Chapter 6

Rapid Multidimensional NMR: Decomposition Methods and their Applications

Martin Billeter and Doroteya K. Staykova

Department of Chemistry, Biochemistry and Biophysics, Gothenburg University, P.O. Box 462, SE-405 30, Gothenburg, Sweden

6.1 INTRODUCTION

A signal in a typical *N*-dimensional (*N*D) NMR spectrum of a macromolecule can be characterized as a direct product of *N* Lorentzian (or similar) functions, where each function characterizes the chemical shift and linewidth of the signal along one of the dimensions. Consequently, a complete spectrum may be described by a sum of direct products, one for each signal. The number of terms in this sum can be reduced by grouping signals that involve the same nuclei in all but one dimension: All signals of such a group can be described by a single term in the reduced sum, i.e., by a single direct product where one of the functions contains several "peaks". Note that this model characterizing NMR data as direct products does not apply to (random) noise. As an illustration, consider a 3D ^{15}N-nuclear Overhauser effect spectroscopy-heteronuclear single-quantum correlation (^{15}N-NOESY-HSQC) spectrum of a protein: For each amide group a direct product is constructed whose three functions define, respectively, the chemical shift of the amide hydrogen HN, of the amide nitrogen N, and the shifts of all hydrogens that interact with the HN via an NOE (Figure 6.1a).[1] Thus, the first two functions resemble typical NMR lineshapes, while the third one contains several resonances. This relation between 1D functions and *N*D NMR spectra represents a basis for decomposition methods. It provides a grouping of the signals in the NMR spectrum (e.g., all NOEs to an individual HN), and it reduces the spectral complexity to a set of 1D functions that contain one or a small number of resonances.

The following considerations make no assumptions on the form of the above functions. They may be of Lorentzian or similar type, but many other forms such as exponential functions are conceivable, for example, damped oscillations as in time-domain data prior to Fourier transformation (i.e., free induction decays (FIDs)): A single NMR signal in an *N*D

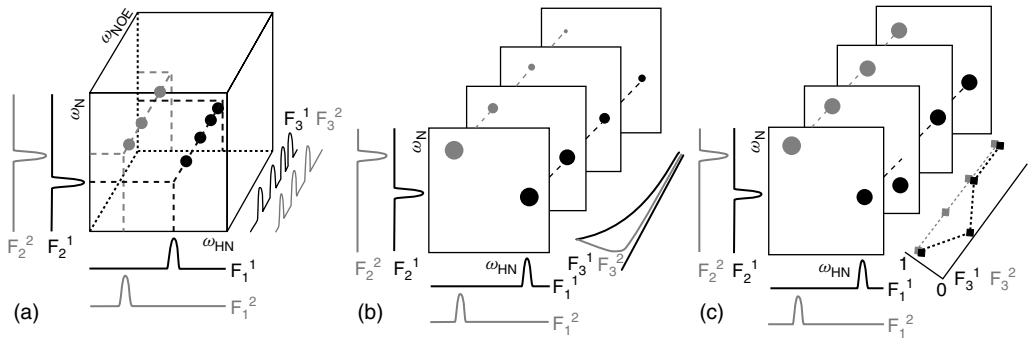

Figure 6.1. Schematic examples of three-way decompositions. The NMR input is represented in the center of each panel either as a cube (3D spectrum) or as a stack of planes (2D spectra). Peaks are depicted by filled circles; dashed lines serve for enhancing visibility. In each panel, two components are shown, in black and gray, respectively. The heavy lines outside the spectra are the shapes of the components annotated by "$\boldsymbol{F}_\mathbf{i}^\mathbf{k}$", where the lower index i refers to the shape number (i.e., the dimension) and the upper index k to the component number. Decompositions are illustrated for (a) a ^{15}N-NOESY-HSQC, (b) a stack of ^{15}N-HSQC-type spectra recorded with an increasing relaxation delay in the pulse sequence, and (c) a stack of ^{15}N-HSQCs recorded for a target protein with different additions of potential ligands. The shapes $\boldsymbol{F}_\mathbf{3}^\mathbf{k}$ in panel (b) describe exponential decays, while the same shapes in panel (c) adopt values of 1 when a peak is present at the position given by the other shapes, and 0 when it has moved owing to ligand binding.

time-domain data set is given by the product of N complex exponentials.[2]

Decomposition methods relying on similar model properties to those outlined above are well studied in mathematics and popular in various fields in chemistry.[3,4] They can easily be generalized from the analysis of conventional NMR spectra to the application to NMR data recorded with *nonuniform sampling* (see Chapters 8 and 9), or as low-dimensional *projections* of a high-dimensional virtual spectrum (see Chapter 5). These two keywords stand for novel techniques of "rapid" NMR experiments that can reduce demands on instrument time by several orders of magnitude.[5,6] Time savings of this extent open an avenue to very high dimensionality (≥5), and more generally to novel types of NMR approaches (e.g., unambiguous backbone assignment of proteins using a single experiment[7]). These rapid techniques exploit the large redundancy present in high-dimensional spectra. Consider a 5D spectrum with 32 points along each indirect dimension and 1024 acquisition points; it would consist of about 10^9 points. Often, however, the number of signals does not exceed a few thousands and each signal can rather accurately be described with not more than 11 parameters (five frequencies, five linewidths, and one amplitude). Thus, the redundancy factor between the spectrum and the information extracted from it encompasses several orders of magnitude.

