Efficient Recording and Processing of Protein NMR Spectra

DANIEL MALMODIN
Department of Chemistry
Göteborg University

GÖTEBORG UNIVERSITY

Efficient Recording and Processing of Protein NMR Spectra

Daniel Malmodin

Institutionen för kemi
Göteborgs Universitet
2006

AKADEMISK AVHANDLING

för filosofie doktorsexamen i kemi som med medgivande av Institutionen för kemi, Göteborgs Universitet, kommer att försvaras offentligt fredagen 7:e april 2006 kl.10.00 i föreläsningssal Nils Nilsson, A2053, Medicinaregatan 3, Göteborg

Fakultetsopponent: Gerhard Wider, Molecular Biology and Biophysics, ETH Zurich

Avhandlingen försvaras på engelska
Göteborg, 2006

ABSTRACT

NMR is a technique with very broad applications within the field of proteins and DNA as it often can provide exclusive information about their; structure, folding properties, mobility, and interactions with other molecules, which is not always possible to measure using other methods. Two major drawbacks with the method are the complex output from the experiments, which usually requires extensive manual investigation before the information content from the experiment can be explained properly, and the traditional choice of data recording where data is recorded on a grid for subsequent Fourier transform to frequency domain, which in combination with the insensitivity of the method itself make experiments very time demanding and sometimes not even practically feasible at all.

This thesis describes methods, implemented in computer programs, which aim to reduce these problems. The first part, "Automated Analysis of regular NMR-spectra", shows that it is possible to assign proteins of the size of the 128 aa protein azurin with existent automated peak picking, and automated assignment program packages, in conjunction with an automatic calibration routine, opening for the possibility to perform studies of mobility, interaction, or structure without having to go through the tedious manual peak picking and assignment procedure first. To show that labelling is not necessary for automated assignment the procedure is also applied on the 29 aa, non-labelled, defensin HNP2 with a weakly bounded ligand. The structure is also solved using the assignments.

Recently, recording of spectra with coupled evolution periods has gained a lot of interest due to its ability to reduce the recording time on the NMR instrument. Unfortunately, the resulting spectra are difficult to interpret due to that sums and differences of nuclei shifts are recorded instead of the nuclei shifts themselves, and that the peak information for every peak is split over many spectra. The second part, "Evaluation of spectra with coupled evolution periods" demonstrates two different procedures on how to calculate the true nuclei shifts and even the full NMR-spectrum from a set of projections from experiments using coupled evolution periods.

KEYWORDS: NMR, automation, peak picking, resonance assignment, coupled evolution periods, convolution, multi-way decomposition, proteins.

ISBN 13: 978-91-628-6750-8
ISBN 10: 91-628-6750-4

Efficient Recording and Processing of Protein NMR Spectra

Daniel Malmodin

Department of Chemistry
Göteborg University
2006

Efficient Recording and Processing of Protein NMR Spectra

Daniel Malmodin
Department of Chemistry
Göteborg University
SE-405 30 Göteborg

Thesis for the degree of doctor of philosophy

ISBN 13: 978-91-628-6750-8
ISBN 10: 91-628-6750-4

Chalmers reproservice
Göteborg 2006

Till Lisa och familjen

ABSTRACT

NMR is a technique with very broad applications within the field of proteins and DNA as it often can provide exclusive information about their; structure, folding properties, mobility, and interactions with other molecules, which is not always possible to measure using other methods. Two major drawbacks with the method are the complex output from the experiments, which usually requires extensive manual investigation before the information content from the experiment can be explained properly, and the traditional choice of data recording where data is recorded on a grid for subsequent Fourier transform to frequency domain, which in combination with the insensitivity of the method itself make experiments very time demanding and sometimes not even practically feasible at all.

This thesis describes methods, implemented in computer programs, which aim to reduce these problems. The first part, "Automated Analysis of regular NMR-spectra", shows that it is possible to assign proteins of the size of the 128 aa protein azurin with existent automated peak picking, and automated assignment program packages, in conjunction with an automatic calibration routine, opening for the possibility to perform studies of mobility, interaction, or structure without having to go through the tedious manual peak picking and assignment procedure first. To show that labelling is not necessary for automated assignment the procedure is also applied on the 29 aa, non-labelled, defensin HNP2 with a weakly bounded ligand. The structure is also solved using the assignments.

Recently, recording of spectra with coupled evolution periods has gained a lot of interest due to its ability to reduce the recording time on the NMR instrument. Unfortunately, the resulting spectra are difficult to interpret due to that sums and differences of nuclei shifts are recorded instead of the nuclei shifts themselves, and that the peak information for every peak is split over many spectra. The second part, "Evaluation of spectra with coupled evolution periods" demonstrates two different procedures on how to calculate the true nuclei shifts and even the full NMR-spectrum from a set of projections from experiments using coupled evolution periods.

KEYWORDS: NMR, automation, peak picking, resonance assignment, coupled evolution periods, convolution, multi-way decomposition, proteins.

LIST OF PUBLICATIONS

I. Fully automated sequence-specific resonance assignments of heteronuclear protein spectra (2003)

Daniel Malmodin, Christina H.M. Papavoine, and Martin Billeter

Journal of Biomolecular NMR 27 (1) 69-79.

II. Signal identification in NMR spectra with coupled evolution periods (2005)

Daniel Malmodin and Martin Billeter

Journal of Magnetic Resonance 176 (1) 47-53.

III. Multiway decomposition of NMR spectra with coupled evolution periods (2005)

Daniel Malmodin and Martin Billeter

Journal of the American Chemical Society 127 (39) 13486-13487.

IV. Robust and Versatile Interpretation of Spectra with Coupled Evolution Periods using Multi-Way Decomposition

Daniel Malmodin and Martin Billeter

Accepted in Magnetic Resonance in Chemistry.

V. Components correlating 13 nuclei from simultaneous multi-way decomposition of several spectra

Daniel Malmodin, Wolfgang Bermel, and Martin Billeter

Manuscript.

Related publication:

High-throughput analysis of protein NMR spectra (2005)

Daniel Malmodin and Martin Billeter

Progress in Nuclear Magnetic Resonance Spectroscopy 46 (2-3) 109-129.

LIST OF ABBREVIATIONS

2D	two-dimensional
3D	three-dimensional
5D	five-dimensional
(5,2)D	five reduced to two dimensional
aa	amino acid
COSY	correlation spectroscopy
DFT	discrete Fourier transform
GFT	G-matrix Fourier transform
HSQC	heteronuclear single quantum coherence
Hz	Hertz
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser enhancement
NOESY	nuclear Overhauser enhancement spectroscopy
ppm	parts per million
TOCSY	total correlation spectroscopy
TWD	three-way decomposition
Å	Ångström

TABLE OF CONTENTS

1. Introduction 1

1.1 NMR spectroscopy 2

Nuclear spin 2

Biomolecular NMR studies 3

NMR signal processing and interpretation 4

2. Methods 9

2.1 Automated analysis of regular NMR spectra 9

Automated peak picking using AUTOPSY 9

Automated calibration using PICS 10

Automated assignment using GARANT 10

2.2 Evaluation of spectra with coupled evolution periods 11

Automated calculation of individual chemical shifts using EVOCOUP 11

Three-way decomposition 13

Decomposition of regular NMR spectra 14

Automated decomposition using PRODECOMP 15

3. Results and discussion 17

3.1 Fully automated analysis of regular NMR-spectra 17

Fully automated resonance assignment of hetero-nuclear protein spectra 18

Fully automated resonance assignment of homo-nuclear protein spectra 20

3.2 Evaluation of spectra with coupled evolution periods 24

Analysis of the HACACONHN spectrum of azurin using the program EVOCOUP 24

Benefits and disadvantages of frequency domain decomposition of sparse recorded spectra 25

Analysis of the HACACONHN spectrum of azurin using the program PRODECOMP 26

A comparison of the results of EVOCOUP and PRODECOMP 27

Correlation of 13 nuclei in one component using PRODECOMP 29

4. Conclusion 31

Acknowledgements 33

References 34

1 INTRODUCTION

Basic research, as well as development of new, clinical, drugs, often relies on good descriptions of molecular mechanisms at the atomic level, therefore, structural characterisation of proteins is of great importance. Still, protein sequence databases like Swiss prot (Boeckmann et al., 2003) grow larger at much higher speed than structural databases like the Protein Data Bank (Berman et al., 2000), since proteins' structures, dynamics and interactions often remain unknown. There are many reasons why this information often is difficult to achieve. The initial steps, cloning, expression and sample preparation has to be successful, as well as the structural analysis itself. The development of all parts of this process is continuously progressing with the goal to gather enough information for being able to determine or predict structural information for all proteins. The methods that are used needs to be highly efficient in order to deliver a high-throughput.

