J Biomol NMR (2008) 42:87–97
DOI 10.1007/s10858-008-9265-z

ARTICLE

Assignment of protein NMR spectra based on projections, multi-way decomposition and a fast correlation approach

Doroteya K. Staykova · Jonas Fredriksson · Wolfgang Bermel · Martin Billeter

Received: 14 May 2008 / Accepted: 11 August 2008 / Published online: 6 September 2008

Abstract We present an approach for the assignment of protein NMR resonances that combines established and new concepts: (a) Based on published reduced dimensionality methods, two 5-dimensional experiments are proposed. (b) Multi-way decomposition (PRODECOMP) applied *simultaneously* to all acquired NMR spectra provides the assignment of resonance frequencies under conditions of very low signal-to-noise. (c) Each resulting component characterizes all spin ½ nuclei in a (doubly-labeled) CβH$_n$–CαH–C′–NH–CαH–CβH$_n$ fragment in an unambiguous manner, such that sequentially neighboring components have about four atoms in common. (d) A new routine (SHABBA) determines correlations for all component pairs based on the common nuclei; high correlation values yield sequential chains of a dozen or more components. (e) The potentially error-prone peak picking is delayed to the last step, where it helps to place the component chains within the protein sequence, and thus to achieve the final backbone assignment. The approach was validated by achieving complete backbone resonance assignments for ubiquitin.

Keywords Projection · Multi-way decomposition · Automated assignment · Algorithm

Electronic supplementary material The online version of this article (doi:10.1007/s10858-008-9265-z) contains supplementary material, which is available to authorized users.

D. K. Staykova · J. Fredriksson · M. Billeter (✉)
Biophysics Group, Department of Chemistry, University of Gothenburg, Box 462, 405 30 Gothenburg, Sweden
e-mail: martin.billeter@chem.gu.se

W. Bermel
Bruker BioSpin GmbH, Silberstreifen, 76287 Rheinstetten, Germany

Introduction

The most limiting aspect of solution NMR spectroscopy of biological systems concerns the complexity of NMR spectra and the difficulty of interpreting these in terms of resonance assignments. Spectral complexity can be overcome by increasing the number of nuclei involved in an experiment so that overlapped signals in a low-dimensional spectrum are spread out along an additional frequency dimension. However, unless the spectral resolution of observed nuclei is compromised time requirement for high dimensionality experiments becomes rapidly prohibitive. The experimental time demands can be shortened by introducing a linear dependence among the evolution times of selected nuclei, which are then observed along the same spectral dimension. Such a "reduced dimensionality" approach allows, for example, the characterization of four nuclei by a 3D data set (Szyperski et al. 1993). The combination of this type of experiments with a variety of analysis approaches has made the concept very popular: In the "GFT NMR" approach, sub-spectra originating from joint sampling of indirect dimensions are linearly recombined and analyzed (Kim and Szyperski 2003). In the "PR" (projection-reconstruction) method, the corresponding full-dimensional spectrum is reconstructed (Freeman and Kupče 2003). Signal peaks may be picked directly in each projection spectrum for further processing by combinatorial or statistical approaches: "PatternPicker" (Moseley et al. 2004), "APSY" (Hiller et al. 2005), "HIFI-NMR" (Eghbalnia et al. 2005). An overview and discussion of the literature of these (and related) tools can be found in a number of recent reviews (Malmodin and Billeter 2005a; Szyperski and Atreya 2006; Yoon et al. 2006). Because the work presented here concerns the analysis of the spectra resulting from any of the above experimental techniques,

but not the techniques themselves, a consideration of differences among the various ways to record such spectra is not relevant here. Therefore, and based on the common underlying projection theorem in Fourier transform (Folland 1992), the term "projection" will be used here to denote spectra in which two or more evolution periods evolve in a dependent manner (Hiller et al. 2005).

Each of the above spectroscopic approaches has been presented together with an interpretation of the resulting spectra (Kim and Szyperski 2003; Freeman and Kupče 2003; Moseley et al. 2004; Hiller et al. 2005; Eghbalnia et al. 2005). Nonetheless, a number of issues concerning the analysis of sets of projected spectra remain, e.g. with respect to optimizing the signal-to-noise ratio (S/N).

Here, we describe a new, comprehensive approach for the analysis of spectra recorded for backbone resonance assignment; the concept is illustrated by application to a small protein. Implementation of this concept within other contexts, e.g. for side chain assignment or extraction of structure information via experiments containing TOCSY or NOESY magnetization transfers, as well as applications to larger proteins is currently in progress. The approach consists of two steps: the earlier presented *"multiway decomposition of NMR spectra with coupled evolution periods"* (Malmodin and Billeter 2005b) and a novel algorithm for sequential assignment optimized for work with components resulting from the first step.

A general consequence of reduced dimensionality approaches is a massive reduction in measurement time, making the signal-to-noise ratio (S/N) a critical entity. In these types of experiments, the already low signal intensity is distributed over peaks in each of the individual projections. However, the positions of all peaks observed for a given spin system are correlated; thus the probability of weak candidate signals is increased if their positions are found to be consistent with related signals in other projections. The following simple consideration on a 5D experiment, represented by 30 2D projections, illustrates some of the implications. Let us assume 100 points in the indirect dimension, white noise only, an S/N ratio making signal intensities comparable to the largest 10% noise intensities, only one expected signal in each projection for every resonance in the acquisition dimension, and that each signal is exactly one point wide. For a non-overlapped resonance in the acquisition dimension, four projections with clear signals would be sufficient to determine the frequencies for the other four nuclei with shifts along the indirect dimension. Taking all combinations of points with sufficiently large intensities (10% of all points) in four projections yields 10^4 potential chemical shift combinations for the four nuclei. However, these four shifts must also correspond to points with sufficiently large intensities in all other 26 projections. Therefore in this thought experiment, the chance of finding a combination of noise intensities yielding a false positive in all 26 spectra is 10^{-22}. Furthermore it has been shown that decompositions of real projections can unambiguously detect true signals even with an S/N close to one, provided that all projections are considered simultaneously (Malmodin and Billeter 2005b, 2006). Thus, our solution to the S/N problem is to always consider all projections simultaneously and to postpone steps such as peak picking until the end of the analysis, i.e. following the sequential assignment.

