteins. With 281 residues (our construct contains an extra methionine at the N-terminus) it has more residues than most structures solved by solid-state NMR thus far. Also, as can be seen in figure 1.5, there are many membrane proteins that have a similar size or smaller as OmpG. The full size of some solid-state NMR structures of multimers might be larger but as we shall see later it is mostly the monomer size that complicates the analysis of solid-state NMR spectra and not so much the molecular mass of the assembly. In this studies, OmpG is reconstituted in native *E. coli* lipids which is, as should be clear from the discussion above, a clear advantage in terms of biological relevance. Other proteins studied by solid-state NMR are often of microcrystalline nature making them more homogeneous. Although state of the art pulse sequences and methodologies were and are developed and first tested on smaller microcrystalline proteins such as the α-spectrin Src-homology 3 domain (SH3) or ubiquitin, it can be advantageous to have a larger and more challenging system to test which of those methodologies lead towards a robust strategy for the study and structure determination of membrane proteins, which has always been one of the goals for the development of solid-state NMR.

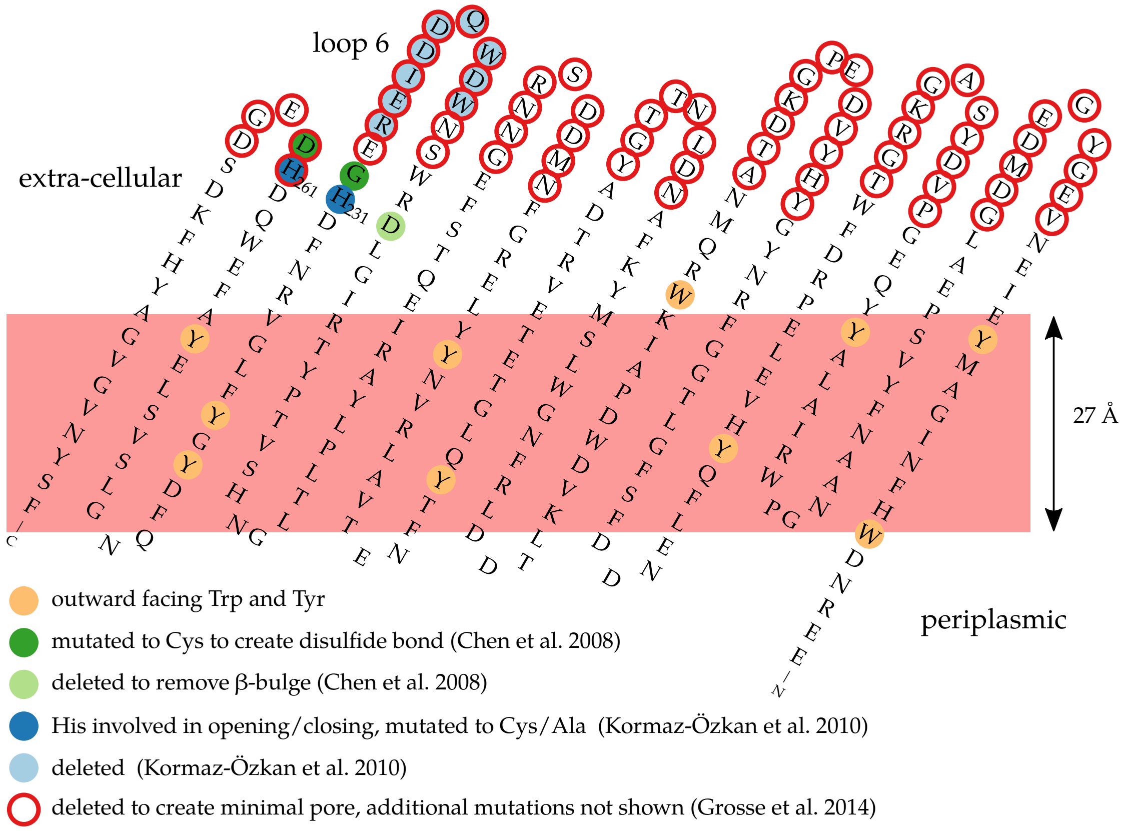

loop 6
extra-cellular
periplasmic
27 Å
H261
H231
outward facing Trp and Tyr
mutated to Cys to create disulfide bond (Chen et al. 2008)
deleted to remove β-bulge (Chen et al. 2008)
His involved in opening/closing, mutated to Cys/Ala (Kormaz-Özkan et al. 2010)
deleted (Kormaz-Özkan et al. 2010)
deleted to create minimal pore, additional mutations not shown (Grosse et al. 2014)

Figure 1.4: Mutations made in different studies investigating the pH-dependent opening/closing mechanism of OmpG. The pink box roughly indicates the thickness and location of the lipid bilayer. See the main text for a detailed review of the effects of the varies mutations.

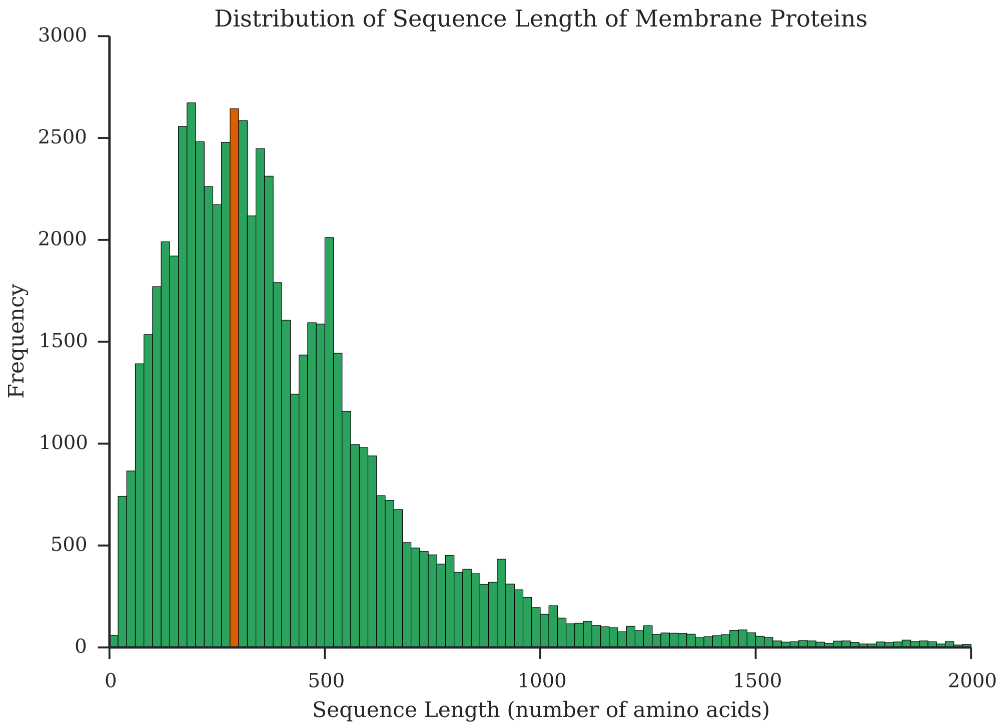


Figure 1.5: Distribution of sequence lengths of all proteins in the uniProt database tagged as a transmembrane protein. The bin in which OmpG (281 residues) falls is highlighted in orange. The amount of proteins with sequence lengths similar to or smaller than OmpG is very large, making OmpG a relevant model system.

Solid-State NMR

NMR spectroscopy is a standard method for the analysis of chemical substances. The source of the signal originates, like in any other form of spectroscopy, by transitions between states that have a difference in energy. In the case of NMR spectroscopy, this energy difference is generated by bringing the sample in a large magnetic field. All atomic nuclei have a quantum mechanical property called "spin". The number of allowed spin states depends on the nuclear spin quantum number I. Many states between I-1 and I+1 in integer steps are allowed. Hence, the spin of isotopes with spin quantum number $\frac{1}{2}$ (for example ^{1}H, ^{13}C and ^{15}N) has only two allowed spin states, $-\frac{1}{2}$ and $+\frac{1}{2}$. When brought into a magnetic field, a difference in energy between these two state is arises, also known as the Zeeman effect. The energy of the spins pointing along the magnetic field is slightly lower than that of those pointed against it. These spin states are generally referred to spin up and spin down, respectively. The energy difference is given by:

$$\Delta E = \hbar\gamma B \tag{1.1}$$

where ħ is the reduced planck's constant (h/2π), B is the magnitude of the magnetic field and γ is the isotope dependent gyromagnetic ratio. Gyromagnetic ratios are listed in table 1.2. As often the case in quantum mechanics, a different but equally valid way to think of it is that the spins are precessing around the magnetic field with a nutation frequency ν (Hz) or ω (rad s^{-1}), which is called the Larmor frequency. Because $\Delta E = h\nu$, the Larmor frequency is:

$$\nu = \frac{\gamma B}{2\pi} \tag{1.2}$$

or in terms of angular frequency ω (rad s^{-1}) simply:

$$\omega = \gamma B \tag{1.3}$$

As can be seen in table 1.2 this frequency is in the radiofrequency range (MHz)

Table 1.2: properties of often used isotopes in NMR spectroscopy. Larmor frequencies ν are given for a magnetic field of 23.5 T which corresponds to a 1 Ghz ^{1}H magnet, which is one of the highest field NMR magnets commercially avaible at the moment. Gyromagnetic ratio and natural abundances from the book of P.J. Hore [72].

	spin quantum number I	$\gamma/10^{7}$T^{-1}s^{-1}	ν/MHz at 1 GHz ^{1}H (23.5 T)	Natural Abundance %
^{1}H	$\frac{1}{2}$	26.75	1000.0	99.985
^{2}H	1	4.11	153.9	0.015
^{13}C	$\frac{1}{2}$	6.73	252.1	1.108
^{15}N	$\frac{1}{2}$	-2.71	101.5	0.37
^{31}P	$\frac{1}{2}$	10.84	406.0	100.0

The energy difference created by even the strongest NMR magnets is only very small compared to the thermal energy at any temperature that is not close to 0 K. Therefor the population difference between spins in the low and high energy state is only very small. For this reason NMR is an inherently insensitive method. A measurable signal can only be generated by the measuring millions of molecules in bulk at the same time. As will be explained later, in one way or another all difficulties with this method lead back to this fact.

Chemical shift

The strength of NMR derives from the fact that the magnetic field perceived by a given spin is not only determined by the external field generated by the magnet. Also the local chemical environment around the spin, to be more precise the surrounding electron cloud, influences the magnetic field perceived by the spin. Therefore every nucleus in a molecule that has a different chemical environment has a different resonance frequency and therefore gives rise to a unique peak in the spectrum. This property is called chemical shift. The difference between the actual resonance frequencies of nuclei in a molecule is, dependent on the nucleus and magnetic field, up to a few tens of kHz. The value for the chemical shift δ is normally not reported in terms of Hz but in parts per million (ppm), which is defined as:

$$\delta = 10^6 \frac{\nu - \nu_{ref}}{\nu_{ref}} \tag{1.4}$$

where ν_{ref} is the resonance frequency of a reference compound. A practical aspect of this measure is that it is independent of the magnetic field of the spectrometer, so that spectra recorded on different instruments can be easily compared. For proteins, there are typical chemical shift ranges for nuclei that are part of different chemical moieties such as methyl groups or aromatic rings. Even if the same type of chemical group is present multiple times in the same molecule, which is generally the case in a large molecule such as a protein, individual peaks belonging to each one of those groups can still be separated from one another, given the resolution of the spectrum is high enough. Because the amount of Hz per ppm is dependent on the magnetic field strength, in addition to the signal to noise, also the resolution of spectra is increased at higher field. Note that the values given in 1.2 for the resonance frequency ν in MHz directly correspond to the number of Hz/ppm on a 1 GHz magnet.

Measurement of the NMR signal

At the start of an NMR experiment the spins are in thermodynamic equilibrium. Because of the Boltzmann distribution, there is a slightly larger number of spins in the low energy state. Therefore, in the case of a spin with a positive gyromagnetic ratio, the net magnetization is pointing along the positive z-axis (by convention the static magnetic field B_0 is oriented along the z-axis), also known as longitudinal magnetization. Due to the angular momentum, for individual spins a component of the magnetization vector is present perpendicular to B_0. But because the direction of these components are randomly distributed, when considering all spins together, there is no net magnetization in the xy-plane. In order to detect the NMR signal, first a radio frequency pulse at the transition frequency of the nucleus of interest (”on resonance”) is given. In practice this is accomplished by applying a current to a coil located around the sample in the magnet. The duration and the amplitude of the pulse determines how many spins change state from the low to the high energy state and vice versa. At a specific combination of duration and amplitude the population difference between the two states can be completely inverted, bringing the net magnetization to the negative z-axis. Such a pulse is referred to as a 180° pulse. When a 90° pulse is applied, which is half of the duration or amplitude of the 180° pulse, the population difference will be 0 and therefore there will be no net magnetization along the z-axis. However, such a pulse does not only cause the spin up and spin down population to be equalized. Because the pulse has a phase, also coherence is generated, meaning that now there is a net magnetization in the xy-plane.

$$B' = \frac{\omega_{res} - \omega_{rf}}{\gamma} \tag{1.5}$$

When ω_{rf} is chosen to be close to ω_{res}, B′ practically becomes 0. In the rotating frame, a magnetic field B_1, induced by a pulse perpendicular to the z-axis at ω_{rf}, will appear static. The direction of B_1 in xy-plane is determined by the phase of the pulse. Because B′ is effectively close to 0, the only magnetic field left is B_1. If a pulse is given along the positive x-axis, the net magnetization originally present along the positive z-axis will start precessing in the yz-plane at the nutation frequency:

$$\omega_{nut} = \gamma B_1 \tag{1.6}$$

initially moving in the direction of the -y axis (following the right-hand rule). After applying a 90° x-pulse this is where the magnetization will be located. Now the magnetization precesses freely through the xy-plane of the rotating frame at ω_{res}-ω_{rf}, or at ω_{res} in xy-plane of the laboratory frame, where ω_{rf} is the carrier frequency and ω_{res} the resonance frequencies of different nulcei in the sample. This rotating magnetization produces an electric field which in turn induces a current in the receiver coil.

It should be noted that the rotating frame is not just a theoretical trick to simplify reasoning about NMR experiments. In the console of the spectrometer, pulses are generated by adding a signal from a synthesizer to a radiofrequency carrier. Furthermore, this carrier is subtracted from the actual detected signal before it is digitized, mostly because it is more achievable to digitize a signal in the kHz range than the MHz range. Therefor the final digital signal that is Fourier transformed to obtain the NMR spectrum actually consist of signals at ω_{res}-ω_{rf}.

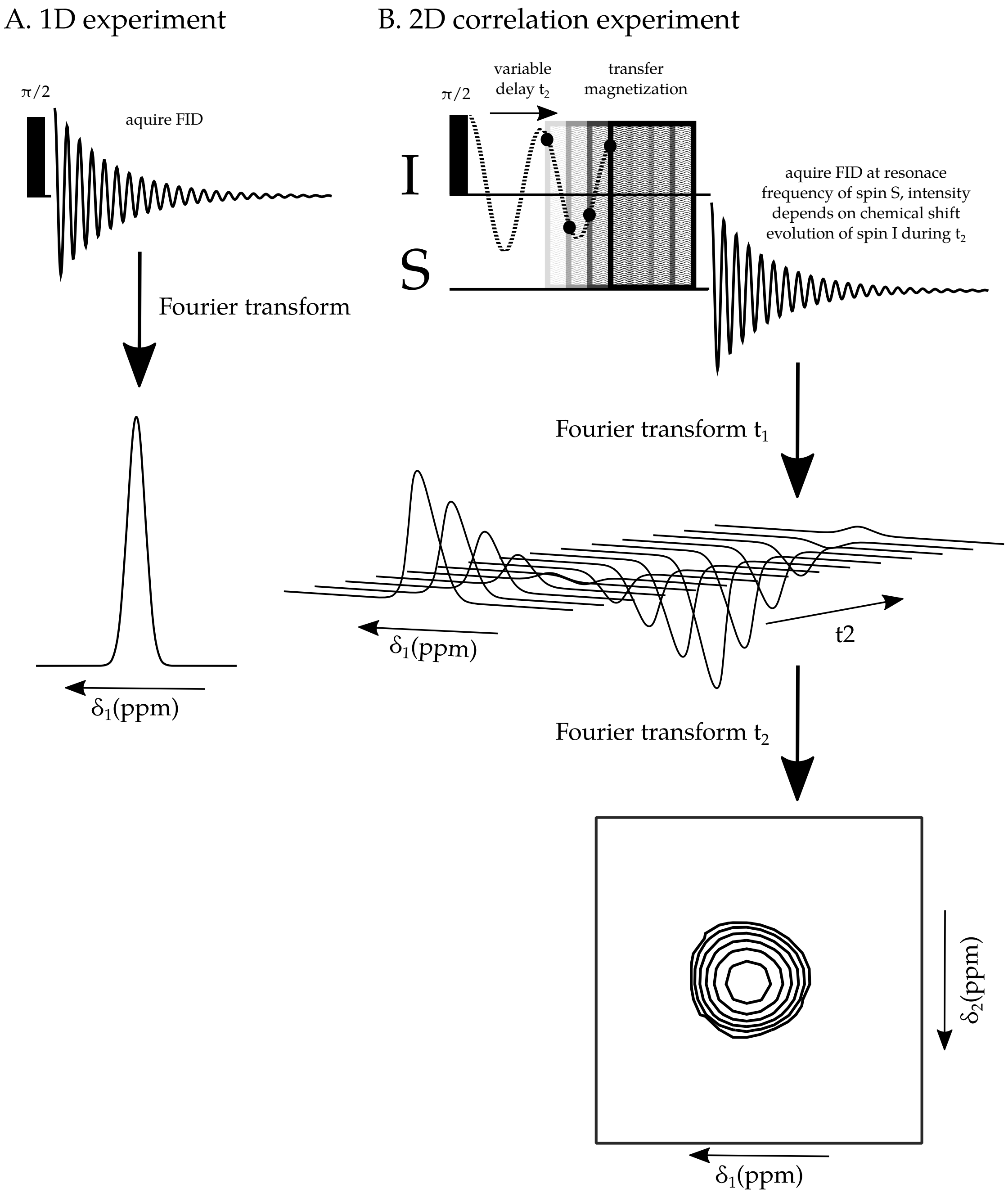


Figure 1.6: Simplified pulse sequences for 1D and multi-dimensional experiments.

Correlation spectroscopy

In a one-dimensional NMR spectrum, peaks are present at the resonance frequencies of all the nuclei in a molecule of a given isotope. However, it can not be known, except for very small compounds, which resonance frequency belongs to which nucleus in the molecule. To resolve this issue, information should be acquired about the relationship between the different resonance frequencies in the spectrum. Such relationships can be used, for example, to look for resonance frequencies belonging to two directly bonded nuclei or two nuclei close together in space. This type of information can be obtained by recording multi-dimensional spectra. In a multi-dimensional NMR spectrum each dimension of a peak corresponds to the resonance frequency of a different nucleus in the molecule. This is achieved by including a magnetization transfer step between nuclei in the pulse sequence, see figure 1.6B. The relation between the different nuclei encoded by the dimensions of one peak, depends on the exact method used to transfer magnetization between those nuclei. Some transfer methods act trough bond, while others act trough space. Magnetization can be transferred between nuclei of the same isotope (homo-nuclear transfers) or between nuclei of different isotopes (hetero-nuclear transfers).

One of the dimensions in a multi-dimensional experiment is aquired directly, in the same way as is done for a one-dimensional experiment. The other dimensions have to be acquired indirectly. This is achieved by recording the same experiment multiple times with a varying delay (t_2 in figure 1.6B). During this delay the magnetization of the indirectly detected nucleus precesses in the xy-plane at the (relative) chemical shift frequency. Depending on the length of t_2, the magnitude of magnetization along the x-axis (or y-axis) is different just before the magnetization transfer (indicated by the black dots in figure 1.6B). Therefore the amplitude of the FID of individual experiments is modulated by the chemical shift frequency of the indirectly detected nucleus. By Fourier transformation of the FID of individual experiments and a subsequent Fourier transformation with respect to t_2, a two-dimensional spectrum is obtained.

This principle can in theory be extended to an infinite number of dimensions. However, the amount of sub-experiments required to acquire an extra dimension goes up exponentially. In addition, due to decay of the signal in the xy-plane (discussed later), individual experiments can not be made too long. The pulse sequence shown in 1.6 are a simplification of sequences used to acquire the spectra used is this thesis. The actual pulse sequences used are shown in figures 2.1, 3.2, 3.3 and 6.2.

By the analysis of the connectivities between different resonance frequencies in the spectra, the mapping between resonance frequencies and nuclei can be determined. This process is known as spectral assignment and will be explained in more detail further on. Furthermore, the distance restraints used to calculate three-dimensional structures of molecules are obtained by interpreting peaks from experiments with a through-space transfer step. In addition to correlating resonance frequencies, multi-dimensional spectra also add resolution, as the peaks are spread through multiple dimensions. For these reasons, multi-dimensional NMR forms the backbone of all applications in structural biology.

Scalar coupling

Scalar coupling is a coupling between spins mediated by the electrons in the chemical bond. When two spins I and S are scalar coupled, the spin state of I will influence the energy levels of the S spin and vice versa. To be more precise, depending on the nature of the molecular orbital, either a parallel or an antiparallel configuration of the spin states of I and S will be lower in energy. In the example of figure 1.7 antiparallel configurations of spin states are favored. Here the effect on the spectrum of S is shown. The scalar coupling has the same effect on the spectrum of I. As can be seen a peak splitting occurs. The volume of both peaks will be the same as the amount of (I) spins in the "spin up" and "spin down" state is (almost) equal. The scalar coupling constant J is defined as the full splitting of the lines, in Hz. The magnitude of the splitting roughly depends on the nuclei that are coupled and by the number of bonds they are separated from one another. This coupling is in general not more than 100 Hz. In contrast to the dipolar coupling (discussed next), scalar coupling is not dependent on the orientation of the molecule with respect to the magnetic field.

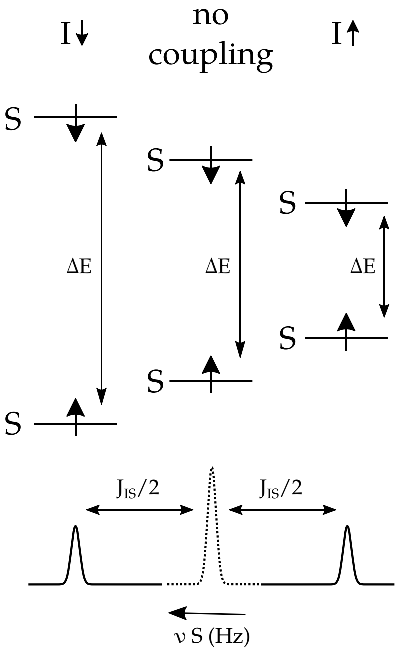


Figure 1.7: Effect of the spin state of I on the resonance frequency of spin S when I an S are scalar coupled. When the I spin is in the "spin up" configuration, the "spin up" configuration of spin S will be higher in energy than when S would not be coupled to I. At the same time, the "spin down" configuration of spin S is lower in energy. Together this leads to a smaller transition energy (right). When spin I is in the spin I is in the "spin down" state, the situation is the exact opposite, leading to a higher transition energy (left). The dotted peak is at the position of the resonance frequency of S in absence of scalar coupling.

Dipolar coupling

A far larger splitting of signals is caused by dipolar coupling. All spins behave as dipoles and therefor create a small magnetic field around themselves. This local field adds to B$_0$, and therefore influences the net magnetic field observed by nearby spins.

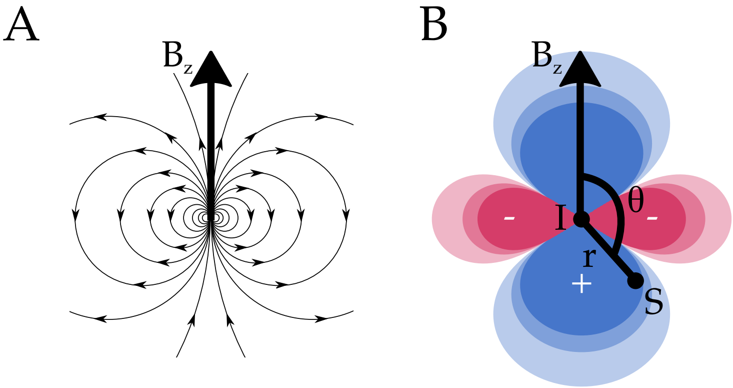


Figure 1.8: Dipolar coupling between two spins I and S. A) Field lines generated by a dipole. B) Colored regions depict the magnitude of the z-component of the magnetic field induced by the dipole I. Spin S also creates a magnetic field but, for clarity, this is not shown. The distance r between the spins and angle θ between the connection vector and the z-axis determine the magnitude of the z-component of the field generated by I at the position of S, and are used in equation 1.7. Depending on whether spin I is in the "spin up" or "spin down" state, the direction of the field lines in A, and thereby the sign of the z-component in B are inverted.

As can be seen in figure 1.8, the magnitude of the z-component of the magnetic field created by spin I at the position of spin S depends on the distance r between the two spins and the angle θ between the vector IS and the z-axis. Depending on whether the I spin is in the "spin up" or "spin down" state, the dipole field will add to or subtract from B$_0$ and the resonance frequency of S will be shifted upfield or downfield. Like as was the case for the scalar coupling, because the number of I spins in the "spin up" and "spin down" state is (almost) equal, there is a symmetric splitting of the signal. Of course, spin I is influenced in the same way by a magnetic field generated by spin S. In the heteronuclear case, the full splitting (in Hz) of the resonance frequencies of I and S is given by the equation:

$$splitting = \frac{\mu_0}{8\pi^2}\frac{\hbar\gamma_I\gamma_S}{r_{IS}^3}(3cos^2\theta - 1) \tag{1.7}$$

where μ_0 is the permeability of the vacuum (4π x 10^{-7} H m^{-1}), γ_I and γ_S are the gyromagnetic ratios of the interacting nulcei. In the homonuclear case this splitting will be larger by a factor $\frac{3}{2}$ [72]. The first part of the equation is the coupling constant d:

$$d = \frac{\mu_0}{8\pi^2} \frac{\hbar\gamma_I\gamma_S}{r_{IS}^3} \tag{1.8}$$

In table 1.3 coupling constants at different values for r_{IS} are given for dipolar coupling between the nuclei most often used in biological NMR. In particular when protons are involved, the dipolar coupling can be very large. Some are larger than the spread in chemical shifts of the involved nuclei (and therefore larger than the typical spectral width).

Table 1.3: Dipolar coupling constants (in Hz) with r of 1,2,4 and 8 Å.

	1Å	2Å	4Å	8Å
^{1}H-^{1}H	120100	15013	1877	235
^{13}C-^{13}C	7602	950	118	15
^{15}N-^{15}N	1233	154	19	2
^{1}H-^{13}C	30216	3777	472	59
^{1}H-^{15}N	12160	1521	190	23

Magic Angle Spinning

In both liquid and solid samples, every molecule in the sample has a different spatial orientation with respect to the direction of the static field of the magnet (unless there is a bias towards certain orientations such as in a crystal). For this reason the local magnetic field observed by a given nucleus changes from molecule to molecule. The chemical shift will be different for each orientation since the orientation of the asymmetric electron cloud with respect to an observed spin and B_0 will be different. This is called chemical shift anisotropy (CSA). The magnitude of the CSA is dependent on the magnetic field and is generally in the order of 10^4 Hz. Furthermore, the angle θ, introduced in the last section, for a given pair of nuclei I-S is different in every molecule. Hence, also the dipolar coupling is anisotropic and for this reason the full range of values for the splitting caused by the dipolar coupling would be observed. In static solid samples, because of this the resulting spectra are rather featureless and contain little information. The reason why solution NMR spectra do not suffer from this extreme anisotropic line broadening is that the fast isotropic tumbling averages out the anisotropic interactions. I.e. at NMR time scales the local magnetic field perceived by a spin is an average field over all orientations.

In solid state NMR, a similar effect can be created by magic angle spinning (MAS). The magic angle is the angle θ_{magic} for which the term $3cos^2\theta - 1$ in equation 1.7 becomes zero, which is at 54.7°. By spinning the sample around this angle, spin S will rotate through the dipole field created by I and angle θ will be oscillating around θ_{magic} (see figure 1.9B). Therefore the term $3cos^2\theta - 1$ can be rewritten to:

$$3cos^2(sin(Asin\theta_{rot}) + \theta_{magic}) - 1 \tag{1.9}$$

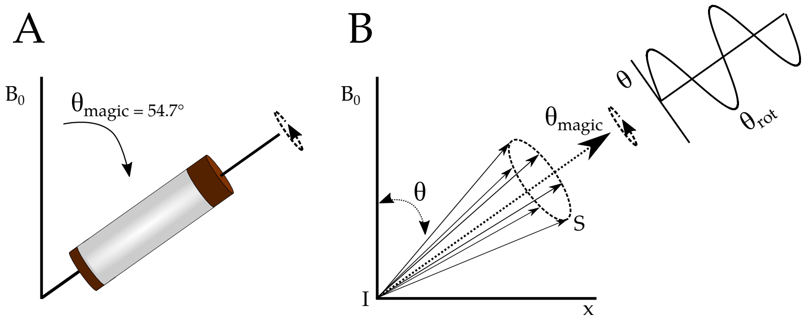


Figure 1.9: A rotor at the magic angle.

where θ_{rot} is the angle of rotation around the magic angle and A is a scaling factor that is the difference of the largest angle θ and θ_{magic}. When the integral is taken over a full rotation:

$$\int_0^{2\pi} (3cos^2(sin(Asin\theta_{rot}) + \theta_{magic}) - 1)d\theta_{rot} = \tag{1.10}$$

$$3cos^2\theta_{magic} - 1 = 0 \tag{1.11}$$

For this reason the average perceived contribution over a full rotation around the magic angle is averaged to zero. MAS averages out the CSA in a very similar way, leaving the same isotropic chemical shift as is measured in solution NMR and sidebands at multiples of the MAS frequency. Note that in contrast to the chemical shift, the dipolar interaction does not have an isotropic part, hence it is fully averaged (assuming the MAS is fast enough on the NMR time scale). Scalar couplings are not averaged since they are orientation independent.

In practice MAS is performed by filling the sample in a small rotor (see figure 1.9A). The rotor has fins, allowing in to be spun in an air stream. To effectively average an interaction, the MAS frequency should be several times larger than the strength of the interaction.

Decoupling

As can be seen in table 1.3 some dipolar interaction can not be averaged out completely at moderate MAS frequencies. For instance, the ^{13}C-detected spectra in chapter 2 are recorded at 12 kHz MAS, but the heteronuclear ^{13}C-^{1}H coupling constants for directly bonded protons (± 1 Å) exceeds this frequency. Therefore, in addition to MAS, rf decoupling is employed. By applying an rf field on the resonance frequency of the nucleus that should be decoupled, the up and down spin states are continuously exchanged, which causes the direction of the dipole field to change around. Also the scalar coupling is decoupled by the constant exchange of spin states. In the simplest form a constant wave (CW) is applied [73]. More advanced sequences, such as two pulse phase modulation (TPPM) and small phase incremental alternation (SPINAL), consist of trains of back-to-back pulses with changing phases [74][75]. The extend of the decoupling is limited by the power of the decoupling pulses (at least at moderate spinning rates). Therefore, effective decoupling requires high power rf, but as a side effect it causes sample heating and degradation, and puts high strain on the equipment.

Relaxation

In NMR two types of relaxation are distinguished: Spin-lattice (T_1) and spin-spin (T_2) relaxation. T_1 relaxation describes the return of transverse magnetization to the thermodynamic equilibrium, where the time T_1 is defined as:

$$M_z(t) = M_{z,eq} - (1 - e^{-t/T1}) \tag{1.12}$$

The T_1 determines the waiting period necessary between individual experiments. Therefor short T1 times are favorable, since more scans can be recorded in the same amount of time.

T_2 descibes the disappearance of magnetization in the transverse plane and is defined as:

$$M_{xy}(t) = M_{xy}(0)e^{-t/T2} \tag{1.13}$$

T_2 might just seem as the inverse of T_1, as relaxation to the thermodynamic equilibrium brings the net magnetization back to the z-axis, moving it away from the xy-plane. However, in general T_1 is not a main contributor to the loss of transverse magnetization. The main contribution to T_2 relaxation is loss of coherence, see figure 1.10. Coherence loss is caused by individual spins from different molecules in the sample precessing at slightly different frequencies. This can be caused by dipolar interactions and CSA that are not fully averaged by the magic angle spinning and decoupling. This is the homogeneous part of T_2. Other contributions are inhomogeneous in nature. Structural inhomogeneity of the protein sample will cause differences in chemical shifts. In addition inhomogeneity of the magnetic field also adds to the T_2 relaxation rates. The homogeneous and inhomogeneous contributions to the T_2 relaxation can be added to obtain T_2*:

$$\frac{1}{T_2^*} = \frac{1}{T_{2,homogeneous}} + \frac{1}{T_{2,inhomogeneous}} \tag{1.14}$$

The full width at half height of the peaks is directly proportional to T_2*:

$$linewidth_{\frac{1}{2}} = \frac{1}{\pi T_2} \tag{1.15}$$

The homogeneous part of T_2 can be measured with a spin-echo echo experiment. This experiment consists of the generation of transverse magnetization followed by a delay. In the middle of the delay a 180°-pulse refocuses the inhomogeneous contributions to the T_2. By acquiring the signal after different lengths of the delay, the $T_{2,\text{ homogeneous}}$ can be calculated. Although it is the T_2* that directly affects the line broadening, and thereby the resolution of the spectra, it can be interesting to known whether homogeneous or inhomogeneous contributions govern T_2. Homogeneous contributions can be further repressed by faster spinning and inhomogeneous contributions could be minimized by improved sample preparation.

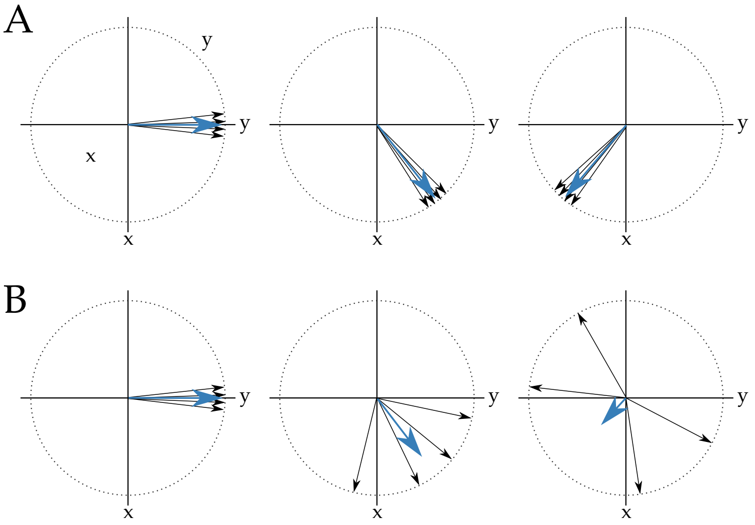


Figure 1.10: Coherence loss in the xy-plane. Black arrows illustrate individual spins, blue arrows show the net magnetization in the xy-plane. Figures from left to right show the evolution over time. In A there is virtually no coherence loss and the magnitude of the net magnetization in the xy-plane stays the same. In B individual spins precess through the xy-plane at different angular velocities causing the net magnetization in the xy-plane to decrease over time.

Transfer of magnetization

The pulse sequence in 1.6B contains a magnetization transfer step. As mentioned, there are multiple ways to transfer magnetization between nuclei. Besides the need to transfer magnetization to correlate different nulcei, as discussed in the section about correlation spectra, often a magnetization transfer is employed to transfer the larger polarization of high-γ nuclei (^{1}H) to other nuclei to enhance the signal to noise. In solution NMR *I*nsensitive *n*uclei *e*nhanced by *p*olarization *t*ransfer (INEPT), which is a scalar coupling based method, is mostly used for this cause. In solid-state NMR this is mostly accomplished using cross-polarization (CP). Here the transfer methods used in the rest of the thesis will be discussed briefly.

Cross-polarization

Cross-polarization (CP) between two unlike nuclei I and S can occur when the splitting in energy levels of the I and S spin are made equal or the difference between them is a multiple of the MAS frequency:

$$\omega_I - \omega_S = \pm n\omega_r \tag{1.16}$$

This condition is known as the Hartmann-Hahn matching condition [76]. When this condition is met, energy can be exchanged between the two nuclei through a dipolar interactions. Since spins I and S have different gyromagnetic ratios in this case, it is impossible to meet this condition in the static B_0 field. However Hartmann-Hahn matching can be performed in the B_1 fields generated by two simultaneous rf-pulses on the resonance frequencies of spins I and S:

$$\gamma_I B_{1,I} - \gamma_S B_{1,S} = \pm n\omega_r \tag{1.17}$$

To keep the direction of the B_1 field constant in the rotating frame, the pulses should be a so-called spin-lock pulses. A spin-lock pulse is just a pulse that is in phase with the transverse magnetization. To transfer magnetization from I to S, first a normal 90° pulse is given on I. As described before, when this pulse was applied along the x-axis, the magnetization of I is moved to the -y axis. The following spin-lock pulse on I should therefor be given along y or -y. A pulse with any other phase would move the magnetization away from the xy-plane. The phase of the spin-lock pulse on S is arbitrary, because there is no pre-existing transverse S magnetization. By adjusting the amplitudes of the two spin-lock pulses the Hartmann-Hahn condition can be met.

Because there is a large difference in the chemical shifts of the CO and Cα (~180 ppm vs. ~60 pmm), a ^{15}N-^{13}C CP condition can be made specific for either one of them [77]. This is called specific-CP and is achieved by applying a selective, low-power, spin-lock pulse on either the CO or Cα resonance frequency. As will be discussed later, the ability to move magnetization from the backbone ^{15}N specifically to the Cα of same residue or to the CO of the preceding residue is very important for the resonance assignments of proteins.

PDSD and DARR

In order to transfer magnetization from one nucleus to another with conservation of energy, the splitting of energy levels of the two involved nuclei should be equalized. I.e. there needs to be a match in the resonance frequency of the two nuclei, much the same as a pulse needs to be on resonance in order to excite a specific nucleus. The Hartmann-Hahn matching condition in the previous section is an example of this and uses rf pulses to create a matching condition between two nuclei with different energy levels. In the case of two nuclei of the same isotope, the resonance frequencies are not very far apart. When assuring that the linewidths are (temporarily) broad enough, there is a partial overlap between the resonance frequencies that allow the transfer of magnetization, with conservation of energy. To accomplished this for a magnetization transfer between two ^{13}C nuclei, the ^{1}H decoupling can simply be turned off during a mixing period. This is know as proton driven spin diffusion (PDSD) [78][79]. Before the ^{1}H decoupling is switched off, the ^{13}C-magnetization in the transverse plane is stored on the z-axis by giving one 90°-pulse, to prevent T2 relaxation. Depending on the duration of the mixing period, the magnetization spreads further through space. To overcome larger energy differences, for instance between CO and methyl carbons, or at higher magnetic field strength, this method can be made more efficient with a technique called dipolar assisted rotational resonance (DARR)[80][81]. During DARR, the protons are irradiated with a field that one or two times the MAS frequency.

RFDR

Radio frequency driven recoupling (RFDR) is a homonuclear magnetization transfer method that employs a series of rotor synchronized 180°-pulses [82]. The pulses counteract the averaging by MAS and thereby partially reintroduce anisotropic interactions that allow transfer of magnetization between like nuclei. Like DARR, the magnetization is transferred through space. In this studies RFDR is used to obtain spectra that provided distance restraints between amide protons in the backbone of OmpG. In figure 6.2 the two pulse sequence used for this purpose are shown. The RFDR mixing period is shown between square brackets.

Scalar transfers

The scalar coupling can be used for homonuclear and heteronuclear magnetization transfers. Insensitive nuclei enhanced by polarization transfer (INEPT) is a heteronuclear transfer method and is one of the main building blocks of solution NMR experiments and is sometimes used in experiments in the solid state aswell. In these studies, scalar coupling based ^{13}C-^{13}C transfer steps are used in the proton detected pulse sequences described in chapter 3 and illustrated in figures 3.2 and 3.3. Such homonuclear scalar coupling based transfers basically employ the same pulse sequence as INEPT, with the only difference that all pulses are applied to the same isotope.

To transfer magnetization from spin I to spin S, first the magnetization of I in brought into the transverse plane. This can be done by applying a 90°-pulse to spin I. In the case of the homonuclear ^{13}C-^{13}C transfers in chapter 3, this is done by a ^{1}H-^{13}C CP step. Subsequently a delay of duration 1/(2J) is given (where J is the scalar coupling constant) with a 180°-pulse on both spins I and S in the middle of this period. During the delay, the scalar coupling evolves. The 180°-pulse on I refocuses the chemical shift evolution, while the simultaneous 180°-pulse on S prevents the scalar coupling to be refocused as well. At the end of the delay a antiphase spin state $2I_xS_z$. By two simultaneous 90°-pulses on I and S, the state 2 $2I_zS_y$ is created.

Sequential Assignment of solid state NMR spectra

In most NMR studies very little information can be obtained before the chemical shifts of the nuclei that are interesting in the context of the biological question are known. Sometimes those are only a few, for instance when one knows on forehand which residue plays an important role in a biological process. However if the goal is to calculate the structure and study the overall dynamics of the protein a fairly complete mapping between resonance frequencies and nuclei in the molecule has to be present [83]. This mapping process is referred to as sequential assignment and often is the most time-consuming part of an NMR study. The general idea behind sequential assignment methodologies is the following: the graph that arises from a set of correlation spectra is mapped on the molecular topology. In most (but not all) methodologies to find this mapping, the process is divided into two steps. In the first step, parts of the total signal pattern are identified that correspond to individual residues. In a second step, connectivities are found between these signal patterns and a larger signal pattern that belongs to a set of sequentially connected residues is created. Because the nuclei in different amino acids give rise to a different combination of chemical shifts, sometimes referred to as "fingerprint patterns", the sets of signals can be classified down to a few or sometimes even one type of amino acid. When this is done, the larger pattern can be mapped to a subsequence in the protein that matches these possible residue type assignments.

From here on I will call the collection of resonance frequencies that belong to one residue a spin system. The term "spin system" is often used in NMR in a somewhat less confined sense, meaning a set of resonances that are in some way influenced by one another. However, since this thesis will deal with sequential assignment for a large extent, it is good to have a defined way to describe this object. Also the CCPNMR Analysis software, that is used to analyze NMR spectra, uses the term as I just defined it.

In order to get a unique match between the potential residue types of a sequential stretch of spin systems and a subsequence in the protein, the stretches should in general be long enough. I.e. the longer a connected stretch is, the higher the chance there is only one possible location along the protein sequence were this stretch fits. Of course it highly depends on the length of the protein sequence how many spin system have to be sequentially connected before a unique match along the protein sequence can be found. In figure 1.11 the fraction of unique subsequences of length 1 (just one amino acid), 2 and 3 are plotted vs. the length of the protein. As can be seen, even for very large proteins, connected stretches of 3 spin systems can in theory be uniquely matched to a subsequence of the protein in the majority of cases. This is of course under the assumption that each spin system can be uniquely typed to one amino acid type, which is not the case in practice since some amino acids give rise to very similar signal sets. Therefore in practice often somewhat longer stretches need to be generated before a unique match to a subsequence in the protein can be found. Furthermore, the more resonance frequencies of a spin system are known (showing a larger part of the fingerprint pattern), the more specifically the amino acid type can be predicted. Especially ^{13}C chemical shifts in the side-chain are very good indicators of amino acid type. Which resonances can be accessed is closely related to the types of experiments performed.

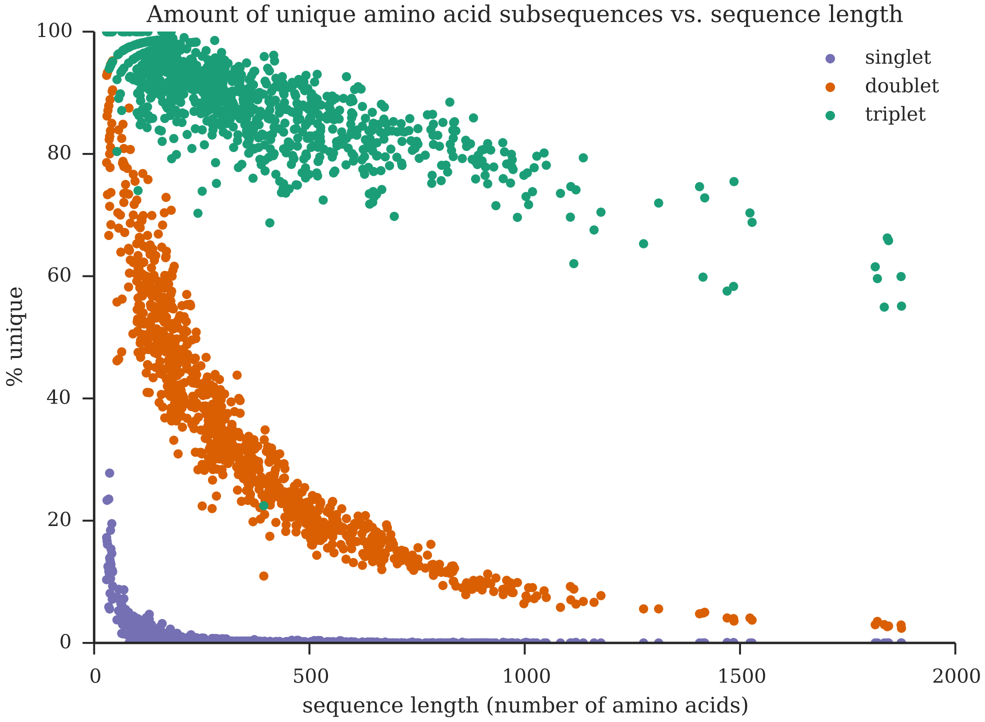


Figure 1.11: Percentage of subsequences that is only present in the sequence once. Purple, orange and green colors correspond to subsequences of length 1,2 and 3 respectively. This plot is made using 1000 membrane protein sequences from the uniProt database. Every point represents one protein. As expected, the amount of single amino acids that only appear in the sequence once very quickly drops off with increasing sequence length. At the other side, even for the largest proteins still more than half of all triplets (subsequences of length 3) is unique in the sequence. Of course, just because the subsequence is unique does not necessarily mean that the subsequence can be distinguished from all other subsequences based on the chemical shifts of these residues.

For the assignment of larger proteins both steps in the assignment process become more difficult. The bottleneck is the chemical shift degeneracy. The larger the amount of NMR active nuclei in a protein, the smaller the average spacing between the resonance frequencies will be. This causes two problems that are very closely related, but should be considered both distinctly:

1. Spectral crowding, the overlap of signals in spectra.
2. Ambiguity of cross-peak assignment.

The first problem means that spectra get very hard to interpret if signals are piled on top of each other. The second problem maybe needs a little more explanation. In order to connect spin systems to form a sequential stretch, cross peaks between them in correlation spectra are needed. However, if for each dimension of a cross-peak there is a large number of nuclei with a matching resonance frequency, it becomes unclear which resonances are really correlated by the peak. This makes it very hard two prove to spin systems are sequentially connected. It is not so much that this type of peak is necessarily really overlapping with another peak, it is the overlap of assignment possibilities. Of course, instead of just one cross-peak, a full pattern of cross-peaks between the different resonances in the two spin systems is used to establish a sequential connection. However, if the assignment ambiguity of most of the cross-peaks is too large, even the patterns of cross-peaks become ambiguous.

Of course, in the end both these problems boil down to linewidth. If linewidths were infinitely narrow, none of these problems would exist. There would be no signal overlap and every cross-peak could be explained by the correlation of spins of exactly the same number as its dimensions. If this were the case, there would be a very clear one to one mapping between the signal set and the molecular topology.

To combat the problem of spectral crowding and assignment ambiguity of cross-peaks, in general three approaches can be taken:

1. Reduce the number of signals in the signal set by isotopically labeling only a subset of the nuclei in the protein.
2. Spread the signal set over more dimensions.
3. Decrease linewidths.

All three methods have been used to facilitate the sequential assignment of OmpG. Over the last few years, solid state NMR has seen a significant progress in the method development for sequential assignment. This is directly reflected by the different experiments and assignment strategies used during these studies. A first chapter will explain the efforts we did to assign OmpG using ^{13}C-detected NMR, because that is how this was generally done at the time this project started off. Afterward, the progress that was made in this project using ^{1}H-detected experiments will be shown. Although it might seem that the use of the latter method makes ^{13}C-detection redundant, the results from the ^{13}C-detected experiments are not just included for chronological completeness. As will be clear, a lot of information obtained from the ^{13}C-detected spectra is still very valuable and, although not impossible, harder to access using ^{1}H-detection. Spectra from both methods have been used in conjunction and in a complementary fashion. Therefore, in chapter 4, the combination of both types of spectra will be discussed. In chapter 5, a computational tool we specifically designed to aid the sequential assignment of solid-state NMR data is discussed. Finally, in chapter 6, the OmpG structure we were able to calculate is presented.

References

[1] D. W. Deamer. “The First Living Systems: A Bioenergetic Perspective.” *Microbiology and Molecular Biology Reviews* 61.2 (Jan. 1997), pp. 239–261.

[2] A. G. Cairns-Smith. *Genetic Takeover and the Mineral Origins of Life*. Cambridge; New York: Cambridge University Press, 1982.

[3] E. V. Koonin and W. Martin. "On the Origin of Genomes and Cells within Inorganic Compartments". *Trends in Genetics* 21.12 (Dec. 2005), pp. 647–654. DOI: 10.1016/j.tig.2005.09.006.

[4] G. Jékely. "Did the Last Common Ancestor Have a Biological Membrane?" *Biology Direct* 1 (2006), p. 35. DOI: 10.1186/1745-6150-1-35.

[5] A. Y. Mulkidjanian, M. Y. Galperin, and E. V. Koonin. "Co-Evolution of Primordial Membranes and Membrane Proteins". *Trends in Biochemical Sciences* 34.4 (Apr. 2009), pp. 206–215. DOI: 10.1016/j.tibs.2009.01.005.

[6] W. Dowhan. "MOLECULAR BASIS FOR MEMBRANE PHOSPHOLIPID DIVERSITY:Why Are There So Many Lipids?" *Annual Review of Biochemistry* 66.1 (1997), pp. 199–232. DOI: 10.1146/annurev.biochem.66.1.199.

[7] T. H. Haines. "Do Sterols Reduce Proton and Sodium Leaks through Lipid Bilayers?" *Progress in Lipid Research* 40.4 (July 2001), pp. 299–324. DOI: 10.1016/S0163-7827(01)00009-1.

[8] E. A. Dennis. "Lipidomics Joins the Omics Evolution". *Proceedings of the National Academy of Sciences* 106.7 (Feb. 2009), pp. 2089–2090. DOI: 10.1073/pnas.0812636106.

[9] P. T. Ivanova, S. B. Milne, D. S. Myers, and H. A. Brown. "Lipidomics: A Mass Spectrometry Based Systems Level Analysis of Cellular Lipids". *Current Opinion in Chemical Biology*. Omics/Biopolymers/Model Systems 13.5–6 (Dec. 2009), pp. 526–531. DOI: 10.1016/j.cbpa.2009.08.011.

[10] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, and S. Subramaniam. "LMSD: LIPID MAPS Structure Database". *Nucleic Acids Research* 35.suppl 1 (Jan. 2007), pp. D527–D532. DOI: 10.1093/nar/gkl838.

[11] G. van Meer, D. R. Voelker, and G. W. Feigenson. "Membrane Lipids: Where They Are and How They Behave". *Nature Reviews Molecular Cell Biology* 9.2 (Feb. 2008), pp. 112–124. DOI: 10.1038/nrm2330.

[12] M. Kates. "The Phytanyl Ether-Linked Polar Lipids and Isoprenoid Neutral Lipids of Extremely Halophilic Bacteria". *Progress in the Chemistry of Fats and other Lipids* 15.4 (Jan. 1977), pp. 301–342. DOI: 10.1016/0079-6832(77)90011-8.

[13] J. Peretó, P. López-García, and D. Moreira. "Ancestral Lipid Biosynthesis and Early Membrane Evolution". *Trends in Biochemical Sciences* 29.9 (Sept. 2004), pp. 469–477. DOI: 10.1016/j.tibs.2004.07.002.

[14] T. Kimura, W. Jennings, and R. M. Epand. "Roles of Specific Lipid Species in the Cell and Their Molecular Mechanism". *Progress in Lipid Research* 62 (Apr. 2016), pp. 75–92. DOI: 10.1016/j.plipres.2016.02.001.

[15] H. Nikaido. "Molecular Basis of Bacterial Outer Membrane Permeability Revisited". *Microbiology and Molecular Biology Reviews* 67.4 (Jan. 2003), pp. 593–656. DOI: 10.1128/MMBR.67.4.593-656.2003.

