Thesis for the degree of doctor of Philosophy

Algorithms for analysis of NMR projections: Design, implementation and applications

Jonas Fredriksson

Department of Chemistry
University of Gothenburg
Sweden

2011

UNIVERSITY OF GOTHENBURG

Department of Chemistry
Göteborg University
SE-405 30 Göteborg

ISBN 978-91-628-8277-8

Printed by Chalmers Reproservice
Göteborg 2011
Sweden

Abstract

With an increasing rate of protein expressions the need for fast protein characterization has become more important. Protein NMR has long been an important contributor for protein characterization; being one of a few techniques that can study proteins at atomic resolution in their native state. Whitin recent years faster experimental and processing methods have emerged that are now becoming routine. This thesis describes algorithms for automatic backbone assignment and validation of structure information by using projection experiments together with a decomposition method. Projection experiments reduce measurement time for multidimensional spectra thus making it possible to obtain very high dimensional spectral information in a fraction of the time required for a conventional experiment. By combining different experiments backbone, side chain and NOE information can be obtained. A set of software tools for automatic backbone characterization where developed from the implementation of different algorithms in conjunction with different proteins and projection experiments. Testing and refinement of the different tools resulted in a robust characterization method well suited for different proteins. Possible future projects are expanding the methods to side chain and structure determination making the characterization more complete.

KEYWORDS: NMR, projection experiments, decomposition, algorithm, automatic assignment, proteins, NOESY, reduced dimensionality, peak picking.

List of publications

1. Assignment of protein NMR spectra based on projections, multi-way decomposition and a fast correlation approach. D.K. Staykova, J. Fredriksson, W. Bermel, M. Billeter, J Biomol NMR. 42 (2008) 87-97.
2. PRODECOMPv3: decompositions of NMR projections for protein backbone and side-chain assignments and structural studies. D.K. Staykova, J. Fredriksson, M. Billeter, Bioinformatics. 24 (2008) 2258-2259.
3. Multi-way Decomposition of Projected Spectra obtained in Protein NMR M. Billeter, D.K. Staykova, J. Fredriksson, W. Bermel, Proc. Appl. Math. Mech., 7 (2007) 1110103-1110104
4. Parameter Estimation of Multidimensional NMR Signals Based on High-Resolution Subband Analysis of 2D NMR Projections, I.Y.H. Gu, M. Billeter, R. Sharafy, V.A. Sorkhabi, J. Fredriksson, D.K. Staykova, IEEE International Conf. Acoustics, Speech and Signal Processing, (ICASSP 2009), 497-500
5. Automated Protein Backbone Assignment using the Projection-Decomposition Approach. J. Fredriksson, W. Bermel, D.K. Staykova, M. Billeter, Manuscript
6. Structural characterisation of a histone domain via projection-decomposition, J. Fredriksson, W. Bermel, M. Billeter, Manuscript

Contribution report

Paper 1: Translation from the original fast-nnls matlab code and efficiency tests, implementation of the first projection experiments at The Swedish NMR Center

Paper 2: Algorithm improvements: speed and robustness

Paper 3: Contribution to the mathematical problem presentation

Paper 4: Transformation of the original NMR spectra into a suitable form for the ESPRIT algorithm. Implementing this algorithm in python for further analysis.

Paper 5: Implementing more experiments for protein characterization (assignment and structure), running of most of the experiments, development of new assignments algorithms.

Paper 6: Project design and complete analysis of the data

Abbreviations

fast-nnls	fast- non-negative least square
nD	n dimensional
PRODECOMP	Projection Decomposition
SHABBA	Shape Backbone Analysis
PDB	Protein Data Bank
TOCSY	total correlation spectroscopy
NOESY	nuclear Overhauser enhancement spectroscopy
NOE	nuclear Overhauser enhancement
NMR	Nuclear Magnetic Resonance
GFT	G-matrix Fourier Transform
HSQC	heteronuclear single quantum coherence

Table of contents

Introduction

NMR is a versatile method for protein characterization[1,2]. With an arsenal of various experimental methods it is possible to explore different properties of a protein, and with disordered proteins NMR is sometimes the only possibility. There is a wide array of different NMR experiments that focus on different parts of protein properties. Two of the most important properties is structure determination and drug discovery, and protein research depends heavily on structure information about proteins[3]. Three methods exists for protein structure determination, X-ray crystallography, NMR and Electron Microscopy. Of these three methods is X-ray crystallography the dominant method which represent 86.9% of the structures deposited in the PDB[4]. Structures deposited from NMR experiments in PDB stands for 12.4% and electron microscopy for less than 1%. Prerequisites to get structural information from X-ray crystallography are obtaining crystals that diffract at high resolution, which can be a challenge in several cases. Further, membrane proteins pose great challenge for crystallization and the crystallized protein may not be in their native state[5]. Electron microscopy does not require crystals but suffers from low atomic resolution making it more suitable for obtaining larger overall structure information of different biological species. Solution NMR on the other hand, provides an excellent way to obtain not only structure at atomic resolution, but also dynamics of proteins in solution, which can be used to study ligand interaction and kinetics behavior to name a few examples. However, applications of experimental NMR methods for protein structure determination are limited by protein size and spectral dispersion/ resolution although continuous development is done to extend the maximum size of measurable proteins[6]. Processing multidimensional NMR experiments can be very time consuming and expensive ^{13}C and ^{15}N isotopes for protein labeling are also required to obtain individual assignments. With the advent of high-throughput methods, for bacterial over expression of proteins and cell free expression systems, large scale production of labeled proteins at a shorter time period have been enabled and in conjunction with this have high throughput

methods been developed[7]. Still the need for rapid protein characterization has become urgent[8].

Protein NMR

For characterization of proteins with NMR, a series of multidimensional spectra are recorded to obtain assignments of different spin systems. These assignments form the basis for further analysis like structure calculation and dynamic studies. 2 dimensional experiments are divided into 4 parts: (i) preparation of the sample where all spins are returned to their equilibrium state (ii) evolution where chemical shifts are encoded (iii) mixing time, where magnetization is transferred from one spin to another and (iv) detection of the final FID. This is extended to higher dimensional experiments by adding more evolution and mixing time steps. Magnetization is transferred through chemical bonds by scalar J couplings over one or more bonds or by dipolar couplings through space. During the evolution period of the indirect dimension, the evolution time t_1 is increased with Δt steps altogether sampled with m points. Increased number of indirect dimensions also increases the number of m points that have to be recorded for every indirect dimension. This gives a measurement of $2^{N-1}*m^{N-1}$ complex points for a N-dimensional experiment[9]. Increasing dimensionality in experiment can solve some of the overlapping problems that exists for larger proteins but longer experiment times put an upper limit for higher dimensionality experiments. In traditional multidimensional experiments, evolution periods are varied by a time delay. This time delay is increased by Δt steps and varied independently for every added dimension[10]. This creates long experimental time for higher dimensional experiments and puts a practical limit on the number of dimensions that can be recorded. Multidimensional experiments are required for almost all proteins due to the high overlap of proton peaks in a 1D spectra. By increasing the dimensionality of the experiments, the resonance frequencies of ^{1}H, ^{13}C and ^{15}N can be separately determined. For unlabeled protein samples 2 dimensional experiments are recorded to obtain individual protein assignments. For ^{13}C and ^{15}N labeled samples, 3-dimensional experiments are routinely used for resolving

spectral overlap[11]. However for larger proteins and proteins with severe overlap (such as sequence repetition, molten globule and partially unfolded) it is essential to increase the resolution further or add different experiments to obtain more complete assignments and thus prolong measurement time. Examples of multidimensional experiments are HNCA, HN(CO)CA, HNCO, HN(CA)CO, HN(CA) HA and HCACO all used for obtaining assignments of backbone residues[12] and HCCH-TOCSY to obtain assignments of side chains. NOESY experiments are used for obtaining structure information and together with backbone and side chain assignments it can be used for a structure determination. Higher dimensional experiments are usually built from lower dimensionality experiments by adding a magnetization path for the added nuclei. In projection experiments additional magnetization paths are added to an existing experiment and linear dependencies are set between selected evolution periods creating high dimensional experiments that uses a fraction of the time taken to measure the original experiment. An example of this is the HBHACBCACONH experiment where a HAHB magnetization path are added to the 4D experiment. This was then used as a projection experiment in one of the two backbone experiments used in this study. The time taken for measurement of protein depend on the number of indirect points measured and the number of scans for increasing signal to noise ratio and the duration of one scan. With an absolute lower theoretical threshold for signal to noise of one scan for signal detection and a duration time of one second per scan, a 2D experiment with 60 complex points would take 120 seconds to acquire i.e $2^{N-1}*m^{N-1}$ where m is the number of complex points in the indirect dimension, N is the dimension of the experiment and 2^{N-1} is for quadrature detection. Therefore a 5D experiment with 30 complex points would take $16*30^4$ seconds which correspond to 5 month and an experiment with N>5D would take several years which is not practical. If the number of points in the indirect dimension is increased, time required to collect data for higher dimensional experiment increases even more dramatically[13]. This creates a conflict between the need for fast experimental time on one hand and better resolution on the other hand. Different methods have been developed for overcoming this problems as outlined below[14,15,16].

Fast NMR

Fast NMR refers here to NMR techniques that significantly reduces measurement time in protein NMR experiments[17,18,19]. Several different experimental and processing methods have been developed to reduce measurement time. Examples these are non-uniform data sampling[20,21], single scan spectroscopy[22,23], HIFI NMR[24], projection reconstruction[25], Hadamard spectroscopy[26], GFT[27,28], Filter Diagonalization Method[29,30], APSY[31], maximum entropy[32] and multiway decomposition. There has also been improvement in hardware to decrease measurement time[33]. Non uniform sampling is a method where the number of points sampled in time domain are much less than with uniform sampling thus reducing measurement time considerably. This is a somewhat general term and includes nonlinear sampling as well as projection experiments. Non linear sampling is a method that records a small optimally selected fraction of the experimental data points. The data is then used for reconstructing the spectra. There exists different sampling schemes but sampling only a fraction of the points substantially decreased measurement time. An iterative procedure is used to increase the number of points until the reconstruction of the spectra is the same as the original[34]. The resulting spectra can then be peak picked[35]. Random sampling[36] are also used in time domain data acquisition and processed with multidimensional Fourier transform. These data are used in an iterative algorithm for artifact suppression. Peak picking are then done with statistical methods[37]. In single scan spectroscopy the indirect time variable is replaced by spatial encoding of the spin interactions using gradient pulses. The gradient pulses creates different excitation in different slices of the sample. This gives different evolution times in the sample that can be detected with a single scan in the 2D case. The 2D data set can then be reconstructed[38]. The HIFI NMR method uses two measured orthogonal 2D planes as starting planes and then measures tilted angles of planes adaptively until the model dose not improve. Peak picking is done using a statistical algorithm on the planes avoiding reconstruction of the 3D spectra. Maximum entropy is a reconstruction tool that can transform non uniform

sampled data without loosing too much information. Hadamard spectroscopy tries to record only narrow frequency intervals instead of the whole spectral width. This can then be used for several regions and then get the same information as in the full spectra, at least for smaller proteins and a decrease of measurement time is also gained. Filter diagonalization is a method that analysis time domain signals and give frequencies, amplitudes and line width, making it a suitable replacement for Fourier Transformation. Projection reconstruction techniques uses projection angles when recording spectra instead of recording the whole time domain grid thus reducing the dimensionality of the experiment. These can then be analyzed in different ways. In APSY several projections are recorded and peak picked iteratively using combinatorial procedures. Another approach is to make decompositions of the projections and make peak picking on the resulting shapes, thus avoiding peak picking in the projections. Finally, GFT is a method used in conjunction with reduced dimensionality spectra and was one of the first methods in projection NMR. Reduced dimensionality is achieved by coupling evolution steps in the indirect dimension together and making them dependent instead of independent. Frequencies in the indirect dimension are then not consisting of one nucleus but instead of a linear combination of these. By multiplying time domain data with a G-matrix and then Fourier Transform the result is a number of lower dimensional spectra that contains different linear combinations of nucleus in the indirect dimension. These are often redundant in information and are used to determine the different frequencies of the nuclei in the indirect dimension.

Presentation of the thesis

The following thesis will describe methods developed in this project and applications to a selected number of proteins. For completeness the following description covers all algorithms relevant to this project and therefore contributions from Daniel Malmodin, Wolfgang Bermel (BRUKER company) Doroteya Staykova are in part included.