Recording linear combinations of chemical shifts may require either choosing a large spectral width in the indirect dimension or some post-processing (typically, identification of aliased peaks followed by their unfolding). The former choice may negatively affect the resolution, measured as data points per ppm, for all nuclei involved in the linear combination. On the other hand, the number of peaks in these spectra is often related to the number of amino acid residues in the protein. Thus, choosing the spectral width in the projected dimension to correspond to those used for the individual nuclei rather than to the combination of their frequencies would often yield a peak density comparable to that in the ^{15}N-HSQC in spite of folding. Decompositions can be designed to accept this type of aliasing in the input projections, and still deliver correct chemical shift information without any special treatment of the folded peaks during or after the calculations; this may be helpful, for example, when reconstructing spectra (Malmodin and Billeter 2006).

As shown below, each component resulting from multiway decomposition contains complete information about the set of all (about 10) nuclei of the corresponding protein fragment, unambiguously characterizing all its chemical shifts. The second, novel step, performing the sequential assignment, is designed to make full use of this information. It is non-iterative, which makes it both robust and fast, requiring a few seconds only.

The goals set for the approach and its application to backbone assignment of proteins presented below include: reliable characterization of all spin ½ nuclei in (doubly-labeled) $C\beta H_n$–$C\alpha$H–C′–NH–$C\alpha$H–$C\beta H_n$ fragments as 9-dimensional components derived from simultaneous decomposition of all input spectra; optimal treatment of data with respect to S/N and resolution; and efficiency and robustness combined with user friendliness.

Methods

NMR spectroscopy

Using a 600 MHz proton resonance frequency spectrometer with a triple resonance (HCN) z-gradient room

temperature probe, two experiments were performed on a 2 mM solution of ubiquitin in 90%H_2O/10%D_2O at 303 K, pH 6.6. The first was a reduced dimensionality version of a 5D HBHACBCACONH (Grzesiek and Bax 1993; Muhandiram and Kay 1994; Szyperski et al. 2002; Atreya and Szyperski 2004; Szyperski and Atreya 2006), where the magnetization is transferred according to:

$$\mathrm{H}\beta\mathrm{H}\alpha\,(t_1) \rightarrow \mathrm{C}\beta\mathrm{C}\alpha\,(t_2) \rightarrow \mathrm{C}\alpha \rightarrow \mathrm{C'}\,(t_3) \rightarrow \mathrm{N}\,(t_4) \rightarrow \mathrm{HN}\,(t_5),$$

yielding for each HN resonance frequency chemical shift information on α- and β-nuclei of the preceding residue. Indirect evolution periods are implemented as constant time schemes (Bax et al. 1979; Bax and Freeman 1981; Powers et al. 1991), except for a semi-constant time scheme for the t_1 evolution period (Grzesiek et al. 1993; Grzesiek and Bax 1993; Logan et al. 1993). Several data sets were recorded by introducing linear dependences between selected evolution periods. For each of the corresponding indirect dimensions, a complex data point was recorded with a 90° phase shift of the pulse prior to the evolution period. For n indirect dimensions, this results in 2^n points per delay setting. Linear combinations were calculated between 2^{n-1} data points resulting in 2^{n-1} separate sub-spectra.

The second experiment was a modified version of the complementary HBHACBCANH experiment (Grzesiek and Bax 1992). The modification consisted of an additional N to C′ transfer leading to a magnetization transfer as follows:

$$\mathrm{H}\beta\mathrm{H}\alpha\,(t_1) \rightarrow \mathrm{C}\beta\mathrm{C}\alpha\,(t_2) \rightarrow \mathrm{C}\alpha \rightarrow \mathrm{N} \rightarrow \mathrm{C'}(t_3) \rightarrow \mathrm{N}(t_4) \rightarrow \mathrm{HN}\,(t_5),$$

showing for each HN resonance frequency chemical shift information on α- and β-nuclei of both a given residue and its preceding residue. With the extra C′ chemical shift labeling (implemented as a real time evolution) in the second experiment, chemical shifts of three nuclei are common to both experiments (C′, N, HN). All evolution periods in the experiments are either not sampled, or sampled between $t = 0$ and $t = t_{max}$, where t_{max} is the maximal evolution time predefined for a given nucleus; sampling the first two evolution periods means that $t2 = t1 * t2_{max}/t1_{max}$ corresponding to a projection angle of 45°. In general only angles of 0, 90 or ±45° were used. All two-dimensional sub-spectra consist of 60 complex data points in the indirect dimension; zero-filling was applied to double the number of points in both dimensions. Inspection of the first spectrum, which corresponds to the ^{15}N-HSQC, indicated the necessity for a constant phase correction in the direct dimension; identical corrections were subsequently applied to all spectra. (See Supplementary Material for more experimental details.)

Multi-way decomposition

The following model applies to 2-dimensional projections P_m (m enumerates the projections) of a N-dimensional spectrum S (Malmodin and Billeter 2005b, 2006):

$$P_m(\omega,\omega_N) = \sum_k \left[(F_1^k * \ldots F_{N-1}^k)(\omega) \otimes F_N^k(\omega_N)\right] + \varepsilon \qquad (1)$$

where the symbol "*" refers to the convolution operation and ⊗ is a direct product between one-dimensional vectors. Each term k in the sum is called a component that is described by the one-dimensional vectors F_i, i = 1…N, referred to in the following as *shapes* (Orekhov et al. 2001). The residual ε collects data that do not follow this assumption (typically noise). The coordinates in the projections are usually ω_{HN} in the direct dimension, and ω, which represents for each projection a different linear combination of frequencies. Decomposition in this case means determining all shapes F_i such that the sum on the right side of Eq. 1 optimally approximates the experimental input of projections P_m on the left side, thus minimizing the residual ε. The shapes F_i characterize the resonances of all nuclei involved; they may also be used to reconstruct the full-dimensional spectrum S.