The two most common methods to obtain structural information are crystallography (e.g. Drenth, 1994) followed by NMR (e.g. Levitt, 2001; Cavanagh et al., 1996). The two methods complement each other well since it has turned out that their success rate for small proteins do not correlate (Snyder et al., 2005). Also, by NMR it is possible to measure dynamics and interactions in ways not possible in crystallography. Although NMR is a very useful tool for protein structural analysis it has some features making it less attractive for normal, and in particular for high-throughput, protein studies; the information content is very high but at the same time often difficult to interpret and requires extensive manual investigation before the experiment can be properly explained, and the insensitivity of the method often makes the experiments very time demanding and sometimes not even practically feasible at all. These problems have been a part of NMR spectroscopy since the early days of the field and each time progress is made in this area it benefits protein NMR studies.

1.1 NMR spectroscopy

Nuclear spin

Nuclear Magnetic Resonance is a physical phenomenon describing the interaction between nuclear spins and a magnetic field. The spin of a nucleus has an integer or half integer value I and the nucleus itself has 2I+1 energy eigenstates with energies E_m according to

$$E_m = -m(h/2\pi)\gamma B_0 \qquad (1.1)$$

where m = -I, -I+1, ..., +I, h is Planck's constant $6.626 \cdot 10^{-34}$ Js, γ is the gyromagnetic ratio of the nucleus and B_0 the strength of the surrounding static magnetic field. The state of the nucleus is usually a superposition of its eigenstates.

Apart from deuterium, nuclei with I = 1 or more are not used in NMR studies of proteins because of their complicated transitions between states and their rather fast relaxation due to quadrupole moment interaction. Neither are nuclei with zero spin since they are always in the same state. This leaves nuclei with spin I = ½. The most important is the most common isotope (~100%) of hydrogen, ^{1}H, followed by the less usual isotopes of carbon (1.1%) and nitrogen (0.37%), ^{13}C and ^{15}N. The low natural abundance of the latter two is often not sufficient for NMR analysis and enrichment, "labelling", is therefore needed.

For nuclei with I = ½, the energy gap between the two eigenstates is

$$\Delta E = (h/2\pi)\gamma B_0, \qquad (1.2)$$

and transitions between these involve absorption or emission of electromagnetic radiation with angular, also called Larmor, frequency

$$\omega_0 = -\gamma B_0. \qquad (1.3)$$

Biomolecular NMR studies

At room temperature, the distribution of spins between the states follows Boltzmann statistics. In an NMR spectrometer with a high magnetic field the occupancy of the lower energy state is only slightly larger than in the higher. Still, this small macroscopic magnetisation is sufficient for measurements by applying pulses of a magnetic field perpendicular to the main field and measure the sinusoidal response when the magnetisation precess back towards equilibrium.

Currents due to electrons very close to the nuclei slightly change the local magnetic fields around these. The effective field for a nucleus is therefore

$$B_{eff} = (1-\sigma)B_0 \qquad (1.4)$$

where σ is called the shielding factor. Since the Larmor frequency is proportional to the magnetic field, each nucleus, also the ones with identical γ will have their own frequency

$$\omega = (1-\sigma)\,\omega_0, \qquad (1.5)$$

making it possible to obtain spectra, which in frequency domain give individual signals for the nuclei. Due to practical reasons, these frequencies are usually compared to a reference signal and the chemical shift

$$\delta = (\omega - \omega_{ref})/\,\omega_0 = \sigma_{ref}-\sigma, \qquad (1.6)$$

measured in parts per million, ppm, is used rather than the Larmor frequency itself.

Today, most NMR experiments correlate two or more different nuclei and are recorded in two or more dimensions. This allows not only individual characterisation of the shifts, but also the possibility to determine which nuclei are covalently bound to each other from scalar coupling interaction, or close in space from the nuclear Overhauser effect (NOE) which is used for structure determination.

Three more examples of valuable NMR techniques, not used though in this thesis, are residual dipolar coupling, relaxation, and interaction measurements. From

measurements of residual dipolar couplings between pairs of nuclei it is possible to decide the angle between a line through each pair, and a particular axis of the protein. The flexibility in various parts of a protein is important for its function and can be determined by relaxation measurements. Proteins interact with other proteins and molecules. In the development of new drugs it is essential to know these processes in detail. From a set of NMR experiments, where the protein is mixed with other proteins or molecules, it is possible to measure interactions with atomic precision.

When possible, labelling of the heavy atoms is of great help, especially for larger proteins. Its use has many advantages; by backbone experiments it is possible to record spectra that connect proton shifts in different residues without the need for NOESY experiments, it is possible to have evolution times also on the heavy nuclei and thereby reducing the risk of missing peaks due to overlap, and by selective labelling of different sets of heavy atoms it is even possible to reduce the number of peaks compared to original spectra reducing the complexity further and making the technique practical for even larger proteins.

NMR signal processing and interpretation

NMR is a low-sensitivity method. A fundamental relation between signal and noise is

$$S/N \propto NQ\gamma^{5/2}B_0^{3/2}T_2^{1/2}T^{-3/2}(t_{max}/T_c)^{1/2}, \qquad (1.7)$$

where N is the number of nuclear spins, Q is the quality factor of the probe coil, γ is the gyromagnetic ratio, B_0 is the static magnetic field strength, T_2 is the transverse relaxation time constant, T is the temperature, t_{max} is the total acquisition time, and T_c is the total time between acquisitions (Cavanagh et al., 1996). Therefore the signal is enhanced by a high concentration of the sample, a high magnetic field, a low temperature, and a long experimental time. There is a limit for how much these parameters can be increased though, and the signals are therefore often weak.

The signal strength decades as the spins precess back to equilibrium. This relaxation imposes a limited time for how long the nuclei actually are recorded, with the result that the signal frequencies cannot be determined exactly since they will have a natural linewidth

$$L = R_2/\pi, \quad (1.8)$$

where R2 is the relaxation rate of the nuclei.

Within the theoretical and practical limits of sensitivity and relaxation, the spectroscopist aims to get the information from the NMR measurements that best answers the questions asked. Of course, first an appropriate experiment must be chosen. Considerable efforts have been, and still are, done to develop pulse sequences to fulfil this purpose. The spectroscopist also must decide what information should be recorded, to what extent, and in what fashion, in order to get a satisfying result.

Today, the most popular way to sample protein NMR spectra is to record the nuclei spins such that the shifts can be immediately determined from peak shifts in the spectra. Often each type of nucleus shift is recorded in a separate dimension. The pulse sequence of the experiment is repeated and the output recorded over a large regular grid of evolution times and after the experiment is finished the data is transformed to frequency domain by successive discrete Fourier transforms (DFT), one for each dimension. Every step is linear in this process and therefore easy and robust.

A problem is that the digital resolution and the spectral width for each dimension are inversely proportional to the sampling interval of that dimension. Therefore, in the strict sense, a more narrow digital resolution and a constant spectral width can only be obtained by an increase of recorded data points. This is not entirely true in practice though, often it is possible in time domain to make a plausible guess about the values immediately after the recorded ones with a method called linear prediction making the digital resolution better. For cosmetic reasons it is popular to damp the signal at one end of the spectrum prior to the DFT and therefore errors in some predicted values will be even less crucial.

Both the DFT and linear prediction are one-dimensional procedures applied on two- or more dimensional problems. There are some methods like three-way decomposition (TWD) (Orekhov et al., 2001; Ibraghimov, 2002), maximum entropy reconstruction (Hoch and Stern, 2001; Stern et al., 2002), and the filter diagonalisation method (FDM2k) (Mandelshtam, 2001) that are multidimensional to

their nature. When using these methods it is possible to leave out some of the data with long evolution times in the experiment and thereby reduce the recording time without sacrificing resolution. TWD is described in some detail in Methods. It is worth mentioning though that TWD does not only manage missing data points and is good at removing noise, it also describes the spectrum as a sum of components which in essence means that it performs an automated assignment step, i.e. TWD does not only reconstruct data but it also makes an interpretation of the spectrum. In maximum entropy reconstruction, the spectrum in frequency domain for which a certain entropy function is at maximum, is reconstructed. The filter diagonalisation method does an explicit interpretation of the spectrum as a sum of decaying exponentials and determines these in terms of frequencies, phases and damping.