Methods

Reduced dimensionality experiments are usually derived from traditional NMR experiments[39,40]. In traditional experiments incremental time steps in the independent dimensions are varied independently. In reduced dimensionality experiment the evolution periods in two or more dimensions are sampled jointly. This is achieved by using a linear dependency between selected evolution periods expressed as a ratio between two delays. This ratio between fixed evolution periods determines the projection angle which can be set from -90 to 90 degree angles[41]. In figure 1 is a projection shown in the shaded plane. The blue peak at position ω_1, ω_2, ω_{HN} is projected 45 degrees to both ω1 and ω2. This gives a frequency of $\omega=\omega_1+\omega_2$ in the indirect dimension with a projection coordinate of (ω,ω_{NH}) in the projection plane.

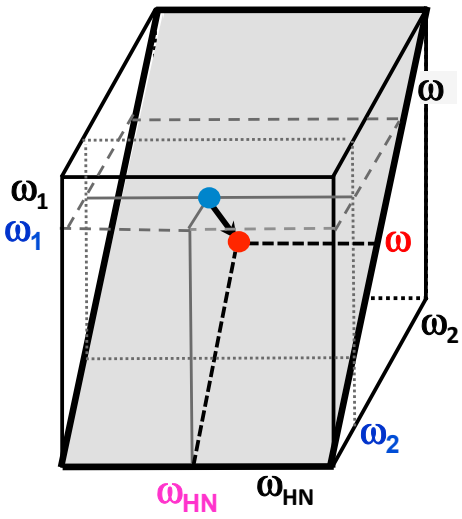

Figure 1. *A projection of 45 degrees gives a linear projection of* $\omega=\omega_1+\omega_2$ *in the 2D plane.*

Depending on the experiment, projection angles used in this study where either 0, ±45 or 90 degrees. 0 or 90 degrees correspond to a 2D projection with one nuclei in the indirect dimension while ±45 degrees projections gives linear combinations between two or more nucleus in the indirect dimension with either positive or negative combinations. Coupling of the different evolution periods reduces measurement time drastically for multidimensional experiments. Measurement time for a corresponding 3D experiment with 100 complex points in the indirect dimension would take approximately 11 h assuming 1 second for every scan. With a projection from 3D to 2D keeping a minimal of 4 planes, ω_1, ω_2, $\omega_1+\omega_2$ and $\omega_1-\omega_2$, would take 13 minutes. This time saving becomes even more enhanced for projected 4D and 5D experiments. The output from these experiments are 2D projection planes, where one peak corresponds to either a single nucleus in indirect dimension or several different nucleus expressed as different linear combinations. The number of planes

recorded depends of the number of indirect dimensions: 13 planes for a 4D experiment and 40 planes for a 5D experiment. All planes are not necessary for the analysis, planes that provide additional information but not unique information can be omitted to save additional measurement time and computational time.

Projection experiments

DIfferent types of projection experiments where developed from conventional higher dimensional protein experiments. All pulses was developed in collaboration with Wolfgang Bermel and was tested and developed on different spectrometers at BRUKER and at the Swedish NMR center. The projection experiments can be grouped into three categories: backbone, TOCSY and NOESY types where various 4D or 5D magnetization paths exists within every group. For backbone characterization mainly two projection experiments have been used in this study based on the following conventional experiments: HAHBCACBCONNH[42] and HAHBCACBNNH[43]. These are referred as backbone experiments. For the first experiment, magnetization transfer path is from residue i-1, while the second experiment transfer magnetization from residue i. The first experiment corresponds to a 5D and the second to a 4D. They complement each other giving frequencies from both the previous residue i-1 and the current residue i. Common nuclei for both residues are N and NH as shown in figure 2. The magnetization path of the two backbone experiments are marked with green and brown. Also shown in the figure are two NOESY experiments, ^{13}C-HSQC-NOESY-^{15}N-HSQC, and ^{15}N-HSQC-NOESY-^{15}N-HSQC, marked by red and orange dotted lines. Backbone magnetization from i-1 starts at the Hα/β nuclei on the previous residue i-1. Then it's transferred via coupling constants to Cα/β nuclei and CO nuclei. Nitrogen is the last nuclei in the indirect dimension and detection is done on the amid proton. The other backbone experiment transfer magnetization from Hα/β on residue i over Cα/β to N and with a final detection on the amid proton as shown as brown lines in figure 2. The two NOSEY experiments shown in figure 2 starts at the

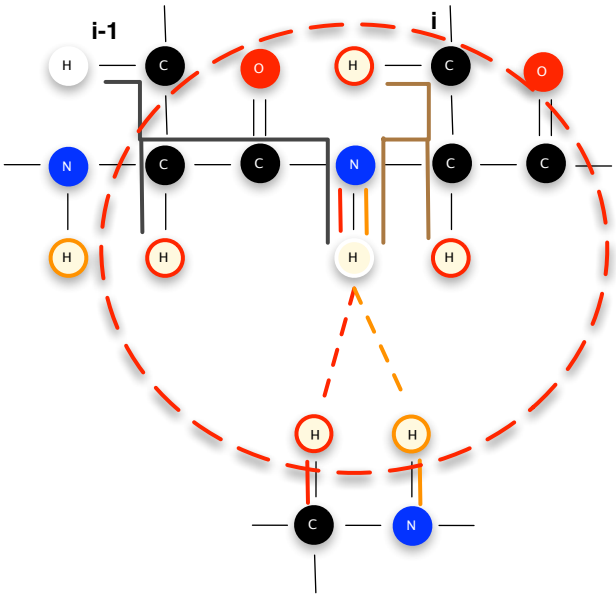

Figure 2. *Magnetization paths for two projection backbone experiments and two projection NOESY experiments. The gray and the brown lines describe backbone magnetization from residue i and i-1. Dotted red and orange lines describes N-edited NOESY and C-edited NOESY.*

nitrogen atom transferring magnetization to the amid proton. Then magnetization is transferred through space with dipolar coupling to either amid protons or protons bound to carbon atoms. 5D NOESY variants also exists where magnetization includes either the carbonyl carbon or the Cα carbon.

All projection experiments can be combined in different ways giving the possibility to use combinations that gives the best result on the given protein depending on what type of information that is required. An example of such combinations has been demonstrated on a Histone domain where five different experiments were used to cover backbone, HCCH-TOCSY and ^{13}C-HSQC-NOESY-^{15}N-HSQC, and ^{15}N-HSQC-NOESY-^{15}N-HSQC. The resulting decomposition of the projections gives components that contains shapes. The decomposition of these five experiments gave 15 dimensional components. One component is shown in figure 3. The left panel show nine shapes that correspond to both backbone experiments. Shape C', Cα/β and Hα/β correspond to residue i-1, in this case D102. The rest of the shapes in the left panel are from residue i, F103. This gives connection information later used by the correlation program for correlating components. The right panel shows TOCSY and NOESY shapes. The TOCSY shapes are from residue i-1. The four remaining shapes comes from the two types of NOESY experiments mentioned above. These experiments have a NOE peak for the amid proton to either HC_{noesy} or HN_{noesy} and these are either bound to C aliphatic or N atoms.

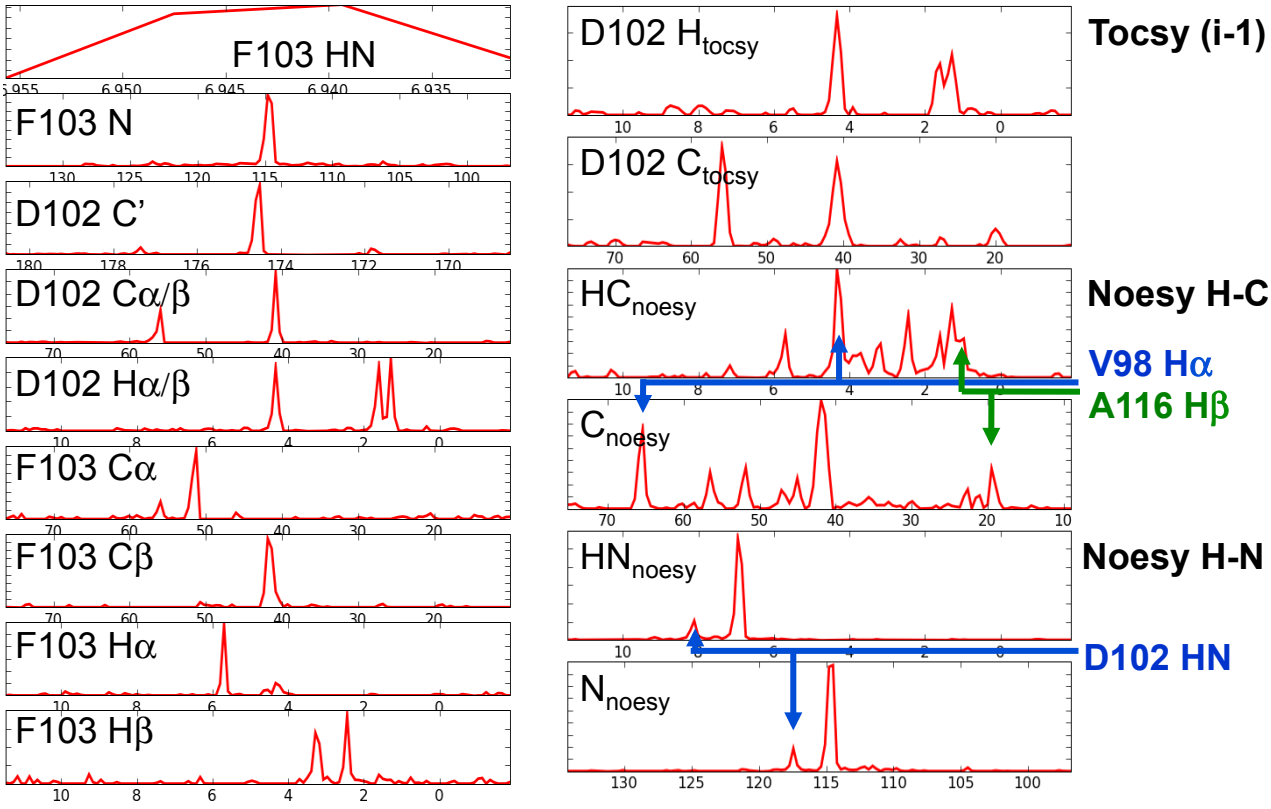

Figure 3. *Example of a 15D component resulting from decomposition of projections selected form five different experiments: two experiments targeting the backbone with scalar couplings, one experiment involving TOCSY transfers for side-chain assignments, and two involving NOESY transfers. The left pane shows shapes for the neighbouring backbone nuclei; the top two shapes on the right provide information on HCCH-TOCSY. The last four shapes provide NOEs to spatially neighbouring H-C and H-N groups, respectively. The blue and green arrows in the third and fourth shapes on the right identify long-range NOE.*

Shown in the HC_{noesy} and C_{noesy} shapes are two NOE peaks and the corresponding carbon atoms from the ^{13}C-HSQC-NOESY-^{15}N-HSQC experiment marked with two arrows. The last two shapes in the right pane shows the connection from the previous residue D102 with F103 from the ^{15}N-HSQC-NOESY-^{15}N-HSQC marked with an arrow. The strongest peak in the HN_{noesy} shape is from the same residue as expected while the second strongest comes from the previous residue in the chain. Both NOESY experiments together with side chain and backbone assignment can be used for structure calculation.

Materials

In this study four proteins where mainly used: Ubiquitin, Histone, Azurin and MMP20. Ubiquitin[44] is a 76 residue (8.6 kDa) protein found in many tissues where It is responsible for protein degradation in the cell. Measuring temperature was 303K conducted on a 600 MHz BRUKER magnet. Two projection experiments where used (paper 1). The Histone domain contains 93 residues[45]. All experiments for histone was conducted on a 600MHz magnet with a temperature of 298 Kelvin. Note that this temperature was 10 Kelvin over the recommended temperature which created a partly unfolding state resulting in shift degeneracy. This behavior was already present at room temperature and was enhanced when measured with higher temperature (paper 5). Several projection experiments where done including backbone, TOCSY and NOESY type experiments. Azurin is a 128 residue blue copper protein that transports electrons and it is found in many bacteria[46]. All experiments on Azurin was conducted on a BRUKER 600MHz magnet with a measurement temperature of 303K. Several different pulse sequences was tested and developed on Azurin at the Swedish NMR center and at BRUKER. MMP20 is a 160 residue protein that regulates tooth enamel formation[47]. All experiments for MMP20 where done on a 900MHz cryoprobe magnet with a measurement temperature of 298K at the CERM lab (www.cerm.unifi.it/home/). All programing development and implementation was done on a Linux workstation with two dual core opteron AMD processors and with 6 GB memory.