PRODECOMP (Malmodin and Billeter 2005b, 2006) performs a simultaneous decomposition on all projections for a selected interval along the direct dimension. Each resulting component corresponds to one resonance in the direct dimension, i.e. one HN nucleus, and all nuclei connected to it by magnetization transfer according to the pulse sequence. It consists of shapes that resemble one-dimensional spectra with one or a few signals. In the present application, the full-dimensional spectrum S would have nine dimensions. Therefore, shapes are given for the following nine nuclei or groups of nuclei: one shape for HN along the direct dimension with a width corresponding to the selected interval, and individual shapes for N, C′($i-1$), Cα/β($i-1$), Hα/β($i-1$), Cα, Cβ, Hα and Hβ, each with a width identical to the full width along the indirect dimension of the input projections. Nuclei of the residue preceding the HN are indicated by "($i-1$)"; for better readability the indication "(i)" for nuclei of the same residue as HN is omitted. Shapes exhibit between one or a few signals, which provide the resonance frequencies of the nuclei involved, and for each component an individual set of shapes is obtained. The shapes for Cα/β($i-1$) and Hα/β($i-1$) exhibit individual signals for both the α and β nuclei; similarly two signals may be observed for $-CH_2$ moieties. Often, the Cα, Cβ, Hα and Hβ shapes show

somewhat weaker signals for the corresponding nuclei of residue $i - 1$.

All experimental projections P_m were collected in the form of planes of a single 3D matrix. The input for PRODECOMP consists of a file defining the individual projections, i.e. describing the linear combinations of chemical shifts appearing along the indirect dimension, the above mentioned 3D matrix with all these spectra, the selection of an interval along the direct dimension in data points, the number of resonances (number of components) expected in this interval, a regularization factor and the number of iterations to be performed. Tykhonov-type regularization (Tykhonov and Samarskij 1990) was used to avoid the occurrence of components with highly differing amplitudes, although this type of problem was not observed in the present applications. An earlier systematic study showed the influence of moderate regularization on the resulting components is very limited (Luan et al. 2005). (See Supplementary Material for more details.)

Sequence-specific assignment

A novel algorithm named SHABBA (SHape Analysis for BackBone Assignments) was designed for the assignment of the backbone resonances on the basis of the shapes derived from the decomposition. The approach consists of several steps described below, each of which is implemented as a Python routine. The first step is interactive; all others are executed without user intervention. Step 3 requires the input of one parameter, although other parameters may be changed, the user is not expected to do so. The following short description outlines the concept of SHABBA (more details are given in Supplementary Material) (1) First, a visual inspection is performed on all resulting components with the help of a component plotting routine. At this stage, components describing fragments centered on the same HN moiety are sorted out; this occurs when overlapping intervals are selected along the direct dimension. (2) Components describing glycine residues are identified on the basis of the missing signals for Cβ and Hβ. This automatic step also corrects for the fact that glycine α-nuclei appear in the shapes for β-nuclei. (3) Comparisons are made among all pairs of components by calculating correlations between the shapes involving the α- and β-nuclei of the same residue as the HN from one component with the shapes involving the α- and β-nuclei of the residue preceding the HN from the other component. High correlations indicate that the two components characterize neighboring $C\beta H_n$–CαH–C′–NH–CαH–$C\beta H_n$ fragments with a $C\beta H_n$–CαH moiety in common. When starting the routine, the user has to provide the only parameter needed in SHABBA: a minimal factor by which a correlation has to exceed alternative correlations to be accepted. This step yields *chains* of sequentially ordered components that are interrupted when no correlation is sufficiently larger than all alternatives; this occurs, for example, when encountering a proline, but may also happen when the input is insufficiently unique because of frequency overlap, low S/N or other problems. (4) For each component, the chemical shift of Cβ is determined by peak picking in the one-dimensional shapes for Cβ; similarly the Cα chemical shifts are determined. (5) Parallel to this, a list of Cβ and Cα chemical shifts is prepared for all residues in the protein sequence with the statistically expected values according to amino acid type as reported in the BioMagResBank (Seavey et al. 1991). A SHABBA routine then "glides" the sequence of the chemical shifts for each component chain along the protein sequence, calculating for each position the RMSD between the shifts extracted from the components and expected in the protein sequence. Each chain of components, starting with the longest one, is placed in the protein sequence where it best matches as indicated by the lowest RMSD. (6) As a last step, all shapes are peak picked in order to get a complete list of backbone chemical shifts. The 1D peak picker used is the same as in step 4; it needs no parameters, however it is aware of how many peaks are expected in different shape types. Input to steps 1–4 and 6 are the shapes from the decomposition; step 5 works on database information and the output of step 4.

Results

The resonance assignment illustrated below on spectra recorded for ubiquitin consists of two major steps (Fig. 1). Multi-way decomposition by PRODECOMP (Malmodin and Billeter 2005b, 2006) represents a general approach applicable also to other sets of projections (e.g. for side chain assignments or extraction of structural data). The second step, SHABBA (SHape Analysis for BackBone Assignments), is specific for the sequential assignment of the protein backbone. However, it can be used to analyze components resulting from the decomposition of projections from backbone experiments that differ from the 5D pulse sequences described in Methods (for example from the experiment presented in Hiller et al. 2007).

Input projections

The input projections describe a network of spins connected by magnetization transfers, or spin systems, that each starts from an HN nucleus and involves all nuclei of a $C\beta H_n$–CαH–C′–NH–CαH–$C\beta H_n$ fragment surrounding the initial HN. Therefore, the decomposition yields components that characterize large spin systems with up to 13 nuclei as illustrated in Fig. 2. It is important to note that the complete spin systems characterized by parallelograms in

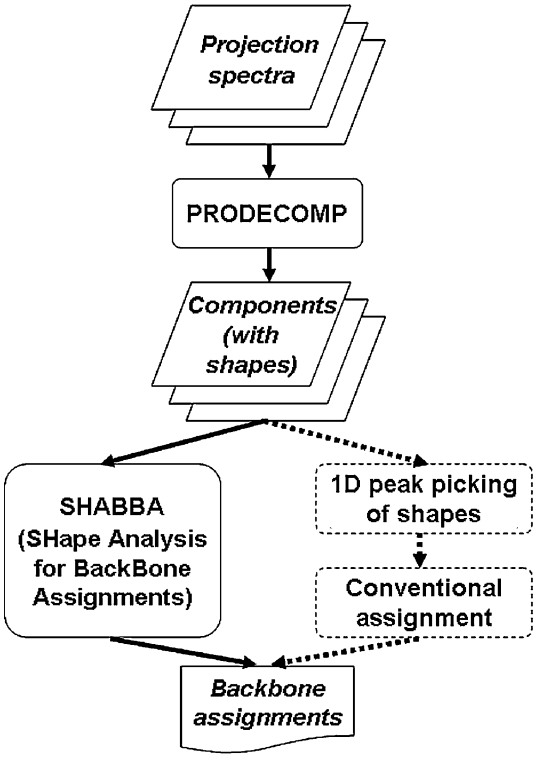

Fig. 1 Flow-chart of the resonance assignment procedure based on multi-way decomposition with PRODECOMP and analysis of the resulting components and shapes for sequential assignment of backbone and CβH$_n$ groups with SHABBA. (Note that for the assignment of side chain chemical shifts and the extraction of structural data, PRODECOMP can be used as is.)