Although similar experiments have been performed before (Bodenhausen and Ernst, 1982), recently a new variant of recording, called GFT-spectroscopy (Kim and Szyperski, 2003) or projection-reconstruction spectroscopy (Kupce and Freeman, 2003), and which we call spectra with coupled evolution periods, has gained a lot of attention. By simultaneous increments of nuclei shifts in the indirect dimension and hyper complex recording, it is possible, after different schemes of DFT, to obtain a set of low dimensional spectra with sums and differences of the correlated spins instead of a "normal" full dimensional spectrum as described above. The advantage is that many spins can be correlated and measured with a good line width and digital resolution without having to record a very large number of data points. The low dimensional spectra can be viewed upon as projections from different angels of the full dimensional spectrum and are either analysed individually, and the results from the different angels compared (Kim and Szyperski, 2003; Moseley et al., 2004; Hiller et al., 2005; Paper II), or all at once by reconstruction of the full dimensional spectrum or a decomposition of this (Kupce and Freeman, 2005; Coggins et al., 2005; Paper III-V).

Almost all protein NMR-spectrum analysis requires assignment of nuclei shifts prior to further investigation. The traditional reference procedure involves two steps; peak picking in a set of spectra, followed by assignment of the nuclei. The assignment is based on that both the sequence of the protein is known and that the outcome of the experiments in terms of magnetic transfer, peaks, can be predicted. Although peak shifts cannot in advance be perfectly predicted these give good aid when assigning peaks to particular nuclei since the peak shifts usually do not differ a lot from random coil values. The final assignment should preferably identify all experimental peaks

with expected peaks. Although there are some programs available both for peak picking and assignment (for review, see e.g. "Related Paper", Malmodin and Billeter, 2005), still most spectroscopists do these steps in a mostly manual fashion.

2 METHODS

2.1 Automated analysis of regular NMR spectra

The normal procedure of NMR analysis, starting from a set of recorded spectra and with assigned peak lists as the final result, can be divided into three distinct parts. First the peaks in each spectrum are picked, then the peak lists from the spectra are compared and calibrated to each other, and finally all peak lists are considered simultaneously and the peaks are assigned to specific nuclei. Therefore, three different programs, each doing one of these tasks fully automatically, can be though of as sub-routines to a process that fully automatically delivers assigned peak lists directly from a set of spectra.

The already existing programs AUTOPSY (Koradi et al., 1998) and GARANT (Bartels et al., 1996, 1997) where chosen to do the first and last part of this work. The second step, the calibration, is not needed if all the spectra are recorded under exactly the same conditions. This is often difficult to guarantee due to the fact that different pulse sequences heats the sample differently etc. and therefore the peak lists are calibrated prior to the assignment. A program called PICS, which does this automatically, was written for this reason. The three programs AUTOPSY, PICS, and GARANT were then run in sequence.

Automated peak picking using AUTOPSY

A variety of automatic algorithms are available for picking peaks in NMR spectra. The common goal for all peak pickers is to find all "true peaks" without picking false positives. Two major problems for both spectroscopists and automated peak pickers are; how to judge between very small peaks and noise, and how to extract the correct individual peaks from crowded regions.

The program AUTOPSY picks peaks based primarily on symmetry and size (Koradi et al., 1998). In regions of the spectrum with signal intensities sufficiently larger than noise it searches for symmetric parts and identifies these as peaks. The user provides lower boundaries on the size of these regions and peaks, as well as

symmetry requirements on the peaks. When a peak is identified it is subtracted from the spectrum and the search continues. An option for the user, after the first search for peaks, is to use the picked peaks and search the spectrum ones more with the restriction that a new peak must have some lineshapes identical to some already picked peak. This reduces the risk of missing peaks, especially in NOESY type of spectra where some peaks often are very small but where larger peaks guides the spectroscopist in the search for smaller ones. Although AUTOPSY previously had been tested only on 2D spectra, in terms of user friendliness, 3D spectra are equally easy to process.

Automated calibration using PICS

The program PICS, implemented in MATLAB code (The MathWorks, Inc.), makes the calibration more reliable. It searches for an optimal calibration by matching all peak positions in one spectrum on top of all peak positions in the other spectrum for relevant dimensions. The user defines maximum expected shift differences for peaks in different spectra originating from the same nuclei. The program then calibrates the axes in the first spectrum in an iterative fashion. First, it subtracts all peak positions in the first spectrum from all peak positions in the second spectrum. Second, for all differences with absolute values smaller or equal to the user-defined values it identifies the median difference value in each dimension. Third, it changes the scales of the first spectrum by adding constant offsets in each dimension such that the median difference values become zero. Then, PICS starts over again at the first step and iterates until the shift differences are equally distributed around zero. The program also presents histograms of the remaining shift differences giving the spectroscopist an idea about the precision of the calculation of the peak shifts.

Automated assignment using GARANT

GARANT is an automatic assignment program that tries to optimally map a set of assigned, theoretically predicted peak list without peak coordinates, on unassigned, experimental peak lists (Bartels et al., 1996, 1997). The program has a library with already defined experiments and uses this and the sequence of the protein to construct theoretical peak lists. In addition to the theoretically predicted peak lists, the

program uses a library file with expected random coil shifts for all nuclei in the different types of amino acids.

The output is a set of merged peak lists with the assignments taken from the predicted lists and the peak shifts from the experimental peak lists. GARANT also provides quantitative numbers for each nucleus telling how sure it is of the assignment and sometimes it also suggests alternative assignments.

The spectroscopist can, by changing the original code, rather easy add experiments to the already existing list. Therefore the assignment process itself does not restrict the choice of experiments as long as they are recorded in the traditional, full dimensional way. The algorithm is generic and the user defines the size of the populations.

2.2 Evaluation of spectra with coupled evolution periods

Automated calculation of individual chemical shifts using EVOCOUP

Although peak picking works the same way in spectra with coupled evolution periods, and in normal full dimensional spectra, the assignment process is different. The peak shifts do not immediately give the shifts of the nuclei, but sums and differences of these. Using peak lists from a set of such spectra, the program EVOCOUP written in MATLAB code (The MathWorks, Inc.) calculates the peak shifts of the individual nuclei, which later can be used for traditional assignment.

Written in matrix form, the vector w containing n, true, chemical shifts, is determined from the vector p, with ideally m peak coordinates observed in m different spectra, and a (m*n)-matrix A describing the linear combinations specific to a given experiment

$$A\,w \sim p. \qquad (2.1)$$

The rank of A has to be at least n and therefore the number m of equations has to be at least as large as the number of unknown chemical shifts n. Typically A is over-determined and equation 2.1 can in general not be strictly fulfilled. Instead, an in the

least square sense optimal solution w' of w can be obtained by calculating the pseudoinverse A' of A using singular value decomposition and multiplying this matrix with the coordinates of the projected peaks

$$w' = A' p. \qquad (2.2)$$

From equation 2.1 and the calculated values w' it is possible to calculate approximate values p', for the input vector p. The difference between p and p' is a good internal measure of the reliability of w' and also the set of peaks p.

EVOCOUP analyses one small region in the traditionally recorded direct dimension at a time. The program uses all peaks from all spectra within the chosen region and tries to match these with possible nuclei shift. When there is no overlap in the traditionally recorded region this is trivial. More interesting are situations with many spin systems overlapping in this dimension. Since the number of peaks in each spectrum then can be large, it would be very computer demanding to calculate all peak combinations for all spectra. Also, it cannot be taken for granted that all peaks will be detected in the first peak-picking step. The peak lists can have false, as well as lacking, peaks. EVOCOUP tries to quickly find the correct peak combinations from the definition of A, by discriminating between different solution vectors in the following way:

1. If possible, from the list of experimental peaks p_{exp}, a not already tested peak list p_{temp}, is initialised with n+1 peaks from different spectra, which together has rank n.
2. A peak list p' is calculated for the dimensions of p_{temp}, using p_{temp} and equations 2.2 and 2.1.
3. $|p_{temp} - p'|$ is calculated to check that all chosen peaks are consistently reproduced.
a If yes, the list p_{cons} is set equal to p_{temp}.
b If not, and if a p_{cons} exists and includes an acceptable number of peaks, this list is moved to a final peak list p_{final} and the peaks are also removed from p_{exp}.
The program starts over at 1.
4. It is checked if there are remaining spectra r that include peaks.
a. If yes, $|p_{exp}(r)-p_{calc}(r)|$ is calculated for these, where $p_{calc}(r)$ is a calculation of possible peak shifts in r using p_{cons}. If there are one or more peaks in $p_{exp}(r)$ that are consistent with p_{calc}, the program picks the peak with the smallest absolute value in the calculation and adds it to p_{temp}, and goes back to the second step.
b. If no, same as 3b.