[16] J. Gidden, J. Denson, R. Liyanage, D. M. Ivey, and J. O. Lay Jr. "Lipid Compositions in Escherichia Coli and Bacillus Subtilis during Growth as Determined by MALDI-TOF and TOF/TOF Mass Spectrometry". *International Journal of Mass Spectrometry*. A Collection of Invited Papers Dedicated to Michael T. Bowers on the Occasion of his 70th Birthday 283.1–3 (June 2009), pp. 178–184. DOI: 10.1016/j.ijms.2009.03.005.

[17] T. A. Garrett, A. C. O'Neill, and M. L. Hopson. "Quantification of Cardiolipin Molecular Species in Escherichia Coli Lipid Extracts Using Liquid Chromatography/Electrospray Ionization Mass Spectrometry". *Rapid Communications in Mass Spectrometry* 26.19 (Oct. 2012), pp. 2267–2274. DOI: 10.1002/rcm.6350.

[18] C. Sohlenkamp and O. Geiger. "Bacterial Membrane Lipids: Diversity in Structures and Pathways". *FEMS Microbiology Reviews* 40.1 (Jan. 2016), pp. 133–159. DOI: 10.1093/femsre/fuv008.

[19] A. Lehninger, D. Nelson, and M. Cox. *Lehninger Principles of Biochemistry*. W H Freeman & Company, 2008.

[20] S. H. White, A. S. Ladokhin, S. Jayasinghe, and K. Hristova. "How Membranes Shape Protein Structure". *Journal of Biological Chemistry* 276.35 (Aug. 2001), pp. 32395–32398. DOI: 10.1074/jbc.R100008200.

[21] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. “Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete genomes1”. *Journal of Molecular Biology* 305.3 (Jan. 2001), pp. 567–580. DOI: 10.1006/jmbi.2000.4315.

[22] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins. “How Many Drug Targets Are There?” *Nature Reviews Drug Discovery* 5.12 (Dec. 2006), pp. 993–996. DOI: 10.1038/nrd2199.

[23] M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal. “Drug—target Network”. *Nature Biotechnology* 25.10 (Oct. 2007), pp. 1119–1126. DOI: 10.1038/nbt1338.

[24] S. White. *Membrane Proteins of Known 3D Structure http://blanco.biomol.uci.edu/mpstruc/.*

[25] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. “The Protein Data Bank”. *European Journal of Biochemistry* 80.2 (Nov. 1977), pp. 319–324. DOI: 10.1111/j.1432-1033.1977.tb11885.x.

[26] J.-J. Lacapère, E. Pebay-Peyroula, J.-M. Neumann, and C. Etchebest. “Determining Membrane Protein Structures: Still a Challenge!” *Trends in Biochemical Sciences* 32.6 (June 2007), pp. 259–270. DOI: 10.1016/j.tibs.2007.04.001.

[27] E. M. Landau and J. P. Rosenbusch. “Lipidic Cubic Phases: A Novel Concept for the Crystallization of Membrane Proteins”. *Proceedings of the National Academy of Sciences* 93.25 (Oct. 1996), pp. 14532–14535.

[28] M. Caffrey. “A Comprehensive Review of the Lipid Cubic Phase or in Meso Method for Crystallizing Membrane and Soluble Proteins and Complexes”. *Acta Crystallographica. Section F, Structural Biology Communications* 71.Pt 1 (Jan. 2015), pp. 3–18. DOI: 10.1107/S2053230X14026843.

[29] C. R. Sanders and G. C. Landis. “Reconstitution of Membrane Proteins into Lipid-Rich Bilayered Mixed Micelles for NMR Studies”. *Biochemistry* 34.12 (Mar. 1995), pp. 4030–4040. DOI: 10.1021/bi00012a022.

[30] J. M. Glück, M. Wittlich, S. Feuerstein, S. Hoffmann, D. Willbold, and B. W. Koenig. “Integral Membrane Proteins in Nanodiscs Can Be Studied by Solution NMR Spectroscopy”. *Journal of the American Chemical Society* 131.34 (Sept. 2009), pp. 12060–12061. DOI: 10.1021/ja904897p.

[31] A.-C. Milazzo, A. Cheng, A. Moeller, D. Lyumkis, E. Jacovetty, J. Polukas, M. H. Ellisman, N.-H. Xuong, B. Carragher, and C. S. Potter. “Initial Evaluation of a Direct Detection Device Detector for Single Particle Cryo-Electron Microscopy”. *Journal of Structural Biology* 176.3 (Dec. 2011), pp. 404–408. DOI: 10.1016/j.jsb.2011.09.002.

[32] X. Li, P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng. “Electron Counting and Beam-Induced Motion Correction Enable near-Atomic-Resolution Single-Particle Cryo-EM”. *Nature Methods* 10.6 (June 2013), pp. 584–590. DOI: 10.1038/nmeth.2472.

[33] X. Li, S. Q. Zheng, K. Egami, D. A. Agard, and Y. Cheng. “Influence of Electron Dose Rate on Electron Counting Images Recorded with the K2 Camera”. *Journal of Structural Biology* 184.2 (Nov. 2013), pp. 251–260. DOI: 10.1016/j.jsb.2013.08.005.

[34] R. D. Zorzi, W. Mi, M. Liao, and T. Walz. “Single-Particle Electron Microscopy in the Study of Membrane Protein Structure”. *Microscopy* (Oct. 2015), dfv058. DOI: 10.1093/jmicro/dfv058.

[35] K. R. Vinothkumar. “Membrane Protein Structures without Crystals, by Single Particle Electron Cryomicroscopy”. *Current Opinion in Structural Biology* 33 (Aug. 2015), pp. 103–114. DOI: 10.1016/j.sbi.2015.07.009.

[36] S. J. Tilley, E. V. Orlova, R. J. C. Gilbert, P. W. Andrew, and H. R. Saibil. “Structural Basis of Pore Formation by the Bacterial Toxin Pneumolysin”. *Cell* 121.2 (Apr. 2005), pp. 247–256. DOI: 10.1016/j.cell.2005.02.033.

[37] L. Wang and F. J. Sigworth. “Structure of the BK Potassium Channel in a Lipid Membrane from Electron Cryomicroscopy”. *Nature* 461.7261 (Sept. 2009), pp. 292–295. DOI: 10.1038/nature08291.

[38] S. H. White and W. C. Wimley. “MEMBRANE PROTEIN FOLDING AND STABILITY: Physical Principles”. *Annual Review of Biophysics and Biomolecular Structure* 28.1 (1999), pp. 319–365. DOI: 10.1146/annurev.biophys.28.1.319.

[39] W.-M. Yau, W. C. Wimley, K. Gawrisch, and S. H. White. “The Preference of Tryptophan for Membrane Interfaces”. *Biochemistry* 37.42 (Oct. 1998), pp. 14713–14718. DOI: 10.1021/bi980809c.

[40] J. A. Killian and G. von Heijne. “How Proteins Adapt to a Membrane–water Interface”. *Trends in Biochemical Sciences* 25.9 (Sept. 2000), pp. 429–434. DOI: 10.1016/S0968-0004(00)01626-1.

[41] W. C. Wimley. “Toward Genomic Identification of β-Barrel Membrane Proteins: Composition and Architecture of Known Structures”. *Protein Science : A Publication of the Protein Society* 11.2 (Feb. 2002), pp. 301–312.

[42] A. G. Lee. “How Lipids and Proteins Interact in a Membrane: A Molecular Approach”. *Molecular BioSystems* 1.3 (Sept. 2005), pp. 203–212. DOI: 10.1039/B504527D.

[43] H.-X. Zhou and T. A. Cross. “Influences of Membrane Mimetic Environments on Membrane Protein Structures”. *Annual review of biophysics* 42 (2013), pp. 361–392. DOI: 10.1146/annurev-biophys-083012-130326.

[44] O. S. Andersen and R. E. Koeppe. “Bilayer Thickness and Membrane Protein Function: An Energetic Perspective”. *Annual Review of Biophysics and Biomolecular Structure* 36.1 (2007), pp. 107–130. DOI: 10.1146/annurev.biophys.36.040306.132643.

[45] J. Ellena, P. Lackowicz, H. Mongomery, and D. Cafiso. “Membrane Thickness Varies Around the Circumference of the Transmembrane Protein BtuB”. *Biophysical Journal* 100.5 (Mar. 2011), pp. 1280–1287. DOI: 10.1016/j.bpj.2011.01.055.

[46] A. R. Battle, P. Ridone, N. Bavi, Y. Nakayama, Y. A. Nikolaev, and B. Martinac. “Lipid–protein Interactions: Lessons Learned from Stress”. *Biochimica et Biophysica Acta (BBA) - Biomembranes*. Lipid-protein interactions 1848.9 (Sept. 2015), pp. 1744–1756. DOI: 10.1016/j.bbamem.2015.04.012.

[47] C. Bechara and C. V. Robinson. “Different Modes of Lipid Binding to Membrane Proteins Probed by Mass Spectrometry”. *Journal of the American Chemical Society* 137.16 (Apr. 2015), pp. 5240–5247. DOI: 10.1021/jacs.5b00420.

[48] N. P. Barrera, M. Zhou, and C. V. Robinson. “The Role of Lipids in Defining Membrane Protein Interactions: Insights from Mass Spectrometry”. *Trends in Cell Biology* 23.1 (Jan. 2013), pp. 1–8. DOI: 10.1016/j.tcb.2012.08.007.

[49] A. G. Lee. “Lipid–protein Interactions in Biological Membranes: A Structural Perspective”. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1612.1 (May 2003), pp. 1–40. DOI: 10.1016/S0005-2736(03)00056-7.

[50] H. Palsdottir and C. Hunte. “Lipids in Membrane Protein Structures”. *Biochimica et Biophysica Acta (BBA) - Biomembranes*. Lipid-Protein Interactions 1666.1–2 (Nov. 2004), pp. 2–18. DOI: 10.1016/j.bbamem.2004.06.012.

[51] C. Hunte. “Specific Protein–lipid Interactions in Membrane Proteins”. *Biochemical Society Transactions* 33.5 (Oct. 2005), pp. 938–942. DOI: 10.1042/BST0330938.

[52] P. L. Yeagle. “Non-Covalent Binding of Membrane Lipids to Membrane Proteins”. *Biochimica et Biophysica Acta (BBA) - Biomembranes*. Membrane Structure and Function: Relevance in the Cell’s Physiology, Pathology and Therapy 1838.6 (June 2014), pp. 1548–1559. DOI: 10.1016/j.bbamem.2013.11.009.

[53] W. C. Wimley. “The Versatile β-Barrel Membrane Protein”. *Current Opinion in Structural Biology* 13.4 (Aug. 2003), pp. 404–411. DOI: 10.1016/S0959-440X(03)00099-X.

[54] Y. Gu, H. Li, H. Dong, Y. Zeng, Z. Zhang, N. G. Paterson, P. J. Stansfeld, Z. Wang, Y. Zhang, W. Wang, and C. Dong. “Structural Basis of Outer Membrane Protein Insertion by the BAM Complex”. *Nature* 531.7592 (Mar. 2016), pp. 64–69. DOI: 10.1038/nature17199.

[55] J. W. Fairman, N. Noinaj, and S. K. Buchanan. “The Structural Biology of β-Barrel Membrane Proteins: A Summary of Recent Reports”. *Current Opinion in Structural Biology*. Engineering and design / Membranes 21.4 (Aug. 2011), pp. 523–531. DOI: 10.1016/j.sbi.2011.05.005.

[56] S. A. Shahid, B. Bardiaux, W. T. Franks, L. Krabben, M. Habeck, B.-J. van Rossum, and D. Linke. “Membrane-Protein Structure Determination by Solid-State NMR Spectroscopy of Microcrystals”. *Nature Methods* 9.12 (Dec. 2012), pp. 1212–1217. DOI: 10.1038/nmeth.2248.

[57] R. Misra and S. A. Benson. “A Novel Mutation, Cog, Which Results in Production of a New Porin Protein (OmpG) of Escherichia Coli K-12.” *Journal of Bacteriology* 171.8 (Aug. 1989), pp. 4105–4111.

[58] D. A. Fajardo, J. Cheung, C. Ito, E. Sugawara, H. Nikaido, and R. Misra. “Biochemistry and Regulation of a Novel Escherichia Coli K-12 Porin Protein, OmpG, Which Produces Unusually Large Channels”. *Journal of Bacteriology* 180.17 (Sept. 1998), pp. 4452–4459.

[59] S. Conlan, Y. Zhang, S. Cheley, and H. Bayley. “Biochemical and Biophysical Characterization of OmpG: A Monomeric Porin”. *Biochemistry* 39.39 (Oct. 2000), pp. 11845–11854.

[60] M. Behlau, D. J. Mills, H. Quader, W. Kühlbrandt, and J. Vonck. “Projection Structure of the Monomeric Porin OmpG at 6 Å resolution1”. *Journal of Molecular Biology* 305.1 (Jan. 2001), pp. 71–77. DOI: 10.1006/jmbi.2000.4284.

[61] G. V. Subbarao and B. van den Berg. “Crystal Structure of the Monomeric Porin OmpG”. *Journal of Molecular Biology* 360.4 (July 2006), pp. 750–759. DOI: 10.1016/j.jmb.2006.05.045.

[62] Ö. Yildiz, K. R. Vinothkumar, P. Goswami, and W. Kühlbrandt. “Structure of the Monomeric Outer-membrane Porin OmpG in the Open and Closed Conformation”. *The EMBO Journal* 25.15 (Aug. 2006), pp. 3702–3713. DOI: 10.1038/sj.emboj.7601237.

[63] B. Liang and L. K. Tamm. “Structure of Outer Membrane Protein G by Solution NMR Spectroscopy”. *Proceedings of the National Academy of Sciences* 104.41 (Sept. 2007), pp. 16140–16145. DOI: 10.1073/pnas.0705466104.

[64] W. Delano. “The PyMOL Molecular Graphics System” (2002).

[65] L.-Q. Gu, O. Braha, S. Conlan, S. Cheley, and H. Bayley. “Stochastic Sensing of Organic Analytes by a Pore-Forming Protein Containing a Molecular Adapter”. *Nature* 398.6729 (Apr. 1999), pp. 686–690. DOI: 10.1038/19491.

[66] H. Bayley and P. S. Cremer. “Stochastic Sensors Inspired by Biology”. *Nature* 413.6852 (Sept. 2001), pp. 226–230. DOI: 10.1038/35093038.

[67] M. Chen, S. Khalid, M. S. P. Sansom, and H. Bayley. “Outer Membrane Protein G: Engineering a Quiet Pore for Biosensing”. *Proceedings of the National Academy of Sciences* 105.17 (Apr. 2008), pp. 6272–6277. DOI: 10.1073/pnas.0711561105.

[68] F. Korkmaz-Özkan, S. Köster, W. Kühlbrandt, W. Mäntele, and Ö. Yildiz. “Correlation between the OmpG Secondary Structure and Its pH-Dependent Alterations Monitored by FTIR”. *Journal of Molecular Biology* 401.1 (Aug. 2010), pp. 56–67. DOI: 10.1016/j.jmb.2010.06.015.

[69] F. Korkmaz, S. Köster, Ö. Yildiz, and W. Mäntele. “In Situ Opening/Closing of OmpG from E. Coli and the Splitting of β-Sheet Signals in ATR–FTIR Spectroscopy”. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 91 (June 2012), pp. 395–401. DOI: 10.1016/j.saa.2012.01.025.

[70] W. Grosse, G. Psakis, B. Mertins, P. Reiss, D. Windisch, F. Brademann, J. Bürck, A. Ulrich, U. Koert, and L.-O. Essen. “Structure-Based Engineering of a Minimal Porin Reveals Loop-Independent Channel Closure”. *Biochemistry* 53.29 (July 2014), pp. 4826–4838. DOI: 10.1021/bi500660q.

[71] T. Zhuang, C. Chisholm, M. Chen, and L. K. Tamm. “NMR-Based Conformational Ensembles Explain pH-Gated Opening and Closing of OmpG Channel”. *Journal of the American Chemical Society* 135.40 (Oct. 2013), pp. 15101–15113. DOI: 10.1021/ja408206e.

[72] P. J. Hore. *Nuclear Magnetic Resonance*. Oxford ; New York: Oxford University Press, U.S.A., June 1989.

[73] F. Bloch. “Theory of Line Narrowing by Double-Frequency Irradiation”. *Physical Review* 111.3 (Aug. 1958), pp. 841–853. DOI: 10.1103/PhysRev.111.841.

[74] A. E. Bennett, C. M. Rienstra, M. Auger, K. V. Lakshmi, and R. G. Griffin. “Heteronuclear Decoupling in Rotating Solids”. *The Journal of Chemical Physics* 103.16 (Oct. 1995), pp. 6951–6958. DOI: 10.1063/1.470372.

[75] B. M. Fung, A. K. Khitrin, and K. Ermolaev. “An Improved Broadband Decoupling Sequence for Liquid Crystals and Solids”. *Journal of Magnetic Resonance* 142.1 (Jan. 2000), pp. 97–101. DOI: 10.1006/jmre.1999.1896.

[76] S. R. Hartmann and E. L. Hahn. “Nuclear Double Resonance in the Rotating Frame”. *Physical Review* 128.5 (Dec. 1962), pp. 2042–2053. DOI: 10.1103/PhysRev.128.2042.

[77] M. Baldus, A. T. Petrovka, J. Herzfeld, and R. G. Griffin. “Cross Polarization in the Tilted Frame: Assignment and Spectral Simplification in Heteronuclear Spin Systems”. *Molecular Physics* 95.6 (Dec. 1998), pp. 1197–1207. DOI: 10.1080/00268979809483251.

[78] N. Bloembergen. “On the Interaction of Nuclear Spins in a Crystalline Lattice”. *Physica* 15.3 (May 1949), pp. 386–426. DOI: 10.1016/0031-8914(49)90114-7.

[79] N. M. Szeverenyi, M. J. Sullivan, and G. E. Maciel. "Observation of Spin Exchange by Two-Dimensional Fourier Transform 13C Cross Polarization-Magic-Angle Spinning". *Journal of Magnetic Resonance (1969)* 47.3 (May 1982), pp. 462–475. DOI: 10.1016/0022-2364(82)90213-X.

[80] K. Takegoshi, S. Nakamura, and T. Terao. "13C–1H Dipolar-Assisted Rotational Resonance in Magic-Angle Spinning NMR". *Chemical Physics Letters* 344.5–6 (Aug. 2001), pp. 631–637. DOI: 10.1016/S0009-2614(01)00791-6.

[81] K. Takegoshi, S. Nakamura, and T. Terao. "13C–1H Dipolar-Driven 13C–13C Recoupling without 13C Rf Irradiation in Nuclear Magnetic Resonance of Rotating Solids". *The Journal of Chemical Physics* 118.5 (Feb. 2003), pp. 2325–2341. DOI: 10.1063/1.1534105.

[82] A. E. Bennett, R. G. Griffin, J. H. Ok, and S. Vega. "Chemical Shift Correlation Spectroscopy in Rotating Solids: Radio Frequency-driven Dipolar Recoupling and Longitudinal Exchange". *The Journal of Chemical Physics* 96.11 (June 1992), pp. 8624–8627. DOI: 10.1063/1.462267.

[83] J. Jee and P. Güntert. "Influence of the Completeness of Chemical Shift Assignments on NMR Structures Obtained with Automated NOE Assignment". *Journal of Structural and Functional Genomics* 4.2-3 (June 2003), pp. 179–189. DOI: 10.1023/A:1026122726574.

Chapter 2

Assignment Using ^{13}C-detected Experiments

Introduction

Types of Experiments

As explained in the general introduction (chapter 1), the assignment process is generally split up into two steps. The first involves the grouping resonance positions belonging to individual residues into spin systems. In ^{13}C-detected solid state NMR this is generally done by evaluation of 2D ^{13}C-^{13}C and NCACX spectra with a short DARR mixing time of about 20 to 50 ms. In these type of spectra only intra-residual cross-peaks are expected. In the second step, these spin systems are connected sequentially, which can be done by analyzing NCOCX spectra that connect the ^{15}N resonance of one residue to the ^{13}C resonances in the previous residue int the protein sequence (N-terminal side). In addition, 2D ^{13}C-^{13}C correlation spectra with longer DARR mixing time of around 150 to 200 ms are used in this step since they contain a lot of short range sequential cross peaks. In figure 2.1 the pulse sequences and the corresponding magnetization transfer pathways for the 2D ^{13}C-^{13}C DARR and NCACX and NCOCX experiments are shown.

There are several advantages in analyzing a combined dataset of ^{13}C-^{13}C and through-backbone NCA/NCO type of spectra. A benefit of the through-backbone experiments is that there is a sense of directionality. If two spin-systems are connected in these spectra one always knows which one is the first in the sequence, making mapping to a subsequence in the protein easier. In contrast, a purely through-space 2D ^{13}C-^{13}C experiments is not directional. At the other side, in 2D ^{13}C-^{13}C correlations very specific cross peaks between the less degenerate side chain resonances can be found. These types of cross-peaks are not present in the NCACX and NCOCX spectra because one of the

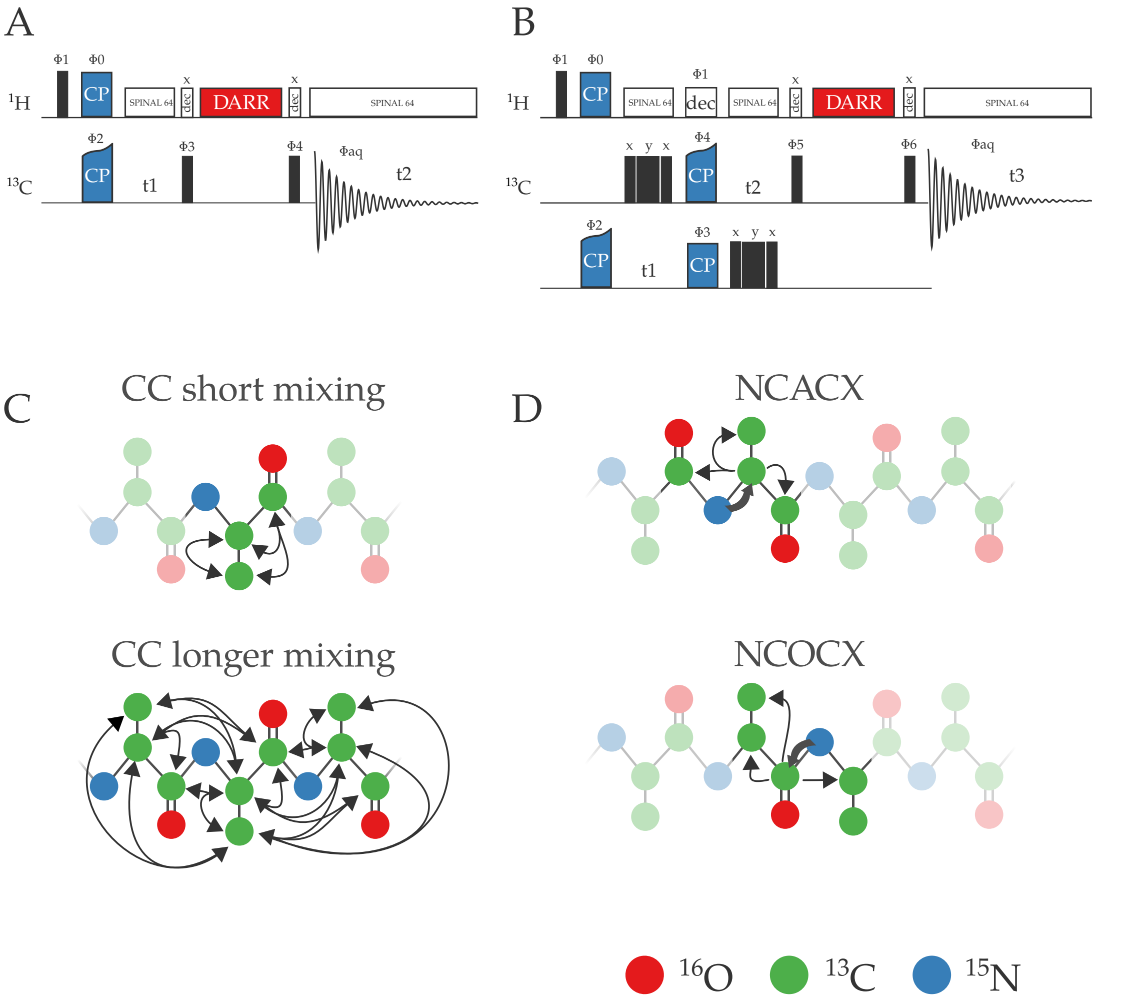


Figure 2.1: Pulse sequences and magnetization transfer schemes in two carbon detected experiments. A) Pulse sequence of 2D CC correlation using DARR. B) Pulse sequence for both NCACX and NCOCX, which are basically identical. Only the exact CP condition is different, resulting in specific transfer from N to Cα or from N to CO. Narrow black rectangles indicate 90°-pulses. Wider black rectangles indicate 180°-pulses. Blue shapes indicate CP steps. Red ractangle indicates DARR mixing. White rectangles indicate decoupling. C) Magnetization transfers of 2D CC correlations. In spectra with a short mixing time (50 ms) only cross-peaks will arise that correlate two nuclei in the same residue. If the mixing time is increased (150-400 ms)long range correlations can be observed. D) Magnetization transfers demonstrating how a sequential walk can be performed using NCACX and NCOCX spectra. Thin arrows indicate magnetization transfer between carbons by DARR. Thick arrow indicates the ^{15}N-^{13}C CP. Protons are left out of this figure for clarity.

^{13}C-dimensions encodes either the Cα or the CO.

When performing a backbone walk using NCACX and NCOCX spectra, the ^{15}N chemical shift is used as a pivot and therefore the usefulness of these spectra is highly dependent on the ^{15}N chemical shift dispersion. Since the ^{15}N T2 relaxation times are very short for the OmpG samples this dispersion is not very high. Additionally, the short T2 times also adversely influence the efficiency of the ^{15}N-^{13}C cross-polarization step and the signal to noise in these experiments is in general lower than in the ^{13}C-^{13}C experiments. There are other ^{13}C-detected experiments possible that complement the NCACX and NCOCX, such as CANCO, CANcoCA and CANCOCX, that circumvent role of the ^{15}N chemical shift as the sole pivot [1][2][3][4]. However these experiments incorporate yet an additional ^{13}C-^{15}N transfer step which decreases the signal to noise even further, making them only applicable to highly ordered samples that have longer T_2 and more efficient ^{15}N-^{13}C transfers. These experiments have only been successfully acquired on small or micro-crystalline proteins and OmpG is neither small nor micro-crystalline.

Isotope Labeling Schemes

Both 2- and 3-dimensional ^{13}C-detected spectra of uniformly ^{13}C-^{15}N labeled OmpG are very crowded, and therefore very hard to assign, for example see figure 2.2. Before I joined the project, different paths were explored to simplify the spectra in order to find starting points for the assignment. There are two main possibilities to do this. The first is to use spectroscopic techniques that reduce the amount of cross peaks in the spectra by specifically selecting resonances based on there spectroscopic or chemical properties. For instance, a route explored in the early stages of the project was to use spectral editing to specifically select methyl resonances [5]. The second strategy that was explored was to reduce the amount of peaks in the spectra by producing a set of selectively labeled samples [6][7]. The labeled samples that were produced can basically be divided into three groups:

1. Amino acid type specific samples: only a subset of the amino acids are ^{15}N, ^{13}C labeled.
2. Uniform 1,3- or 2-glycerol labeled samples: this labeling is produced by feeding bacteria with glycerol as the sole carbon source, labeled either on the 1st and 3rd position or just on the 2nd position.
3. Amino acid type specific 1,3- or 2-glycerol labeled samples: only a subset of the amino acids is labeled with the 1,3 or 2-glycerol labeling pattern.

Forward labeled schemes

To produce the first group of labeling schemes forward labeling is used. This is conceptually the most straight-forward method and involves adding a set of labeled amino acids to an otherwise unlabeled feedstock. The combinations of amino acids that can be labeled together is restricted by the amino acid metabolism. Since it is not possible to suppress certain metabolic routes completely,

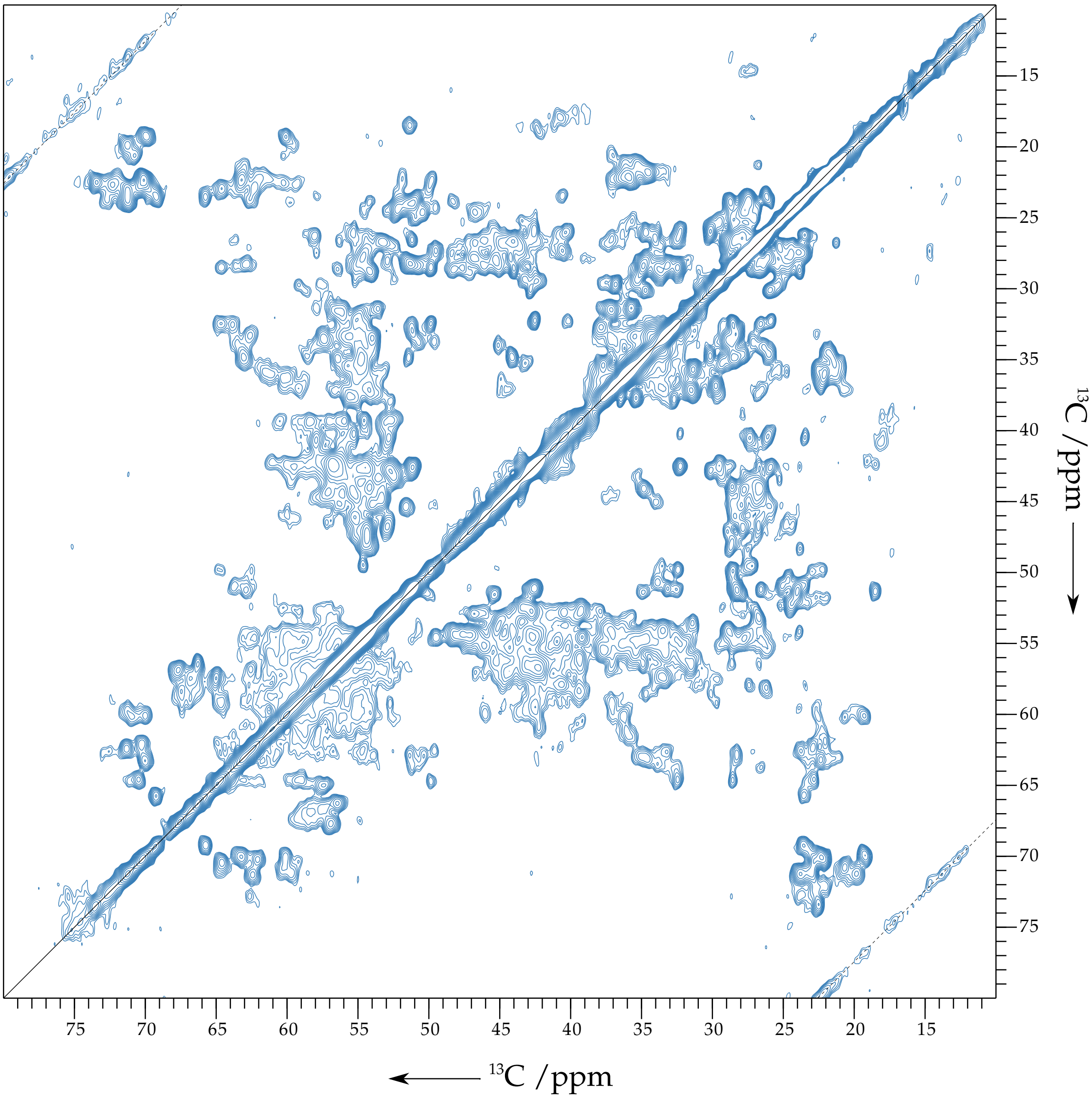


Figure 2.2: 2D ^{13}C-^{13}C DARR spectrum with 25 ms mixing of uniformly labeled OmpG at 400 MHz.

one has to choose a set of amino acids that are either related in the metabolism, form an endpoint in the metabolism (like tyrosine or lysine), or of which the production/use is easily suppressed.

The forward labeling schemes have advantages and disadvantages. The big advantage is that the amino acids that are labeled are uniformly labeled. As a consequence, spectra of these samples contain the full intra-residual peak pattern, which is extremely helpful for grouping resonances into spin systems. This is a lot harder in the spectra of 1,3- or 2-glycerol based labeled samples. This is especially true for the amino acids in group I in figure 2.3. In this group there is only one isotopomer per amino acid in the 1,3-glycerol sample and one isotopomer in 2-glycerol sample, that are completely complementary. Hence, for any given residue two distinct peak patterns arise, but the information that connects these two patterns to one and the same residue is absent. Therefore, to be able to generate spin systems, it is necessary to have a set of samples that is uniformly ^{13}C/^{15}N-labeled on the residue level, preferably with as little as possible overlap of the intra-residual peaks of different amino acids.

1,3- and 2-glycerol labeling

These two labeling schemes are produced by using glycerol that is either labeled on the extreme two carbons (1,3) or the middle carbon (2) as the sole carbon source during protein expression [8][9]. The labeling patterns produced in this fashion are shown in figure 2.3. This labeling scheme has been succesfully used in the assignment and structure calculation of SH3 and αB-crystallin [10][11]. Examples of how to use these labeling patterns to assign large proteins and specifically OmpG have been illustrated in the publication by Higman et al. in 2009 [7].

Apart from decreasing the amount of signals in the spectra, these labeling schemes also produce narrower lines because most directly bound carbon nuclei are not labeled in the same isotopomer. This reduces the remaining ^{13}C-^{13}C homonuclear dipolar coupling and the J-coupling (that are not removed by MAS), which in turn causes lines to be narrower. Additionally, long-range cross-peaks and cross-peaks between sequential residues are easier to obtain, which is particularly important for assignments. For the same reason, these type of labeling schemes are really useful for generating distance restraints used in structure calculations.

As indicated before, the downside of the 1,3- and 2-glycerol labeled samples is that they are not well suited to generate spin systems, because a lot of the intra-residual peaks are missing. At the other side, for exactly the same reason, inter-residual cross-peaks that would otherwise be overlapped by intra-residual peaks can now often be resolved.

Along the same lines, glycerol labeled samples are of limited use in 3 dimensional NCACX and NCOCX spectra. To make a sequential walk, it is essential that the CO peak is present in the strip from the NCACX, so that that the connecting strip in the NCOCX (at the ^{15}N chemical shift of the following residue) can be found. In exactly the same way, when walking “backwards” it is necessary that the NCOCX strip contains the Cα peak, so that the connecting NCACX strip can be found. Since directly bound carbons are almost never simultaneously labeled in these samples (figure 2.3), these peaks are often absent. Therefor spectra of samples that are uniformly labeled on the residue level are always necessary in conjunction to spectra of glycerol labeled samples.

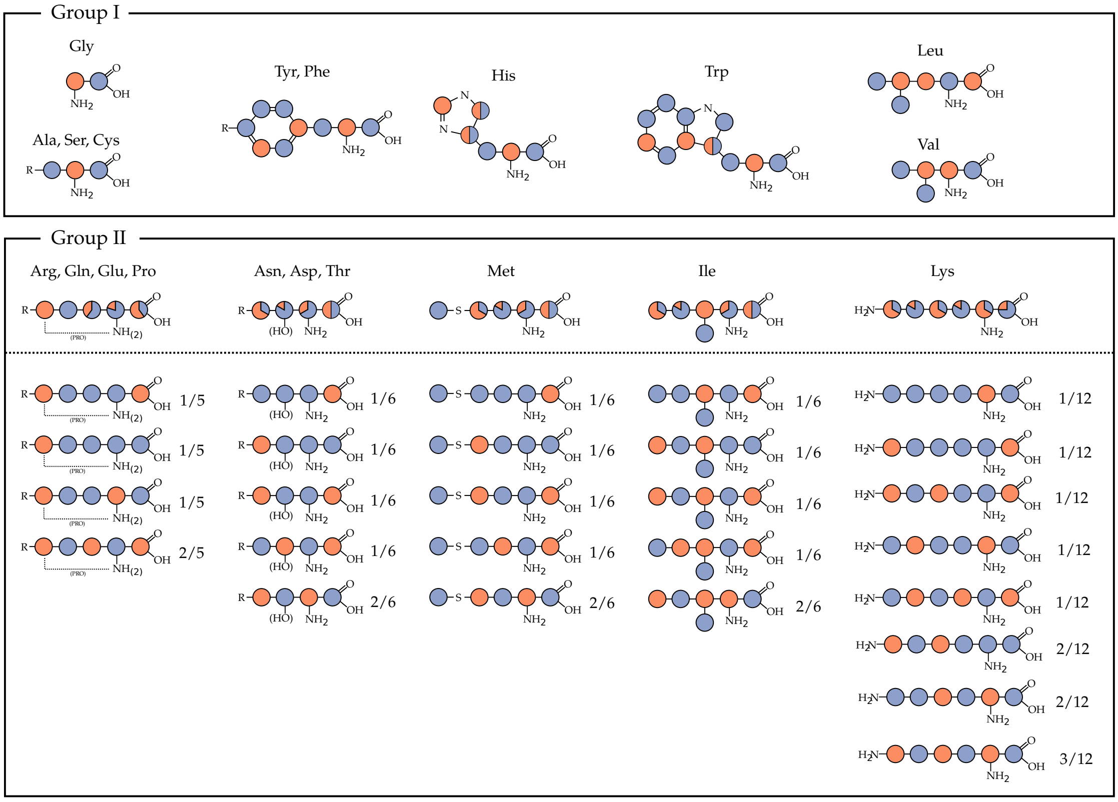


Figure 2.3: Labeling patterns in 20 amino acids when 1,3-glycerol (blue) or 2-glycerol (red) are used as feedstock. Group I consists of amino acids for which there is exactly one isotopomer. For the amino acids in group II, there are multiple isotopomers. Above the dotted line, the average labeling of those amino acids is shown, while underneath the line the individual isotopomers are shown together with the fractions in which they are present.

Amino acid specific 1,3- or 2-glycerol labeling

The third group of labeled samples is produced using reverse labeling. In this case *E. coli* is grown on an isotopically labeled feedstock, here 1,3- or 2-glycerol, and all amino acids that should not be labeled are added in unlabeled form to suppress their metabolism. This technique was pioneered by Hong and Jakes and one of the labeling schemes used here, 2-SHLYGWAVF, is basically identical to the labeling scheme introduced by them as TEASE (**te**n **a**mino acid **s**elective and **e**xtensive labeling), where the tenth amino acid is cysteine, which is not present in OmpG [12]. Using this strategy, two sets of amino acids were produced. The first set of amino acids, SHLYGWAFV, consists out of the

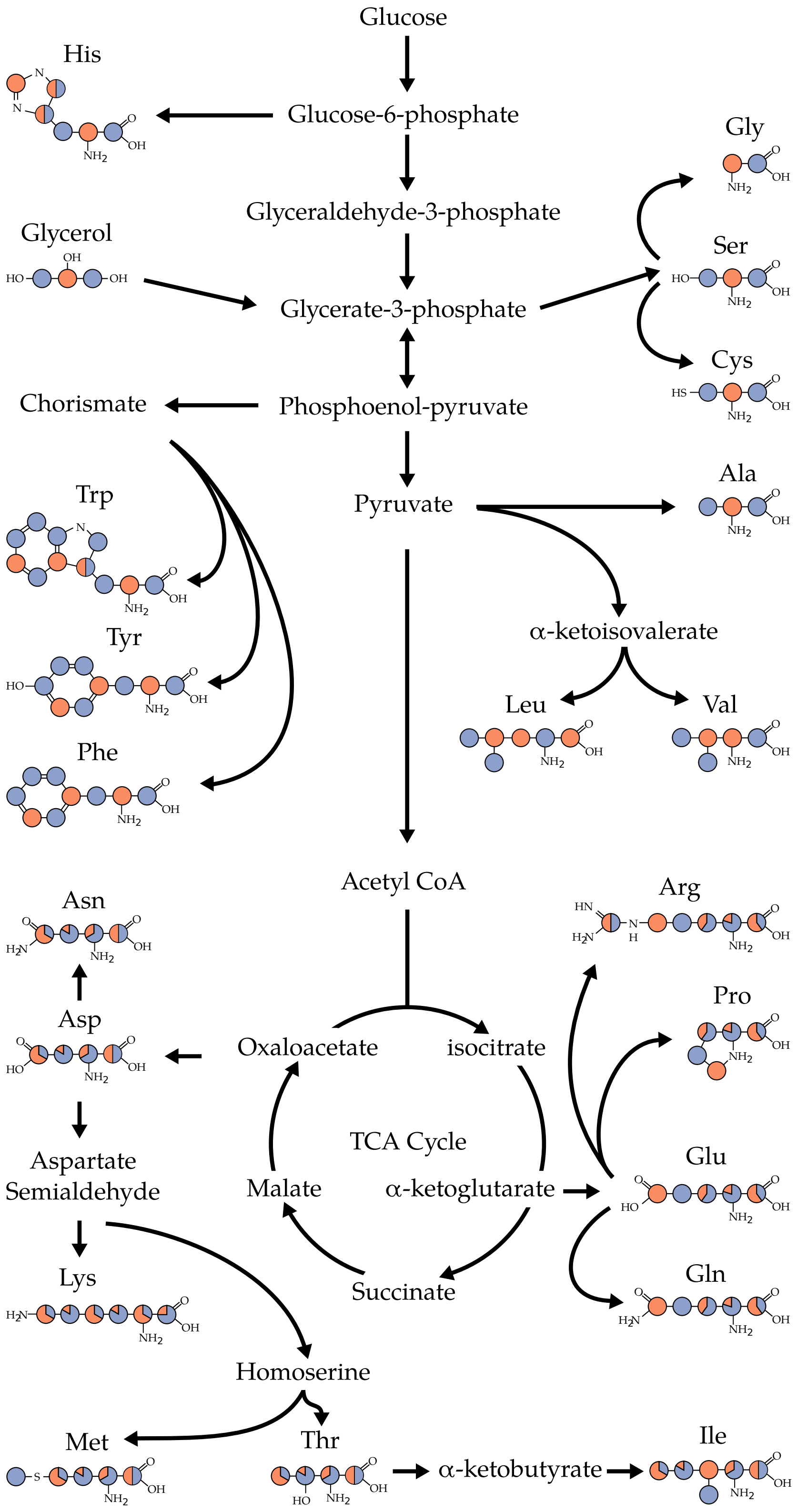


Figure 2.4: Amino acid metabolism leading to the different labeling patterns when 1,3- or 2-labeled glycerol is used as the sole carbon source.

amino acids produced in the glycolysis and the pentose phosphate pathway, see figure 2.4. The other set of amino acids, TEMPQANDSG corresponds to the amino acids produced in the citric acid (TCA) cycle (minus lysine, isoleucine and arginine) plus alanine, glycine and serine. For each of the two sets amino acids, SHLYGWAFV and TEMPQANDSG, two samples were produced using the 1,3- and 2-glycerol labeling strategy, leading to a total of four labeled samples: 2-SHLYGWAFV, 1,3-SHLYGWAFV. 2-TEMPQANDSG and 1,3-TEMPQANDSG.

Co-labeling fraction

To evaluate which peaks are expected in the spectra of labeled samples. The co-labeling fraction of the nuclei on the magnetization transfer pathway has to be evaluated. The co-labeling fraction of a set of nuclei is the fraction of molecules in the sample for which all nuclei in the set are labeled simultaneously. For the labeling schemes described here that are not based on 2- or 1,3-glycerol labeling, the co-labeling fraction of a set of nulcei is either 1 or 0 (disregarding natural abundance of isotopes). For example, in the RIGA(S) sample (introduced later) the co-labeling of isoleucine-Cα and alanine-Cβ is 1 and the co-labeling of isoleucine-Cα and tyrosine-Cβ is 0. For the 2- and 1,3-glycerol based labeling schemes, the individual isotopomers of the amino acids in group II (figure 2.3) have to be taken into account. For example, in a 1,3-glycerol labeled sample, threonine Cα and CO are simultaneously labeled in only 1 out of 6 isotopomers Cα and CO, making the intra-residual co-labeling fraction for these two nuclei 1/6 (0.17). For inter-residual correlations the average labeling over all isotopomers for the nuclei in the set can be multiplied. For example, the inter-residual co-labeling fraction in a 1,3-glycerol labeled sample of the Cα of one threonine and the CO of another threonine in the sequence is $\frac{4}{6}.\frac{3}{6} = \frac{1}{3}$. The co-labeling fraction is used to calculate expected peak patterns in the CCPNMR Analysis plug-ins described in this chapter and in chapter 5 and to generate correct assignment options for ambiguous distance restraints discussed in chapter 6.

Combinations of residues in residue specific labeling schemes

As can be seen in 2.4, the combinations of amino acids that can be labeled simultaneously in a labeling scheme are defined by the bacterial metabolism. For the freedom that is left, a concession has to be made between two major conflicting interests. At the one hand the crowding in the resulting spectra should be reduced as much as possible. On the other hand, as many as possible neighboring residues should be co-labeled in at least one of the labeling schemes. For example, alanine is co-labeled with every other amino acid, except for lysine, in at least one of the labeling schemes (figure 2.5). That means that there will almost always be one or more spectra were the cross peaks between a sequential stretch involving an alanine can be observed, thereby enabling the assignment of this stretch. At the other hand, proline and tyrosine (as an example) are not co-labeled in any of the residue specific labeling schemes, so whenever there is a proline-tyrosine pair in the sequence, the more crowded spectra from non-residue specific labeled samples have to be used to find the cross peaks connecting them.

As discussed in the previous chapter, it is preferable to be able to connect at least three spin systems to unambiguously assign them to a unique subsequence in the protein. By having a set of labeling schemes with a certain overlap, it is possible to analyze spectra in parallel to find the sequential cross peaks in order to produce longer stretches of connected spin systems. In figure 2.6, such stretches are hight-lighted on the OmpG sequence. Whenever the color changes there is a “dead end”, where no residue specific labeling scheme connects two neighboring residues. On average, a given residue in the sequence is part of a stretch of 5.5 residues, which allows an unambiguous assignment in many cases.

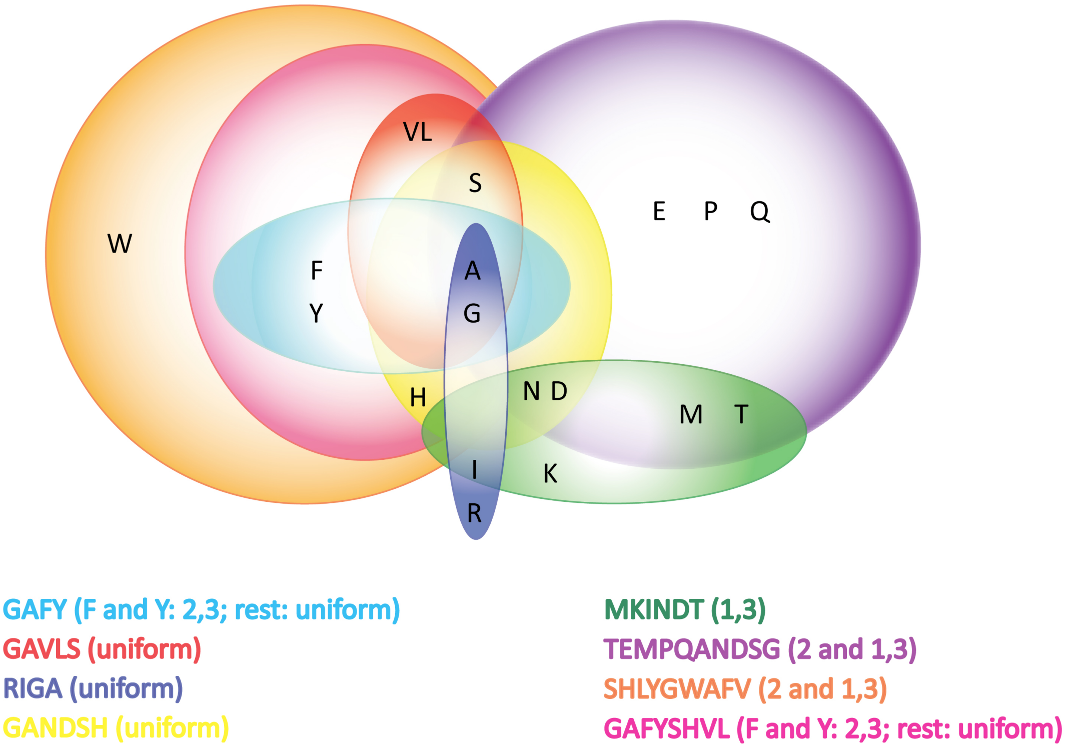


Figure 2.5: Venn-diagram illustrating the overlap between the different labeling schemes that were produced of OmpG. Every amino acid present in the OmpG sequence is at least labeled in one labeling scheme. Many of the residue types are present in multiple labeling schemes. This feature is important because the combination of labeling schemes can then be exploited to connect longer stretches in the protein sequence.

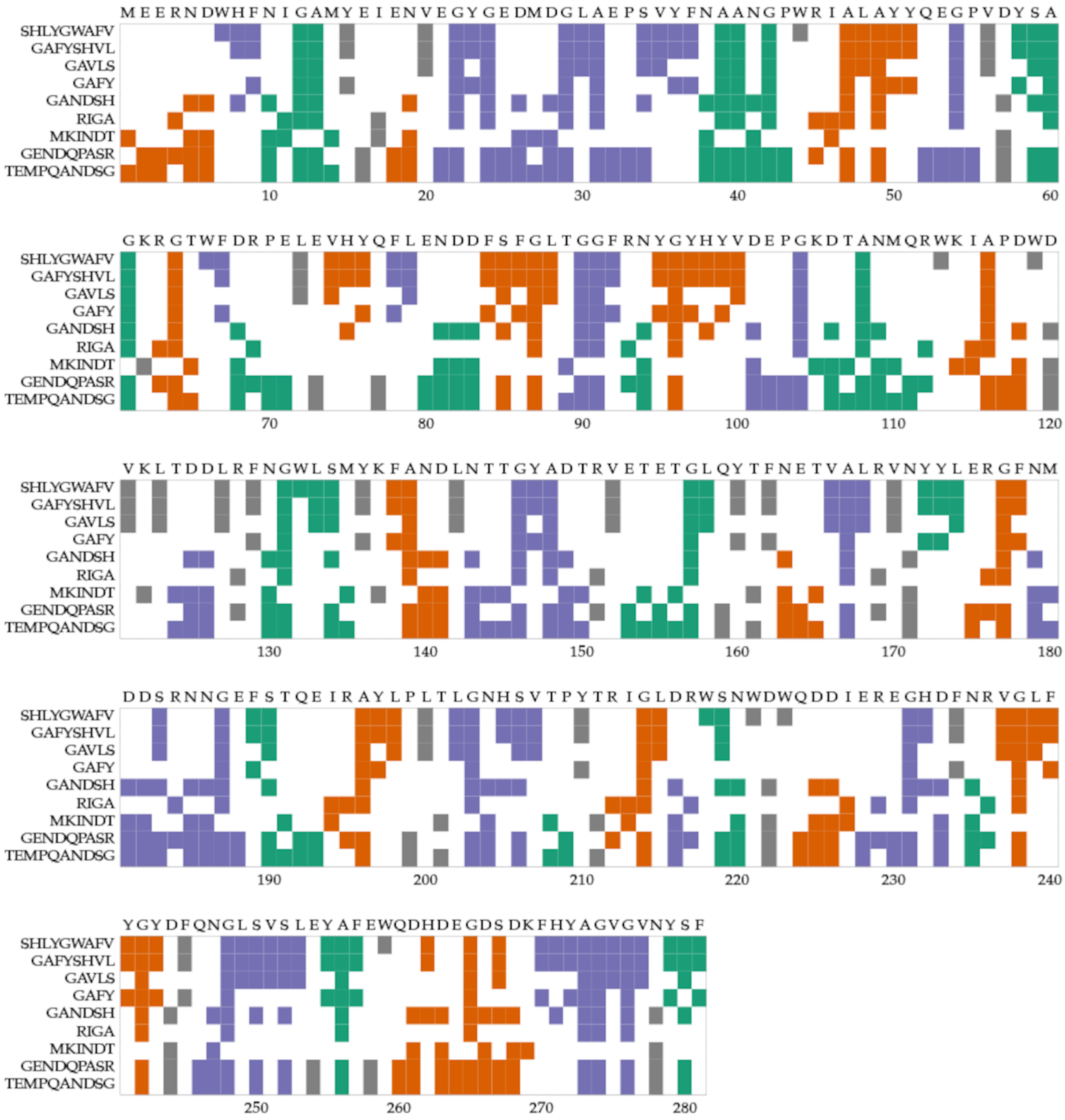


Figure 2.6: All amino acid selective labeling schemes used for the sequential assignment of OmpG depicted on the sequence. Highlighted rectangles indicate in which labeling schemes the residue is labeled. Colored (green, orange and purple) clusters of rectangles indicate that a sequential walk is possible without using the more crowded spectra of non-residue specific labeling schemes. Individual colors do not have any special meaning. A sequential walk is possible when two sequential residues are co-labeled in at least one labeling scheme. Grey rectangles indicate that the residue is not co-labeled with any of its two neighboring residues. The average cluster length is 3.0 and on average a given residue is part of a cluster of length 5.5.