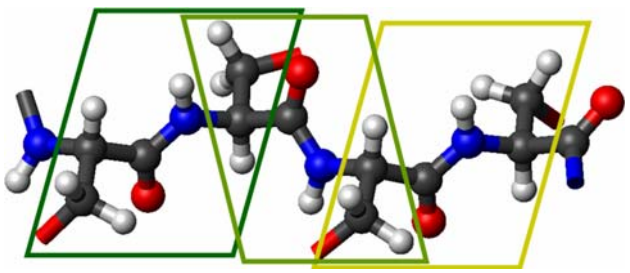

Fig. 2 Protein fragment with the backbone and CβH$_n$ groups of four residues. The parallelograms define individual spin systems described as single components in the output of PRODECOMP. Overlap of the rectangles contains nuclei that appear in two components and can be used to make sequential assignments

Table 1 Overview of input spectra with corresponding projection schemes[a]

Input spectra[b]	N[c]	C′ (i − 1)	Cα/β (i − 1)	Hα/β (i − 1)	Cα	Cβ	Hα	Hβ
1	+							
2		+						
3–4	+	+/−						
5–8	+	+/−	+/−					
9–10	+		+/−					
11–14	+	+/−		+/−				
15–18;	+	+/−			+/−			
31–34	+	+/−				+/−		
19–20;	+				+/−			
35–36	+					+/−		
21–24;	+	+/−					+/−	
37–40	+	+/−						+/−
25–26;	+						+/−	
41–42	+							+/−
27–28;		+					+/−	
43–44		+						+/−
29–30;		+			+/−			
45–46		+				+/−		

[a] The "+" and "−"signs in the table define the linear combination of chemical shifts (corresponding to a projection) along the indirect dimension. For example, spectrum 1 has only ω_N along this dimension, spectrum 2 has only $\omega_{C'(i-1)}$, spectrum 3 has $\omega_N + \omega_{C'(i-1)}$ etc. The table entry "+/− "indicates that two projections are available, one for each sign. Thus, the indirect dimension in spectrum 3 is $\omega_N + \omega_{C'(i-1)}$ while for spectrum 4 it is $\omega_N - \omega_{C'(i-1)}$

[b] Input spectra are enumerated from 1–46, where spectra 31–46 are copies of spectra 15–30 after sign inversion for all data points. Because only positive intensities are considered by PRODECOMP, only signals involving Cα or Hα nuclei are used from spectra 15–30, while only signals involving Cβ or Hβ nuclei are used from spectra 31–46 (see text)

[c] The column headings correspond to the shape definitions of the decomposition. Shapes are identified by nuclei names, but the shape axes are frequencies; thus "N" stands for a shape with ω_N along its axis. "$i - 1$" refers to nuclei of the residue preceding the HN, whose shift is recorded along the direct dimension. Nuclei without such a label belong to the same residue as the HN. Because no immediate distinction is possible between Cα and Cβ signals of residue $i - 1$, these appear in the same shape "Cα/$\beta(i - 1)$", and similar for Hα and Hβ of residue $i - 1$. For residue i this distinction is possible due to the different signs of the signals in the spectra, and therefore individual shapes are assigned

Fig. 2 are optimized in single calculations that use *all* input projections simultaneously. This yields a tight relation among all chemical shifts of the nuclei in a fragment.

In total, 30 projections were recorded in two sets of experiments using conventional instrumentation. Table 1 provides an overview of the input spectra by characterizing the linear combinations of nuclear shifts in each projection. A total of 14 projections were used from the first experiment, and 16 projections from the second (see Methods). Signals in the spectra 15–30, which involve the resonance frequencies of the CαH and CβH$_n$ groups from the same residue as the HN nucleus, have positive or negative intensity, depending on whether CαH or CβH$_n$ is involved. Because the current version of PRODECOMP ignores negative intensity in the input spectra, it considers only information regarding CαH from these spectra. Input of the same spectra after inversion of all signs provides information on CβH$_n$ to PRODECOMP; these "inverted"

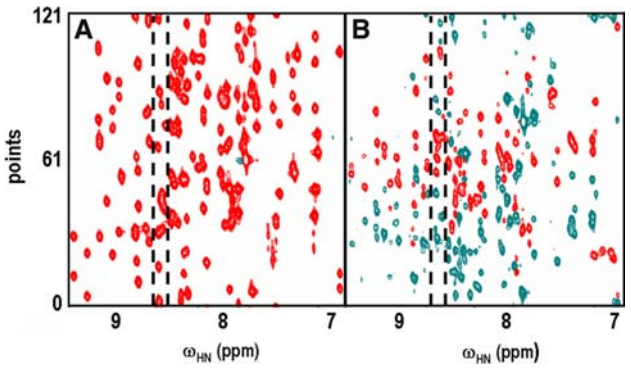

Fig. 3 Two of the 46 spectra representing the input for the PRODECOMP decomposition. (**a**) Spectrum with the linear combination ω_N–$\omega_{C'(i-1)} + \omega_{H\alpha/\beta(i-1)}$ along the indirect dimension (spectrum 12 in Table 1). (**b**) Spectrum with the linear combination $\omega_N + \omega_{C'(i-1)}$–$\omega_{C\alpha/\beta}$ along the indirect dimension (spectrum 17 in Table 1). Red contour lines indicate positive intensity, blue contour lines negative intensity. The dashed lines outline a region discussed in the text and Fig. 4. Note that these spectra contain more than one peak per component