The "acceptable number of peaks" in 3b is to begin with a rather tough criterion; one peak from every spectrum has to be included. When all combinations have been tried this is relaxed by one peak and the process starts over again and so on. The user decides the lower number of projected peaks, which has to be included in the final peak list as belonging to the same, true peak.

Three-way decomposition

Three-way decomposition (TWD), also referred to as parallel factor analysis (PARAFAC) (Bro, 1997), and canonical decomposition (Caroll and Chang, 1970), is a mathematical approximation procedure where a 3D matrix is decomposed into a sum of components where each component is defined as the direct product of a set of one dimensional vectors, also called shapes. Historically the method has been used for relatively small problems in psychometrics (Caroll and Chang, 1970) and chemometrics (reviewed in Journal of Chemometrics, 2000, Vol. 14, No. 3). A recently published book describes the method itself (Smilde et al., 2004).

Almost always it is valid to describe evolution and delay periods in NMR experiments with average Liouvillian operators, and then so, there is a direct connection between TWD and NMR experiments making it very attractive to describe NMR spectra accordingly (Orekhov et al., 2001; Billeter and Orekhov, 2003).

Mathematically, the approximation can be formulated as

$$S = \sum_{n=1}^{R} a_n F1_n \otimes F2_n \otimes F3_n \qquad (2.3)$$

with S representing the 3D input spectrum, and R the number of components. F1, F2, and F3, are matrices where the columns correspond to different shapes. Often the shapes are normalized in the output and the amplitude of each component is shown in the number a_n.

The largest integer k for which a matrix X is universal k-column independent is called the k-rank of X and is denoted by k_x. A sufficient, and for less than four components necessary, condition for TWD to give unique parameter estimates is

$k_1+k_2+k_3>=2R+2$ (2.4)

where k_1, k_2 and k_3 are the k-ranks of the component matrices F1,F2 and F3 (described further in e.g. Smilde et al., 2004).

For example, this requirement is not met in the case of 2D spectra. A 2D spectrum can be thought of as a 3D spectrum but with only one index in the third dimension. The corresponding normalised component matrix F3 will then be a column vector where all the R columns are equal to one and therefore also the k-rank will be one for this matrix. Since the k-ranks of the other two component matrices are not larger than R, equation 2.4 is not fulfilled.

The same is true for a 3D spectrum, where some components effectively only has two dimensions, i.e. a 3D spectrum with two or more components having identical or nearly identical shapes in one dimension. The effect is called "mixing" and can be explained from

$$X = F_1 A F_2^T = F_1 A U^{-1} U F_2^T = (F_1 A U^{-1})(U F_2^T) = \tilde{F}_1 \tilde{F}_2^T \quad (2.5)$$

Where X represents the 2D input spectrum, F_1 and F_2 are the component matrices along the two dimensions, respectively, and A is the diagonal matrix of component amplitudes. The equation shows that it is always possible to find an arbitrary, unitary matrix U with the same size as A and obtain another description of X by defining the shapes differently.

Decomposition of regular NMR spectra

The goal with the decomposition is to find a solution that minimizes

$$\min(|S - \sum_{n=1}^{R} a_n F1_n \otimes F2_n \otimes F3_n|^2). \quad (2.6)$$

There are different options how to solve equation 2.6. The method that has been applied for NMR studies in our group is called alternating least squares and is probably the most common one due to that it performs well and is easy to implement.

Though convergence to the global minimum is not guaranteed (Henrion, 2000), in practical cases, according to the experiences in the group, the minimization converges almost always to an acceptable solution. The actual procedure is iterative and solves one index for one shape and for all components at a time holding the other shapes constant. More explicitly, first all values of F2 and F3 are initialised. Then F1 is calculated, starting from the beginning of F1 to the end, each index is calculated for all components simultaneously. After that F1 has been calculated, F2 is calculated in the same way, and finally F3. The iteration then starts over with F1 again.

In order to make the iteration scheme more robust in respect to convergence in NMR spectra, it is possible to change the equation to minimize from equation 2.6 to

$$\min\left(|S-\sum_{n=1}^{R} a_n F1_n \otimes F2_n \otimes F3_n|^2 + \lambda \sum_{n=1}^{R} a_n^2\right). \qquad (2.7)$$

This is called Tikhonov regularization, and λ the Tikhonov regularization factor (Tikhonov and Samarskij, 1990). A too large value of λ can of course not be justified because then a wrong equation would be minimized. But, also for a rather small value on λ, the added part work as a weighting factor and penalty on components with very high amplitudes. This forces the components to describe actual intensities in the input rather than just compensate large positive values in some components with large negative values in others, which is typical for the "mixing" described above.

TWD can be applied both in time and frequency domain and is over all very general since the only assumptions are the model itself and user provided values for R and λ.

Automated decomposition using PRODECOMP

The method can also be used for spectra with coupled evolution periods and a very similar scheme as in normal TWD, and was implemented in the program PRODECOMP written in MATLAB code (The MathWorks, Inc.). For example three indirect dimensions can be recorded in this way. Omitting the direct dimension, which can be treated with outer products as described above, the relevant equation describing the m:th spectrum (projection) becomes

$$P_m = \sum_{n=1}^{R} F1_n(c_{1m}) * F2_n(c_{2m}) * F3_n(c_{3m}), \quad (2.8)$$

where "*" means convolution, the different c_{im} reflect the effect of a given type of coupling in spectrum m on dimension i, and the equation to minimize is,

$$\min(|\sum_{m=1}^{M}(P_m - \sum_{n=1}^{R} F1_n(c_{1m}) * F2_n(c_{2m}) * F3_n(c_{3m}))|^2). \quad (2.9)$$

One difference compared to the full dimensional version is that more than one spectrum is used instead of only one, another that the spectra are described as convolutions of the shapes rather than outer products. Explicitly this means that the alternating least squares iteration scheme is slightly modified. Still F1 is calculated first, followed by F2 and F3. But instead of calculating one index of these at a time, all values of F1 are calculated simultaneously, and then all values for F2 and at last all values for F3 before starting over with F1 again. Just like for regular spectra a regularisation factor λ can be used also in equation 2.9. In addition to this, the shapes are always assumed to have a specific sign (Bro and DeJong, 1997).

In practice, PRODECOMP is easy to use. The spectroscopist gives as input, the spectra, a matrix describing the particular convolutions in each spectrum in terms of shapes, the number of components, a value λ, which data points in the direct dimension of the spectra to be analysed, and how many iterations the program shall use. In return the program delivers the components with appropriate shapes. As a validation of the result, the original spectra are compared to back calculations from the shapes.

3 RESULTS AND DISCUSSION

NMR spectra are often very complex and time demanding to analyse. Also, the recording time is often substantial. These features of NMR make the method less attractive since the well-established idea that it is of great importance to learn as much as possible about as many proteins as possible to better understand life, develop medical drugs etc. requires high-throughput tools. This thesis describes methods that reduce these problems. Paper I, and a separate example for the defensin HNP2, show that existent computer programs and a new calibration program reliably can handle the complex output of sets of NMR experiments by correctly peak picking and assigning these automatically. Four papers, Paper II, III, IV and V, demonstrate different procedures on how to calculate the true nuclei shifts and even the full NMR-spectrum from spectra recorded using the very time saving method of coupled evolution periods.

3.1 Fully automated analysis of regular NMR-spectra

The peak picking program AUTOPSY (Koradi et al., 1998) and the assignment program GARANT (Bartels et al., 1996, 1997) had already before our study proved to work very nicely. But AUTOPSY had only been rigorously tested for 2D spectra, and GARANT had not been tested on automatically picked peaks. Therefore it was interesting to investigate these programs and the calibration program PICS, combined performance on both 3D and 2D spectra. For the first time a sequential assignment of a protein was made using fully automated tools.

