Journal of Biomolecular NMR (2005) 33: 1–14
DOI 10.1007/s10858-005-1363-6

Article

Optimization of resolution and sensitivity of 4D NOESY using multi-dimensional decomposition

T. Luan[a], V. Jaravine[a], A. Yee[b], C. H. Arrowsmith[b] & V. Yu. Orekhov[a,*]
[a]*The Swedish NMR centre at Göteborg University, Medicinaregatan 5C, P.O. Box 465, 40530, Göteborg, Sweden;* [b]*Northeast Structural Genomics Consortium, Ontario Cancer Institute, 610 University Avenue, Toronto, Ontario, CanadaM5G 2M9*

Received 25 March 2005; Accepted 14 July 2005

Key words: NMR, NOESY, non-uniform sampling, structural genomics

Abstract

Highly resolved multi-dimensional NOE data are essential for rapid and accurate determination of spatial protein structures such as in structural genomics projects. Four-dimensional spectra contain almost no spectral overlap inherently present in lower dimensionality spectra and are highly amenable to application of automated routines for spectral resonance location and assignment. However, a high resolution 4D data set using conventional uniform sampling usually requires unacceptably long measurement time. Recently we have reported that the use of non-uniform sampling and multi-dimensional decomposition (MDD) can remedy this problem. Here we validate accuracy and robustness of the method, and demonstrate its usefulness for fully protonated protein samples. The method was applied to 11 kDa protein PA1123 from structural genomics pipeline. A systematic evaluation of spectral reconstructions obtained using 15–100% subsets of the complete reference 4D ^{1}H–^{13}C–^{13}C–^{1}H NOESY spectrum has been performed. With the experimental time saving of up to six times, the resolution and the sensitivity per unit time is shown to be similar to that of the fully recorded spectrum. For the 30% data subset we demonstrate that the intensities in the reconstructed and reference 4D spectra correspond with a correlation coefficient of 0.997 in the full range of spectral amplitudes. Intensities of the strong, middle and weak cross-peaks correlate with coefficients 0.9997, 0.9965, and 0.83. The method does not produce false peaks. 2% of weak peaks lost in the 30% reconstruction is in line with theoretically expected noise increase for the shorter measurement time. Together with good accuracy in the relative line-widths these translate to reliable distance constrains derived from sparsely sampled, high resolution 4D NOESY data.

Abbreviations: 3DD and 4DD – MDD for three and four dimensions; 4D – four-dimensional; LP – linear prediction; MDD – multi-dimensional decomposition; N_c – number of spectral components; S/N – signal-to-noise

Introduction

Substantial advances in biomolecular NMR have been achieved over the past several decades due to contributions from many research groups. As in many other spectroscopic methods, these efforts have mostly focused on improving spectral sensitivity and resolution. A number of important advances have been achieved including isotope labelling techniques and improved methodologies for pulse sequence design (Bax, 1994; Yamazaki

*To whom correspondence should be addressed. E-mail: orov@nmr.se

et al., 1994; Goto and Kay, 2000; Fernandez and Wider, 2003; Tugarinov et al., 2004a). These methods, together with the development of new hardware, now allow the routine acquisition of sensitive spectra of large biomolecules at sub-millimolar concentrations. The problem of signal overlap in crowded spectra can be alleviated by using modern spectrometers with higher magnetic fields, isotope spectra editing, and decreasing natural line widths (Pervushin et al., 1997; Tugarinov et al., 2004b). Nonetheless, the most efficient way for improving resolution in NMR is to increase the spectra dimensionality (Ernst, 1992; Bax, 1994; Wuthrich, 2003). However, exponential increase in measurement time due to the increased dimensionality of spectra hinders the practical use of 3D and 4D spectra and prohibits spectroscopy in higher dimensions. Significant reduction in acquisition time of a multi-dimensional NMR experiment is a prerequisite for its routine usage. Another requirement is to achieve a better sensitivity-resolution-time balance for 3D (4D) spectroscopy, similar to that achieved by contemporary 2D (3D) spectroscopy. The ability to achieve high resolution while avoiding the time barrier in higher dimensionality NMR experiments would significantly facilitate protein spectral assignment and structure determination. This, in particular, would further advance the generally acknowledged role of NMR in structural genomics (Christendat et al., 2000; Kennedy et al., 2002; Szyperski et al., 2002; Yee et al., 2003; Peti et al., 2004).