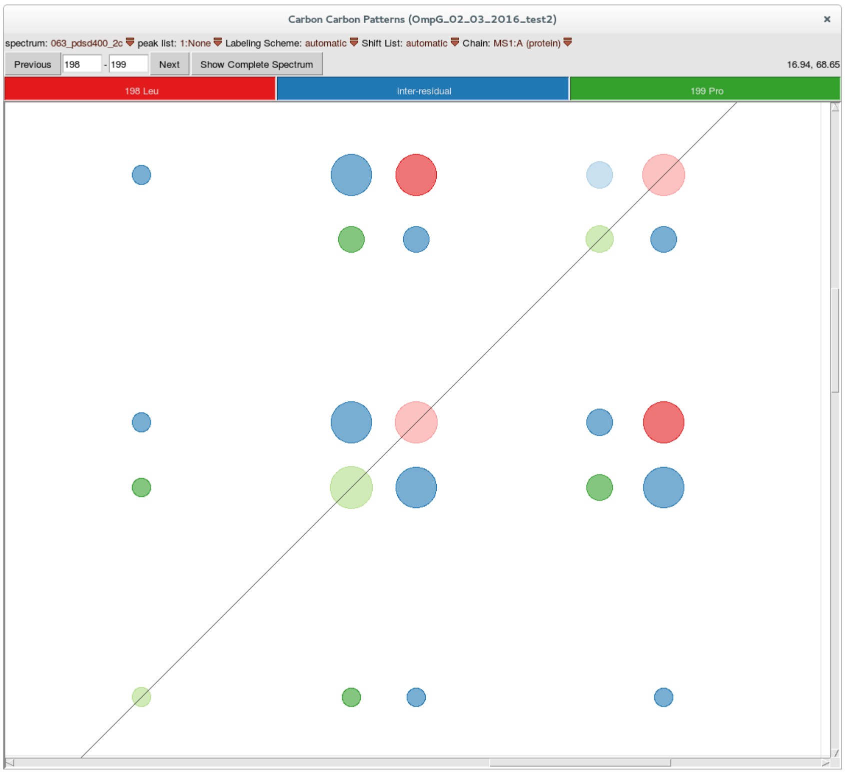
A CCPNMR Analysis plug-in for the visualization of cross peak patterns

When using various labeling schemes it can be very instrumental to visualize the cross-peak patterns that are expected, especially for glycerol based labeling schemes. Although it is possible to infer these patterns from the diagrams of the isotopomer schemes, it is more intuitive to work with the expected peak patterns as this directly combines the co-labeling of nuclei with the expected chemical shifts. The supporting material of the paper of Higman et al. 2009 contains visualizations of the expected correlation patterns for the 1,3- and 2-glycerol labeling schemes [7]. Inspired on these type of diagrams I wrote a plug-in for CCPNMR Analysis to automatically generate expected cross-peak pattern in 2D ^{13}C-^{13}C correlations for arbitrary labeling schemes, see figure 2.7. Integrating these kind of diagrams within the Analysis software has several advantages:

1. Not only intra-residual, but also expected inter-residual cross-peak patterns can be shown for any combination of two residues. Since this gives rise to about 400 combinations, it has a clear advantage to be able to do this “on screen”.
2. The location of expected peaks can be based on assigned chemical shifts, if present. If they are not present average values from the refDB are used, which is a carefully re-referenced subset of the bmrb [13]. To avoid confusion, when hovering over a peak in the diagram an indication is shown telling which of its dimensions are based on assigned chemical shifts and which are based on average shifts. Furthermore, the mouse position in the diagram is mirrored by the cross-hairs position in the spectra, so the actual peaks can be found.
3. The assignment status of peaks in a spectrum is indicated by dark/light coloring in the diagram. This is really helpful as it gives a quick overview of the completeness of the assignment of a peak pattern. It is hard to get this type of overview just by looking at the spectra or in peak tables.

The two selected residues do not have to be sequential and therefor this plug-in can in principle also be used to visualize and find long-range cross-peaks between any two residues. Also the expected peaks for the whole spectrum can be shown at once, which can be useful when considering which labeled samples to produce in the future.

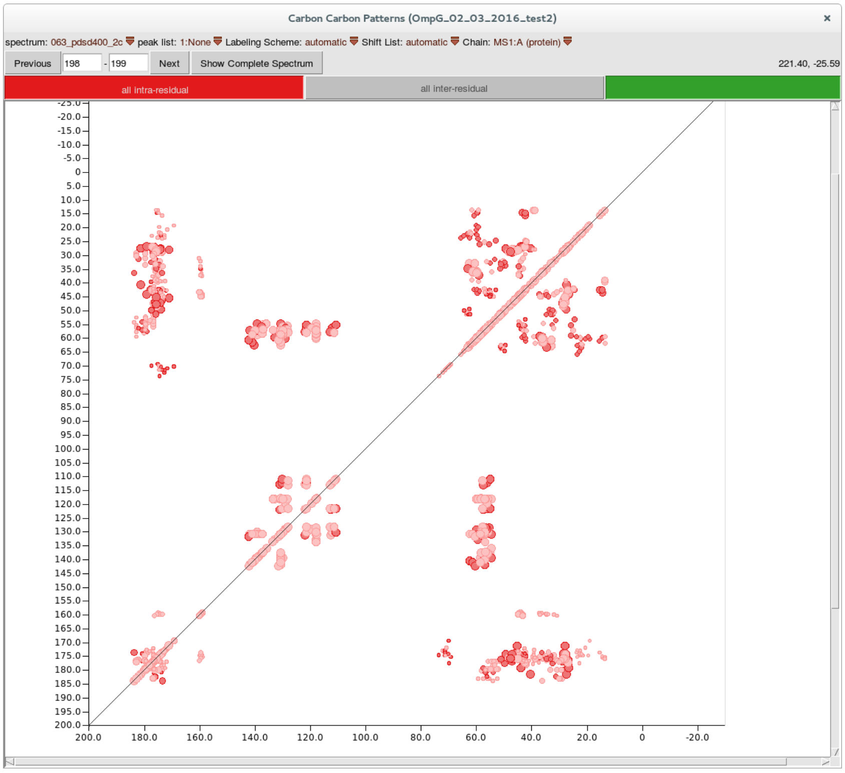
CCPNMR Analysis has very good support for configuring custom labeling schemes [14]. For each amino acid a set of isotopomers can be configured. Also a labeled sample can be created based on these labeling schemes. This labeled sample can in turn be connected to an experiment. All necessary information about the labeling schemes is directly taken from the project and therefor basically any scheme can be visualized using this plug-in. Also all other information, like residue sequences, chemical shifts and peak assignments are pulled directly from the project. It is straight-forward to open this plug-in in CCPNMR Analysis and no real installation is necessary. It can be downloaded from https://github.com/jorenretel/ccpnmr-cc-patterns, where more detailed instructions are given.


Carbon Carbon Patterns (OmpG_02_03_2016_test2)

spectrum: 063_pdsd400_2c peak list: 1:None Labeling Scheme: automatic Shift List: automatic Chain: MS1:A (protein)

198 - 199

16.94, 68.65


Carbon Carbon Patterns (OmpG_02_03_2016_test2)

spectrum: 063_pdsd400_2c peak list: 1:None Labeling Scheme: automatic Shift List: automatic Chain: MS1:A (protein)

198 - 199

221.40, -25.59

Figure 2.7: CCPNMR plug-in that helps visualizing expected sub-patterns in ^{13}C-^{13}C correlation spectra of labeled samples. The size of the circles represent the co-labeling of the two correlated nuclei. Top: expected intra-residual peak pattern for the residues leucine 198 (red) and proline 199 (green) and sequential cross-peaks (blue) in 2-glycerol labeled OmpG. Bottom: all expected intra-residual peaks in 2-glycerol labeled OmpG, for clarity in the figure, sequential peaks have been unselected.

Results and Discussion

Most of the labeled samples were already made by Matthias Hiller before I joined the project. The GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$SHVL and GENDQPASR samples were produced by Gregorio Guiseppe de Palma while I was already involved in this project. The TEMPQANDSG and SHLYGWAFV samples were expressed by Matthias Hiller and reconstituted into lipids by me. Additionally, an initial assignment was already made based on the spectra recorded on the GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$(S), GAVLS(W$_{\alpha,\beta,\gamma}$), RIGA(S), GANDSH(LV) and uniformly 2- and 1,3-glycerol labeled samples [6][7]. They are briefly reviewed in this chapter, because they form the basis for further assignments made in this work. All names of labeling schemes directly reflect the amino acids are labeled. Between brackets are amino acids that were unintentionally labeled due to metabolic scrambling. In table 2.1 all amino acid selective labeling schemes produced for the assignment of OmpG listed.

Table 2.1: Amino acid selective ^{13}C labeled samples produced for the assignment of OmpG. Amino acids between brackets were not intended to be labeled (in the case of the forward labeled schemes these amino acids were not added labeled to the growth medium. In the case of the reverse labeled schemes, these amino acids were added unlabeled to the growth medium). In all labeling schemes, all residues in the sequence are ^{15}N labeled.

labeling scheme	labeled residues	sequential pairs
forward labeled		
GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$(S)	94	33
GAVLS(W$_{\alpha,\beta,\gamma}$)	97	32
RIGA(S)	77	17
GANDSH(LV)	142	70
GENDQPASR	157	74
GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$SHVL(W$_{\alpha,\beta,\gamma}$)	144	76
reverse labeled		
2-TEMPQANDSG	162	84
1,3-TEMPQANDSG	162	84
2-SHLYGWAFV(QENDT)	238	201
1,3-SHLYGWAFV	144	76
1,3-MKINDT	82	23

Forward labeled schemes

All these samples contain labeled glycine and alanine. Furthermore, labeled serine is present in all samples, including GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$(S) and RIGA(S) for which serine was not added labeled to the feedstock, because serine is metabolically closely related to glycine. Alanine and serine Cα-Cβ peaks are well resolved in uniformly labeled OmpG and therefor obviously also in residue selective labeled samples.

GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$(S)

In this labeling scheme, the phenylalanine and tyrosine where ^{13}C-labeled only on the Cα and Cβ nuclei. This was done because their fast relaxing aromatic rings can act as a magnetization sink. By not labeling the aromatic ring, the phenylalanine and tyrosine Cα-Cβ peaks are higher in intensity and better defined than in spectra of uniformly labeled samples. This labeling strategy has been described in Hiller et al. 2008 [6].

GAVLS(W$_{\alpha,\beta,\gamma}$)

This labeling scheme yields complete cross peak patterns for the targeted amino acids. In addition, for tryptophan the Cα-Cβ peaks are visible along with correlations to the carbonyl region. However, no signals were present in the aromatic region. The labeling of tryptophan C′, Cα and Cβ and the backbone nitrogen atom arise from the last step of the tryptophan synthesis in which serine is used to build this part of the molecule. The additional tryptophan labeling turned out to be an advantage since the tryptophan Cα-Cβ peaks are seperated from the rest of the intra-residual peaks, while these peaks are not present in any of the other residue specific labeling schemes. The leucine Cα-Cβ peaks are well resolved, while they would overlap in a uniformly labeled sample with the Cα-Cβ peaks of aspartic acid and asparagine and partially with those of tyrosine and phenylalanine. The leucine Cα-Cγ peaks are freed from partial overlap with intra-residual peaks from glutamine/glutamic acid and lysine. The Valine Cα-Cγ1/Cγ2 resonances are not overlapped any longer with those of threonine Cα-Cγ. Also the Cα-Cβ peaks of valine are singled out in these spectra but this is a feature that is also present in the 2-glycerol labeled samples.

RIGA(S)

Complete cross-peak patterns can be observed for all labeled amino acids, although the isoleucine peaks are slightly lower in intensity than for instance the alanines. This might be due to an insufficient amount of labeled isoleucine to suppress the metabolism from threonine. In a uniformly labeled sample of OmpG, isoleucine Cα-Cβ peaks are almost completely covered by the Cα-Cβ peaks from phenylalanine and tyrosine. Furthermore, the intra-residual peak pattern of arginine is not overlapping with the intra-residual peaks from glutamine, glutamic acid, lysine and methionine in the RIGA(S) sample.

GANDSH(LV)

Complete cross-peak patterns for all labeled amino acids can be observed. However, the asparagine and aspartic acid peaks are very low in intensity. Additionally, the full intra-residual peak patterns of leucine and valine are present, although they are very low in intensity. The histidine cross-peaks are well resolved. The amount of useful inter-residual peaks in the GANDSH spectra with longer mixing times was lower than expected.

GENDQPASR

For this labeling schemes, the protocol of Tong et al. was followed to suppress isotope scrambling/dilution of Asn, Asp, Gln and Glu [15]. In this case the M9 medium was supplemented with inhibitors against the aspartate transaminase, aspartate ammonium lyase, ß-alanine-pyruvate-transaminase, glutamine synthase and with an excess of unlabelled amino acids. ^{13}C-^{13}C DARR spectra of this sample show the complete side chain signal patterns for all the amino acids that intended to be labeled. No additional isotope scrambling was found.

GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$SHVL(W$_{\alpha,\beta,\gamma}$)

This labeling scheme is basically a combination of the GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$(S) and GAVLS(W$_{\alpha,\beta,\gamma}$) schemes with the addition of histidine. All intra-residual peak patterns of the amino acids that were intended to be labeled are present. Additionally, small Trp Cα-Cβ peaks were observed as was also the case for the GAVLS(W$_{\alpha,\beta,\gamma}$) sample. The unique feature of this labeling scheme is that because the aromatic rings of phenylalanine, tyrosine and tryptophan are not labeled, the only signals in the aromatic region of the spectra are from histidines.

Residue selective 1,3- and 2-glycerol labeling schemes

2-TEMPQANDSG

The most remarkable feature of this labeling scheme is that in the spectral region between 42 and 49 ppm there are basically only chemical shifts from glycine Cα, with the exception of the Cβ shifts of aspartic acid and asparagine. However, in the latter two residues, the Cβ is only labeled in one third of the isotopomers and therefore give rise to less intense cross-peaks. Therefore, all large (inter-residual) cross peaks in this region correlate glycine with one of the other labeled residues, see figure 2.8. This is a great advantage since in general glycine is the most frequently occurring amino acid in membrane integrated β-barrels (27 glycines in OmpG) and the third most frequently occurring amino acid in membrane spanning α-helices [16]. Knowledge of which spin systems potentially neighbor a glycine facilitates assignment. In comparison, in the spectra of the uniform 2-glycerol sample this chemical shift region is overlapped by the shift of the leucine Cβ's and partially by the lysine Cε and arginine Cδ shifts.

Because there are only four clusters of intra-residual peaks in this spectrum belonging to proline Cα-Cδ, proline Cβ-Cδ, methionine Cα-Cγ and Threonine Cα-Cγ2, a lot of "spectral" space is left for inter-residual cross peaks. Furthermore, because many Cα resonances are removed compared to the uniformly 2-glycerol labeled sample, the Cα-Cα region of the spectrum is less crowded and easier to interpret.

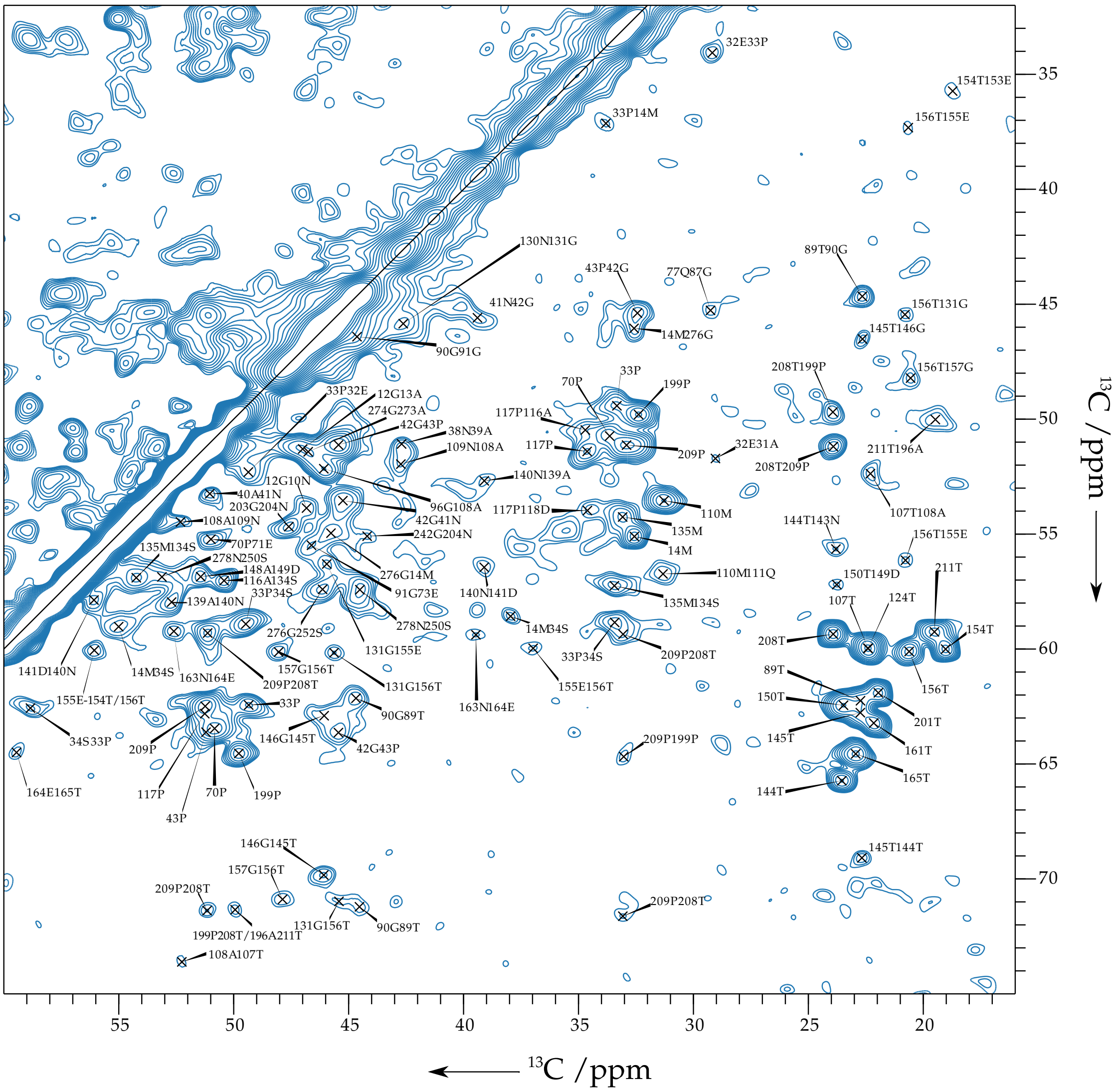


Figure 2.8: 2D ^{13}C-^{13}C correlation spectrum of 2-TEMPQANDSG labeled OmpG, with 400 ms DARR mixing.

1,3-TEMPQANDSG

Like the 2-TEMPQANDSG labeling scheme, 1,3-TEMPQANDSG does not reduce substantially the crowding of intra-residual peaks in comparison to its uniformly glycerol labeled counterpart. However, the inter-residual peaks are easier to interpret because of the reduced chemical shift degeneracy in the spectrum. For instance, in the spectra of both the uniformly 1,3-glycerol and the 1,3-TEMPQANDSG sample, a lot of well dispersed cross-peaks are found in the methyl region around 20 ppm. However, in the spectrum of the 1,3-TEMPQANDSG sample, the amount of possible resonances that could be correlated by these peaks is reduced. The same is true for the inter-residual cross-peaks on the serine and threonine Cβ frequencies. Also here, the cross-peaks are well dispersed in spectra of both the uniformly 1,3-glycerol and 1,3-TEMPQANDSG sample. Although the assignment of the first dimension of these cross peaks (corresponding to serine or threonine) would be rather straight-forward in both spectra, the number of assignment options for the second dimension is greatly reduced in spectra of the 1,3-TEMPQANDSG sample. In table 2.2 the spectral region up to 80 ppm is divided in 4 major regions and for each region the resonances of the amino acids that are removed and left over are indicated.

Table 2.2: Shifts being removed in the 1,3-TEMPQANDSG labeling scheme compared to a uniformly 1,3-glycerol labeled sample for different regions of the spectrum.

region	removed	left
+/- 20 ppm	LeuCδ, ValCγ, IleCγ/Cδ	AlaCβ, ThrCγ2, MetCε
25 - 35 ppm	LysCβ/Cγ, ArgCβ/Cγ, TrpCβ, IleCγ1, HisCβ	ProCβ/Cγ, MetCβ/Cγ, GluCβ/Cγ, GlnCβ/Cγ
35 - 45 ppm	PheCβ, TyrCβ, LysCε	AsnCβ, AspCβ
50 - 80 ppm	LysCα, LeuCα, IleCα, ArgCα	ThrCα/Cβ, MetCα, GluCα, GlnCα, AspCα, AsnCα, ProCα, SerCβ

2-SHLYGWAFV(QENDT)

In theory the only intra-residual peaks in this labeling scheme are those of leucine (Cβ-Cγ, C-Cβ and C-Cγ) and valine (Cα-Cβ). In practice, however, threonine (Cα-Cγ), glutamine/glutamic acid (Cα-Cδ, Cβ-Cδ) and possibly asparagine/aspartic acid (Cα-Cγ) peaks are present in this spectrum as well. This indicates that the metabolism of amino acids produced in the TCA cycle were insufficiently suppressed. However, we did not observe intra-residual peaks of amino acids like proline, methionine, isoleucine, lysine and arginine. This might be explained by the relative frequencies of the amino acids in the OmpG sequence. The amino acids for which peaks are present are roughly those that present in larger numbers in the OmpG sequence: threonine (15), glutamic acid (23), glutamine (8), asparagine (21) and apartic acid (27) versus proline (8), methionine (6), isoleucine (7), lysine (6) and arginine (16). Therefore, to fully suppress the metabolism towards the amino acids produced in the TCA cycle, higher amounts of unlabeled amino acids may be needed for those residue types that are more frequent in the protein sequence. This idea is further supported by comparing the differences in peak volume of the glutamic acid and threonine peaks in the 2-SHLYGWAFV spectra to those in the 2-TEMPQANDSG spectra. In the 2-SHLYGWAFV sample, the glutamic acid peak volumes are relatively higher than those of threonine, indicating that the unlabeled glutamic acid (23 amino acids in the sequence) was exhausted quicker than the unlabeled threonine (15 amino acids in the sequence).

Even though this scrambling was present, these spectra turned out to be useful. Many inter-residual peaks are present in the Cα-Cα region of the spectra, which assisted the sequential assignment and the generation of distance restraints for the structure calculation. Also, inter-residual cross-peaks in the aromatic region of the 2-SHLYGWAFV spectra recorded with longer mixing times were highly valuable since their ambiguity for assignment is greatly reduced in comparison to those in the same region of the uniform 2-glycerol spectra.

1,3-SHLYGWAFV

The spectra of this sample look like spectra of a partially unfolded protein (data not shown). Something might have gone wrong during the preparation of this sample. Therefore, this spectrum was not analyzed further.

1,3-MKINDT

Another sample produced using the reverse labeling strategy is 1,3-MKINDT. The expected peak patterns for the labeled amino acids is present. However, the signal sets for isoleucine and methionine are very weak. No isotope dilution to other amino acids was observed, though. The inter-residual peaks were not helpful for the sequential assignment. However, a useful feature of this labeling scheme is that the asparagine and aspartic acid Cα-Cβ peaks have a high intensity and are well-resolved. This is a very welcome feature since these signals have low intensity and partially overlap with the leucine Cα-Cβ peaks in the GANDSH(LV) spectra.

Conclusion

Using these labeling schemes 84 residues could be assigned with high certainty. In particular, the labeling schemes for which only a few amino acids are labeled, such as GAF$_{\alpha,\beta}$Y$_{\alpha,\beta}$(S), GAVLS(W$_{\alpha,\beta,\gamma}$) and RIGA(S) are very instrumental for finding starting points. From these starting points larger stretches can be found by using labeling schemes with more simultaneously labeled amino acids. In contrast, also the GANDSH(LV) and 1,3-MKINDT samples contain a relatively small number of labeled amino acids, but these samples turned out to be less helpful in terms of defining sequential connectivities as the three earlier mentioned samples. However, as will be discussed in chapter 4, ^{13}C-^{13}C spectra with short mixing time are used as reference spectra to find Cα-Cβ peaks of histidine (GANDSH(LV)) and asparagine/aspartic acid (1,3-MKINDT) spin systems.

The spectral quality of the uniformly 1,3- and 2-glycerol, 1,3- and 2-TEMPQANDSG and 2-SHLYGWAFV(QENDT) labeled samples was very high. Because larger numbers of simultaneously labeled amino acids were present, more peaks can be found in these spectra connecting spin systems. In particular, resolved correlations involving glycines in DARR spectra of the 2-TEMPQANDSG labeled sample with long mixing of 400 ms turned out to be exceptionally helpful. The 2-SHLYGWAFV(QENDT) sample contained valuable inter-residual peaks. However, this labeling scheme would have been of higher value if the complementary 1,3-SHLYGWAFV sample would have been of good quality and the isotope scrambling would have been suppressed better. Furthermore, many sequential correlations between residues labeled in 2-SHLYGWAFV(QENDT) could be assigned previously (before I joined the project) using the $\mathrm{GAF}_{\alpha,\beta}\mathrm{Y}_{\alpha,\beta}$(S) and $\mathrm{GAVLS}(\mathrm{W}_{\alpha,\beta,\gamma})$ samples.

In general, the robustness of the assignment strategy based on ^{13}C-detected experiments was severely hampered by the short ^{15}N T_2, which resulted in low ^{15}N dispersion in NCOCX and NCACX spectra. Therefore most assignments rely heavily on 2D ^{13}C-^{13}C correlations. Because of this difficulty it proved hard to assign a large enough part of the OmpG sequence to calculate a structure. As will be discussed in the next chapter, ^{1}H-detected experiments were crucial to extend the assignment. The ^{13}C-detected spectra of the different labeling schemes were used in conjunction with these newer ^{1}H-detected spectra to give access to side-chain chemical shifts. This will be the topic of chapter 4.

Materials and Methods

Materials

Isotopically labeled amino acids or labeled glucose, glycerol and labeled NH_4Cl were purchased from SIGMA-ALDRICH and Cambridge Isotopes Laboratories Inc., respectively. Dodecyl-β-D-maltoside was purchased from Glycon and E.coli total lipid extract from Avanti Polar lipids.

Protein expression and purification

Forward labeling strategy

Cell mass is predominately grown on 4 litres of unlabeled rich medium allowing rapid growth to high cell densities. Upon reaching optical cell-densities of ~0.5-0.7 (measured at 600 nm), cells were pelleted by centrifugation. The cells were then washed and pelleted using a 1 x M9 salt solution to exclude all nitrogen and carbon sources. The cell pellet was re-suspended in 1 L isotopically labelled M9 minimal medium containing 100mg of labeled and unlabelled amino acids, 2 g of unlabeled glucose and 0.5 g of ^{15}N-NH$_4$Cl. After one hour of incubation, protein expression was induced by the addition of isopropylthio-β-D-galactoside (1 mM IPTG). Cells were harvested after 3 hours by centrifugation.

Reverse labeling strategy

The expression protocol is nearly the same as described above with the following exception: after washing and pelleting, the cell pellet was re-suspended in 1 L isotopically labeled M9 minimal medium containing 50 mg of ^{15}N-labelled amino acids, 2 g of 1.3- or 2-^{13}C labeled glycerol and 0.5 g of ^{15}N-NH$_4$Cl. Protein purification, refolding and 2D crystallization were carried out as described previously [17].

NMR experiments

2D ^{13}C-^{13}C DARR spectra were recorded on a Bruker narrow-bore 900 MHz spectrometer equipped with a 3.2 mm triple-resonance probe (Bruker, Karlsruhe, Germany). The MAS-frequency was set to 13 kHz and the sample temperature was set to 280 K. The typical $\pi/2$-pulses were 3-3.5 μs for ^{1}H and 5 μs for ^{13}C. ^{1}H/^{13}C cross-polarization (CP) contact time was 1.5 ms, with a constant radio-frequency (r.f.) field of 58.5 kHz on proton and with a carbon lock-field ramped linearly around the Hartmann-Hahn n=1 matching condition (50% ramp, optimized experimentally). SPINAL64 decoupling with a power level of 90kHz was used during indirect and direct chemical shift evolutions. DARR mixing with durations of 20 ms, 200 ms and 400 ms were used for the forward labeled OmpG-samples. 50 ms, 200 ms and 400 ms were used for reverse-labelled OmpG-samples. The carrier frequency was placed at 100 ppm. The time domain data matrix of each experiment was 512 (t1) × 2048 (t2) points, with t1 and t2 increments of 10 μs and 16 μs, respectively. 96 or 160 scans per point were recorded with a recycle delay of 3 s, resulting in total acquisition times of ~ 42 or 68 hours, respectively. Data were processed with shifted Sinebell (t1) and Lorentzian-to-Gaussian (t2) apodization functions and zero filled to 4096 (t1) × 8192 (t2) points using Topspin version 2.1 (Bruker, Karlsruhe, Germany). The carbon chemical shifts were indirectly referenced to 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS) through the ^{13}C adamantane downfield peak resonating at 40.48 ppm.

3D NCACX and NCOCX spectra were recorded on a Bruker 400 MHz wide bore spectrometer (Bruker, Karlsruhe, Germany). The MAS-frequency was set to ν_R 8 kHz and the sample temperature was set to 280 K. Typical $\pi/2$-pulses were 3-3.5 μs for ^{1}H, 5 μs for ^{13}C and 7 μs ^{15}N. ^{1}H/^{15}N cross-polarization (CP) contact time was 1.5 ms, with a constant rf-field of 55 kHz on proton and with a nitrogen lock-field ramped linearly around the Hartmann-Hahn n=1 matching condition (70% ramp, optimized experimentally). The ^{15}N carrier frequency was set to 120 ppm. Following the evolution of nitrogen, adiabatic CP was employed to selectively transfer magnetization from ^{15}N to Cα or CO. For NCA-type experiments the ^{13}C carrier-frequency was placed at 55ppm. The rf-field strengths N-Cα transfer were optimized around 3/2 ω_R (for Cα) and 5/2 ω_R (for nitrogen). For NCO-type experiments the ^{13}C carrier-frequency was placed at 170 ppm. The rf-field strengths during the N-CO transfer were optimized around 7/2 ω_R (for CO) and 5/2 ω_R (for nitrogen). In both cases, the N-C CP- contact time was optimized between 3 and 5 ms. For ^{13}C-^{13}C mixing, DARR irradiation was used with a duration of 20, 50, 100, 200 and 400 ms, depending on the labeling scheme. During all evolution periods, proton decoupling was applied, using SPINAL64 (90 kHz) [18]. The 3D data sets were recorded using evolution times of 6.8 and 6.4 ms in t1 and t2, respectively. Each FID was averaged from 96 scans yielding a total measurement time of ~4 ½ days per spectrum.

References

[1] W. T. Franks, K. D. Kloepper, B. J. Wylie, and C. M. Rienstra. "Four-Dimensional Heteronuclear Correlation Experiments for Chemical Shift Assignment of Solid Proteins". *Journal of Biomolecular NMR* 39.2 (Aug. 2007), pp. 107–131. DOI: 10.1007/s10858-007-9179-1.

[2] A. Schuetz, C. Wasmer, B. Habenstein, R. Verel, J. Greenwald, R. Riek, A. Böckmann, and B. H. Meier. "Protocols for the Sequential Solid-State NMR Spectroscopic Assignment of a Uniformly Labeled 25 kDa Protein: HET-s(1-227)". *ChemBioChem* 11.11 (July 2010), pp. 1543–1551. DOI: 10.1002/cbic.201000124.

[3] L. J. Sperling, D. A. Berthold, T. L. Sasser, V. Jeisy-Scott, and C. M. Rienstra. "Assignment Strategies for Large Proteins by Magic-Angle Spinning NMR: The 21-kDa Disulfide-Bond-Forming Enzyme DsbA". *Journal of Molecular Biology* 399.2 (June 2010), pp. 268–282. DOI: 10.1016/j.jmb.2010.04.012.

[4] C. Shi, H. K. Fasshuber, V. Chevelkov, S. Xiang, B. Habenstein, S. K. Vasa, S. Becker, and A. Lange. "BSH-CP Based 3D Solid-State NMR Experiments for Protein Resonance Assignment". *Journal of Biomolecular NMR* 59.1 (Mar. 2014), pp. 15–22. DOI: 10.1007/s10858-014-9820-8.

[5] S. Jehle, M. Hiller, K. Rehbein, A. Diehl, H. Oschkinat, and B.-J. van Rossum. "Spectral Editing: Selection of Methyl Groups in Multidimensional Solid-State Magic-Angle Spinning NMR". *Journal of Biomolecular NMR* 36.3 (Sept. 2006), pp. 169–177. DOI: 10.1007/s10858-006-9078-x.

[6] M. Hiller, V. A. Higman, S. Jehle, B.-J. van Rossum, W. Kühlbrandt, and H. Oschkinat. "[2,3-13C]-Labeling of Aromatic ResiduesGetting a Head Start in the Magic-Angle-Spinning NMR Assignment of Membrane Proteins". *Journal of the American Chemical Society* 130.2 (Jan. 2008), pp. 408–409. DOI: 10.1021/ja077589n.

[7] V. A. Higman, J. Flinders, M. Hiller, S. Jehle, S. Markovic, S. Fiedler, B.-J. van Rossum, and H. Oschkinat. "Assigning Large Proteins in the Solid State: A MAS NMR Resonance Assignment Strategy Using Selectively and Extensively 13C-Labelled Proteins". *Journal of Biomolecular NMR* 44.4 (July 2009), pp. 245–260. DOI: 10.1007/s10858-009-9338-7.

[8] D. M. LeMaster and D. M. Kushlan. "Dynamical Mapping of E. Coli Thioredoxin via 13C NMR Relaxation Analysis". *Journal of the American Chemical Society* 118.39 (Jan. 1996), pp. 9255–9264. DOI: 10.1021/ja960877r.

[9] M. Hong. "Determination of Multiple φ-Torsion Angles in Proteins by Selective and Extensive 13C Labeling and Two-Dimensional Solid-State NMR". *Journal of Magnetic Resonance* 139.2 (Aug. 1999), pp. 389–401. DOI: 10.1006/jmre.1999.1805.

[10] F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein, and H. Oschkinat. "Structure of a Protein Determined by Solid-State Magic-Angle-Spinning NMR Spectroscopy". *Nature* 420.6911 (Nov. 2002), pp. 98–102. DOI: 10.1038/nature01070.

[11] S. Jehle, P. Rajagopal, B. Bardiaux, S. Markovic, R. Kühne, J. R. Stout, V. A. Higman, R. E. Klevit, B.-J. van Rossum, and H. Oschkinat. “Solid-State NMR and SAXS Studies Provide a Structural Basis for the Activation of αB-Crystallin Oligomers”. *Nature Structural & Molecular Biology* 17.99 (Sept. 2010), pp. 1037–1042. doi: 10.1038/nsmb.1891.

[12] M. Hong and K. Jakes. “Selective and Extensive 13C Labeling of a Membrane Protein for Solid-State NMR Investigations”. *Journal of Biomolecular NMR* 14.1 (May 1999), pp. 71–74. doi: 10.1023/A:1008334930603.

[13] H. Zhang, S. Neal, and D. S. Wishart. “RefDB: A Database of Uniformly Referenced Protein Chemical Shifts”. *Journal of Biomolecular NMR* 25.3 (Mar. 2003), pp. 173–195. doi: 10.1023/A:1022836027055.

[14] T. J. Stevens, R. H. Fogh, W. Boucher, V. A. Higman, F. Eisenmenger, B. Bardiaux, B.-J. van Rossum, H. Oschkinat, and E. D. Laue. “A Software Framework for Analysing Solid-State MAS NMR Data”. *Journal of Biomolecular NMR* 51.4 (Sept. 2011), pp. 437–447. doi: 10.1007/s10858-011-9569-2.

[15] K. I. Tong, M. Yamamoto, and T. Tanaka. “A Simple Method for Amino Acid Selective Isotope Labeling of Recombinant Proteins in E. Coli”. *Journal of Biomolecular NMR* 42.1 (Sept. 2008), pp. 59–67. doi: 10.1007/s10858-008-9264-0.

[16] M. B. Ulmschneider and M. S. P. Sansom. “Amino Acid Distributions in Integral Membrane Protein Structures”. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1512.1 (May 2001), pp. 1–14. doi: 10.1016/S0005-2736(01)00299-1.

[17] M. Hiller, L. Krabben, K. R. Vinothkumar, F. Castellani, B.-J. van Rossum, W. Kühlbrandt, and H. Oschkinat. “Solid-State Magic-Angle Spinning NMR of Outer-Membrane Protein G from Escherichia Coli”. *ChemBioChem* 6.9 (Sept. 2005), pp. 1679–1684. doi: 10.1002/cbic.200500132.

[18] B. M. Fung, A. K. Khitrin, and K. Ermolaev. “An Improved Broadband Decoupling Sequence for Liquid Crystals and Solids”. *Journal of Magnetic Resonance* 142.1 (Jan. 2000), pp. 97–101. doi: 10.1006/jmre.1999.1896.

Chapter 3

Assignment Using ^{1}H-detected Experiments

Introduction

There are good reasons to detect ^{1}H as opposed to ^{13}C, which was the most common detection method in solid-state NMR until recently. Since the gyromagnetic ratio of ^{1}H is 8 and 31 times than those of ^{13}C and ^{15}N respectively, the signal to noise in proton detected spectra is higher. In addition, protons add an additional observable nucleus, which provides an independent dimension to multi-dimensional spectra, increasing resolution and making assignment strategies more robust. Because the application of solid-state NMR to biological samples is always sensitivity limited, proton detection was badly needed to make larger systems accessible. The major challenge with the inclusion of protons is their large linewidth caused by the strong ^{1}H-^{1}H dipolar couplings (see table 1.3). There are two main strategies that can be employed to reduce the ^{1}H linewidths:

1. Reduction of the amount of ^{1}H in the sample by perdeuterating the protein and subsequently reintroduction of small amounts of protons at the exchangeable sites, see figure 3.1.
2. Spinning of small diameter rotors (<2 mm) at higher MAS frequencies (>40 kHz).

Early proton-detected experiments were performed on dipeptides and small proteins using the first strategy [1][2][3][4]. It was shown that by drastically reducing the amount of back-exchanged protons from 100% to 10% ^{15}N-^{1}H correlation spectra of the small protein SH3 could be acquired with high resolution [5]. Experiments employing proton detection enabled the assignment of backbone ^{1}H, ^{15}N and ^{13}C resonances which, in combination with restraints obtained from additionally protonated methyl groups, yielded the structure of SH3 at low back-exchange levels (10 and 25% respectively) and moderate MAS rates [6][7]. The optimal amount of reprotonation at 24kHz MAS was found to be around 30% [8]. A first ^{15}N-^{1}H correlation of OmpG was measured at this back-exchange level and 20kHz MAS [9].

Recently, probes spinning small diameter rotors at faster MAS rates (40-60kHz) have become commercially available. At these high spinning rates, high-resolution spectra can be recorded with full reprotonation of the exchangeable sites of otherwise perdeuterated proteins [10][11]. Although the sample volume in these rotors is smaller, loss of signal to noise is balanced by several factors. First of all, the higher content of exchanged protons increases the number of protons available for detection. Second, the filling factor of the coils for smaller rotors plays a role. For instance, ^{15}N-^{1}H correlations of SH3 measured in 1.9 mm rotors spinning at 40 kHz with 60% back-exchanged protons and in 1.3 mm rotors spinning at 60 kHz with 100% back-exchanged protons are very similar both in signal to noise and resolution. However, for 3D experiments that contain a scalar transfer step, the signal to noise ratio in the 1.3 mm rotors is higher [12]. At ultra-fast spinning conditions low power hetero-nuclear decoupling can be used, which reduces strain on the instrumentation and sample heating [13][14][15]. Furthermore, ^{1}H-^{13}C cross-polarization conditions become available that selectively transfer magnetization to either the carbonyl or aliphatic carbons [16].

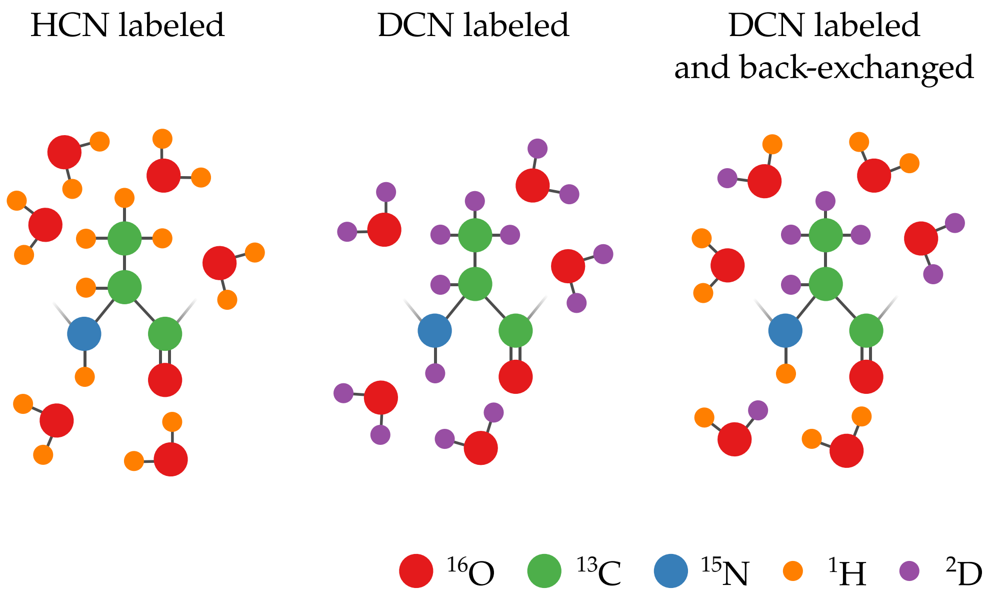


Figure 3.1: Left: fully ^{1}H/^{13}C/^{15}N labeled protein as used in combination with the ^{13}C-detected experiments in the previous chapter. Middle: protein expressed using perdeuterated (^{2}H/^{13}C/^{15}N) medium followed by back-exchange procedure that reintroduces protons. Right: the perdeuterated protein from the figure in the middle after the back-exchange procedure, as used in ^{1}H-detected experiments. In the case of the OmpG samples, the back-exchange procedure is performed before refolding. By choosing the H_2O:D_2O ratio in the buffer a defined proton content can be obtained at the exchangeable sites.

Various groups have introduced experiments employing proton detection to acquire sequential assignments and have applied them successfully to a number of different systems: micro-crystalline (SH3, GB1, Human Superoxide Dismutase, DsbA and β2m), α-synuclein fibrils, sedimented viral capsids, a secretion needle from salmonella and membrane proteins (the conductance domain from influenza A M2, and DsbB) [17][18][19][20][21]. Also structures have been calculated using ^{1}H-^{1}H distances acquired at fast MAS, for example GB1 and Human Superoxide Dismutase [17][18][22].

For the assignment of OmpG, two sets of three types of experiments were recorded, as illustrated in figures 3.2 and 3.3, at 60 kHz MAS. These are the pulse sequences presented in the papers of Barbet-Massin et al. in 2013 and 2014 [23][19]. The first set, consisting of the hCANH, hCOcaNH and hcaCBcaNH experiments, correlates each ^{1}H-^{15}N pair in the backbone to the chemical shift of the Cα, CO and Cβ respectively, within the same residue. The second set correlates the ^{1}H-^{15}N pairs to the Cα, CO and Cβ frequencies of the preceding residue. To achieve the sequential assignments, strips are generated corresponding to the chemical shifts of one ^{1}H-^{15}N pair. By finding two strips in which the ^{13}C chemical shifts from the first set of spectra match the ^{13}C chemical shifts of the second set, a sequential connection can be established. When several strips are placed in order, a match to a part of the protein sequence can be found, based on the possible amino acid types that are allowed for the combinations of Cα and Cβ chemical shifts.

The hcoCAcoNH, hCOcaNH, hcaCBcaNH and hcaCBcacoNH pulse sequences make use of scalar-coupling based transfer steps to transfer magnetization between the carbons. To evolve the scalar coupling, relatively long delays τ of $1/(4J_{CC})$ are needed. This equals 4.5 ms for a Cα-CO and 7.1 ms for a Cβ-Cα transfer (with $J_{C\alpha CO}$=55Hz and $J_{C\alpha C\beta}$=35Hz). For the transfer of magnetization between the CO and Cα in the hcoCAcoNH and the Cα and Cβ in the hcaCBcaNH and hcaCBcacoNH experiments out-and-back schemes are used [23]. This means that instead of using a CP transfer directly from proton to the carbon measured in one of the indirect dimensions (which would be the case in hCAcoNH, hCBcaNH and hCBcacoNH experiments), the magnetization is transferred first to the neighboring carbon on the magnetization transfer pathway. Both the out-and-back and "normal" variant of the experiment contain 4 delays τ. The advantage of the out and back scheme is that the transverse magnetization during these delays is on the slower relaxing nucleus (Cα in the case of a Cα-Cβ transfer and CO in the case of a Cα-CO transfer) instead of on each nucleus for 2 out of 4 delays. Only just before the acquisition of the carbon dimension, the spin of the faster relaxing nucleus is brought into the transverse plane. Because of the inhomogeneous nature of the sample ^{13}C T$_2$ times are shorter in OmpG than in more structured proteins. The bulk T$_2$ time for Cα was measured to be around 8 ms. The CO T$_2$ was not measured but is likely to be a factor of 3-4 larger. Therefore this experimental scheme dramatically increased the sensitivity of these experiments.

This set of spectra is conceptually very similar to the basic set of spectra used for the assignment of solution NMR data although there are a few differences. Because in solution NMR the ^{15}N-^{13}C transfer is achieved by INEPT instead of CP and the N-Cα_i and N-Cα_{i-1} scalar couplings are similar in size, the HNCA experiment in solution NMR normally includes both the Cα_i and Cα_{i-1} peak. In

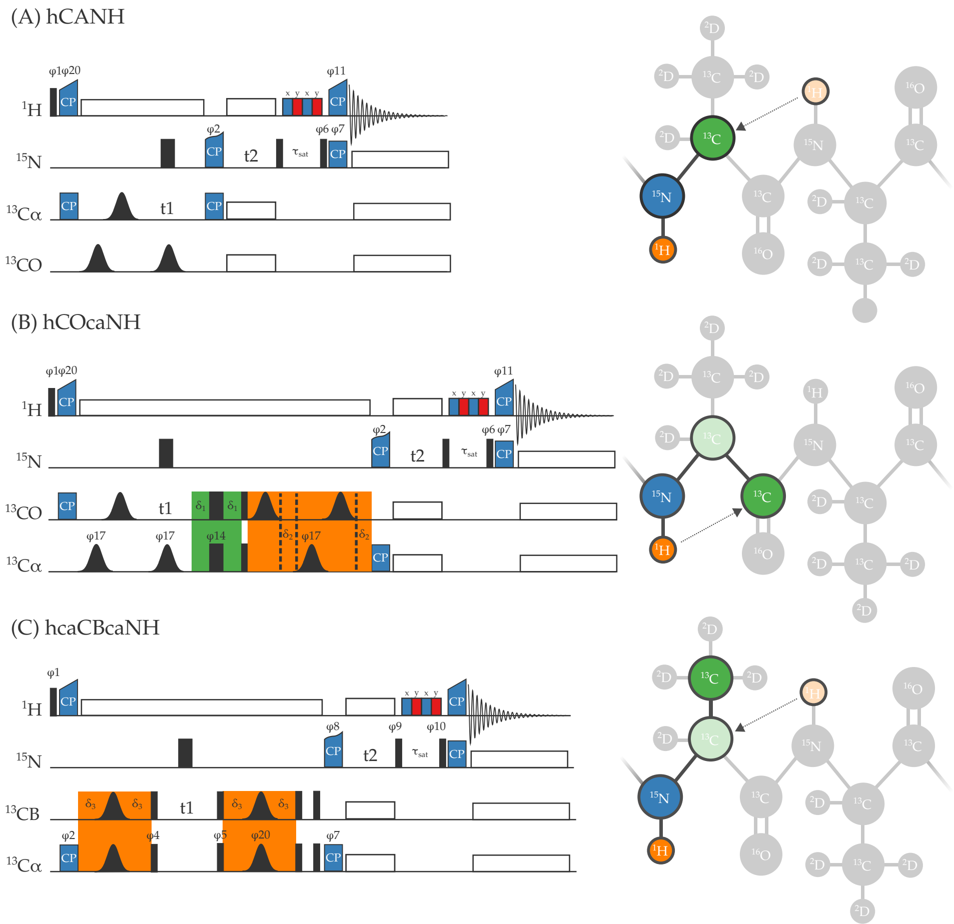
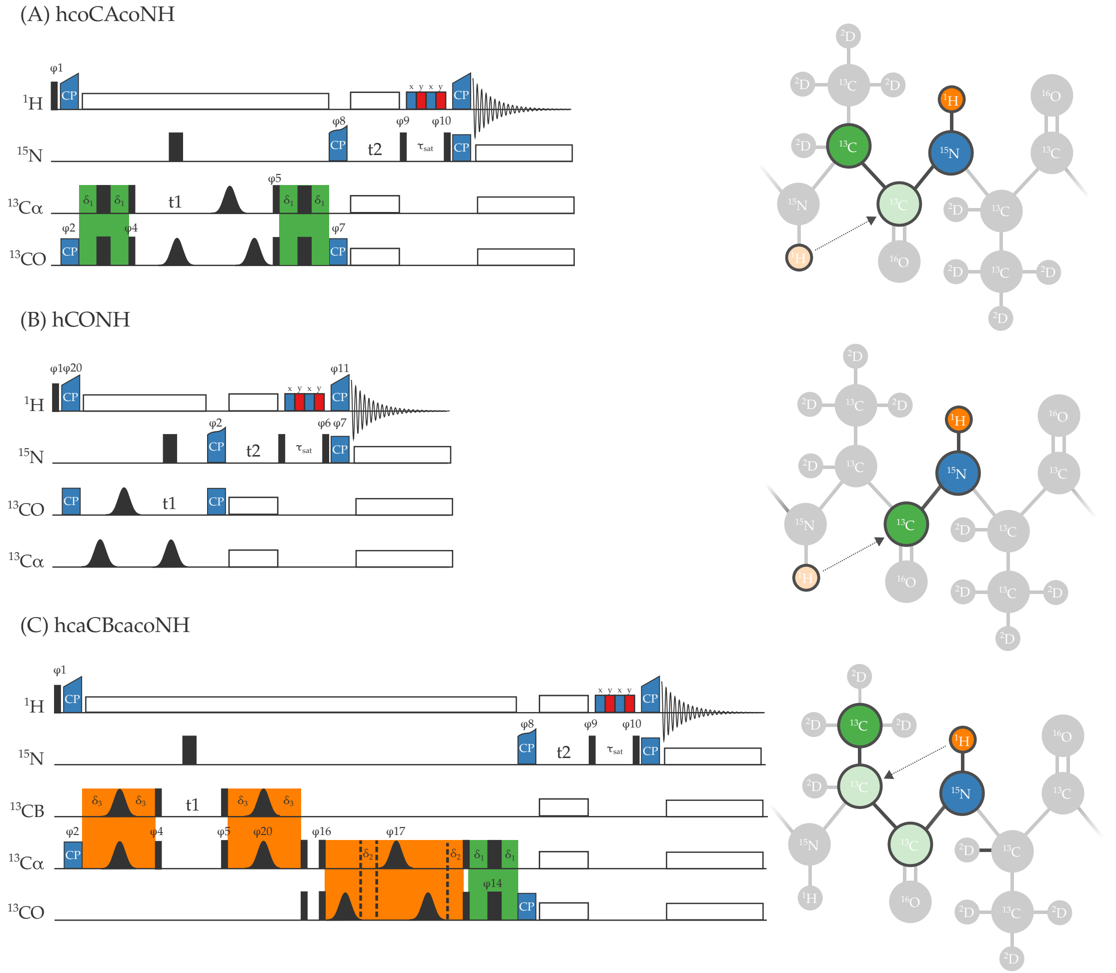


Figure 3.2: Pulse sequences (left) for ^{1}H-detected experiments correlating the amide ^{15}N and ^{1}H with intra-residual carbons together with a schematic depiction of the magnetization transfer pathway. In the pulse sequences, narrow black rectangles indicate 90°-pulses. Wider black rectangles indicate 180°-pulses. Blue shapes, marked CP depict cross-polarization steps. Gaussian shapes indicate band selective 180° pulses. White rectangles depict decoupling. The blue/red striped rectangles indicate the MISSISSIPPI water suppression sequence. The green and orange areas indicate scalar based ^{13}C-^{13}C transfers, where green indicated that the transverse magnetization is on the CO and orange means the transverse magnetization in on Cα. In the schematic depictions of the magnetization transfer pathway, dark colors (dark green for ^{13}C, blue for ^{15}N and dark orange for ^{1}H) show the measured nuclei. Light green indicates a ^{13}C that is present on the magnetization transfer pathway but not measured. The dotted arrow shows the first ^{1}H-^{13}C cross-polarization step. Light orange indicates the first excited proton, in the case this proton is not the detected proton. Pulse sequences by Barbet-Massin et al. [23][19]



(A) hcoCAcoNH

(B) hCONH

(C) hcaCBcacoNH

Figure 3.3: Pulse sequences (left) for ^{1}H-detected experiments correlating the amide ^{15}N and ^{1}H with inter-residual carbons together with a schematic depiction of the magnetization transfer pathway. Shapes and colors have the same meaning as in figure 3.2. Pulse sequences by Barbet-Massin et al. [23][19]

the solid-state version of the experiment, the hCANH only includes the Cα_i peak, which is advantageous since this reduces signal overlap. Also in solution NMR, the Cα-Cβ scalar transfer is generally evolved only half-way to create an HNCA/CB experiment were the Cβ peaks are negative.