Ideally, the chemical shifts should be the same in all spectra recorded. Unfortunately this is not always the case, for example it is difficult to keep the temperature exactly the same between different experiments since different pulse sequences heat the sample to different extents. Also more everyday reasons, like restricted measurement time on the instrument forcing the spectroscopist to record spectra at different times and possibly on different machines, influences negatively on the possibility to create the same experimental conditions for all spectra. Therefore, prior to the assignment of peaks from a set of various spectra, it is important to make sure that the shifts do not differ between the spectra, or at least to have some idea how

large the deviations might be. Also, it is a good idea to minimize systematic differences of the shifts between spectra. The spectroscopist often does this by arbitrarily looking at a small number of peaks, which can be assumed to origin from the same nuclei in two or more dimensions. By visual inspection the shift differences of these peaks are minimized in the dimensions in common by adding appropriate constants to the scales.

This procedure is of course quite rough and does not guarantee that the overall difference is minimized, nor does it give any precise idea about the remaining, not systematic, differences between the peaks in the different spectra. The program PICS turned out to be useful for doing this step. It is automatic and therefore not very time consuming for the user. It provides a more reliable calibration since all peaks are considered simultaneously, and it presents histograms showing the remaining differences between peaks that have been calibrated. The last feature was very useful when defining accuracies of the shifts in GARANT.

Fully automated resonance assignment of hetero-nuclear protein spectra

Paper I describes the use of AUTOPSY, PICS, and GARANT, used in conjunction to assign a doubly labelled sample of the 128 aa blue cupper protein azurin (Karlsson et al., 1989). The aim of the study was to investigate how well a fully automated peak picking and assignment procedure could perform in terms of correct and wrong assignments.

Peaks from five 3D spectra, HNCACB, HNHA, HCCH-TOCSY, ^{15}N-NOESY-HSQC, and ^{13}C-NOESY-HSQC, and two 2D spectra, ^{15}N-HSQC and ^{13}C-HSQC were picked using AUTOPSY. A more rigorous investigation of the ^{15}N-HSQC- and HNCACB-spectra shows that for reasonable combinations of the required symmetry and minimal size values AUTOPSY does a very good job. For a wide range of input combinations only some true peaks are missed and artefacts picked. Therefore the input values were set uniformly for all spectra except for the NOESY where somewhat smaller peaks were allowed. A quick, visual inspection in the spectra to confirm that the resulting peak lists are all right, i.e. that obvious peaks are picked but not too many artefacts, is sufficient for getting a very good peak list.

In the next step the peak lists were calibrated and filtered by PICS. The ^{15}N-HSQC was chosen as starting point. For the HNCACB, HNHA, and ^{15}N-NOESY-HSQC, the ω_{HN} and ω_N axes were calibrated to the initial spectrum, while preserving the ^{1}H-^{1}H diagonal in the latter two spectra. In order to proceed with the HCCH-TOCSY and ^{13}C-NOESY-HSQC, an artificial peak list with ω_{HC} and ω_C entries was created by combining the HNCACB and HNHA peak lists, allowing calibration of both the ^{1}H and the ^{13}C dimension. Finally the ^{13}C-HSQC was calibrated with respect to ^{13}C-NOESY-HSQC. The histograms after each calibration show distinct distributions around zero for the peak shift differences and were a good aid in the assignment step. After the calibration, the peak lists from the 3D spectra were also filtered from peaks that lacked corresponding peaks in the better-resolved 2D spectra.

The assignment was done by letting GARANT run ten times starting with ten different values as seeds for the random number generator but with the same five filtered 3D peak lists and always 100 in population size. With the histograms from PICS in consideration the accuracy both within and between spectra were set equal to one data point but not less than 0.01 ppm. GARANT was considered to have made an assignment when it assigned the same shift to the same nuclei in at least 6 of the calculations.

The final result for protons shows that 123 of 123 HN-, 134 of 139 HA-, 48 of 57 methyl-, 9 of 21 ring-, and 185 of 288 other (CH, CH_2)-, and 14 of 24 NH_2-protons, were assigned to the correct shift. The corresponding values for incorrect assignments are 0, 3, 4, 5, 16, and 4. The result for the carbon-bound protons in the group "others" is 222 correctly assigned protons out of 288 protons if only the correct residue but not necessarily the correct proton is stipulated for a right answer. This can be justified since only TOCSY- and NOESY-type of spectra were used, lacking the explicit information given in a COSY-type spectrum, which only correlates pairs of protons with three or less covalent bounds in between. Despite the lack of specific spectra for the nuclei in side chain rings the corresponding shifts were still to some extent assigned, but the result would most probably have been better with additional spectra. Five residues with long side chains contain two incorrect assignments, but otherwise no clustering of errors in the 3D structure is observed.

In conclusion, the combined use of the earlier presented programs AUTOPSY and GARANT, together with PICS, works well as a tool for fully automated assignment.

The result from the presented example can be expected to be sufficient for 3D-structure determination. Future improvements may include a better direct connection between GARANT and the data by not only letting the program consider peak shifts in peak lists but also quality parameters, volumes etc. for each peak.

Fully automated resonance assignment of homo-nuclear protein spectra

As a part of a collaboration with the Institute für Physiologische Chemie, Tierarztliche Fakultät, Munich, and the Bijvoet Center for Biomolecular Research, Utrecht University, the fully automated assignment approach using AUTOPSY (Koradi et al., 1998), PICS, and GARANT (Bartels et al., 1996, 1997) was applied on spectra from the 29 aa, non-labelled, defensin Human Neutrophil peptide 2 (HNP2) in the presence of lactose. The assignments were then used for structure determination with CYANA (Güntert et al., 1997).

Defensins are peptides with a characteristic three-cysteine bridges framework, and are antimicrobial effectors of innate immunity. They occur in phagocytes, body fluids as well as in epithelia and contribute to host defence against bacterial, fungal and viral infections. Defensins are particularly abundant and widely distributed in various animal species including humans. A number of defensins bind carbohydrates specifically allowing identification and subsequently destruction of their targets. In order to get an insight into the interactions between defensins and sugars using the docking program HADDOCK (Dominguez et al., 2003), structure determination is a prerequisite.

Two spectra of HNP2 in presence of lactose were used in the assignment, one 2D-TOCSY and one 2D-NOESY. Both spectra were obtained on a 750 MHz Bruker NMR spectrometer. First, the spectra were processed with the NMRPipe software (Delaglio et al., 1995) using twofold zero filling along both dimensions, and then converted to the XEASY (Bartels et al., 1995) format. The size of the final spectra was 640*1902 data points covering the range between -1.33ppm and 10.66ppm in both dimensions.

AUTOPSY peak lists were generated as for azurin. The same macro was used for both spectra and included that peaks should only be picked in regions with a signal at least a factor 1.2 times the local noise level, both the regions and the peaks had to be at least 2*2 data points in size, and the peaks highest point had to be at least 2

times the local noise level. The peaks were expected to have a high symmetry with the symmetry requirement set to 0.15. The resulting peak list for the TOCSY spectrum included 827 peaks and the peak list for the NOESY spectrum 1547 peaks.

PICS first verified that the picked diagonal peaks were distributed evenly along the diagonals and then the TOCSY peak list was calibrated to the NOESY peak list. As for azurin, GARANT was run 10 times with the population size for the genetic algorithm set to 100. The accuracy of peak positions both within and between spectra was set to 0.01ppm. Assignments that were repeated in at least six runs were kept, the rest were considered not trustworthy.

The structure calculation was made in a partially manual way. CYANA was not given the assigned proton list from GARANT and the NOESY peak list from GARANT. Instead, the long distance NOEs in the NOESY peak list were assigned manually using the proton list from GARANT. The final peak list included 402 assigned peaks, which were translated into 277 upper distance constraints. In addition to these, another nine upper and lower constraints were added describing the covalent bonds in the three cysteine bridges. 200 structures were calculated and the best 10 were kept.

Since the NMR investigation of HNP2 was based on a 2D-TOCSY and a 2D-NOESY spectrum only, this made the assignment much more difficult than if spectra from a labelled sample had been available. Despite the lack of scalar couplings connecting the spin systems in different residues GARANT managed to sequentially assign almost all proton shifts. The assignments were then compared with a manual assignment of the spectra based on an earlier manual assignment of HNP2 in complex with another ligand.

Probably due to relaxation effects the peaks of cys1 must be small and were therefore not detected and the shifts of this residue not assigned. GARANT correctly assigned all the 27 other ω_{HN}, and 26 out of 28 ω_{HA}. The two missed shifts, for pro6 and trp25, were too close to the waterline for being detected by AUTOPSY. 39 out of 41 ω_{HB} belonging to residues 2-29 were totally correctly assigned. The assignment of the last two, the ω_{HB} for cys3, are strictly speaking not totally correct since they are assigned to the same frequency at approximately 2.35ppm but it is possible to see slightly different shifts in the spectra, not detected by AUTOPSY which picks one

peak instead of two. In addition to the described shifts GARANT correctly assigned 45 side chain protons. Some assignments in aromatic rings remain a bit unclear, possible due to relaxation effects caused by interaction with the ligand. The ligand also contributed with peaks in the spectrum making it harder to assign.

