1 Introduction

1.1 Structural Genomics

Structural genomics approaches aim at identifying the three-dimensional structure of all protein folds encoded in the genome of an organism. A prerequisite to achieving such an ambitious goal was the decrease of time required to determine the structure of each single protein. Many steps in the course of protein structure determination such as cloning, protein expression and protein purification could be automated adopting technology that was invented for the genome sequencing projects. At the same time new methods were developed to meet the needs of a high-throughput project. For both, X-ray crystallography and protein nuclear magnetic resonance (NMR) spectroscopy, bottlenecks in the structure determination process had to be overcome. Data acquisition and evaluation had to be standardised and streamlined. In addition new ways of recording and automatically analysing the structural data were invented to further enhance the speed of the structure determination process.

Structure determination by NMR represents a particular challenge for automation. Nevertheless, it is used by several structural genomics consortia. Much work has been devoted to the determination of structures using NMR and the development of new methods to automate and streamline the structure determination process. Many new ways to analyze NMR-data have been proposed in the last years, some of which will be discussed in the chapters 1.3 and 1.4.

In the following section of the introduction, the aims of structural genomics and the strategies for target selection will be described. The possible role of NMR in this context, its particular strengths and weaknesses will be presented to provide the context of the structure determination projects discussed in chapter 3 of this thesis.

1.1.1 Aims of Structural Genomics

With the completion of the big genome sequencing projects a vast quantity of sequence data became available in the late nineties. Unfortunately, the coverage of the sequence space by protein structures has increased much more slowly. Although the number of structures deposited in the protein data bank (PDB) has substantially grown in recent years, it is far from covering the proteome of even the simplest organism. At the same time, the current knowledge of protein folds is far from complete. Many new structures that are added to the PDB have a high similarity to already existing entries, even more have the same basic fold as other already described proteins.

Thus the idea was born to systematically explore the structures of the proteome of one or more organisms with the final goal of a complete coverage. The structural genomics efforts aim at providing a comprehensive data basis for all structure driven research in biology and medicine. It may help to understand the biology of pathogenic organisms and provide the information required for structure based drug design (Hol, 2000). But not only the protein structures themselves are to be expected to be of great use to the scientific community. Ideally, the initiatives would promote the implementation of automation in protein structure determination and optimization to the same level as the genomics initiatives have done for the sequencing of genomes. As a prerequisite for all structure determination efforts large numbers of expression clones would be generated that could aid all biophysical or biological research even if the particular construct failed to produce a structure (Hol, 2000). Finally, apart from the practical benefits, the data obtained by structural genomics might provide the basis to answer one of the most fundamental unresolved questions in protein science (Burley et al., 1999): to link the amino acid sequence of a protein to its three dimensional structure. Ever since the discovery that all information for protein folding is contained in the primary structure (Anfinsen et al., 1955), the nature of this ‘folding code’ has puzzled the scientific community. The availability of a structure to almost all known sequences might help to recognize and understand this code.

The determination of all protein structures for every protein from every organism of interest is a daunting task. It would also lead to a high degree of redundancy, since many proteins will be similar in sequence within the same or between different organisms. Hence, the aim of the structural genomics projects is not to determine all structures from all proteins but to determine a sufficient number of structures to be able to deduce the structures of all other proteins from these structures with a given accuracy.

Basically two different factors have to be considered when estimating the number of proteins to be determined (Vitkup et al., 2001). One is the degree of sequence similarity between the sequence in question and the one with a known structure needed to build a sufficiently accurate structural model. This also requires a definition of the term "sufficiently accurate". The other is the level of family coverage that is to be achieved. Since there are many outliers in the protein sequence space that have no apparent relation to the sequence of other proteins, much effort would have to be directed into the determination of structures, from which no other protein could be modelled. If 10 % of the most divergent sequences are disregarded, the number of structures to be determined can be reduced by a factor of two to four, depending on the level of similarity assumed to be necessary for successful modelling. Vitkup et al. extrapolate from experience obtained in the CASP-modelling competitions (Moult et al., 2005) that this level of sequence identity is about 30-35 %. In the CASP trials this is the minimum level of sequence identity from which the sequence alignment tends to be correct. Below this level of identity, the alignments often have significant errors leading to large errors in the final predicted model. Vitkup et al. predict that with a modelling threshold of 30 % about 8000 to 10.000 individual structures would have to be determined to cover the sequence space of all non-membrane proteins in the Pfam database.

This prediction shows that the choice of target structures is extremely important. In an ideal situation, all structural genomics initiatives would coordinate their choice of

targets to ensure that those structures, from which most other structures can be modelled, would be determined first. In this approach representative members of large protein families would be prioritised. Such a "greedy" approach would lead to the fastest progress towards the "completion" of the structural genomics aims. In reality the coordination between the different structural genomics initiatives is less optimal. The "greedy approach" is only one way to prioritize the targets for structure determination. Other strategies could aim for "new folds" to maximize the knowledge of protein shapes or the determination of the structure for sequences that lack similarity to others, which would maximize the knowledge of protein structures in general (Brenner, 2000). Brenner predicts that once the major goals of structural genomics have been achieved, the same highly automated and broad approach could be used to study the structures of all proteins related to a specific biological function or context.

1.1.2 NMR in Structural Genomics

Since NMR spectroscopy has been established as a method for protein structure determination in the late 80's, it has contributed about 13 % of the structures deposited in the PDB by September 14th 2005 (Berman et al., 2000). Thus it is reasonable to assume that NMR will also contribute its share of structures to the structural genomics projects. Compared to X-ray crystallography, NMR has a number of advantages and disadvantages in this context. One of the most prominent advantages of NMR is that it does not require the protein under investigation to form crystals. Thus, the investigation can be started as soon as soluble and folded protein has been obtained. If the protein is folded and sufficient structural data can be obtained, it is possible in almost every case to solve the structure. On the other hand, the time for data acquisition and evaluation can be very long compared to crystallography. Data acquisition takes traditionally about three to four weeks of spectrometer time. Recent methods and equipment, which will be discussed in chapter 1.3.2, make it possible to reduce this time substantially. The evaluation of the data, including resonance assignment and the structure calculation process, may take more than two months,

depending on the size and fold of the protein and the quality of the spectra. Recently, improvements have been achieved in this area, which will be discussed in chapter 1.3.3. Perhaps the biggest disadvantage of NMR is the limit in protein size. Although systems with a molecular weight of ~100 kD can be studied (Riek et al., 1999), standard NMR structure determination is still limited to proteins with a molecular weight not larger than 25-30 kD. The long data analysis time seems to be the major disadvantage when compared to crystallography. Nevertheless, if also the time that might be needed for crystallization trials is considered, NMR might even be faster. Thus, structural genomics initiatives with a considerable NMR component such as the RIKEN consortium tend to solve structures of small, well folded non-aggregating proteins by NMR spectroscopy.

Besides the use for structure determination, NMR may have other potential roles in structural genomics projects. Perhaps the most obvious is the use as an analytical tool to determine the characteristics and quality of protein preparation (Page et al., 2005). Since the distribution of chemical shifts is very indicative of the "foldedness" of a protein, NMR can yield information about the state of the protein without any assignments. Page et al. use NMR to group purified NMR samples into four groups that range from completely structured monomeric over possibly aggregated to mostly unfolded preparations.

This analytical power stems from the fact that the ^{1}H chemical shifts of methyl and amide groups of proteins are very sensitive to the protein fold. Methyl-groups in an aqueous environment show chemical shifts of ~1 ppm. In the hydrophobic core of proteins methyl-groups are often closely packed next to aromatic rings hence the methyl resonances may occur up to –2 ppm due to ring current shifts. The ^{1}H chemical shift of amide groups of unfolded peptides is between 8 and 8.5 ppm. In both regions of the spectrum only very few other resonances may overlap with the resonances of interest. When the protein is folded, both hydrogen-bonds and the packing in the hydrophobic core of the protein alter the amide chemical shifts. These are then spread

out between 9.5 and 6 ppm. Thus the distribution of peaks in this area can report the "foldedness" of the protein (Figure 1.1). When carefully analysed, the line width of the signals can also indicate the quality of the protein preparation. Since the line width is dependent on the rotational motion of the protein in solution, aggregates show broader lines than a monomeric protein of given size. However, other parameters also influence the line width of NMR spectra, making this kind of analysis very difficult.