In comparison to ^{13}C-detected experiments, the assignment strategy is enormously simplified. As discussed before, in the 3-dimensional ^{13}C detected experiments NCACX and NCOCX, the pivot along which a strip representing one spin system is connected to its sequential neighbor is the backbone ^{15}N chemical shift. In the set of ^{1}H-detected experiments, this pivot is dispersed by the chemical shift of its directly bound proton. Therefore, it is in most cases clear which peaks from the 6 experiments belong to one ^{1}H-^{15}N combination. As a results, before any strips have been matched, the possible amino acid types of two sequential spin systems can be deduced. In the ^{13}C-detected 3D experiments at least two strips have to be matched to do the same. When considering a strip in the NCACX one can not tell just on the basis of the ^{15}N chemical shift which NCOCX strip is connecting to its N-terminal neighbor. The two strips have to match based on corresponding peaks in both strips. Only after this match has been done, the possible residue types of two sequential spin systems can be deduced. Because of the reduced overlap in the ^{1}H-detected strips, it is often possible for multiple strips to be matched with a relatively high degree of certainty before the stretch of connected spin systems has been matched to a specific sequence of the protein.

Results and discussion

There is still a lot of overlap in the ^{1}H-^{15}N correlation relative to a similar spectrum in solution NMR (see figure 3.4). In the 3-dimensional spectra, it is easier to distinguish individual ^{1}H-^{15}N pairs. This is because most peaks do not overlap in the ^{13}C dimension, thereby making it possible to see the exact peak maxima. Both H-CO and H_Cα CP conditions were 87 kHz

Using the hCANH, hcoCAcoNH, hCONH, hCOcaNH, hcaCBcaNH and hcaCBcacoNH spectra, 151 strips could be assigned corresponding to 55% of the sequence when prolines are excluded, see figure 4.2. In addition, some of the ^{13}C chemical shifts of another 16 residues, including 6 prolines (prolines do not have an amide proton and therefore do not give rise to a strip in the hCANH, hCOcaNH and hcaCBcaNH spectra), could be determined based on peaks in the hcoCAcoNH, hCONH and hcaCBcacoNH spectra. Interestingly only a few strips were left unassigned and therefore the signal set for a large part of the sequence seems to be missing. Because the ^{1}H-^{15}N-correlation is too crowded, the ^{13}C-^{15}N-projection of the hCANH, shown in figure 3.5, was used as a reference spectrum. Although some of the peaks in this projection still correspond to multiple peaks in the ^{1}H-dimension, it is clear that the vast majority of peaks present are assigned.

In figure 3.6 a representative stretch of strips is shown. Note that towards the end of this stretch, which is also the end of the assignment for this part of the sequence, peak intensities decrease. This is especially the case for the peaks in the Cβ-correlated spectra. The decline in peak intensity towards

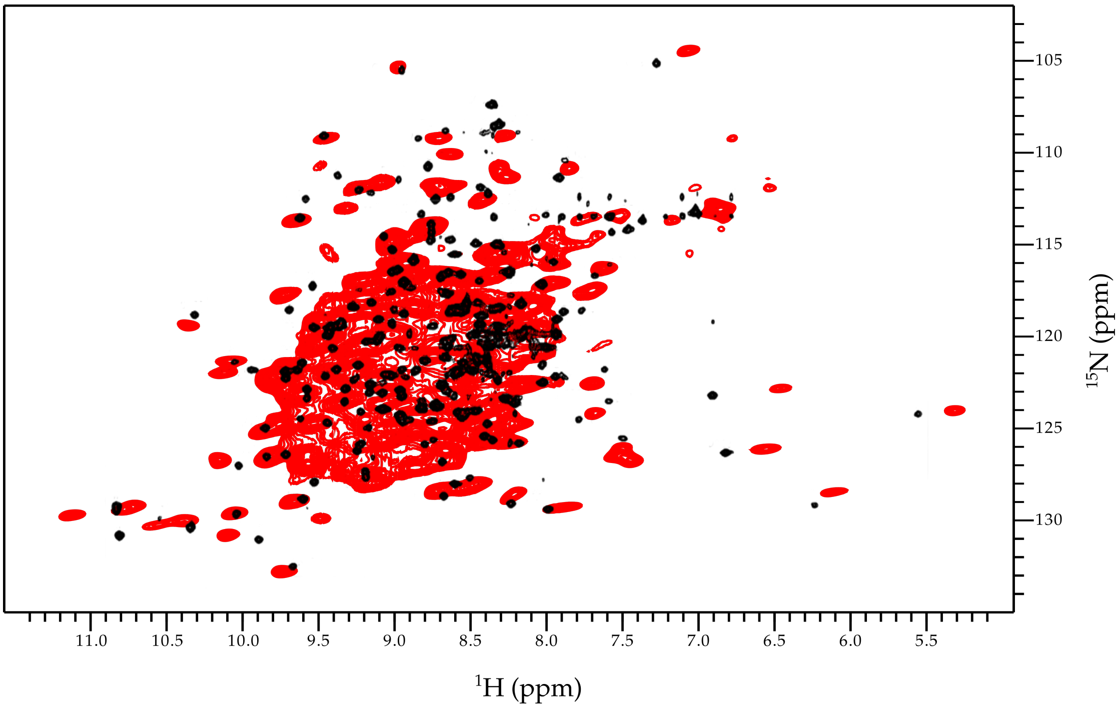


Figure 3.4: Overlay of an HN correlation spectrum acquired using solid state NMR (red), and solution NMR TROSY spectrum on OmpG in detergent micelles (black). The solution spectrum is a modified copy of the second figure in the paper of Lukas K. Tamm and coworkers describing the solution structure of OmpG [24]. Besides the difference in linewidth between the two spectra, there are also peaks present in the solution spectrum that are absent in the solid state spectrum. These peaks correspond mostly to the flexible loops on the extra-cellular side of OmpG and some to the shorter turns on the intra-cellular side.

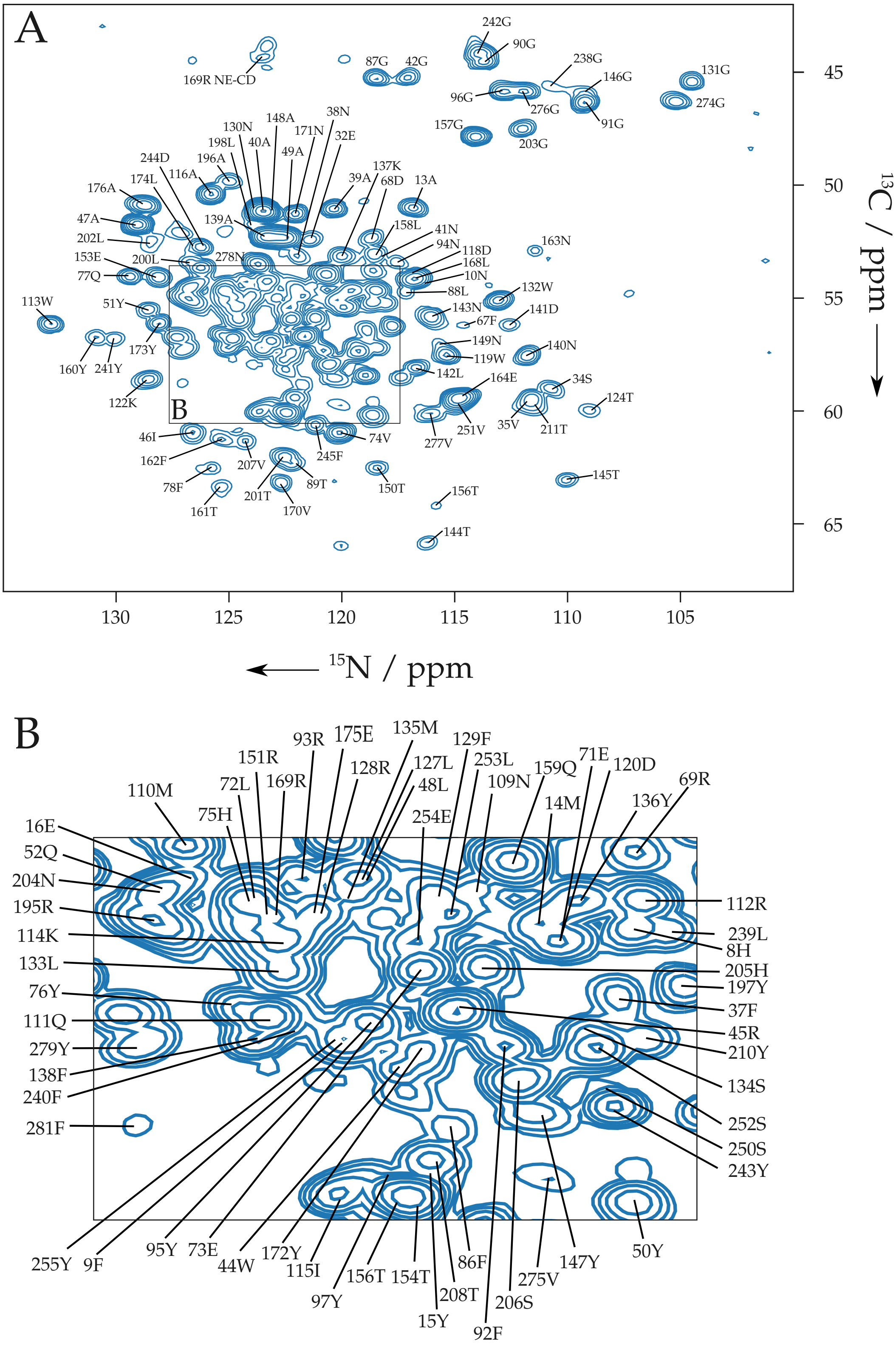


Figure 3.5: Assignment shown on the CN projection of the hCANH spectrum.

the end of assigned stretches seems to be a trend over the whole sequence, as the lack of sufficient cross peaks is the direct cause for the assignments stopping. In figure 3.6 signal intensities of the hCANH, hcoCAcaNH, hcaCBcaNH and hcaCBcacoNH spectra are shown for the assigned residues in the sequence. As can be seen, there are very large differences in the peak intensities in the Cβ-correlated spectra, whereas the Cα correlated spectra are more consistent. The residue numbers in the plot correspond to the residue of the carbon that is measured. It can be observed that there is a strong correlation between the signal intensities of the hcaCBcaNH and the hcaCBcacoNH. This indicates that it is not the efficiency of the CP steps that is determining the signal intensity, but the ^{13}C-relaxation during the delays τ that allow the scalar coupling to evolve (which are much longer than the acquisition time). As explained in the introduction, the transverse magnetization during these delays is on the Cα in the case of the hcaCBcaNH and hcaCBcacoNH (for the last experiment at least in all the delays for the Cα-Cβ transfer and half of the time during the Cα-CO transfer). In the hcoCAcoNH, the transverse magnetization during these delays is on the CO. In this light it is also interesting to note that some hCOcaNH peaks are missing towards the end of the assigned stretch in figure 3.6. Also during this experiment, the transverse magnetization is on Cα for half of the scalar transfer step. However, the pattern here is not so clear, as the quality of this spectrum is slightly lower than expected. For example, also the hCOcaNH peak in the 46 isoleucine strip is missing while the hcaCBcaNH peak is large. We did not repeat the experiment to obtain a better spectrum, because the CO dispersion is not very high and therefore this spectrum is less valuable for assignments than the Cα and Cβ correlated experiments. For this reason, the CO-correlated spectra have been excluded from the plot in figure 3.7. It is likely that structural inhomogeneity and slow motions in the large extracellular loops OmpG are causing large variations in T2 relaxation times. This is discussed further in the next chapter when comparing the signal sets in the ^{1}H- and ^{13}C- datasets.

The peaks in the hCANH spectrum have a signal to noise around 10, which is far from excessive. Therefore it might be interesting to signal-average the hCANH experiment longer to see how many peaks are still hidden below the noise level. Also experiments with dipolar transfer schemes should be tried [25][20]. On well-behaved proteins such as SH3 scalar coupling based homo-nuclear transfer schemes proved slightly more efficient than dipolar based ones. However on proteins such as OmpG with T_2' times that are about three times shorter than in SH3 (8 vs. 25 ms for Cα) DREAM transfers might lead to less signal loss.

When just focusing on the signal loss during the scalar transfer blocks, in principle only the homogeneous component of the T_2 should play a role. Because these scalar transfer blocks are spin-echo sequences the inhomogeneous component is, at least partially, refocused by the π-pulse. What is not refocused is the coherence loss caused by motions in the μs-ms timescale. In table 3.1 linewidth and measured T_2 relaxation times are listed. By comparing linewidths and measured $T_{2,\mathrm{hom}}$, the inhomogeneous contribution to T_2 can be estimated using equation 1.14. As can be seen, the homogeneous component of T_2 is still very considerable. Even higher MAS frequencies will decrease the homogeneous component of T_2, and therefore will probably lead to better signal to noise. Progress towards higher spinning rates is still being made, and MAS of proteins close to a 100 kHz in 0.8 mm rotors has been reported [26].

Table 3.1: Measured and approximated T_2 relaxation times. Linewidth and homogeneous bulk T_2 times were measured (bold), all other values have been calculated based on these. The measurements of the ^{15}N linewidth are based on a selection of isolated peaks in the ^{15}N-^{1}H correlation spectra, recorded with a long enough acquisition time to avoid truncation of the signal, processed without window functions.

nucleus	**LW (Hz)**	T2* (ms)	**$T2_{hom}$(ms)**	$T2_{inhom}$ (ms)	LW_{nat} (Hz)
^{1}H	60 - 90	5.3 - 3.5	8	15.7 - 6.3	20 - 50
^{15}N	30 - 50	10.0 - 6.3	19	24.0 - 9.5	13 - 33
$^{13}C\alpha$	-	-	8	-	-

Another interesting observation that can be made by looking plots in figure 3.7 is that there is an alternating “zig-zag” pattern between high and low signal to noise within a sequential stretch. This can be seen for the residues 44 to 49 and 112 to 117. Upon comparison to the structure calculated later, the side chains of the residues with higher signal to noise point into the pore, whereas the side chains of the residues with lower signal to noise point into the lipid bilayer. This may indicate that the lipids add disorder, increasing relaxation. It is also the case that the lipids are not deuterated here. In previous studies, the effect of deuterated versus protonated lipids on the linewidths in OmpG was compared and it was found that at 20 kHz MAS and 30% back-exchanged protons, deuteration of the lipids decreased the ^{1}H-linewidth by 25%, while the effect on ^{15}N-linewidth was not significant [9]. The effect on the ^{13}C linewidth was not measured in that study.

Conclusion

The introduction of ^{1}H-detected experiments highly simplified the assignment procedure for OmpG. The addition of the ^{1}H dimension adds enough dispersion to generate unique strips for most residues. The strip based approach is intuitive and because there is one strip per residue it is easier to estimate how much of the total residue sequence is represented by the spectra than in ^{13}C-^{13}C correlations. It became clear that the signals for a large part of the sequence are missing. Whereas there is sufficient signal intensity in the middle of the assigned stretches, the signal intensity decreases towards the extremities of these stretches. A correlation between the intensities of the peaks in the hcaCBcaNH and hcaCBcacoNH measuring the same Cβ chemical shift (but other ^{1}H and ^{15}N chemical shifts) indicates that most signal is lost during the evolution of the scalar coupling in the ^{13}C-^{13}C transfer steps and not during the CP steps.

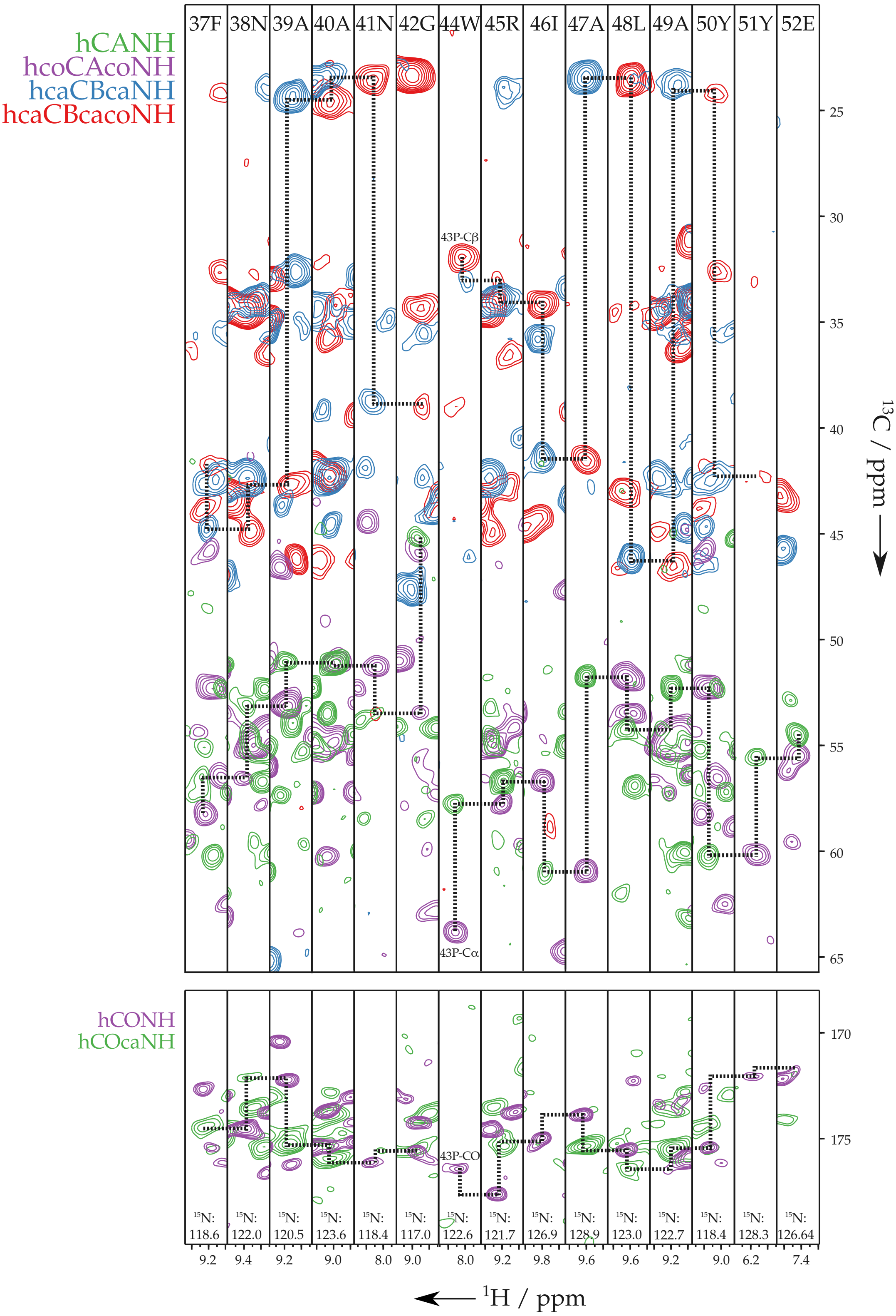


Figure 3.6: Strip plots from the backbone walk from phenylalanine 37 to Glutamine 52. Residue 42 is a proline and therefore this strip is not present. The correlations to the carbon nuclei can be observed in the strip of tryptophan 44. Notice how the signal intensities, especially of the peaks in the longer experiments like the hcaCBcacoNH, drop off towards the end of this sequential stretch and eventually completely disappear in the strip of Glutamic acid 52, which is the last assigned residue on this strand of the β-sheet.

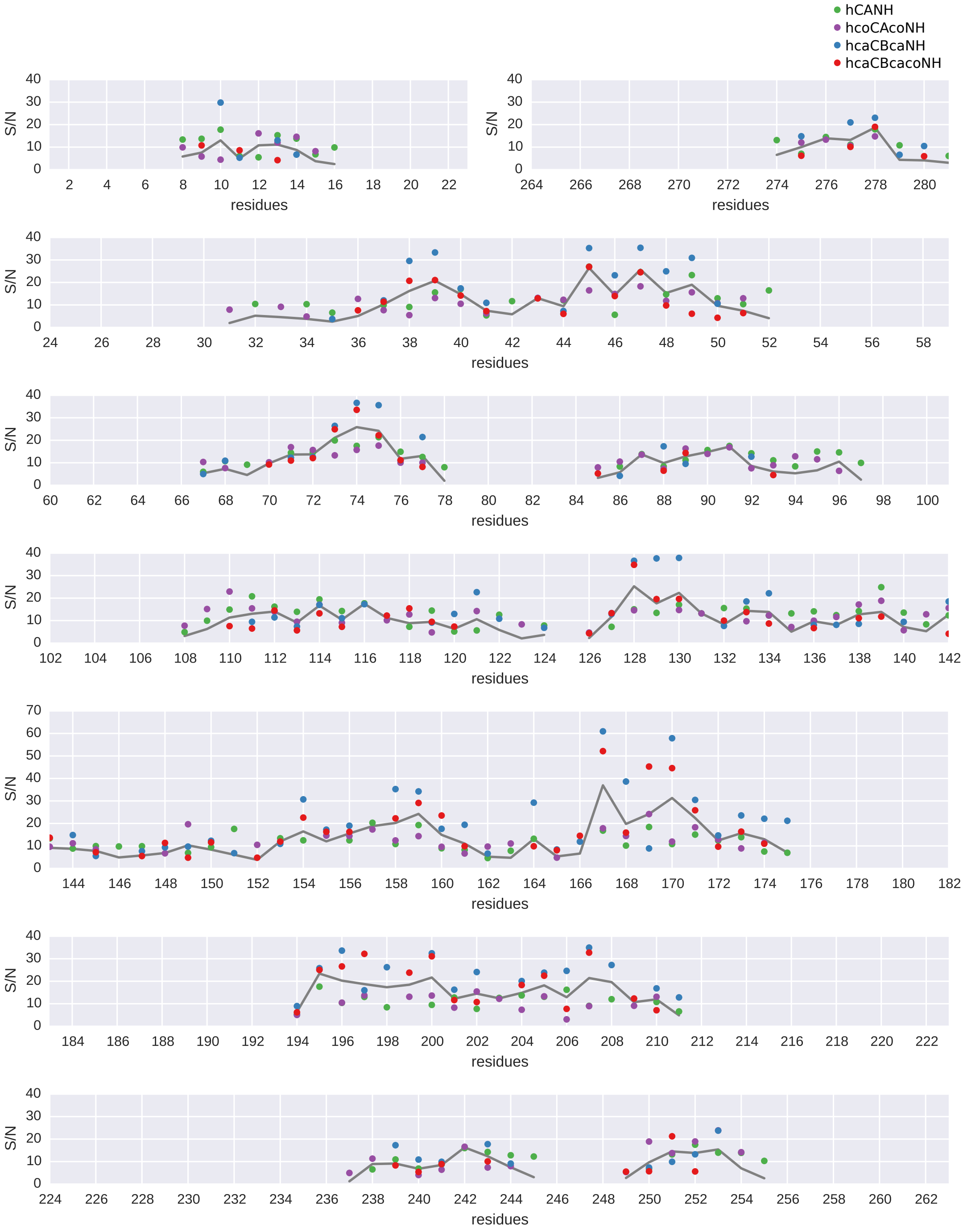


Figure 3.7: Signal intensities in the hCANH, hcoCAcoNH, hcaCBcaNH and hcaCBcacoNH versus sequence. Every panel represents two strands in the β-sheet connected by an intracellular turn in the structure, except for the two panels on the top which represent the first and last strand of the sequence. For all peaks, residue indices are based on the location of the measured carbon. I.e. peak intensities in the hcoCAcoNH and hcaCBcacoNH correspond to strips at the ^{15}N-^{1}H position of index+1. Noise level (1) is defined as one standard deviation of noise intensity calculated by CCPN analysis by taking 10 subsets of 1000 random samples in a spectrum and choosing the smallest subset.

Materials and methods

Sample preparation

Samples were prepared in the same way as the fully protonated samples used for the ^{13}C-detected experiments described in previous chapter, with a few exceptions [27]. The M9 minimal medium was perdeuterated. For the samples used to record the hCANH and hcoCAcoNH shown here the refolding buffer contained a mixture of 70:30 H_2O:D_2O. For the samples used in all other experiments, the refolding buffer did not contain D_2O. Reconstitution in lipid bilayers of all samples was performed in buffer containing H_2O. After reconstitution, the samples were pelleted and incubated in an MES buffer with pH 6.3. In the case of the hCANH and hcoCAcoNH samples, this buffer contained a mixture of 70:30 H_2O:D_2O. All samples were reconstituted in total lipid extract (Avanti Polar Lipids), except for the sample used to measure distance restraints (see chapter 6). For this sample, the polar lipid extract (Avanti Polar Lipids) was used. No differences where observed in ^{1}H-^{15}N and hCANH spectra recorded on this sample and the other samples.

NMR experiments

Experiments were recorded on an Bruker Avance III 800 MHz ^{1}H larmor frequency spectrometer at 60 kHz MAS using a triple-resonance HCN 1.3 mm probe. The temperature of the VT gas flow was set to 230 K, which roughly corresponds to a sample temperature of 300 K. 90°-pulses were 2.33 μs (107 kHz) for ^{1}H, 2.75 μs (90 kHz) for ^{13}C and 6.3 μs (40 kHz) for ^{15}N. H-Cα and H-CO CP conditions had a contact time of 2250 μs, with the ^{1}H spin lock centered at 87 kHz with a 30% ramp and the ^{13}C spin lock at a constant amplitude of 29 kHz. CO-N and CA-N CP conditions had a contact time of 12 ms and a constant amplitude spin lock of 14 kHz on ^{13}C and a tangent-modulated amplitude spin lock centered at 14 kHz on ^{15}N. ^{15}N-^{1}H CP was achieved using a contact time of 900 μs. The ^{1}H spin lock was centered at 81 kHz with a 30° ramp. The spin lock on ^{15}N had a constant amplitude of 39 kHz. Half echo delays τ were set to 4.7 ms for scalar transfers between Cα and CO and 7.2 ms for scalar transfers between Cα and Cβ. Selective π-pulses used during the second half of the Cα-CO in the hcaCBcacoNH and hCOcaNH experiments were Gaussian-cascade Q3 pulses [28]. The pulses on CO were 350 μs with an amplitude of 1.25 kHz. The pulse on Cα was 1 ms with an amplitude of 0.14 kHz. Selective Q3 pulses during the indirect acquisition of ^{13}C in the hCANH, hCONH, hCOcaNH an hcoCAcoNH had a duration of 350 μs and an amplitude of 1.01 kHz, for both CO- and Cα-selective pulses. Selective Q3 π-pulses on the entire aliphatic region during the Cα-Cβ scalar transfer blocks in the hcaCBcaNH and hcaCBcacoNH experiments were 200 μs with an amplitude of 3.37 kHz. Water suppression was achieved using the MISSISSIPI sequence without homospoil gradients [29]. Swept-low-power TPPM was used for ^{1}H decoupling and WALTZ-16 for ^{15}N and ^{13}C decoupling duing ^{1}H-detection [30][31]. All spectra were acquired using States-TPPI in the direct dimensions to obtain pure-phase line shapes and phase discrimination [32]. Table 3.2 lists acquisition times in the indirect ^{13}C and ^{15}N dimensions and total duration of the experiments.

Table 3.2: Acquisition parameters for the six ^{1}H-detected experiments used for the assignment of OmpG.

experiment	^{13}C aq (ms)	^{15}N aq (ms)	scans	duration
hCANH	8.6	14.4	12	1d15h
hcoCAcoNH	6.6	14.4	36	3d18h
hcaCBcaNH	3.0	5.0	64	2d20h
hcaCBcacoNH	3.0	5.4	128	5d23h
hCONH	10.0	14.0	8	16h
hCOcaNH	10.0	9.8	64	3d15h

References

[1] B. Reif, C. P. Jaroniec, C. M. Rienstra, M. Hohwy, and R. G. Griffin. “1H–1H MAS Correlation Spectroscopy and Distance Measurements in a Deuterated Peptide”. *Journal of Magnetic Resonance* 151.2 (Aug. 2001), pp. 320–327. DOI: 10.1006/jmre.2001.2354.

[2] B. Reif and R. G. Griffin. “1H Detected 1H,15N Correlation Spectroscopy in Rotating Solids”. *Journal of Magnetic Resonance* 160.1 (Jan. 2003), pp. 78–83. DOI: 10.1016/S1090-7807(02)00035-6.

[3] V. Chevelkov, B. J. van Rossum, F. Castellani, K. Rehbein, A. Diehl, M. Hohwy, S. Steuernagel, F. Engelke, H. Oschkinat, and B. Reif. “1H Detection in MAS Solid-State NMR Spectroscopy of Biomacromolecules Employing Pulsed Field Gradients for Residual Solvent Suppression”. *Journal of the American Chemical Society* 125.26 (July 2003), pp. 7788–7789. DOI: 10.1021/ja029354b.

[4] E. K. Paulson, C. R. Morcombe, V. Gaponenko, B. Dancheck, R. A. Byrd, and K. W. Zilm. “Sensitive High Resolution Inverse Detection NMR Spectroscopy of Proteins in the Solid State”. *Journal of the American Chemical Society* 125.51 (Dec. 2003), pp. 15831–15836. DOI: 10.1021/ja037315+.

[5] V. Chevelkov, K. Rehbein, A. Diehl, and B. Reif. “Ultrahigh Resolution in Proton Solid-State NMR Spectroscopy at High Levels of Deuteration”. *Angewandte Chemie International Edition* 45.23 (June 2006), pp. 3878–3881. DOI: 10.1002/anie.200600328.

[6] R. Linser, U. Fink, and B. Reif. “Proton-Detected Scalar Coupling Based Assignment Strategies in MAS Solid-State NMR Spectroscopy Applied to Perdeuterated Proteins”. *Journal of Magnetic Resonance* 193.1 (July 2008), pp. 89–93. DOI: 10.1016/j.jmr.2008.04.021.

[7] R. Linser, B. Bardiaux, V. Higman, U. Fink, and B. Reif. “Structure Calculation from Unambiguous Long-Range Amide and Methyl 1H-1H Distance Restraints for a Microcrystalline Protein with MAS Solid-State NMR Spectroscopy”. *Journal of the American Chemical Society* 133.15 (Apr. 2011), pp. 5905–5912. DOI: 10.1021/ja110222h.

[8] Ü. Akbey, S. Lange, W. T. Franks, R. Linser, K. Rehbein, A. Diehl, B.-J. van Rossum, B. Reif, and H. Oschkinat. “Optimum Levels of Exchangeable Protons in Perdeuterated Proteins for Proton Detection in MAS Solid-State NMR Spectroscopy”. *Journal of Biomolecular NMR* 46.1 (Aug. 2009), pp. 67–73. DOI: 10.1007/s10858-009-9369-0.

[9] R. Linser, M. Dasari, M. Hiller, V. Higman, U. Fink, J.-M. Lopez del Amo, S. Markovic, L. Handel, B. Kessler, P. Schmieder, D. Oesterhelt, H. Oschkinat, and B. Reif. “Proton-Detected Solid-State NMR Spectroscopy of Fibrillar and Membrane Proteins”. *Angewandte Chemie International Edition* 50.19 (May 2011), pp. 4508–4512. DOI: 10.1002/anie.201008244.

[10] D. H. Zhou, G. Shah, M. Cormos, C. Mullen, D. Sandoz, and C. M. Rienstra. "Proton-Detected Solid-State NMR Spectroscopy of Fully Protonated Proteins at 40 kHz Magic-Angle Spinning". *Journal of the American Chemical Society* 129.38 (Sept. 2007), pp. 11791–11801. DOI: 10.1021/ja073462m.

[11] J. R. Lewandowski, J.-N. Dumez, Ü. Akbey, S. Lange, L. Emsley, and H. Oschkinat. "Enhanced Resolution and Coherence Lifetimes in the Solid-State NMR Spectroscopy of Perdeuterated Proteins under Ultrafast Magic-Angle Spinning". *The Journal of Physical Chemistry Letters* 2.17 (Sept. 2011), pp. 2205–2211. DOI: 10.1021/jz200844n.

[12] A. J. Nieuwkoop, W. T. Franks, K. Rehbein, A. Diehl, Ü. Akbey, F. Engelke, L. Emsley, G. Pintacuda, and H. Oschkinat. "Sensitivity and Resolution of Proton Detected Spectra of a Deuterated Protein at 40 and 60 kHz Magic-Angle-Spinning". *Journal of Biomolecular NMR* 61.2 (Feb. 2015), pp. 161–171. DOI: 10.1007/s10858-015-9904-0.

[13] M. Ernst, A. Samoson, and B. H. Meier. "Low-Power Decoupling in Fast Magic-Angle Spinning NMR". *Chemical Physics Letters* 348.3–4 (Nov. 2001), pp. 293–302. DOI: 10.1016/S0009-2614(01)01115-0.

[14] M. Kotecha, N. P. Wickramasinghe, and Y. Ishii. "Efficient Low-Power Heteronuclear Decoupling in 13C High-Resolution Solid-State NMR under Fast Magic Angle Spinning". *Magnetic Resonance in Chemistry* 45.S1 (Dec. 2007), S221–S230. DOI: 10.1002/mrc.2151.

[15] S. Laage, J. R. Sachleben, S. Steuernagel, R. Pierattelli, G. Pintacuda, and L. Emsley. "Fast Acquisition of Multi-Dimensional Spectra in Solid-State NMR Enabled by Ultra-Fast MAS". *Journal of Magnetic Resonance* 196.2 (Feb. 2009), pp. 133–141. DOI: 10.1016/j.jmr.2008.10.019.

[16] S. Laage, A. Marchetti, J. Sein, R. Pierattelli, H. J. Sass, S. Grzesiek, A. Lesage, G. Pintacuda, and L. Emsley. "Band-Selective 1H-13C Cross-Polarization in Fast Magic Angle Spinning Solid-State NMR Spectroscopy". *Journal of the American Chemical Society* 130.51 (Dec. 2008), pp. 17216–17217. DOI: 10.1021/ja805926d.

[17] D. Zhou, J. Shea, A. Nieuwkoop, W. Franks, B. Wylie, C. Mullen, D. Sandoz, and C. Rienstra. "Solid-State Protein-Structure Determination with Proton-Detected Triple-Resonance 3D Magic-Angle-Spinning NMR Spectroscopy". *Angewandte Chemie International Edition* 46.44 (Nov. 2007), pp. 8380–8383. DOI: 10.1002/anie.200702905.

[18] M. J. Knight, A. L. Webber, A. J. Pell, P. Guerry, E. Barbet-Massin, I. Bertini, I. C. Felli, L. Gonnelli, R. Pierattelli, L. Emsley, A. Lesage, T. Herrmann, and G. Pintacuda. "Fast Resonance Assignment and Fold Determination of Human Superoxide Dismutase by High-Resolution Proton-Detected Solid-State MAS NMR Spectroscopy". *Angewandte Chemie International Edition* 50.49 (Dec. 2011), pp. 11697–11701. DOI: 10.1002/anie.201106340.

[19] E. Barbet-Massin, A. J. Pell, J. S. Retel, L. B. Andreas, K. Jaudzems, W. T. Franks, A. J. Nieuwkoop, M. Hiller, V. Higman, P. Guerry, A. Bertarello, M. J. Knight, M. Felletti, T. Le Marchand, S. Kotelovica, I. Akopjana, K. Tars, M. Stoppini, V. Bellotti, M. Bolognesi, S. Ricagno, J. J. Chou, R. G. Griffin, H. Oschkinat, A. Lesage, L. Emsley, T. Herrmann, and G. Pintacuda. "Rapid Proton-Detected NMR Assignment for Proteins with Fast Magic Angle Spinning". *Journal of the American Chemical Society* 136.35 (Sept. 2014), pp. 12489–12497. DOI: 10.1021/ja507382j.

[20] D. H. Zhou, A. J. Nieuwkoop, D. A. Berthold, G. Comellas, L. J. Sperling, M. Tang, G. J. Shah, E. J. Brea, L. R. Lemkau, and C. M. Rienstra. "Solid-State NMR Analysis of Membrane Proteins and Protein Aggregates by Proton Detected Spectroscopy". *Journal of Biomolecular NMR* 54.3 (Sept. 2012), pp. 291–305. DOI: 10.1007/s10858-012-9672-z.

[21] V. Chevelkov, B. Habenstein, A. Loquet, K. Giller, S. Becker, and A. Lange. "Proton-Detected MAS NMR Experiments Based on Dipolar Transfers for Backbone Assignment of Highly Deuterated Proteins". *Journal of Magnetic Resonance* 242 (May 2014), pp. 180–188. DOI: 10.1016/j.jmr.2014.02.020.

[22] M. J. Knight, A. J. Pell, I. Bertini, I. C. Felli, L. Gonnelli, R. Pierattelli, T. Herrmann, L. Emsley, and G. Pintacuda. "Structure and Backbone Dynamics of a Microcrystalline Metalloprotein by Solid-State NMR". *Proceedings of the National Academy of Sciences* 109.28 (Oct. 2012), pp. 11095–11100. DOI: 10.1073/pnas.1204515109.

[23] E. Barbet-Massin, A. J. Pell, K. Jaudzems, W. T. Franks, J. S. Retel, S. Kotelovica, I. Akopjana, K. Tars, L. Emsley, H. Oschkinat, A. Lesage, and G. Pintacuda. "Out-and-Back 13C–13C Scalar Transfers in Protein Resonance Assignment by Proton-Detected Solid-State NMR under Ultra-Fast MAS". *Journal of Biomolecular NMR* 56.4 (June 2013), pp. 379–386. DOI: 10.1007/s10858-013-9757-3.

[24] B. Liang and L. K. Tamm. "Structure of Outer Membrane Protein G by Solution NMR Spectroscopy". *Proceedings of the National Academy of Sciences* 104.41 (Sept. 2007), pp. 16140–16145. DOI: 10.1073/pnas.0705466104.

[25] R. Verel, M. Ernst, and B. H. Meier. "Adiabatic Dipolar Recoupling in Solid-State NMR: The DREAM Scheme". *Journal of Magnetic Resonance* 150.1 (May 2001), pp. 81–99. DOI: 10.1006/jmre.2001.2310.

[26] V. Agarwal, S. Penzel, K. Szekely, R. Cadalbert, E. Testori, A. Oss, J. Past, A. Samoson, M. Ernst, A. Böckmann, and B. H. Meier. "De Novo 3D Structure Determination from Sub-Milligram Protein Samples by Solid-State 100 kHz MAS NMR Spectroscopy". *Angewandte Chemie International Edition* 53.45 (Nov. 2014), pp. 12253–12256. DOI: 10.1002/anie.201405730.

[27] M. Hiller, L. Krabben, K. R. Vinothkumar, F. Castellani, B.-J. van Rossum, W. Kühlbrandt, and H. Oschkinat. "Solid-State Magic-Angle Spinning NMR of Outer-Membrane Protein G from Escherichia Coli". *ChemBioChem* 6.9 (Sept. 2005), pp. 1679–1684. DOI: 10.1002/cbic.200500132.

[28] L. Emsley and G. Bodenhausen. "Gaussian Pulse Cascades: New Analytical Functions for Rectangular Selective Inversion and in-Phase Excitation in NMR". *Chemical Physics Letters* 165.6 (Feb. 1990), pp. 469–476. DOI: 10.1016/0009-2614(90)87025-M.

[29] D. H. Zhou and C. M. Rienstra. "High-Performance Solvent Suppression for Proton Detected Solid-State NMR". *Journal of Magnetic Resonance* 192.1 (May 2008), pp. 167–172. DOI: 10.1016/j.jmr.2008.01.012.

[30] J. R. Lewandowski, J. Sein, M. Blackledge, and L. Emsley. "Anisotropic Collective Motion Contributes to Nuclear Spin Relaxation in Crystalline Proteins". *Journal of the American Chemical Society* 132.4 (Feb. 2010), pp. 1246–1248. DOI: 10.1021/ja907067j.

[31] A. J. Shaka, J. Keeler, T. Frenkiel, and R. Freeman. "An Improved Sequence for Broadband Decoupling: WALTZ-16". *Journal of Magnetic Resonance (1969)* 52.2 (Apr. 1983), pp. 335–338. DOI: 10.1016/0022-2364(83)90207-X.

[32] D. Marion, M. Ikura, R. Tschudin, and A. Bax. "Rapid Recording of 2D NMR Spectra without Phase Cycling. Application to the Study of Hydrogen Exchange in Proteins". *Journal of Magnetic Resonance (1969)* 85.2 (Nov. 1989), pp. 393–399. DOI: 10.1016/0022-2364(89)90152-2.

Chapter 4

Combining Information Obtained from ^{13}C- and ^{1}H-detected Experiments

Access to side-chain chemical shifts in perdeuterated proteins

For the sequential assignment of proteins, the knowledge of side-chain chemical shifts other than the Cβ has many advantages. Because these shifts vary more widely between the different amino acids, residue typing is simplified which in turn helps mapping stretches of connected spin systems to a subsequence of the protein. Also, side-chain chemical shifts are needed to generate the necessary distance restraints for structure calculation of most proteins. In perdeuterated proteins, two related problems arise that complicate the acquisition of the spectra necessary to assign the side-chain resonances. First of all, the protons in close proximity to the carbon nuclei in the side-chain are removed making ^{1}H-^{13}C cross-polarization less efficient. Second, mixing schemes that are based on the re-introduction of ^{1}H-^{1}H dipolar couplings such as DARR and PDSD decrease in efficiency as MAS rate increases.

The simplest solution to the first problem is to use direct ^{13}C excitation. This leads to lower signal to noise compared to ^{1}H-^{13}C CP in a fully protonated protein due to the lower gyromagnetic ratio of ^{13}C. Also the ^{13}C T1 relaxation times are long and therefore a long recycle delay is necessary. However this can be mitigated by adding a paramagnetic relaxation enhancer to the sample [1][2]. An alternative solution is to transfer magnetization from deuterium to carbon. Although deuterium has a slightly lower gyromagnetic ratio than ^{13}C itself (4.11x10^{7} versus 6.73x10^{7} T^{-1}s^{-1}) its T_1 time is very short allowing for very fast repetition of experiments. Initial deuterium excited experiments have been performed on SH3, ubiquitin and OmpG [3][4]. An additional benefit is that the deuterium double quantum frequency was measured in an indirect dimension, adding dispersion to the spectra. This technique is promising but not very mature yet.

There are multiple possible mixing schemes that can transfer magnetization between carbons in the side-chain that do not rely on protons such as DREAM, RFDR and TOBSY [5][6][7][8][9]. The ideal experiment would be an experiment similar to the NH-TOCSY used in solution NMR which correlates a complete set of side-chain ^{13}C resonances to the ^{15}N-^{1}H pair in the backbone [10][11]. Indeed, *Linser* performed a conceptually similar experiment on SH3 in the solid state using a combination of a long-range ^{1}H-^{13}C CP step and direct ^{13}C excitation to transfer sufficient magnetization to the carbons and using TOBSY as the mixing sequence [2]. Here the ^{13}C magnetization is transferred back to the backbone using another long-range ^{1}H-^{13}C CP instead of a ^{15}N-^{13}C CP, which positively influenced the signal to noise but comes at the cost of losing some specificity for intra-residual peaks. In a larger protein, this will be more problematic because signal overlap and ambiguity are increased. It would be interesting to see how well a similar sequence based on deuterium excitation works.

Another approach which gives access to both assignments and distance restraints is to use a sample that is sparsely and randomly protonated at non-exchangeable sites [12][13][14]. Alternatively, the introduction of proton labeled methyls, as is often done in solution NMR of large proteins with long correlation times, gives access to structural restraints and has also been applied in the solid state [15][16][17][18].

Combining information from protonated and deuterated samples

To perform the assignment of OmpG, a combination of ^{13}C- and ^{1}H-detected experiments was used. Since a large set of fully protonated and residue specifically labeled samples and corresponding spectra already existed this was the most straight-forward approach. The three nuclei that are measured in both the ^{1}H- and ^{13}C-detected experiments, the N, Cα and Cβ, can be used to connect data from both these types of samples. The ^{13}C chemical shifts found in a strip of the ^{1}H detected spectra can be used to find the Cα-Cβ cross peak 2D ^{13}C-^{13}C correlation spectra. If there is enough dispersion in the ^{13}C-^{13}C spectrum this peak can be easily found uniquely which will identify the rest of the ^{13}C chemical shifts of the side-chain. The NCACX spectra were also used for this purpose. In this case, the 2D ^{13}-^{13}C spectra were of superior quality and proved more useful. An added advantage of using both protonated and deuterated samples is that both ^{1}H-^{1}H restraints from ^{1}H-detected RFDR experiments and ^{13}C-^{13}C restraints from PDSD or DARR experiments could be used during the structure calculation.

Isotope Shift

The replacement of protons by deuterons alters the ^{13}C chemical shifts by up to a full ppm. Therefore, to compare data from protonated and deuterated samples, this deuterium isotope shift must be corrected for. This deuterium shift has been described before and quantified by solution NMR spectroscopists [19][20][21]. The magnitude of the shift can be approximated by the following equation:

$$\Delta C(D) =^1 \Delta C(D)d_1 +^2 \Delta C(D)d_2 +^3 \Delta C(D)d_3 \tag{4.1}$$

Here d_1, d_2 and d_3 are the amount of deuterons one, two and three bonds away from the carbon nucleus of interest. For all amino acid types, except Glycine, more deuterons are surrounding the Cβ than the Cα and therefore the Cβ shifts are more affected. One of the residue types with the largest isotope shifts is leucine, as can be seen in figure 4.1. Both *Venters et al.* and *Maltsev et al.* determined the factors ^{i}ΔC(D) experimentally, but got slightly different values [20][21]. As argued by *Maltsev et al.* this is just an estimate, as the real values also heavily depend on the local structure. Indeed the study of *Maltsev et al.* used α-synuclein, which is an intrinsically disordered protein where *Venters et al.* used human carbonic anhydrase I, which is mostly β-sheet with some small α-helices.

The values found by these studies can be used to approximate the isotope shift and will in most cases be good enough to connect the resonances in proton and carbon detected spectra. A more exact calculation of the isotope shift does not seem possible for now and if possible in the future it will most probably involve at least secondary structure information like φ and χ angles, which are most likely not known at the stage of sequential assignment.

Strategy for combining spin systems

Residue specifically labeled samples ease the process of matching Cα and Cβ chemical shifts in protonated and deuterated samples enormously because of reduced overlap in the Cα-Cβ regions of these spectra. Due to the large number of labeling schemes produced, for nearly every residue type there is a spectrum in which the Cα-Cβ cross-peaks are well resolved, see table 4.1. It is a lot less error-prone and faster to find a Cα-Cβ peak in the ^{13}C-detected 2D spectra from a strip in the ^{1}H-detected data than the reverse procedure. The reason for this is that when starting from a peak in the ^{13}C-detected 2D spectra all planes in the ^{1}H-detected 3D spectra have to be checked for a fitting Cα-Cβ combination. An alternative to going through all planes in the 3D spectra is to use two windows, one displaying hCANH and the other the hcaCBcaNH, with the ^{13}C-dimension in the z-direction. By setting the first window to the Cα chemical shift and the second to the Cβ chemical shift, peaks can be found that are present in both displayed ^{1}H-^{15}N planes. This technique is complicated and since the Cα and Cβ chemical shifts are not exactly known (because of the isotope shift), prone to mistakes.

The most linear approach to sequential assignment would be to first finish the backbone and Cβ chemical shift assignment using the ^{1}H detected strip matching approach described earlier, and afterward find the ^{13}C side-chain chemical shifts using ^{13}C-detected spectra. However, whether it is easier to first sequentially assign a strip in the ^{1}H-detected data to a residue and then find its ^{13}C detected counterpart or the other way around depends completely on the situation. If the Cα-Cβ combination in the deuterated sample corresponds to a resolved region of the uniformly labeled 2D ^{13}C-^{13}C spectra, it is favorable to find the corresponding spin system in the ^{13}C-detected data first since the improved residue typing decreases the number of options for the sequential assignment. In addition, inter-residual cross peaks in the ^{13}C-^{13}C correlation can further confirm that two strips are really a sequential match.

Table 4.1: For every amino acid there is a labeled sample where the intra-residual peaks are best resolved. For some residue types, multiple spectra could be used as a reference spectrum in which case the one specifically used in this study is listed. For methionine, there is no labeling scheme in which the Cα-Cβ is separated well from other peaks. Therefore the Cα-Cγ peak is used. The Cγ chemical shift can not be observed in ^{1}H-detected spectra and only the Cα shift can be used directly. Further support for the assignment is given by sequential cross peaks.

amino acid	sample	comment
alanine	RIGA(S)	
asparagine / aspartic acid	1,3 MKINDT	
glutamine /glutamic acid	1,3-TEMPQANDSG	
phenylalanine / tyrosine	GAFY	
glycine	RIGA(S)	Cα-CO peak is used
histidine	GANDSH	
isoleucine	RIGA(S)	
lysine	1,3-MKINDT	
leucine	GAVLS(W)	
methionine	2-TEMPQANDSG	Cα-Cγ peak is used
proline	1,3-TEMPQANDSG	
arginine	RIGA(S)	
serine	RIGA(S)	
threonine	1,3-TEMPQANDSG	
valine	GAVLS(W)	
tryptophan	GAVLS(W)	

However, if the Cα-Cβ combination in a strip of the deuterated sample corresponds to a very crowded area of the uniformly labeled ^{13}C-^{13}C spectrum, it can sometimes be easier to sequentially assign the spin system purely based on matching strips in the ^{1}H-detected spectra first. This is because crowding in the uniformly labeled sample means that for a degenerate Cα-Cβ combination it is not known yet in which spectrum of which residue-specifically labeled sample to look for the Cα-Cβ peak. Of course finding the i+1 strip is also harder for very degenerate Cα-Cβ combinations, but at least the peak positions in the ^{13}C-dimensions of the matching strips should fit almost perfectly because of the lack of isotope shift within spectra recorded using the same sample. Also, when the hcaCBcacoNH and hcoCAcoNH peaks in the strip correspond to a less degenerate Cα-Cβ combination the assignment can be easily extended in the N-terminal direction. After the sequential assignment of a particular spin system is done, it is a lot easier to find the corresponding

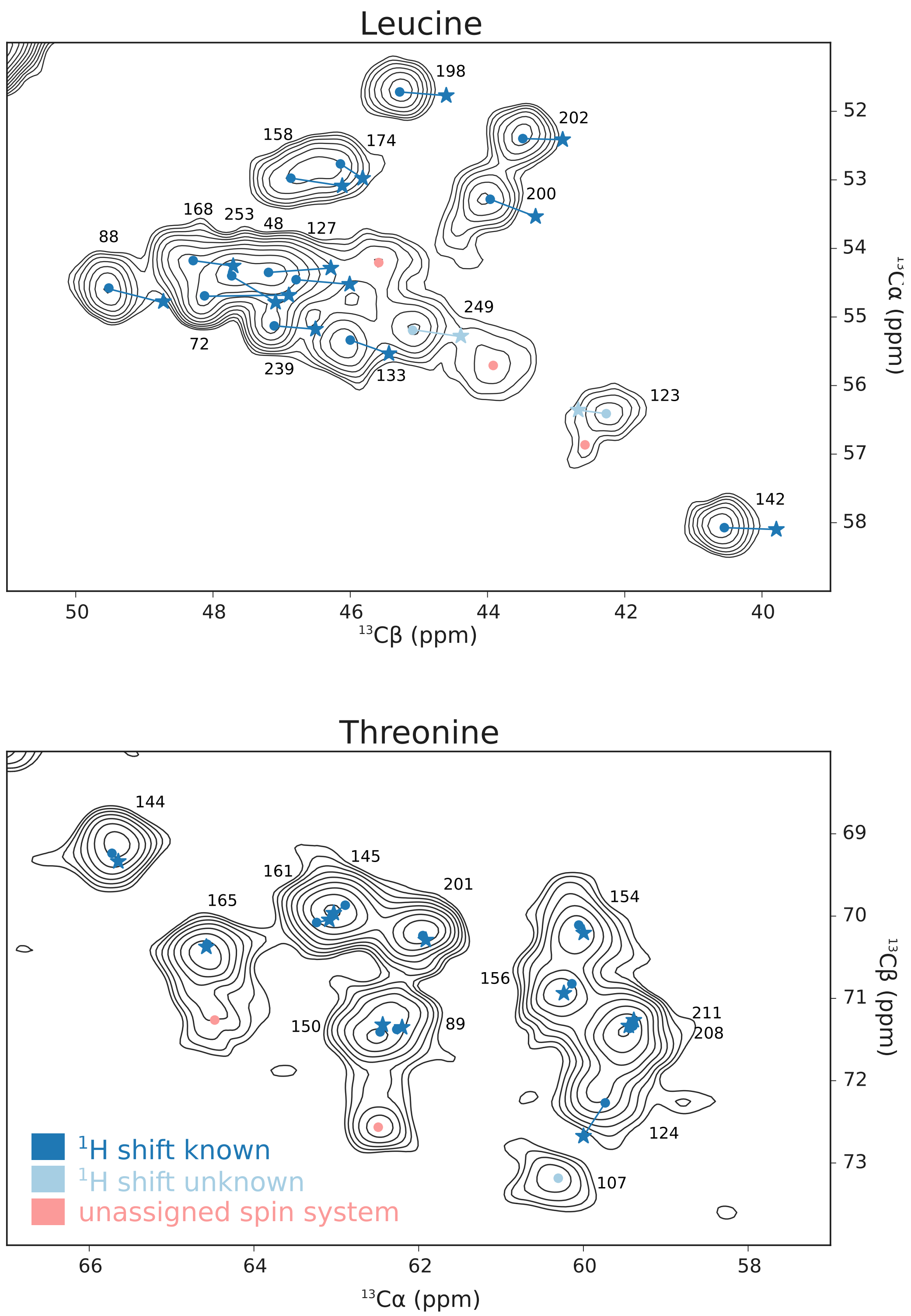


Figure 4.1: Mapping Cα-Cβ combinations from the experiments on deuterated samples on 2D ^{13}C-^{13}C DARR spectra. For leucine the GAVLS(W) spectrum (20 ms DARR) is used; for the threonine the 1,3-TEMPQANDSG (50 ms DARR). Positions of Cα-Cβ chemical shifts in deuterated and protonated samples are depicted by stars and circles respectively. Blue colors indicate assigned residues. Light blue indicates that no strip in the hCANH, hcaCBcaNH or hCOcaNH could be found and therefore the ^{1}H shift is unknown. Pink peaks are unassigned and no strip could be found in the ^{1}H-detected data.