In total, this was a very encouraging result. Although the peptide is small and its spectra clear some shifts are very close to each other making the assignment difficult (Figure 1). It clearly shows that the combination of both automated peak picking and assignment is fruitful also for non-labelled proteins.

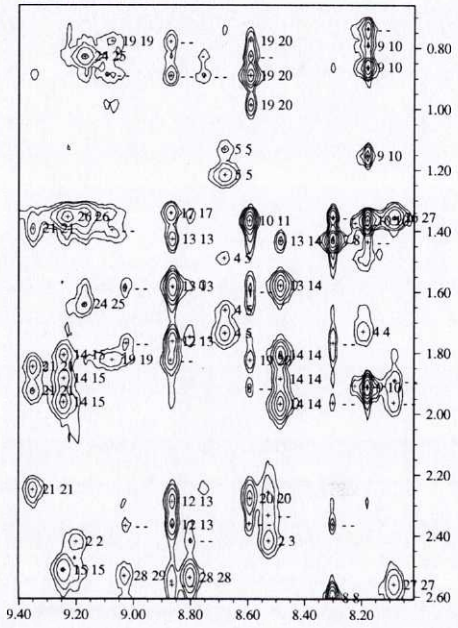

Figure 1. *A small region of the NOESY spectrum for HNP2 is displayed. The units are in ppm. The small crosses show where AUTOPSY has picked peaks. The numbers are GARANT assignments; only the residues and not the particular protons are written out. Despite that the shifts for some nuclei are very similar, e.g. HN11 and HN20 at approximately 8.60ppm and HN13 and HN17 at approximately 8.88ppm, GARANT still manages to assign the peaks correctly.*

The main reason why only approximately a quarter of the NOESY peaks were assigned is that AUTOPSY picks peaks both above and below the diagonal and the manually assigned peaks are only above the diagonal. Also, AUTOPSY was set to pick peaks on and close to the diagonal and most of these peaks were not assigned.

The distance constraints from the NMR experiment together with distance constraints from the covalent bonds were good enough to produce 10 structures with an average of 4 upper limit violations per structure with the cut off set at 0.5Å, and the worst violation being 0.69Å. The van der Waals limit, set at 0.2Å was violated on average once in every structure and the worst result was 0.27Å. The average backbone RMSD to mean is 0.76±0.17Å and the global displacement of the backbone varies between 0.31Å and 1.11Å and is shown in Figure 2. Since the spectra do not allow stereo specific assignments of the side chains the upper limits from the NOESY spectrum are less strict than otherwise possible and therefore the side chain positions in the structures are less precise. The global displacement of the heavy atoms in the side chain varies between 0.39Å and 2.00Å (Figure 2). The peptide interacts with ligands in the region between residue 1, 20, and 25.

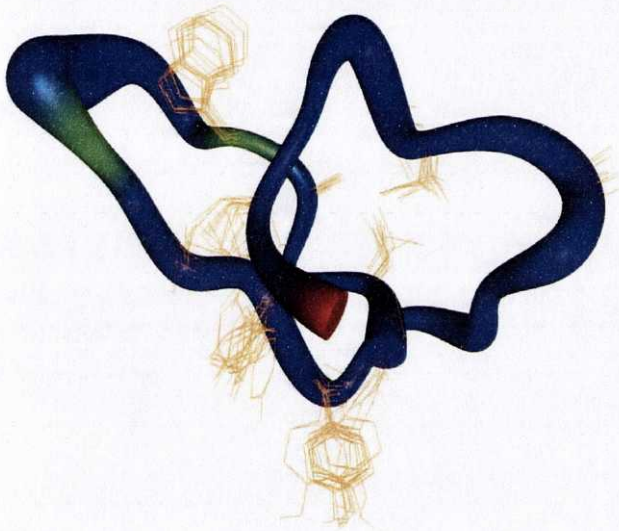

Figure 2. *The structure of HNP2. The average global backbone displacement is shown as a function of the thickness of the "sausage", the most narrow part at residues 26-28 corresponds to approximately 0.4Å and the thickest part at residues 20-21 to approximately 1.1Å. All ten structures for side chains with global displacements smaller than 1.0Å are explicitly drawn. Ligands bind in the region between residues 1, 20, and 25.*

3.2 Evaluation of spectra with coupled evolution periods

In Paper II the program EVOCOUP, and in Paper III and IV the program PRODECOMP, are tested on the same (5,2)D HACACONHN-spectrum of the 128 aa blue cupper protein azurin. From the (5,2)D HACACONHN experiment 12 2D spectra are obtained, with peaks at frequencies corresponding to $\omega_N \pm \omega_{CO} \pm \omega_{CA}$ and $\omega_N \pm \omega_{CO} \pm \omega_{CA} \pm \omega_{HA}$ in the indirect dimension, and ω_{HN} in the direct dimension. In addition to the HACACONHN-spectrum also an N15HSQC-spectrum was used in the test of EVOCOUP. The common goal for the both approaches was to obtain a five dimensional peak list representing the N-, CO-, CA-, HA- and HN-shifts of the spin systems in the experiment.

Analysis of the HACACONHN spectrum of azurin using the program EVOCOUP

Since EVOCOUP relies on peak lists these had to be picked first. Then, EVOCOUP analysed these and produced 5D peak lists of the spin systems, which were compared with the known shifts of azurin.

AUTOPSY was used to pick peaks in the twelve HACACONHN-spectra and the N15HSQC spectrum. EVOCOUP used the resulting peak lists and scanned the spectra with intervals of 0.02 ppm in steps of 0.01 ppm, with the internal consistency criterion, maxdiff, set to ~18.25Hz in the indirect dimension, which for example is equal to 0.3 ppm when regarding the ^{15}N-shifts. The lowest acceptable number of peaks for consistently defining a peak position was set to eight.

EVOCOUP successfully assigned all but four of the 123 backbone spin systems. No false positives were found. Probably due to the rather unusual ω_{HN} shift of Asn38, ω_{HN} = 11.36 ppm, its HACACONHN-peaks were very small and not picked by AUTOPSY. The remaining three spin systems, Phe15, Val31, and His35, all occur in crowded regions around ω_{0HN} = 8.70 and ω_{HN} = 8.95 ppm. Peaks of these spin systems overlap in the indirect dimension in some spectra with peaks from other spin systems with very similar ω_{HN} shifts. The three spin systems did not have eight unique peaks, and peaks they shared with other spin systems were assigned to these other systems, therefore the former were missed.

For five of the 11 glycines only one ω_{HA} was detected. The reason for this was the same as for the three spin systems mentioned above. In the N15HSQC and the $\omega_N \pm \omega_{CO} \pm \omega_{CA}$ spectra one of the two glycine peaks in the spin system is assigned first, and therefore, when not all 8 peaks are available in the remaining 8 spectra, one glycine peak will remain unassigned.

The examples above show that EVOCOUP has a problem when dealing with overlap due to that it only uses picked peaks once. One way to solve this could be to allow for multiple assignments of peaks. Unfortunately this is a less attractive solution for two reasons. The major one is that in a crowded region with many spin systems the calculations would take a very long time. The second is that the exclusion of peaks makes successive assignments with smaller numbers of supporting peaks more reliable. If the peaks are not removed sometimes new but false assignments are possible by replacing only a few peaks in an earlier correct assignment with an even lower number of peaks from other spin systems.

Benefits and disadvantages of frequency domain decomposition of sparse recorded spectra

The time domain spectrum of an experiment with coupled evolution periods can be thought of as an extremely sparse spectrum where only one or more hyper complex diagonals of various angels are recorded. In principle, this is sufficient for processing the spectrum using the sparse version of TWD (Ibraghimov, 2002) since data points are recorded in all rows for all dimensions. But this cannot be easily done because of two reasons. If only a few sets of diagonals are available, as for example in the (5,2)D HACACONHN experiment, each time increment in the indirect dimension lacks any connection to the others. Decomposition of a spectrum with two or more components having identical shapes in the direct dimension is therefore not possible. The explanation is that the components with identical shapes in the direct dimension, can interchange all data in the shapes in the indirect dimension for any recorded time increment without influencing the equation to be minimised but totally corrupting the result. The other reason is, with equation 2.4 in mind, that the in some sense separate experiments for different time increments in the indirect dimension are too small for determination of more than, at the very best, three components.