The prohibitively long acquisition time of a high-dimensional NMR experiment is due to a large number of uniformly sampled points in several indirect dimensions. A 4D of the same experiment type as a 3D contains the same number of spectral peaks but takes much more measurement time. Thus, if adequate sensitivity of an experiment is achieved for the 3D version, there would be a large redundancy in the 4D, in proportion to the number of points in the extra dimension. The situation is known as so-called ''sampling limited'' regime. For a given experiment type reduction of the measurement time can be achieved by reducing number of sampled points either by truncating acquisitions in the indirect dimensions or by omitting some points inside a full FID. In both cases experimental sensitivity roughly scales as a square root of the number of retained points. However, in the first case spectral resolution is largely sacrificed. In the second case, which is referred to as non-uniform or sparse sampling (Hoch and Stern, 1996; Rovnyak et al., 2004), the resolution can be retained. To summarize, by omitting enough points one can scale down the experimental time of a 4D to equal the one of a 3D, while keeping the resolution and most of the sensitivity of the 3D. This generally solves the time problem. Further improvement can be achieved if the sampling points are selected in an optimal way, e.g. by using ideas of matched acquisition (Barna et al., 1987; Schmieder et al., 1994).

Spectra recorded in the non-uniform mode cannot be directly processed by the regular Discrete Fourier Transform (DFT). However, there are at least two methods: Maximum Entropy reconstruction (ME) (Laue et al., 1986; Hoch and Stern, 2001) and multi-dimensional decomposition (MDD), which have been proven successful in dealing with non-uniformly sampled data. ME was recently applied to a suite of non-uniformly sampled 3D triple resonance experiments for protein backbone assignment (Rovnyak et al., 2004). MDD has been demonstrated for a representative region of a 3D ^{15}N NOESY-HSQC of 14 kDa protein (Orekhov et al., 2003) and applied for analysis of 4D ^{13}C Methyl-NOESY spectrum of deuterated 82 kDa protein MSG (Tugarinov et al., 2005). In these cases a measurement time saving of 70–80% was achieved without sacrificing spectral resolution and sensitivity. In other words, most of available resolution was obtained, while the sensitivity corresponded to the actual measurement time.

Unlike maximum entropy reconstruction method, which has been developed over the past 20 years, MDD is yet to be developed to such a level. The justification for further development of the MDD method comes from its ability of correctly reproducing crowded spectra with high-dynamic range of intensities. These are encountered for example in the NOESY type experiments, which are the major source of information for NMR structure calculations. It is worth mentioning that for the experiments of the NOESY-HSQC type, the most natural and preferable dimensionality is four, where both interacting protons can be identified based on the generally good signal dispersion in a 2D HSQC spectrum. Two and three-dimensional NOESY schemes often result in ambiguous peak assignments even for non-overlapped signals.

So far the following questions were not addressed: if the method is applicable to fully protonated protein samples exhibiting large number of signals in 4D ^{13}C NOESY spectrum; how accurate are the spectral intensities in the different spectral regions; are there false peaks in the spectral reconstruction; how does the sensitivity scale with the amount of missing data; is the method robust with respect to variations of the input data and parameters? Here we attempt to address these questions and further develop the MDD methodology to the level of ready-to-use processing protocols. The method is demonstrated for a uniformly labelled 106-residue protein taken as a typical representative from a structural proteomics pipeline. (Christendat et al., 2000; Gerstein et al., 2003).

Theory

MDD is a mathematical concept for approximation of three- or higher dimensional matrix by a product of one-dimensional vectors (Kruskal, 1977).

For the three-dimensional case the MDD can be formulated as follows. Given a matrix $\boldsymbol{S}$ with elements $s_{k,m,n}$ ($k = 1 \ldots K$, $m = 1 \ldots M$, $n = 1 \ldots N$), find numbers a^{β} and normalized vectors $\boldsymbol{F1}^{\beta}$, $\boldsymbol{F2}^{\beta}$, $\boldsymbol{F3}^{\beta}$ with elements $f1_k{}^{\beta}$, $f2_m{}^{\beta}$ and $f3_n{}^{\beta}$, respectively, such that the following norm becomes minimal.

$$|\boldsymbol{G} \bullet [\boldsymbol{S} - \Sigma_{\beta}(a^{\beta}\boldsymbol{F1}^{\beta} \otimes \boldsymbol{F2}^{\beta} \otimes \boldsymbol{F3}^{\beta})]|^2 + \lambda\Sigma_{\beta}(a^{\beta})^2 \quad (1)$$

Here, the symbol $\otimes$ denotes tensor product operation; the matrix $\boldsymbol{S}$ corresponds, for example, to an experimental three-dimensional NMR spectrum in time or frequency domain. In the case of sparse sampling only a fraction of elements in $\boldsymbol{S}$ is measured and the matrix $\boldsymbol{G}$, which contains elements $g_{k,m,n} \in \{0,1\}$, indicates the absence or presence of a particular data point. Accordingly, the symbol $\bullet$ describes element-wise multiplication of matrices. The last term represents a Tikhonov regularization, which is parameterized with the factor λ (Tikhonov and Samarakij, 1990) and may be used for improving the convergence of the MDD algorithm (Ibraghimov, 2002).