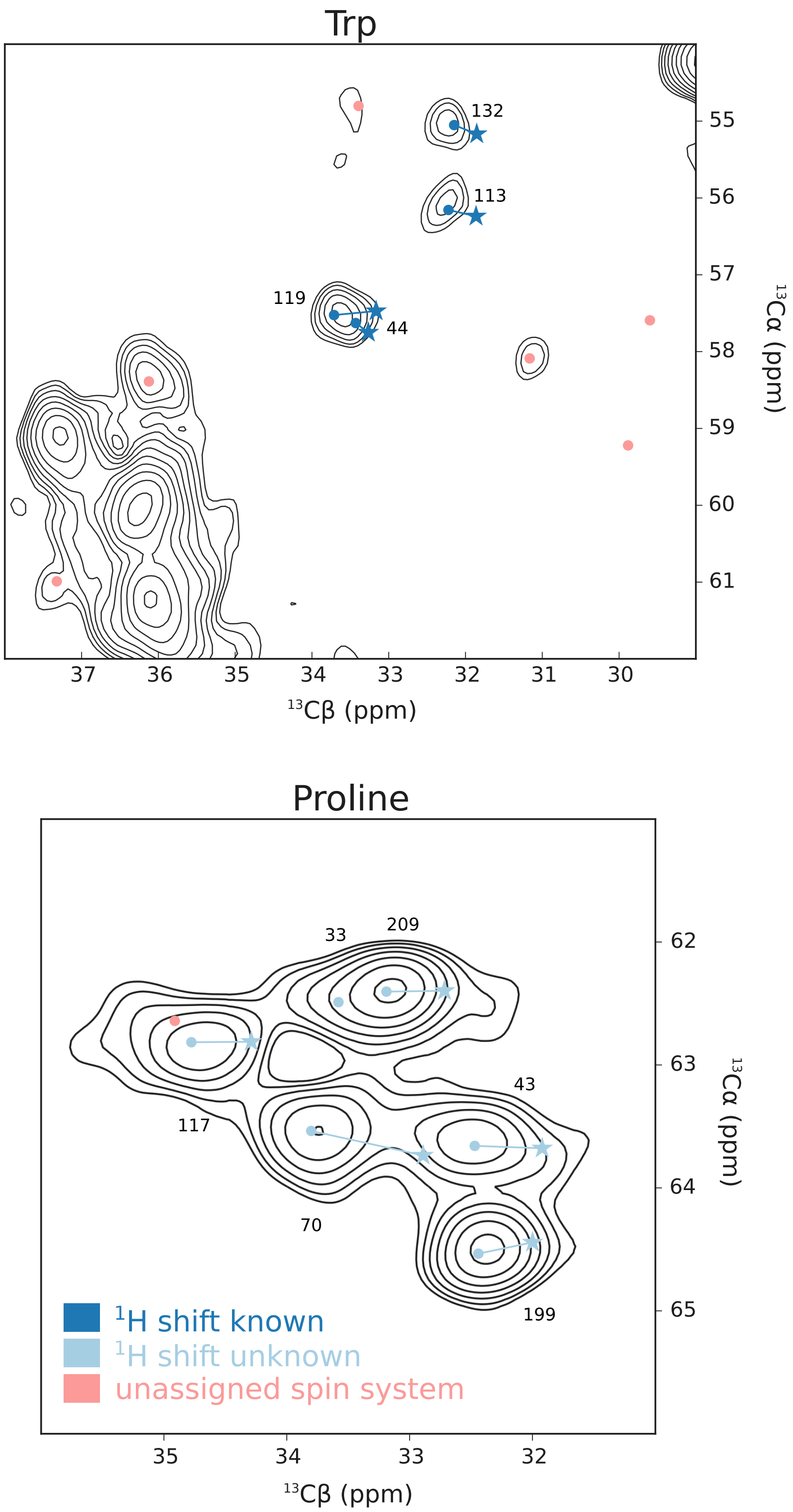


Figure 4.2: Mapping Cα-Cβ combinations for tryptophan and proline. These are the same two spectra as shown in figure 5.1. GAVLS(W) was used for tryptophan and 1,3-TEMPQANDSG for prolines. Colors also have the same meaning. Two of the tryptophan peaks are very close to the noise and therefore below the contour level drawn here.

^{13}C-detected spin system since now the residue type is known, which limits the choice between possible spin systems and it is clear which of the residue specific labeled ^{13}C-^{13}C correlations to use to find the matching Cα-Cβ peak. If it is still not clear which Cα-Cβ peak should be chosen, the exact resonance frequencies in the protonated samples can be found by looking at sequential cross-peaks instead of just at the Cα-Cβ peaks. When the strip has already been sequentially assigned, this becomes a lot more trivial since often the correct ^{13}C chemical shifts of the neighboring spin system in the protonated sample are already known. In practice, there is no sharp distinction between the two strategies, since they can basically be used at the same time. In solution NMR the situation is similar. Often the entire backbone is assigned first before the TOCSY spectra are used to find the side-chain chemical shifts. However, in many cases, they are consulted during the sequential assignment process to aid residue typing.

Final extent of the assignment

By combining data from the ^{1}H- and ^{13}C-detected spectra, a coherent assignment could be found for a bit less than 60% of the residues, see table 4.2. As can be seen in this table, there are some assigned residues for which the ^{1}H and ^{15}N chemical shifts are not assigned. Often these residues are the first residue in an assigned stretch and therefore the ^{13}C chemical shifts are known from the hcaCBcacoNH and hcoCAcoNH from the next residue in the stretch. This is also the case for all prolines and for example leucines 149 and 123 shown in figure 4.1. Also there are a few residues that only have assignments in the ^{1}H-detected data. In this case, it was very hard to determine exactly where the corresponding shifts in the ^{13}C-detected spectra were. This was the case for some of the glutamic acid and glutamine residues and residues where only the Cα-peak was found in the ^{1}H detected data and not the Cβ-peak, which makes finding the corresponding spin system in the ^{13}C-detected data harder. Appendix A contains the full chemical shift list, where the Cα and Cβ shifts given for both the protonated and deuterated samples.

As can bee seen by comparing figures 4.1 and 4.2 to figure 4.3 peaks are present in the ^{13}C-^{13}C correlations for nearly all leucine, threonine tryptophan and proline residues in the sequence. There are however, some peaks in the ^{13}C-^{13}C spectra for which no strip could be found in the ^{1}H-detected data (colored light blue and pink in figures 4.1 and 4.2). With the exception of the prolines, these residues had a comparatively low signal intensity and often unregular lineshapes in the ^{13}C-detected data. Only the tryptophan peak at position 36/58 ppm (x-dimension/y-dimension) is larger but could not be assigned anyway. Since inter-residual cross peaks in the DARR spectra with longer mixing times are about ten times weaker than the intra-residual peaks, no sequential cross peak pattern could be found to allow the assignment of the left-over unassigned spin systems. It is unfortunate that these peaks are not present, preventing a complete assignment. However, as is discussed later in chapter 6, inter-residual cross-peaks between unassigned residues lead to wrong distance restraints, which is one of the largest challenges during structure calculation. In that context, it is an advantage that most of these peaks are absent.

Table 4.2: Extend of the assignment in the shift lists based on the carbon and proton detected spectra.

	of assigned residues	of all residues
Carbon Detected		
Residues	165/170 (97%)	165/281 (59%)
N backbone	124/170 (73%)	124/281 (44%)
C aliphatic	443/485 (91%)	443/781 (57%)
C aromatic	58/227 (26%)	58/341 (17%)
C carbonyl	127/204 (62%)	127/360 (35%)
CA	163/170 (96%)	163/281 (58%)
CB	145/156 (93%)	145/254 (57%)
CO (backbone)	117/170 (69%)	117/281 (42%)
Proton Detected		
Residues	167/170 (98%)	167/281 (59%)
H backbone	151/164 (92%)	151/273 (55%)
N backbone	151/170 (89%)	151/281 (54%)
CA	167/170 (98%)	167/281 (59%)
CB	131/156 (84%)	131/254 (52%)
CO (backbone)	131/170 (77%)	131/281 (47%)

As can be seen in figure 4.3 almost all missing assignments cluster near the extracellular part of the protein or in the intra-cellular turns. The crystal structure, the solution NMR structure and the structure calculated during this work all share this same basic topology. Such a topology can also be predicted for beta-barrels with programs such as PRED-TMBB, see figure 6.1 [22]. The parts on the extra-cellular side that could not be assigned here fit very well to where there are flexible loops in the solution structure of *Liang and Tamm* [23]. This explains the heavy broadening of both the peaks in the ^{1}H- and ^{13}C-detected data shown here. Although more residues could be assigned in the solution state, hardly any distance restraints could be found in these extracellular loops. While the crystal structure for this part the beta-barrel extends further into the extra-cellular space, his is most likely due to crystal artifacts. The fact we don't see these signals in our lipid preparation is further evidence that the native structure is better represented by the NMR structures.

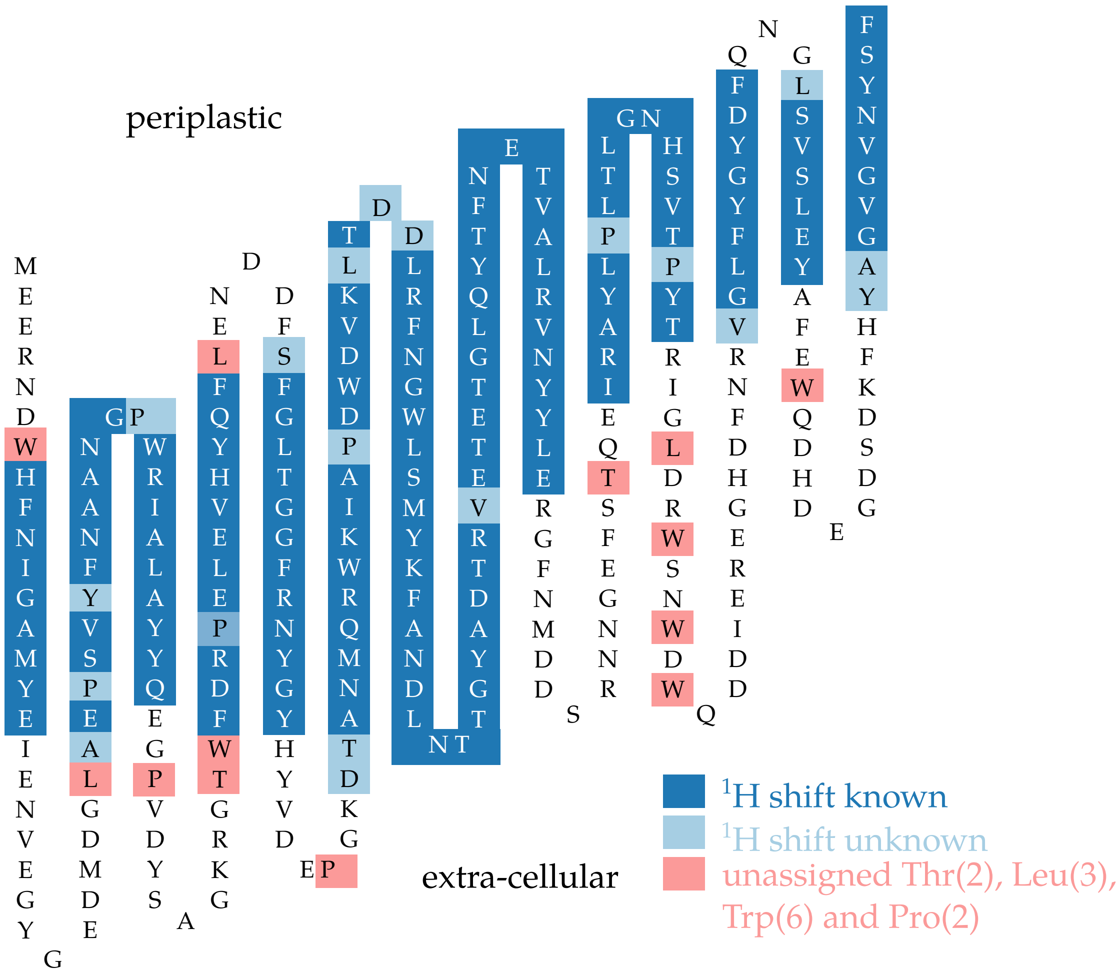
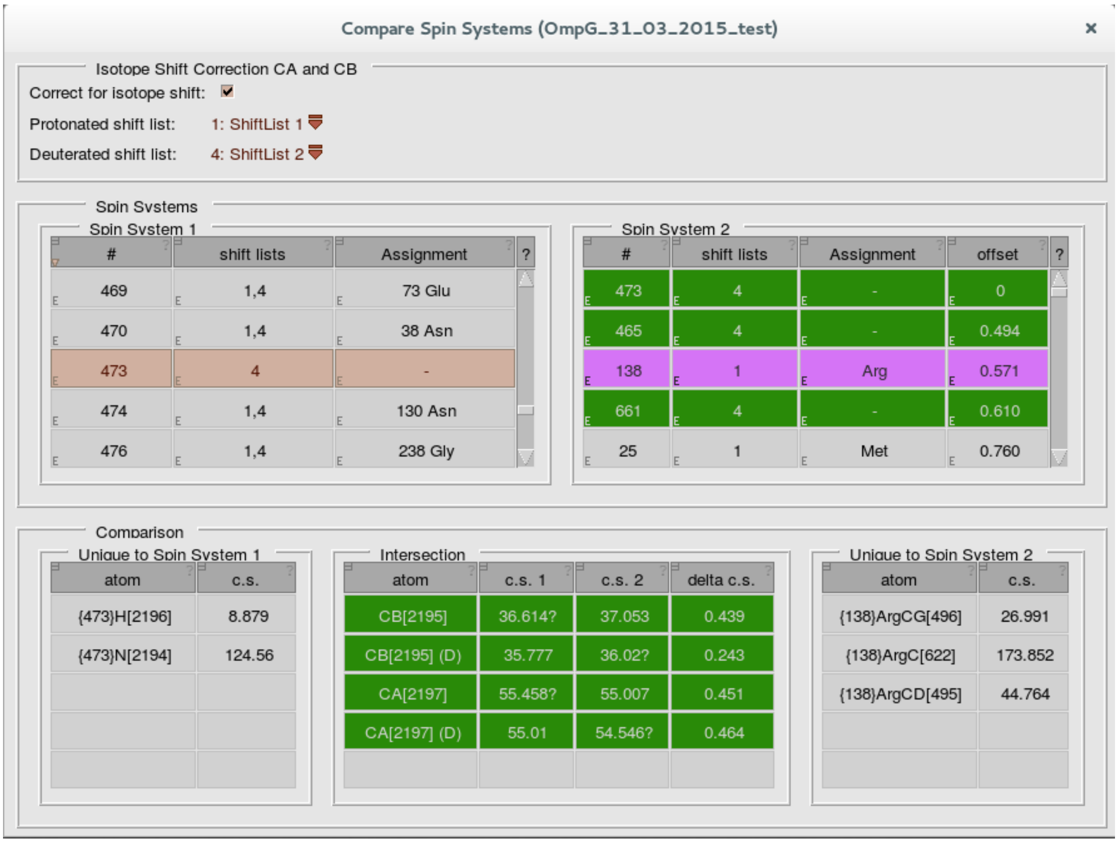


Figure 4.3: Assigned status of residues on the topology of OmpG. Colors correspond to the colors used in figure 5.1 and 5.2: blue labeled residues are assigned. Light blue indicates that the ^{1}H chemical shift of this residue is not known. Often these residues are the first in a connected stretch and their ^{13}C chemical shifts are known from the hcaCBcacoNH, hcoCAcoNH and hCONH peaks in the strip the next residue in the sequence. Residues in red highlight the unassigned residues corresponding to the unassigned spin systems in figure 5.1 and 5.2: three leucines, two threonines, six tryptophans and two prolines. There is only one unassigned proline spin system in figure 5.2, the chemical shifts corresponding to a remaining proline spin system could not be easily determined because of signal overlap.

A CCPNMR Analysis plug-in for comparing spin systems

As mentioned before, the program CCPNMR Analysis was used to do these assignments. During the assignment process described above one will very likely end up with two sets of spin systems, one from the ^{1}H detected data and one from the ^{13}C detected data, because initially the mapping between the two sets of data is not known. It was important within CCPNMR Analysis for ^{1}H- and ^{13}C-detected spectra to be connected to separate shift lists to prevent internally averaging the two shifts into one main shift, which would reduce functionality for all parts of the program that rely on shift matching in some way. To make the process of matching up and merging the two sets of spin systems I wrote a simple CCPN analysis plug-in, see figure 4.4. In it, two spin systems can be compared to one another. As a measure of how comparable the two spin systems are, the root mean square deviation between the corresponding shifts is calculated. If the shifts from protonated and deuterated samples are divided into two different shift lists, a correction based on the values reported by *Maltsev et al.* can be applied. Using this tool to find similar spin systems in the ^{1}H- and ^{13}C-detected datasets can be an alternative to searching for Cα-Cβ cross-peaks in the spectra. This tool could also be useful in other scenarios where spin systems have to be compared. It can be downloaded at https://github.com/jorenretel/compare_spinsystems.



Compare Spin Systems (OmpG_31_03_2015_test)

Isotope Shift Correction CA and CB

Correct for isotope shift: ☑

Protonated shift list: 1: ShiftList 1

Deuterated shift list: 4: ShiftList 2

Spin Systems

Spin System 1

#	shift lists	Assignment
469	1,4	73 Glu
470	1,4	38 Asn
473	4	-
474	1,4	130 Asn
476	1,4	238 Gly

Spin System 2

#	shift lists	Assignment	offset
473	4	-	0
465	4	-	0.494
138	1	Arg	0.571
661	4	-	0.610
25	1	Met	0.760

Comparison

Unique to Spin System 1

atom	c.s.
{473}H[2196]	8.879
{473}N[2194]	124.56

Intersection

atom	c.s. 1	c.s. 2	delta c.s.
CB[2195]	36.614?	37.053	0.439
CB[2195] (D)	35.777	36.02?	0.243
CA[2197]	55.458?	55.007	0.451
CA[2197] (D)	55.01	54.546?	0.464

Unique to Spin System 2

atom	c.s.
{138}ArgCG[496]	26.991
{138}ArgC[622]	173.852
{138}ArgCD[495]	44.764

Figure 4.4: Graphical User Interface of the CCPN Analysis plug-in that helps to compare spin systems to each other. In the tables at the top, the two spin systems that should be compared are selected. The three tables at the bottom show the resonances unique to the first spin system, the resonances that are assigned to the same type of nuclei and the resonances unique to the second spin system respectively. In this case, a spin system created based on the proton detected data (left side) is compared to one that was created using carbon detected data (right side). Discrepancy of values in delta chemical shift in the intersections table are due to the fact that for untyped spin systems an average isotope shifts correction is used where an amino acid specific one is used for the typed spin system on the right.

References

[1] R. Linser, V. Chevelkov, A. Diehl, and B. Reif. "Sensitivity Enhancement Using Paramagnetic Relaxation in MAS Solid-State NMR of Perdeuterated Proteins". *Journal of Magnetic Resonance* 189.2 (Dec. 2007), pp. 209–216. DOI: 10.1016/j.jmr.2007.09.007.

[2] R. Linser. "Side-Chain to Backbone Correlations from Solid-State NMR of Perdeuterated Proteins through Combined Excitation and Long-Range Magnetization Transfers". *Journal of Biomolecular NMR* 51.3 (Aug. 2011), pp. 221–226. DOI: 10.1007/s10858-011-9531-3.

[3] V. Agarwal, K. Faelber, P. Schmieder, and B. Reif. "High-Resolution Double-Quantum Deuterium Magic Angle Spinning Solid-State NMR Spectroscopy of Perdeuterated Proteins". *Journal of the American Chemical Society* 131.1 (Jan. 2009), pp. 2–3. DOI: 10.1021/ja803620r.

[4] D. Lalli, P. Schanda, A. Chowdhury, J. Retel, M. Hiller, V. A. Higman, L. Handel, V. Agarwal, B. Reif, B. van Rossum, Ü. Akbey, and H. Oschkinat. "Three-Dimensional Deuterium-Carbon Correlation Experiments for High-Resolution Solid-State MAS NMR Spectroscopy of Large Proteins". *Journal of Biomolecular NMR* 51.4 (Oct. 2011), pp. 477–485. DOI: 10.1007/s10858-011-9578-1.

[5] R. Verel, M. Ernst, and B. H. Meier. "Adiabatic Dipolar Recoupling in Solid-State NMR: The DREAM Scheme". *Journal of Magnetic Resonance* 150.1 (May 2001), pp. 81–99. DOI: 10.1006/jmre.2001.2310.

[6] E. H. Hardy, R. Verel, and B. H. Meier. "Fast MAS Total Through-Bond Correlation Spectroscopy". *Journal of Magnetic Resonance* 148.2 (Feb. 2001), pp. 459–464. DOI: 10.1006/jmre.2000.2258.

[7] J. Leppert, O. Ohlenschläger, M. Görlach, and R. Ramachandran. "Adiabatic TOBSY in Rotating Solids". *Journal of Biomolecular NMR* 29.2 (June 2004), pp. 167–173. DOI: 10.1023/B:JNMR.0000019248.48726.ff.

[8] V. Agarwal and B. Reif. "Residual Methyl Protonation in Perdeuterated Proteins for Multi-Dimensional Correlation Experiments in MAS Solid-State NMR Spectroscopy". *Journal of Magnetic Resonance* 194.1 (Sept. 2008), pp. 16–24. DOI: 10.1016/j.jmr.2008.05.021.

[9] K.-Y. Huang, A. B. Siemer, and A. E. McDermott. "Homonuclear Mixing Sequences for Perdeuterated Proteins". *Journal of Magnetic Resonance* 208.1 (Jan. 2011), pp. 122–127. DOI: 10.1016/j.jmr.2010.10.015.

[10] G. T. Montelione, B. A. Lyons, S. D. Emerson, and M. Tashiro. "An Efficient Triple Resonance Experiment Using Carbon-13 Isotropic Mixing for Determining Sequence-Specific Resonance Assignments of Isotopically-Enriched Proteins". *Journal of the American Chemical Society* 114.27 (Dec. 1992), pp. 10974–10975. DOI: 10.1021/ja00053a051.

[11] S. Grzesiek, J. Anglister, and A. Bax. "Correlation of Backbone Amide and Aliphatic Side-Chain Resonances in 13C/15N-Enriched Proteins by Isotropic Mixing of 13C Magnetization". *Journal of Magnetic Resonance, Series B* 101.1 (Feb. 1993), pp. 114–119. DOI: 10.1006/jmrb.1993.1019.

[12] S. Asami, P. Schmieder, and B. Reif. "High Resolution 1H-Detected Solid-State NMR Spectroscopy of Protein Aliphatic Resonances: Access to Tertiary Structure Information". *Journal of the American Chemical Society* 132.43 (Nov. 2010), pp. 15133–15135. DOI: 10.1021/ja106170h.

[13] S. Asami and B. Reif. "Assignment Strategies for Aliphatic Protons in the Solid-State in Randomly Protonated Proteins". *Journal of Biomolecular NMR* 52.1 (Dec. 2011), pp. 31–39. DOI: 10.1007/s10858-011-9591-4.

[14] S. Asami, K. Szekely, P. Schanda, B. H. Meier, and B. Reif. "Optimal Degree of Protonation for 1H Detection of Aliphatic Sites in Randomly Deuterated Proteins as a Function of the MAS Frequency". *Journal of Biomolecular NMR* 54.2 (Aug. 2012), pp. 155–168. DOI: 10.1007/s10858-012-9659-9.

[15] V. Tugarinov, V. Kanelis, and L. E. Kay. "Isotope Labeling Strategies for the Study of High-Molecular-Weight Proteins by Solution NMR Spectroscopy". *Nature Protocols* 1.2 (July 2006), pp. 749–754. DOI: 10.1038/nprot.2006.101.

[16] V. Agarwal, A. Diehl, N. Skrynnikov, and B. Reif. "High Resolution 1H Detected 1H,13C Correlation Spectra in MAS Solid-State NMR Using Deuterated Proteins with Selective 1H,2H Isotopic Labeling of Methyl Groups". *Journal of the American Chemical Society* 128.39 (Oct. 2006), pp. 12620–12621. DOI: 10.1021/ja064379m.

[17] R. Linser, B. Bardiaux, V. Higman, U. Fink, and B. Reif. “Structure Calculation from Unambiguous Long-Range Amide and Methyl 1H-1H Distance Restraints for a Microcrystalline Protein with MAS Solid-State NMR Spectroscopy”. *Journal of the American Chemical Society* 133.15 (Apr. 2011), pp. 5905–5912. doi: 10.1021/ja110222h.

[18] M. Huber, S. Hiller, P. Schanda, M. Ernst, A. Böckmann, R. Verel, and B. H. Meier. “A Proton-Detected 4D Solid-State NMR Experiment for Protein Structure Determination”. *ChemPhysChem* 12.5 (Apr. 2011), pp. 915–918. doi: 10.1002/cphc.201100062.

[19] P. E. Hansen. “Isotope Effects in Nuclear Shielding”. *Progress in Nuclear Magnetic Resonance Spectroscopy* 20.3 (1988), pp. 207–255. doi: 10.1016/0079-6565(88)80002-5.

[20] R. A. Venters, B. T. Farmer II, C. A. Fierke, and L. D. Spicer. “Characterizing the Use of Perdeuteration in NMR Studies of Large Proteins:13C,15N and1H Assignments of Human Carbonic Anhydrase II”. *Journal of Molecular Biology* 264.5 (Dec. 1996), pp. 1101–1116. doi: 10.1006/jmbi.1996.0699.

[21] A. S. Maltsev, J. Ying, and A. Bax. “Deuterium Isotope Shifts for Backbone 1H, 15N and 13C Nuclei in Intrinsically Disordered Protein -Synuclein”. *Journal of biomolecular NMR* 54.2 (Oct. 2012), pp. 181–191. doi: 10.1007/s10858-012-9666-x.

[22] P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas. “PRED-TMBB: A Web Server for Predicting the Topology of β-Barrel Outer Membrane Proteins”. *Nucleic Acids Research* 32.suppl 2 (Jan. 2004), W400–W404. doi: 10.1093/nar/gkh417.

[23] B. Liang and L. K. Tamm. “Structure of Outer Membrane Protein G by Solution NMR Spectroscopy”. *Proceedings of the National Academy of Sciences* 104.41 (Sept. 2007), pp. 16140–16145. doi: 10.1073/pnas.0705466104.

Chapter 5

(Semi-) Automatic Assignment of Solid-State NMR spectra

Introduction

As described before the assignment process can be divided in two steps. First, resonances belonging to the same residue are grouped into spin systems. In the ^{13}C-detected set of experiments through-space ^{13}C-^{13}C correlation spectra with short mixing times and NCACX spectra are used for this purpose. In the ^{1}H-detected assignment suite the hCANH, hCOcaNH and hcaCBcaNH fulfill this function. The second step of the assignment process involves the sequence-specific assignment of these spin systems, in which each system is assigned to a specific residue in the protein sequence. This is generally done by evaluating through-space ^{13}C-^{13}C correlation spectra with longer mixing times, NCOCX spectra, NCACX spectra with longer mixing times or the ^{1}H-detected hcoCAcoNH, hCONH and hcaCBcacoNH spectra.

Whereas the first step in this process is often relatively straight-forward, the sequential assignment step is non-trivial and time consuming. Often information from a large set of different spectra has to be combined to come to a solution. In practice, a series of hypotheses are made about two spin systems being a sequential pair in the sequence. Subsequently the chemical shifts of the resonances within the spin system are used (either by looking them up in a resonance list or by setting visual rulers on intra-residual peaks) to search for peaks that support this hypothesis. When spectra with different labeling schemes are used, an extra step is introduced where one determines which peaks are expected in which spectra. This is a repetitive effort that distracts from and confuses the actual sequential assignment. Furthermore, when assigning spectra it is tempting to focus on specific peak patterns for spin systems relevant to the current sub-hypothesis one tries to prove. However, as one pattern leads to another, each new sequential connectivity adds new (and often multiple) hypotheses that have to be tested in parallel with originals; after several steps, the amount of possibilities explodes and one might find himself wandering away from the initial hypothesis that was to be tested. To get a better overview of the information present in a large set of spectra and to easy the process of evaluating alternative hypotheses for the assignment of spin systems, we designed a plug-in for CCPNMR Analysis specifically to help with assignments in solid-state NMR.

CCPN analysis already has a graphical tool that interactively helps with back-bone walks [1]. This tool works very well with the typical spectra in solution NMR. However, it is not designed with the combinations of experiments typically used in solid state NMR in mind (with the exception of the newer proton detected experiments, which yield spectra very similar to the solution NMR spectra used for backbone walks). With the existing tool a strip from a "query" spectrum is selected and subsequently the program suggests multiple strips from a "match" spectrum that could belong to the neighboring residue. Subsequently, when some strips are placed in order, fitting sequence fragments are suggested along with their likelihood. Unfortunately, not all types of spectra fit well in this approach, for instance, typical solid state NMR experiments, where magnetization between atoms of two residues is transferred through-space, instead of through a specific path over the backbone. Peak patterns in spectra that have more than one dimension in which more than one atom site is measured, such as 2D ^{13}C-^{13}C correlations, can not be easily visualized as strips. Furthermore, for these types of spectra, it is mostly not possible to differentiate between "query" and "match" spectra. Therefore, a similar tool with less restrictions to experiment types would be very helpful in the sequential assignment process of solid state NMR spectra.

The new tool calculates the expected peak pattern for each neighboring pair in the sequence based on the magnetization transfer pathway of the experiment and the labeling scheme of the sample. These patterns are checked against the corresponding peak lists using the chemical shifts from every possible combination of spin systems that can be assigned to this pair. Subsequently, this information is used to search for a globally optimal sequential assignment using a combined Monte Carlo / simulated-annealing procedure in a way similar to the algorithm described by Tycko and coworkers for uniformly labeled spectra [2][3]. The principle is based on a simple scoring mechanism: the more (experimental) peaks that connect two spin systems show up and the better these peaks fit to the chemical shifts of the resonances within those spin systems, the more likely the hypothesized connection is real. The algorithm can be used in combination with virtually any labeling scheme, magnetization transfer pathway, through-bond or through-space, and is not limited by the dimensionality of the correlation spectra.

Many different algorithms for automatic sequential assignment have been published before but are never used [4]. Many programs require of the input specifically formatted tables. This discourages users from even trying out a program if expectations are already low. A general problem is that most algorithms are implemented as stand-alone applications. This often invokes a complicated work-flow in which several peak and shift lists from the program that is used to do the analysis of the NMR spectra have to be exported and converted to specially formatted tables that are expected by the routine. Afterwards often the results have be imported back into the analysis program so they can be carefully judged. This is not necessarily bad if it has to be done only once. However, for complex assignment projects, with a work-flow in which new data and assignments are added incrementally, this becomes cumbersome. Also the type of data these stand-alone applications output is somewhat problematic. Although mostly an output is generated containing some confidence measure for the individual assignments, it is hard to import more than the consensus assignment back into the analysis program. Ideally one would like to compare the information that supports different assignment alternatives on the fly, even the ones that were never chosen by the algorithm, and cherry-pick the assignments that are believed to be correct.

Integrating the algorithm within a program that is used for spectral analysis (in this case CCPN analysis) allows for a natural interaction with the data and lowers the energy barrier for a user to try out the algorithm. All relevant information can be automatically fetched from the Analysis project. All other information specific to the optimization procedure is configured in the Graphical User Interface (GUI) of the plug-in. The output is also shown in the GUI, to allow for quick comparison of alternative assignments. Peaks can be directly navigated to in the spectra and assignments can be transferred to the project one by one or all at once. Additionally, CCPN analysis has great support for labeled samples. To correctly calculate the expected peak pattern, integration of labeling schemes is essential. If a simple input table would be used to define which nuclei are labeled and which ones are not (instead of a data model, allowing more complexity, like CCPN provides), often detailed information about the exact makeup of a labeling scheme is lost. This is especially true when labeled samples are used that are composed of different isotopomers, like those based on 1,3- and 2-glycerol used in this studies, or glucose labeling [5]. In these cases the expected peak pattern can only be calculated by correctly evaluating the simultaneous labeling of nuclei in individual isotopomers. Otherwise excellent tools that do not support this can not be used if the dataset includes these type of labelling schemes.

Two algorithms worth mentioning explicitly, GAMES_ASSIGN and solid-state FLYA, have been respectively created especially and customized especially for assignment of solid-state NMR spectra [6][7][8]. These algorithms do not require the creation of spin systems as a starting point for the assignment. This can be advantageous since wrongly configured spin systems will inevitably lead to wrong assignments. However they both suffer from some of the general drawbacks described before, some of which could be solved by a better integration with a spectral analysis program like CCPN Analysis. GAMES_ASSIGN does not support ^{1}H-detected spectra and neither of them seems to support complex labeling schemes in a straight-forward matter.

Description of the algorithm

In figure 5.2 an overview is given of the different steps in the algorithm, which are discussed in more detail in the following paragraphs.

1 Input of data

In the first step the input to the algorithm is gathered. In principle most information can be automatically loaded from the CCPN project:

- primary sequence
- labeling schemes
- experiment types describing the magnetization transfer pathways
- peak lists
- spin systems
- shift lists
- assignment tolerances
- previously made sequential, tentative, and non-sequential amino acid type assignments

Which data is accessed by the algorithm can be configured by the user. This can be useful when the algorithm is used to confirm manually made assignments in an unbiased way without changing the CCPN analysis project.

Spin systems in CCPN Analysis can basically have one of to following five levels of assignment: 1) The most definite form of assignment is of course when a spin system is sequentially assigned to a specific residue in the sequence. 2) One level of assignment lower, a spin system can be assigned "tentatively" to multiple residues but it is not known which of those residues is the correct one. 3) Then there are spin systems that are residue typed, but no information about sequential assignment is present. 4) Also multiple residue types can be set for a spin system. This option is not standardly accessible in the GUI of CCPN Analysis, but is present in the API and called ResidueTypeProbs. An additional plug-in is provided to set this property. This feature is useful because often a residue type can be narrowed down to two very closely related residue types like asparagine and aspartic acid. Therefore, in our opinion, this feature should become standardly accessible in the CCPN Analysis GUI. 5) The spin systems with the lowest level of assignment are those for which not any form of sequential assignment nor residue type information is available.

All levels of assignment may be used by the algorithm, if wanted. If residue type information is not available or not used, the residue typing algorithm already included in CCPN will be used to classify the spin systems to residue types. Because in most cases the classification is not definite, the user can set a threshold score above which residue types will considered a possibility. There is one more type of assignments that can be, namely the assignment of peak dimensions to specific resonances. If this information is chosen to be used, the only possible assignment of a peak dimension that is considered is the present assignment.

Resonances within spin systems should be assigned to an atom type, Cα for example. If a resonance does not have an atom type assignment it is not possible to map it to a dimension of an expected peak and therefore it will be ignored. Spin systems without any resonances will be ignored as they are generally used in CCPN Analysis as placeholders for unassigned residues. A subset of spectra can be selected using the GUI. In the Analysis project, each spectrum should have peak lists, a labelling scheme and an experiment type describing the magnetization transfer steps and dimensions of the experiment.

The spectra should be peak picked. In many experiments, most notably in experiments with a though-space step like ^{13}C-^{13}C or NCACX spectra, signal sets that originate from intra-residual, sequentially and long-distance correlations between nuclei are mixed. Although the program simulates and finds intra-residual peaks, the optimization procedure naturally only relies on peaks that form sequential links between two spin systems. To prevent the algorithm from misinterpreting intra-residual peaks as sequential peaks, these peaks should either not be picked in the spectra that are to be used by the algorithm or all their dimensions should be properly assigned to resonances from the same spin system, tipping the program off that the peak should not be interpreted as a possible sequential peak. Because grouping resonances into spin systems is a prerequisite for the algorithm, it is in principle already known which peaks are intra-residual. Assigning intra-residual peaks to known spin systems can be performed using a short run of the algorithm and subsequently letting it assign all intra-residual peaks.

Furthermore the molecular chain (the primary sequence) has to be selected. Optionally, parts of the sequence that should not be considered for assignment can be entered. This last option can be used if it is clear that parts of the molecule can not be seen because of dynamics or for instance incomplete back-exchange of protons.

2 Evaluate possible mapping between spin systems and residues

On the basis of the different levels of assignment described above and which of that information should be used, for each spin system a set of possible residue assignments is created. Based on these sets, it can already be determined for each spin system with which other spin systems it could in principle exchange sequential assignments (under the condition that both spin systems are assigned to residues in the intersection of the two sets). This information is used during the Monte Carlo procedure to pick which two spin systems to exchange. Also, at this stage “joker” spin systems are introduced to make sure that always a spin system can be assigned to every residue. This is important since the Monte Carlo procedure was designed to select two spin systems and exchange their residue assignments rather than the other way around. For this reason all residues should at any stage of the optimization procedure have a spin system assigned to them, because once that would not be case no spin system will ever be assigned to it. Joker spin systems have a residue type assignment. For each amino acid, as many joker spin systems are generated as there are residues in the sequence of that type.

3 Predict peak pattern

The intra-residual and sequential peak patterns are predicted based on the molecular topology, the experiment graph and the isotope labeling scheme. Each spectrum in a CCPN project is connected to an experiment type. Normally, the user is prompted to set the experiment type for each new spectrum that is loaded into the CCPN project. When this was not done, it can be set afterwards and is essential for this algorithm to work. In the CCPN data model each experiment type is connected to an experiment graph. This graph describes which magnetization transfers happen during the experiment and how parts of the experiment map to dimensions in the spectrum. For each magnetization transfer information is present about which types of nuclei take part, whether the transfer is through-space or through-bond and whether a transfer from a nucleus to itself can happen. The specified atom types are not restricted to just isotopes but can be more specific, only aromatic carbons for instance. Together with the molecular topology, the graph can be walked recursively to generate a list of expected peaks for virtually any correlation experiment. This list is then filtered by the labeling scheme. For each peak the co-labeling fraction over all nuclei on the magnetization transfer pathway is calculated. Only if the co-labeling fraction exceeds a user defined variable the peak is retained. By default this minimal co-labeling fraction is 0.1.

4 Match predicted and experimental peaks

Now that a peak pattern is predicted for each spectrum, this can be matched with the peaks in the peak lists corresponding to those spectra. The positions of the expected peaks can be determined using the chemical shifts assigned to resonances in the spin systems. Of course we don't know yet at this moment which spin systems is in which position on the sequence as that is the purpose of this algorithm. Therefor the possible mapping between residues and spin systems determined in step 2 is used. For every two sequentially neighboring residues A and B, all combinations of spin systems A′ and B′ are used to search the peak lists, see figure 5.1 A. The position of the expected sequential peaks will be different for each combination of A′ and B′. To find out which peaks in the peaklists match to the expected pattern a chemical shift tolerance is used in each dimension of the spectrum. For each dimension of each spectrum a assignment tolerance can be set in CCPN Analysis, and those tolerances are used here as well.

The more of the expected sequential cross peaks are present in the peak lists, the likelier it is that a specific combination A′-B′ is indeed a sequential pair. It is also important how well the actual peak positions fit the expected positions. To do this, for each matched peak a simple function is used that assigns an energy between 0 and -1 depending on the difference between cross-peak position and expected peak position. That value is then multiplied with the square of the number of resonances involved with the expected peak to acknowledge that peaks in higher dimensional spectra carry more weight. This number is mostly equal to the number of dimensions, except for partially diagonal peaks. For instance the $N_{\text{resonances}}$ for a diagonal peak in a NCOCX spectrum would be 2 instead of

3, since it contributes less prove for a sequential connection between two spin systems than an off-diagonal peak. Furthermore the energy is normalized by the symmetry of the spectrum the peak is in. A 2D ^{13}C-^{13}C correlation for instance has two sets of crosspeaks on each side of the diagonal, making the symmetry 2. All together the energy contribution of one peak can then be expressed as:

$$E_{peak} = max(-1, \sum_{n=1}^{N_d} (\frac{\Delta\delta_n}{t_n}^2 - 1)\frac{1}{N_d(1-k^2)})\frac{N^2_{resonances}}{symmetry}$$

Where N_d is the number of dimensions, $\Delta\delta_n$ is the difference between the shift of the peak and the shift from the shiftlist in the n-th dimension, t_n is the tolerance in the n-th dimension and k is the fraction of the tolerance window that has a flat bottom. This last value is set by default on 0.4. The flat bottom was introduced to prevent over-interpreting small differences between the peak and the expected position. The energy then goes up gradually and becomes 0 for peaks that are all the way in the corner of the tolerance window, see figure 5.1 B.

As discussed before chemical shifts can differ between spectra depending on the sample and experimental conditions such as temperature and isotope shifts. If these kind of differences are present it is important that spectra are connected to different shift lists. The correct shift list is then used by this algorithm to perform this matching step.

Besides from sequential cross-peaks, also intra-residual cross-peaks are matched. They do not play a role during the optimization of the sequential assignment as they carry no sequential information. However, it is useful to collect these peaks as well, since they can be used for a quick assignment of peaks in new spectra to already known spin systems.

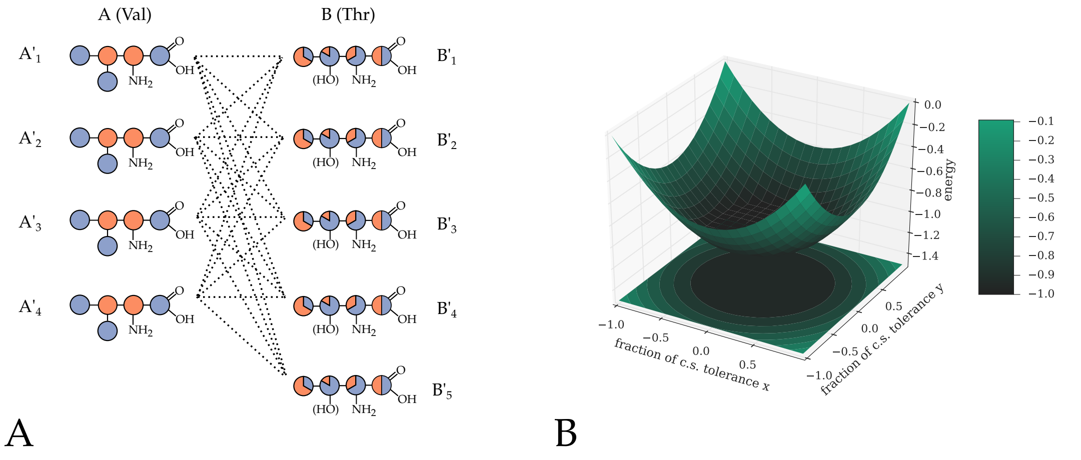


Figure 5.1: The expected peaks can be matched to picked peaks in the spectra. Therefor the chemical shifts of all combinations of spin systems A', B' that can be assigned to sequential residues A and B can be used to predict the location of the peaks (A). How well the real fits the predicted peak location is scored by a flat bottom scoring function (B).

5 Temporarily remove a fraction of the cross peaks

The optimization procedure that follows is repeated multiple times to create an ensemble of possible sequential assignment, that can be later on compared to one another. If the assignment of a spin system to a residue in the sequence stays the same with different subsets of the peaks used, this might be a good indication that this assignment is correct. In each run a new randomly selected part of the data will be removed before the optimization starts. This is optional as the fraction can be set to 0. Also without removing cross-peaks the result of the optimization will likely be a little different every time depending on how well defined the energy minimum is.

6 Generate a random starting assignment

A random assignment is generated that is consistent with the possible mapping between spin systems and residues determined in step 2. Every residue is assigned to one spin system. This can also be a joker spin system. Not every spin system necessarily has a residue assignment, because the total of real and joker spin systems is larger than the number of residues in the sequence. A spin system is never assigned to more than one residue at the same time.

7 Optimization of the sequential assignment using a simulated annealing / Monte Carlo procedure

For each step in the Monte Carlo procedure two spin systems are selected to exchange residue assignments. In practice this is done by first randomly choosing one spin system, independent on whether it is assigned to a residue or not. Then from the more selective list of spin systems this spin system could ever exchange with, as determined in step 2, randomly one other spin system is chosen. Before the change is attempted, a check is performed to assure that the change would not produce an assignment that is inconsistent with the possible mapping between spin systems and residues. Now the change in energy can be calculated corresponding to the attempt. Therefor the energy of the individual links between the two spin systems and the current neighboring spin systems in the sequence has to be calculated. If both spin systems were assigned to residues this would be 4 links both in the old and the new situation. The energy of one link can be defined as:

$$E_{link} = \sum_{n=1}^{N_p} \frac{E_{peak,n}}{degeneracy_n} N_{resonances,total}$$

where N_p is the number of peaks. $E_{peak,\,n}$ is the peak score of the n-th peak determined in step 4. The degeneracy is the amount of different assignments the peak has at the current point of the minimization. The peak energies are normalized by this value, because if a peak already has a lot of assignments it is not very relevant in proving this link between two spin systems is correct. $N_{resonances\ total}$ is the total amount of unique resonances playing a role in the assignments of all peaks. The difference in energy is now simply:

$$\Delta E = E_{links,new} - E_{links,old}$$

The change will be accepted when the Monte Carlo criterium is fulfilled:

$$e^{\frac{-\Delta E}{kT}} \geq random(0 \rightarrow 1)$$

During the procedure the temperature is lowered in according to an annealing schedule, making it harder for assignment changes that increase the energy to be accepted in later stages of the optimization.

Steps 5, 6 and 7 are repeated a chosen number of times to generate an ensemble of solutions that can be compared in the graphical user interface.

The graphical user interface

A relatively simple GUI was created to choose the data to be used, configure the algorithm and display the results. In the first tab, see figure 5.3 A, spectra and corresponding peak lists and whether to use the connected labeling scheme are selected. In the second tab (figure 5.3 B), the parameters discussed in step 1 can be set. Also the residue range can be set here, as it can be useful to exclude parts of the sequence from the optimization if it is known that these parts do not give rise to peaks in the spectra. Furthermore, the cooling regime and the amount of steps per temperature point and the total amount of runs can be configured. When the algorithm is started, the energy after each temperature step is shown in a plot for all annealing runs.

After all annealing runs are completed the results will be shown, see 5.4 C. The user can walk through the sequence and see a subsequence of five residues at a time. For each residue the spin system selected in a given run is shown. The five tables below summarize the overall outcome of all runs. For each residue all spin systems are shown that could be assigned to this residue consistent with the mapping performed in step 2. For each spin system the percentage of runs is shown in which it was selected as the assignment to the residue. When clicking on one of the “links” buttons in between the buttons representing the residues, all peaks will be shown that were found to connect the two selected spin systems. These are the found peaks on which the algorithm based its decisions. When clicking on the residue button itself all found intra-residual peaks will be shown for the selected spin system. Another row of residue buttons can be used to configure a self defined assignment, independent of the annealing procedure, and check the information supporting that assignment. The advantage of having the assignment procedure integrated within CCPN Analysis is that it is possible to automatically navigate to a selected peak in the table. This makes checking by eye whether the peak pattern is indeed good prove for the assignment suggested by the algorithm easy and fast. It is also possible to automatically navigate to the expected peak positions of peaks that were not found. This is very important to form an opinion about the correctness of the assignment. In this way it

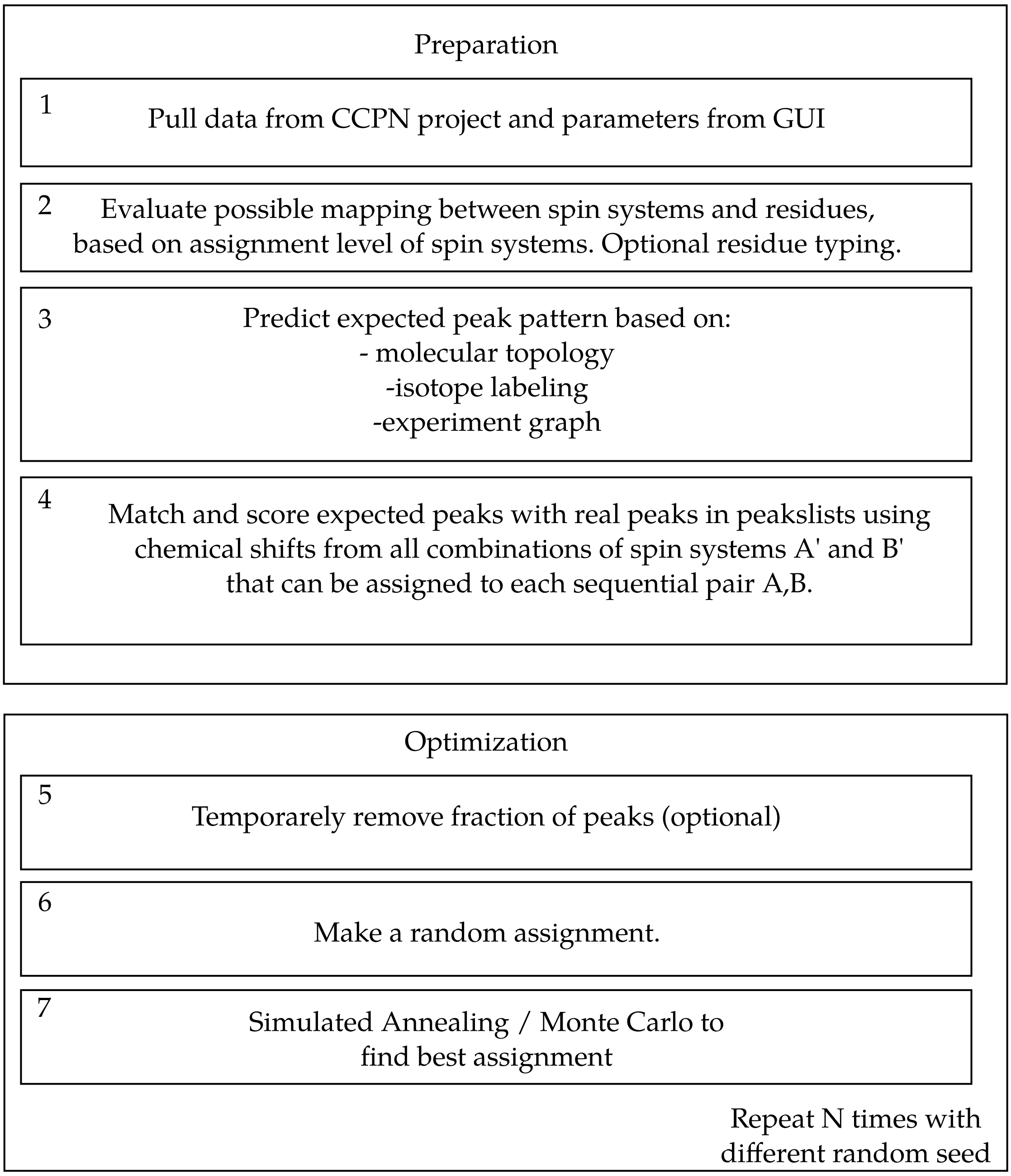

Preparation
1
Pull data from CCPN project and parameters from GUI
2
Evaluate possible mapping between spin systems and residues,
based on assignment level of spin systems. Optional residue typing.
3
Predict expected peak pattern based on:
- molecular topology
-isotope labeling
-experiment graph
4
Match and score expected peaks with real peaks in peakslists using
chemical shifts from all combinations of spin systems A' and B'
that can be assigned to each sequential pair A,B.
Optimization
5
Temporarely remove fraction of peaks (optional)
6
Make a random assignment.
7
Simulated Annealing / Monte Carlo to
find best assignment
Repeat N times with
different random seed

Figure 5.2: Steps in the (semi-) automatic assignment algorithm. The preparation stage is executed only once. The Optimization stage can be repeated many times to obtain a set of different solutions on which statistics are based and that can be compared to one another later.