Results and discussion

The overall goal of this project was to implement and develop software tools for analyzing projection experiments on different proteins. Different projection experiments where tried on different proteins for experimental development and to investigate how different proteins affected the analysis of the decomposition. The different projection experiments where mainly done on ubiquitin, azurin, histone and MMP20, four proteins with increased complexity. The projection experiments that was used where combined in different ways to obtain optimal experimental results depending on the type of protein used and the type of experiment suitable for the analysis. The analysis and development part of the project resulted in various algorithms that where implemented providing an set of software tools. The result was PRODECOMP-SHABBA, two sets of programs for automated backbone assignments of projection experiments. One of the first implementation of the decomposition algorithm was tried on two 5 dimensional projection experiments characterizing $C\beta H_n$-$C\alpha H$-C'-NH-$C\alpha H$-$C\beta H_n$ on double labeled ubiquitin. For the analysis of the projections, the first version of SHABBA was implemented. SHABBA correlated the resulting components from the decomposition by using $C\alpha_i/C\beta_i$, $C\alpha_{i-1}/C\beta_{i-1}$ and $H\alpha_i/H\beta_i$, $H\alpha_{i-1}/H\beta_{i-1}$ shifts from current (i) and previous residue (i-1). These resulting chains where then used on statistical shift data to make a sequential assignment. A final peak picking resulted in a complete and correct backbone assignment (paper 1). To be able to use the software on larger datasets an improved implementation of PRODECOMP was done that reduced the amount of memory needed and decreased computational time. This version of PRODECOMP was implemented in python and a graphical user interface was added (paper 2). The mathematical background for PRODECOMP was presented in paper 3 together with a flowchart describing the algorithm and an application example. In paper 4 the 2D LS-ESPRIT method was tried together with projection data to estimate frequencies and damping factors in time domain data. The method was tested and verified on a ^{15}N-HSQC projection plane. In paper 5 four different proteins where used for further improvement of the SHABBA algorithm. The result was an improved version with a novel

assignment procedure and improved peak pickers for backbone characterization. The previous results for ubiquitin could be reproduced and also result from the three other proteins where presented. NOESY type projection experiments on the histone protein domain where tried and the resulting decompositions contained enough information to be comparable to a published histone structure (paper 6).

Overall algorithm

The overall algorithm from recording experiments to the final output of a backbone assignment or distance list is described in figure 1. Protein experiments are recorded first with coupled evolution periods to reduce measurement time. The resulting time domain data are then preprocessed resulting in 2D data sets. These are then Fourier transformed resulting in a number of 2D projection planes each with different linear combinations of frequencies from the nuclei in the indirect dimension. All or a selection of the planes are then used as input for PRODECOMP. An interval list defining the number of residues is also required and it can be either done manually or with the help of a program. The interval list is defined from a ^{15}N-HSQC spectra where every interval should contain one peak from the ^{15}N-HSQC spectra that correspond to a residue and should be as small as possible. The selected projections spectra together with the interval list are then used for the simultaneous decomposition calculation by PRODECOMP resulting in components containing shapes. Every component correspond to the residue defined in the interval list and contains different shapes describing the frequencies of the nucleus involved in the experiment. The

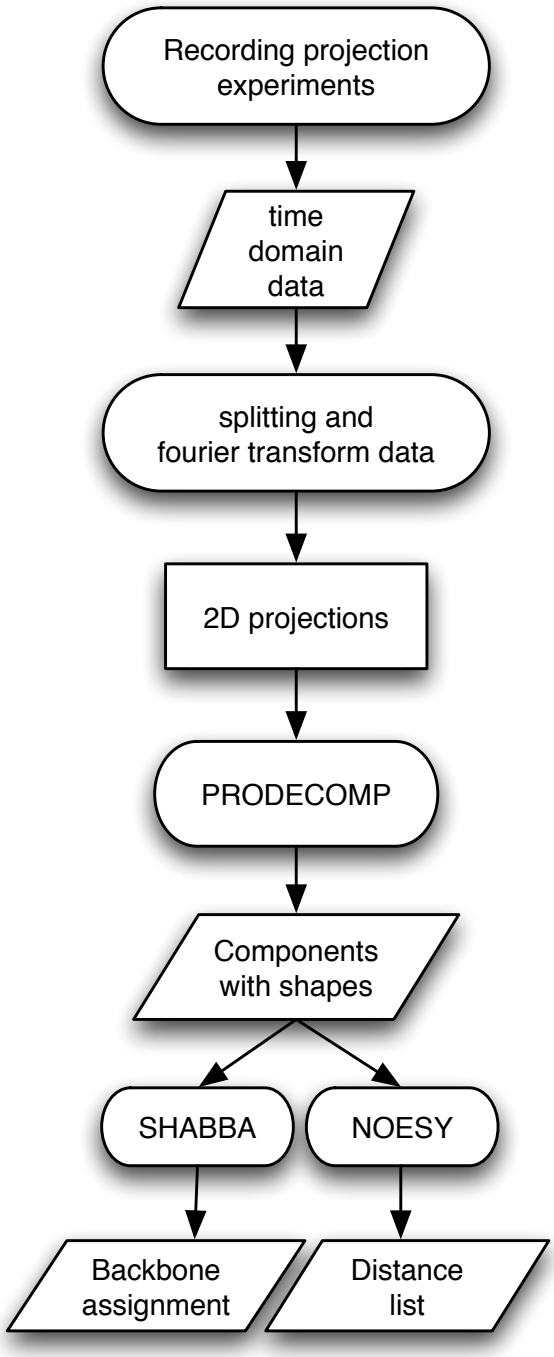

Figure 1. *Flowchart for the overall algorithm for backbone characterization or distance list output.*

resulting shapes are then used for either backbone characterization or NOESY analysis. The backbone analysis is done with the SHABBA software package that correlates the components and make a final backbone assignment. The NOESY analysis uses a program together with a short distance list that assign and verifies that enough information is contained in the shapes for a structure elucidation.

Projection decomposition approach

Time domain data in a multidimensional NMR experiment can be expressed as[48]:

$$S(t_1,t_2...t_N) = \sum_k f_1^k(t_1) \otimes f_2^k(t_2)... \otimes f_N^k(t_N) \quad (1)$$

Here the time domain signal in different dimensions are expressed as the sum over all components. Every component k contains Kronecker products of functions describing the time signal. Fourier transform over all signals in eq. 1 gives the corresponding spectra in frequency domain[49]:

$$S(\omega_1,\omega_2....\omega_N) = \sum_k f_1^k(\omega_1) \otimes f_2^k(\omega_2)... \otimes f_N^k(\omega_N) \quad (2)$$

Here the N-dimensional spectra is described as a sum of Kronecker products between the components of the spectra. This equation is an extension of a method called three-way decomposition (TWD)[50] and have been implemented in NMR[51]. Components are one or several peaks present in the experiment. Here every component k consists of the Kronecker product of different one dimensional vectors describing the different resonance frequencies in the left side of equation 2. These vectors are called shapes and they correspond to the different resonances of the different nuclei in the experiment. In projection experiments indirect evolution periods are coupled, meaning that time increments in the indirect dimensions are dependent. This means that an experiment with M indirect dimensions can be projected from N dimensions to N-M+1 dimensions. These projection experiments can then be described as in equation 3:

$$P_m(\omega,\omega_N) = \sum_k (F_1^k * F_2^k ...)(\omega) \otimes F_N^k(\omega_N) \quad (3)$$

Here P_m represent one 2D projection spectra with frequencies ω representing the indirect dimension and ω_N representing the direct dimension. For every projection m there exists a specific linear combination of nuclei and this linear combination is represented as shapes F_1,F_2...F_{N-1} in the right side of equation 3. The summation goes over all components k where one component now consists of N-1 shapes describing the indirect frequencies and one direct shape F_N normally represented by the amid proton. The indirect dimension consists of convolutions between the different shapes marked with the convolution operator '*'. Different convolutions can be combined and described in equation 3. In every projection spectrum one peak corresponds to either one nuclei or a linear combination of two or more nuclei in the indirect dimension. Peaks in the indirect dimension can be folded because of limited spectral width. Decomposition can resolve folded peaks correctly thus avoiding the need for larger spectral width that would reduce resolution. By using equation 3 it is also possible to reconstruct the projection and therefore check for consistency between the calculated spectra and the measured spectra. The reconstruction is a part of the iterative procedure to obtain the closest solution to the optimization problem by finding the minimal difference between the calculated projection and the measured projection:

$$\min\left(\sum_m^M (P_m(\omega,\omega_N) - \sum_k (F_1^k * F_2^k ...)(\omega) \otimes F_N^k(\omega_N))^2\right) \quad (4)$$

The minimization procedure calculates first F_1 keeping all other indirect shapes fixed. Then F_2 is calculated with the rest of the shapes are fixed. The whole minimization procedure is repeated for all shapes thus minimizing all shapes simultaneously for all projections. This will in effect distribute all signals over all projections and also increase the possibility to resolve peaks that are very weak which is important in projection experiments. To improve the convergence a Tikhonov regulation factor[52,53] can be added to eq. 4.

PRODECOMP

PRODECOMP, **Pro**jection **Decomp**osition, decomposes projection experiments described in eq. 3. The output are vectors called shapes describing the different frequencies in the experiment. The algorithm (paper 3) is described in figure 2 on the next page. The flowchart shows the decomposition of one interval consisting of three loops and where every pair of components are optimized. When all three loops have finished the output consists of shapes from one component. The algorithm is then repeated for the next interval until all components have been calculated. Input to the algorithm consists of projection experiments and an interval list. Individual projection planes can also be excluded from the analysis, to reduce computational time. This was done for the backbone analysis in paper 1 and in paper 6.

The interval list contains an interval for every residue present in the experiment and can either be determined manually or by a peak picker from a normal ^{15}N-HSQC and compared to a projection ^{15}N-HSQC to remove side chains and to see wether there exists weak peaks. The intervals are defined in points from the direct

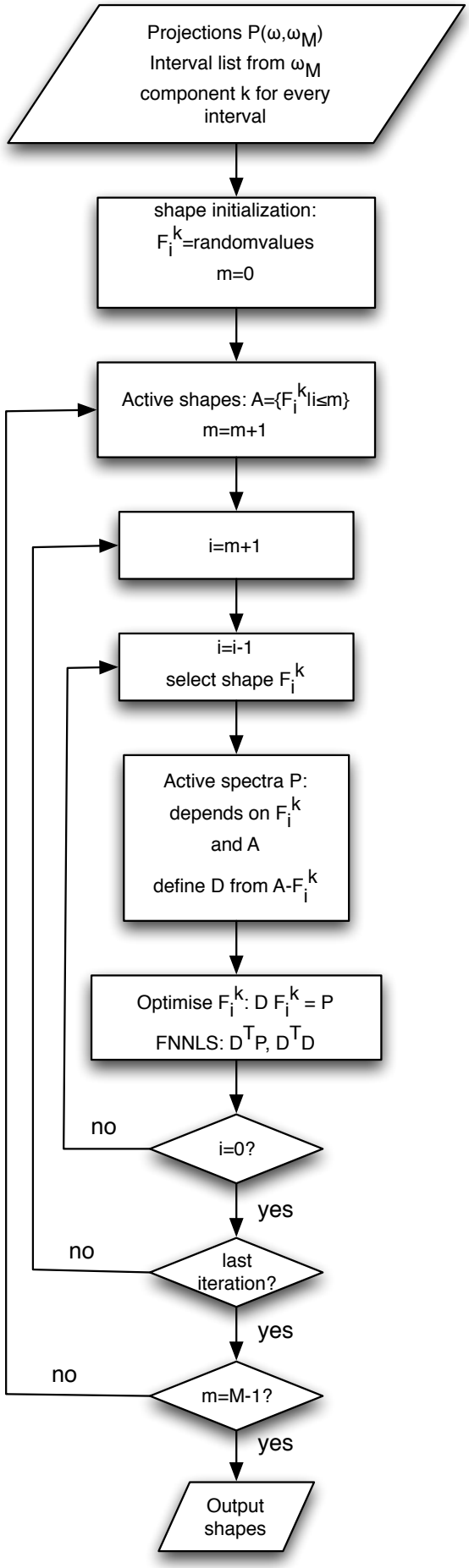

Figure 2. Flowchart for the PRODECOMP algorithm. Input are different projections from experiments and a list of intervals from an HSQC spectra. Components are defined from the interval list. Every shape in a component is initialized with random values as starting values. The first loop defines a set A of active F_i shapes starting with the first shape and then adding more to A as i increases until all shapes for component k are added. The direct dimension is always active. The next loop includes the next active shape. The third loop selects the current shape and defines a set of active spectra that contains the active shape. A square matrix D is defined as the row shifted shapes that correspond to the convoluted nucleus in the experiment except shape F_i that is going to be determined. The D matrix together with shape F_i and P is now used as input for the FNNLS algorithm. After determination of shape F_i all the previous shapes are optimized the same way. Then m is increased and another shape is added and optimized against all others. In the optimization step the every shape is optimized against every relevant spectra thus drastically decreasing the chance for a false positive. After the third loop another interval is calculated until all intervals have been decomposed. The resulting set of shapes can then be used for further analysis depending on the experiment.