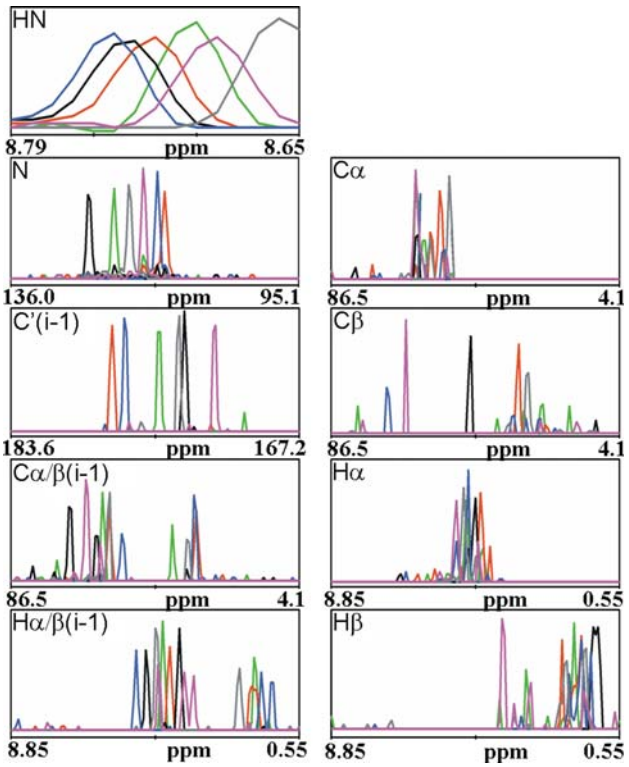

Fig. 4 PRODECOMP output for the decomposition of the region indicated in Fig. 3 by dashed lines. For each of the six components, which were needed for this region, the shapes are given in a different color. Referring to the final assignment, the color code is blue for Thr 7, green for Thr 14, black for Leu 15, grey for Glu 18, red for Glu 34 and magenta for Thr 66. Shapes are given for the nuclei or nuclei groups listed on top of Table 1 as well as for HN. For better clarity, the shapes for Glu 18 are also given in Fig. 5 omitting the shapes of the other components. Intensity units along the vertical axes are arbitrary

projections are numbered 31–46 in Table 1. As a result, chemical shifts for CαH nuclei can be discriminated from those of the $C\beta H_n$ nuclei. This is not the case for the CαH and $C\beta H_n$ groups of the residue preceding the HN nucleus, whose resonance frequencies are recorded in spectra 5–14 (Table 1). Figure 3 shows two of the input projections with the following linear combinations of chemical shifts along the indirect dimension: ω_N–$\omega_{C'(i-1)} + \omega_{H\alpha/\beta(i-1)}$ corresponding to spectrum 12 in Table 1 (Fig. 3a), and $\omega_N + \omega_{C'(i-1)} - \omega_{C\alpha/\beta}$ corresponding to spectrum 17 in Table 1 (Fig. 3b).

PRODECOMP decompositions

Each individual decomposition calculation is thus based on the simultaneous use of 46 input spectra (spectra 1–30 as is plus spectra 15–30 after sign inversion). Because of the joint appearance of $C\alpha(i-1)$ and $C\beta(i-1)$ nuclei, and $H\alpha(i-1)$ and $H\beta(i-1)$ nuclei, the 13 nuclei depicted in Fig. 2 determine eight shapes in the resulting components corresponding to eight types of nuclei listed as column headings in Table 1, plus a shape for HN (Figs. 4 and 5). These shapes are optimized during the decomposition to fit *all* input spectra simultaneously; for example, the two shapes for N and $C'(i-1)$ in a component must be consistent with 37, respectively 35, of the spectra listed in Table 1.

PRODECOMP was applied to a series of intervals along the ω_{HN} axis. These intervals were selected based on the extent of separation and overlap of the HN chemical shifts. They include between one and seven HN chemical shifts with a spectral width of up to 0.2 ppm. In total, 26 different decomposition calculations were run yielding 72 spin systems. As an example, the location of an interval in a crowded region comprising six HN chemical shifts, and therefore yielding six components, is indicated in the two spectra of Fig. 3. The shapes of the resulting components for this interval are shown in Fig. 4, where different colors stand for different components, i.e. spin systems. The first panel shows the shapes along the directly detected dimension ω_{HN}; its width corresponds to the width of the interval chosen, i.e. 14 data points or slightly over 0.1 ppm. The following eight panels show the shapes for the nuclei involved in the indirect dimension of the input spectra in the order listed on top of Table 1, and their width is the same as the spectral width along this dimension. Panels 2 and 3 with shapes for N and $C'(i-1)$ contain as expected exactly one signal per spin system; residual noise is significantly smaller than the signals. Therefore these shapes unambiguously characterize the chemical shifts for the nuclei N and $C'(i-1)$. The next two panels contain the resonances of $C\alpha(i-1)$ and $C\beta(i-1)$, respectively $H\alpha(i-1)$ and $H\beta(i-1)$, i.e. of nuclei in the residue preceding the HN of panel 1. Depending on the type of

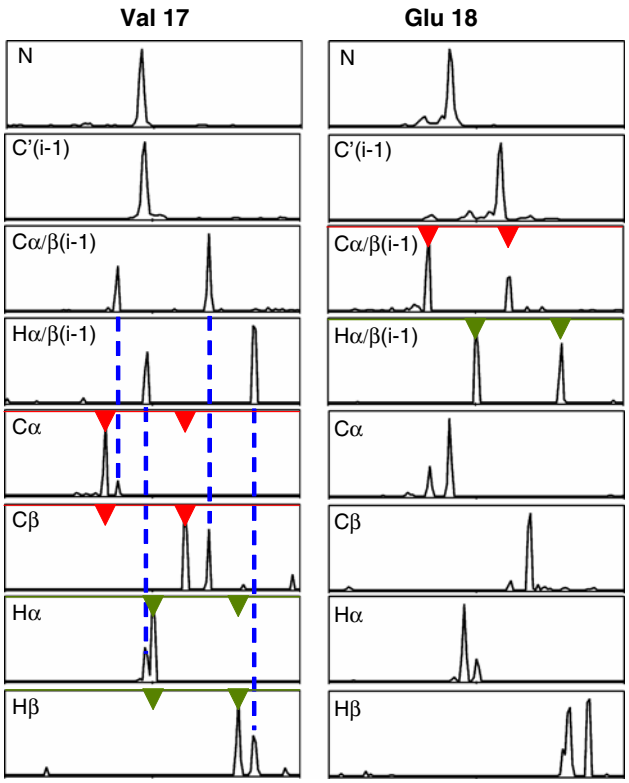

Fig. 5 Output shapes from PRODECOMP for two components describing residues that are adjacent in the protein sequence (shapes for ω_{HN} are not shown). The right component for Glu 18 is the same as the grey component in Fig. 4. The blue dashed lines indicate the presence of, typically weaker, peaks for the $i - 1$ nuclei in the shapes for the i nuclei. The red arrows on the right point at the peaks for C$\alpha(i - 1)$ and C$\beta(i - 1)$ in the component for Glu 18. For the left component the same positions are marked in the shapes for Cα and Cβ again with red arrows illustrating the correlations between the two components; similar correlations are identified with green arrows for α- and β-hydrogens. These correlations allow unambiguous identification of the left component as being succeeded by the right component. Multiple application leads eventually to complete sequential assignment (see text). See Fig. 4 for frequency units for all shapes