6.2 THE DECOMPOSITION MODEL

The mathematical formulation of the above model relating a high-dimensional NMR data set and 1D functions forms a basis for several analysis tools that rely on general decomposition methods (see the book by Smilde, Bro and Geladi,[8] and references therein):

$$\boldsymbol{S}(x_1, x_2, \ldots, x_N) \approx \sum_{k=1}^{K} a_k \cdot \left(\boldsymbol{F}_1^k(x_1) \otimes \boldsymbol{F}_2^k(x_2) \ldots \otimes \boldsymbol{F}_N^k(x_N)\right) \tag{6.1}$$

$\boldsymbol{S}(x_1, x_2, \ldots, x_N)$ stands for a general ND NMR data set (e.g., time domain or frequency domain). The sum runs over K terms, henceforth called *components*; $\boldsymbol{F}_\mathbf{i}^\mathbf{k}(i = 1, \ldots, N, k = 1, \ldots, K)$ are 1D functions, henceforth called *shapes*, which are connected by the direct product operator ⊗, and the numbers a_k are amplitudes.[1] Because NMR data sets consist of a finite number of discrete measurement points, $\boldsymbol{S}(x_1, x_2, \ldots, x_N)$ is best written as an ND matrix and the shapes $\boldsymbol{F}_\mathbf{i}^\mathbf{k}$ as vectors. Decomposing an input spectrum $\boldsymbol{S}$ into K components, i.e., replacing the ND matrix $\boldsymbol{S}$ by the NK shapes $\boldsymbol{F}_\mathbf{i}^\mathbf{k}$ according to equation (6.1), compresses the data size from p^N (p is the number of points along each axis, assumed here to be equal for all dimensions) to $K(1 + Np)$ numbers. It usually also provides a logical grouping of the signals into

components (in the example of Figure 6.1a, a 3D ^{15}N-NOESY-HSQC, all NOEs involving a given HN are grouped into one component), and it simplifies the NMR data by yielding 1D shapes that are easy to analyze. However, spectra also contain noise and artifacts that do not follow the model of equation (6.1), thus the "≈" sign in this equation. Assuming that most of the spectral intensity originates from true signals, an optimal decomposition of the input is defined by finding vectors $\boldsymbol{F}_{\mathbf{i}}^{\mathbf{k}}$ that minimize the argument of the following expression:

$$\min \left[\left\| \boldsymbol{S}(x_1, x_2, \ldots, x_N) - \sum_{k=1}^{K} a_k \cdot \left(\boldsymbol{F}_1^k(x_1) \otimes \boldsymbol{F}_2^k(x_2) \ldots \otimes \boldsymbol{F}_N^k(x_N) \right) \right\|_2 + \lambda \sum_{k=1}^{K} {a_k}^2 \right] \tag{6.2}$$

where the norm of a matrix $\boldsymbol{M}$ is given by $\|M\|_2 = \sum \left[(M)_{x_1 x_2 \ldots x_N} \right]^2$. The last term in expression (6.2) is a Tikhonov regularization, which helps avoiding large variations in the component amplitudes a_k; it is controlled by the Tikhonov factor λ.[9] The number of components, K, can often be estimated from inspection of simpler spectra; in many cases it corresponds to the number of peaks in a ^{15}N-HSQC spectrum. Alternatively, several values of K can be tested to see when an increase of K no longer lowers the value calculated with expression (6.2).[10]

It has been shown that decompositions achieved by minimization of expression (6.2) are unique for (nondegenerate) situations with $N \geq 3$.[11] The name "three-way decomposition" refers to this fact; however, the number three is in no way an upper limit, and increasing the dimensionality will only improve convergence in near-degenerate situations. Decompositions, i.e., fitting of the vectors $\boldsymbol{F}_{\mathbf{i}}^{\mathbf{k}}$ to minimize expression (6.2), are normally achieved by the method of alternating least squares: a set of parameters, e.g., the shape vectors in all but one dimension, is held fixed while the remaining parameters are optimized in a least-square sense, and the choice of free parameters is cyclically permuted so that eventually all parameters are optimized. Other approaches such as eigenvalue-based optimizations are also conceivable.[8]

In typical decompositions of NMR data recorded for proteins, a component is defined by a characteristic atom, often an HN, and by the magnetization transfer paths between this and other atoms. Thus, the group of nuclei involved in a component is reminiscent of a "spin system", and indeed, in many situations, a better understanding of the decomposition result may be obtained by replacing the term *component* by *spin system*; the above example of a 3D ^{15}N-NOESY-HSQC would, however, require a generalization of the term *spin system* due to the inclusion of NOEs.

6.2.1 Three-way Decompositions

Decompositions based on the above model were introduced to NMR recently and applied to various types of data. This section addresses some practical concepts of decomposition of NMR data. Conventional NMR spectra rarely exceed three dimensions due to the prohibitive experiment time required to acquire higher dimensional spectra with sufficient resolution. The decomposition of a conventional 3D ^{15}N-NOESY-HSQC is illustrated in Figure 6.1(a). All three dimensions are frequency axes, and the number of components corresponds closely to the number of HN resonances, i.e., to the number of residues in the protein. Applying the decomposition to intervals selected along the HN-axis rather than the entire spectral range increases the efficiency of the calculations. This splitting into several tasks is allowed since spin systems centered on HN nuclei with very different resonance frequencies do not interfere with each other; overlapping HN resonances must, however, be decomposed simultaneously. The decomposition results also allow quantitative extraction of distance restraints for structure determination. A straightforward approach consists of 1D peak picking and intensity estimation in the shapes along the NOE axis ($\boldsymbol{F}_{\mathbf{3}}^{\mathbf{1}}$ or $\boldsymbol{F}_{\mathbf{3}}^{\mathbf{2}}$ in Figure 6.1a). This is warranted by the fact that the decomposition maintains relative intensities of NOEs within a component. NOEs from different components may then be calibrated by assuming that all diagonal peaks have the same size.[12]