dimension and should be as small as possible to avoid overlap. Ideally, every interval defined in the direct dimension should represent one peak in a ^{15}N-HSQC spectra. This is normally achievable in less dense regions of the spectra but can be more challenging in crowded regions depending on the protein. Every interval has a number of components that is set equal to the number of peaks in the interval. Additional components can be added if there is a lot of noise in the interval or if there is a lot of overlap in the direct dimension therefore making it hard to distinguish between two or more peaks. This was more frequent for the azurin, histone and MMP20 proteins than for ubiquitin. The reason for this was that these spectra contained more overlap and different signal intensities that required more components in the analysis (paper 5). The intervals are then used for calculation of the corresponding shapes from the selected components. An example of an interval list can be seen in figure 3. Those projections that have more than one nuclei in their indirect dimension are convoluted which means that every single peak in those spectra correspond two or more convoluted frequencies as described in formula 3. When all shapes are known a reconstruction can be done to compare with the original spectra and then calculate a residual. This residual is then used as an optimization criteria and it is used for minimizing the differences between the reconstructed spectra and the measured.

In projection experiments the signal intensity for one nucleus is usually spread over all spectra containing the nucleus giving a low signal to noise in the projections. By simultaneously analyzing all spectra the signal intensity can be preserved. This can be illustrated from the following example: consider 15 projections from a 5D projection experiment with 100 points in each projection. Every projection corresponds to one equation in a system of linear equations. Each signal is represented by one point. Let signal to noise be close to one and lets consider only 20% largest positive points as potential signals, that is 10 points for every projection. If we would consider only the first four equations there would be 10^4 solutions. However a solution is only valid if it is satisfies also the other 11 equations. For each equation there is a 10% chance that one of the

random solutions of the first four equations is satisfied. Thus the chance for large noise points to give a consistent signal in all 15 equations is $10^4/10^{11}=10^{-7}$. With several experiment optimized simultaneously the chance for a false peak identification is very low as shown above because all signals has to be matched in every projection in the optimization.

A user interface was developed for the prodecomp algorithm (paper 2). The interface was written in TCL/TK and it's available at www.lundberg.gu.se/nmr/. Figure 3 shows an example of the input intervals for azurin. All intervals are defined by points in the direct dimension and every interval has a number of components. The number of iterations can be changed and the regularization factor.

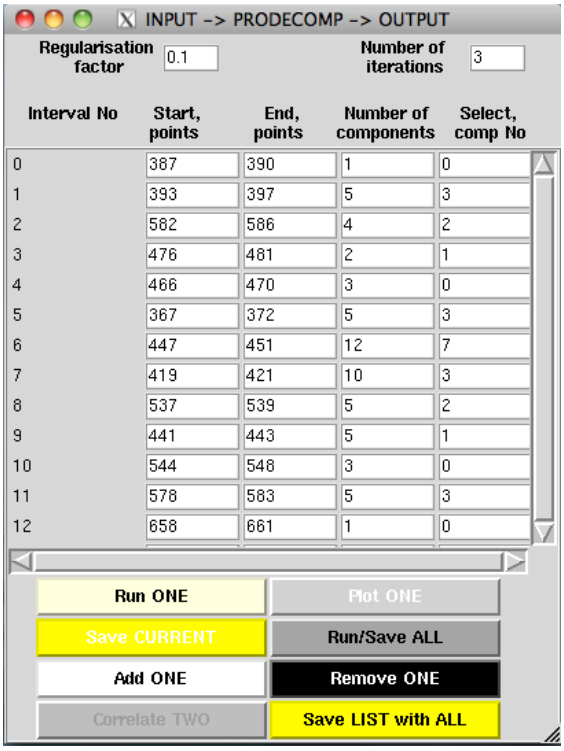

Figure 3. *Graphical input for decomposition calculations. Every peak in a ^{15}N-HSQC corresponds to an interval defined points in the direct dimension, defined in the first two columns. the next columns indicates how many components that should be used for the calculation. When the calculation is done one component is selected that represent the residue as seen in the last column. One interval at the time can be calculated or the whole list can be sequentially calculated. Every interval can be plotted and an interval can be added or deleted.*

Backbone analysis

The resulting components from a decomposition of backbone experiments are used in several steps before a final backbone assignment can be done. All correlations and assignments are based on the shapes in the components. The resulting decomposition from a projection experiment contains shapes describing different frequencies of the nuclei involved. An example of two components S66 and G67 of azurin resulting from decomposition of two backbone experiments described earlier with magnetization transfer $C\beta H_n$-CαH-C'-NH-CαH-$C\beta H_n$ are shown in figure 4. Every component contains 9 shapes describing the involved nuclei. Note that in figure 4 the shape describing the direct dimension NH is omitted. The shapes $C\alpha/\beta_{i-1}$ and $H\alpha/\beta_{i-1}$ are shifts from the previous residue in the sequence. The arrows in figure 4 between S66(i-1) and G67(i) shows how shapes $C\alpha/\beta_{i-1}$ and $H\alpha/\beta_{i-1}$ in G67 have the same shifts as the Cα, Cβ and Hα, Hβ shapes of S66. This indicates a correlation between the two sequentially connected residues that can be used for a sequential assignment. The $C\alpha/\beta_{i-1}$ and $H\alpha/\beta_{i-1}$ shifts can also be present in the same component as indicated with dotted lines in the left pane. In the right pane shifts for Cα and Hα are missing in the corresponding shape. This is because glycine lacks Cβ and Hβ signals and the Cα and Hα signals in glycine have the same phase as resonances involving Cβ and Hβ in all other residues. This is common in many triple resonance experiments.

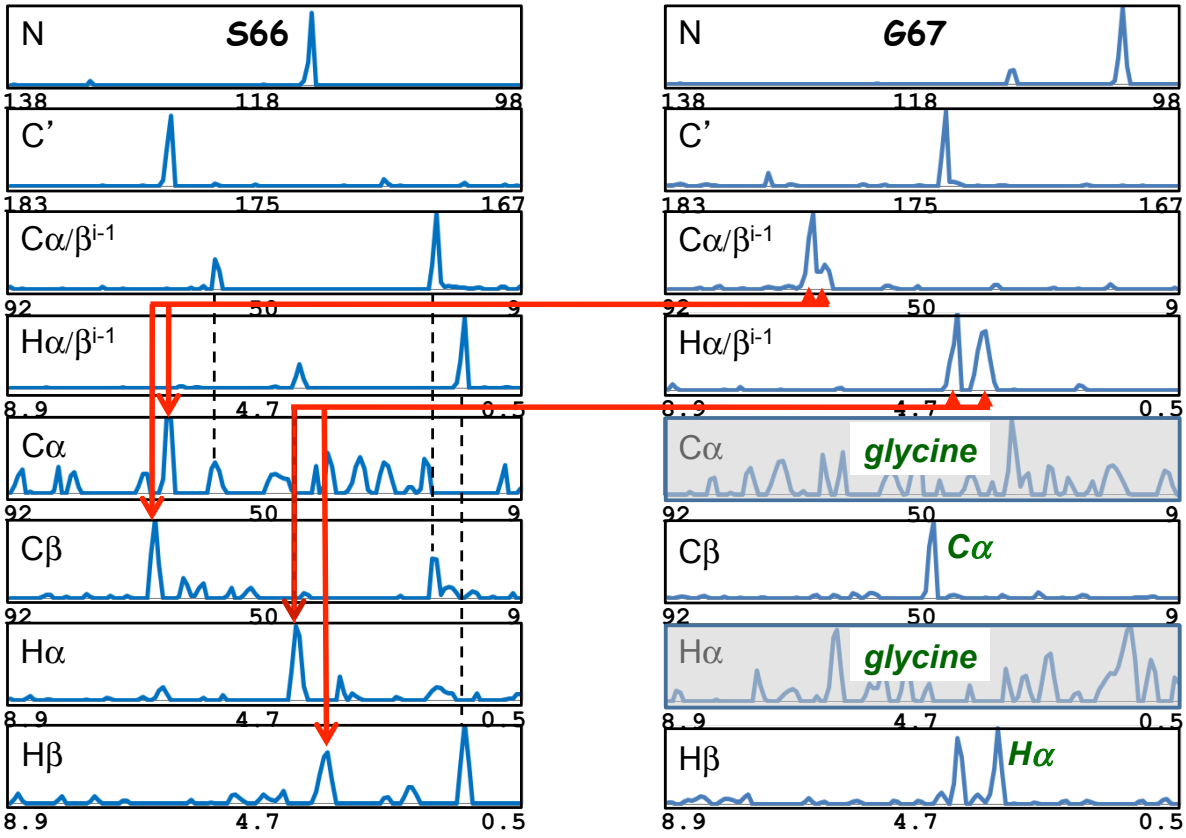

Figure 4. *Two 9 dimensional components from decomposition of azurin showing residue S66 and G67. The arrow shows the same shifts for and Cα/β$_{i-1}$ Hα/β$_{i-1}$ in residue G67 to Cα$_i$, Cβ$_i$ and Hα$_i$, Hβ$_i$ in S66. Note that the HN shape is omitted in this figure and that glycine lacks peaks in the Cα and Hα shape.*

SHABBA

SHABBA, **Sha**pe **B**ack**b**one **A**nalysis, uses shapes from PRODECOMP as input to make a backbone assignment. Several intermediate steps are required using different programs. All steps described are implemented in different python programs except the sliding part which was implemented in the Fortran language. The overall procedure is to correlate components from PRODECOMP resulting in chains of components and then slide them over the sequence comparing peak picked Cβ values with statistical values. Different length of the chain is compared and the position with the lowest RMSD is a candidate for sequence assignment. When all chains have been assigned a final peak picking procedure gives the final assignment.

The first version of SHABBA was used in paper 1 to make a backbone assignment of ubiquitin. This version used a correlation procedure that gave chains as an output. These chains where then peak picked with respect to Cβ and Cα. The result was then used for comparison with statistical data making a sequential assignment. A final peak picker was then used to give a complete assignment. This first version gave good results for Ubiquitin (paper 1) but for larger or otherwise more challenging proteins an improved version had to be developed (paper 6). The algorithm for the improved version is described in figure 5. Input are components from decomposed projection experiments that describes backbone frequencies. An automatic glycine detection is done on the components by identifying missing Cβ and Hβ signals in the shapes. The user has also the option to manually inspect the shapes and add or remove suggested glycines. The loop in figure 5 that follow after the initialization step describes how the chains are calculated and slided with different parameters. Every iteration in the loop is indicated by a iteration variable which is initialized to zero. When the loop starts all components are used for a correlation calculation. The correlation calculation is done by