MDD has been introduced as a tool for data analysis in the early seventies under various names such as parallel factor analysis, canonical decomposition or three-way decomposition. Theoretical considerations concerned notably questions of uniqueness of optimal approximations, generalization for higher dimensions, and development of efficient algorithms for solving the least square minimization problem defined by Equation 1. Notably, an efficient algorithm for dealing with large fraction of missing data for matrices of any dimensionality higher or equal to three has been recently introduced (Ibraghimov, 2002). This algorithm is implemented in the mddNMR software, which is used in this work. Since the advanced mathematical details are outside of the scope of this work, we present only a brief outline here.

The sum in Equation 1 represents the fundamental model assumption of MDD: direct products of one-dimensional vectors are sufficient to describe all features of a high-dimensional matrix. In the following we refer to $\boldsymbol{S}$ as the (input) spectrum and to the entities in the sum over β as (output) *amplitudes* a^{β} and *shapes* $\boldsymbol{F1}^{\beta}$, $\boldsymbol{F2}^{\beta}$ and $\boldsymbol{F3}^{\beta}$, while the summation terms are called *components*. Note that while the input sparse data matrix $\boldsymbol{S}$ may lack many entries, the shapes $\boldsymbol{F1}^{\beta}$, $\boldsymbol{F2}^{\beta}$ and $\boldsymbol{F3}^{\beta}$ representing the output of the MDD are complete, allowing reconstruction of a full matrix. The only non-restraining condition for the completeness of the output shapes is that every plane in $\boldsymbol{S}$ contains at least one data point. The amplitudes a^{β} result from the use of normalized shapes $\boldsymbol{F1}^{\beta}$, $\boldsymbol{F2}^{\beta}$ and $\boldsymbol{F3}^{\beta}$. The summation index β runs over the number of components used for the decomposition. The range for this index depends on the type of spectrum. Notably the number of components may be significantly less than the number of peaks in a multi-dimensional experiment. This is true, for example, for the 3D ^{1}H–^{15}N NOESY-HSQC spectrum (Orekhov et al., 2003), where one component comprises all peaks, including diagonal, sharing ^{1}H and ^{15}N frequencies of one amide group. Thus, the number of components for this spectrum is defined by the number of peaks in a corresponding HSQC spectrum.

Relation between NMR signals and MDD

MDD and multi-dimensional NMR spectroscopy are intimately related, as can be shown by deducing the model assumption for the MDD (Equation 1) from the description of NMR experiments. This connection can be shown on a general theoretical

level; however, in the following we demonstrate this on an example of the 4D experiment used for this work, 4D HCCH-NOESY. During the evolution delays t_1 and t_2 initial proton magnetization of an aliphatic group CH_n ($n = 1,2,3$) is labeled by the proton and carbon chemical shifts; subsequently the magnetization is transferred via ^1H–^1H NOE cross-relaxation to another CH_n group, where it is labeled by the second carbon chemical shift (evolution time t_3) and finally detected as a transverse proton magnetization during the acquisition time t_4. Considering flow of magnetization in this experiment, the first and second CH_n groups are denoted as the origin and destination, respectively. If all the origin (destination) CH_n groups are enumerated by indices α (β), the 4D model spectrum $\boldsymbol{S'}$ (to be distinguished from the experimental spectrum $\boldsymbol{S}$) in the time and frequency domain is given by

$$S'(t_1,t_2,t_3,t_4) = \Sigma_{\alpha\beta} A_{\alpha\beta} L^{\alpha}(t_1) L^{\alpha}(t_2) L^{\beta}(t_3) L^{\beta}(t_4) \tag{2a}$$

$$S'(\omega_1,\omega_2,\omega_3,\omega_4) = \Sigma_{\alpha\beta} A_{\alpha\beta} L^{\alpha}(\omega_1) L^{\alpha}(\omega_2) \times L^{\beta}(\omega_3) L^{\beta}(\omega_4) \tag{2b}$$

where summations run over all pairs α, β of interacting aliphatic groups; $A_{\alpha\beta}$ represent amplitudes of the peaks, and $L(t)$ and $L(\omega)$ denote their normalized line-shapes in the time and frequency representations. While the amplitudes $A_{\alpha\beta}$ are defined by the amount of initial magnetization and efficiencies of the magnetization transfer during fixed delays for J-coupling evolutions and NOE cross-relaxation, etc., the shapes L are determined by the resonance frequencies and magnetization decays exhibited during the evolution delays t_1–t_4. Notably, Equations 2a and 2b make no specific assumptions about the functional form of the line shapes and, consequently, look formally the same in the time and frequency domains.