The fact that no assumption is made on the form of the $\boldsymbol{F}_{\mathbf{i}}^{\mathbf{k}}$ shapes in expression (6.2) opens the possibility for a series of other applications. Relaxation data on proteins are often collected as stacks of 2D ^{15}N-HSQC-type spectra. A varying time delay in the pulse sequence defines a third axis perpendicular to

the two frequency axes of the HSQCs.[10,13] Decomposition of this 3D data set yields for each N–HN moiety a component with two shapes along the two frequency axes, providing chemical shifts for HN and N, and a third shape along the relaxation time axis (Figure 6.1b). Fitting an exponential decay function to the latter vector yields an accurate estimation of the relaxation time. Because the decomposition fits frequency and relaxation time shapes simultaneously, spectral features such as linewidths are automatically preserved among the different HSQCs. This supports the separation of overlapping peaks in the HSQC spectra, providing better volume estimates.[13]

In the above applications, the units in all dimensions (or for all shapes) are physical entities such as frequencies or time, and the vectors are digitized forms of continuous functions. None of these properties are, however, required for the shapes according to formulas (6.1) and (6.2). In drug discovery applications, a stack of ^{15}N-HSQC spectra is often recorded for a target protein in the absence or presence of different potential ligand molecules. The effect of binding of a molecule to the protein is a dislocation in the corresponding ^{15}N-HSQC spectrum of peaks belonging to protein residues near the binding site. Thus, for a given residue, a signal is either present or absent at the location of the signal for the free protein. In a decomposition of a set of HSQCs recorded for a target protein in varying mixtures with potential ligand molecules, this two-valued function (1 for presence and 0 for absence of a peak) forms the third shape besides frequency shapes for HN and N (Figure 6.1c).[14]

6.2.2 Model for Projected Data

The right-hand side of equation (6.1) is a full description of all signals found in an *N*D NMR spectrum, from which it is straightforward to derive all parameters of interest such as individual resonance frequencies, but also a complete characterization of the *N*D peaks, with coordinates and lineshapes in all dimensions as well as intensities. The term on the left-hand side represents the corresponding regular *N*D spectrum. With increasing values of N it becomes difficult to record such a full-dimensional spectrum due to impossibly long experiment times and excessive data sizes. Modifying equation (6.1) by replacing the left-hand side with more realistic experimental data, while keeping all the shapes on the right-hand side, may open an avenue to the determination of the same type of data as would be found in a full *N*D spectrum with $N > 3$, but with drastic reductions in experiment time and data size. In reduced dimensionality approaches, all N frequencies of a high-dimensional spectrum are reported in lower dimensional spectra.[15] In the particular case of 2D projections of an *N*D spectrum, and with the directly detected dimension unaffected, the frequencies observed in the $N - 1$ indirect dimensions are combined in a linear fashion on the single projected dimension. To better understand the projection concept, consider a 3D spectrum represented by a cube with three frequency axes, where the third dimension is the directly detected one denoted as ω_{HN} (Figure 6.2a). A peak in such a spectrum can be described with its Cartesian coordinates (Ω_1, Ω_2, Ω_{HN}). A 2D projection of this 3D spectrum with the directly detected dimension unaffected can be any plane including the ω_{HN}-axis; a diagonal between the ω_1-axis and the ω_2-axis is chosen in Figure 6.2(a). The projection coordinate is thus $\omega = \omega_1 + \omega_2$, and the above peak is projected to (Ω, Ω_{HN}), where $\Omega = \Omega_1 + \Omega_2$. Generalizing to an *N*D spectrum $\boldsymbol{S}(\omega_1, \omega_2, \ldots, \omega_N)$ with the last dimension being the directly detected one, a projection corresponding to the example in Figure 6.2(a)

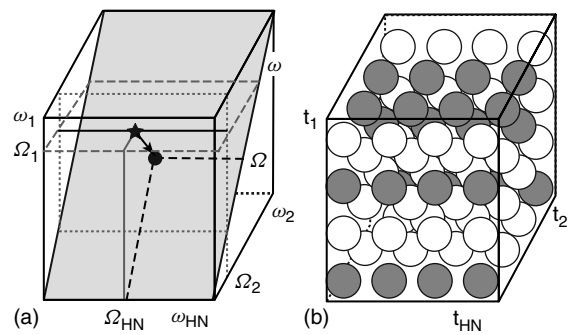

Figure 6.2. Illustrations of partial sampling. (a) Diagonal 2D projection of a 3D spectrum. The peak at (Ω_1, Ω_2, Ω_{HN}) in the full-dimensional spectrum (star) is projected to the diagonal plane (shaded) with the same ω_{HN}-axis and a new axis ω tilted at 45° to both ω_1 and ω_2. The projected peak (filled circle) is given by ($\Omega = \Omega_1 + \Omega_2$, Ω_{HN}) in the projected plane. (b) Individual data points in a 3D time-domain data set with coordinates t_1, t_2, and t_{HN} are represented by spheres. While a conventional sampling would require all spheres to be measured, random sparse sampling records only data at the gray spheres. Note that the acquisition time t_{HN} is not affected by the sparse sampling.

would map a peak at $(\Omega_1, \Omega_2, \ldots, \Omega_N)$ to (Ω, Ω_N) with $\Omega = \Omega_1 + \Omega_2 + \cdots + \Omega_{N-1}$; other projections would result in different linear combinations of Ω_1 to Ω_{N-1}.