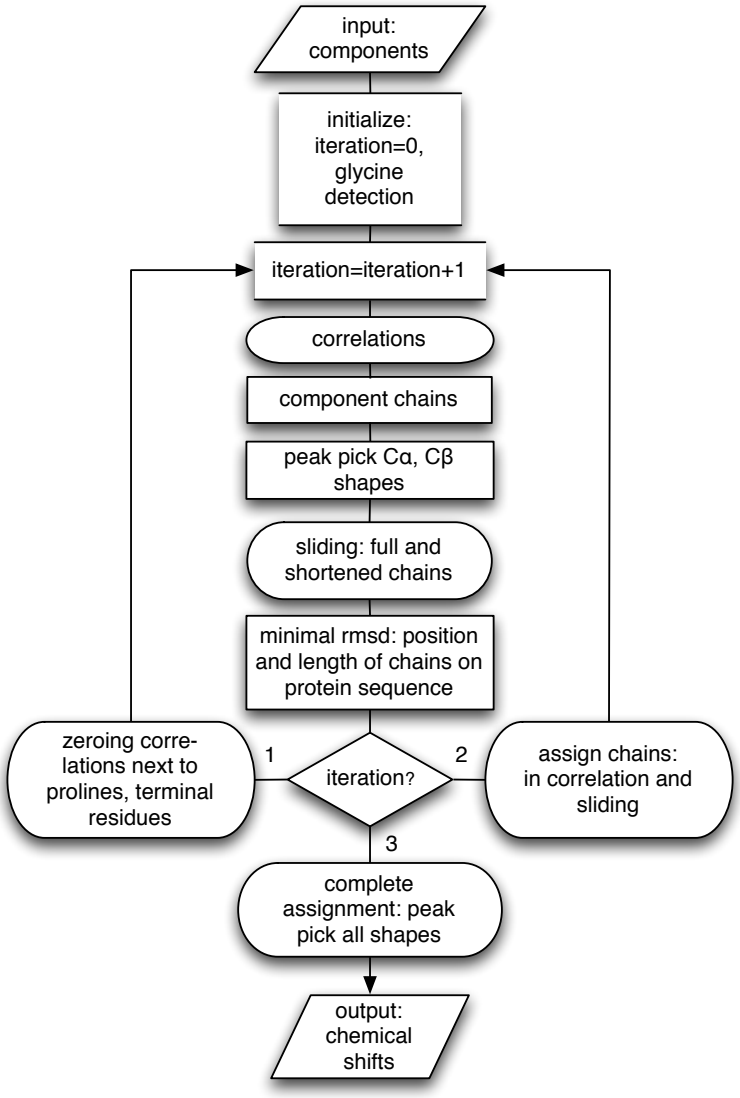

Figure 5: *Flowchart for backbone assignment using decomposed projections. Output is a chemical shift list.*

comparing all pairs of components with regard to common Cα, Cβ and Hα, Hβ shapes from the i residue and shapes Cα/β and Hα/β from the i-1 residue as described previously. These shapes can then be used to correlate two neighboring components. All component correlation pairs form a square matrix where all column entries are from the i-1 component and all row entries are from the i component. Every element in the matrix has a correlation value. The correlation value is calculated by adding Cα and Cβ shapes together for the component. The resulting shape is then compared to the Cα/β shapes of the other component and the same is done for Hα and Hβ. This repeated for all of the rest of the components. When all correlation values have been calculated in the matrix four rules are applied on the resulting correlation matrix:

1. Set all correlations that are negative to zero. Correlations are defined to be positive.

2. Set diagonal values to zero. Diagonal values represent a correlation from a component to itself which is not realistic.

3. Remove one of the entries that have the lowest value of pairs that are symmetric with respect to the diagonal. This avoids circular connections

4. Set all elements on row y and column x to zero that are lower than the maximum value. The highest correlation is assumed to be the correct one. If only one element is left for the row and column then its unique and considered a correlation.

The final step is to remove all correlations that are under 20%. This means that if the correlation was below this value the correlation was to weak to be considered as a candidate for sequential assignment. When all rules have been applied a set of chains are returned that are used in the sliding step. Ideally the chain should only be broken by a proline, giving a minimum number of chains from the correlation. The correlation procedure is repeated two more times with different

pre set values for different element in the matrix. These values are coming from the next sliding procedure.

Sliding

The resulting chains from the correlation calculation have their Cα and Cβ shapes peak picked before the sliding step. This peak picker uses information from the current residue and the previous one. The peak picker removes intensities from the shapes that are under a noise level. It then removes all peaks in the Cα and Cβ shapes that correspond to the previous residue indicated by dotted lines in residue S66 in figure 4 to avoid false positives. Then all shifts that are not whitin a statistical range are removed. The final peak is then peak picked using a three step procedure. The resulting shift list is then used for a sliding procedure where all chains are slided over the protein sequence. Every residue in the sequence has a statistical value for the Cα and Cβ collected from the BMRB[47] database witch is compared to the value of the peak picked values in the chain by calculating a RMSD value for all shifts in the chain. The loop in the flowchart of figure 5 consists of three iterations. In the first step every chain with more than five components is slided over the sequence. For every position a RMSD value is calculated separately for Cα and for Cβ. Normally Cβ shifts have a wider spread than Cα shifts making them more suitable for RMSD comparison. Cα values are nevertheless used for supporting information. Prolines have a penalty factor added which will increase RMSD when a chain is slided over to detect where a chain should be stopped. This is then repeated for the same chain but with the end component removed. This procedure is repeated until the length of the chain is 6 residues. All Cβ RMSD values for every length of the selected chain is compared and ordered. The position that gives the lowest RMSD value for the specific length of chain is then recorded. If the position is directly after a proline or the N-terminus or directly before a proline or a C-terminus the correlation is zeroed to indicate that a component cannot have a correlation to a proline or the terminal ends of the sequence. This procedure is then done for all of the rest of the chains that have a length over 5 components.

When entering iteration 2 the correlation calculation is repeated again now with the added cuts in iteration 1. After the correlation calculation the sliding step is repeated with the same parameters as before using the resulting chains from the correlation calculation. After the sliding and RMSD comparison new chains are now fixed internally by setting the correlation between them to one. This means that no other components are able to replace a component within a chain, i.e the chains are 'fixed'. What is left now are chains with a length under 6 that have to be placed in the sequence.

In step 3 a final correlation calculation is done and the rest of the chains are slided over the sequence. These last chains have a short length that give them a high probability to be placed in many positions because less unique RMSD values. By assigning all other chains this probability will decrease giving only a few positions left to position them. The resulting small chains are then placed in the right position on the sequence. The final step is then to do a final peak picking over the sequence, giving a final peak list. The peak picker uses both residue i and residue i-1 for peak picking. It also uses residue specific statistics to increase the chance for a correct assignment.

An example of an sliding result is shown in figure 6 for azurin (paper 5). The first three rows display a component chain of length 16 and a possible position on the protein sequence (residue numbers and names). Row 4 lists the Cβ chemical shifts peak picked in the components. Row 5 contains the statistical Cβ values from BMRB for the protein sequence. For each component-residue pair the shift difference is used to calculate the RMSD value for the chain. As can be seen in figure 6, the first 10 pairs yield small differences resulting in a small RMSD (0.9) for this partial chain. However, adding the following 6 pairs increases the RMSD to 18.3. complete chain is fitted over the right position up to residue 72. The whole chain has a RMSD of 18.3 By removing one component at the time and calculating new RMSD values for every new length until the length is 6 a better

comp	59	60	61	62	63	64	65	66	67	68	53	54	55	56	57	58
res	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78
	GLY	MET	ALA	SER	GLY	LEU	ASP	LYS	ASP	TYR	LEU	LYS	PRO	ASP	ASP	SER
	0.	33.	19.	64.	0.	42.	40.	32.	39.	?	28.	0.	32.	33.	?	?
	0.	33.	19.	64.	0.	42.	41.	33.	41.	39.	56.	57.	no	41.	41.	64.

rmsd: 0.9; res: 63-72; comp: 59-68

rmsd:18.3; res: 63-78; comp: 59-68,53-58

Figure 6. *Example of fitting a chain from azurin into the sequence. Cβ shifts from components 68, 57 and 58 have not been detected and do not contribute to the RMSD calculation. Residue 75 is a proline giving a high penalty to the RMSD calculation. Zero ppm is given for glycine.*

NOESY

Structure determination of proteins are in many cases the final goal of a protein characterization. The type of experiment can be either 4D or 5D NOESY or both with different magnetization paths. The decomposition is done with PRODECOMP and the resulting shapes from these experiments provide information about HN and CH NOE distances in the protein and can be used for an structural analysis together with additional assignments. An example of two shapes from the

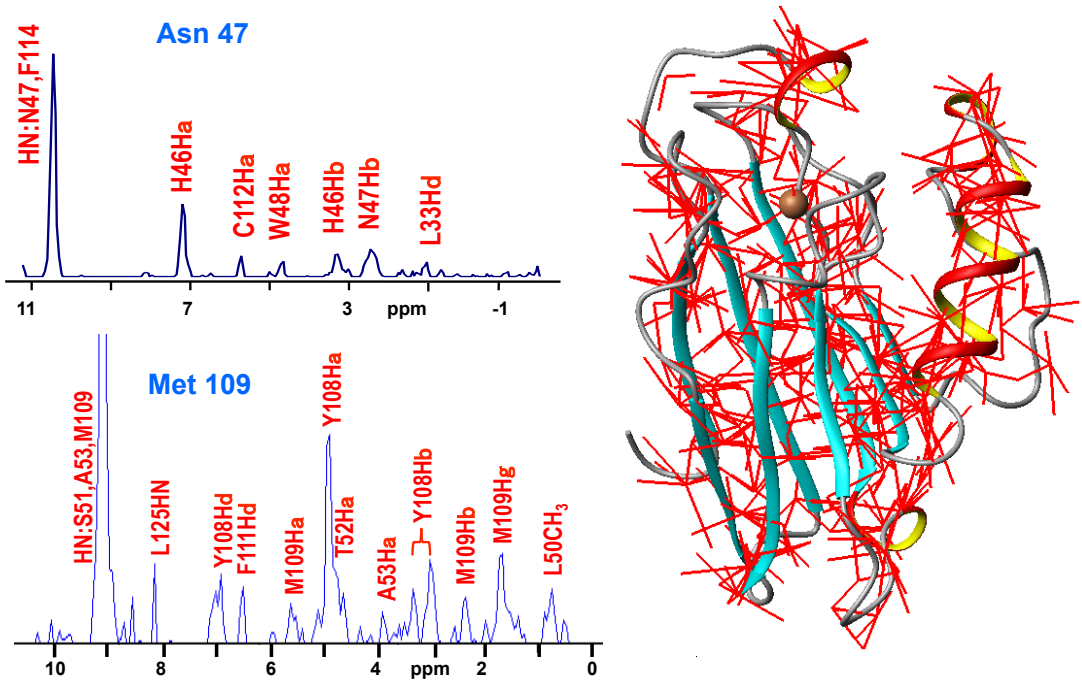

Figure 7. *Two shapes from two 4D NOESY experiments containing NOE distances from amid protons to protons. All distances are marked in the structure.*

decomposition of two NOESY experiments is shown in figure 7 for azurin. Shown in the two shapes in figure 7 are all assigned NOE signals with some signals very close to the noise level. Long distance NOE are also present. The shapes in the figure where assigned using a reference list[55] and using distances from a published[56] PDB (4azu) structure. The assigned distances are marked in the structure with lines. The mean backbone RMSD to the x-ray structure was 1.3 Å and with side chain 1.5 Å. This shows that the structure information given from the two NOESY projection experiments were consistent with the known PDB structure.

Another approach involving NOESY projection experiments is to use a combination of NOESY experiments and backbone experiments to obtain sequential correlation. This is illustrated in figure 8 where two components from three experiments are shown. The left component shows shapes from two backbone experiments, while the right component shows shapes from one of the backbone experiments together with a projection HSQC-NOSEY-HSQC experiment.

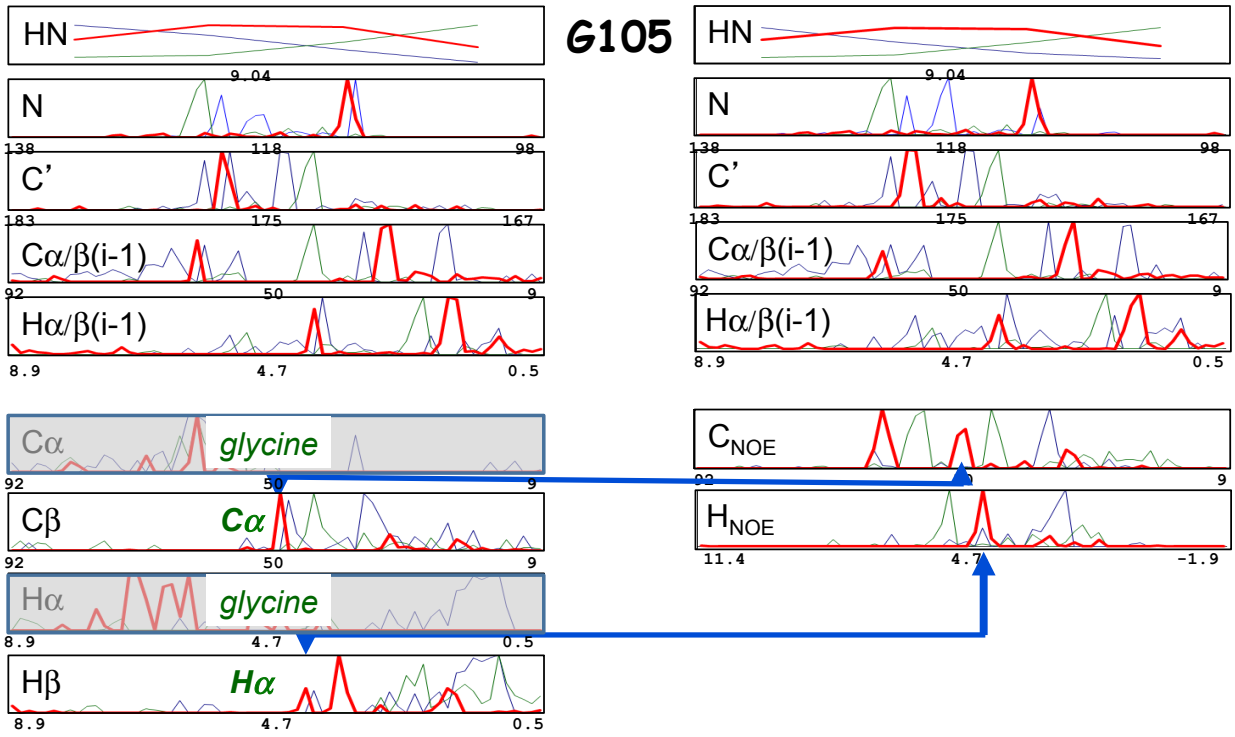

Figure 8. Sequential NOE connectivities in azurin indicated by two arrows from C_{NOE} and H_{NOE} to the Cα and Cβ shape. With this a sequential assignment can be done.