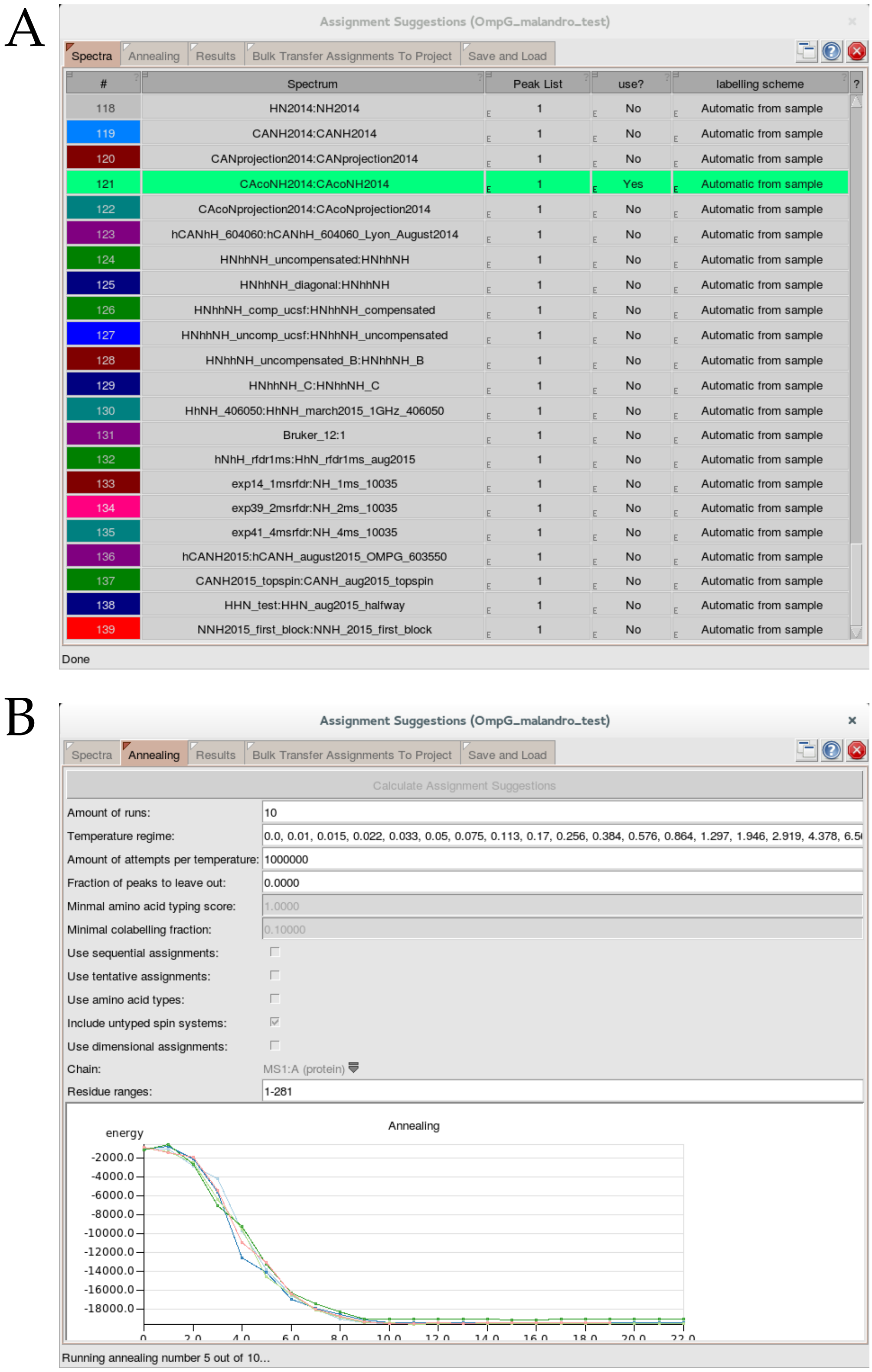
is possible to check whether the peak is really absent, or it was just not picked. If the user agrees with certain assignments, they can be transferred to the CCPN Analysis project. There are basically two types of assignment to be considered here. Assignment of spin systems to residues and assignment of resonances to peak dimensions. Both can be done independently. None of the suggested assignment are transferred to the project just by running the optimization procedure, as this would change the CCPN Analysis project, possibly against the will of the user. This would be very hard in an import-export based routine, as the results already have to be imported (thereby changing the CCPN Analysis project), before they can be evaluated. With the plug-in presented here, assignments the user agrees on can be cherry-picked and transferred to the project individually. If the user wants to transfer the assignments to the project in bulk anyway, this is possible in the last tab, see figure 5.4 B. Because every run of the annealing generates a different possible sequential assignment, the user has to choose which one to the project. It is also possible to only transfer assignment for those spin systems that were selected in a certain threshold percentage of all runs. This threshold has to be set higher than 50% as it is otherwise unclear which spin system to choose if there could be two spin systems exceeding the threshold otherwise.

Implementation details

The plug-in for analysis was written in python, making extensive use of the Python API of CCPN analysis [9]. Parts of the code that needed to be executed faster to make a lot of Monte Carlo attempts in a reasonable amount of time was written in Cython. Cython is used to generate Python extensions that are compiled to C, which in turn is compiled to byte code, making the execution a lot faster. [10] The pseudo random number generator used is a Mersenne twister [11]. At first the linear congruential generator from the c standard library was used. But this showed, as widely known, not to give random enough numbers and thereby skewed the results. The python mersenne twister was re-implemented in Cython by Josh Ayers [12].

Performance of the algorithm

To evaluate the algorithm, automatically generated assignments were compared with the manual assignments that were previously made for the α-Spectrin SH3 Domain and the Yersinia enterocolitica adhesin A (YadA) and OmpG. For SH3, two different tests were ran. First, with a sub-set of carbon detected spectra used for the original assignment and structure determination [13][14]. And second, with a set of proton-detected spectra, recorded at 40 kHz MAS more recently [15]. For YadA a set of ^{13}C-detected spectra of a uniformly labeled sample was avaible [16][17]. The original assignments where used to generate spin systems as this step is not part of the algorithm. Also intra-residual peaks were assigned. For OmpG, both ^{1}H- and ^{13}C-detected spectra were used in conjunction.



A
Assignment Suggestions (OmpG_malandro_test)
Spectra
Annealing
Results
Bulk Transfer Assignments To Project
Save and Load

#	Spectrum	Peak List	use?	labelling scheme
118	HN2014:NH2014	1	No	Automatic from sample
119	CANH2014:CANH2014	1	No	Automatic from sample
120	CANprojection2014:CANprojection2014	1	No	Automatic from sample
121	CAcoNH2014:CAcoNH2014	1	Yes	Automatic from sample
122	CAcoNprojection2014:CAcoNprojection2014	1	No	Automatic from sample
123	hCANhH_604060:hCANhH_604060_Lyon_August2014	1	No	Automatic from sample
124	HNhhNH_uncompensated:HNhhNH	1	No	Automatic from sample
125	HNhhNH_diagonal:HNhhNH	1	No	Automatic from sample
126	HNhhNH_comp_ucsf:HNhhNH_compensated	1	No	Automatic from sample
127	HNhhNH_uncomp_ucsf:HNhhNH_uncompensated	1	No	Automatic from sample
128	HNhhNH_uncompensated_B:HNhhNH_B	1	No	Automatic from sample
129	HNhhNH_C:HNhhNH_C	1	No	Automatic from sample
130	HhNH_406050:HhNH_march2015_1GHz_406050	1	No	Automatic from sample
131	Bruker_12:1	1	No	Automatic from sample
132	hNhH_rfdr1ms:HhN_rfdr1ms_aug2015	1	No	Automatic from sample
133	exp14_1msrfdr:NH_1ms_10035	1	No	Automatic from sample
134	exp39_2msrfdr:NH_2ms_10035	1	No	Automatic from sample
135	exp41_4msrfdr:NH_4ms_10035	1	No	Automatic from sample
136	hCANH2015:hCANH_august2015_OMPG_603550	1	No	Automatic from sample
137	CANH2015_topspin:CANH_aug2015_topspin	1	No	Automatic from sample
138	HHN_test:HHN_aug2015_halfway	1	No	Automatic from sample
139	NNH2015_first_block:NNH_2015_first_block	1	No	Automatic from sample

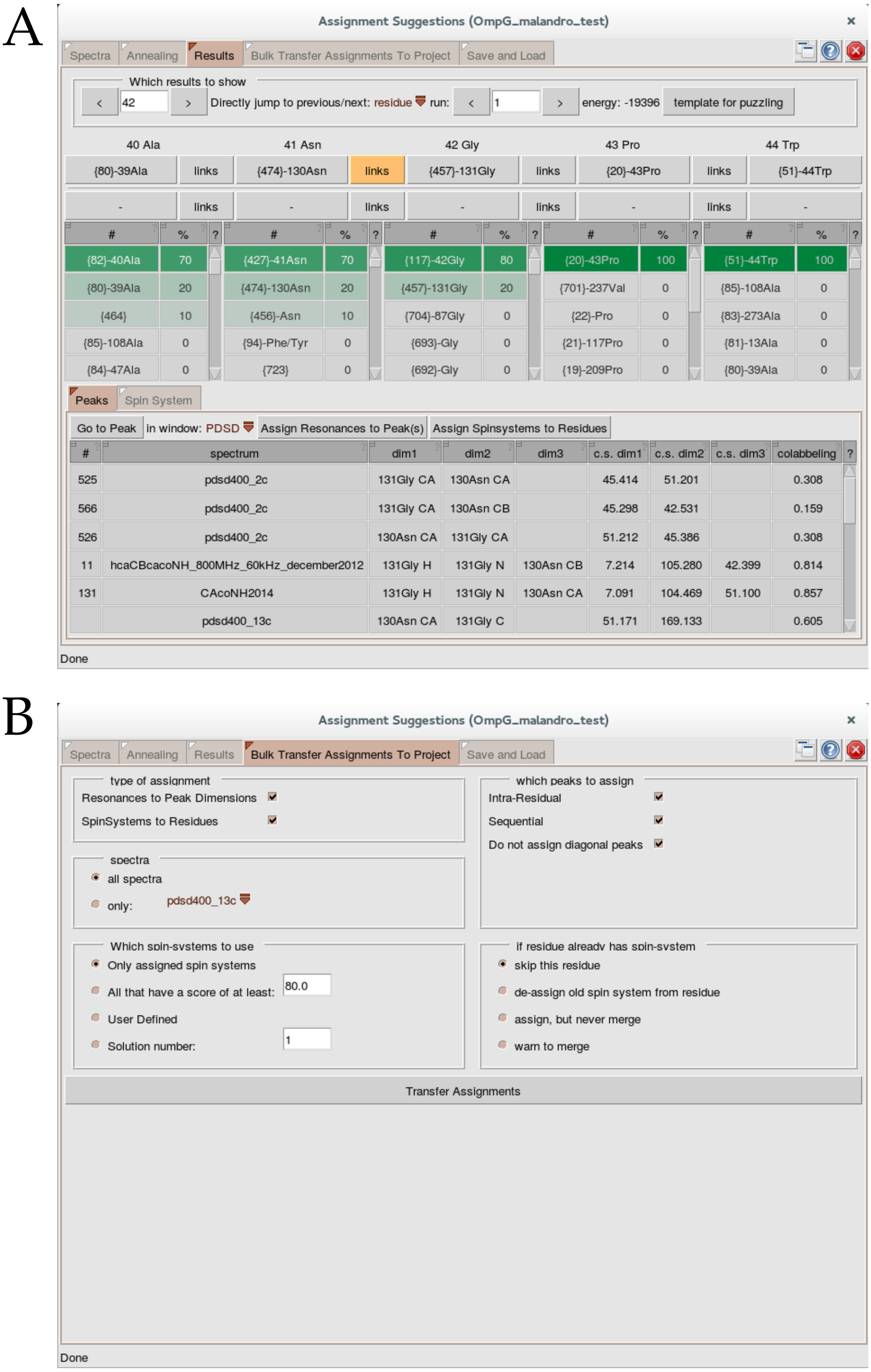
Done

B
Assignment Suggestions (OmpG_malandro_test)
Spectra
Annealing
Results
Bulk Transfer Assignments To Project
Save and Load
Calculate Assignment Suggestions

Amount of runs:	10
Temperature regime:	0.0, 0.01, 0.015, 0.022, 0.033, 0.05, 0.075, 0.113, 0.17, 0.256, 0.384, 0.576, 0.864, 1.297, 1.946, 2.919, 4.378, 6.5
Amount of attempts per temperature:	1000000
Fraction of peaks to leave out:	0.0000
Minmal amino acid typing score:	1.0000
Minimal colabelling fraction:	0.10000
Use sequential assignments:	□
Use tentative assignments:	□
Use amino acid types:	□
Include untyped spin systems:	☑
Use dimensional assignments:	□
Chain:	MS1:A (protein)
Residue ranges:	1-281

Running annealing number 5 out of 10...

Figure 5.3: Graphical User Interface of the (semi-)automatic assignment algorithm. A) a subset of spectra can be selected to be used by the routine. B) a number of settings can be configured controlling which information in the CCPN project is used by the algorithm. Also parameters controlling the annealing process are set here. The graph at the bottom shows the progress of the annealing procedure.


A
Assignment Suggestions (OmpG_malandro_test)
Spectra
Annealing
Results
Bulk Transfer Assignments To Project
Save and Load
B

Which results to show

< 42 > Directly jump to previous/next: residue run: < 1 > energy: -19396 template for puzzling

40 Ala		41 Asn		42 Gly		43 Pro		44 Trp
{80}-39Ala	links	{474}-130Asn	links	{457}-131Gly	links	{20}-43Pro	links	{51}-44Trp
-	links	-	links	-	links	-	links	-

#	%	#	%	#	%	#	%	#	%
{82}-40Ala	70	{427}-41Asn	70	{117}-42Gly	80	{20}-43Pro	100	{51}-44Trp	100
{80}-39Ala	20	{474}-130Asn	20	{457}-131Gly	20	{701}-237Val	0	{85}-108Ala	0
{464}	10	{456}-Asn	10	{704}-87Gly	0	{22}-Pro	0	{83}-273Ala	0
{85}-108Ala	0	{94}-Phe/Tyr	0	{693}-Gly	0	{21}-117Pro	0	{81}-13Ala	0
{84}-47Ala	0	{723}	0	{692}-Gly	0	{19}-209Pro	0	{80}-39Ala	0

Peaks | Spin System

Go to Peak in window: PDSD Assign Resonances to Peak(s) Assign Spinsystems to Residues

#	spectrum	dim1	dim2	dim3	c.s. dim1	c.s. dim2	c.s. dim3	colabbeling
525	pdsd400_2c	131Gly CA	130Asn CA		45.414	51.201		0.308
566	pdsd400_2c	131Gly CA	130Asn CB		45.298	42.531		0.159
526	pdsd400_2c	130Asn CA	131Gly CA		51.212	45.386		0.308
11	hcaCBcacoNH_800MHz_60kHz_december2012	131Gly H	131Gly N	130Asn CB	7.214	105.280	42.399	0.814
131	CAcoNH2014	131Gly H	131Gly N	130Asn CA	7.091	104.469	51.100	0.857
	pdsd400_13c	130Asn CA	131Gly C		51.171	169.133		0.605

Done

Assignment Suggestions (OmpG_malandro_test)

type of assignment
- Resonances to Peak Dimensions ☑
- SpinSystems to Residues ☑

spectra
- ● all spectra
- ○ only: pdsd400_13c

Which spin-systems to use
- ● Only assigned spin systems
- ○ All that have a score of at least: 80.0
- ○ User Defined
- ○ Solution number: 1

which peaks to assign
- Intra-Residual ☑
- Sequential ☑
- Do not assign diagonal peaks ☑

If residue already has spin-system
- ● skip this residue
- ○ de-assign old spin system from residue
- ○ assign, but never merge
- ○ warn to merge

Transfer Assignments

Done

Figure 5.4: A) The results are shown in 5 tables, representing 5 consecutive residues in the sequence. In each table all spin systems that can be assigned to that particular residue are listed. When selecting two spin systems for two sequential residues, all peaks that connect these spin systems are listed in the table at the bottom. Assignments can be inspected here and individually transferred to the project. B) Assignments can also be transferred in bulk to the project. In order to do so, the user should indicate which assignments exactly as multiple annealing runs were performed. One of the possibilities is to only assign those spin systems that are assigned in a threshold fraction of all annealing runs.

SH3 ^{13}C-detected spectra

To test the performance of the algorithm on spectra of specifically labeled samples, 5 spectra of SH3 were used: 2D ^{13}C-^{13}C DARR spectra with a mixing time of 300 ms and NCOCX spectra with a mixing time of 50 ms of both 2- and 1,3-glycerol labeled samples and a NCACX with a mixing time of 200 ms of the 2-glycerol labeled sample. Spectra were peak picked automatically just above the noise, where the diagonal was excluded. In the 1,3 glycerol 2D ^{13}C-^{13}C correlation spectrum 168 out of 512 peaks were assigned as intra-residual. In the 2-glycerol 2D ^{13}C-^{13}C correlation spectrum these were 91 out of 449 peaks and in the NCACX 53 out of 271 peaks. In the two NCOCX spectra, there are no purely intra-residual peaks since the backbone nitrogen of residue i is correlated to the carbonyl of residue i-1. The tolerances of all ^{13}C dimensions was set to 0.3 and 0.4 for all ^{15}N dimensions. All spin-systems were typed automatically as part of the procedure, were the minimal type score was set to 1%. The system converged in 22 temperature steps with 100,000 Monte Carlo attempts per temperature point. For all but the first 5 residues, residues 47, 48 and 62 a unique spin system was assigned, see figure 5.5A. All unassigned residues were also unassigned in the original assignment except for residue 62, of which only the backbone nitrogen was defined in the original assignment.

SH3 ^{1}H-detected spectra

From the ^{1}H-detected spectra at 40 kHz MAS a HNCO and a HNcoCA were used as input to the algorithm, since these spectra contain sequential cross peaks. Spin systems were generated containing HN, N, CO, CA and CB resonances, by evaluating HNCA, HNcaCB and HNCO spectra. Automated peak picking yielded 52 peaks in the HNCO and 118 peaks in the HNcoCA. The higher than expected amount of peaks in the HNcoCA can be explained by the presence of the CA$_{\mathrm{i}}$ peak in a lot of strips where only a CA$_{\mathrm{i\text{-}1}}$ peak is expected. No repicking was performed to change this situation. Possible amino acid types were determined during the procedure as mentioned before, with a minimal type score of 1%. The algorithm already gives good results with 22 temperature steps and 100,000 attempts per step. However, when the amount of steps was increased to 1000,000 a bigger amount of the runs found the same final energy. This can partially be explained by the fact that the search-space is bigger due to less exclusive residue typing than is possible with ^{13}C-detected data. Now spin systems only contain CA and CB carbon resonances that are relevant for residue typing, in contrast to the more fully configured spin-systems containing more side-chain carbon resonances in the previous example. For all residues the most frequent chosen assignment was the one that agrees with previously made manual assignments, except for proline 20, see figure 5.5B. On this residue a joker spin system was placed in more than 90% of the runs, meaning that the algorithm could not assign it. No connecting peaks could be found to Arginine 21.

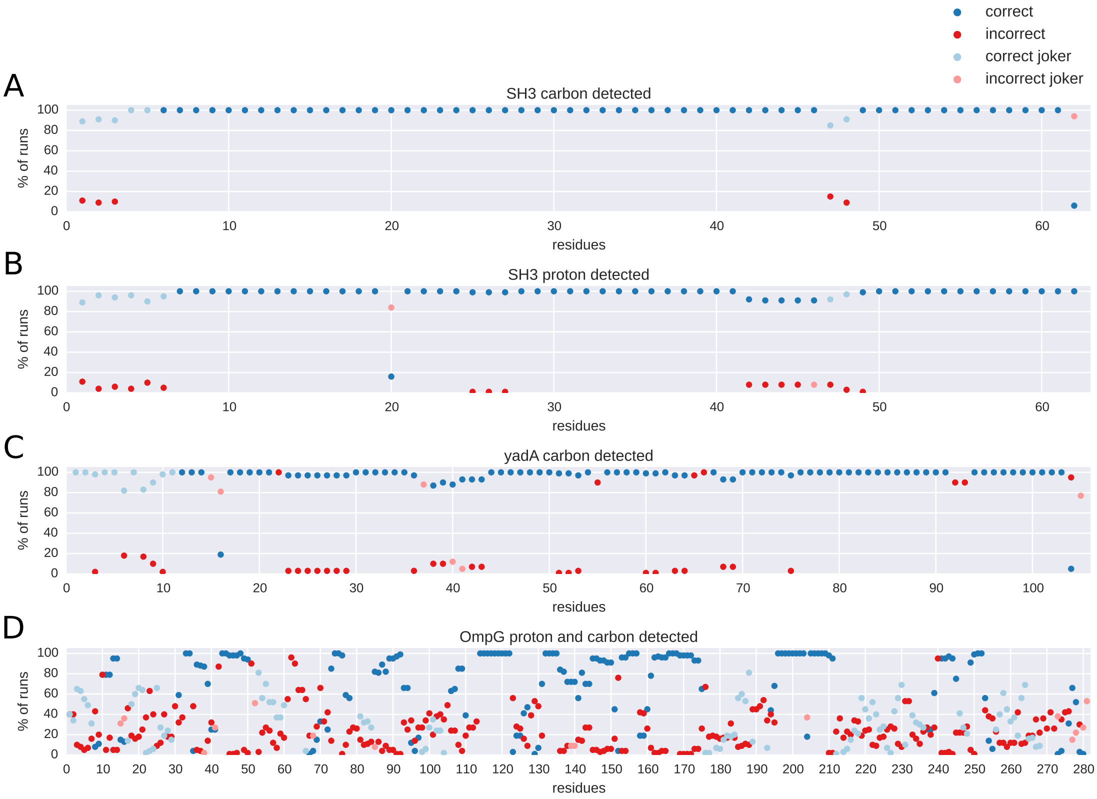


Figure 5.5: Correctness of proposed sequential assignment of residues in three different proteins. The y-axis corresponds to the percentage of the ensemble of solutions in which a certain assignment was chosen. Assignments corresponding to previously made manual assignments are shown in blue. In red the most selected assignment is shown that did not correspond to the manual assignment. Light colors correspond to joker spin systems. A correctly placed joker is a joker placed on residue that was not assigned manually either.

YadA ^{13}C-detected spectra

To test whether the algorithm was still of use for more challenging systems, YadA was used. Only two spectra were used for this optimization: a 2D ^{13}C-^{13}C correlation spectrum, and a NCOCX, both recorded with a DARR mixing period of 200 ms and on a uniformly ^{13}C, ^{15}N labeled sample. As with SH3, spin-systems were created using the published chemical shifts and the intra-residual peaks in the 2D carbon-carbon correlation were assigned to spin-systems by doing a short run of the algorithm. Two alanine spin-systems (corresponding to residues 82 and 88) scored very low (less than 0.2%) for alanines in the residue-typing procedure as their Cβ shifts were slightly more downfield than expected. This was a clear case where human intervention was needed and these two spin systems were typed by hand. This directly reveals one of the weakest spots in the procedure. If the correct residue-type is not in the set of possibilities, a correct assignment of the spin system is not possible, which possibly leads to more errors. For the rest of the spin systems, a set of possible residue-types was determined automatically with a cut-off at 1% as described before. Tolerances were set to 0.3 and 0.4 for carbon and nitrogen respectively. 100 independent runs with 22 temperature points and 1000,000 attempts per temperature point were performed. As can be observed in 5.5C there are several differences between the manual assignment and the most frequently chosen assignment by the algorithm. There are basically 3 differences. In all cases the spin system corresponding to asparagine 55 in the original assignment is assigned to aspartic acid 22 and in most cases also visa versa. Assigning these residues manually was also very difficult. A collection of other sequential peaks from other connections are misinterpreted by the algorithm to be the connection between 21 and 55. Furthermore a large part of the signal set connecting asparagine 55 to its neighbors is missing because the residue is located in a loop between the beta-sheet and alpha-helix where a lot of line-broadening is observed. The second difference is that alanine 37 is assigned to glutamic acid 104 in almost all cases, leaving alanine 37 with a joker. Alanine 37 is part of a region of the protein dubbed the "ASSA" region, which is a flexible hinge playing a role in the autotransport mechanism of this protein. Because of the flexibility, signals are not present or broadened, explaining why the algorithm could not find the correct assignment. The algorithm could find the correct assignment only when the tolerances were increased or when the spin system was hand-typed to Alanine. A third difference is that two out of the three serine-serine pairs (residues 65-66 and 92-93 respectively) are assigned differently. Some less severe issues includes placing jokers on 15 and 16, which basically indicates that the right solution could not be found, but also no erroneous solution is proposed.

OmpG ^{1}H- and ^{13}C-detected spectra

For the assignment of OmpG seven spectra in total were selected. Five of those were ^{13}C-^{13}C correlation spectra with a DARR mixing period of 400 ms of the 1,3-glycerol, 2-glycerol, 1,3-TEMPQANDSG, 2-TEMPQANDSG and 2_SHLYGWAFV samples. These spectra are used for the structure calculation described later as well and the same peak lists were used as were used to generate distance restraints. The used ^{1}H-detected spectra were the hcaCBcacoNH and hcoCAcoNH spectra as these contain good sequential data. The spin systems consisted of all the spin systems that were generated as described in previous chapters, including spin systems that could not be assigned to residues. This last category are mostly spin systems corresponding to intra-residual peak pattern that could be observed in the ^{13}C-^{13}C correlations but could not be sequentially assigned, such as the two threonine and three leucine spin systems correponding to the residues shown in red in figure 4.3. Spin systems contain both chemical shift information on shifts in fully protonated and perdeuterated samples as described in the chapter about connecting assignments in ^{13}C- and ^{1}H-detected spectra. Also here 100 independent runs with 22 temperature points and 1000,000 attempts per temperature point were performed. The results on OmpG are a lot less clear than in the other two examples, see figure 5.5D. Only 41 spin systems where unanimously and correctly assigned to a specific residue. These residues correspond to parts of the spectra with good signal to noise. Another 51 residues was assigned correctly in over 80% of the optimization runs, and another 25 in over 50% giving good hints for the possible assignment.

Effect of missing peaks on the accuracy of proposed assignments

To simulate the effect of incomplete data, an increasing number of randomly selected peaks were excluded from the ^{13}C-detected dataset of SH3. This dataset was chosen for this purpose because it contains a lot of redundant information. The algorithm was tested on this reduced dataset and the amount of correctly and incorrectly assigned spin-systems were determined. Because the quality of the results is influenced by the subset of peaks that happens to be excluded, this procedure was repeated 10 times for each datapoint. The averages are shown in figure 5.6. Each execution of the algorithm consisted of 100 annealing runs in the same fashion as described before (22 temperature steps of 100.000 Monte Carlo attempts). As can be seen, the algorithm is tolerant against the exclusion of peaks. This can partially be explained by the small search space for a 62 residue protein like SH3 and the excellent dispersion of the spectra. Furthermore, it should be noted that in a real "bad" dataset the absent peaks are not randomly distributed over the primary sequence. Additionally, generation of spin systems in a "bad" dataset would be relatively hard.

Conclusion

A tool was created to help with the assignments in solid-state NMR. The main goal for the creation of this tool was to give an overview of the peak patterns over a large set of different types of experiments with various labeling schemes, connecting spin systems in a sequential matter. For every combination of two spin systems that can be assigned to two sequential residues, the expected peak patterns in different spectra is shown along with the information which of those peaks are present in the corresponding peak lists. Additionally, a global optimization procedure was added. This

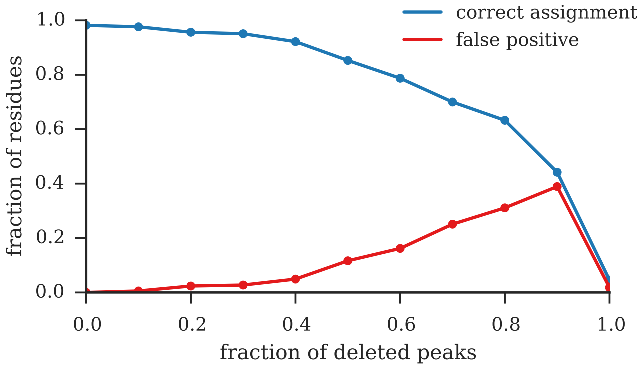


Figure 5.6: Amount of correct assignments and false positives as a function of the amount of deleted peaks in ^{13}C detected SH3. A (false) positive is defined here as any spin system being assigned to a residue in over 70% of the runs. When all peaks are removed (1.0 on the x.axis), the set of generated assignments become random and therefore no spin systems are assigned to a specific residue in over 70% of the runs, resulting in neither correct assignments nor false positives.

optimization routine is similar to automated assignment algorithms written before by other groups, and shows to be relatively reliable for the assignment of small proteins and to give valuable hints towards the correct assignment in larger proteins. Integration with CCPN Analysis allows access to valuable information that has been configured in the CCPN data model. All this information, such as the detailed composition of labeling schemes, the graphs describing magnetization transfers in experiments and the different levels of assignment already present for spin systems can be accessed in a more straight-forward way than when an import-export procedure is necessary, and without loss of information. With the tool presented here, the results of the optimization procedure can be evaluated in an interactive way and only the results that are believed to be correct can be accepted. Furthermore, all alternative assignment possibilities of spin systems to residues can be evaluated.

Download Information

Download and installation instructions for the assignment plug-in can be found at https://github.com/jorenretel/Malandro. The additional plug-in to set the residueTypeProp property of spin systems can be found at https://github.com/jorenretel/ccpnmr-residueTypeProbs-editor.

References

[1] T. J. Stevens, R. H. Fogh, W. Boucher, V. A. Higman, F. Eisenmenger, B. Bardiaux, B.-J. van Rossum, H. Oschkinat, and E. D. Laue. "A Software Framework for Analysing Solid-State MAS NMR Data". *Journal of Biomolecular NMR* 51.4 (Sept. 2011), pp. 437–447. DOI: 10.1007/s10858-011-9569-2.

[2] R. Tycko and K.-N. Hu. "A Monte Carlo/Simulated Annealing Algorithm for Sequential Resonance Assignment in Solid State NMR of Uniformly Labeled Proteins with Magic-Angle Spinning". *Journal of magnetic resonance (San Diego, Calif. : 1997)* 205.2 (Aug. 2010), pp. 304–314. DOI: 10.1016/j.jmr.2010.05.013.

[3] K.-N. Hu, W. Qiang, and R. Tycko. "A General Monte Carlo/Simulated Annealing Algorithm for Resonance Assignment in NMR of Uniformly Labeled Biopolymers". *Journal of Biomolecular NMR* 50.3 (June 2011), pp. 267–276. DOI: 10.1007/s10858-011-9517-1.

[4] P. Guerry and T. Herrmann. "Advances in Automated NMR Protein Structure Determination". *Quarterly Reviews of Biophysics* 44.03 (Aug. 2011), pp. 257–309. DOI: 10.1017/S0033583510000326.

[5] M. Schubert, T. Manolikas, M. Rogowski, and B. H. Meier. "Solid-State NMR Spectroscopy of 10% 13C Labeled Ubiquitin: Spectral Simplification and Stereospecific Assignment of Isopropyl Groups". *Journal of Biomolecular NMR* 35.3 (July 2006), pp. 167–173. DOI: 10.1007/s10858-006-9025-x.

[6] E. Schmidt and P. Güntert. "A New Algorithm for Reliable and General NMR Resonance Assignment". *Journal of the American Chemical Society* 134.30 (Aug. 2012), pp. 12817–12829. DOI: 10.1021/ja305091n.

[7] E. Schmidt, J. Gath, B. Habenstein, F. Ravotti, K. Székely, M. Huber, L. Buchner, A. Böckmann, B. H. Meier, and P. Güntert. "Automated Solid-State NMR Resonance Assignment of Protein Microcrystals and Amyloids". *Journal of Biomolecular NMR* 56.3 (May 2013), pp. 243–254. DOI: 10.1007/s10858-013-9742-x.

[8] J. T. Nielsen, N. Kulminskaya, M. Bjerring, and N. C. Nielsen. "Automated Robust and Accurate Assignment of Protein Resonances for Solid State NMR". *Journal of Biomolecular NMR* 59.2 (May 2014), pp. 119–134. DOI: 10.1007/s10858-014-9835-1.

[9] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, and E. D. Laue. "The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline". *Proteins: Structure, Function, and Bioinformatics* 59.4 (June 2005), pp. 687–696. DOI: 10.1002/prot.20449.

[10] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn, and K. Smith. "Cython: The Best of Both Worlds". *Computing in Science & Engineering* 13.2 (Mar. 2011), pp. 31–39. DOI: 10.1109/MCSE.2010.118.

[11] M. Matsumoto and T. Nishimura. "Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator". *ACM Trans. Model. Comput. Simul.* 8.1 (Jan. 1998), pp. 3–30. DOI: 10.1145/272991.272995.

[12] J. Ayers. *https://bitbucket.org/joshayers/cythonlib*. https://bitbucket.org/joshayers/cythonlib.

[13] J. Pauli, M. Baldus, B. van Rossum, H. de Groot, and H. Oschkinat. "Backbone and Side-Chain 13C and 15N Signal Assignments of the α-Spectrin SH3 Domain by Magic Angle Spinning Solid-State NMR at 17.6 Tesla". *ChemBioChem* 2.4 (Apr. 2001), pp. 272–281. DOI: 10.1002/1439-7633(20010401)2:4<272::AID-CBIC272>3.0.CO;2-2.

[14] F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein, and H. Oschkinat. "Structure of a Protein Determined by Solid-State Magic-Angle-Spinning NMR Spectroscopy". *Nature* 420.6911 (Nov. 2002), pp. 98–102. DOI: 10.1038/nature01070.

[15] A. J. Nieuwkoop, W. T. Franks, K. Rehbein, A. Diehl, Ü. Akbey, F. Engelke, L. Emsley, G. Pintacuda, and H. Oschkinat. "Sensitivity and Resolution of Proton Detected Spectra of a Deuterated Protein at 40 and 60 kHz Magic-Angle-Spinning". *Journal of Biomolecular NMR* 61.2 (Feb. 2015), pp. 161–171. DOI: 10.1007/s10858-015-9904-0.

[16] S. A. Shahid, S. Markovic, D. Linke, and B.-J. van Rossum. "Assignment and Secondary Structure of the YadA Membrane Protein by Solid-State MAS NMR". *Scientific Reports* 2 (Nov. 2012). DOI: 10.1038/srep00803.

[17] S. A. Shahid, B. Bardiaux, W. T. Franks, L. Krabben, M. Habeck, B.-J. van Rossum, and D. Linke. “Membrane-Protein Structure Determination by Solid-State NMR Spectroscopy of Microcrystals”. *Nature Methods* 9.12 (Dec. 2012), pp. 1212–1217. doi: 10.1038/nmeth.2248.

Chapter 6

Structure Calculation

The experimental restraints used for the structure calculation of OmpG exist of two types: torsion angle restraints that are predicted based on chemical shifts and distance restraints based on cross-peaks in through-space correlation spectra. This second group of restraints can in turn be subdivided into a group of distance restraints obtained from ^{1}H-detected experiments and another group that is based on ^{13}C-detected experiments. To automatically produce lists of distance restraints, peak positions are matched with chemical shifts. As described earlier, all ^{1}H-detected experiments were performed on perdeuterated and back-exchanged samples, whereas all ^{13}C-detected experiments were performed on fully protonated samples. Because of the large isotope shift between these two different samples, two different shift lists were used to produce the distance restraints. Also spectra of poor quality that were present in the CCPNMR Analysis project were excluded from shift averaging before starting the shift matching procedure. Because of chemical shift overlap, this peak matching procedure does not produce unambiguous restraints between two nuclei in the protein for most peaks. Therefor sets of ambiguous distance restraints (ADRs) are generated. Each ADR basically consists of a list of possible assignments of a cross-peak in the spectrum. These assignment options are referred to as restraint items, or short items. The ADRs were disambiguated using ARIA (Ambiguous Restraints for Iterative Assignment) [1][2]. This program calculates the structure in a number of iterations. In each iteration an ensemble of structures is calculated based on the ADRs. After each iteration the assignment options that are unlikely to be correct based on the average distances in the highest energy structures of this temporary ensemble are removed. As ADRs become less ambiguous the calculated structures converge and vice versa. After a first round of structure calculations using ARIA, hydrogen bond restraints can be added between residues that are in the right conformation in the β-strands.

Torsion Angle Restraints

128 φ/ψ torsion angles (256 in total) where predicted using the program TALOS+ [3][4]. In figure 6.1 the secondary structure that corresponds to these torsion angles is shown along the OmpG sequence. As expected the largest part of the assigned residues are predicted to be in a β-sheet conformation. These results can be compared to a prediction of the topology done purely on the basis of the amino acid sequence by a program called PRED-TMBB [5]. This tool is specifically designed for β-barrels and predicts which part of the molecule is part of the transmembrane β-sheet, intra-cellular turn and extra-cellular loop. Because the algorithm is based on machine learning, we verified with the author of the program, that previously calculated OmpG structures were not part of the training data, which was not the case. It can be observed that the two predictions align fairly well. Where PRED-TMBB predicts a turn, the chemical shifts are more coil-like. In these turns also a lower random coil index (RCI) value can be observed, indicating a less ordered part of the molecule. As discussed before, the missing assignments cluster largely in the extra-cellular loops.

Restraints based on ^{1}H-detected through-space correlation experiments

To obtain a set of distance restraints using ^{1}H-detection, through-space experiments were recorded on the perdeuterated samples where the exchangeable sites were 100% back-exchanged for protons. Two spectra where recorded for this purpose: an hNHH and an hNhhNH, both using cross-polarization for transfers between proton and nitrogen and a 2 ms RFDR (radio frequency driven recoupling) mixing step to transfer magnetization between the protons. Also an extra hCANH was recorded in the same measurement block. The peaks in this spectrum were assigned (based on the previously made sequential assignment) to obtain a chemical shift list that corresponds well to the through-space spectra. This was necessary because for the moment it is hard to exactly control the temperature in fast spinning samples. These kind of slightly different experimental conditions always introduce small chemical shift differences, that are unfavorable for a correct outcome of the shift matching procedure.

Since most of the proton sites in the molecule are deuterated, most peaks present in the two through-space spectra are peaks correlating one amide group to another. In both spectra, strips can be drawn at the ^{15}N and ^{1}H chemical shifts of one amide group. In general such strips contain, besides a diagonal peak, one big and often one or two smaller cross-peaks. Since the correlation pathway of the hNhhNH experiment guarantees that both interacting protons are part of an NH-group this spectrum is a bit cleaner. If both spectra are evaluated together there are four peaks indicating the proximity of two NH groups. An example of such a set of four peaks correlating two amide groups is shown in figure 6.3. In the case of an anti-parallel β-sheet, the strongest off-diagonal peak is almost always correlating two amide groups facing each other from neighboring strands in the sheet.

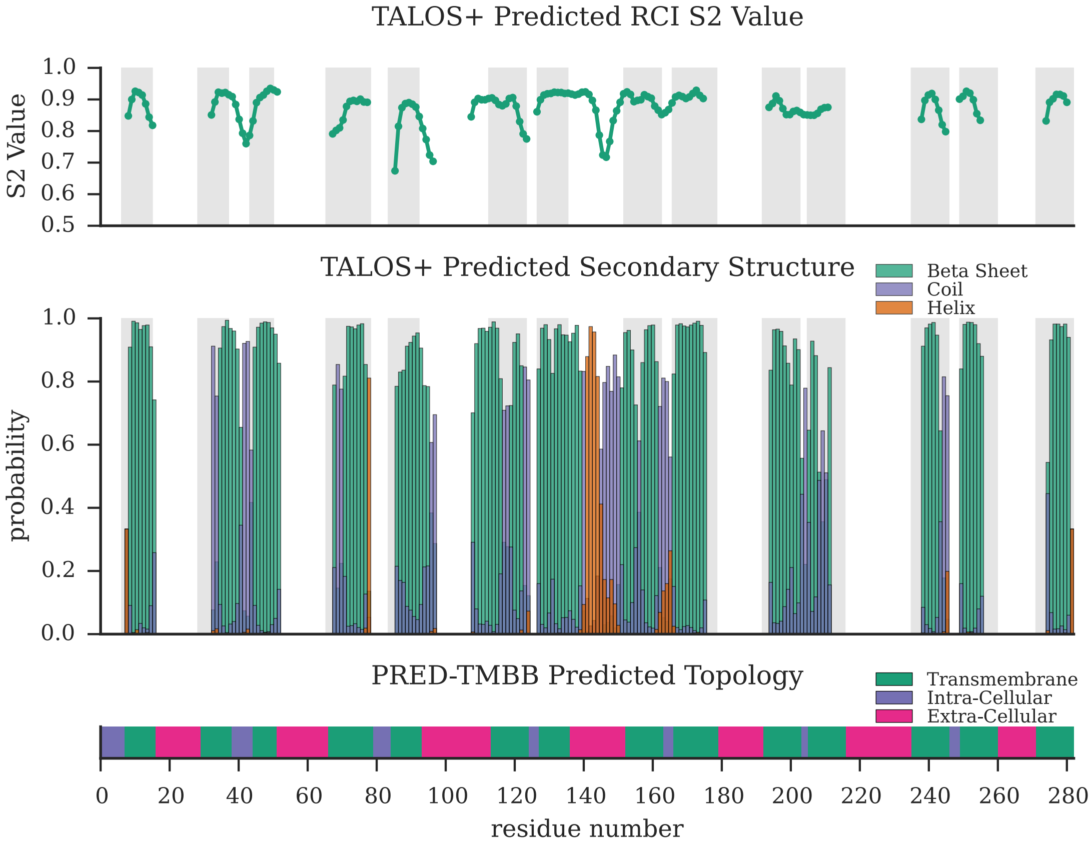


Figure 6.1: Prediction of the secondary structure of OmpG by TALOS+ and PRED-TMBB. TALOS+ uses the secondary chemical shifts of assigned residues to search a database for triplets in the sequence of high resolution structures with similar secondary chemical shifts to predict φ/ψ torsion angles. PRED-TMBB is an algorithm that solely relies on the sequence and predicts which parts of the sequence are intra-cellular, extra-cellular and transmembrane given the molecule is a transmembrane β-barrel. Grey blocks in TALOS+ plots correspond to areas predicted as transmembrane by PRED-TMBB.

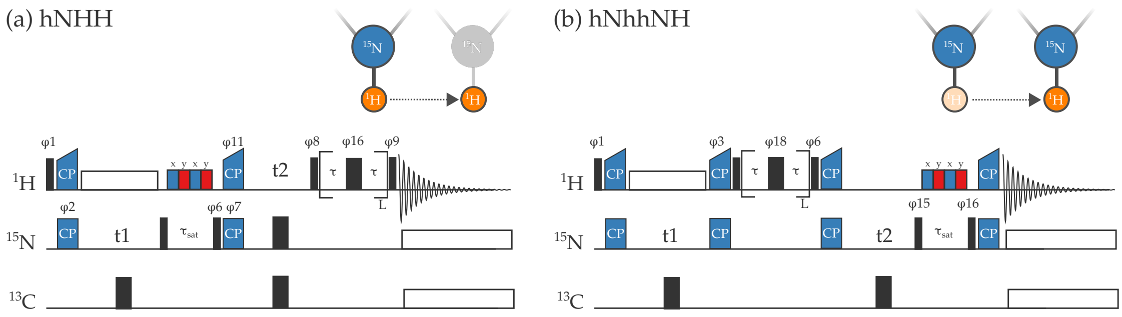


Figure 6.2: Pulse sequences of the hNHH (a) and hNhhNH (b) experiments. Diagrams above the pulse sequences illustrate the nuclei of which the chemical shift is measured. Blue and dark orange nuclei are measured, while the light orange ^{1}H in the diagram of figure b is on the magnetization transfer pathway but not measured. Periods between square brackets are RFDR transfer blocks. Phase cycle: (a) φ1 = 1 3, φ2 = 1, φ6 = 0 0 2 2, φ12 = 1, φ7 = 1, φ8 = 2, φ9 = 0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3, φ16 = 0 1 0 1 1 0 1 0, φ11 = 1, φrec = 0 2 2 0 1 3 3 1 2 0 0 2 3 1 1 3; (b) φ1 = 1 3, φ3 = 1 1 3 3, φ6 = 1 1 1 1 3 3 3 3, φ16 = 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3, φ15 = 1, φ18 = 0 1 0 1 1 0 1 0, φrec = 0 2 2 0 2 0 0 2 2 0 0 2 0 2 2 0. All other pulses have phase 0.

These two residues are involved in two hydrogen bonds between one another's carbonyl-oxygens and amide protons. On average, these two amide protons are only 3.1 Å separated from one another, see figure 6.8B. The smaller peaks are often correlations to the amide groups of the neighboring residues in the same strand or to the amide group of the residue following the directly hydrogen bonded residue in the opposing strand. As can be seen in figure 6.8C, a very specific alternating pattern of cross-peaks between residues is expected, connecting two strands in the β-sheet, skipping residues in between which are facing towards the interface with other strands in the sheet. The mixing time of 2 ms is relatively short, so that it was possible to distinguish between the short (over-the-strand) distance and the correlations between more distant protons. The optimal mixing time was determined by recording several 2D hNhH spectra with different mixing times. Some peaks that are well separated in the NH correlation, like those of tryptophan 113 and glutamic acid 155, could be used to monitor the relative sizes of the main cross-peak (over-the-strand) and the additional smaller peaks. At longer mixing times the smaller cross-peaks increase in size, while the main cross-peak become smaller. Since the smaller cross-peaks are often sequential, and therefor of little structural value, we chose to run the 3D experiments with a mixing time that gave maximum intensity for the main cross-peaks.

Matching peak dimensions to chemical shifts

To generate ADRs based on peak positions, a set of chemical shift tolerances has to be defined. These tolerances were set to 0.4 ppm for the ^{15}N dimensions, 0.1 ppm for the indirectly detected ^{1}H dimensions and 0.07 ppm for the directly detected ^{1}H dimensions. The slightly smaller tolerance on the directly directed acquired dimension is possible because this dimension is better digitized. When

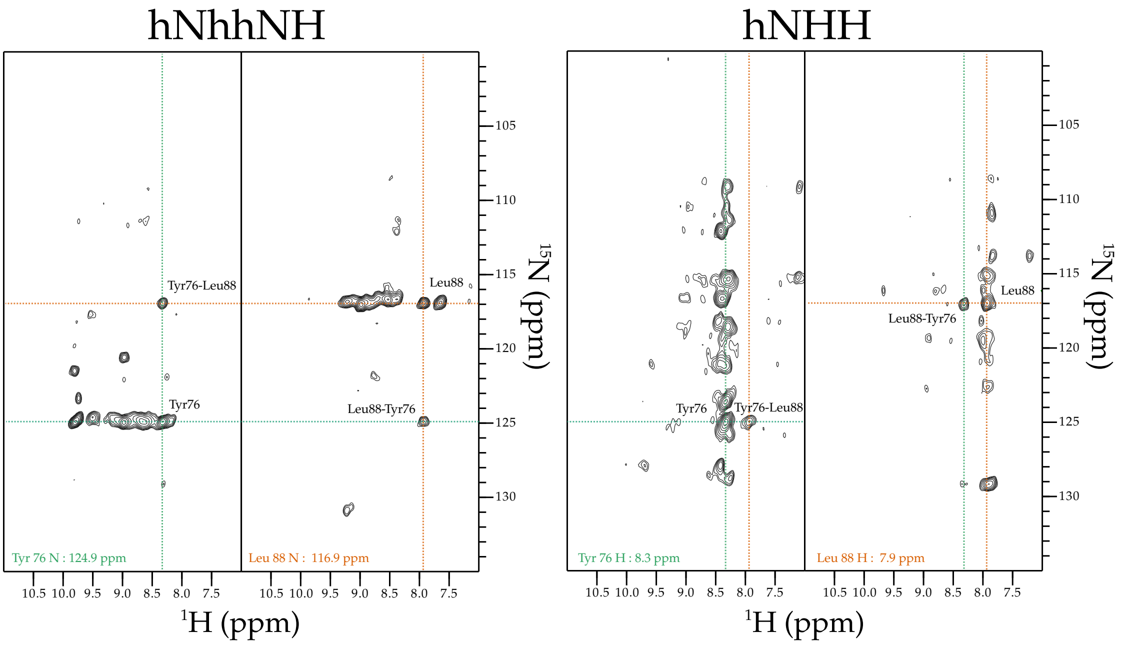


Figure 6.3: Strips from the hNhhNH and hNHH spectra, illustrating the cross-strand inaction between the backbone amide groups of tyrosine 76 and leucine 88. Orange lines correspond to the ^{1}H and ^{15}N chemical shifts of leucine 88. Green lines correspond to the ^{1}H and ^{15}N chemical shifts of tyrosine 76. A total of 4 cross-peaks is present at the intersections of orange and green lines. While the peaks at the intersections of lines with the same color are the diagonal (intra-residual) peaks.

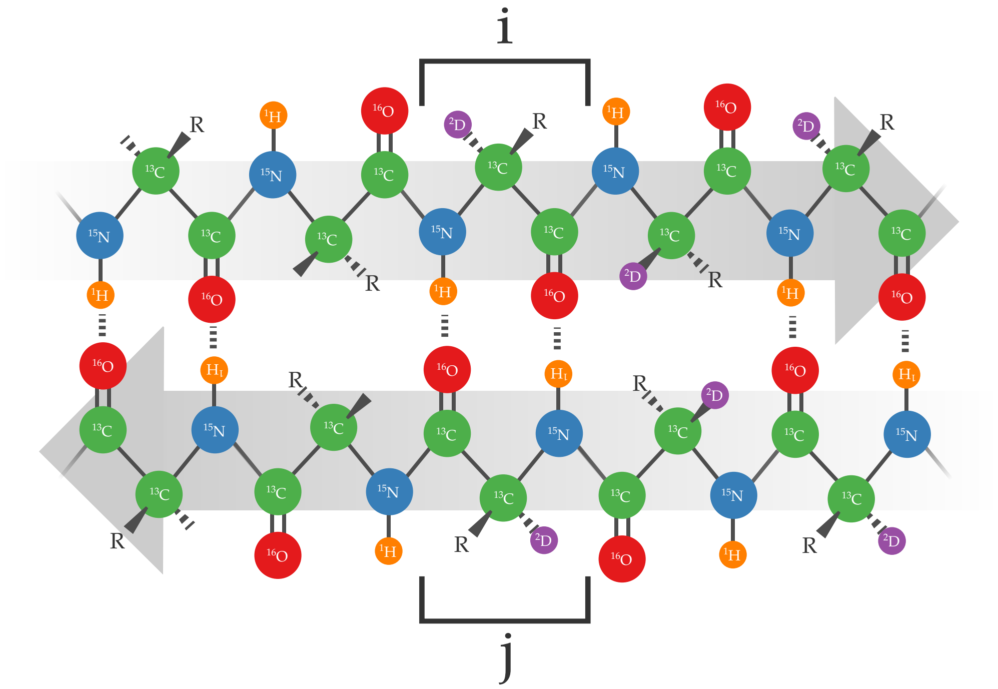


Figure 6.4: Orientation of strands in an anti-parallel β-sheet. Dotted vertical lines indicate hydrogen bonds. Residues i and j correspond correspond to figure 6.8.

using these tolerances directly, the window around the actual peak position in which assignments are accepted is square or cubical depending on the dimensionality. When considering the chemical shifts for each dimension together, in the corners of this square (such as the possition of D in 6.5) a combination of chemical shift that is still accepted as a possible assignment is further away from the actual peak position as a combination that is closer to one of the mid-lines of the square. In other words, when using tolerances like this on a multidimensional spectrum, some restraint items are created where all chemical shifts are really far away from the actual peak position. When assigning a peak by hand these kind of combinations would probably be discarded as they are less likely to be correct than a combination where only one of the dimensions is close to the edge of the tolerance.