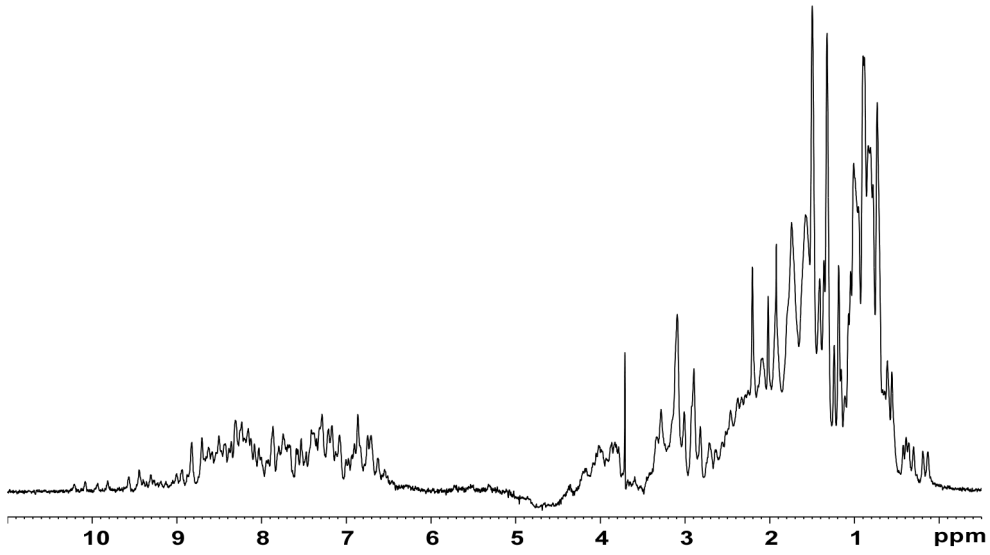

Figure 1.1 ^{1}H-NMR spectrum of CI-B8
The Spectrum shows the characteristics of a folded protein. Characteristic are the peaks below 0.5 ppm and beyond 9 ppm that only occur in folded proteins through the packing of methyl-groups against aromatic sidechains or the involvement of NH-protons in hydrogen-bonds, respectively.

While the use of ^{1}H-NMR for determining the quality of a protein preparation is straightforward, it could also be used in much earlier steps of the purification process. The high abundance of heterologously expressed proteins in *E.coli* cell lysates, in which these proteins can represent large fractions of the dry mass, permits the detection of a single protein in the complex mixture of the lysates soluble fraction (Almeida et al., 2001). With this method, some characteristic features of the protein can be detected even before affinity-purification.

In addition to the actual structure determination, NMR can also provide other structural parameters difficult to obtain by other methods. Chemical shift perturbations allow to identify the interaction sites of proteins with each other or with smaller molecules (Shuker et al., 1996). Using residual dipolar coupling (RDC) measurements, also the structure of complexes, for which the structures of the individual subunits are available, can be solved with amazing precision (Clore et al., 2003). These methods show promising applications for NMR in the functional characterization of proteins.

Furthermore, NMR can be a very powerful tool for fast fold identification (Prestegard et al., 2001). Based only on a backbone assignment, which can be obtained from relatively few and sensitive experiments and a number of RDC-measurements, NMR can provide fairly accurate structures that report the fold of regular backbone structures and their topology. If the aim of a structural genomics project is to determine the fold of a very large number of proteins it may not be necessary to determine these folds at high resolution. If a high resolution structure is not the primary goal, the traditional NOE-based NMR approach that delivers such structures can be bypassed and faster methods that provide only data with low structural resolution such as RDCs or chemical shifts can be used.

1.2 Protein NMR

1.2.1 Introduction to Protein Solution NMR Methods

A thorough introduction to state of the art protein NMR methods is beyond the scope of this work and the field has developed into a complexity that could barely be covered in a single dedicated textbook. However, a short introduction to some key principles of NMR and biological solution NMR will be given in this chapter to provide the basis for the work presented here. These basic principles will also include the scheme which is used to identify protein NMR methods and the strategies that lead to resonance and NOE assignments of protein spectra.

All nuclei with a Spin I ≠ 0 show an alignment of their nuclear spins in an external magnetic field B_0. For I = ½ there are two alignment possibilities: one parallel and the other anti-parallel to the magnetic field. The energy difference ΔE between these alignments is given in equation 1.1 and depends on the strength of the external field B_0 and the type of the nucleus, in particular its gyromagnetic ratio γ_N.

$$\Delta E = \gamma_N \hbar B_0 \tag{1.1}$$

This energy difference can be exploited for spectroscopy using electromagnetic fields in the MHz range (eq. 1.2).

$$\Delta E = h\nu \tag{1.2}$$

Each spin has an angle to the magnetic field vector and precesses around this axis with a characteristic frequency. The dependency of this Lamor-frequency ω_0 on the magnetic field is given by equation 1.3.

$$\omega_0 = \gamma B_0 \tag{1.3}$$

The fact that the resonance frequency is dependent on the magnetic field has led to the convention of referencing all frequencies against an internal standard. This relative frequency is called the chemical shift given in parts per million (ppm). The sensitivity of the chemical shift towards the strength of the magnetic field makes it a very interesting tool to analyse biological macromolecules. The magnetic field experienced by a particular nucleus in a compound depends on both its chemical bonds (its direct electron environment) and the distribution of electrons in its vicinity. Therefore, chemical shifts of protons in biological macromolecules are differentiated to a great extent. Not only do different types of protons in different amino acids exhibit characteristic chemical shifts but also the unique environment in the structure of the

macromolecule influences these shifts. This makes it possible to distinguish most of the resonances that occur e.g. in a protein and to exploit the shifts for structure determination purposes (Neal et al., 2003).

Modern NMR spectroscopy is implemented as Fourier-transformation spectroscopy. Rather than the absorption of irradiation energy, the Lamor-frequency of the nuclei is measured directly. This is mainly due to the fact that the energy difference between the two states for spin ½ nuclei is very small. This results in a small population difference, which makes a direct measurement of the absorbed energy difficult. Fourier-transform NMR spectroscopy exploits the Lamor-precession of the nuclei and the most basic experiment can be described as follows: In equilibrium with an external magnetic field all spins precess around an axis parallel to B_0 (which by convention is the z-axis of the coordinating system used to describe NMR-experiments), resulting in a net magnetisation. A short radio-frequency pulse can turn the net-magnetisation along z into the xy-plane, where the spins still precess with their Lamor-frequencies around z. With a stationary coil one now can measure a current resulting from this oscillating net magnetization. Since the spins slowly return to equilibrium, the transverse magnetization is lost over time. The resulting signal is a dampened sinus, the free induction decay or FID, from which the contained frequencies can be extracted by Fourier-transformation. Usually the 'excitation' pulse is a μs-short rectangular pulse that excites frequencies over a range of 10 kHz. Consequently, a simple spectrum can be recorded using a pulse sequence as depicted in Figure 1.2 A. Two relaxation constants describe the loss of the xy-magnetisation. The spin-lattice relaxation T_1 is due to energy transmission to the environment and describes the return of the z-magnetisation. The second relaxation time constant T_2 results from exchange spins within the xy-plane that leads to a loss of the coherence. This relaxation time constant also depends on the rotational correlation time of the molecule τ_C as approximately expressed in equation 1.4.

$$5 \cdot 10^{-9} \cdot T_2 \approx \frac{1}{\tau_C} \tag{1.4}$$

The T_2 relaxation time constant is in general smaller than T_1 (eq. 1.5).

$$T_2 \leq T_1 \tag{1.5}$$

The dephasing of the transverse magnetization is also influenced by differences in Lamor-frequency that together with T_2 leads to an observable transverse relaxation time T_2^*. The line-width of the NMR signal depends on this observable rate (eq. 1.6).

$$\Delta \nu_{1/2} = \frac{1}{\pi T_2^*} \tag{1.6}$$

Since τ_C depends on the hydrodynamic radius of the molecule (eq. 1.7) which again is proportional to its size, the size of the molecule becomes limiting in NMR spectroscopy of biological macromolecules.

$$\tau_c = \frac{\eta V_{hyd}}{kT} \tag{1.7}$$

The third important effect in NMR spectroscopy is the coupling between spins. The direct interaction of neighbouring spins, mediated by the electron spins of the chemical bonds, is called 'scalar' or J-coupling. It leads to a splitting of lines in the spectra and is independent on B_0. Therefore it is usually given in Hz. It depends on the angle and the number of bonds between the coupling nuclei, since it is mediated by the binding electrons. It therefore contains important structural parameters. The scalar coupling can be exploited for NMR experiments in which it is either measured (Figure 1.2 A, B) or used to transfer magnetisation between nuclei (Figure 1.2 D). Two types of experiments are of major importance for protein NMR. The first is the ^{1}H-^{1}H-COSY

(Figure 1.2 B), in which coupled pairs of protons give rise to crosspeaks. The second is the INEPT (insensitive nucleus enhancement by polarisation transfer) transfer (Figure 1.2 D) in which magnetisation is transferred from protons to other spin ½ nuclei exploiting the J-coupling (Morris et al., 1979).