The benefit of making the decomposition in frequency domain is that it is possible to assume a specific sign of the shapes (Bro and DeJong, 1997). This is a very strong requirement and by doing this, the first problem is solved; the use of frequency domain data with the assumption that shapes are of a specific sign "connects the independent measurements in time domain together". This advantage is also its disadvantage. The cost of the transformation to frequency domain is that the calculations take more time since not only one index for all shapes in one dimension is calculated in each step but all indices for all shapes in one dimension. The sparse version of TWD would, if applicable, be very fast for spectra with coupled evolutions. The other problem that the number of independent data points must be sufficient for determination of correct shapes remains and is solved only by using more input data. Problems of this kind are described below.

Analysis of the HACACONHN spectrum of azurin using the program PRODECOMP

PRODECOMP works immediately on spectra and does not require prior peak picking. Instead peak picking was performed on the output of PRODECOMP and the shifts of these peaks were compared to the known shift values of azurin.

Since only spectra with peak positions at $\omega_N \pm \omega_{CO} \pm \omega_{CA}$ and $\omega_N \pm \omega_{CO} \pm \omega_{CA} \pm \omega_{0HA}$ in the indirect dimension were used it was not possible to choose ω_N, ω_{CO}, ω_{CA}, and ω_{HA} as shape definitions. The reason is that spectra at angles $\pm\alpha$ only, cannot provide unique lineshape information on the basic axes unless the peaks are assumed not to be folded. Instead, $\omega_N+\omega_{CO}+\omega_{CA}$, $2\omega_{CO}$, $2\omega_{CA}$, and ω_{HA} were used. In practice, this means that all line shape information for N, CO, and CA will end up in the $\omega_N+\omega_{CO}+\omega_{CA}$-shape and the $2\omega_{CO}$-, and $2\omega_{CA}$-shapes will describe how much the first shape moves between different spectra. In theory this implies that the other two shapes will be Dirac-like functions with shifts corresponding to twice the chemical shifts of CO, and CA, i.e. that half their spectral widths are lost due to the lack of basic spectra.

The calculation of ω_N, which was not performed explicitly by PRODECOMP for the two papers, but after peak identification using Lorentzian fits in the shapes and

subtraction of half the peak shift values in $2\omega_{CO}$ and $2\omega_{CA}$ from $\omega_N+\omega_{CO}+\omega_{CA}$, also suffers the same loss of spectral width.

Lorentzian fits were done for all shapes and the results compared to known chemical shift values. PRODECOMP successfully assigned all but five of the 123 backbone spin systems. No false positives were found. The remaining spin systems, in residues Asp11, Val31, Ala53, Asp69, and Phe97, occur at 8.4ppm < ω_{HN} < 9.0ppm where the signal density is highest. In addition, the spin systems of four of the eleven glycines were not complete.

One interesting effect of PRODECOMP's very general description of the shapes is that peaks that are folded in the input spectra do not cause a problem. This was actually the case for some glycines.

A comparison of the results of EVOCOUP and PRODECOMP

The straightforward peak information in the full dimensional spectrum is lost and the intensity of one peak is shared between peaks in many spectra when using experiments with coupled evolution periods. Conceptually, EVOCOUP and PRODECOMP deal with this in two different ways.

EVOCOUP picks the peaks in the spectra and deals with them in the same way as peak lists from different normal, full dimensional, spectra are dealt with, i.e. from chemical shifts. The difference is that the peak shifts do not immediately give the shifts of the individual nuclei making the assignment technically different. The advantage and disadvantage of the pick picking approach are also the same as for peak picking and consecutive assignment in sets of full dimensional spectra. Peak picking is a very definite procedure, either a peak is picked or it is not and the decision is made on only a part of the total information content available. When peaks are correctly identified, peak picking is a very effective filter that makes further processing much easier since less data has to be handled. The problem arises when true peaks are not picked or artefacts are picked. The strong non-linearity of the peak picking process hinders later steps to recover excluded information. Also, the reliability of the peak lists is easily over exaggerated in later steps giving picked artefacts undeserved support. Reflecting the uncertainty in the peak picking with

quality parameters etc. and keeping not only the peak shifts but also the peak shapes in the later processing steps reduce these problems. But the principal problem remains as long as all data is not considered simultaneously.

PRODECOMP deals with the problems of peak picking simply by skipping it. Instead, the peaks are picked in the output from the computation. The obvious problem of this method is how to correctly calculate the large number of data required to describe the full dimensional spectrum and not only peak shifts. The answer is that the complexity of the full dimensional NMR spectrum is not comparable to its size. Still, the method will present a good result only if a sufficient number of spectra are used.

In practice, neither of the two programs managed to extract all the spin systems from the spectra. But they did not deliver false positives either which would have been worse. The investigations show that both approaches are reliable and that the results can be trusted. Noteworthy is that the missed spin systems, except for Val31, were not the same in the two approaches showing that the information is there. It is a matter of getting it.

A major reason for that spin systems were missed is most likely the very small number of spectra. In retrospect, the approximately 14 hours of measuring time should have been spent differently. The signal to noise is good and AUTOPSY would definitely be able to pick peaks with half the size compared to now except for the Asn38, which it did not find anyway. Therefore the recording time prior to the use of EVOCOUP could easily have been decreased to a quarter of the time and probably even more compared to now. The other three quarters could have been used for other spectra.

Even more could have been done to optimise the input to PRODECOMP. To begin with PRODECOMP does not require as large sweep widths as EVOCOUP since PRODECOMP manages folding. PRODECOMP also, because of its ability to treat all input data simultaneously, handles signal to noise problems much better than AUTOPSY. Already with twelve spectra this is obvious for Asn38. The larger the number of spectra, the larger the advantage using PRODECOMP instead of AUTOPSY due to that both PRODECOMP and AUTOPSY for each added spectrum gets more data but only AUTOPSY is requested to deliver more data. Therefore the signal to noise could have been decreased more for PRODECOMP than AUTOPSY allowing even more spectra or simply a shorter recording time.

Correlation of 13 nuclei in one component using PRODECOMP

In Paper V, a slightly modified PRODECOMP is tested on 24 different spectra from four different experiments, (5,2) HBHACONH, (4,2) CBCACONH, (5,2) HBHACANH and (4,2) HBHANHGP, of the 76-residue protein ubiquitin, providing components that each describes spin systems with up to 13 nuclei: 2HB(i-1), CB(i-1), HA(i-1), CA (i-1), CO(i-1), N, HN, CA, HA, CB, 2HBs. As in the previous example, the direct dimension holds the ω_{HN}. The indirect dimensions correspond to $\omega_N \pm \omega_{CO}(i-1) \pm \omega_{Cx}(i-1) \pm \omega_{Hx}(i-1)$, $\omega_N \pm \omega_{CO}(i-1) \pm \omega_{CX}(i-1)$, $\omega_N \pm \omega_{CX} \pm \omega_{HX} \pm \omega_{CA}$ and $\omega_N \pm \omega_{CX} \pm \omega_{HX}$, where X stands for either A or B. In theory, this set of spectra is sufficient for backbone assignment.

For optimal use of the data PRODECOMP was modified to handle different number of shapes in different dimensions but for the same component. The advantage with this in the particular case was that otherwise two components would have been needed to describe all the shapes, and all but the 2HB(i-1), CB(i-1), HA(i-1) and CA (i-1) shapes would be present in two different components, since the total system of the 13 nuclei can be described as a branched system where the 2HB(i-1) and the CB(i-1) nuclei is one branch and the HA(i-1) and the CA(i-1) nuclei another, and the CO, N, and HN nuclei are the trunk. A similar branching is found in residue i. For the same reason as in the example with azurin, shapes with shift combinations, in this case N+CO(i-1)+Cx(i-1), N+Cα+Hα and N+Cβ+Hβ, were required, and half the sweep width was lost in the output here as well. This would have been avoided if ^{15}N-HSQC, (3,2)D HNCO, etc. had been used as well which is to recommend.

The very large number of shapes in each component make sequential assignment of these robust since up to five different nuclei connect them. One way to do this is of course to pick the peaks in the shapes followed by normal assignment. A more challenging idea would be to first decompose all spectra into components and from direct comparison of the shapes, by direct products or similar, sequentially assign the components rather than the peak shifts. The shifts could then be determined in a later step or possibly by defining the shapes as probability scores for different shift values and compare the shapes with random coil shifts etc. during the assignment.