To reduce the number of unlikely assignment options, an extra rule was applied. In both the hNHH and the hNhhNH, 2 out of 3 dimensions directly correspond to one bonded ^{15}N-^{1}H pair, i.e. one peak in the ^{15}N-^{1}H correlation. Therefor these 2 chemical shifts can be considered together when mapping chemical shifts to peak positions. In practice the euclidean distance between the peak position and the combination of the ^{15}N and ^{1}H chemical shifts normalized by the shift tolerances is calculated. All assignment options that are closer than half of the distance between the center and the corner of the normalized tolerance square (i.e. $\sqrt{2}/2$) are accepted indiscriminately (A and B in figure 6.5). All assignment options that are outside of this circle are only accepted when there is not an assignment option that is twice closer to the actual peak position. Therefor option C in 6.5 would only be accepted in the absence of option A. On the dimension of the through space correlated nucleus this rule was not applied. A similar technique is used by the CANDID routine in CYANA, where the likelihood of a peak assignment is made dependent on the closeness in chemical shift match [6]. In the case of CYANA this happens before every iteration of the structure determination/cross-peak assignment protocol. Here it is done only once before the ambiguous distance restraints enter the ARIA protocol.

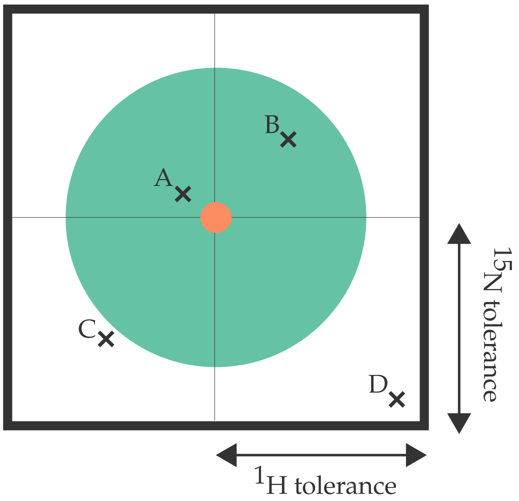


Figure 6.5: For the directly bonded ^{1}H and ^{15}N dimensions of the hNHH and hNhhNH spectra, the distance between the actual peak position (orange circle) and the combinations of ^{15}N-^{1}H chemical shifts of one amide group is calculated. These distances are normalized by the shift tolerances. Within the turquoise circle, corresponding to half of the distance between the peak position and the corners of the square, all possible assignments (A and B) are accepted. Outside of this circle assignment possibilities are only accepted in the absence of an assignment possibility twice closer to the peak position. For instance option C would only be accepted if A would not be present.

Using redundancy to disambiguate restraints

As detailed before there are in principle four cross-peaks forming a network that correlates the same two amide groups. This redundancy can be used to decrease the ambiguity of automatically generated ADRs, before the structure calculation. A CCPNMR macro script was used to determine for which items of each ADR all 4 peaks were present (giving rise to three other ADRs that also have the correlation between these amide hydrogens as one of their items). For restraints that had one or more of such items, all other items that had a "symmetry" of 2 or less (instead of 4) were removed. In these cases all restraint items with a symmetry of 3 were preserved, as to not remove a possibly correct item just because one peak is missing. In figure 6.6 it can be seen that after applying this operation, the amount of restraints that become unambiguous or only have two items left is drastically increased while the amount of restraint with ten or more items decreases. When plotting the resulting restraints on a residue interaction matrix the pattern expected for β-sheets, lines of interactions perpendicular to the diagonal, can already be seen (figure 6.6). Based on this pattern the peak assignment could in principle be continued manually. However, we decided to give ARIA the task of further disambiguating the remaining ambiguous restraints. This procedure is similar to a feature present in the CANDID routine, where it is applied before every iteration of structure calculation in the protocol [6]. In the case here, it allows ARIA to find the correct global fold already in the first iteration of the protocol, which helps to disambiguate the more ambiguous restraints in the next iterations.

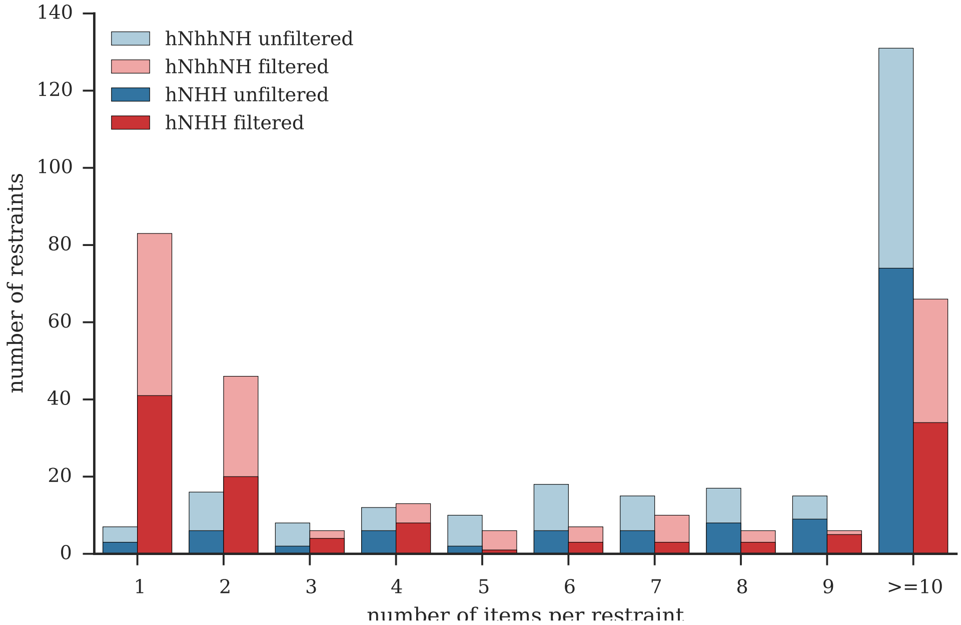


Figure 6.6: Ambiguity of restraints based on the hNhhNH and hNHH spectra. Blue bars correspond to restraints that are automatically created by matching chemical shifts to peak dimensions. Red bars represent the same restraint set, but after applying a filter that selects restraints items for which all four expected peaks are present in the two spectra. This operation effectively decreases the amount of restraints with very high ambiguity and in both spectra about a third of the restraints becomes unambiguous (1 item per restraint). Light and dark color represent the hNhhNH and hNhhNH spectra, respectively.

Distance classes

In the dipolar-based transfer experiments used in solid-state NMR distances can not be extracted from peak volumes with the same amount of precision as in solution NMR. Therefor a very crude division into two upper bound distance classes was done. Peaks were sorted from high to low intensity. Starting from the most intense peaks in the list, peaks were classified to correspond to a short

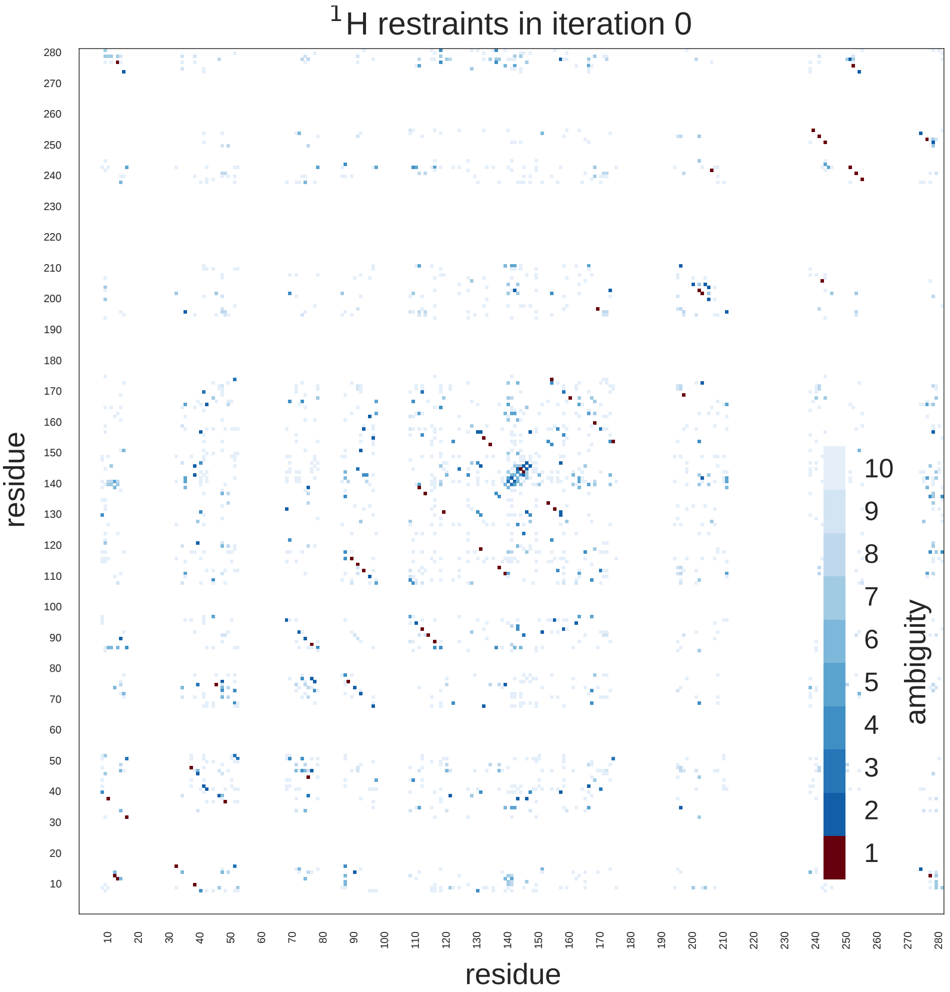


Figure 6.7: Residue interaction matrix for ^{1}H-^{1}H ADRs entering the ARIA protocol (before any disambiguation by ARIA). The color indicates the ambiguity of the least ambiguous restraint present for the interaction between two residues. Interactions between two residues for which an unambiguous restraint is present are colored red. Patterns perpendicular to the diagonal, indicating an anti-parallel β-sheet, can already be observed.

distance (3.5 Å) until the first peak was encountered that was not the largest in its strip (and therefore not corresponding to the shortest over-the-strand distance). All peaks with a intensity equal or lower to this peak are given a more generous upper bound of 5.5 Å. Lower bounds were set to 1.0 Å in both cases.

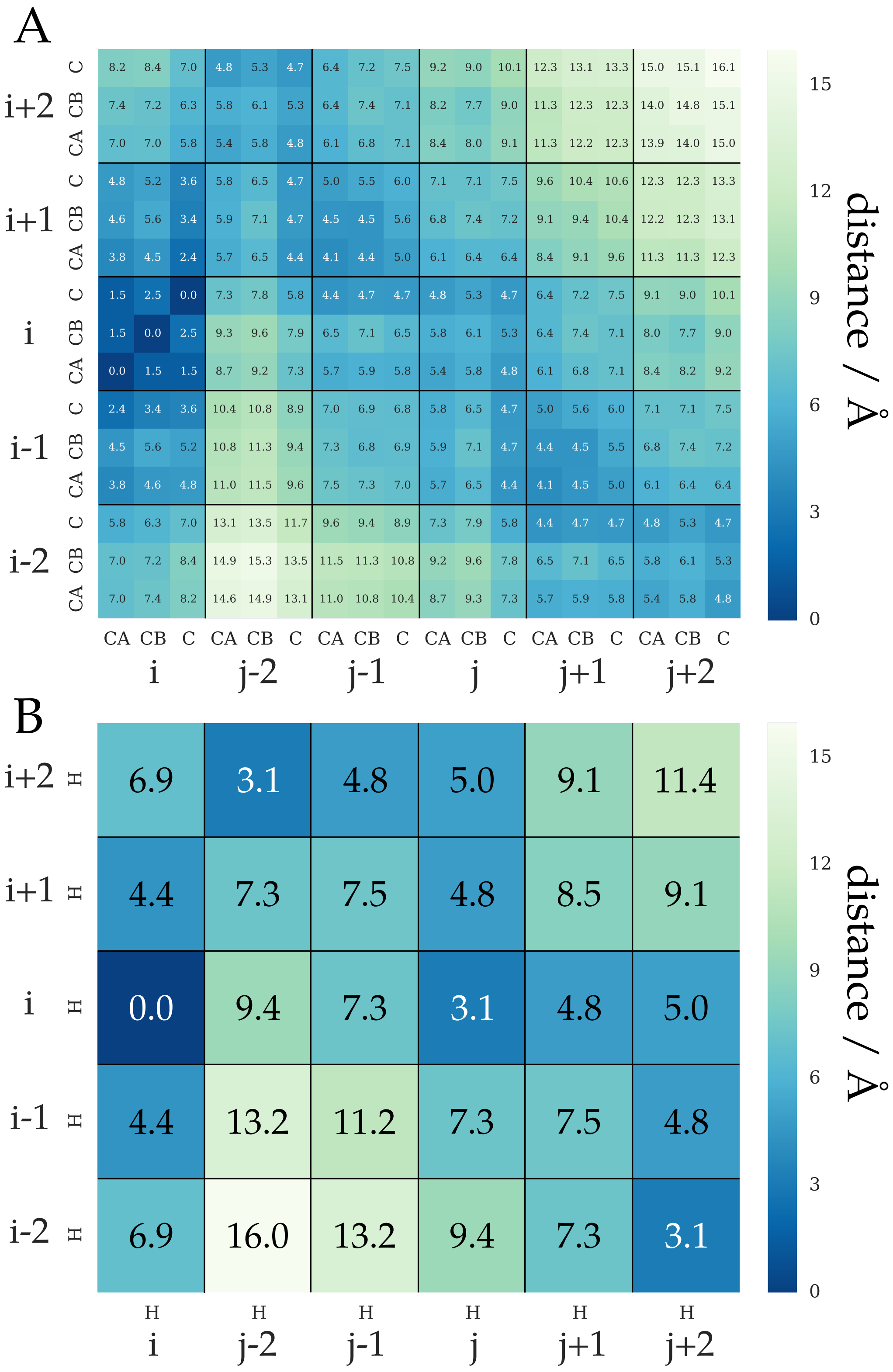
Restraints based on ^{13}C-detected through-space correlation experiments

Another set of distance restraints was based on a set of five 2D 13^{C}-^{13}C correlations with 400 ms of DARR mixing. Only peaks in the aliphatic region of the spectra were picked. The reason for this is that the chemical shift assignment for this region is relatively complete in comparison to other regions of the spectra (at least for the resonances within assigned residues (91%), see table 4.2). This is important since the biggest bottle-neck in structure calculation is incomplete resonance assignment. It has been shown in solution NMR studies that the resonance assignment should be at least 90% complete to produce reliable structures using automated NOE assignment [7]. The completeness of assignment over the whole sequence is far below that (57%) but, as argued before, inter-residual cross-peaks are expected to be absent for the unassigned parts of the sequence. Intra-residual peaks were avoided in the peak picking. This was done by comparing the used spectra with spectra that were recorded using a shorter mixing time complemented by knowledge of which regions in the spectra simply can not contain intra-residual cross-peaks, see figures 6.9 and 6.10. As discussed weak intra-residual signals are present in the ^{13}C-^{13}C correlations corresponding to unassigned spin systems. By not picking the intra-residual signal set it is avoided that incorrect ADRs are generated based on these peaks.

ARIA can either use lists of ADRs as input or peak lists accompanied by a chemical shift list. In the last case ARIA performs the shift-matching itself. Here lists of ADRs were produced using CCPNMR Analysis because of the build-in support for labeling schemes. Restraints were produced by shift-matching with a tolerance of 0.4 ppm in both dimensions and only assignment possibilities were generating for which the co-labeling fraction of the two correlated carbons exceeded 0.1. All ADRs based on the ^{13}C-detected spectra were put in a single distance class with a lower bound of 1.5 Å and an upper bound of 8.0 Å.

Structure calculation protocol

For the structure calculation and disambiguation of the ADRs the standard ARIA protocol was used with a few alterations detailed below. As in the default protocol 9 iterations (0-8) were done, followed by a refinement in DMSO. In each iteration 192 structures were calculated and the 15 lowest energy subset of those structures was used to disambiguate the assignment of the ADRs for the next



A

		i			j-2			j-1			j			j+1			j+2		
		CA	CB	C	CA	CB	C	CA	CB	C	CA	CB	C	CA	CB	C	CA	CB	C
i+2	C	8.2	8.4	7.0	4.8	5.3	4.7	6.4	7.2	7.5	9.2	9.0	10.1	12.3	13.1	13.3	15.0	15.1	16.1
	CB	7.4	7.2	6.3	5.8	6.1	5.3	6.4	7.4	7.1	8.2	7.7	9.0	11.3	12.3	12.3	14.0	14.8	15.1
	CA	7.0	7.0	5.8	5.4	5.8	4.8	6.1	6.8	7.1	8.4	8.0	9.1	11.3	12.2	12.3	13.9	14.0	15.0
i+1	C	4.8	5.2	3.6	5.8	6.5	4.7	5.0	5.5	6.0	7.1	7.1	7.5	9.6	10.4	10.6	12.3	12.3	13.3
	CB	4.6	5.6	3.4	5.9	7.1	4.7	4.5	4.5	5.6	6.8	7.4	7.2	9.1	9.4	10.4	12.2	12.3	13.1
	CA	3.8	4.5	2.4	5.7	6.5	4.4	4.1	4.4	5.0	6.1	6.4	6.4	8.4	9.1	9.6	11.3	11.3	12.3
i	C	1.5	2.5	0.0	7.3	7.8	5.8	4.4	4.7	4.7	4.8	5.3	4.7	6.4	7.2	7.5	9.1	9.0	10.1
	CB	1.5	0.0	2.5	9.3	9.6	7.9	6.5	7.1	6.5	5.8	6.1	5.3	6.4	7.4	7.1	8.0	7.7	9.0
	CA	0.0	1.5	1.5	8.7	9.2	7.3	5.7	5.9	5.8	5.4	5.8	4.8	6.1	6.8	7.1	8.4	8.2	9.2
i-1	C	2.4	3.4	3.6	10.4	10.8	8.9	7.0	6.9	6.8	5.8	6.5	4.7	5.0	5.6	6.0	7.1	7.1	7.5
	CB	4.5	5.6	5.2	10.8	11.3	9.4	7.3	6.8	6.9	5.9	7.1	4.7	4.4	4.5	5.5	6.8	7.4	7.2
	CA	3.8	4.6	4.8	11.0	11.5	9.6	7.5	7.3	7.0	5.7	6.5	4.4	4.1	4.5	5.0	6.1	6.4	6.4
i-2	C	5.8	6.3	7.0	13.1	13.5	11.7	9.6	9.4	8.9	7.3	7.9	5.8	4.4	4.7	4.7	4.8	5.3	4.7
	CB	7.0	7.2	8.4	14.9	15.3	13.5	11.5	11.3	10.8	9.2	9.6	7.8	6.5	7.1	6.5	5.8	6.1	5.3
	CA	7.0	7.4	8.2	14.6	14.9	13.1	11.0	10.8	10.4	8.7	9.3	7.3	5.7	5.9	5.8	5.4	5.8	4.8

B

		i	j-2	j-1	j	j+1	j+2
		H	H	H	H	H	H
i+2	H	6.9	3.1	4.8	5.0	9.1	11.4
i+1	H	4.4	7.3	7.5	4.8	8.5	9.1
i	H	0.0	9.4	7.3	3.1	4.8	5.0
i-1	H	4.4	13.2	11.2	7.3	7.5	4.8
i-2	H	6.9	16.0	13.2	9.4	7.3	3.1

Figure 6.8: Average distances between C', $C\alpha$ and $C\beta$ nuclei (A), and between the amide protons in the backbone (B) between residues on positions -2 to 2 in the sequence relative to residues i and j, where i and j are the residues labeled as such in figure 6.4. The first column of B and C represent distances within the same strand, all other distances are between the two strands in the sheet. The shortest ^{13}C-^{13}C distance is between $C\alpha_{i-1}$-$C\alpha_{j+1}$ or $C\alpha_{j-1}$-$C\alpha_{i+1}$ which is 4.1 Å on average. The ^{1}H-^{1}H distances show a distinct pattern where smallest distance is between the amide protons of residue i and j (3.1 Å). The connection between residues i and to both j+1 and j+2 is a lot shorter than between i and j-1.

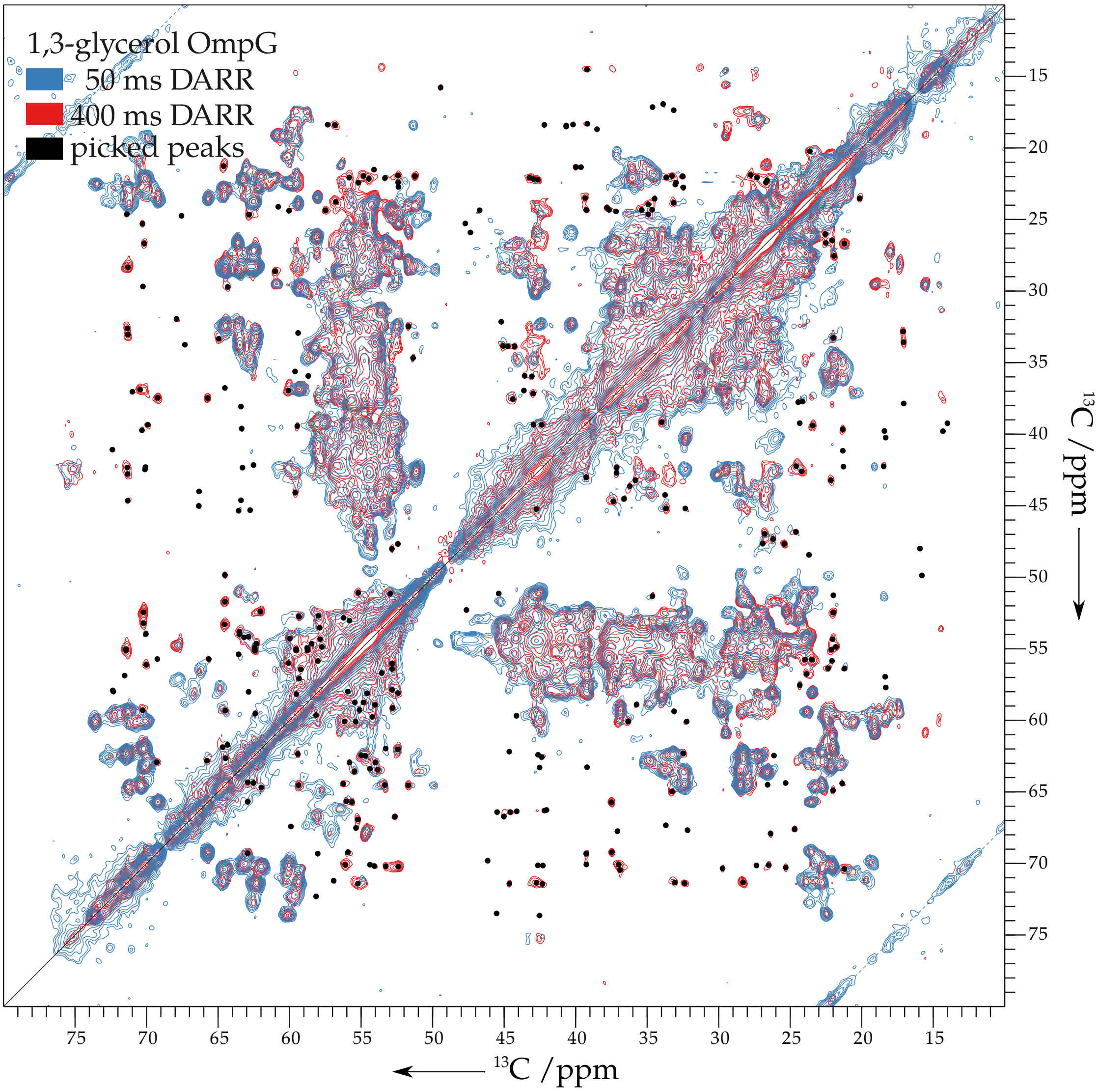


Figure 6.9: Peaks picked in the 400 ms 1,3-glycerol labeled OmpG spectrum. The 50 ms spectrum is superimposed.

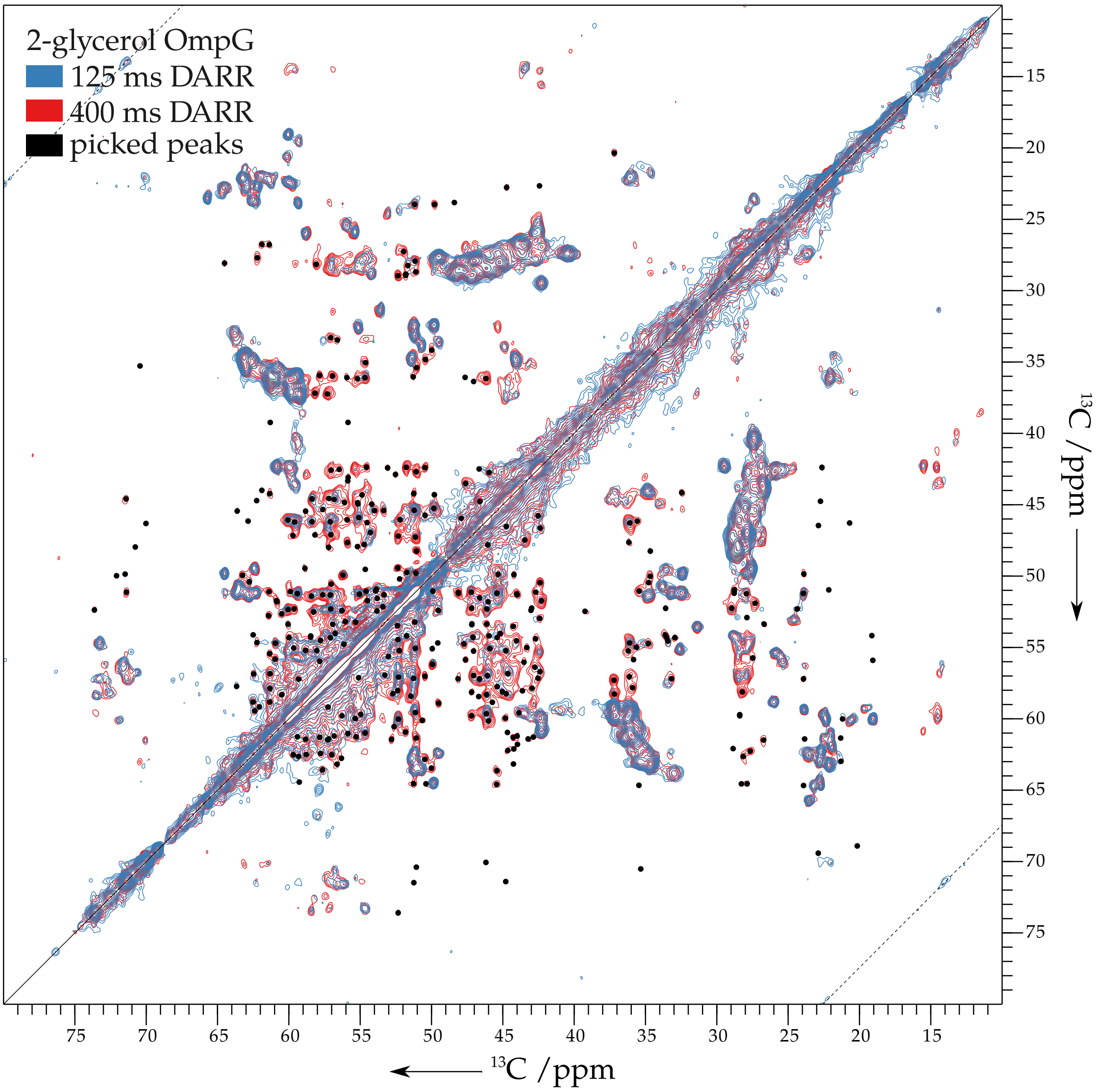


Figure 6.10: Peaks picked in the 400 ms 2-glycerol labeled OmpG spectrum. The 125 ms spectrum is superimposed.

iteration. The new ramachandran potential included in new distributions of ARIA was employed to generate the structures using both torsion angle dynamics and cartesian dynamics. The relevant settings for the ARIA protocol and the structure generation steps that were used are shown in table 6.1.

The most challenging part of the current structure calculation was to reduce the effect of ADRs where a correct item is not present. These type of restraints can be generated if noise or artifacts are present in the peak lists on which the ADRs are based. In the case of this structure calculation, the most likely source of these restraints is the incompleteness of the resonance assignment. For instance, when a cross-peak is present to a nucleus that is not assigned and at the same frequency there are one or more other, incorrect, assignment possibilities, an ADR will be generated with several items except for the correct one. One such a distance restraint can already cause the calculation to converge to a wrong structure. This was not such an issue in the ^{1}H detected spectra, since only a few ^{15}N-^{1}H combinations were left unassigned. Because the amount of peaks in the proton detected spectra is a lot smaller than in the ^{13}V-detected spectra and there is exactly one strip per residue it is more straight-forward to detect which peaks correlate an unassigned nucleus. There are some unassigned side-chain protons left at exchangeable sides. However, their chemical shifts are often distinct and not overlapped by other chemical shifts. Therefor cross-peaks to these nuclei could be easily recognized and removed from the peak list.

In the ^{13}C-detected spectra however, this problem is more severe. First of all, the amount of peaks in these spectra is larger. And second, there are missing ^{13}C assignments even in the parts of the protein that are structured. In these cases the lack of assignment is not caused by missing signals but by the ambiguity in the spectra. The unassigned shifts do not differ in any way from the assigned ones, making it hard to remove the peaks giving rise to incorrect ADRs before the structure calculation.

Because the ^{1}H-detected restraints between amide protons are very appropriate for constraining the backbone conformation of a protein that is almost entirely β-sheet, the first 4 iterations (0-3) of the protocol solely rely on these restraints. Already after the first iteration the lowest energy structures clearly show the the shape of the β-barrel, see figure 6.11. Only starting from iteration 4 the ^{13}C-^{13}C distance restraints were added. All incorrect ADRs that not fit within the violation tolerance to at least half of the lowest energy structures in the previously calculated ensemble are then rejected by ARIA's violation analysis. The default violation tolerances for each iteration were used, which is 1.0 Å in iteration 4.

In addition, restraint combination was employed to reduce the destructive effect of the presence of incorrect ADRs. Restraint combination was first introduced in CYANA and later implemented in ARIA as well [6]. The basic idea behind this strategy is to combine the restraint items of the two ADRs stemming from two unrelated peaks into a single new ADR. Because the amount of erroneous ADRs is normally small compared to the amount of correct ADRs the chance that the newly generated ADR still does not contain at least one correct item is decreased. Two strategies for the combination of restraints are implemented in both CYANA and ARIA: combining two ADRs to create one new ADR, or combining four ADRs to create four new combined ADRs. The last option was chosen because it keeps the amount of restraints the same and it is the most widely used strategy. Restraint combination was only applied to the ^{13}C-detected restraints and was enabled from the moment they enter the calculation (iteration 4) until iteration 6. In the last iterations (7 and 8) the violation tolerance is small by default (0.1 Å) effectively removing any of the restraints that do still not fit the previously calculated ensemble.

Because the structure already converged quite well in the very first iteration, the structure does not notably improve until iteration 7, see figure 6.11. Because the partial assignment threshold is initially reduced very slowly, the algorithm only becomes more discriminatory between restraint items with similar average distances in the structural ensemble in the later iterations. The influence on the convergence of the OmpG structure of both the moment at which the ^{13}C-^{13}C restraints enter the protocol and the number of iterations in which restraint combination is applied should still be thoroughly studied.

Table 6.1: Settings used in ARIA for the structure calculation of OmpG. For a detailed overview of the used restraints, see figure 6.13.

structure generation	
structure engine	CNS
potential	ramachandran
TAD high temperature	20,000 K
TAD time step factor	9.0
cartesian high temperature	3000 K
time step	0.003
final temperature cool stage 1	1000 K
steps in cool stage 1	100,000
final temperature cool stage 2	50 K
steps in cool stage 2	100,000
high temperature steps	20,000
refine steps	8000
protocol	
number of iterations	9
number of structures calculated per iteration	192
number of lowest energy structures used	15
first iteration with ^{1}H-^{1}H restraint	s 0
first iteration with ^{13}C-^{13}C restrai	nts 4
4 to 4 restraint combination on ^{13}C-^{13}C restraints in iterations	4-6
Merging method all other restraints/iterations	standard
final refinement	DMSO

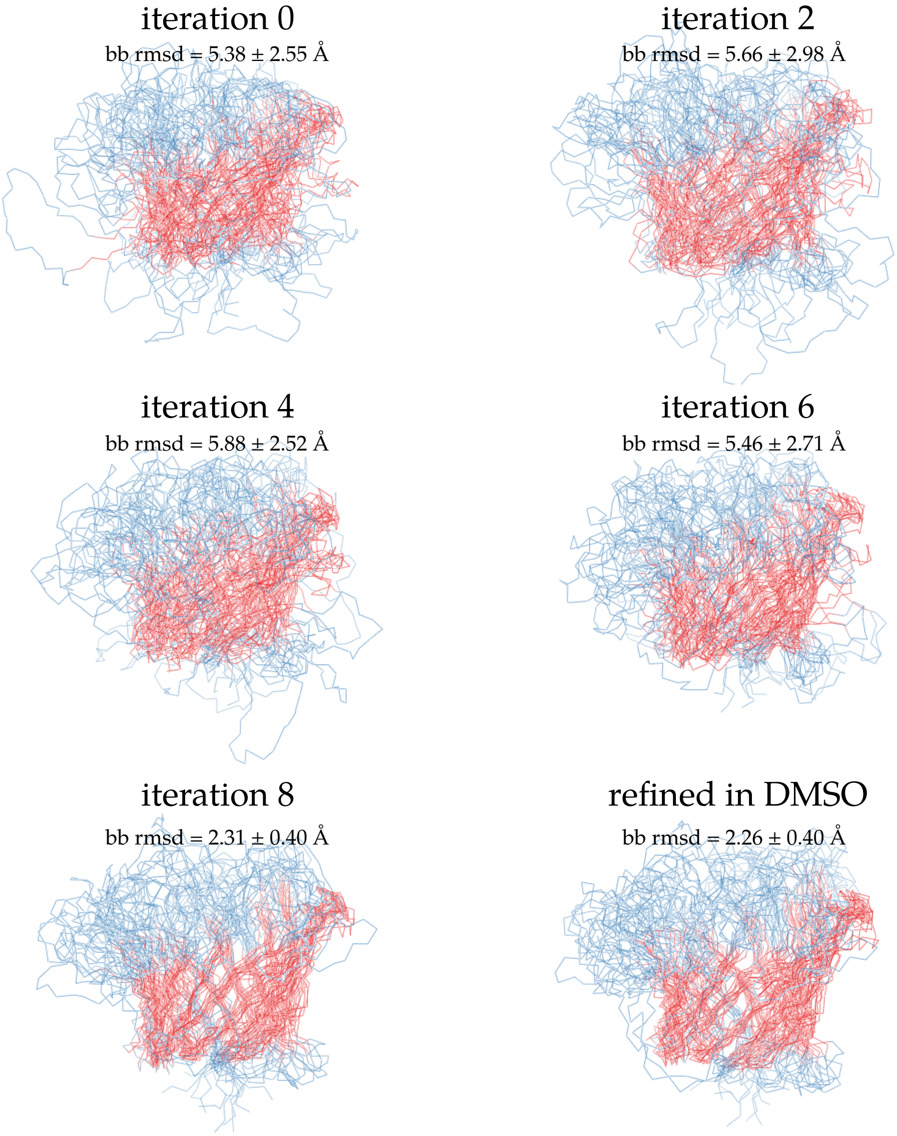


Figure 6.11: The 15 lowest energy structures at the end of every second ARIA iteration.

Hydrogen Bond Restraints

No hydrogen bond restraints were added in the initial structure calculations of OmpG. This was done because no experiments were performed to directly observe hydrogen bonds. However, after an initial structure is obtained, the hydrogen bonding pattern in the β-sheet is obvious and these type of restraints can be added. Co-linear hydrogen bond restraints were created between every two residues for which the predicted dihedral angles indicated beta-sheet and for which cross-peaks appear in the ^{1}H-detected spectra. Co-linear hydrogen bond restraints are basically distance restraints, one between the H and O and one between the N and the O. This makes these restraints very powerful, as they effectively constraint the HN bond vector. Every two residues facing each other from opposite strands interact in two hydrogen bonds. For both of these two bonds a co-linear hydrogen bond restraint is introduced. 92 co-linear restraints (184 restraints in total) were produced using CCPN Analysis. The lower and upper bound for the H-O bound is 1.73 and 2.7 respectively. For the N-O distances these were 2.516 and 3.927. These are the default values.

Structure

The calculated structure shows the expected 14-strand β-barrel, figure 6.12. The backbone rmsd of this structure was 1.6 Å in the β-sheet region and 4.9 Å for all residues. As there are no restraints present for a large parts of the extra-cellular residues, they form unstructured loops in this model. In table 6.13 an overview is presented of the final assignment of the ADRs by ARIA and quality measures on the resulting ensemble of 15 structures. Almost all peaks in the ^{1}H-detected spectra were unambiguously assigned by ARIA, while a large part of the ^{13}C-^{13}C restraints remain ambiguous. Exact counts for restraints stemming from different spectra are shown in this table. Because there are identical peaks on both sides of the diagonal and peaks in different spectra correlating the same nulcei (for instance the 2-glycerol and 2-TEMPQANDSG spectra), a count is given for the amount of unique restraints in the set. Over the whole dataset, there are 196 unique long-range distance restraints, from which 131 could be assigned unambiguously by ARIA. The classification of ambiguous restraints into distance ranges was based on the restraint item with the shortest range. Therefor there are also long-range contributions present for some of the restraints classified as medium and sequential restraints. In order to not over-represent certain restraints, ARIA merges restraints that are containing the same set of restraint items. Therefor the amount of unique restraints is the "true" set of restraints on which the structural models are based. In figures 6.14 and 6.15 the assignment of the ADRs are shown on an interaction matrix for the respectively the ^{1}H-^{1}H restraints and the ^{13}C-^{13}C restraints.

Some of the restraints in the ^{13}C detected data were assigned to intra-residual correlations although the peak picking was performed in such a way to avoid intra-residual peaks. Peaks that were assigned as intra-residual were close to an intra-residual peak and therefor the shift-matching included this as one of the possibilities. To prevent this an extra filter could be applied during the shift-matching in future structure calculations, but was not done here yet. Although a few restraints were lost, the quality of the structure is not degraded as intra-residual distances are mostly well below the 8 Å upper bound independent of the local geometry.

Structure validation has been performed using the iCing server which analyses violations and runs several programs that check the normality of the structure [8]. One violation over 0.3 Å (0.31Å) was present in one of the 15 structures of the ensemble. It should be noted that the PROCHECK results, indicating how well the torsion angles in the structures fit to different regions in the ramachandran plot, are naturally almost perfect because a ramachandran potential has been used for the calculation of this structure. They have been included for completeness. The WHATIF RMS Z-scores show that there is a lower variability in the bond angles, side-chain planarity and amount of cis-conformations of the peptide bond (ω angles) than is expected from a database of high resolution x-ray structures. These are general problems in NMR structures and are caused by the way the force-field used for structure calculation operates and not so much by the dataset of this particular structure [9]. The unusual inside/outside distribution can be explained by the fact OmpG is a membrane protein.

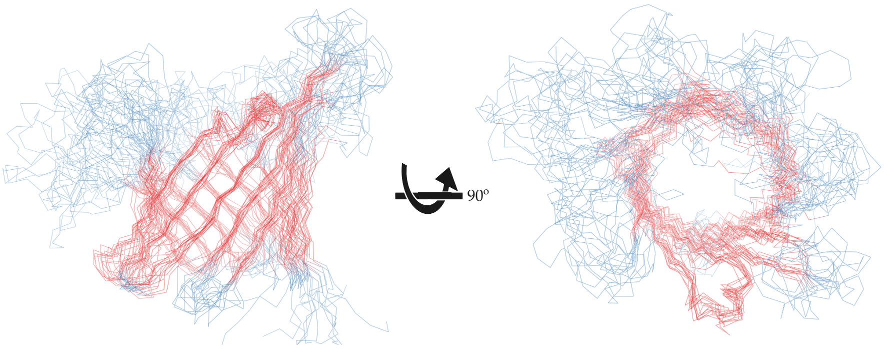


Figure 6.12: The 15 lowest energy structures in iteration 8 of the ARIA procedure when adding hydrogen bond restraints and using the ramachandran potential. Figure produced using pymol [10].

Remaining ambiguity of the ^{13}C-^{13}C restraints

488 of the restraints based on the peaks in the ^{13}C-^{13}C correlation spectra remained ambiguous at the end of ARIA protocol. In figure 6.16 the distribution of ^{13}C-^{13}C restraints over the different types of samples and their ambiguity is shown. The number of ADRs that remain ambiguously is relatively large. The remaining level of ambiguity for most ADRs is only 2, 3 or 4 though. The number of ambiguously assigned cross-peaks could potentially be reduced by manual inspection. However,

distance restraints	**all**	**^{1}H-detected**	hNHH	hNhhNH	**^{13}C-detected**	2-glycerol	1,3-glycerol	2-TEMPQANDSG	1,3-TEMPQANDSG	2-SHLYGWAFV
input to ARIA										
total	**1501**	249	122	127	1252	355	312	135	232	218
unambiguous	**105**	83	41	42	22	0	0	4	14	4
rejected during protocol	**155**	8	5	3	147	15	30	13	49	40
assignment in ARIA iteration 8										
total	**1346**	241	117	124	1105	340	282	122	183	178
distance class 1.0-3.5 Å	**139**	139	66	73						
distance class 1.0-5.5 Å	**102**	102	51	51						
distance class 1.5-8.0 Å	**1105**				1105	340	282	122	183	178
intra-residual	**59**	0	0	0	59	33	6	10	2	8
sequential	**788**	62	36	26	726	231	199	77	130	89
medium-range (2 ≤ | i-j | <5)	**141**	13	6	7	128	28	43	6	20	31
long-range (| i-j | ≥ 5)	**358**	166	75	91	192	48	34	29	31	50
unambiguous	**841**	224	106	118	617	165	125	102	124	101
unique	**765**	107	77	78	658	237	199	81	126	132
uique long-range	**196**	57	44	53	139	37	30	21	26	40
unique unambiguous long-range	**131**	54	42	52	77	11	17	15	17	28
violations in dmso refined OmpG[a]										
> 0.3 Å	**1**	0	0	0	0	0	1	0	0	0
> 0.5 Å	**0**	0	0	0	0	0	0	0	0	0

torsion angle restraints		**backbone rmsd**[b]		**PROCHECK**	
φ/ψ angles	128 (256 total)	β-sheet residues	1.57 ±0.45 Å	ramachandran	
rmsd	0.880 ± 0.098	vs. 2IWW (x-ray)	2.34 ±0.29 Å	core:	93.7%
violation count per model		vs. 2IWV (x-ray)	2.36 ±0.31 Å	allowed:	5.5%
violations > 1°	24.9 ±3.8	vs. 2F1C (x-ray)	2.29 ±0.30 Å	generous:	0.4%
violations > 3°	5.8 ±2.0	vs. 2JQY (solution NMR)	2.27 ±0.33 Å	disallowed:	0.5%
violations > 5°	1.5 ±0.7	β-sheet + turn residues	1.92 ±0.40 Å		
violations > 10°	0	all residues	4.89 ±0.54 Å		

hydrogen bond restraints		**WHATIF**	
92 colinear restraints (184 total)		Structure Z-scores, positive is better than average:	
violation > 0.3 Å	0	1st generation packing quality	0.356 ±1.201
		2nd generation packing quality	1.706 ±1.403
		Ramachandran plot appearance	0.499 ±0.246
		chi-1/chi-2 rotamer normality	3.047 ±0.529
		Backbone conformation	0.547 ±0.206
		RMS Z-scores, should be close to 1.0:	
		Bond lengths	0.971 ±0.001
		Bond angles	0.318 ±0.003 (tight)
		Omega angle restraints	0.723 ±0.035 (tight)
		Side chain planarity	0.327 ±0.026 (tight)
		Improper dihedral distribution	0.398 ±0.007
		Inside/Outside distribution	1.214 ±0.015 (unusual)

Table 6.13: Statistics on the restraints and quality metrics on the 15 lowest energy structures. All quality measures correspond to the structure refined in DMSO. Structure validation was performed using the iCing server [8] from which the PROCHECK [11] and WHATIF [12] were obtained. More precise counts for specific restraint subsets were obtained using a CCPNMR Analysis macro. a) Numbers are over the complete ensemble. 1 violation was present in 1 of the 15 models. b)Alignment of models within the ensemble and with structures 2IWW and 2IWV [13], 2F1C [14] and 2JQY [15] were calculated using biopython [16]. β-sheet residues are 8-16, 34-41, 44-51, 70-78, 85-95, 110-122, 127-139, 151-161, 167-175, 194-202, 205-211, 238-244, 249-255 and 274-280. Turn residues are 42-43, 79-84, 123-126, 162-166, 203-204 and 245-248.

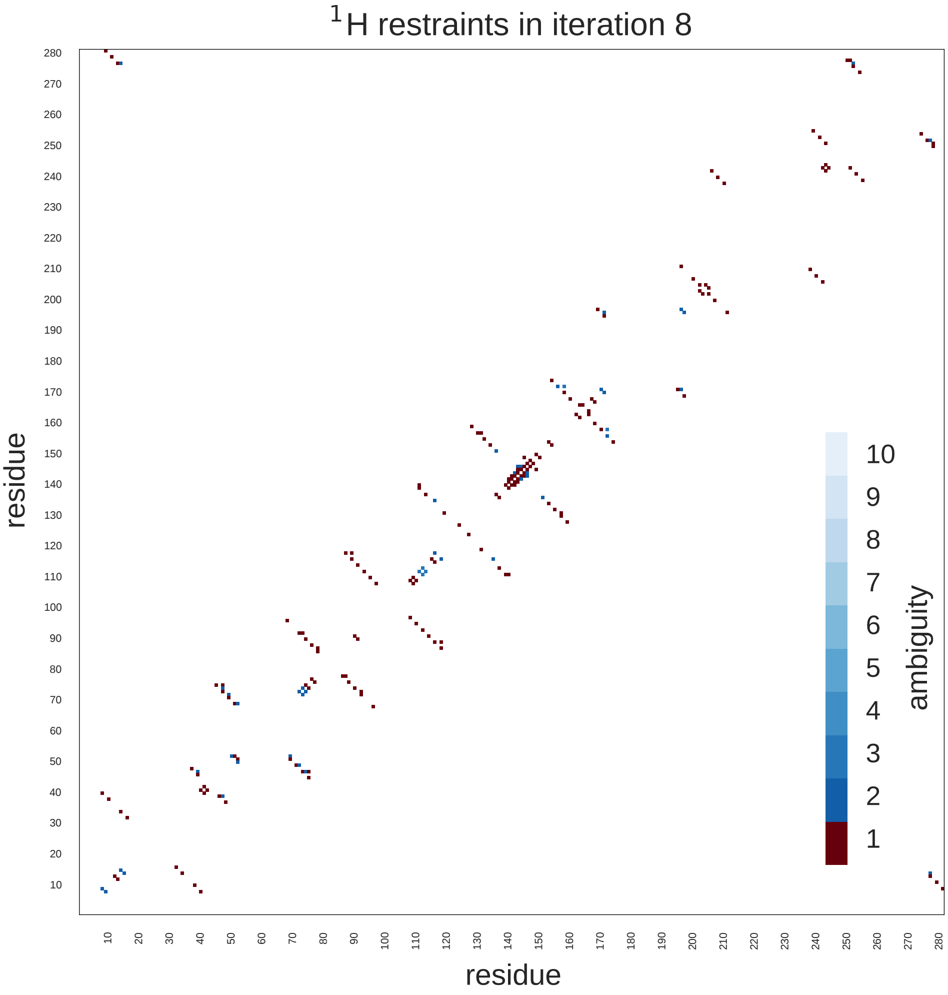


Figure 6.14: Assignment of ^{1}H-^{1}H ADRs in iteration 8 of the ARIA protocol. The color indicates the ambiguity of the least ambiguous restraint present for the interaction between two residues. Interactions between two residues for which an unambiguous restraint is present are colored red. A clear alternating pattern can be seen for the β-sheets. 11 restraints in the ^{1}H-^{1}H restraint set were left ambiguous at the end of the ARIA procedure.

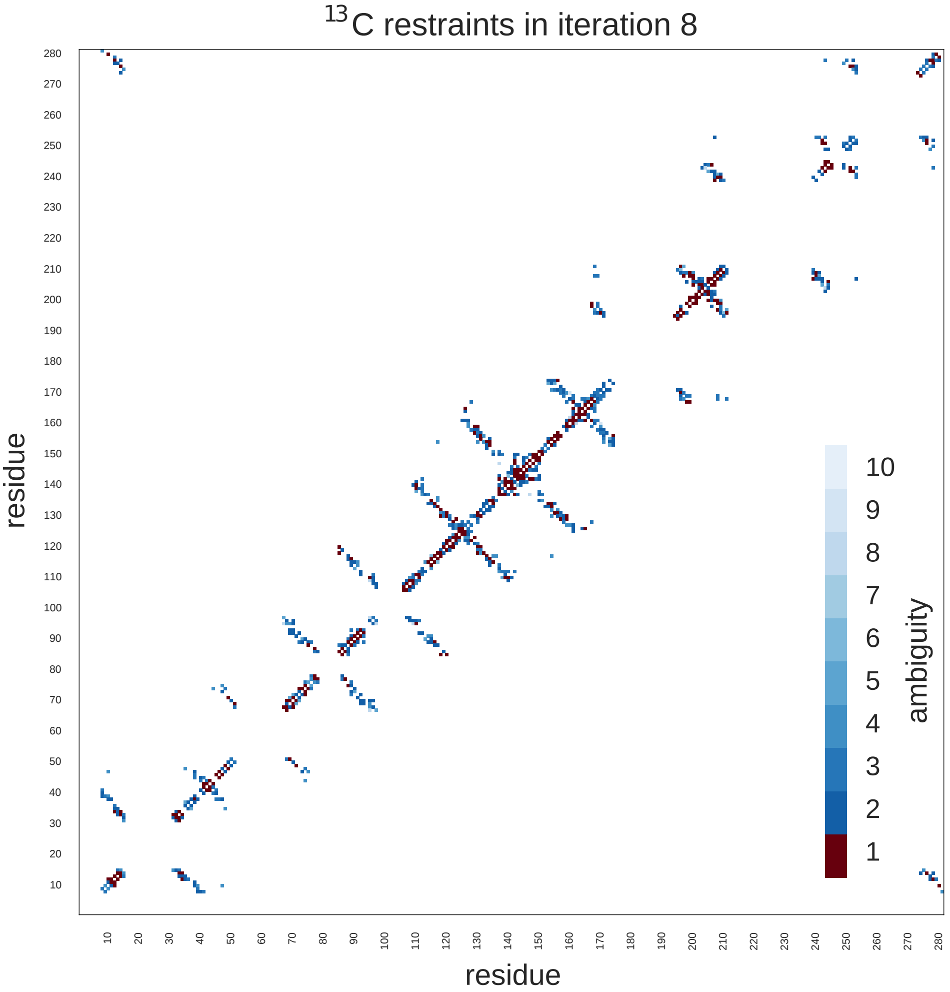


Figure 6.15: Assignment of ^{13}C-^{13}C ADRs in iteration 8 of the ARIA protocol. The color indicates the ambiguity of the least ambiguous restraint present for the interaction between two residues. Interactions between two residues for which an unambiguous restraint is present are colored red. 488 restraints in the ^{13}C-^{13}C restraint set were left ambiguous at the end of the ARIA procedure.

it is hard to define solid criteria on which basis to do this. Furthermore, for overlapped peaks, an ambiguous assignment is actually the correct assignment as there are multiple contributions to these peaks. Since there is a relatively high degree of peak overlap in these spectra, the large number of ADRs that remain ambiguous reflects the data.