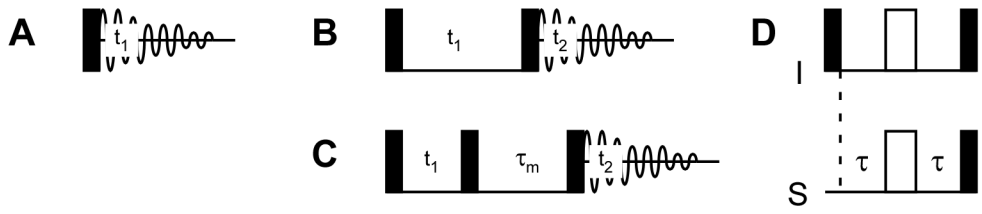

Figure 1.2 Basic 1D and 2D ^{1}H-experiments and the INEPT building block
Three simple ^{1}H-NMR experiments are shown in A through C. **(A)** the simplest Fourier transform NMR experiment consists of a single 90° pulse followed by a detection time. **(B)** ^{1}H-^{1}H-COSY. **(C)** ^{1}H-^{1}H-NOESY. **(D)** The INEPT building block is the basis of heteronuclear experiments. After initial excitation the 180° pulses followed by the 90° pulses transfer I_x into S_y magnetisation, if τ is set to 0.25 J_{IS}. Black rectangles represent 90° hard pulses, open rectangles 180° pulses.

The interaction through space is called the 'nuclear Overhauser effect' or NOE. It is a consequence of dipole-dipole interactions of two spins close in space. The NOE of a spin I, η_I is given as the change in signal intensity when a nearby transition equilibrium of a spin is disturbed (eq. 1.8).

$$\eta_I = \frac{I - I_{eq.}}{I_{eq.}} \tag{1.8}$$

This is perhaps the most important effect in biological NMR since under 'extreme narrowing conditions' ($1/\tau_C >> \omega_0$) the cross-relaxation-rate R_{ij} between two spins is, besides being dependent on other factors, proportional to the distance between the nuclei (eq. 1.10). This allows to determine distances between protons that are close in space and to calculate the structure of a molecule from a large number of such distances. Although other NMR parameters can be used in structure calculations, the

NOE is the most important one. The simplest ^{1}H-^{1}H-experiment measuring such cross-relaxation rates is given in Figure 1.2 C.

$$R_{ij} = \frac{1}{10}\gamma^4\hbar^2\frac{1}{r_{ij}^{-6}}\left(-\tau_c + \frac{6\tau_c}{1+(2\omega_0\tau_c)^2}\right) \tag{1.10}$$

Although most of the main elements that constitute biological macromolecules have spin ½ isotopes, only the natural abundance of ^{1}H is high enough to be effectively used for NMR spectroscopy without labelling the molecule of interest. The isotopes ^{13}C and ^{15}N occur at only 1 % and 0.2 % natural abundance, respectively. To be able to profit from the carbon and nitrogen chemical shifts to assign and to separate the hundreds of ^{1}H-resonances in a biological macromolecule the isotopes ^{13}C and ^{15}N are enriched to almost 100 % in the molecules. This is achieved by growing the bacteria expressing the molecules on media containing ^{13}C-glucose and $^{15}NH_4Cl$ as the sole carbon and nitrogen sources. In such a 'fully labelled' molecule also the J-coupling between protons and nitrogen or carbon and between nitrogen and carbon can be exploited for NMR spectroscopy.

Since the gyromagnetic ratios of ^{13}C and ^{15}N are only 10 % and 25 % of that of ^{1}H, heteronuclear NMR experiments are usually started with ^{1}H excitation and the magnetisation is then transferred to other nuclei in subsequent steps. One such transfer element is the INEPT element (Figure 1.2 D). It exploits the larger polarisation of a sensitive (high γ) spin to excite a directly bound neighbour. This is used in the hetero-nuclear single quantum coherence (HSQC) experiment which allows correlating each proton to either its directly bound carbon or nitrogen atom (Figure 1.3).

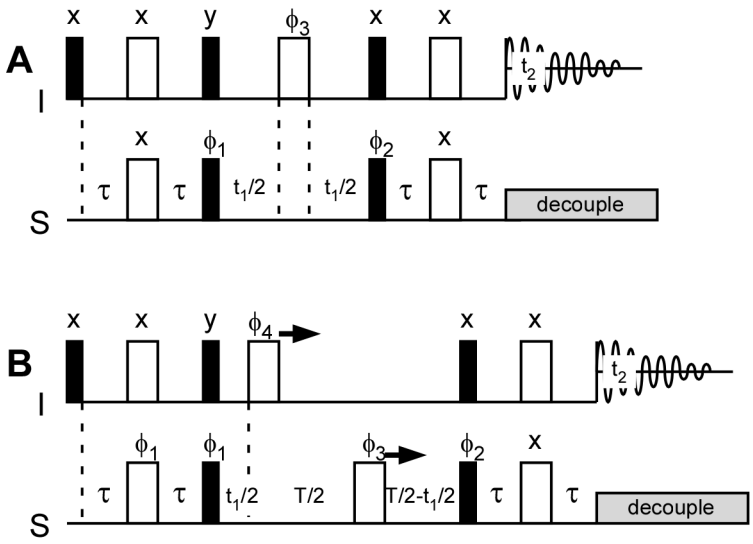

Figure 1.3 Pulse sequence for HSQC and CT-HSQC

(A): Pulse sequence for the HSQC experiment. After the initial excitation the magnetisation is transferred using a INEPT building block and the chemical shift of the heteronucleus is recorded. Then another INEPT block transfers the magnetisation back to the protons, where it is detected. **(B)**: CT-HSQC. This pulse-sequence basically results in the same spectrum but removes homonuclear coupling during t_1. Black rectangles represent 90° hard pulses, open rectangles 180° pulses. Phase-cycles are (A): Φ1: x, -x; Φ2: 2(x), 2(-x); Φ3: 4(y), 4(-y); Receiver: 2(x, -x, -x, x). (B): Φ1: x, -x; Φ2: 8(x), 8(-y); Φ3: 2(x), 2(y), 2(-x), 2(-y); Φ4: 16(y) 16(-y); Receiver: 2(x, -x, -x, x), 2(-x, x, x, -x)

The resolution in the indirect dimension of these experiments is limited by the number of time increments that are recorded. An additional factor that strongly contributes to the apparent line-widths in ^{13}C-HSQC experiments is the homonuclear ^{13}C-^{13}C ^{1}J-coupling. While heteronuclear couplings can be removed with 180° pulses in the middle of the evolution period, it is difficult to suppress homonuclear couplings in an traditional HSQC experiment. One way to remove homonuclear couplings is to introduce a constant time evolution period (Figure 1.3 B). Here the overall evolution time is constant and only two 180° pulses are moved in t_1 increments. This results in the same spectrum as the HSQC sequence (Figure 1.3 A) but removes the homonuclear couplings. Apart from echo-effects the signal intensity is constant, allowing extensive linear prediction of further points (Led et al., 1991). The major disadvantage is the limited number of possible t_1-increments that can be recorded within the length of the overall time-period.

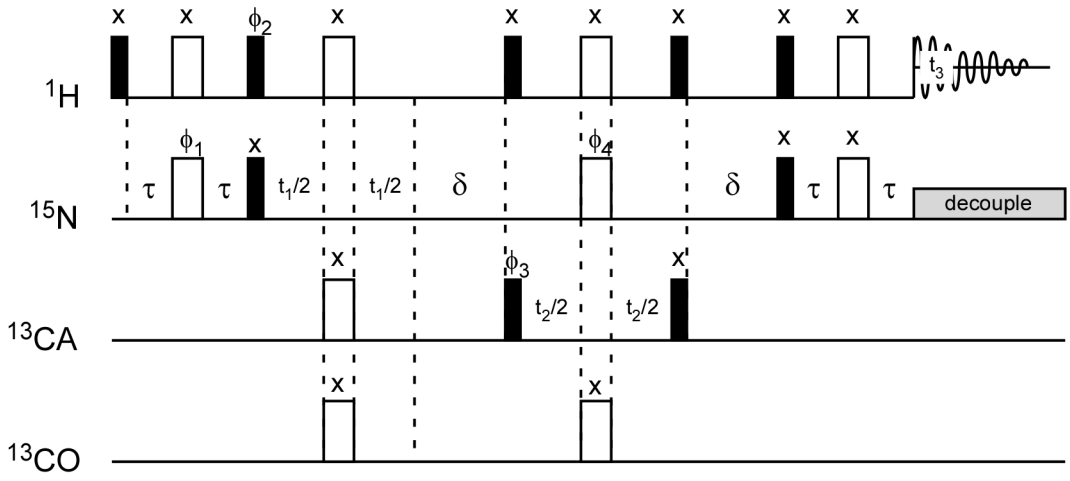

Figure 1.4 Simple implementation of an HNCA

Simple implementation of an HNCA-pulse sequence. This implementation is the simple combination of a ^{1}H-^{15}N-HSQC and a ^{15}N-^{13}C-HMQC, showing paradigmatically the 'building block' approach in multidimensional heteronuclear NMR. Today, more sophisticated implementations of the HNCA have been published (Grzesiek et al., 1992a) that circumvent some of the problems of this example. The black rectangles represent 90° hard pulses, open rectangles 180° pulses. Phase-cycle: Φ1: x, -x; Φ2: y, -y; Φ3: 2(x), 2(-x); Φ4: 4(x), 4(y), 4(-x), 4(-y); Receiver: x, -x, -x, x, -x, x, x, -x

NMR spectroscopy is not limited to two dimensions. In principle experiments of any dimensionality could be recorded simply by introducing additional evolution times into the pulse-sequence. A simple 3D experiment is the HNCA, which may be used to illustrate the approach towards multidimensional NMR. Here a ^{1}H-^{15}N-HSQC is interleaved with a ^{15}N-^{13}C-HMQC yielding the 3D-HNCA (Figure 1.4). In practice, the number of dimensions is limited by two factors. First the overall relaxation rate that limits the maximum length of the pulse-sequence after which it is still possible to record a signal and second by the total measurement time that is needed to sample the increments for each dimension.