For 2D projections of *N*D spectra, Fourier theory implies that equation (6.1) translates to

$$\begin{aligned} \boldsymbol{P}_m(\omega, \omega_N) \\ \approx \sum_{k=1}^{K} a_k \cdot (\hat{p}_{m1}[\boldsymbol{F}_1^k(\omega_1)] * \hat{p}_{m2}[\boldsymbol{F}_2^k(\omega_2)] \ldots \\ * \hat{p}_{mN-1}[\boldsymbol{F}_{N-1}^k(\omega_{N-1})] \otimes \boldsymbol{F}_N^k(\omega_N)) \end{aligned} \quad (6.3)$$

where $\boldsymbol{P}_{\mathbf{m}}(\omega, \omega_N)$ are 2D projections enumerated by the index m, and * is the convolution operator.[16,17] The operators $\hat{p}_{mi}$ affect the shapes $\boldsymbol{F}_{\mathbf{i}}^{\mathbf{k}}(i = 1, \ldots, N, k = 1, \ldots, K)$ in a manner depending on the projection chosen. With the restriction of projection angles to 0° or $\pm 45^\circ$ as valid in the applications discussed below, the effect of these operators is simply to either remove a shape from the equation (in the case of 0° where a dimension is excluded from the projection) or invert the direction of the shape (in the case of -45°).[17] As long as enough experimental information is provided by the projections $\boldsymbol{P}_{\mathbf{m}}(\omega, \omega_N)$ on the left-hand side of equation (6.3), all values in the *NK* shapes can be determined with a suitable optimization algorithm. Again, the presence of noise is considered by minimizing the difference of the two sides of the (approximate) equation (6.3). The shapes $\boldsymbol{F}_{\mathbf{i}}^{\mathbf{k}}$ contain all the expected information (chemical shifts, lineshapes, and intensities of peaks), but inserting these $\boldsymbol{F}_{\mathbf{i}}^{\mathbf{k}}$ into equation (6.1) would also allow the back-calculation of a clean reconstruction of $\mathbf{S}(\omega_1, \omega_2, \ldots, \omega_N)$; one may also choose to omit selected dimensions obtaining subspaces, or to drop selected components yielding a spectrum with a well-characterized subset of peaks.

A suitable tool for the minimization according to equation (6.3) is the fast nonnegativity-constrained least squares (FNNLS) algorithm.[18] An intrinsic feature of this algorithm is the assumption of nonnegativity of all parameters in the shapes $\boldsymbol{F}_{\mathbf{i}}^{\mathbf{k}}$, and thereby also of all intensities in the input projection spectra $\boldsymbol{P}_{\mathbf{m}}(\omega, \omega_N)$. In spectra with negative peaks, the latter often have a somewhat different interpretation than the positive ones; typical examples occur in experiments involving magnetization transfers to the α- and β-nuclei in proteins, where the former may yield positive peaks and the latter negative ones. Such spectra may be used twice in the decomposition, the original, and a copy after inversion of the sign for all data points, allowing for separate shapes for the α- and β-nuclei.[19] With the simultaneous use of sizable amounts of data from dozens of projections, the sizes of intermediate matrices in the algorithm may become problematic, and proper modifications help avoid memory problems and significantly speed up the calculations.[20]

6.2.3 Model for Sparse Data

Another concept for reducing experiment time and data size while keeping the high dimensionality is the recording of sparse data, where only a (often randomly chosen) subset of the FIDs are collected (Figure 6.2b). For this type of application, equation (6.1) needs to be modified in two respects: the decomposition calculation is performed in the time domain, since nonuniformly sampled data cannot be subjected to discrete FT, and the experimental input (left-hand side of equation (6.1)) is incomplete. (The term *nonuniform sampling* is often used in a general sense, including the sparse type of sampling addressed in this section as well as the projections introduced above and other nonstandard sampling schemes.) The modified model for 3D data is described by

$$\begin{aligned} \boldsymbol{S}(t_1, t_2, t_3) \approx \boldsymbol{G}(t_1, t_2, t_3) \bullet \sum_{k=1}^{K} a_k \cdot (\boldsymbol{F}_1^k(t_1) \\ \otimes \boldsymbol{F}_2^k(t_2) \otimes \boldsymbol{F}_3^k(t_3)) \end{aligned} \quad (6.4)$$

where t_1 and t_2 are evolution times and t_3 is the acquisition time.[21,22] The matrix $\boldsymbol{G}(t_1, t_2, t_3)$ contains values of 1 for all combinations (t_1, t_2) for which an FID was recorded and 0 otherwise; the product $\bullet$ implies that each element of $\boldsymbol{G}$ is multiplied by the corresponding element of the matrix defined by the sum. Missing FIDs in $\boldsymbol{S}$, which are replaced by zeros, do not contribute when the difference between the left- and right-hand sides of equation (6.4) is calculated and minimized. The choice of the extent of sparseness must ensure that the information on the left-hand (experiment) side of equation (6.4) is sufficient to allow determination of all parameters on the right-hand side. An additional (but hardly limiting) constraint is that each multiple of the dwell time for each time axis occurs at least once in the selection of FIDs.[21] Again, back-calculation with equation (6.4) (without the matrix $\boldsymbol{G}$) allows for

reconstruction of the complete time-domain data set, which is suitable for regular FT. The model can also be generalized to $N > 3$ dimensions.[23,24]

6.3 APPLICATIONS TO CONVENTIONAL NMR DATA

In this section, we address selected aspects of three-way decompositions to protein NMR data with three examples: the earlier mentioned 3D ^{15}N-NOESY-HSQC (Figure 6.1a), the determination of relaxation times, and the detection of ligand binding to a target protein.