residue $i - 1$, one or two, respectively two to three, signals are expected, and indeed observed. The remaining four panels provide shapes for the nuclei Cα, Cβ, Hα and Hβ of the residue of the HN atom. These shapes often contain also signals for the nuclei of the same type but from residue $i - 1$. The green component (later assigned to Thr 14) exhibits a shape pattern typical for pure noise in the Cβ panel, i.e. it misses the Cβ resonance; this is the only instance where a resonance for a nucleus other than Hβ is missing. Because this Cβ frequency is also given by the component for residue 15 in the shape with C$\alpha(i - 1)$ and C$\beta(i - 1)$, this does not impede the further steps nor does it prevent a complete resonance assignment. Figure 5 (right panels) provides a better resolved view for the component with grey shapes in Fig. 4 (later assigned to Glu 18).

Assignment with SHABBA

After interactively sorting out of components that were found twice owing to overlapping intervals along the direct dimension, SHABBA identified 72 components for ubiquitin. This is the same as the number expected (76 residues minus 3 prolines minus the N-terminal residue). Seven components lack observed β-nuclei (only noise in the corresponding shapes) and were assigned to glycines. However, since ubiquitin has only six glycines, one of these glycine components must represent a conformational or structural inhomogeneity (see below). In addition, it turned out that Glu 24 could not be assigned, because no corresponding component could be found. Visual inspection of the input spectra showed that the regions, where the signals would be expected according to chemical shift information from the BioMagResBank (Seavey et al. 1991), were empty.

The sequential connection of components relies on the fact that each component provides shapes for the CαH-CβH$_n$ moiety of two neighboring residues (Fig. 2). Thus, the shapes for C$\alpha/\beta(i - 1)$ and H$\alpha/\beta(i - 1)$ (see red and green arrows in Fig. 5) of one component should correlate with the ones for Cα, Cβ, Hα and Hβ of another component, if the latter describes a residue preceding the one characterized by the former component. Correlations were calculated among all pairs of components and tabulated. High correlations were identified by removing entries in this table that were smaller than 0.75 times the largest entry on the same row or column; this factor represents the only required parameter from the user when making assignments with SHABBA, and any value between 0.67 and 1.0 will result in a correct assignment of ubiquitin. A portion of the full $72 \cdot 72$ table is given as Table 2; it contains correlations (represented as percentages) for the components that were assigned to residues 2–18. The remaining entries provide a clear sequential connection, yielding a *chain* with 17 components. Even for a correlation in Table 2 with values around 50%, all "competitors" on the same row or column are smaller than 0.75 times this correlation. They reflect situations where all correlations to a certain component are reduced: not only the correct one but also its "competitors". In Table 2 small values result from missing resonances for β-nuclei of threonines (Hβ of Thr 7, 9, 12, and Cβ of Thr 14). Automated inspection of the complete table with all 72 spin systems yielded five chains of components plus a single glycine component. These chains are delimited by prolines, the chain termini as well as residue 24, which is missing in the spectra (see above). The additionally observed component, which resembles a glycine, exhibited a strong correlation to the component for Arg 74. This spin system thus competes with that for Gly 75, and it is the cause for disrupting the chain of components after Arg 74. Signals for this spin system were present in all projections corresponding to

Table 2 Final correlations (in %) between all pairs of components corresponding to the spin systems for residues 2–18[a]

k\k – 1	13	5	8	2	17	6	7	14	15	12	18	4	3	16	10	9	11
13										59							
5												92					
8							48										
2																	
17														72			
6		74															
7						83											
14	83																
15								50									
12																	92
18					90												
4													79				
3				80													
14									92								
10																41	
9			91														
11															98		

[a] Correlations are listed following processing of the initial correlation table. Components are sorted according to their ω_{HN}. Entries were calculated for the entire table with 72 components, but only those corresponding to residues 2–18 are shown. Correlations are between the shapes for Cα, Cβ, Hα and Hβ for component $k-1$, listed on top of the table, and the shapes for C$\alpha/\beta(i-1)$ and H$\alpha/\beta(i-1)$ for component k, listed to the left of the table. For orientation purposes only, the lists on top and to the left of the table correspond to the residue numbers in the final sequence-specific assignment

frequencies $\omega_{HN} = 7.95$ ppm, $\omega_N = 115.9$ ppm, and independent measurements showed that the sample was partially degraded. The chains of components also provide chemical shifts of the α and β-nuclei of the spin systems immediately preceding each chain, e.g. for the first residue, the missing residue 24 or the prolines. This allows also filling in values for shifts that were missing in the corresponding i shapes (as for Thr 14).

In the next step, the chemical shifts of the Cα and Cβ nuclei are listed for each chain of components. Similarly, a list of Cα and Cβ chemical shifts is prepared for all residues in the protein sequence with the statistically expected values according to amino acid type as reported in the BioMagResBank (Seavey et al. 1991). For all possible alignments of each chain of components to the protein sequence, SHABBA then calculates the RMSD between the component shifts and the statistically expected shifts. This systematic comparison is illustrated in Fig. 6a, where the RMSD values for all alignments are plotted with different colors for different chains of components. Each point on a curve defines a start point for the chain and indicates how well the chain matches the expected shifts with this starting point. Points with an RMSD below 10 ppm are shown as squares. For all chains, these squares indicate the best fit and at the same time the correct starting point for each chain. For this assignment, the correlation coefficient between the observed Cβ chemical shifts (Fig. 6b) and the expected chemical shifts according to the BioMagResBank is 0.99; for the Cα chemical shifts the corresponding correlation coefficient is 0.90.