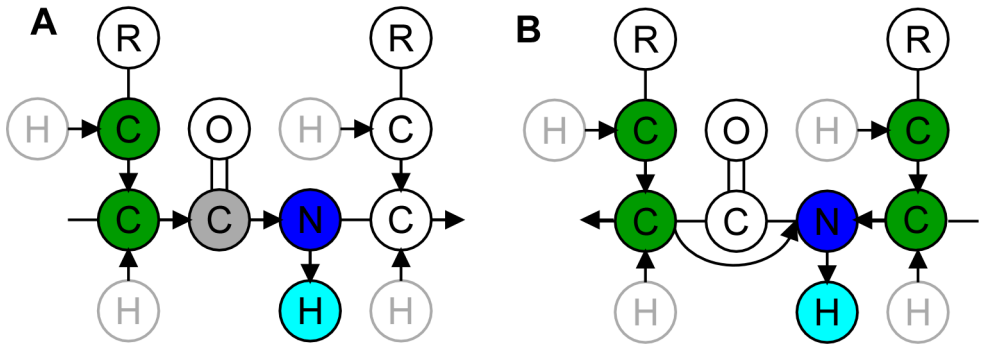

Figure 1.5 Magnetisation flow in the CBCA-pair of 3D experiments
(A) CBCA(CO)NH, **(B)** CBCANH. Detected nuclei are coloured in green, blue and cyan. Protons that are excited with the first pulse are grey. Carbons, explicitly used to transfer magnetisation but not detected are coloured in grey. Arrows give the direction of the transfers.

Since there are many pulse-sequences and phase cycles that lead to the same spectrum it has become a tradition to name the NMR-experiments used for protein NMR after the nuclei involved (Sattler et al., 1999). The nuclei are quoted in the order in which the magnetisation is transferred. Nuclei of which no chemical shift is recorded but that are used to transfer magnetisation are given in parenthesis. The proton, which is excited at the beginning of the experiment, is usually omitted. As examples the HN(CO)CA and CA(CO)NH will be discussed here. In both experiments, the amide ^{1}H and ^{15}N resonances of one residue are correlated with the C^{α} resonance of the preceding amino acid. In the HN(CO)CA the magnetisation is transferred first from the amide-proton to the nitrogen, then to the carbonyl-carbon and finally to the C^{α}. After the C^{α} chemical shift evolution, the magnetisation is transferred back to the nitrogen for recording its chemical shift and finally back to the amide-proton, where the FID is recorded. In the case of the CA(CO)NH, the experiment is started with excitation of H^{α}-magnetisation, which is transferred to the C^{α}, afterwards it takes the same way as in the second half of the HN(CO)CA.

The assignment of the protein resonances is usually based on three dimensional heteronuclear methods. A complete assignment is then obtained in two steps. In the first step, the sequential connectivities are established and the protein backbone resonances are assigned. In the second step the resonances of the amino acid sidechains

are correlated to the backbone and the direct connectivities are assigned. This allows the assignment of all sidechain resonances with the exception of the aromatic rings.

For the sequential backbone assignment ^{15}N-HSQC based pairs like the CBCA(CO)NH and the CBCANH (Grzesiek et al., 1992b; Grzesiek et al., 1992a) are used. In the CBCA(CO)NH, the resonances of C^{α} and C^{β} of the preceding or 'i-1' amino acid are correlated to the N and H of the own or 'i' amino acid. In the CBCANH, both the i-1 and the i C^{α} and C^{β} are correlated to the NH_i (Figure 1.5). Thus to every NH_i cross-peak in the ^{15}N-HSQC a spin system consisting of two pairs of carbon resonances are known including information about the origins of the carbon resonances. This allows establishing a sequence of spin systems, since the same combination of C^{α} and C^{β} chemical shifts that is found as 'own' for one amino acid can be found as 'preceding' for another. Characteristic chemical shifts like the C^{β} of alanine, serine and threonine allow to align the ordered spin systems to the protein sequence. Similar pairs can be obtained for the CO- and H^{β}/H^{α}-resonances. These pairs can provide additional information to solve ambiguities in the CBCA-pair.

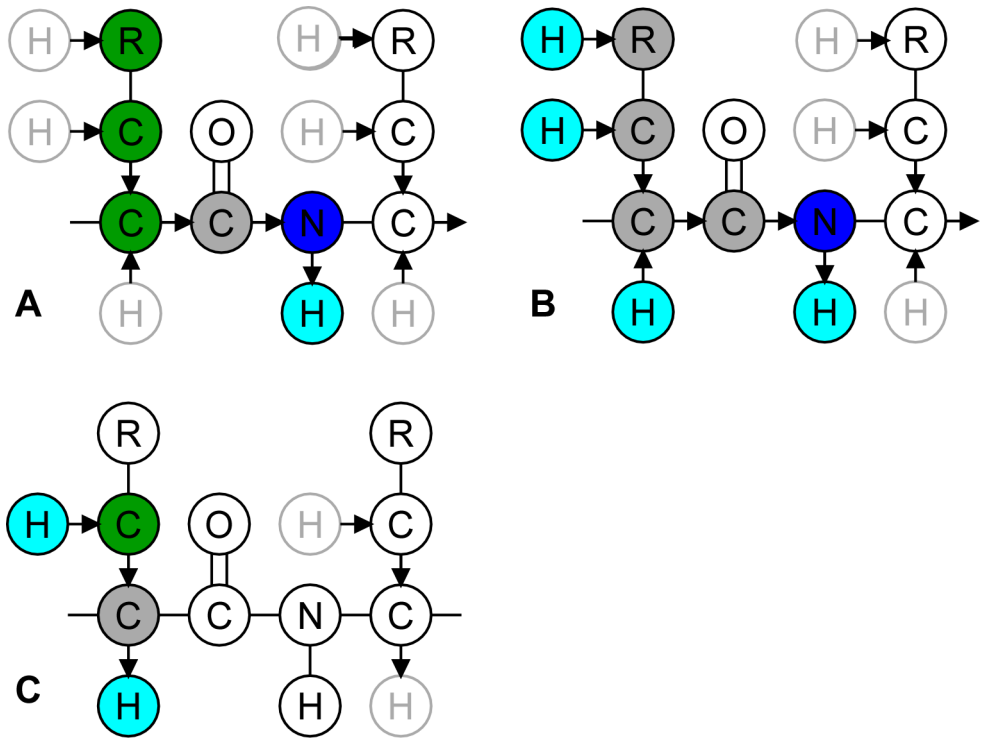

Figure 1.6 Magnetisation flow in the experiments used for sidechain assignment.
(A) CC(CO)NH-TOCSY, **(B)** H(CCCO)NH-TOCSY, **(C)** HC(C)H-COSY. Detected nuclei are coloured in green, blue and cyan. Protons that are excited with the first pulse are grey. Carbons, explicitly used to transfer magnetisation but that are not detected are coloured in grey. Arrows give the direction of the transfers.

The remaining resonances of the protein sidechain can be assigned with a strategy based on TOCSY-type 3D experiments (Figure 1.6 A, B). In the H(CCCO)NH-TOCSY and the CC(CO)NH-TOCSY (Montelione et al., 1992) all proton and carbon resonances from the sidechain are correlated to the NH of the following amino acid. An HC(C)H-COSY (Figure 1.6 E) then allows to establish three-bond connectivities between the protons and thus the order of the resonances in the sidechain. This approach fails for the sidechains that contain ring systems. The resonances of these amino acids can be assigned using the 3D-^{13}C-HMQC-NOESY in which the protons of the same sidechain should show strong cross-peaks due to the small distances between them.

Once the assignment of all protein resonances is established, structural restraints can be collected from NOE-based spectra. Although a 2D-^{1}H-^{1}H-NOESY contains in principle all information about proton-proton contacts present in the protein, its

evaluation is often hampered by severe overlap of the proton resonances. This overlap can be partially resolved by the use of 3D-^{15}N-HSQC- and 3D-^{13}C-HMQC-NOESY experiments that provide all NOE cross-peaks of a proton identified by its ^{1}H and ^{15}N or ^{13}C chemical shift. 4D-NOESY spectra like the 4D-^{15}N^{13}C-HSQC-NOESY-HMQC would in principle provide an even better resolution of the overlap of the proton resonances. In practise, this advantage is negligible since these spectra can be recorded only at a very low experimental resolution in the indirect dimensions due to their enormous requirements on experiment time.