4 CONCLUSION

This thesis includes one part where two software packages, the peak picking program AUTOPSY (Koradi et al., 1998) and the resonance assignment program GARANT (Bartels et al., 1996, 1997), are tested in conjunction with the in house developed calibration program PICS. An example for the 128 aa ^{15}N- and ^{13}C-labelled protein azurin shows that the method successfully assigns almost all nuclei the correct shifts and therefore is applicable and useful for proteins of this size. Another example demonstrates for a 29 aa non-labelled defensin (HNP2) that the three programs delivers an assignment good enough for structure calculation and that labelling therefore is not a strict requirement.

The other part treats evaluation aspects of data from spectra with coupled evolution periods. Using two different in house developed programs, EVOCOUP and PRODECOMP, it is shown that it is possible to obtain almost complete and totally accurate peak lists from these experiments, either by picking the peaks in the lower dimensional spectra followed by appropriate analysis of the peaks, or by reconstruction of a decomposition of the full spectrum and pick the peaks in the decomposition. The second approach is also shown to work for sets of spectra that can be used for sequential assignment.

ACKNOWLEDGEMENTS

To:

Martin, my supervisor, the man who is the definition of knowledge and professionalism combined with patience and kindness, always willing and able to help,
Cissi and **Aleks**, my former PhD-student colleagues in the group, who made weekdays funnier than weekends,
Hanna, **Caroline**, **Jenny**, **Ulrika** and **Anders**, the rest of the gang soon to become doctors and who's moral cannot be questioned, -Hänsyn och ansvar!,
Lars N, **Bruno**, **Ann-Cathrine**, **Britt** and **Rigmor**, who make everything but me work,
Vladislav and **Göran**, the Swedish NMR center gurus, who give you good advice, good courses, good food, and good company on regular basis,
Tineke and **Charlotta**, who has documented teaching skills in NMR for dummies,
Therese, **Karin** and **Helena**, who could stand having me as a room mate for a while, although not for as long as Cissi,
Jonas and **Daniel**, who have the future ahead,
Mamma, **Pappa**, **Hanna**, **Matilda** and **Olof**, who always give me the best support, when needed, you are there,
ALL OTHERS, who make my life enjoyable in general,
and
Lisa, who makes my life fabulous and special.

Thank You!

REFERENCES

Bartels, C., Xia, T.H., Billeter, M., Güntert, P., Wüthrich, K.,The program Xeasy for computer-supported NMR spectral-analysis of biological macromolecules,

Bartels, C., Billeter, M., Güntert, P., Wüthrich, K., Automated sequence-specific NMR assignment of homologous proteins using the program GARANT, J. Biomol. NMR 7 (1996) 207-213.

Bartels, C., Güntert, P., Billeter, M., Wüthrich, K., GARANT - A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra, J. Comput. Chem. 18 (1997) 139.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig H., Shindyalov, I.N.,Bourne, P.E., The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235.

Billeter, M., Orekhov, V.Y., Three-way decomposition and nuclear magnetic resonance. In: Sloot, P., Abramson, D., Bogdanov, A., Dongarra, J., Zomaya, A., Gorbachev, Y. (Eds.), Computational Science, Springer, St. Petersburg, 2003, 15.

Bodenhausen, G., Ernst, R.R., Direct determination of rate constants of slow dynamic processes by two-dimensional "accordion" spectroscopy in nuclear magnetic resonance, J. Am. Chem. Soc. 104 (1982) 1304.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res., 31 (2003) 365.

Bro, R., PARAFAC. Tutorial and applications., Chemometrics Intel. Lab. Sys. 38 (1997) 149.

Bro, R., DeJong, S., A fast non-negativity-constrained least squares algorithm, J. Chemometrics 11 (1997) 393.

Carroll, J.D., Chang, J.J., Analysis of individual differences in multidimensional scaling via N-way generalization of "Eckert-Young" decomposition. Psychometrika. 35 (1970) 283.

Cavanagh, J., Fairbrother, W.J., Palmer III, A.G., Skelton, N.J., Protein NMR spectroscopy: principals and practice, Academic press, San Diego, 1996.

Coggins, B.E., Venters, R.A., Zhou, P., Filtered backprojection for the reconstruction of a high-resolution (4,2)D CH3-NH NOESY spectrum on a 29 kDa protein, J. Am. Chem. Soc. 127 (2005) 11562.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes, J. Biomol. NMR 6 (1995) 277.

Dominguez, C., Boelens, R., Bonvin, A.M., HADDOCK: a protein-protein docking approach based on biochemical or biophysical information, J. Am. Chem. Soc. 125 (2003) 1731.

Drenth, J., Principles of protein X-ray crystallography, Springer, New York, 1994.

Güntert, P., Mumenthaler, C., Wüthrich, K., Torsion angle dynamics for NMR structure calculation with the new program DYANA 273 (1997) J. Mol. Biol. 283.

Henrion, R., On global, local and stationary solutions in three-way data analysis, J. Chemometrics 14 (2000) 261.

Hiller, S., Fiorito, F., Wüthrich, K., Wider, W., Automated projection spectroscopy (APSY), Proc. Natl. Acad. Sci. USA 102 (2005) 10876.

Hoch, J.C. , Stern, A.S., Maximum entropy reconstruction, spectrum analysis and deconvolution in multidimensional nuclear magnetic resonance, Methods in Enzymology, Pt A, vol. 338, Academic Press, San Diego, 2001, 159.

Ibraghimov, I., Application of three-way decomposition for matrix compression, Numer. Linear Algebr. Appl. 9 (2002) 551.

Karlsson, B.G., Pascher, T., Nordling, M., Arvidsson, R.H., Lundberg, L.G., Expression of the blue copper protein azurin from Pseudomonas aeruginosa in Escherichia coli, FEBS Lett. 246 (1989) 211.

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

Koradi, R., Billeter, M., Engeli, M., Güntert, P., Wüthrich, K., Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY, J. Magn. Reson. 135 (1998) 288.

Kupce, E., Freeman, R., Reconstruction of the three-dimensional NMR spectrum of a protein from a set of plane projections, J. Biomol. NMR 27 (2003) 383.

Kupce, E., Freeman, R., Fast multidimensional NMR: radial sampling of evolution space, J. Magn. Reson., 173 (2005) 317.

Levitt, M.H., Spin dynamics: basics of nuclear magnetic resonance, Wiley, Chichester, 2001.

Malmodin, D., Billeter, M., High-throughput analysis of protein NMR spectra, Prog. Nucl. Magn. Reson. Spectrosc. 46 (2005) 109.

Mandelshtam, V.A., FDM: the filter diagonalization method for data processing in NMR experiments, Prog. Nucl. Magn. Reson. Spectrosc. 38 (2001) 159.

Moseley, H.N.B., Riaz, N., Aramini, J.M., Szyperski, T., Montelione, G.T., A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra, J. Magn. Reson. 170 (2004) 263.

Orekhov, V.Y., Ibraghimov, I.V., Billeter, M., MUNIN: A new approach to multi-dimensional NMR spectra interpretation., J. Biomol. NMR 20 (2001) 49.

Smilde, A., Bro, R., Geladi, P., Multi-way analysis, Wiley, Chichester, 2004.

Snyder, D.A., Chen, Y., Denissova, N.G., Acton, T., Aramini, J.M., Ciano, M., Karlin, R., Liu, J.F., Manor, P., Rajan, P.A., Rossi, P., Swapna, G.V.T., Xiao, R., Rost, B., Hunt, J., Montelione, GT, Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination, J. Am. Chem. Soc. 127 (2005) 16505.

Stern, A.S., Li, K.B., Hosh, J.C., Modern spectrum analysis in multidimensional NMR spectroscopy: Comparison of linear-prediction extrapolation and maximum-entropy reconstruction, J. Am. Chem. Soc. 124 (2002) 1982.

Tikhonov, A.N., Samarskij, A.A., Equations of mathematical physics, Dover, New York, 1990.

På grund av upphovsrättsliga skäl kan vissa ingående delarbeten ej publiceras här.
För en fullständig lista av ingående delarbeten, se avhandlingens början.

Due to copyright law limitations, certain papers may not be published here.
For a complete list of papers, see the beginning of the dissertation.

GÖTEBORGS UNIVERSITET

GÖTEBORG
UNIVERSITY

Faculty of Science

ISBN 13: 978-91-628-6750-8
ISBN 10: 91-628-6750-4