6.3.1 3D ^{15}N-NOESY-HSQC

This 3D spectrum contains a large number of NOEs that define short distances between amide hydrogen atoms, HN, on one side, and all kinds of other hydrogen atoms on the other. Frequencies for these hydrogen atoms are found on axes labeled ω_{HN} and ω_{NOE} in Figure 6.1(a); in addition, for each HN, the frequency of the nitrogen to which it is bound is given on the axis labeled ω_N. All NOEs to a particular amide hydrogen are localized on one line along the ω_{NOE} axis defined by the resonances of the amide hydrogen and nitrogen on the other two axes; this provides a natural grouping of the NOEs. Decomposition according to expression (6.2) follows this scheme by using one component per amide group, resulting in a shape for the HN, one for the amide nitrogen, and one that contains all resonances of hydrogen atoms interacting via an NOE with the HN (Figure 6.1a). Note that this grouping of all NOEs to a given HN into one component is not imposed, but is purely a consequence of the algorithm trying to minimize the residual in expression (6.2): with a given number of components, more intensity can be incorporated into the model when using components that can accommodate several signals.

Using previously assigned amide resonances, a 3D ^{15}N-NOESY-HSQC for the 14 kDa protein azurin could be decomposed yielding a component for each backbone amide group, i.e., for each amino acid residue except prolines.[12] About 95% of the NOEs obtained from 1D peak picking of the corresponding shapes refer to short distances in the crystal structure of azurin; and, vice versa, about 95% of these short distances could be related to observed NOEs. As shown by residue pairs with strong overlap for both the amide hydrogen and nitrogen resonances (differences of 0.01 and 0.3 ppm, respectively),[1] and by simulations,[25] the model profits strongly from its inherent use of lineshapes when decomposing overlapping signals, and from grouping the NOEs to the various amide hydrogen atoms.

6.3.2 Relaxation

In typical relaxation measurements by NMR, the intensities of peaks, corresponding to amide groups, are followed through a series of relaxation-modulated ^{15}N-HSQC spectra. The conventional approach of determining peak intensities independently for each spectrum, followed by fitting of exponential curves, suffers from difficulties in extracting intensities for weak peaks, e.g., in the spectra with longest relaxation, as well as from randomly distributed

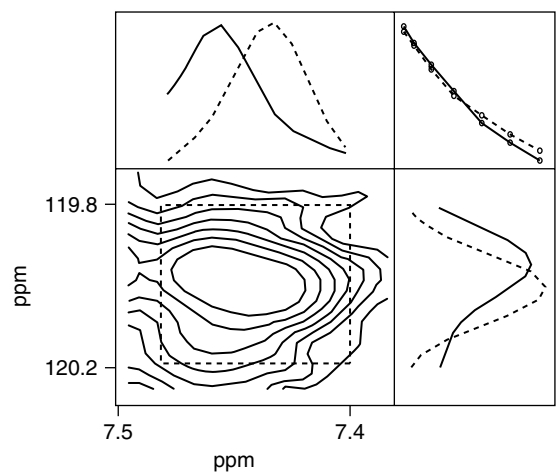

Figure 6.3. Calculation of T_1 for two signals with strong overlap in the ^{15}N-HSQC for the 370 residue protein MBP.13 The lower left panel displays a region from the spectrum with shortest relaxation delay. The directly adjacent panels show the resulting shapes along the frequency axes for a decomposition with two components of the region enclosed with dashed lines. The upper right panel is a plot of exponential curves fitted to the shapes in the third dimension (compare also with Figure 6.1b). The resulting relaxation times were 1177 ± 19 and 1357 ± 39 ms. Solid and dashed lines in the latter three panels describe shapes for residues 55 and 195, respectively.

artifacts. Applying the decomposition model as explained above, shapes for the exponential decay are a direct result (Figure 6.1b). Furthermore, the lineshapes for hydrogen, and for nitrogen, for a given amide group are forced to be the same in all spectra; this makes the extraction of signal intensities from the spectra with the longest relaxation delays more reliable. The joint use of all spectra in the decomposition allows the separation of strongly overlapping peaks based on their different relaxation times. Thus, the decomposition for all 241 assigned backbone amide groups in the 370 residue long maltose binding protein (MBP) resulted in systematically lower error bounds for the relaxation times, better discrimination in the case of spectral overlap, and overall more residues that are accessible to relaxation measurements (Figure 6.3).[13]

6.3.3 Drug Discovery

The application of the decomposition model of expression (6.1) in drug discovery helps to monitor chemical shift perturbations due to the binding of ligands in spectra recorded for a target protein (Figure 6.1c). Here, the third shape is a function with two values: 1 if a peak is present at the location where it is found in most spectra, and 0 if the peak has been shifted to a different location, leaving the old location empty. The need to use the same shapes for the hydrogen and the nitrogen resonances for all spectra again enables the identification of peaks in situations with spectral overlap and/or artifacts. Tests on a small enzyme and simulations showed the applicability of the method to sets of more than 50 spectra.[14] The procedure involves essentially only one step, the decomposition, and avoids the necessity for peak picking, making it robust against low signal to noise and the presence of artifacts.

6.4 APPLICATIONS TO "RAPID" NMR DATA

Probably the best applications of the combination of "rapid" NMR techniques with decomposition analyses are found in protein studies. Investigations of internal dynamics and interaction with other molecules have been mentioned above in the context of more conventional NMR approaches. Structural studies, e.g., in genomics approaches, are likely to require a further increase of efficiency, achievable with the help of projections or other nonuniform sampling techniques. While resonance assignments may be considered an intermediate spectroscopic result, they nonetheless represent the foundation for most protein studies. On this basis, goals with more obvious interest such as determination of 3D structures, typically based on constraints on interatomic distances extracted from NOE-type experiments, can be formulated.