Another source of structural information are residual dipolar couplings (RDCs) obtained from samples in which the protein is partially aligned to the external magnetic field B_0. They can be measured as altered splitting of the lines due to J-coupling and provide the orientation of the connection vectors of the coupling atoms relative to the alignment vector of the molecule. The alignment of the protein in solution to B_0 can be achieved by adding lipid bicelles or phages. Stretched gels that partially limit the diffusion of the molecule are an alternative method of alignment. The splittings due to the coupling constants like the $^{1}J_{NH}$ are relatively easy to measure since the coupling information has to be actively suppressed in many of the NMR methods. In some cases it is advisable to use pulse sequences designed to extract J-coupling data like IPAP or ECOSY based techniques (Ding et al., 2002). It has been shown that RDC data can greatly improve the quality of an NMR structure, if used as additional constraints in the structure calculation (Cornilescu et al., 2000). In some cases it is even possible to determine the fold of a protein from RDC data alone (Valafar et al., 2004).

1.2.2 Structure Calculation from NMR Data

As opposed to X-ray crystallography, NMR does not provide data that is directly related to the spatial coordinates. However, it provides distance information for isolated pairs of protons or information about dihedral-angles. To solve the structure of the molecule, these constraints are used in a molecular dynamics (MD) simulation

usually following a simulated annealing protocol. In MD simulations , the equations of motion for all atoms (eq. 1.11) are solved using a specified force-field that determines the energy contribution of all parameters (eq. 1.12).

$$\frac{d^2\vec{r}_i}{dt^2} = -\frac{c}{m_i}\frac{\partial}{\vec{r}_i}E_{hybrid} \tag{1.11}$$

$$\begin{aligned} E_{hybrid} &= \sum_l w_l E_l \\ &= w_{bond}E_{bond} + w_{angle}E_{angle} + ... \end{aligned} \tag{1.12}$$

The force-field E_{hybrid} contains potentials representing the restraints derived from the NMR data in addition to other molecular parameters like bond-lengths and angles. The simulated annealing protocol helps to overcome energy-barriers during the search for the global energy minimum by raising the temperature to about 1500 K and slowly reducing to room temperature. Usually hundreds of such runs are calculated from random starting structures and an ensemble of possible solutions is reported as the resulting 'structure'. One implementation to solve NMR structures in this manner is XPLOR-NIH (Schwieters et al., 2003).

A major difficulty in the assignment of ^{1}H-resonances in the NOESY spectrum is the overlap of proton resonances. While the resolution in the direct dimension can be increased by applying 3D-NOESY methods, it remains likely that several assignment options prevail for a single peak in the indirect dimension of the spectrum. To circumvent the problem of assigning such peaks to a single proton pair, ambiguous distance restraints (ADRs) were developed (Linge, 2000) allowing the use of restraints with several assignment options. The ADR combines all possible distances that can be assigned to a single peak into an effective distance D (eq. 1.13).

$$D \equiv \left(\sum_{a=1}^{N} d_a^{-6} \right)^{-1/6} \tag{1.13}$$

If a constraint is generated such that D stays within the distance range derived from the volume of the peak, the information can be used for structure calculation, even if it is not possible to assign the peak unambiguously. This technique allows to calculate reasonable initial structures from highly ambiguous data. Subsequent rounds of NOE assignment then allow to remove the ambiguous assignments on the basis of these initial structure. ADRs also are the basis for many computer programs for automated NOE assignments discussed in chapter 1.3.3.

1.3 Improvement of NMR Structure Determination by Automation and Integration

The success of the genome sequencing projects, through which high-throughput methods were introduced into life-sciences, led to changes in protein structure determination. Automated platforms for cloning and DNA-purification were available, thus large scale protein expression, purification, crystallisation and even structure determination seemed only a matter of integrating the already existing systems. In this chapter, an introduction to the advances that have been made in these fields will be given.

1.3.1 Sample Preparation

Structural genomics, especially of human proteins, often starts with testing many expression constructs. The use of robotics in this field has allowed to conduct expression tests and the initial characterization of the expressed protein at medium to larger scales (Scheich et al., 2003). The application of such systems allows testing many different proteins in a reasonable amount of time or systematically testing different expression constructs for a single protein or domain.

Aditionally, the cost-effectiveness of the labelling of proteins expressed in *E. coli* has been improved considerably. Based on the observation that the exhaustion of nutrients comes with a sharp drop in oxygen consumption (Cai et al., 1998) it was possible to develop new labelling protocols. The cells are first grown to high densities using unlabelled nutrients while the labelled nutrients, such as ^{13}C-glucose or ^{15}N-NH_4Cl are added prior to induction. Compared to traditional protocols in which the growth of the bacteria already takes place in labelled M9 media, the new protocols allow producing more labelled protein per labelled nutrient.

1.3.2 Data Acquisition

During the progress of the structural genomics efforts the NMR methodology has been improved, primarily to reduce measurement time. In principle, two different parameters can determine the time needed to record an NMR experiment. One is the ratio of signal to noise (S/N) determining the number of 1D experiments (scans) that are needed to obtain a sufficiently strong signal. If this limits the time allotted for an experiment it is called “sensitivity limited”. In other cases a far larger number of scans has to be recorded than needed for S/N, e.g. to obtain sufficient resolution in the indirect dimensions. This is called the “sampling limited regime”.

One of the most important improvements that lead to shorter measurement times was the development of cold-probes (Varian) or cryo-probes (Bruker). In these probeheads, the circuitry for the coils is cooled to approximately 30 K using pressurized helium reducing electric noise to an absolute minimum. Using this technology the S/N ratio can be improved up to eightfold. Unfortunately, modern multidimensional heteronuclear experiments are often “sampling limited”. Thus the improvement of S/N by cryo-probe technology can not always be completely translated into shorter measurement times. However, in some cases, the measurement times may be reduced by a factor of two or four.

Independently from these hardware developments, pulse-sequences have been developed trying to circumvent the problems in the "sampling limited" regime. These approaches commonly try to reduce the experimental time by parallel evolution of the chemical shifts of different nuclei. The GFT (G-matrix / Fourier transform) approach (Kim et al., 2003) allows the reduction of the dimensionality of the recorded spectrum. In this way, for example the 3D-HNCO can be recorded as a 2D-spectrum. The information from one or more indirect dimensions is encoded as peak multiplets in lower dimensions and can be translated into normal chemical shifts. This approach allows to either reduce the measurement time of sampling limited common NMR experiments or to design experiments for chemical shift assignment correlating four or five chemical shifts, which would otherwise be impossible to record as 4D or 5D experiments due to the excessive measurement times required. A similar approach is 'High-Resolution Iterative Frequency Identification' or HIFI-NMR (Eghbalnia et al., 2005). Here a 3D spectrum is reconstructed from tilted 2D planes. Since experiments with parallel evolution times are used, it is possible to record planes with a 'tilt-angle' different from 0° and 90° by choosing a ratio for the lengths of the increments. The method is implemented using an algorithm that calculates the tilt angle of the next recorded plane in order to maximize the information added by this experiment. If the additional information drops below a certain threshold, the experiment is stopped. Since the 2D planes are recorded with resolutions far beyond what is feasible in 3D experiments, the time reduction is not as drastic as expected but still in the order of a factor of ten.

A different method to reduce the measurement times in the sampling "limited regime" refines the analysis of the recorded data beyond Fourier transformation. Since many of the properties of the signals in frequency space are principally known, it is possible to achieve resolutions similar to FT from fewer timepoints using FDM analysis (Chen et al., 2004). This approach does not differ in the pulse sequences or in the resulting spectra from traditional FT-NMR spectroscopy. Hence, it can easily be introduced into existing data acquisition strategies.

The power combining these modern methods was recently demonstrated by the North-Eastern Structural Genomics Consortium (Liu et al., 2005). Application of GFT-NMR methods and automated assignment allowed to reduce the time for NMR structure determination including the measurement time to an average of fifteen days.

1.3.3 Assignment / Structure Calculation

The analysis of NMR data has always been the major bottleneck in NMR structure determination. Naturally, great effort has been dedicated in most structural genomics initiatives to the development of software that automates this process. The analysis of NMR spectra is perceived as a logical puzzle that in principle could be solved by appropriate computational approaches. Over the years many efforts have been made to develop such software, but data imperfection has so far hampered these developments. Nevertheless, there are promising approaches to automate resonance assignment or NOE-assignment/structure calculation. This section will not provide a complete overview, but insights into some paradigmatic approaches that might be employed to automate the process in the future.