As the very last step, one-dimensional peak picking of all shapes provides a complete table of chemical shifts, representing a near-complete resonance assignment of the backbone and the CβH$_n$ groups. Besides interruptions at prolines, this assignment misses chemical shifts for N and HN of residue 24, whose resonances could not be found in the spectra. Additionally, the chemical shifts of Hβ of Thr 7 and Cα and C$'$ of Ile 23 are missing. The correctness of the backbone assignment resulting from use of PRODECOMP and SHABBA is unambiguously demonstrated by statistical tests that compare it to independently obtained assignments, in particular with the chemical shifts for ubiquitin deposited as entries 6457 and 6466 at the BioMagResBank. (More details about the use of PRODECOMP and SHABBA as well as the statistical tests for the correctness of the assignments are given in Supplementary Material.)

Discussion and conclusions

As pointed out earlier (Malmodin and Billeter 2006), multiway decomposition profits from several advantages. (1) It is based on a mathematical model that can directly be related to NMR theory. (2) By interpreting all input spectra

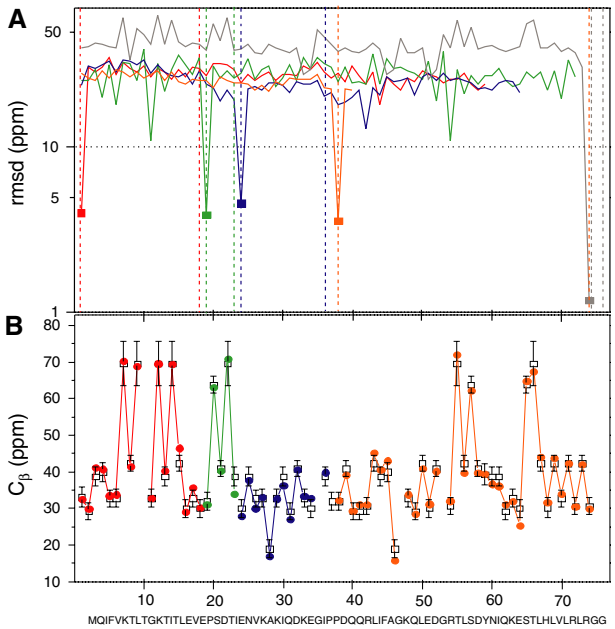

Fig. 6 Placement in the protein sequence of five chains of component (color coded in red, green, blue, orange and grey) from correlation calculations using Cα- and Cβ- chemical shifts. (**a**) The different curves show for each chain the RMSD between observed and statistically expected (according to residue type) Cα- and Cβ-chemical shifts, plotted in a logarithmic scale, for all possible positions of the chain in the protein sequence. Values below 10 are highlighted by squares. For each chain, dashed vertical lines mark the position of the first and last chain component when the chain is located in the sequence with minimal RMSD. (**b**) Cβ chemical shifts of the chains (colored curves) when these are located according to minimal RMSD, and statistically expected Cβ chemical shifts according to residue type (empty squares; values from the BioMagResBank)

simultaneously it takes advantage of the distribution of information among the various projections, substantially reducing the sensitivity problems that would arise if individual projections are analyzed (e.g. peak picked) separately. The thought experiment in the Introduction as well as experimental data on very weak signals (Malmodin and Billeter 2005b, 2006) indicate the robustness of the approach in situations with S/N close to one. (3) Aliasing caused by linearly combining several chemical shifts along the indirect dimensions requires no additional processing during or after the decomposition (Malmodin and Billeter 2006). (4) Spectral reconstructions of various types of subspaces of the corresponding full-dimensional spectrum are possible without introduction of artifacts.

Current limitations of the approach presented here include the inability of the present code to handle projection angles other than 0, ±45 and 90°. It was shown earlier that multi-way decomposition in principle can use spectra with other selected projection angles such as 18.4, 26.6, 63.4 or 71.6° (Malmodin and Billeter 2006). Arbitrary projection angles may be implemented using interpolation techniques, but in contrast to all angle values mentioned above this has not been tested. Another current drawback is the need to select intervals along the acquisition dimension for the decomposition. This is mainly a consequence of the considerable memory use and the strong dependence of calculation times on the number of components; reducing memory needs and speeding up of the algorithm is currently being pursued in our lab.

The first step of the approach described here for resonance assignment of the complete backbone and $C\beta H_n$ nuclei consists of multi-way decomposition (Malmodin and Billeter 2005b, 2006) applied to 30 spectra. This decomposition is very general and the same tool, PRODECOMP, may be used for projections recorded for achieving side chain assignments or extracting structural data. The second step, SHABBA is more specific to sequential backbone assignments. Corresponding tools for side chain assignments or extracting structural data are currently implemented and tested. Backbone assignments with PRODECOMP decompositions and SHABBA require user interference only for defining the run-time parameters for PRODECOMP (see Supplementary Material for more details), entering a factor to identify unique correlations among components in SHABBA, and for removing output components that are copies of others; automation of this latter step is in progress.

The novel tool, SHABBA, which analyzes the components that result from decomposition with PRODECOMP relies on direct analysis of spectroscopic data as represented by the shapes. Correlations among shapes are used to identify the signal similarities observed for sequential components where for example the Cα shape of one component and the $C\alpha(i-1)$ shape of another describe the same nucleus. This more direct approach is less prone to errors introduced by other analysis steps such as peak picking, which becomes only necessary after sequentially connecting the spin systems. Most of the chains of components obtained from the correlation analysis are more than long enough to allow unique positioning in the protein sequence based on the Cβ and Cα shifts as shown in Fig. 6. This is due to the unique shifts of serines, threonines, alanines and glycines, the latter being identified by their missing shift for β nuclei. Even shorter chains such as the one comprising residues 20–23 or the one with the C-terminal two glycines show a rather unique pattern; in any case, they can be unambiguously positioned once the larger ones leave only a few possible gaps. This redundancy when positioning the chains of fragments in the protein sequence allows for additional gaps, i.e. shorter and more fragments, which would be the first consequence of poorer spectral data when signals become undetectable.