With the possibility to record very high-dimensional NMR data using various forms of nonuniform sampling, one may define large groups of atoms whose frequencies are characterized by one or a few signals in a single experiment. Within the decomposition formalism, each group corresponds to one component. Figure 6.4 illustrates on a schematic protein fragment several examples of atom groups, for which all resonance frequencies are determined by a single component resulting from the decomposition of a single set of projections.

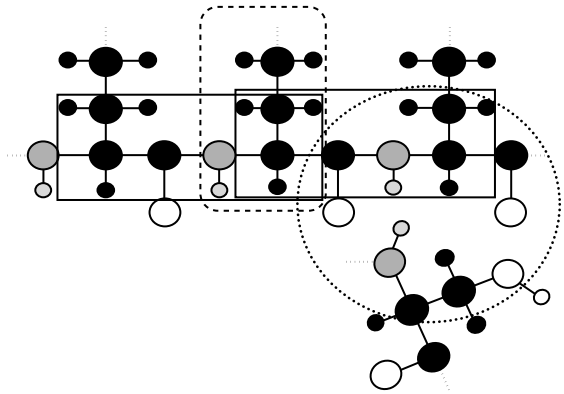

Figure 6.4. Schematic protein fragment with a stretch of three residues (top, backbone running horizontally) and a spatially close further residue. Carbon atoms are shown as black spheres, nitrogen atoms in gray, and oxygen atoms in white, while hydrogen atoms are smaller but have the same shading as the heavy atoms to which they are attached. The different contours enclose examples of atom groups that can be characterized by single decompositions: solid rectangles enclose two identical atom groups for backbone assignment, the dashed rectangle for side chain assignment, and the dotted circle for NOE observation. Note the extensive overlap between the two backbone groups as well as between either the side chain or the NOE-based group and the nearest backbone group.

Thus, the two rectangles with solid contours each include about a dozen spin-1/2 nuclei (in a doubly labeled sample depending on the residue type) from experiments proposed for backbone assignment; these are discussed below in more detail. The dashed contour outlines a group of atoms useful for side chain assignments while the dotted (circular) contour reflects a NOESY-type experiment for structure determination. Note the extensive overlap between the two backbone groups as well as between either the side chain or the NOE-based group and the nearest backbone group, which is made possible by the large number of atoms they contain, i.e., by the use of high-dimensional NMR data.

6.4.1 Projections

Decomposition according to equation (6.3) of projection spectra is illustrated with data from 5D experiments involving magnetization transfers along the backbone (including the $C\beta H_n$ moieties) with projection angles of 0°, $\pm 45^\circ$, and 90°. Resonances for the atom group of each rectangular box with solid contours in Figure 6.4 are observed using two experiments based on the following magnetization transfers:[19]

$$\begin{aligned} &H\beta H\alpha_{i-1}(t_1) \rightarrow C\beta C\alpha_{i-1}(t_2) \rightarrow C\alpha_{i-1} \\ &\rightarrow C'_{i-1}(t_3) \rightarrow N(t_4) \rightarrow HN(t_5) \end{aligned} \tag{6.5}$$

yields for each HN chemical shift information on the α- and β-nuclei of the preceding residue; and

$$\begin{aligned} &H\beta H\alpha(t_1) \rightarrow C\beta C\alpha(t_2) \rightarrow C\alpha \rightarrow N \rightarrow C'_{i-1}(t_3) \\ &\rightarrow N(t_4) \rightarrow HN(t_5) \end{aligned} \tag{6.6}$$

yields for each HN chemical shift information on the α- and β-nuclei of the own residue. For this choice, the two experiments have chemical shifts of three nuclei in common: C′, N, and HN. Moreover, the common directly detected dimension HN (denoted ω_N in equation (6.3)) allows for *simultaneous* decomposition of the projections from both experiments, ensuring a tight relation among the shapes for *all* nuclei in $C\beta H_n{-}C\alpha H{-}C'{-}NH{-}C\alpha H{-}C\beta H_n$ fragments centered around the HN nuclei (Figure 6.4). The resulting decomposition is a complete description of all chemical shifts of the fragment in a 9D space with axes defined according to the frequencies of the different nuclei types (Figure 6.5). Since components centered on neighboring HN nuclei have shapes for a $C\alpha H{-}C\beta H_n$ moiety in common, correlation calculations can reliably connect components in a sequential manner.[19] While this connection fails around prolines, and possibly in a few additional cases, the emerging chains of sequentially connected components are sufficiently long to be placed unambiguously in the protein sequence with the help of their $C\beta$ and $C\alpha$ chemical shifts using statistically expected shift values for each amino acid type, as reported, for example, in the BioMagResBank.[26]

6.4.2 Nonuniform Sampling

For decompositions according to equation (6.4), the sampled points are selected using a random number generator (in contrast to radial sampling as in projections) but with an exponential bias favoring shorter acquisition times; this denser distribution at the beginning of the decaying signal was shown to improve sensitivity compared to unbiased random sampling.[27] An interesting option regarding the extent of data to be recorded, a crucial issue in "rapid" NMR, concerns concurrent acquisition and data analysis, allowing data acquisition to be ended as soon as the analysis of the data collected thus far indicates acceptable results. Both projection spectroscopy[28] and sparse sampling in time domain[23,24] lend themselves to this time-saving approach. The concept of sparse sampling combined with simultaneous data processing is illustrated in Figure 6.6. In many applications, e.g., for the assignment of the protein backbone, a ^{15}N-HSQC may often be used to estimate the number of peaks expected in the experiments using sparse sampling. Thus, in the example of Figure 6.6, it is assumed that knowledge gained prior to the nonuniform sampling experiments indicates that four peaks are expected. After starting the experiment, the data acquired so far are decomposed, and the resulting components are used to construct a complete time-domain data set, which is transformed and peak picked. Figure 6.6 illustrates a possible progression: in this case, sampling only 1% of the time-domain points yields merely noise; after 5% sampling a first peak appears, and with 15% sampling the expected four peaks are observed. Thus, the decomposition and analysis at this stage warrant that the experiment be ended, with a resulting time-saving of 85%.