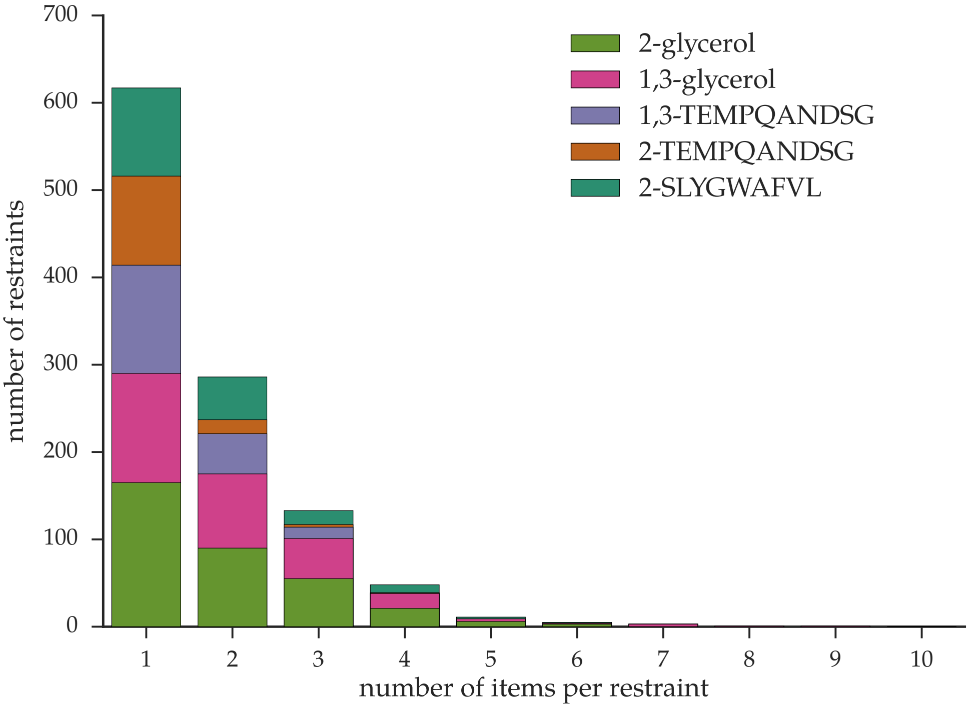


Figure 6.16: Ambiguity in the restraints in of the 5 ^{13}C-detected restraint sets after disambiguation by ARIA. Restraints with only one item are unambiguous.

Rejected restraints

In total 155 ADRs were rejected during the ARIA protocol, which is about 10% of the total number of ADRs (see table 6.13). There is no assignment for the corresponding peaks that fits the calculated structure. As argued before, it is very clear that the major reason for the absence of a correct restraint item is directly caused by missing assignments. In this perspective, the number of rejected peaks is actually lower than would be expected purely from statistics: with 10% of unassigned resonances in the aliphatic region, one would expect about 20% of the peaks to be rejected because one of the dimensions can not be correctly assigned.

In addition, a second explanation for these unexplained peaks is the potential presence of intermolecular contacts, since the packing of protein in the sample is relatively dense (with a protein to lipid ratio of 2:1 in terms of weight). Since the ambiguity of the restraints that were rejected is relatively high, it is difficult to find out whether this is indeed the case or not. To illustrate this, in figure 6.17 a residue interaction matrix is shown for the rejected ADRs. As can be seen, no specific regions in this matrix are represented stronger than would be expected, and therefore it is hard to tell whether there are specific parts in the structure in close proximity of a neighboring OmpG molecule. Also most inter-molecular contacts would be expected between the large side-chains of the aromatic residues. Because only the aliphatic part of the spectra was used for the structure calculation, these contacts are largely absent from the calculation.

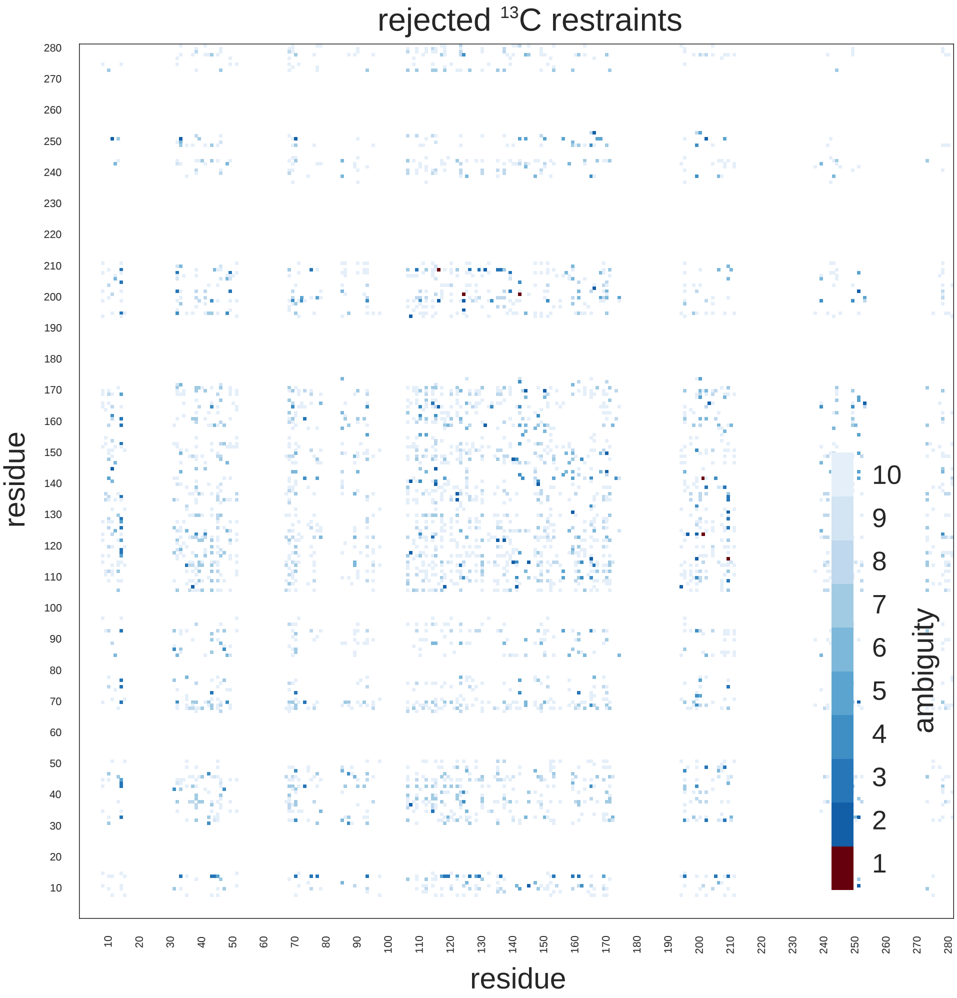


Figure 6.17: Residue interaction matrix for rejected restraints. There is no pattern

Comparison to crystal and solution NMR structures

As can be seen in table 6.13, the β-sheet region of the structure calculated here is in fairly good agreement with the structures determined by x-ray crystallography and solutions NMR, with a rmsd of around 2.3 Å. In contrast, the current structure strongly deviates from the crystal structures in the extra-cellular part of the molecule. Whereas here flexible loops are found, in the crystal structures the β-sheet continues almost entirely from the bottom to the top of the barrel. As discussed in the introduction this is likely caused by crystallization artifacts.

The solid state NMR structure is very similar to the solution NMR structure. The extend of the β-sheet is almost identical for most strands. The largest difference between the solid and solution structure is shown in figure 6.18: between strands 9 and 10 an additional set of NOE's between two pairs of NH groups could be observed in the liquid state. Hence, also two extra hydrogen bond restraints were added. In the solid state however, the corresponding stretch of residues (191 Thr, 192 Gln and 193 Glu) in strand 10 was not assigned. Therefore, no restraints are present between residues pairs 191 Thr-175 Glu and 193 Glu-173 Tyr. Thr 191 is one of the two unassigned threonines shown in figures 4.1 and 4.3. Because threonines are in general easy to assign, because of their distinct finger print pattern, it is clear that the signal pattern necessary for the assignment was really absent. The result of these missing assignments and restraints is that the β-sheet extends less far on strand 10 in the solid state structure.

Materials and Methods

Sample preparation

The sample used was prepared as described in chapter 3 at pH 6.3 with 100% back-exchanged protons, with the only exception that the polar lipid extract (Avanti Polar Lipids) was used instead of the total lipid extract. No differences were observed between the ^{1}H-^{15}N and hCANH spectra recorded on this new sample and the sample used for the assignment experiments.

NMR experiments

Experiments were recorded on an Bruker Avance III 1000 MHz ^{1}H larmor frequency spectrometer at 60 kHz MAS using a triple-resonance HCN 1.3 mm probe. The temperature of the VT gas flow was set to 230 K, which roughly corresponds to a sample temperature of 300 K. 90°-pulses were 2.5 µs (100 kHz) for ^{1}H, 3.5 µs (71 kHz) for ^{13}C and 5.5 µs (45 kHz) for ^{15}N. CP steps from ^{1}H to ^{15}N had a duration of 700 µs. The ^{1}H spin lock amplitude was centered on 8 kHz with a 30% linear ramp. The ^{15}N spin lock field had a constant amplitude of 32 kHz. The CP steps from ^{15}N to ^{1}H had a duration of 300 µs. The ^{1}H spin lock field amplitude was centered on 5 kHz with a

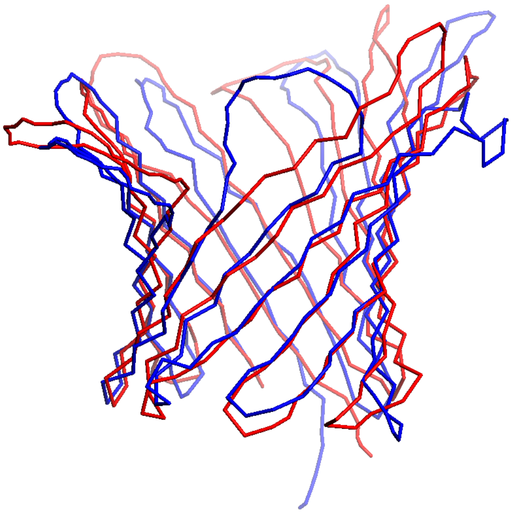


Figure 6.18: Overlay of aligned average solid-state (blue) and solution (red) NMR structures. The largest difference between the two structures is shown in the foreground. The beta-sheet is extended further in the solution model. An additional two long range hydrogen bonds are present in the solution structure. A stretch of three residues (191 Thr, 192 Gln, 193 Glu) showing these connections to the preceding strand in the solution spectra could not be assigned in the solid state. Figure produced using pymol [10].

30% linear ramp. The ^{15}N spin lock field had a constant amplitude of 34 kHz. Water suppression was achieved using the MISSISSIPI sequence without homospoil gradients [17]. Swept-low-power TPPM was used for ^{1}H decoupling and WALTZ-16 for ^{15}N and ^{13}C decoupling during ^{1}H-detection [18][19]. All spectra were acquired using the States-TPPI in the direct dimensions to obtain pure-phase line shapes and phase discrimination [20]. For the hNHH experiment the acquisition times in the indirect dimensions were set to 4.7 and 12.1 ms for ^{1}H and ^{15}N, respectively, with 8 scans per increment and a total experiment time of 3 days. For the hNhhNH experiment, the acquisition time for the ^{15}N dimension acquired before the through-space transfer, the acquisition time was set to 15.4 ms. The acquisition time of the second ^{15}N dimension, that corresponds to the ^{15}N in the same amide group as the correlated ^{1}H, was set to 10.7 ms. The number of scans per increment was 16 and the total experiment time 7 days. The hCANH measured for calibration of the chemical shifts was recorded in the same fashion as descried in chapter 3.

References

[1] J. P. Linge, M. Habeck, W. Rieping, and M. Nilges. "ARIA: Automated NOE Assignment and NMR Structure Calculation". *Bioinformatics* 19.2 (Jan. 2003), pp. 315–316. doi: 10.1093/bioinformatics/19.2.315.

[2] W. Rieping, M. Habeck, B. Bardiaux, A. Bernard, T. E. Malliavin, and M. Nilges. "ARIA2: Automated NOE Assignment and Data Integration in NMR Structure Calculation". *Bioinformatics* 23.3 (Jan. 2007), pp. 381–382. doi: 10.1093/bioinformatics/btl589.

[3] G. Cornilescu, F. Delaglio, and A. Bax. "Protein Backbone Angle Restraints from Searching a Database for Chemical Shift and Sequence Homology". *Journal of Biomolecular NMR* 13.3 (Mar. 1999), pp. 289–302. doi: 10.1023/A:1008392405740.

[4] Y. Shen, F. Delaglio, G. Cornilescu, and A. Bax. "TALOS+: A Hybrid Method for Predicting Protein Backbone Torsion Angles from NMR Chemical Shifts". *Journal of Biomolecular NMR* 44.4 (June 2009), pp. 213–223. doi: 10.1007/s10858-009-9333-z.

[5] P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas. "PRED-TMBB: A Web Server for Predicting the Topology of β-Barrel Outer Membrane Proteins". *Nucleic Acids Research* 32.suppl 2 (Jan. 2004), W400–W404. doi: 10.1093/nar/gkh417.

[6] T. Herrmann, P. Güntert, and K. Wüthrich. "Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA". *Journal of Molecular Biology* 319.1 (May 2002), pp. 209–227. doi: 10.1016/S0022-2836(02)00241-3.

[7] J. Jee and P. Güntert. "Influence of the Completeness of Chemical Shift Assignments on NMR Structures Obtained with Automated NOE Assignment". *Journal of Structural and Functional Genomics* 4.2-3 (June 2003), pp. 179–189. doi: 10.1023/A:1026122726574.

[8] J. F. Doreleijers, W. F. Vranken, C. Schulte, J. L. Markley, E. L. Ulrich, G. Vriend, and G. W. Vuister. "NRG-CING: Integrated Validation Reports of Remediated Experimental Biomolecular NMR Data and Coordinates in wwPDB". *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D519–D524. doi: 10.1093/nar/gkr1134.

[9] C. A. Spronk, S. B. Nabuurs, E. Krieger, G. Vriend, and G. W. Vuister. "Validation of Protein Structures Derived by NMR Spectroscopy". *Progress in Nuclear Magnetic Resonance Spectroscopy* 45.3-4 (Dec. 2004), pp. 315–337. doi: 10.1016/j.pnmrs.2004.08.003.

[10] W. Delano. "The PyMOL Molecular Graphics System" (2002).

[11] R. A. Laskowski, J. A. C. Rullmann, M. W. MacArthur, R. Kaptein, and J. M. Thornton. "AQUA and PROCHECK-NMR: Programs for Checking the Quality of Protein Structures Solved by NMR". *Journal of Biomolecular NMR* 8.4 (Dec. 1996), pp. 477–486. DOI: 10.1007/BF00228148.

[12] G. Vriend. "WHAT IF: A Molecular Modeling and Drug Design Program". *Journal of Molecular Graphics* 8.1 (Mar. 1990), pp. 52–56. DOI: 10.1016/0263-7855(90)80070-V.

[13] Ö. Yildiz, K. R. Vinothkumar, P. Goswami, and W. Kühlbrandt. "Structure of the Monomeric Outer-membrane Porin OmpG in the Open and Closed Conformation". *The EMBO Journal* 25.15 (Aug. 2006), pp. 3702–3713. DOI: 10.1038/sj.emboj.7601237.

[14] G. V. Subbarao and B. van den Berg. "Crystal Structure of the Monomeric Porin OmpG". *Journal of Molecular Biology* 360.4 (July 2006), pp. 750–759. DOI: 10.1016/j.jmb.2006.05.045.

[15] B. Liang and L. K. Tamm. "Structure of Outer Membrane Protein G by Solution NMR Spectroscopy". *Proceedings of the National Academy of Sciences* 104.41 (Sept. 2007), pp. 16140–16145. DOI: 10.1073/pnas.0705466104.

[16] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics". *Bioinformatics* 25.11 (Jan. 2009), pp. 1422–1423. DOI: 10.1093/bioinformatics/btp163.

[17] D. H. Zhou and C. M. Rienstra. "High-Performance Solvent Suppression for Proton Detected Solid-State NMR". *Journal of Magnetic Resonance* 192.1 (May 2008), pp. 167–172. DOI: 10.1016/j.jmr.2008.01.012.

[18] J. R. Lewandowski, J. Sein, M. Blackledge, and L. Emsley. "Anisotropic Collective Motion Contributes to Nuclear Spin Relaxation in Crystalline Proteins". *Journal of the American Chemical Society* 132.4 (Feb. 2010), pp. 1246–1248. DOI: 10.1021/ja907067j.

[19] A. J. Shaka, J. Keeler, T. Frenkiel, and R. Freeman. "An Improved Sequence for Broadband Decoupling: WALTZ-16". *Journal of Magnetic Resonance (1969)* 52.2 (Apr. 1983), pp. 335–338. DOI: 10.1016/0022-2364(83)90207-X.

[20] D. Marion, M. Ikura, R. Tschudin, and A. Bax. "Rapid Recording of 2D NMR Spectra without Phase Cycling. Application to the Study of Hydrogen Exchange in Proteins". *Journal of Magnetic Resonance (1969)* 85.2 (Nov. 1989), pp. 393–399. DOI: 10.1016/0022-2364(89)90152-2.

Chapter 7

General Discussion and Outlook

The goal of these studies was to develop a new solid-state NMR methodology for the structure determination of membrane proteins in native lipid bilayers. OmpG serves as a good model system for the development of such new methods. Because of its size and the non-crystalline nature of the sample. It provides the challenges that make it necessary to explore different strategies for sequential assignment and structure calculation.

Different labeling strategies were explored to be able to achieve sequence-specific assignments using ^{13}C-detected experiments. This strategy gave access to a "starting assignment". However, the use of ^{1}H-detected experiments was absolutely necessary to arrive at a complete assignment of the folded part. The assignment strategy based on ^{1}H-detected experiments turned out to be more robust since the addition of an independent nucleus and thereby a spectral dimension contributed enormously to the dispersion of the signals.

The successful application of ^{1}H-detection to a variety of systems, among which OmpG, already caused ^{1}H-detection to become more common and will undoubtedly become the standard detection method in solid-state NMR within the next few years. 3 years ago only a few laboratories owned an ultra fast spinning probe, however, now the field is quickly adapting. Still, the methodology requires further improvements. For example, it would be interesting to compare 3D spectra recorded with different ^{13}C-^{13}C transfer methods, as it seems that a substantial amount of signal is lost during the scalar coupling based methods used so far in this project. Sequences optimized for proteins with short T_2 need to be developed. Furthermore, ^{1}H-detected experiments that give access to sidechain chemical shifts should be tested. In our case, it was possible to assign OmpG ^{13}C sidechain chemical shifts with the aid of ^{13}C-detected experiments that made use of a completely different set of isotope labeled samples as those that were used for the ^{1}H-detected experiments. It would be favorable to only use one sample to obtain the spectra necessary for the sequential assignment and the assignment of sidechain chemical shifts. As discussed in chapter 4, a number of different ^{1}H-detected approaches have been developed to access sidechain chemical shifts. However, more experiments are necessary to evaluate approach is most effective for a membrane protein preparation like OmpG.

For the structure calculation it was very important that the restraint sets based on ^{1}H-detected and ^{13}C-detected experiments could enter the calculation at different stages. In addition, restraint combination was crucial to minimize the negative effect of ambiguous restraints that did not contain the correct assignment option. Because OmpG is a β-barrel, ^{1}H-^{1}H restraints between amide groups in the backbone play a large role in defining the contacts between the individual β-strands. Therefore, the correct topology of the strands is already found in the first iteration of the ARIA protocol. This in turn restricts the assignment options of the ^{13}C-^{13}C restraints that enter the calculation at a later stage. When determining the structure of proteins with other topologies, restraints between sidechains are more important for finding the correct relative orientation between different structural elements. In this case, the structure calculation protocol described here, using only ^{1}H-^{1}H restraints between exchangeable protons in the first iterations, may need to be modified. To be able to let the ^{13}C-^{13}C restraints enter the calculation in the very first iteration, the fraction of assigned ^{13}C chemical shifts should be higher. An alternative would be to measure ^{1}H-^{1}H restraints between labeled methyl groups. Yet another option would be to use even higher MAS rates, so that fully protonated protein samples can be studied.

The use of fully protonated proteins and faster MAS frequencies (~100 kHz) might also be necessary for proteins that can not (in contrast to OmpG) be easily refolded. In these proteins, the exchange of deuterons by protons within secondary structure elements will not be efficient as all amide protons will be involved in a hydrogen bond. It might be an option to exploit deuterium double quantum chemical shifts in this case.

Besides using OmpG as a system to further improve solid-state NMR methodology, the derived structure is of high interest with regards to addressing some of the open questions surrounding the pH-dependent opening and closing mechanism of this porin. The derived structure, like the solution structure, contains large flexible loops on the extra-cellular side. This contrasts with the crystal structures for which the β-sheet extends far beyond the membrane interface, stabilized by crystal contacts. It could be investigated whether more signals, and therefore more structure content, are present in spectra of OmpG mutants for which the spontaneous gating is minimized, as described in the general introduction.

During the assignment process, which was by far the most time consuming part of these studies, it was important to be aided by good software. For this reason, the comprehensive data model provided by CCPNMR was of great value. Furthermore, it is very important to be able to get a fast overview of the assignments that have been made and of those that are still missing. Because methodologies in solid-sate NMR are quickly evolving, the possibility to extend CCPNMR Analysis by writing macros is very useful. This can be used to provide computational tools that fit newly developed methodology. The plug-ins for CCPNMR Analysis written in the context of these studies proved to be very helpful during the assignment process.

Appendix A

Chemical Shifts

#	aa	H	N	C	C(D)	CA	CA(D)	CB	CB(D)	other
8	His	8.79	118.54	174.14	174.51	55.00	54.96	31.76	-	CG:131.26, CD2:120.99
9	Phe	9.30	123.48	-	174.38	57.12	57.32	43.36	43.02	CG:139.70
10	Asn	8.69	116.65	173.88	173.70	54.27	54.23	44.09	43.79	ND2:115.80, HD22:6.71, CG:178.50, CG(D):178.49
11	Ile	9.46	116.61	175.45	175.43	59.66	59.45	43.48	43.00	CG1:28.00, CG2:18.90, CD1:14.46
12	Gly	7.83	110.65	170.37	170.44	47.12	46.93			
13	Ala	8.56	116.64	176.09	175.96	51.06	51.08	24.12	23.73	
14	Met	9.39	120.01	-	-	55.09	55.05	37.87	37.34	CG:32.67, CE:17.07
15	Tyr	9.29	122.46	174.86	-	59.58	59.54	42.93	-	CG:132.80
16	Glu	7.48	126.09	-	-	-	54.24	-	-	
31	Ala	-	-	178.90	178.97	51.83	51.82	22.06	-	
32	Glu	8.76	121.47	-	-	52.42	52.42	29.17	-	CG:32.40
33	Pro		141.11	175.85	176.01	62.50	62.45	33.58	-	CG:27.14,CD:49.47
34	Ser	9.45	110.56	173.66	172.75	58.83	58.97	67.42	-	
35	Val	9.05	111.50	-	-	59.55	59.48	36.17	36.21	CG2:21.30
36	Tyr	-	-	-	-	57.06	57.83	42.37	41.70	CG:130.70
37	Phe	9.23	118.65	-	174.50	56.48	56.47	45.31	44.81	CG:139.46
38	Asn	9.35	122.09	172.09	172.18	53.13	53.11	42.73	42.69	ND2:109.23, HD21:5.74, CG(D):175.25
39	Ala	9.13	120.45	175.31	175.33	51.08	51.07	25.18	24.48	
40	Ala	9.01	123.63	175.94	176.05	51.27	51.17	24.06	23.55	
41	Asn	8.09	118.36	175.45	175.59	53.23	53.20	39.30	38.88	
42	Gly	8.94	116.95	174.15	174.03	45.40	45.29			
43	Pro		137.23	176.41	176.43	63.66	63.68	32.47	31.92	CG:26.62, CD:51.19
44	Trp	8.06	122.58	177.62	177.61	57.63	57.74	33.43	33.27	CG:112.64, CD2:131.13, CZ3:121.90
45	Arg	9.23	121.66	175.01	174.98	56.57	56.54	34.88	34.07	CG:28.65, CD:44.04, CZ:159.58

#	aa	H	N	C	C(D)	CA	CA(D)	CB	CB(D)	other
46	Ile	9.77	126.91	173.59	173.86	60.92	60.88	42.33	41.46	CG1:29.52, CG2:19.09, CD1:15.53
47	Ala	9.60	128.89	175.58	175.45	51.72	51.62	24.09	23.46	
48	Leu	9.62	123.03	176.55	176.53	54.34	54.27	47.18	46.28	CG:28.92, CD1:25.98
49	Ala	9.20	122.72	175.47	175.42	52.29	52.28	24.52	23.93	
50	Tyr	9.09	118.41	-	172.07	60.07	60.07	42.18	42.09	CG:129.07
51	Tyr	6.14	128.30	-	171.80	55.39	55.50	43.41	43.31	CG:128.42
52	Gln	7.44	126.64	-	-	-	54.55	-	34.86	
67	Phe	8.77	114.63	-	173.78	56.33	56.34	42.12	42.20	CG:139.11
68	Asp	9.03	118.60	-	-	52.19	52.31	42.79	42.35	
69	Arg	8.28	118.60	174.45	-	53.76	53.80	36.48	-	CG:26.89, CD:44.81, CZ:159.36
70	Pro		139.69	176.03	176.09	63.54	63.74	33.80	32.89	CG:28.49, CD:50.82
71	Glu	8.96	119.63	-	-	55.28	55.39	-	35.67	
72	Leu	8.96	125.08	174.73	174.73	54.66	54.71	47.98	46.89	CG:28.43, CD2:26.40
73	Glu	9.74	121.95	-	-	55.90	55.97	-	34.71	CG:37.82
74	Val	9.40	120.10	173.63	173.59	61.05	60.92	36.02	35.39	CG2:22.10
75	His	9.78	125.16	172.72	172.56	54.59	54.71	32.58	32.71	CG:130.90, CD2:122.14, CE1:135.92
76	Tyr	8.31	125.20	173.04	173.42	57.03	56.63	42.54	42.70	CG:129.79
77	Gln	7.88	129.04	173.79	173.67	54.09	54.05	28.78	28.21	CG:33.50, CD:179.66
78	Phe	8.22	125.75	-	-	62.48	62.52	41.22	-	CG:140.43
85	Ser	-	-	-	-	58.00	57.83	66.74	67.03	
86	Phe	8.67	121.77	-	-	58.66	58.83	45.02	44.68	CG:139.32, CE*:129.91
87	Gly	8.80	118.46	170.20	170.43	45.44	45.23			
88	Leu	7.95	117.12	174.15	174.25	54.68	54.78	49.47	48.72	CG:27.61, CD1:24.09
89	Thr	8.54	121.84	174.56	174.72	62.27	62.19	71.38	71.35	CG2:22.70
90	Gly	9.60	113.64	172.95	172.83	44.67	44.58			
91	Gly	9.43	109.20	170.36	170.33	46.49	46.40			
92	Phe	9.23	120.61	173.19	173.31	57.25	57.35	44.43	43.75	CG:138.57
93	Arg	7.65	124.03	175.13	174.99	54.50	54.34	37.32	36.82	CG:26.55, CD:44.11, CZ:159.21
94	Asn	7.65	117.28	-	-	-	53.46	-	-	
95	Tyr	9.53	122.93	-	175.21	56.91	56.91	39.14	-	CG:133.59
96	Gly	9.26	112.35	172.04	171.94	46.06	45.95			
97	Tyr	8.55	122.82	-	-	59.47	59.64	42.46	-	CG:130.90
106	Asp	-	-	-	-	56.32	-	42.65	-	CG:180.43
107	Thr	-	-	174.52	-	60.16	60.25	73.59	-	CG2:22.37
108	Ala	8.12	120.76	175.77	176.17	52.19	52.26	24.21	-	
109	Asn	10.06	121.40	171.96	172.10	54.38	54.55	42.61	43.17	
110	Met	8.60	126.23	175.71	175.52	53.53	53.70	37.36	36.92	CG:31.38, CE:14.44
111	Gln	9.68	124.66	-	175.18	56.88	56.82	30.89	30.84	CG:36.84
112	Arg	8.52	118.36	175.05	175.20	54.84	54.69	37.05	36.16	CG:26.99, CD:44.76, CZ:159.77
113	Trp	9.68	132.86	175.06	175.31	56.16	56.23	32.22	31.86	CG:112.20, CD2:130.16, CZ3:121.53
114	Lys	9.17	124.43	-	175.15	55.31	55.40	40.27	39.33	CG:25.92, CD:32.31, CE:42.49

#	aa	H	N	C	C(D)	CA	CA(D)	CB	CB(D)	other
115	Ile	8.68	123.65	174.83	175.27	59.81	60.08	42.38	41.68	CG1:29.51, CG2:18.42, CD1:14.59
116	Ala	9.10	125.71	176.80	176.56	50.44	50.38	24.67	23.83	
117	Pro		143.54	176.39	176.08	62.83	62.81	34.78	34.29	CG:28.25, CD:51.34
118	Asp	8.80	116.82	173.93	173.90	54.05	54.20	45.17	45.05	CG:180.79
119	Trp	8.25	115.28	175.38	175.51	57.54	57.44	33.71	33.16	CG:112.92, CD1:124.87
120	Asp	8.37	119.57	-	174.97	55.29	55.55	44.37	43.98	
121	Val	9.79	126.81	-	-	61.31	61.21	36.20	35.81	
122	Lys	8.64	128.38	-	-	58.82	58.67	34.79	33.82	CG:26.00, CD:30.16, CE:42.83
123	Leu	-	-	177.62	177.66	56.37	56.36	42.27	42.68	CG:26.86, CD1:21.17
124	Thr	8.68	109.36	172.87	-	59.73	59.98	72.27	72.67	CG2:22.57
125	Asp	-	-	-	-	58.08	-	41.18	-	CG:180.08
126	Asp	-	-	-	173.97	54.36	54.41	44.11	44.04	
127	Leu	7.71	122.79	175.17	175.15	54.40	54.52	46.71	46.01	CG:27.86, CD1:25.14
128	Arg	9.24	124.16	174.65	174.44	54.88	54.98	34.77	34.54	CZ:159.79
129	Phe	9.34	122.06	173.79	173.61	54.60	54.90	42.39	42.34	CG:139.31
130	Asn	9.00	123.79	173.98	173.99	51.18	51.15	42.57	42.38	
131	Gly	7.13	104.70	169.12	169.14	45.79	45.71			
132	Trp	6.90	113.24	175.16	174.85	55.05	55.17	32.14	31.85	NE1:129.70, HE1:10.01, CG:110.77, CD2:130.18, CE3:120.58, CZ3:121.41
133	Leu	8.71	124.69	173.56	173.67	55.37	55.49	45.98	45.43	CG:28.41, CD1:26.58
134	Ser	10.28	119.15	173.06	172.92	57.06	57.29	67.63	67.63	
135	Met	8.80	123.95	-	-	54.54	54.41	-	-	CG:33.31, CE:17.07
136	Tyr	9.36	119.57	-	-	54.66	54.80	42.38	42.27	CG:130.77
137	Lys	8.93	119.88	-	-	53.08	53.18	35.70	34.80	CG:24.58, CD:29.74, CE:42.43
138	Phe	9.10	124.90	-	-	57.04	57.24	43.01	41.78	CG:141.86
139	Ala	9.22	123.33	175.48	175.58	52.30	52.37	24.25	24.04	
140	Asn	8.59	111.28	172.60	172.45	57.53	57.57	39.19	38.43	CG:179.16
141	Asp	8.39	112.18	177.66	177.80	56.22	56.28	39.53	-	
142	Leu	8.29	116.59	181.34	181.38	58.05	58.09	40.53	39.79	CG:27.42, CD1:23.55
143	Asn	8.93	115.87	177.11	177.08	55.73	55.80	37.51	37.41	
144	Thr	7.62	116.19	175.20	175.42	65.72	65.63	69.25	69.34	CG2:23.57
145	Thr	8.60	110.15	177.41	177.41	62.89	63.03	69.87	69.97	CG2:22.86
146	Gly	8.27	109.14	174.12	174.11	46.18	46.11			
147	Tyr	7.09	120.29	-	-	58.40	58.58	39.35	39.19	CG:130.75
148	Ala	8.19	122.78	179.02	178.94	51.36	51.35	18.39	18.07	
149	Asp	8.26	115.48	176.72	176.82	56.88	56.91	42.22	41.71	CG:180.83
150	Thr	8.44	118.09	172.76	173.37	62.47	62.44	71.42	71.32	CG2:23.57
151	Arg	8.94	124.85	174.38	173.80	55.04	54.91	32.08	31.40	CG:26.74, CD:44.04, CZ:159.74
152	Val	-	-	173.19	173.57	59.85	59.96	35.82	35.68	
153	Glu	9.11	127.77	173.56	173.67	54.06	54.10	36.36	36.25	CG:36.92, CD:183.77
154	Thr	9.02	122.10	169.32	169.43	60.06	59.99	70.11	70.21	CG2:19.12
155	Glu	5.43	123.92	176.17	175.89	56.16	56.23	31.33	31.00	CG:37.01, CD:182.24
156	Thr	8.94	122.69	171.80	171.85	60.13	60.25	70.82	70.94	CG2:20.65

#	aa	H	N	C	C(D)	CA	CA(D)	CB	CB(D)	other
157	Gly	8.73	113.81	172.06	172.02	48.05	47.90			
158	Leu	8.53	118.44	175.95	176.08	52.93	53.09	46.85	46.11	CG:27.95, CD2:24.33
159	Gln	9.10	120.29	174.79	174.78	54.23	54.02	33.37	32.70	
160	Tyr	9.92	130.77	174.47	174.70	56.64	56.59	42.43	42.22	CG:128.86, CE*:118.27
161	Thr	8.33	124.30	174.11	173.82	63.24	63.06	70.09	70.05	CG2:22.16
162	Phe	8.42	125.66	176.25	176.53	61.40	61.36	39.55	39.71	CG:140.97
163	Asn	8.24	111.28	173.81	174.17	52.76	52.96	39.37	38.50	
164	Glu	8.93	114.64	176.59	176.56	59.34	59.43	29.82	29.12	CG:36.98
165	Thr	8.76	116.01	173.98	173.91	64.57	64.58	70.36	70.38	CG2:22.79
166	Val	7.96	118.92	173.87	174.17	61.40	61.55	35.39	35.01	CG2:21.23
167	Ala	8.19	128.85	172.84	172.97	51.00	50.86	23.46	23.23	
168	Leu	9.06	116.78	175.62	175.68	54.21	54.26	48.28	47.70	CG:28.70, CD2:27.71
169	Arg	9.42	124.62	175.72	175.68	54.87	54.83	34.74	34.19	NE:123.69, HE:7.13, CG:27.76, CD:44.23, CZ:159.67
170	Val	8.94	122.64	172.94	172.99	63.15	63.06	34.70	34.00	CG2:21.92
171	Asn	9.64	121.89	175.20	174.72	51.22	51.33	43.03	42.89	
172	Tyr	9.69	122.30	-	173.08	57.43	57.51	42.07	41.86	
173	Tyr	8.49	127.92	172.02	172.10	55.81	56.13	42.71	43.14	CD*:132.88
174	Leu	7.56	126.58	173.92	174.17	52.66	52.99	46.12	45.82	CG:28.47, CD1:24.20
175	Glu	8.76	124.20	-	-	-	55.08	-	32.83	
194	Ile	8.84	117.56	175.22	175.45	59.37	59.58	39.41	39.18	
195	Arg	9.48	126.35	173.46	173.61	55.10	55.13	34.00	33.36	CG:28.62, CD:44.83, CZ:159.72
196	Ala	8.57	124.50	175.49	175.42	49.88	49.91	22.20	22.11	
197	Tyr	9.64	117.50	174.70	174.55	56.17	56.27	43.34	42.69	CG:129.02, CE*:119.61
198	Leu	8.35	123.73	171.11	-	51.72	51.75	45.28	44.60	CG:27.94, CD2:26.10
199	Pro		134.46	177.55	177.67	64.53	64.44	32.45	32.00	CG:28.47, CD:49.83
200	Leu	9.51	127.18	179.14	178.94	53.28	53.55	43.97	43.30	CG:26.73
201	Thr	9.60	122.03	174.02	174.11	61.94	61.91	70.24	70.29	CG2:21.99
202	Leu	8.46	128.36	176.40	176.49	52.38	52.41	43.46	42.90	CG:26.84, CD1:25.34
203	Gly	8.57	111.68	176.04	176.00	47.59	47.53			
204	Asn	9.38	126.73	174.43	174.41	54.73	54.74	39.28	38.94	
205	His	8.53	121.16	176.14	176.16	55.83	55.89	33.32	33.00	CG:137.33, CD2:120.81, CE1:137.72
206	Ser	9.40	120.26	173.11	173.06	57.81	57.96	64.96	65.23	
207	Val	8.73	124.32	175.15	175.21	61.29	61.30	35.87	35.05	
208	Thr	9.98	121.83	173.58	-	59.44	59.38	71.37	71.27	CG2:23.91
209	Pro		141.20	175.95	176.11	62.40	62.40	33.19	32.72	CG:28.28, CD:51.18
210	Tyr	9.02	118.76	-	172.95	57.65	57.13	42.67	42.56	CG:130.73
211	Thr	9.01	111.45	-	-	59.40	59.47	71.34	71.34	CG2:19.67
237	Val	-	-	-	173.45	60.60	61.09	-	-	
238	Gly	8.26	110.28	170.51	170.44	46.12	45.85			
239	Leu	9.21	117.91	174.26	174.47	55.13	55.18	47.11	46.50	CG:27.84, CD2:26.73
240	Phe	8.26	124.23	-	173.56	57.10	57.11	42.63	42.49	CG:140.00
241	Tyr	9.42	130.01	-	173.45	56.78	56.83	43.19	42.70	CG:128.37
242	Gly	8.66	113.97	170.86	170.74	44.49	44.46			
243	Tyr	8.18	118.93	-	173.76	58.27	58.45	42.49	42.10	CG:130.77, CE*:117.78

#	aa	H	N	C	C(D)	CA	CA(D)	CB	CB(D)	other
244	Asp	6.58	125.91	-	-	52.77	52.75	42.86	42.88	
245	Phe	8.67	121.07	-	-	60.54	60.49	39.23	-	CG:142.32, CE*:131.58
249	Leu	-	124.01	175.75	176.05	55.18	55.28	45.14	44.39	CG:28.06, CD2:26.66
250	Ser	9.78	119.57	173.03	172.09	57.29	58.29	68.16	66.82	
251	Val	8.95	115.11	175.34	175.43	59.23	59.38	37.19	36.64	CG1:20.15
252	Ser	9.46	119.44	172.88	172.82	57.06	57.40	67.67	68.04	
253	Leu	9.67	121.81	175.71	175.19	54.56	54.77	47.62	47.09	CG:28.54, CD2:27.09
254	Glu	9.22	122.31	-	-	-	55.43	-	-	
255	Tyr	8.53	123.71	-	-	-	57.06	-	-	
272	Tyr	-	-	172.09	-	-	-	42.03	-	
273	Ala	-	-	175.53	175.89	51.43	51.26	24.19	-	
274	Gly	8.93	105.19	172.77	172.31	46.42	46.37			
275	Val	8.94	119.84	173.53	173.78	59.70	59.72	36.09	35.59	CG1:21.89
276	Gly	9.23	112.05	171.99	172.01	46.18	45.91			
277	Val	8.89	115.95	-	-	60.03	60.18	36.29	35.87	
278	Asn	9.05	123.38	172.95	172.88	53.43	53.50	44.99	44.68	ND2:114.72, HD21:7.25, CG:178.11
279	Tyr	10.09	126.78	-	173.69	57.01	57.30	42.20	41.94	
280	Ser	8.33	123.97	173.89	173.65	56.49	56.44	66.27	66.36	
281	Phe	8.69	127.12	-	-	58.57	58.78	-	-	CG:141.02

Appendix B

Publications

Lalli D, Schanda P, Chowdhury A, **Retel J**, Hiller M, Higman VA, Handel L, Agarwal V, Reif B, van Rossum B, Akbey Ü, Oschkinat H (2011) Three-dimensional deuterium-carbon correlation experiments for high-resolution solid-state MAS NMR spectroscopy of large proteins. Journal of Biomolecular NMR 51: 477–485. doi: http://dx.doi.org/10.1007/s10858-011-9578-1.

Barbet-Massin E, Pell AJ, Jaudzems K, Franks WT, **Retel JS**, Kotelovica S, Akopjana I, Tars K, Emsley L, Oschkinat H, Lesage A, Pintacuda G (2013) Out-and-back 13C13C scalar transfers in protein resonance assignment by proton-detected solid-state NMR under ultra-fast MAS. Journal of Biomolecular NMR 56: 379–386. doi: http://dx.doi.org/10.1007/s10858-013-9757-3.

Barbet-Massin E, Pell AJ, **Retel JS**, Andreas LB, Jaudzems K, Franks WT, Nieuwkoop AJ, Hiller M, Higman V, Guerry P, Bertarello A, Knight MJ, Felletti M, Le Marchand T, Kotelovica S, Akopjana I, Tars K,Stoppini M, Bellotti V, Bolognesi M, Ricagno S, Chou JJ, Griffin RG, Oschkinat H, Lesage A, Emsley L, Herrmann T, Pintacuda G (2014) Rapid Proton-Detected NMR Assignment for Proteins with Fast Magic Angle Spinning. Journal of the American Chemical Society 136: 12489–12497. doi: http://dx.doi.org/10.1021/ja507382j.

Muench F, **Retel J**, Jeuthe S, O h-Ici D, van Rossum B, Wassilew K, Schmerler P,Kuehne T, Berger F, Oschkinat H, Messroghli DR (2015) Alterations in creatine metabolism observed in experimental autoimmune myocarditis using ex vivo proton magic angle spinning MRS. NMR in Biomedicine 28: 1625–1633. doi: http://dx.doi.org/10.1002/nbm.3415.

Summary

The topic of this thesis is the determination of a three-dimensional structure of outer membrane protein G (OmpG) from *E. coli* in its native lipid environment by solid state magic-angle-spinning (MAS) NMR. For this purpose, it was necessary to develop and test new methods for enabling resonance assignments, the collection of distance restraints as well as methods for structure calculation. The lipid bilayer is thought to influence structure and function of membrane proteins, and one of the advantages of solid-state NMR over other methods in structural biology is that this technique allows membrane proteins to be studied in their native environment.

OmpG is a porin in the outer membrane of *E. coli* and has a molecular weight of 34 kDa, with 281 amino acid residues forming a β-barrel composed of 14 strands. The sequence of OmpG is longer than that of most other proteins of which a structure was solved so far by solid state MAS NMR. Since the complexity of NMR spectra increases with the number of amino acids in a protein, OmpG is a challenging system, requiring the development of new assignment strategies. Furthermore, it is an appropriate test system to benchmark the effectiveness of new NMR experiments.

First attempts to assign OmpG were made using ^{13}C-detected spectra, as this was the most common detection method in solid-state NMR at the time this project was started. To reduce spectral overlap and the ambiguity of cross-peak assignment, a set of amino acid-specifically ^{13}C-labeled OmpG samples was produced. This set included samples where only a few amino acids at a time were labeled, such as GAVLS, GAFY and RIGA. The spectra recorded of these samples offered a starting point for the sequential assignment. Other labeling schemes tested were based on the specific carbon labeling pattern obtained when glycerol labeled either on the first and third carbon positions (1,3-glycerol) or the second carbon position (2-glycerol) is used as the sole carbon source during protein expression. Combining the glycerol labeling method with a reverse labeling strategy, two more schemes were produced, in which a specific set of amino acids was labeled following the 1,3- or 2-glycerol labeling pattern: 1,3- and 2-TEMPQANDSG and 1,3- and 2-SHLYGWAFV. These samples allowed the initial assignment to be extended. However, it was not possible to arrive at an assignment to such an extend that structure determination became possible, using ^{13}C-detected experiments alone.

In recent years, solid-state NMR has seen enormous progress. The availability of perdeuterated and partially back-exchanged protein samples and fast spinning probes has opened up the possibility

to detect protons. Using this strategy, a set of spectra was recorded on perdeuterated and partially back-exchanged samples of OmpG, where all amino acids were as well ^{15}N/^{13}C-labeled. The set of ^{1}H-detected experiments allows the assignment of the amide ^{1}H and ^{15}N, Cα and Cβ resonances and greatly simplifies the assignment strategy.

It is a characteristic of the methodology developed in this thesis that data from the deuterated samples were evaluated together with data obtained on amino acid-selectively labeled samples in order to arrive at a trustful assignment. The Cα and Cβ chemical shifts obtained from ^{1}H-detected spectra were used to find corresponding peak patterns in the ^{13}C-detected spectra, giving access to side-chain ^{13}C chemical shifts. The knowledge of side-chain ^{13}C chemical shifts results in more specific amino acid typing, which further facilitated the sequential assignment. Furthermore, cross-peaks in the ^{13}C-detected spectra, provide extra evidence that a sequential assignment is correct. With this combined approach, using information from ^{13}C- and ^{1}H-detected spectra, residues in the regions of the sequence that correspond to the membrane-embedded β-barrel could be assigned.

Using these assignments and distance restraints from a set of through-space correlation experiments between carbons and protons, the structure of OmpG in its native lipid environment could be calculated. Ambiguous distance restraints were disambiguated using the ambiguous restrain iterative assignment (ARIA) protocol. The final structure has a backbone rmsd of ~1.6 Å for the residues in the β-sheets and closely resembles the structure calculated before using solution NMR. The structure deviates from available crystal structures in the extra-cellular part of the protein. In this part of the protein, the structure calculated here shows large flexible loops, like in the solution NMR structure. In contrast, in the crystal structures the β-sheets are extended and on most strands, only small turns, instead of loops, remain. This can be explained by crystal contacts stabilizing the extended structure.

The progress made in our OmpG project allows to design further experiments that yields more insight into the pH-dependent opening and closing mechanism of the porin. Furthermore, the new methodology presented here and verified by the successful structure determination of a membrane protein in its native lipid environment opens up the way to structural investigations of proteins under similar sample conditions.

Zusammenfassung

Ziel der vorliegenden Arbeit ist die Bestimmung der drei-dimensionalen Struktur des aus *E. coli* stammenden Proteins *Outer Membrane Protein G* (OmpG) in seiner nativen Lipidumgebung mittels Festkörper - NMR Spektroskopie. Dazu war es notwendig neue Methoden für die Zuordnung der Resonanzsignale sowie für die Bestimmung von Strukturparametern und Berechnung der Struktur zu entwickeln und zu testen. Da die Lipiddoppelschicht die Struktur und Funktion von Membranproteinen beeinflusst, ist die Festkörper-NMR besonders für Studien dieser Proteine geeignet. Sie ermöglicht im Gegensatz zu anderen Methoden der Strukturbiologie Untersuchungen in der natürlichen Umgebung von Proteinen.

OmpG ist ein Porin in der äußeren Membran von *E. coli*, es besteht aus 281 Aminosäuren die ein aus 14 Strängen aufgebautes β-*barrel* bilden und hat ein Molekulargewicht von 34 kDa. Die Sequenz von OmpG ist länger als die der meisten Proteine deren Struktur bisher mittels Festkörper-NMR gelöst wurde. Die Anwendung von NMR an diesem herausfordernden System erfordert die Entwicklung neuer Zuordnungsstrategien, da die Komplexität von NMR Spektren mit der Anzahl an Aminosäuren eines Proteins zunimmt. Darüber hinaus ist es ein geeignetes Testsystem um die Effektivität neuer NMR Experimente zu validieren.

Erste Versuche die Signale von OmpG zuzuordnen wurden mit ^{13}C-detektierten Spektren unternommen, da diese Detektionsmethode zu Projektbeginn die meistverbreitete in der Festkörper-NMR war. Um spektrale Überlagerungen und die Uneindeutigkeit der Zuordnung von Kreuzsignalen zu reduzieren, wurden verschiedene Aminosäuren-spezifischen ^{13}C-markierte Proben von OmpG hergestellt. Diese umfassen auch solche, in denen nur einige wenige Aminosäuren markiert wurden, beispielsweise GAVLY, GAFY und RIGA. Die Spektren dieser Proben ermöglichten es mit der sequentiellen Zuordnung zu beginnen. Andere verwendete Markierungsprotokolle basierten auf dem resultierenden spezifischen Kohlenstoff-Markierungsmuster in welchen Glycerol, entweder in der ersten und dritten (1,3-glycerol) oder in der zweiten (2-glycerol) Kohlenstoffposition ^{13}C-markiert, als einzige Kohlenstoffquelle während der Proteinexpression eingesetzt wird. Durch Kombination dieser Glycerol-Markierungsmethode und einer umgekehrten Markierungsstrategie konnten zwei weitere Schemata erzeugt werden, in welchen Aminosäuren gemäß den Regeln der 1,3- und 2-glycerol Markierung spezifisch markiert werden konnten: 1,3- und 2-TEMPQANDSG und 1,3-

und 2-SHLYGWAFV. Diese Proben ermöglichten weitere Zuordnungen von Resonanzen. Die Verwendung von ausschließlich ^{13}C-detektierten Experimenten war allerdings nicht ausreichend eine für die Strukturbestimmung hinreichende Anzahl an Signalen zuzuordnen.

In den letzten Jahren wurden enorme Fortschritte in der Weiterentwicklung der Festkörper-NMR erzielt. Die Verfügbarkeit von deuterierten und teilweise reprotonierten Proteinproben und schnell drehenden Probenköpfe hat die Möglichkeit Protonen zu detektieren eröffnet. Dieser Strategie folgend, wurden Spektren von deuterierten und teilweise rück-getauschten OmpG-Proben, in denen alle Aminosäuren zudem ^{15}N/^{13}C markiert wurden, aufgenommen. Der Satz an ^{1}H-detekierten Experimenten ermöglicht die Zuordnung der Amidprotonen und der ^{15}N, Cα und Cβ Resonanzen und vereinfacht die Zuordnungsstrategie deutlich.

Es ist für die in dieser Arbeit entwickelten Methode charakteristisch, dass Daten von deuterierten Proben gemeinsam mit Daten von Aminosäure-spezifisch markierten Proben evaluiert wurden um die zuverlässige Zuordnung zu gewährleisten. Die chemischen Verschiebungen von Cα und Cβ in den ^{1}H-detektierten Spektren konnten dazu verwendet werden korrespondierende Signalmuster in den ^{13}C-detektierten Experimenten zu analysieren, was Zugang zu den chemischen Verschiebungen der Kohlenstoffe der Aminosäure-Seitenketten ermöglichte. Die Kenntnis über diese chemischen Verschiebungen führt zu einer exakteren Typisierung der Aminosäuren, was im Weiteren die sequentielle Zuordnung der Signale erlaubt. Zudem konnten mit Hilfe der Kreuzsignale in den ^{13}C-detektierten Spektren die Korrektheit sequentieller Zuordnungen bestätigt werden. Mit der Kombination aus ^{1}H- und ^{13}C-detektierten Spektren konnten Aminosäure-Reste der Sequenzregion die zu dem in der Membran eingebetteten β-barrel korrespondiert, zugeordnet werden.

Mit Hilfe dieser Zuordnungen und Abstandsmessungen von Korrelationsexperimenten, in denen Magnetisierung durch den Raum zwischen Kohlenstoffen und Protonen ausgetauscht wird, konnte die Struktur von OmpG in seiner nativen Lipidumgebung berechnet werden. Durch die Verwendung des *Ambiguous Restrain Iterative Assignment* (ARIA) Protokolls konnten Abstandsparameter eindeutig zugeordnet werden. Die endgültige Struktur hat ein RMSD des Proteinrückgrates von ~1.6 Å für Aminosäuren in den β-Strängen und ähnelt der Struktur, die bereits mittels Lösungs-NMR gelöst wurde. Die Struktur weicht in den extra-zellulären Bereichen des Proteins von verfügbaren Kristallstrukturen ab. In diesen Bereichen zeigt die hier ermittelte Struktur große flexible loops, ähnlich wie in der mittels Lösungs-NMR bestimmten Struktur. Im Gegensatz dazu sind in den Kristallstrukturen die β-Stränge verlängert und anstatt loops sind nur kurze strukturelle Wendungen festzustellen. Dies kann durch Kristallkontakte, die eine verlängerte Struktur stabilisieren erklärt werden.

Der Fortschritt in diesem OmpG Projekt ermöglicht die Konzeption neuer Experimente, die neue Einblicke in den pH-abhängigen Öffnungs- und Schließmechanismus des Porins bieten können. Darüber hinaus ermöglicht die hier vorgestellte neue Methode, die durch die erfolgreiche Bestimmung der Struktur eines Membranproteins in seiner nativen Lipidumgebung verifiziert wurde, neue strukturelle Untersuchungen von Proteinen unter ähnlichen Probenbedingungen.

Curriculum Vitae

For reasons of data protection, the Curriculum vitae is not published in the online version.