Perhaps the biggest challenge in automating the assignment process of NMR spectra is the assignment of backbone and sidechain resonances. Earlier attempts to solve this problem tried to reproduce the manual assignment process roughly described in chapter 1.2.1 in a fully automated manner. Programs like Catch23 (Oschkinat et al., 1994) or Autoassign (Zimmerman et al., 1997) try to complete the assignment process as far as possible. This was a problem with these early approaches, since it often introduced errors that had to be corrected manually. An obstacle to the practical use of these programs was the complicated setup and the tedious process of verifying the assignments made by the program. Newer approaches like IBIS (Hyberts et al., 2003) and Smartnotebook (Slupsky et al., 2003) try to avoid these difficulties. They do not aim at a complete assignment but try to solve the straight-forward tasks of assigning the residues easy to assign in a transparent manner. The programs are therefore tightly integrated into graphical the NMR analysis platforms, IBIS into

XEASY (Bartels et al., 1995) and Smartnotebook into NMRview (Johnson, 2004). This allows for integration of the automated procedures into the manual assignment process, since the decisions made by the program can be easily visualized. These approaches included also a probability check based on chemical shift distributions to verify the sequential assignment.

These methods do only offer very basic assignments of the protein sidechains. Therefore manual interaction is always required. Perhaps the most complete sidechain assignment is produced by the software IBIS, since it also accepts (H)CC(CO)NH-TOCSY and (H)CCNH-TOCSY experiments that contain information about all sidechain carbons. Unfortunately the latter experiment is very insensitive.

More success in terms of automation was obtained with methods that automatically assign the crosspeaks of NOESY spectra. Perhaps the most basic implementation is ARIA (Linge et al., 2003b). From a list of chemical shift assignments and a peaklist from NOE spectra ARIA generates ADRs (as introduced in chapter 1.2.2) that contain all assignment possibilities for each peak. Using these highly ambiguous restraints, a first ensemble of structures is calculated. Based on these initial structures the possible assignments with the largest distance in the structure are removed from the ADR and another round of structure calculation is performed. A similar approach is followed by CANDID/DYANA (Herrmann et al., 2002b). In addition to the procedures used by ARIA, additional mechanisms are implemented to ensure convergence of structures in the first round of calculations. These mechanisms are ‘network anchoring’ and ‘restraint combination’. In ‘restraint combination’ the restraints for two unrelated peaks are grouped in a single virtual ADR that will be fulfilled by the correct structure but not by other conformations. This greatly decreases the chance that ADR in which none of the assignment options is correct distort the structure. The ‘network anchoring’ method introduces a weighting factor for each restraint that evaluates the consistency with other restraints. This is based on the observation that usually true NMR restraints form a self-consistent network, in which

the information added by a single restraint is also contained in the combination of others. (See also the discussion of QUEEN in chapter 1.4.2). This weighting factor reduces the impact of less supported restraints that might otherwise distort the structure or prevent the detection of the correct fold in the first few steps of the structure calculations.

Both procedures try to remove the likelihood of wrong restraints in the first cycle of structure calculation, to avoid distorting the initial structures, on which all subsequent data analysis is based.

A similar but differently implemented approach to automate NOE-assignment is PASD (Kuszewski et al., 2004). Here the likelihood for a certain assignment to be correct is assessed in subsequent cycles of structure calculations, where all possible assignments for a peak are included as restraints. These assignments are switched on and off in a stochastic manner and the likelihood for an assignment to be correct is calculated. This likelihood then determines the frequency, which a certain assignment is used with. The authors claim that the algorithm tolerates up to 80 % wrongly assigned long range restraints, since the only information that is passed to subsequent iterations is the likelihood for a restraint to be correct. This independence from the structures of the previous calculations makes it much less important to obtain reasonable structures in the first cycle.

The most radical approach to both the resonance- and the NOE-assignment problem is CLOUDS (Grishaev et al., 2002). The authors of this algorithm present a completely new way to calculate NMR structures. In its basic implementation CLOUD assumes that each chemical shift in the ^{1}H-^{1}H-NOESY corresponds to a single proton. A NOESY cross peak therefore determines the distance between two protons. The network of distances that connect each proton to others should in principle allow only a single spatial position for each proton relative to each others, even if the identity of the proton and thus its connection to other atoms via bonds is unknown. Using a

simulated annealing algorithm the authors of CLOUDS were able to determine 'proton-densities', describing the likelihood of a proton to be located at a certain position, into which the covalent structure of the protein could be fitted. This way, they were able to obtain the structure of the protein and, as a side product, also the resonance assignment. This basic approach, although brilliant and effective for small proteins, suffers from a major drawback when applied to 'real world' proteins. It is completely unable to deal with overlap of ^{1}H chemical shifts. To address this issue, BACUS (Grishaev et al., 2004) and ABACUS (Grishaev et al., 2005) were developed, which work on molecular fragments, which represent spin-systems identified by COSY- and TOCSY-type experiments, instead of considering only single protons. These procedures are able to deal with overlap and ambiguities and may be used also for larger proteins.

Another recent development is to integrate the peak-picking procedure into the assignment/structure calculation procedure. ATNOS (Herrmann et al., 2002a) is a peak picking software that interfaces with CANDID. It helps to overcome the sensitivity to either very noisy or incomplete peaklists of ARIA or CYANA-like assignment strategies. Since ATNOS has information on all possible chemical shifts of the protein, it is able to pick only those peaks in the spectra that have chemical shifts corresponding to the assignments. Thus, only peaks that can in principle be assigned get picked reducing the amount of noise in the peaklists without compromising the sensitivity. Otherwise, high quality peaklists would have to be obtained manually, since only visual inspection of the peaks that get picked provide a robust quality control. Automated peak picking routines ususally have no means to determine peaks from noise and tend to be very sensitive to 'noise ridges'.

1.4 Quality Assessment of NMR-Structures

In the past, the acceptance of NMR spectroscopy in protein structure determination suffered from the fact that it does not provide a single structure as a best solution to

explain the data, but a set of possible solutions that are compliant with the data measured. It is impossible to know the "true" structure of a protein in solution, and some differences to crystal structures are expected due to the different states of the proteins. Although it could be shown that most of the structures determined by NMR are essentially the same as those that are determined by X-ray crystallography from the same protein construct, some uncertainty remains regarding the correctness and precision of the NMR structures. In comparison to X-ray crystallography, where the measured refraction pattern is directly dependent on the electron density and thus on the structure, NMR provides only a much more ambiguous set of pair-wise atom distances. While there are methods that allow addressing the questions of accuracy and precision in X-ray crystallography (e.g. the "R-factors", "B-factors" and the "completeness"), it has been difficult to establish similar parameters for NMR protein structure determination. In addition a good overall geometry has to be maintained during the structure calculation and favourable contacts of sidechains which are not included in the NMR data have to be established. Thus NMR structure determination has a third quality parameter, the 'protein likeness' in which it is compared to other well-structured proteins.

1.4.1 Accuracy

There are two main features of NMR protein structure determination that make it difficult to estimate how well the final structure determined explains the measured data. First, there is no straightforward way, a forward model, to describe the relation between the structure and the spectrum which is the data that can be directly measured. Secondly, the structure calculation relies on the assignment of peaks to atoms and the calibration for the translation of peak volumes into distances, which are then used to calculate the structures. Both introduce a bias into the structure calculation as well as into the back-calculation of the spectra from the structure. For a true un-biased back-calculating approach both the chemical shift assignment as well as the NOESY pattern should be predicted from theory. While such a prediction should be in principle possible, the number of parameters that is required to perform these

calculations (such as relaxation rates, shielding properties and coupling constants) make it impossible for analyzing proteins. Many of those parameters are unknown and most of them are very difficult to measure. To extract all these parameters just to be able to tell if the structure determined is correct would cause both measurement and analysis time to increase to unbearable length.

Still, the measure most commonly used for the accuracy of a NMR structure determination is a combination of the restraint violation analysis with basic quality checks. It is assumed that only a correct assignment will lead to a unique ensemble of structures not violating the restraints extracted from the data and containing features comparable to other proteins. However, this approach is only able to differentiate between good and bad structures and contains little information about what to improve in case of a bad one. The violation analysis is problematic. First, the assignments can be consistent but wrong, indicated only by bad normality scores. Second, the violations do usually not only occur at wrong restraints but also at correct ones that contradict wrong ones. Thus, violated restraints indicate areas in the structure where problems occur rather than the problematic restraints directly.

To circumvent the problem of cause and effect while evaluating NMR restraints, a complete cross-validation has been proposed (Brunger et al., 1993). This method randomly excludes a certain part of the distance information and performs the full calculations. If enough test sets are calculated, the impact of a single restraint on structure can be evaluated. Another method that is able to report the impact of single restraints is QUEEN (Qualitative Evaluation of Experimental NMR restraints) (Nabuurs et al., 2003). This method relies on distance matrices and is able to report the relative average information content of each single restraint without taking the structure into account.