The two arrows in figure 8 indicates that the C_{NOE} signal and the H_{NOE} signal in the NOESY shapes are from the same residue. The C_{NOE} shape has the second largest intensity while the largest intensity comes from the i-1 residue which can be seen in the i-1 Cα/β shape. With this information a possible sequential assignment can be done thus making it possible to replace one backbone experiment with a projection HSQC-NOESY-HSQC experiment.

Papers

As described in paper 1, the first version of PRODECOMP and SHABBA was applied on the ubiquitin protein. Here the algorithm was used for analyzing the backbone projections and the resulting correlated components was sequentially assigned using a comparison with statistical data for every residue in the sequence. The complete backbone assignment was done using 30 projections from two backbone experiments covering spins $C\beta H_{n-1}$-CαH-C'-NH-CαH-$C\beta H_n$. Figure 9 shows two projections planes from the ubiquitin experiments showing linear

combinations N-CO-Cα/β_{i-1} and N+CO+Cα/β_i where i is the current residue and i-1 the preceding. The resulting shapes from decomposition of these experiments corresponded to a 9 dimensional experiment.

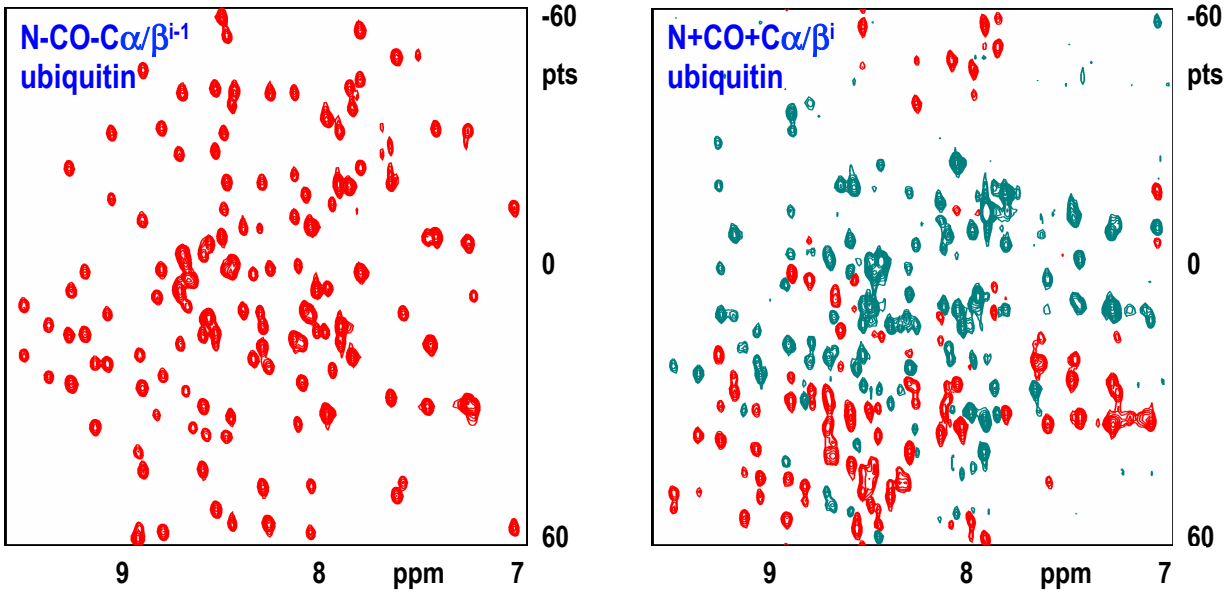

Figure 9. *Two projection planes from the Ubiquitin backbone experiments. Green peaks in the right pane correspond to negative Cβ peaks.*

The left projection shows N-CO-Cα/β_{i-1} combination in the indirect dimension. The right projection shows N+CO+Cα/β_i combination. The green peaks corresponds to negative peaks coming from Cβ in the same residue. Because the fast-NNLS algorithm cannot use negative values as input, projections containing negative peaks are sign inverted and therefore adding an additional 16 projections to the decomposition.

The correlation procedure calculates all correlations and fill the corresponding entry in the correlation table. An example of the correlation calculation are shown In figure 10 where all correlations are displayed for a fragment of size 17. The columns show the i-1 components and the row show the i component. For example, component 5 correlate with component 4 with 92% correlation. To achieve this number, all correlations have been calculated for all components

i / i-1	13	5	8	6	17	2	7	15	14	18	12	4	3	16	10	9	11
13	**35**	-2	-1	-3	-3	-3	1	8	34	23	59	-3	-2	-4	-2	9	-3
5	5	**51**	11	3	-1	10	0	-4	-4	-4	-3	**92**	5	6	-3	2	0
8	17	59	**15**	27	13	-3	48	18	16	-1	0	0	1	35	-2	3	-3
6	28	74	17	**45**	9	2	40	11	11	-2	1	0	-1	33	-2	7	6
17	11	23	12	18	**29**	36	18	0	11	0	2	9	18	72	-2	2	7
2	-2	1	9	19	0	**52**	8	-3	-2	3	8	9	16	14	0	-1	15
7	-3	12	0	83	0	42	**30**	2	-3	-2	6	7	19	13	-2	-2	20
15	18	-3	3	-3	0	-3	12	**13**	50	1	33	-2	0	0	-2	15	0
14	83	39	15	16	3	3	28	-3	**33**	-4	0	12	16	8	-2	7	0
18	-3	2	-1	1	90	-3	-2	22	1	**23**	-1	-4	4	9	-2	0	-4
12	-2	0	33	9	-2	4	0	-3	3	-4	**51**	0	9	0	24	29	92
4	31	17	22	7	-3	7	7	-4	14	-4	2	**31**	79	-2	-2	2	12
3	2	8	1	43	7	80	19	-3	0	26	1	9	**42**	29	-2	0	9
16	-4	14	-4	11	19	-4	-1	92	-4	28	-3	-4	-4	**53**	-2	-1	-3
10	36	0	-1	-2	-2	-2	1	2	29	-4	7	-3	-3	-3	**4**	41	2
9	3	9	91	1	1	6	-1	-4	0	-2	16	17	10	4	26	**23**	29
11	-3	-4	20	-2	-3	-2	-2	-3	-3	-3	12	-3	-2	-4	98	25	**43**

Figure 10. *Correlation table for Ubiquitin showing one chain 2-18 before any application of rules. Bold numbers are final correlations between components.*

pairs. A fraction of these are shown in figure 10. First all entries that have a negative number are replaced with zeros, and all numbers on the diagonal are replaced with zeros. Then all numbers on row 5 and column 4 (component ordering) that are under the maximum correlation value are replaced with zeros, in this case all values on row 5 and column 4 except the maximum value 92. The next step is to remove mirror values to avoid circular connections, in this case the correlation on row 4 and column 5, with a value of 17 which is less than the maximum value. The lower threshold for finally accepting a correlation was 20%, correlations under this value are considered to weak. For ubiquitin the average correct correlation was 79.67% and the average correlation was 5.94% indicating that most correct correlations were strong. This was also seen in paper 5 were the sequential assignment only required one step in the sliding procedure, also an indication of strong correlations.

In paper 2 the PRODECOMP algorithm was translated to python and improved with respect to memory consumption and speed improvement. A large part of the memory consumption was due to large matrixes handling. This was replaced by a tracing method that reduced memory consumption. Normalization of the input

data also reduced computational time making it possible to analyze more projections and also to decompose projections with higher resolution which can be especially important in TOCSY and NOESY projection experiments. As mentioned previous, a graphical user interface was also implemented.

Paper 3 gives the mathematical formulas behind PRODECOMP and a flowchart describing the algorithm. Also an application example is given in the form of a projection decomposition of projections spectra from ubiquitin. The flowchart is described in detail in the PRODECOMP section. The resulting shape in the example contains 15 points in the direct dimension in the NH shape. This illustrates one approach to use broad intervals containing several peaks in the direct dimension. Another approach is to use small intervals covering only one peak. This approach was subsequently used for the rest of the proteins studied.

Paper 4 described how a signal processing method can be used on time domain NMR data for signal parameter estimation. The method was used on a selected projection corresponding to a ^{15}N-HSQC spectra measured from two 5D backbone projection experiments on ubiquitin. By using 2D sub band filters and 2D LS-ESPRIT methods on time domain data the signal estimation showed a clear agreement with the Fourier transformed spectra, se figure 11. The method is promising but needs to be investigated with more proteins. One drawback is that the number of indirect points must be larger than the number of sinusoids describing the signals making it necessary to introduce sub band filters to reduce the spectra into regions where the number of indirect points are larger than the number of peaks.

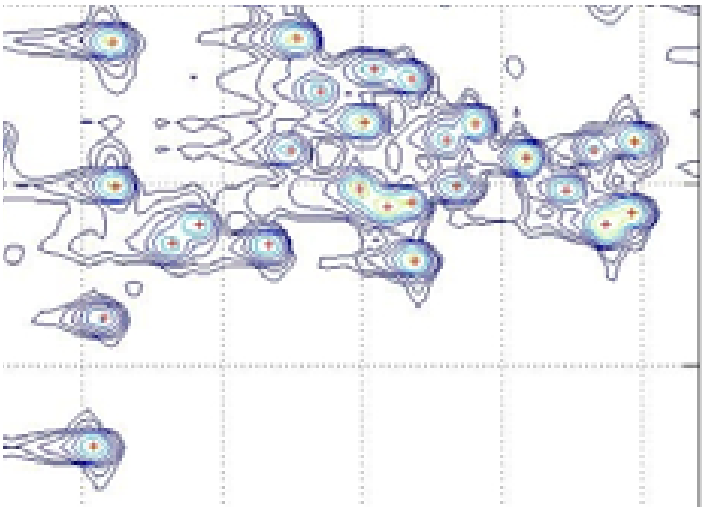

Figure 11. *Part of a corresponding ^{15}N HSQC projection where frequency estimated peaks are marked with a cross, showing good agreement with the Fourier transformed peaks. This area was 95 pts in the direct dimension and 60 pts in the indirect dimension with 26 signals. The number of signals where less than the number of indirect points, a criteria for the algorithm.*

In paper 5 four proteins of varying characteristics, ubiquitin, azurin, histone and MMP20, where used for investigating the limits of the PRODECOMP approach. Two backbone experiments with the following magnetization paths $C\beta H_n$-CαH-C'-NH-CαH-$C\beta H_n$ where used on all four proteins. One result of this study was that the correlation and assignment program SHABBA was improved to handle a wider range of proteins. For ubiquitin the previous result obtained in paper 1 could be reproduced and improved due to the improved software with a 99% complete and correct backbone assignment. The correlations calculation gave 5 chains and these chains needed only one iteration to be sequentially assigned because a high correlation between the components. Only one of the chains had a length that was under six residues thus making it eligible for a second round in the sliding procedure. With just one chain left after the first sliding step a second round was not necessary because it could be assigned directly. 30 backbone projections where used for the ubiquitin analysis and 72 components where the result from the decomposition. One peak was missing in the ^{15}N-HSQC experiment, Glu 24 but was found shifted in the decomposition analysis. An additional glycine was also detected due to partial degradation.