An implementation of the decomposition algorithm for multidimensional sparse data sets, the multidimensional decomposition (MDD) program,[23,24,29,30] has been applied to a variety of data. For the extraction of distance information from a 4D

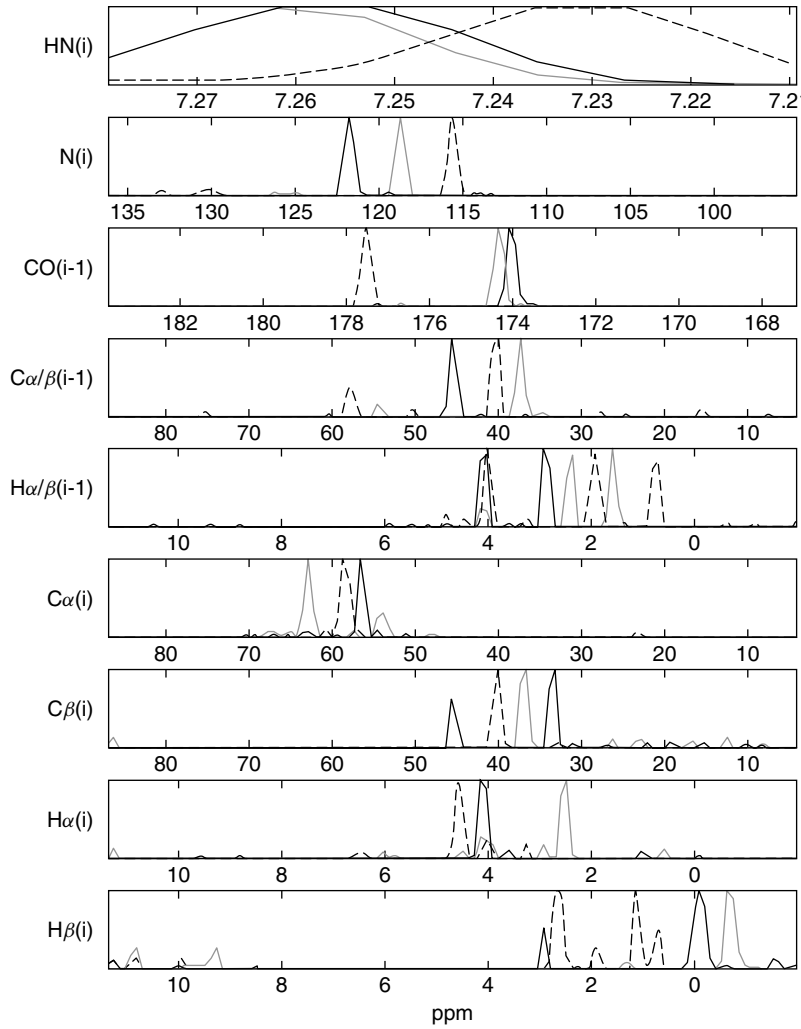

Figure 6.5. Result of a decomposition with three components from 30 projections recorded for the protein ubiquitin.[19] The components with black, gray, and dashed shapes describe residues 59, 11, and 61, respectively. The nine shapes of each component characterize *all* nuclei in $C\beta H_n - C\alpha H - C' - NH - C\alpha H - C\beta H_n$ fragments centered around an HN nucleus (Figure 6.4). All horizontal axes show ppm values for the corresponding nuclei, while the vertical axes show intensities normalized to the maximum of the respective shape.

HCCH-correlation NOESY, 30% sampling proved sufficient to reproduce about 98% of all cross peaks with a correlation of 0.997 between their intensities and those of a conventional spectrum.[29] Tests for backbone assignment using four triple resonance spectra with sparse sampling of 6–9% of these 3D spectra yielded complete assignments for four proteins with molecular weights between 8 and 22 kDa.[30]

6.5 CONCLUSION

Decomposition methods based on the models of equations (6.1), (6.3) or (6.4) provide a number of benefits for the interpretation of NMR data. The close relation between the mathematical model of equation (6.1) and NMR theory assures solid foundations for the approach. In applications reported so far, practically no convergence failures were

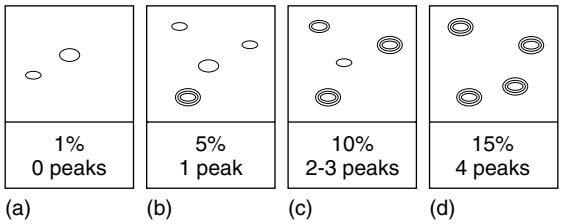

Figure 6.6. Concept of concurrent data recording and analysis. Decomposition and analysis of 1% of the FIDs that would form a conventional time-domain data set yields a spectrum with mostly noise (a). Continued recording provides (b) a first peak when 5% of the data are available, and (c) more peaks with 10% of the data. (d) After 15% of the full data set has been recorded, four peaks are clearly observed, which, in this example, is assumed to be the expected number, and the experiment is stopped.

observed, making the tool very robust and reliable. The simultaneous processing of all experimental data ensures optimal treatment with respect to signal to noise, a critical aspect in "rapid" NMR. Finally, applying equation (6.1) "backwards", i.e., using the shapes resulting from decomposition as input for the calculation of a spectral data set, allows for a full reconstruction, e.g., a full-dimensional spectrum from projections or a fully sampled time-domain matrix from sparsely sampled data. Instead of a full reconstruction, subspaces or subsets are possible by omitting all shapes along selected dimensions or individual components, respectively.