A recent development towards a real R-factor for NMR structure determination is the program RFAC (Gronwald et al., 2000). In this approach the measured NOE data is

compared to the NOE data back-calculated from the structure. It relies on a correct resonance assignment and is therefore in the best case able to measure the quality of the structure calculation alone.

A new approach in the field of NMR-structure evaluation is the so-called RPF-score (Huang et al., 2005). This score is based on information-retrieval theory and allows comparing the information contained in the NOESY spectra with the structural ensemble. The graph of the distance network of proton-proton distances in the structure is compared to the one that can be generated from the NOESY data taking into account all assignment possibilities. The score provides a statistical measure for the agreement of both distance networks.

Residual dipolar couplings (RDCs) provide another very interesting tool for NMR structure quality measurements. Both the molecular alignment orientation of the molecule in the medium and the expected RDC for a given pair of interacting nuclei can be calculated from the structure providing the means for a true R-factor based on the use of RDCs (Clore et al., 1999). Although the structural information of the RDCs could in principle only be used for quality checks on structures calculated from NOE-data alone, RDC data is usually used the to further refine the structures. The quality check can then still be performed in a complete cross-validated manner.

1.4.2 Precision

Perhaps the most difficult question to answer is how the precision, that means the certainty in Å that the positions of the individual atoms can be given with, can be determined. In contrast to X-ray crystallography, where this is a parameter that is inherent to the data set, it is subject to discussion in NMR.

The measure of precision most commonly used is the convergence of the ensemble of structures generated from the MD-simulations. Usually, about 200 structures are

calculated and sorted by their energy according to the force-field. From these 200 structures the 20 with the least energy are then used as the final ensemble and their root mean square deviation (RMSD) from the average structure is reported as the precision of "the NMR-structure". There are many reasons why it is difficult to judge whether this information is representative of the precision of the ensemble or not. Perhaps the biggest problem is that the 20 (or 10 %) best structures in terms of energy might not be representative of the structural ensemble at all. They simply represent those structures that have the lowest energy in the ensemble, which might contain other conformations that are equally well populated but have for some reason a slightly higher energy. In general, the convergence of the final ensemble is more likely to represent the design of the force-field and the simulated annealing algorithm than the precision of the structure.

Some attempts have been made to solve this problem. It has been proposed to report all of the RMDS describing the whole ensemble. This allows to evaluate if the structures reported are only a minority or really representatives of the ensemble. However, this does not yield a single number that makes the precision of the determined structure comparable to other NMR or X-ray determined structures.

Recently a method has been proposed that could provide a general measure for NMR precision (Spronk et al., 2003). The authors proposed that the RMSD of the most divergent ensemble possible not violating the NMR restraints is a far better measure of the precision of the structure determination than the deviations within an ensemble describing the global minimum of the proteins conformational space.

1.4.3 Protein Normality

As opposed to well resolved X-ray structures, NMR structures completely rely on modelling to deduce the local geometry of the connections between atoms. In X-ray crystallography the distribution of atoms in space and thus also their relative positions

are directly related to the measured electron density. In contrast, the positions of all heavy atoms in NMR structure calculations is deduced from the relative positions of the hydrogens with use of a force-field that restricts bond angles and lengths. Thus, NMR structures have a strong model character, requiring careful modelling of all interactions not described in the NMR data. Mainly the angle distributions of dihedral angles in the ensemble and protein normality scores like the one produced by the software WhatIf (Vriend, 1990) are used to judge the 'protein-likeness' of the structures.

Recently, some effort has been dedicated to the development of force-fields for NMR structure determination. It has been widely recognized that there are different requirements for a force-field for NMR structure determination compared to a force-field for molecular dynamics simulations. This is simply due to the fact that in NMR structure determination the goal is usually to obtain a realistic structure as fast as possible, while MD simulations focus on realistic dynamics of the protein. A large contribution to the development of NMR specific force-fields has been made during the development of the automatic NOE assignment software ARIA (Linge et al., 2001). There are, for example, extensive studies about the effect of different sets of parameters concerning the quality, of the obtained structures (Linge et al., 1999), which is here defined as protein normality.

Another important aspect of achieving more realistic protein structures from MD-simulations is the refinement of the structures in explicit solvent. Especially the sidechains of charged amino acids that are extended into the solvent behave more naturally than in vacuum simulations. A very efficient protocol for this procedure (Linge et al., 2003a) has been used to refine a whole database of NMR structures (Nabuurs et al., 2004). In the course of this study it could be shown that the protein normality of all these proteins was increased compared to the originally published structures although exactly the same set of restraints was used.

1.4.4 Discussion

Although many efforts have been made to find a quality measure for NMR structure determination similar to those long established in X-ray crystallography, only little progress has been achieved so far. This is mainly due to the fact that NMR structure determination consists of many error prone interpretation steps. While it should in principle be possible to calculate the parameters determined in these steps from the structure alone, this neither computationally feasible nor as robust as required for a truly independent quality measure.

All quality measures that are available at the moment can be used for specific tasks such as estimating the overall correctness of the assignment or the precision of the generated ensemble even if they have their specific shortcomings. A single true quality measure is not available at the moment. Thus, in this thesis the quality of structures will be mainly measured by traditional means such as convergence of the ensemble and the distribution of the backbone dihedral angles in the Ramachandran-plot. For two of the three proteins water-refinement procedures will be used to improve the overall geometry of the proteins.

1.5 Aims of this Thesis

Structural genomics presents new challenges to NMR structure determination. Although recent developments have accelerated the process considerably, determination of high quality structures by NMR methods still takes months rather than weeks.

The strongest methodological advance came through automated structure calculation / NOE-assignment methods like CANDID/CYANA and ARIA based on ambiguous distance restraints (ADRs). Promising results on test proteins suggested that these methods would play an important role in a structural genomics context. However, it remained to be seen how reliable these methods worked when applied in

an high-throughput manner. Thus, an aim of this thesis was to explore the possibilities of automated NMR structure determination and to define standard procedures that allow to take advantage of these recent developments without compromising the quality of the structures.

During this thesis, the structures of three small protein domains, the BAG-domain from human SODD, the An1-like zinc finger domain from the human hypothetical protein BC018415 and the B8 subunit from Complex I, were to be solved using the new ADR-based NOE-assignment methods. One aim of this thesis was to define a strategy that represents an effective way for high throughput protein structure determination based on the experiences gained from the three structure determination projects. Furthermore, additional NMR or biophysical experiments were to be employed to explore functional hypotheses generated from the structures.

New methods for protein structure determination by NMR did not only increase the speed of the structure determination process but also allow to analyse larger and larger proteins. To aid the assignment process of proteins and complexes too large to be assigned by traditional means, a method to acquire information on amino acid types from ^{13}C-labelling pattern was to be developed using samples that were either labelled with 1,3-^{13}C-glycerol or 2-^{13}C-glycerol. This new method allowed the identification of amino acid types as easily as amino acid-specifically labelled samples, while also providing the sequential information. The method was to be tested on a small test protein, the α-spectrin SH3 domain, and procedures to represent and analyze the data were to be developed.

1.6 References for Chapter 1

Almeida,F.C., Amorim,G.C., Moreau,V.H., Sousa,V.O., Creazola,A.T., Americo,T.A., Pais,A.P., Leite,A., Netto,L.E., Giordano,R.J., and Valente,A.P. (2001). Selectively labeling the heterologous protein in Escherichia coli for NMR studies: a strategy to speed up NMR spectroscopy. J. Magn Reson. 148, 142-146.

Anfinsen,C.B., Harrington,W.F., Hvidt,A., Linderstrom-Lang,K., Ottesen,M., and Schellman,J. (1955). Studies on the structural basis of ribonuclease activity. Biochim. Biophys. Acta 17, 141-142.

Bartels,C., Xia,T.H., Billeter,M., Guntert,P., and Wuthrich,K. (1995). The Program Xeasy for Computer-Supported Nmr Spectral-Analysis of Biological Macromolecules. Journal of Biomolecular Nmr 6, 1-10.

Berman,H.M., Bhat,T.N., Bourne,P.E., Feng,Z., Gilliland,G., Weissig,H., and Westbrook,J. (2000). The Protein Data Bank and the challenge of structural genomics. Nat. Struct. Biol. 7 Suppl, 957-959.

Brenner,S.E. (2000). Target selection for structural genomics. Nat. Struct. Biol. 7 Suppl, 967-969.

Brunger,A.T., Clore,G.M., Gronenborn,A.M., Saffrich,R., and Nilges,M. (1993). Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. Science 261, 328-331.

Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W., and Swaminathan,S. (1999). Structural genomics: beyond the human genome project. Nat. Genet. 23, 151-157.

Cai,M., Huang,Y., Sakaguchi,K., Clore,G.M., Gronenborn,A.M., and Craigie,R. (1998). An efficient and cost-effective isotope labeling protocol for proteins expressed in Escherichia coli. J. Biomol. NMR 11, 97-102.

Chen,J., Nietlispach,D., Shaka,A.J., and Mandelshtam,V.A. (2004). Ultra-high resolution 3D NMR spectra from limited-size data sets. J. Magn Reson. 169, 215-224.