The correlation step presented in paper 1 was sufficient for ubiquitin but it had to be developed further for larger proteins, partly because larger proteins gives

closer correlations increasing the chance for a false correlation. Also the peak picker used where sufficient for ubiquitin but had to be improved to handle different proteins by introducing statistical shifts limits for every residue. As seen in figure 10, most correlations for the ubiquitin fragment is well over 50%. This is a promising result for ubiquitin sized proteins and a future goal is to investigate more proteins whitin this size range.

For azurin, three sliding steps where required because the correlation calculation gave 3 chains that were smaller than 6 residues out of a total of 9 chains. These 3 chains where not considered in the first sliding step. The result of the 3 sliding iterations for azurin are shown in figure 12.

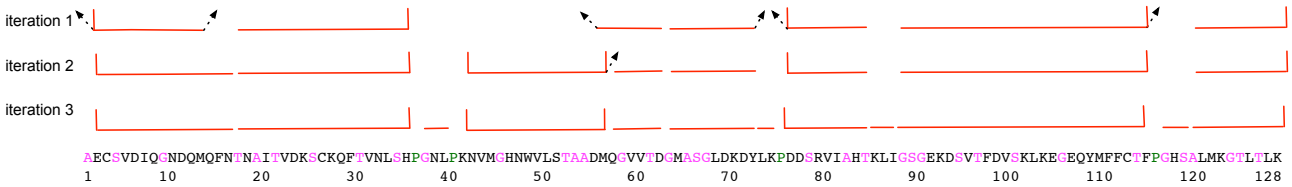

Figure 12. *Result of Backbone the assignment of azurin. The three lines illustrate the results of the three iterations with chains positioned on the sequence by low RMSD marked red. Arrows indicate positions for cutting the chain and vertical lines mark zero correlations.*

The number of residues in azurin are 128 with 4 prolines and 11 glycines. The N-terminus residue is not visible in a ^{15}N-HSQC. The remaining 123 residues where used for correlation after the PRODECOMP and glycine detection step. In iteration one 6 chains out of 9 chains where slided over the sequence. These had a length over 5 components while the rest where less than 6 components and therefore not used in this first iteration. In figure 12, all chains with an arrow are chains that have a sequence that are not sequential with the rest of the of the chain. The leftmost chain is flanked by two additional chains that have the wrong correlation and therefore the total chain have a higher RMSD than the correct chain. Therefore a cut can be placed there. A cut means that the corresponding correlation row or column is set to zero to indicate that no correlation should be accepted i.e that no component can come directly after or before this component.

This cut is then used for the next iteration. Four other components have also one part of the chain that is not correlated correctly, as indicated by the tilting arrows. The second and last chain was correctly assigned. The vertical bars in the line indicate a cut. Five cuts have been added for five of the components. The first cut is from the N-terminus of the chain, indicating that the left part of the chain is wrongly correlated, extended beyond the N-terminus. The remaining cuts come from proline positions and finally the C-terminus position. Chain 3 and 4 came from one chain that contained two low RMSD and therefore both where accepted since they did not overlap.

In the second iteration the modified correlation table was recalculated and the resulting 8 chains that had a length over 5 where peak picked and slided again. As can be seen in figure 12, the third chain have a part that is overlapping the fourth. An additional cut is placed between chains 3 and 4 to avoid the overlap. The result of this iteration was that all 8 chains where assigned to the sequence and locked, meaning that the internal correlation is set to one. No other components can be assigned whitin these chains.

In iteration 3 the last 4 chains had a length of 3 components except one that had a length of 2 components. The final sliding step positioned two of these chains with low RMSD. The remaining two had a length of 2 and 3 and could then be assigned because of their different length in the sequence. The final result of the correlation calculation and the sliding positioned all chains in place and a final peak picking could take place, giving almost complete backbone assignment. The analysis of Azurin showed the importance of a quality criteria giving the user an option to chose to either discard or keep a calculated component. Components with partly wrong shapes can give correlations that are wrong but with a high correlation making these hard to resolve and creating problems in the later sliding procedure. While most components could be used without any further refinement some had to be recalculated after visual inspection to improve the individual shapes. The correlation for some of the difficult shapes was close to 20%, showing the importance to have good and correct shapes for the analysis. The next

protein, Histone, had been measured with a temperature of 298 kelvin which was ten kelvin over the recommended measurement temperature. This condition resulted in shift degeneracy for several peaks reducing the detectable peaks to 74 in a ^{15}N-HSQC. Correlation and sliding iterations of the 74 components that resulted from the decomposition resulted in a final sequence assignment of 7 residue chains. Between these were 7 gaps: 2, 14-16, 24-25, 42-43, 47-48, 67-68 and 81-83 from the missing components. Still 97% of the rest of the observable residues could be assigned. The last protein, MMP20 proved to be the most difficult protein to resolve. With missing residues from convectional NMR experiments several peaks had to be excluded from the analysis. 102 components could be extracted from a selected 127 peaks from a ^{15}N-HSQC. This created gaps that made the assignment difficult. Nevertheless three component chains with a unique Cβ pattern could be identified in the whole sequence. A restricted region of residues 11-60 was also tested resulting in a near complete sequential assignment. The result of this study proved that PRODECOMP-SHABBA is a versatile tool for automated backbone assignment.

In paper 6 results from two types of 4D projection NOESY experiments, ^{15}N-HSQC-NOESY-^{15}N-HSQC and ^{13}C-HSQC-NOESY-^{15}N-HSQC, on histone where used two verify that enough distance information was contained in these two experiment. This information was then used for a structure calculation that was compared to the published structure. When examining projections from both 4D NOSEY experiments on histone it was clear that these two experiments alone could not resolve unique components from only N and NH shifts, especially with a high degree of degraded peaks in histone. Therefore the projections from the two

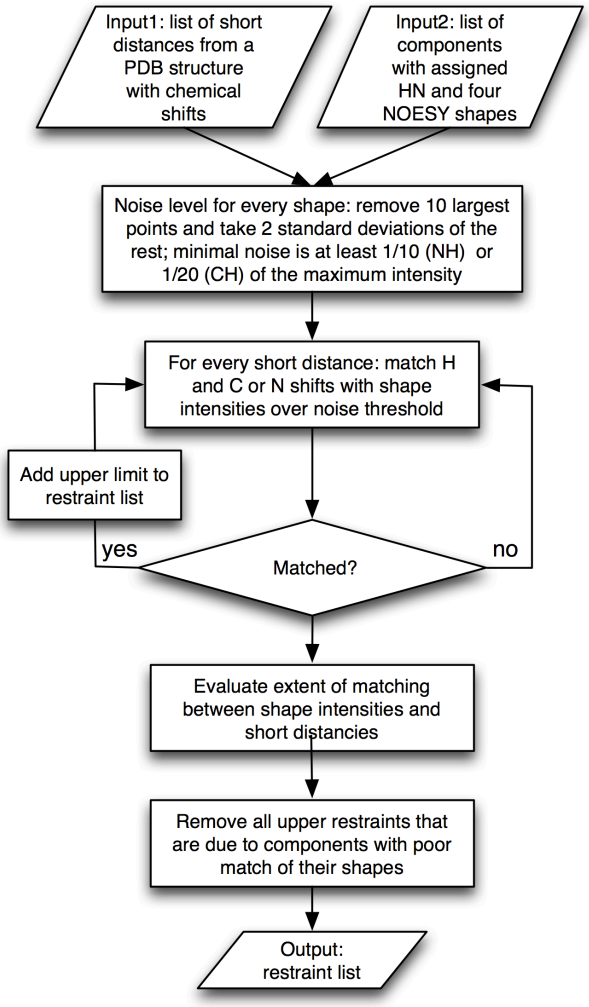

Figure 13: *Algorithm for calculating a restraint list from an input of NOESY shapes and an assigned PDB list.*

NOESY experiments where used together with projection from the two backbone experiments, 5D HBHACBCACONH and 4D HBHACBCANH, used in paper 5 for additional support of the decomposition. The decomposition used also the same 74 intervals that where used in the previous backbone study. A selection of 60 components had to be chosen because of severe overlap. The rest of the intervals considered of flexible N and C terminus ends that had a lot of overlap. Figure 13 describes the algorithm used for calculating an upl list that was used in cyana for a structure calculation.

The input list consist of selected intervals from PRODECOMP containing NOESY shapes and a distance and a reference list. The first step was to assign all PDB distances less that 4.5 Å. A noise threshold was determined for every shape. The next step was to match PDB distances with both HN_{noesy} and HC_{noesy} shapes and the corresponding heavy atom shapes. For the matching, residues 44-116 was used. Those shapes that had less than 50% assignment where then discarded. This test was to remove components that had mostly noise in them. The remaining distances was then written to a distance list. 80% of the points in the HC_{nosey} shapes could be matched to a short distance while 94% of the points in the HN_{noesy} shapes could be matched against a short distance. 64% of the HC short distances could be matched to non noise intensities in both the HC_{nosey} and C_{nosey} shapes, the corresponding for HN short distances was 85%. The upl list

was then used for a CYANA[57] structure calculation using residues 44-116. 200 calculations where performed. Residues 44-49 formed a disordered region. The mean structure compared to the 10 calculated structures had a RMSD of 0.7 Å for residues 50-116. The RMSD for the mean calculated structure compared to the published mean structure was 0.9 Å. The result can be seen in figure 14. Here the blue structure is the calculated mean structure and the red structure is the published structure. Three α-helices and a two-stranded β-sheet are visible.

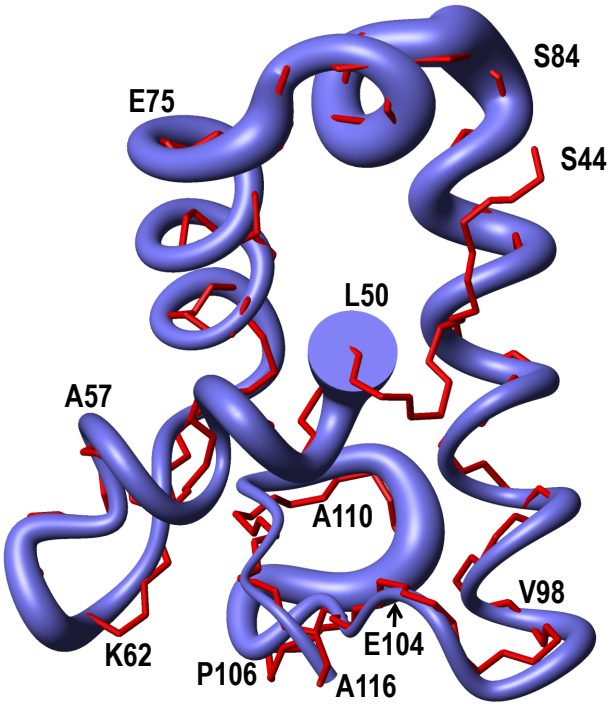

Figure 14: *Calculated structure of the histone domain in blue compared to published structure in red.*

Future improvements

One of the most common sources for wrong components in the interval list is when a strong peak partly overlaps a weaker one. When resolving the weak peak, the strong peak signal intensity will influence the weak peak giving contribution in the shapes. This can partly be avoided by increasing the number of components but it cannot completely solve the problem. Weak peaks have weak signals in all

projections and therefore are hard to resolve. Still, weak signals can give reasonable shapes that can be used for correlation and give correct result by selecting the right interval and the right number of components in the decomposition step. An algorithm for stepwise elimination of signals in the projections thereby eliminating overlapping peaks is an idea that have been discussed and partly implemented.

One of the highest priorities are to implement a quality check that gives information how well the component was resolved. An idea is to have a preliminary peak picker that checks number of peaks in each shape to give an estimation about the correctness of the selected interval. This could also be expanded to remove intervals that are to hard to resolve. These intervals could create problems in the later correlation step giving wrong correlations.

Another source for improvement is to reduce the number of parameters when selecting intervals and number of components for that interval. With 3 degrees of freedom the number of available options for an interval of 5 points with a maximum of 12 components for each interval are 60. Clearly, this is to many options for selecting a suitable interval and components. One approach that is in a developing stage is to reduce the number of points for selecting interval to one, thereby eliminating two degrees of freedom. Then the number of components could be resolved in an iterative manner. Another approach would be to start with a number of points and then reduce them iteratively and together with a peak picker select the interval that has the best result.