Decomposition algorithms also pose a number of challenges when applied to NMR data from proteins. Often, the size of the NMR input is very large. Consider, for example, a 5D experiment with good resolution along all axes; even sparse sampling may result in a considerable data set. However, the problem concerns not so much the input size, but rather the extent of intermediate or final data. Memory demands in the decomposition algorithm for projections have been reported to represent a bottleneck.[20] Also, reconstruction either of time-domain data in the context of sparse sampling, or of frequency domain data from projections, would often yield unacceptably large data sets. Fortunately, full reconstructions can be avoided rather easily in both cases. Algorithm improvements as well as computer hardware development will push back the limits posed by the current bottlenecks and allow for wider applications of decomposition methods in protein NMR.

RELATED ARTICLES IN THE ENCYCLOPEDIA OF MAGNETIC RESONANCE

Analysis of Spectra: Automatic Methods

Data Processing

Maximum Entropy Reconstruction

REFERENCES

1. V. Y. Orekhov, I. Ibraghimov, and M. Billeter, *J. Biomol. NMR*, 2001, **20**, 49.
2. M. H. Levitt, '*Spin Dynamics*', 2nd edn., John Wiley & Sons: Chichester, 2008, p. 74.
3. H. A. L. Kiers, *J. Chemom.*, 2000, **14**, 105.
4. R. Bro, *Crit. Rev. Anal. Chem.*, 2006, **36**, 279.
5. R. Freeman and Ē. Kupče, *J. Biomol. NMR*, 2003, **27**, 101.
6. D. Malmodin and M. Billeter, *Prog. NMR Spectrosc.*, 2005, **46**, 109.
7. F. Fiorito, S. Hiller, G. Wider, and K. Wüthrich, *J. Biomol. NMR*, 2006, **35**, 27.
8. A. Smilde, R. Bro, and P. Geladi, '*Multi-way Analysis*', John Wiley & Sons: Chichester, 2004.
9. A. N. Tikhonov and A. A. Samarskij, '*Equations of Mathematical Physics*', Dover: New York, 1990.
10. D. M. Korzhnev, I. Ibraghimov, M. Billeter, and V. Y. Orekhov, *J. Biomol. NMR*, 2001, **21**, 263.
11. J. B. Kruskal, in '*Multiway Data Analysis*', eds R. Coppi and S. Bolasco, Elsevier: Amsterdam, 1989.
12. A. Gutmanas, P. Jarvoll, V. Y. Orekhov, and M. Billeter, *J. Biomol. NMR*, 2002, **24**, 191.
13. A. Gutmanas, L. Tu, V. Y. Orekhov, and M. Billeter, *J. Magn. Reson.*, 2004, **167**, 107.
14. C. S. Damberg, Y. Orekhov, and M. Billeter, *J. Med. Chem.*, 2002, **45**, 5649.
15. T. Szyperski, G. Wider, J. H. Bushweller, and K. Wüthrich, *J. Am. Chem. Soc.*, 1993, **115**, 9307.
16. D. Malmodin and M. Billeter, *J. Am. Chem. Soc.*, 2005, **127**, 13486.
17. D. Malmodin and M. Billeter, *Magn. Reson. Chem.*, 2006, **44**, S185.
18. R. Bro and S. de Jong, *J. Chemom.*, 1997, **11**, 393.

19. D. K. Staykova, J. Fredriksson, W. Bermel, and M. Billeter, *J. Biomol. NMR*, 2008, **42**. 87.

20. D. K. Staykova, J. Fredriksson, and M. Billeter, *Bioinformatics* [online], 2008, **24**, 2258.

21. V. Yu. Orekhov, I. Ibraghimov, and M. Billeter, *J. Biomol. NMR*, 2003, **27**, 165.

22. M. Billeter and V. Y. Orekhov, in '*Computational Science—ICCS2003*', eds P. M. A. Sloot, D. Abramson, A. V. Bogdanov, J. J. Dongarra, A. Y. Zomaya and Y. E. Gorbachev, Springer: Berlin, 2003, Part I.

23. V. A. Jaravine and V. Y. Orekhov, *J. Am. Chem. Soc.*, 2006, **128**, 13421.

24. V. A. Jaravine, A. V. Zhuravleva, P. Permi, I. Ibraghimov, and V. Y. Orekhov, *J. Am. Chem. Soc.*, 2008, **130**, 3927.

25. T. Luan, V. Y. Orekhov, A. Gutmanas, and M. Billeter, *J. Magn. Reson.*, 2005, **174**, 188.

26. B. R. Seavey, E. A. Farr, W. M. Westler, and J. L. Markley, *J. Biomol. NMR*, 1991, **1**, 217.

27. A. S. Stern, K.-B. Li, and J. C. Hoch, *J. Am. Chem. Soc.*, 2002, **124**, 1982.

28. H. R. Eghbalnia, A. Bahrami, M. Tonelli, K. Hallenga, and J. L. Markley, *J. Am. Chem. Soc.*, 2005, **127**, 12528.

29. T. Luan, V. A. Jaravine, A. Yee, C. H. Arrowsmith, and V. Y. Orekhov, *J. Biomol. NMR*, 2005, **33**, 1.

30. V. A. Jaravine, I. Ibraghimov, and V. Y. Orekhov, *Nat. Methods*, 2006, **3**, 605.