Clore,G.M., and Garrett,D.S. (1999). R-factor, Free R, and Complete Cross-Validation for Dipolar Coupling Refinement of NMR Structures. J. Am. Chem. Soc. 121, 9008-9012.

Clore,G.M., and Schwieters,C.D. (2003). Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1H/15N chemical shift mapping and backbone 15N-1H residual dipolar couplings using conjoined rigid body/torsion angle dynamics. J. Am. Chem. Soc. 125, 2902-2912.

Cornilescu,G., Bax,A., and Case,D.A. (2000). Large Variations in One-Bond 13Calpha-13Cbeta J Couplings in Polypeptides Correlate with Backbone Conformation. J. Am. Chem. Soc. 122, 2168-2171.

Ding,K., and Gronenborn,A.M. (2002). Sensitivity-enhanced E.COSY-type HSQC experiments for accurate measurements of one-bond 15N-1H(N) and 15N-13C' and two-bond 13C'-1H(N) residual dipolar couplings in proteins. J. Magn Reson. 158, 173-177.

Eghbalnia,H.R., Bahrami,A., Tonelli,M., Hallenga,K., and Markley,J.L. (2005). High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. J. Am. Chem. Soc. 127, 12528-12536.

Grishaev,A., and Llinas,M. (2004). BACUS: A Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems. J. Biomol. NMR 28, 1-10.

Grishaev,A., and Llinas,M. (2002). CLOUDS, a protocol for deriving a molecular proton density via NMR. Proc. Natl. Acad. Sci. U. S. A 99, 6707-6712.

Grishaev,A., Steren,C.A., Wu,B., Pineda-Lucena,A., Arrowsmith,C., and Llinas,M. (2005). ABACUS, a direct method for protein NMR structure computation via assembly of fragments. Proteins 61, 36-43.

Gronwald,W., Kirchhofer,R., Gorler,A., Kremer,W., Ganslmeier,B., Neidig,K.P., and Kalbitzer,H.R. (2000). RFAC, a program for automated NMR R-factor estimation. J. Biomol. NMR 17, 137-151.

Grzesiek,S., and Bax,A. (1992a). An efficent experiment for sequential backbone amide assignment of medium sized isotopically enriched proteins. J. Magn Reson. 99, 201-207.

Grzesiek,S., and Bax,A. (1992b). Correlating backbone amide and sidechain resonances in proteins by multiple triple resonance NMR. J. Am. Chem. Soc. 115, 11620-11621.

Herrmann,T., Guntert,P., and Wuthrich,K. (2002b). Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J. Mol. Biol. 319, 209-227.

Herrmann,T., Guntert,P., and Wuthrich,K. (2002a). Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J. Biomol. NMR 24, 171-189.

Hol,W.G. (2000). Structural genomics for science and society. Nat. Struct. Biol. 7 Suppl, 964-966.

Huang,Y.J., Powers,R., and Montelione,G.T. (2005). Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J. Am. Chem. Soc. 127, 1665-1674.

Hyberts,S.G., and Wagner,G. (2003). IBIS--a tool for automated sequential assignment of protein spectra from triple resonance experiments. J. Biomol. NMR 26, 335-344.

Johnson,B.A. (2004). Using NMRView to visualize and analyze the NMR spectra of macromolecules. Methods Mol. Biol. 278, 313-352.

Kim,S., and Szyperski,T. (2003). GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J. Am. Chem. Soc. 125, 1385-1393.

Kuszewski,J., Schwieters,C.D., Garrett,D.S., Byrd,R.A., Tjandra,N., and Clore,G.M. (2004). Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. J. Am. Chem. Soc. 126, 6258-6273.

Led,J.J., and Gesmar,H. (1991). Linear prediction enhancement of 2D heteronuclear correlated spectra of proteins. J. Biomol. NMR 1, 237-246.

Linge,J.P. (2000). New methods for automated NOE assignment and NMR structure calculation.Books on Demand).

Linge,J.P., and Nilges,M. (1999). Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. J. Biomol. NMR 13, 51-59.

Linge,J.P., O'Donoghue,S.I., and Nilges,M. (2001). Automated assignment of ambiguous nuclear overhauser effects with ARIA. Methods Enzymol. 339, 71-90.

Linge,J.P., Williams,M.A., Spronk,C.A., Bonvin,A.M., and Nilges,M. (2003a). Refinement of protein structures in explicit solvent. Proteins 50, 496-506.

Linge,J.P., Habeck,M., Rieping,W., and Nilges,M. (2003b). ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19, 315-316.

Liu,G., Shen,Y., Atreya,H.S., Parish,D., Shao,Y., Sukumaran,D.K., Xiao,R., Yee,A., Lemak,A., Bhattacharya,A., Acton,T.A., Arrowsmith,C.H., Montelione,G.T., and Szyperski,T. (2005). NMR data collection and analysis protocol for high-throughput protein structure determination. Proc. Natl. Acad. Sci. U. S. A 102, 10487-10492.

Montelione,G.T., Lyons,B.A., Emerson,S.D., and Tashiro,M. (1992). An efficient triple resonace experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically enriched proteins. J. Am. Chem. Soc. 114, 6291-6293.

Morris,G.A., and Freeman,R. (1979). Enhancement of Nuclear Magnetic-Resonance Signals by Polarization Transfer. J. Am. Chem. Soc. 101, 760-762.

Moult,J., Fidelis,K., Tramontano,A., Rost,B., and Hubbard,T. (2005). Critical assessment of methods of protein structure prediction (CASP) - round VI. Proteins.

Nabuurs,S.B., Nederveen,A.J., Vranken,W., Doreleijers,J.F., Bonvin,A.M., Vuister,G.W., Vriend,G., and Spronk,C.A. (2004). DRESS: a database of REfined solution NMR structures. Proteins 55, 483-486.

Nabuurs,S.B., Spronk,C.A., Krieger,E., Maassen,H., Vriend,G., and Vuister,G.W. (2003). Quantitative evaluation of experimental NMR restraints. J. Am. Chem. Soc. 125, 12026-12034.

Neal,S., Nip,A.M., Zhang,H., and Wishart,D.S. (2003). Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. J. Biomol. NMR 26, 215-240.

Oschkinat,H., and Croft,D. (1994). Automated assignment of multidimensional nuclear magnetic resonance spectra. Methods Enzymol. 239, 308-318.

Page,R., Peti,W., Wilson,I.A., Stevens,R.C., and Wuthrich,K. (2005). NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. Proc. Natl. Acad. Sci. U. S. A 102, 1901-1905.

Prestegard,J.H., Valafar,H., Glushka,J., and Tian,F. (2001). Nuclear magnetic resonance in the era of structural genomics. Biochemistry 40, 8677-8685.

Riek,R., Wider,G., Pervushin,K., and Wuthrich,K. (1999). Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules. Proc. Natl. Acad. Sci. U. S. A 96, 4918-4923.

Sattler, M., Schleucher, J., and Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Progress in Nuclear Magnetic Resonance Spectroscopy 34, 93-158. 1999.
Ref Type: Generic

Scheich,C., Sievert,V., and Bussow,K. (2003). An automated method for high-throughput protein purification applied to a comparison of His-tag and GST-tag affinity chromatography. BMC Biotechnology 3, 12.

Schwieters,C.D., Kuszewski,J.J., Tjandra,N., and Marius,C.G. (2003). The Xplor-NIH NMR molecular structure determination package. J. Magn Reson. 160, 65-73.

Shuker,S.B., Hajduk,P.J., Meadows,R.P., and Fesik,S.W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. Science 274, 1531-1534.

Slupsky,C.M., Boyko,R.F., Booth,V.K., and Sykes,B.D. (2003). Smartnotebook: a semi-automated approach to protein sequential NMR resonance assignments. J. Biomol. NMR 27, 313-321.

Spronk,C.A., Nabuurs,S.B., Bonvin,A.M., Krieger,E., Vuister,G.W., and Vriend,G. (2003). The precision of NMR structure ensembles revisited. J. Biomol. NMR 25, 225-234.

Valafar,H., Mayer,K.L., Bougault,C.M., LeBlond,P.D., Jenney,F.E., Jr., Brereton,P.S., Adams,M.W., and Prestegard,J.H. (2004). Backbone solution structures of proteins using residual dipolar couplings: application to a novel structural genomics target. J. Struct. Funct. Genomics 5, 241-254.

Vitkup,D., Melamud,E., Moult,J., and Sander,C. (2001). Completeness in structural genomics. Nat. Struct. Biol. 8, 559-566.

Vriend,G. (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 8, 52-6, 29.

Zimmerman,D.E., Kulikowski,C.A., Huang,Y., Feng,W., Tashiro,M., Shimotakahara,S., Chien,C., Powers,R., and Montelione,G.T. (1997). Automated analysis of protein NMR assignments using methods from artificial intelligence. J. Mol. Biol. 269, 592-610.