There is also need for improvement in the speed for the calculation to reduce the time for the calculation of all intervals. One option would be to replace some python functions with C functions using a cType interface between them, thus reducing memory consumption and cpu time. Replacement of in house implementations like fast-nnls with existing functions from different python libraries would also affect performance.

A future development could also be to set up a protocol for fast analysis and assignment of small protein for quick structure determination. This could increase the throughput in structure determination and drug discovery. The user interface could be ported to a web solution making the software suit operating system independent enabling more people to use it and remove the need for software installations. In conjunction with this could a database be connected for saving and handling different data. The following figure describes how a user could submit projection planes to a server that would iteratively calculate different assignment depending on type of experiment and display the result in the same web interface:

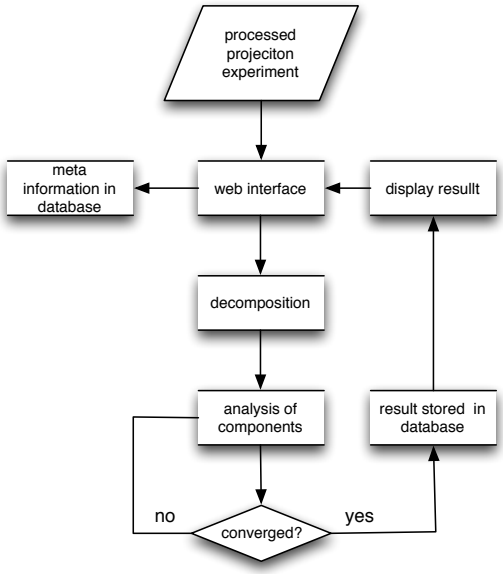

Figure 15: *Flowchart for a future web based projection analysis portal*

Another area of development would be to use experiments that can correlate more than Cα,Cβ and Hα,Hβ peaks between components. This would increase the correlation reliability making it more error prone and thus making the correlation more simple, opening up for a wider range of proteins that can be investigated.

Conclusions

This study showed that projection experiments decomposed with PRODECOMP and analyzed with SHABBA gives an almost complete backbone assignment for small to medium sized well behaved proteins. The algorithm is stable and the method is flexible in the sense that experiments can be custom designed to be used for different protein and there exists a range of different experiments that together can give full characterization information about the protein. A NOESY based study proved that projection experiments contain sufficient structural information to characterize the 3D fold of a protein.

Acknowledgments

Martin Billeter: supervisor with patience, intelligence and kindness. A very rare combination in this field

Daniel Farkas: colleague and friend, making this roller coaster project manageable

Wolfgang Bermel: super spectroscopist who kept the 600 alive while I tried to kill it

Moheb Nayeri: for playing the good cop in the gas lab interrogations

Lars ”LL” Lundvik: for proving that age is really just a number

References

1 M. Foster, C. McElroy, C. Amero, Solution NMR of Large Molecules and Assemblies, Biochemistry, 46 (2007) 331-340

2 K. Wüthrich, Nuclear Magnetic Resonance Spectroscopy of Proteins, Encyclopedia of Life Sciences, (2001) Wiley

3 G. Petsko, D. Ringe, Protein Structure and Function, (2004) New science Press Ltd

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

5 J. Drenth, J. Principles of protein X-ray crystallography, (1994) Springer

6 A. G. Tzakos, R. R. Christy, P. Grace, J. Lukavsky, R. Riek, NMR Techniques for Very Large Proteins and RNAs in Solution, Annu. Rev. Biophys. Biomol. Struct. 35 (2006) 319–342

7 D. Raftery, High-throughput NMR spectroscopy, Anal Bioanal Chem 378 (2004) 1403–1404

8 J. H. Prestegard, H. Valafar, J. Glushka, F. Tian, Nuclear Magnetic Resonance in the Era of Structural Genomics, Biochemistry, 40 (2001) 8677-8685

9 B. Simon, M. Sattler, Speeding Up Biomolecular NMR Spectrscopy, Angewandte Chemie International Edition, 43 (2004) 782-786 Wiley

10 J. Cavanagh, Protein NMR Spectroscopy, second edition. (2007) ELSEVIER

11 H. Oschkinat, C. Griesinger, P. Kraulis, O. Sørensen, R. Ernst, Three-dimensional NMR spectroscopy of a protein in solution, Nature 332 (1988) 374-376

12 M. Sattler, J. Schleucher, C. Griesinger, Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients, Progress in Nuclear Magnetic Resonance Spectroscopy 34 (1999) 93–158

13 G. Rule, T. Hitchens, Fundamentals of Protein NMR Spectroscopy, (2006) Springer

14 Y. Xia, G. Zhu, S. Veeraraghavan, X. Giao, (3,2)D GFT-NMR experiments for fast data collection from proteins, Journal Of Biomolecular NMR, 29 (2004) 467-476

15 E. Kupče, R. Freeman, Projection-Reconstruction Technique for Speeding up Multidimensional NMR Spectroscopy, J. Am. Chem. Soc. 126 (2004) 6429-6440

16 W. Gronwald, H. Kalbitzer, Automated structure determination of proteins by NMR spectroscopy, Progress in Nuclear Resonance Spectroscopy 44 (2004) 33-96

17 D. Malmodin, M. Billeter, High-throughput analysis of protein NMR spectra, Progress in Nuclear Magentic Resonance Spectroscopy 46 (2005) 109-129

18 D. Malmodin, M. Billeter, Robust and versatile interpretation of spectra with coupled evolution periods using multi-way decomposition, Magn. Reson. Chem. 44 (2006) S185–S195

19 M. Billeter, D. K. Staykova, Rapid Multidimensional NMR: Decomposition Methods and their Applications, Encyclopedia of Magnetic Resonance, 9 (2009) Wiley

20 V. Orekhov, V. Jaravine, Analysis of non-uniformly sampled spectra with Multi-Dimensional Decomposition, Progress in Nuclear Magnetic Resonance Spectroscopy Progr. NMR. Spect. (2011) (doi:10.1016/j.pnmrs.2011.02.002)

21 D. Marion, Fast acquisition of NMR spectra using Fourier transform of non-equispaced data, Journal of Biomolecular NMR 32 (2005) 141–150

22 L. Frydman, T. Scherf, A. Lupulescu, The acquisition of multidimensional NMR spectra within a single scan, Proc. Natl. Acad. Sci. USA 99 (2002) 15858

23 L. Frydman, A. Lupulescu, T. Scherf, Principles and Features of Single-Scan Two-Dimensional NMR Spectroscopy, J. Am. Chem. Soc. 125 (2003) 9204-9217

24 H. Eghbalnia, A. Bahrami, M. Tonelli, K. Hallenga, J. Markley, High-Resolution Iterative Frequency Identification for NMR as a General Strategy for Multidimensional Data Collection, J. Am. Chem. Soc. 127 (2005) 12528-12536

25 E. Kupče, R. Freeman, Fast reconstruction of four-dimensional NMR spectra from plane projections, J. Biomol. NMR 28 (2004) 391–395

26 E. Kupče, R. Freeman, Frequency-domain Hadamard spectroscopy, Journal of Magnetic Resonance 162 (2003) 158–165

27 S. Kim, T. Szyperski, GFT NMR, a New Approach To Rapidly Obtain Precise High-Dimensional NMR Spectral Information, J. Am. Chem. Soc. 125 (2003) 1385-1393

28 H. Atreya, T. Szyperski, G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignment, PNAS, 101 (2004) 9642-9647

29 G. Armstrong, B. Bendiak, The single basis filter diagonalization method: A rapid multidimensional data processing scheme, Journal of Magnetic Resonance 174 (2005) 163–170

30 X. Meng, B. Nguyenb, C. Ridgea, A. Shaka, Enhanced spectral resolution by high-dimensional NMR using the filter diagonalization method and “hidden” dimensions, J. Magn. Reson. 196 (2009) 12-22

31 S. Hiller, F. Fiorito, K. Wüthrich, G. Wider, Automated projection spectroscopy (APSY), PNAS, 102 (2005) 10876-10881

32 H. Gzyl, The Method of Maximum Entropy, World Scientific (1995) Singapore

33 Y. Li, T.M. Logan, A.S. Edison, A. Webb, Design of small volume HX and triple-resonance, Journal of Magnetic Resonance, 164 (2003) 128–135.

34 V. Jaravine, V. Orekhov, Targeted Acquisition for Real-Time NMR Spectroscopy, ‚J. Am. Chem. Soc. 128 (2006) 13421-13426

35 L. Wong, J. Masse, V. Jaravine, V. Orekhov, K. Pervushin, Automatic assignment of protein backbone resonances by direct spectrum inspection in targeted acquisition of NMR data, J. Biomol. NMR 42 (2008) 77–86

36 K. Kazimierczuk, J. Stanek, A. Zawadzka-Kazimierczuk, W. Kozminski, Random sampling in multidimensional NMR spectroscopy, Progress in Nuclear Magnetic Resonance Spectroscopy 57 (2010) 420–434

37 J. Stanek, W. Kozminski, Iterative algorithm of discrete Fourier transform for processing randomly sampled NMR data sets, J. Biomol. NMR 47 (2010) 65–77

38 A. Tal, B. Shapira, L. Frydman, Single-Scan 2D Hadamard NMR Spectroscopy, Angew. Chem. Int. Ed. 48 (2009) 2732–2736

39 K. Ding, A. Gronenborn, Novel 2D triple-resonance NMR experiments for sequential resonance assignments of proteins, J. Magn. Reson. 156 (2002) 262-268

40 T. Szyperski, G. Wider, J. Bushweller, K. Wiithrich, Reduced Dimensionality in Triple-Resonance NMR Experiments, J. Am. Chem. Soc. 115 (1993) 9307-9308

41 T. Szyperski, H. Atreya, Principles and applications of GFT projection NMR spectroscopy, Magn. Reson. Chem. 44 (2006) 51-61

42 S. Grzesiek, A. Bax, Amino-acid type determination in the sequential assignment procedure of uniformly C-13/N-15- enriched proteins. J Biomol NMR 3 (1993) 185–204

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

44 A. Hershko, A. Ciechanover, The Ubiquitin System, Annu. Rev. Biochem. 67 (1998) 425–79

45 T. Ali, P. Coles, T. Stevens, K. Stott, J. Thomas, Two Homologous Domains of Similar Structure but Different Stability in the Yeast Linker Histone, Hho1p, J. Mol. Biol. 338 (2004) 139–148

46 G. Anderson, J. Williamsll, R. Hille, The Purification and Characterization of Arsenite Oxidase from Alcaligenes faecalis, a Molybdenum-containing Hydroxylase, The Journal Of Biological Chemistry, 267 (1992) 23674-23682

47 Y. Arendt, L. Banci, I. Bertini, F. Cantini, R. Cozzi, R. Del Conte, L. Gonnelli, Catalytic domain of MMP20 (Enamelysin) – The NMR structure of a new matrix metalloproteinase, FEBS Letters 581 (2007) 4723–4726

48 M.H. Lewitt, Spin Dynamics Basics of Nuclear Magnetic Resonance, (2001) Wiley

49 D. Malmodin, M. Billeter, Multiway Decomposition of NMR Spectra with Coupled Evolution Periods, J. Am. Chem. Soc. Communications, 127 (2005) 13486-13487

50 A. Smilde, R. Bro, P. Geladi, Multi-way analysis, (2004) Wiley

51 V. Orekhov, I. Ibraghimov, M. Billeter, MUNIN: A new approach to multi-dimensional NMR spectra interpretation, J. Biomol. NMR. 20 (2001) 49-60

52 A.N. Tikhonov, A.A. Samarskij, Equations of mathematical physics, (1990) Dover

53 I. Ibraghimov, Application of the three-way decomposition for matrix compression, Numer. Linear Algebra Appl. 9 (2002) 551–565

54 E. Ulrich, H. Akutsu, J. Doreleijers, Y. Harano, Y. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. Schulte, D. Tolmie, R. Wenger, H. Yao, J. Markley, BioMagResBank, Nucleic Acids Research 36 (2007) D402-D408

55 J. Leckner, Ph.D. Dissertation, Chalmers University of Technology, (2001) Göteborg, Sweden

56 H. Nar, A. Messerschmidt, R. Huber, M.Van De Kamp, Crystal structure analysis of oxidized Pseudomonas aeruginosa azurin at pH 5.5 and pH 9.0. A pH-induced conformational transition involves a peptide bond flip. J. Mol. Biol. 221 (1991)765-772.

57 P. Güntert, Automated NMR protein structure calculation. Prog Nucl Magn Reson Spectrosc 43 (2003) 105–125