Each 9-dimensional component, which results from the multi-way decomposition, characterizes in a single step the chemical shifts of all nuclei in the corresponding

component. The set of shifts corresponding to a component is therefore highly unlikely to contain any erroneous frequency in spite of shift degeneracy in the spectra. One should note that the directly detected dimension with ω_{HN} plays no special role in the decomposition; thus even complete overlap in this dimension will not give rise to ambiguities as long as differences are present in at least three other dimensions (Smilde et al. 2004; Orekhov et al. 2001). In the second step, based on correlations among components, spectral descriptions of individual nuclei are also combined to create new, not directly measured correlations between shapes of neighboring components, making this step rather similar to "covariance" (Zhang and Brüschweiler 2004) and "hyperdimensional" (Kupče and Freeman 2006) NMR. Along these lines, simple direct product operations allow artifact-free (re)construction of spectra of nuclei that do not belong to the same component; for example a 2D spectrum of type [$\omega_{HN(i-1)}$, $\omega_{HN(i)}$] could be constructed (which could be extended to higher dimensions with other nuclei such as nitrogens), yielding "hyperdimensional" correlations reminiscent of the 7D NMR data presented in Hiller et al. 2007, where such correlations were however directly observed.

We have illustrated here the decomposition of a set of projection spectra recorded using 5D pulse sequences followed by the complete resonance assignment of the backbone and CβH$_n$ groups. Characterization of large unambiguous spin systems comprising up to 11 chemically different nuclei result directly from the decomposition, and the novel tool SHABBA provides backbone assignments exploiting correlations among the components describing the spin systems. Demonstration and testing of the approach on ubiquitin provided fully correct and 99.5% complete backbone resonance assignments. With the experimental conditions described in Methods, each projection can be recorded in 30 minutes; the decomposition with PRODECOMP required less than 3 h CPU time; total time for manual intervention (interval selection prior to PRODECOMP, visual screening for components that are copies of other or that contain only noise) was about 2 h; and the SHABBA CPU time was on the order of seconds (more details are given in Supplementary Material). While we do not know the limit in terms of protein size, the following considerations indicate that this limit is well beyond 130 residues: For azurin (128 residues) clear components were obtained even for a residue whose signals exhibit a S/N that is about ten times smaller than for all other residues (Malmodin and Billeter 2005b, 2006); using higher field instruments (the present experiments were all performed at 600 MHz) and cryogenic probes will extend the applicability of the approach, as will optimizations of the experiments or the multi-way decomposition algorithm. Currently, we are introducing TOCSY- and NOESY-type evolutions in the experiments in order to obtain side chain assignments and structural information; preliminary results indicate that in the NOESY-type applications most distances shorter than 4.5Å are detectable for 15 kD proteins. The entire tool is available from the authors, but it is also being incorporated into generally available packages such as CCPN (www.ccpn.ac.uk) and is being interfaced to instrument software (TopSpin, Bruker Biospin GmbH).

Acknowledgments This work was supported by research grants from the Swedish Research Council (621-2003-4048) and the EU (LSHG-CT-2005-018988). Data from the BioMagResBank were used.

References

Atreya HS, Szyperski T (2004) G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignment. Proc Natl Acad Sci USA 101:9642–9647

Bax A, Freeman R (1981) Investigation of complex networks of spin-spin coupling by two-dimensional NMR. J Magn Reson 44:542–561

Bax A, Mehlkopf AF, Smidt J (1979) Absorption spectra from phase-modulated spin echoes. J Magn Reson 35:373–378

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

Folland GB (1992) Fourier analysis and its applications. Brooks/Cole, Pacific Grove, CA

Freeman R, Kupče E (2003) New methods for fast multidimensional NMR. J Biomol NMR 27:101–113

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

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

Grzesiek S, Anglister J, Bax A (1993) Correlation of backbone amide and aliphatic side-chain resonances in C-13/N-15 enriched proteins by isotropic mixing of C-13 magnetization. J Magn Reson Series B 101:114–119

Hiller S, Fiorito F, Wüthrich K, Wider G (2005) Automated projection spectroscopy (APSY). Proc Natl Acad Sci USA 102:10876–10881

Hiller S, Wasmer C, Wider G, Wüthrich K (2007) Sequence-specific resonance assignment of soluble nonglobular proteins by 7D APSY-NMR spectroscopy. J Am Chem Soc 129:10823–10828

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

Kupče E, Freeman R (2006) Hyperdimensional NMR spectroscopy. J Am Chem Soc 128:6020–6021

Logan TM, Olejniczak ET, Xu RX, Fesik SW (1993) A general method for assigning NMR spectra of denatured proteins using 3D HC(CO)NH-TOCSY triple resonance experiments. J Biomol NMR 3:225–231

Luan T, Orekhov VY, Gutmanas A, Billeter M (2005) Accuracy and robustness of three-way decomposition applied to NMR data. J Magn Reson 174:188–199

Malmodin D, Billeter M (2005a) High-throughput analysis of protein NMR spectra. Prog Nucl Magn Res Spec 46:109–129

Malmodin D, Billeter M (2005b) Multiway decomposition of NMR spectra with coupled evolution periods. J Am Chem Soc 127:13486–13487

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

Moseley HNB, Riaz N, Aramini JM, Szyperski T, Montelione GT (2004) A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra. J Magn Reson 170:263–277

Muhandiram DR, Kay LE (1994) Gradient-enhanced triple-resonance 3-dimensional NMR experiments with improved sensitivity. J Magn Reson B 103:203–216

Orekhov VY, Ibraghimov IV, Billeter M (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR 20:49–60

Powers R, Gronenborn AM, Clore GM, Bax A (1991) 3-dimensional triple-resonance NMR of C-13/N-15-enriched proteins using constant-time evolution. J Magn Reson 94:209–213

Seavey BR, Farr EA, Westler WM, Markley JL (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1:217–236

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

Szyperski T, Atreya HS (2006) Principles and applications of GFT projection NMR spectroscopy. Magn Reson Chem 44:S51–S60

Szyperski T, Wider G, Bushweller JH, Wüthrich K (1993) Reduced dimensionality in triple-resonance NMR experiments. J Am Chem Soc 115:9307–9308

Szyperski T, Yeh DC, Sukumaran DK, Moseley HNB, Montelione GT (2002) Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. Proc Natl Acad Sci USA 99:8009–8014

Tykhonov AN, Samarskij AA (1990) Equations of mathematical physics. Dover, New York

Yoon JW, Godsill S, Kupče E, Freeman R (2006) Deterministic and statistical methods for reconstructing multidimensional NMR spectra. Magn Reson Chem 44:197–209

Zhang FL, Brüschweiler R (2004) Indirect covariance NMR spectroscopy. J Am Chem Soc 126:13180–13181