Equations 2 show a direct link between the four-dimensional experiment and the MDD in four-dimensions (4DD), where each component or peak is described by its line shapes in all four dimensions. However, exactly because every peak has to be described by an individual component, the 4DD is not suitable for extracting cross-peaks from the 4D NOESY spectra, which contains the diagonal and large number of cross-peaks of significantly different intensities. From our experience only intense signals (e.g. NOESY diagonals) can be reliably identified if the 4DD is used. It is much more attractive to use the 3DD, in which case diagonal signal and all cross-peaks sharing the line shapes of the destination CH_n group β are collected into a single component. Indeed, Equations 2 can be rewritten as:

$$S'(t_1,t_2,t_3,t_4) = \Sigma_{\beta} a_{\beta} L^{\beta}(t_1,t_2) L^{\beta}(t_3) L^{\beta}(t_4) \tag{3a}$$

$$S'(\omega_1,\omega_2,\omega_3,\omega_4) = \Sigma_{\beta} a_{\beta} L^{\beta}(\omega_1,\omega_2) \times L^{\beta}(\omega_3) L^{\beta}(\omega_4) \tag{3b}$$

where normalized two-dimensional shapes $L^{\beta}(t_1,t_2)$ and $L^{\beta}(\omega_1,\omega_2)$ and new amplitudes α_{β} are introduced so that $a_{\beta} L^{\beta}(t_1,t_2) = \Sigma_{\alpha} A_{\alpha\beta} L^{\alpha}(t_1) L^{\alpha}(t_2)$ and $a_{\beta} L^{\beta}(\omega_1,\omega_2) = \Sigma_{\alpha} A_{\alpha\beta} L^{\alpha}(\omega_1) L^{\alpha}(\omega_2)$.

So far the line shapes L were defined as functions of continuous time or frequency arguments. In practice, the evolution times are defined on a grid with regular intervals Δt:

$$t_1 = (p-1)\Delta t_1, \quad t_2 = (q-1)\Delta t_2, \quad t_3 = (m-1)\Delta t_3, \quad t_4 = (n-1)\Delta t_4 \tag{4}$$

where indices p, q, m, n run from 1 to the maximal values P, Q, M, N defined by the maximum acquisition times in dimensions 1–4, respectively. Analogous grid can be envisaged in the frequency domain, although dimension sizes P, Q, M, N could have different values. In addition, for the following we introduce an index k that runs over $K = P^*Q$ values, with k an element of (p, q). Namely, $k = 1$ for $p = 1$, $q = 1$; $k = 2$ for $p = 1$, $q = 2$; … $k = P^*Q$ for $p = P$, $q = Q$. Using indices k,m,n, Equation 3a can be written as:

$$S'_{k,m,n} = \Sigma_{\beta} a_{\beta} f1^{\beta}_k \cdot f2^{\beta}_m \cdot f3^{\beta}_n \tag{5a}$$

where the following substitutions were made $f1^{\beta}_k = L^{\beta}([p-1]\Delta t_1, [q-1]\Delta t_2)$, $f2^{\beta}_m = L^{\beta}([m-1]\Delta t_3)$, $f3^{\beta}_n = L^{\beta}([n-1]\Delta t_4)$. Finally, by introducing vectors $\boldsymbol{F1}^{\beta}$, $\boldsymbol{F2}^{\beta}$, $\boldsymbol{F3}^{\beta}$ with elements $f1^{\beta}_k$, $f2^{\beta}_m$, $f3^{\beta}_n$, respectively, the model spectrum of Equations 2, 3, 5a is presented in the form equivalent to the MDD model used in the least square minimization defined by Equation 1:

$$\boldsymbol{S'} = \Sigma_{\beta} (a^{\beta} \boldsymbol{F1}^{\beta} \otimes \boldsymbol{F2}^{\beta} \otimes \boldsymbol{F3}^{\beta}) \tag{5b}$$

Considering the time grid defined in Equation 4 the input spectrum $\boldsymbol{S}$ can be presented as a four dimensional matrix with elements $s_{p,q,m,n}$ or, alternatively, as a three-dimensional array with elements $s_{k,m,n}$. The later representation, being used together with Equation 5b, allows processing of a 4D spectrum by the three-dimensional version of MDD (3DD). The relation between the 4D spectrum and output shapes, described above, gives a clue for interpretation of the MDD output. In particular, the vector $\boldsymbol{F1}^{\beta}$ corresponds to a two-dimensional ^{1}H–^{13}C correlation spectrum in the time domain, where signal intensities are proportional to efficiencies of the NOE transfer from the protons of all the aliphatic groups α to the destination group β. Relation between the elements of one-dimensional vector $\boldsymbol{F1}^{\beta}$ and the two-dimensional matrix is given by Equations 3–5. Vectors $\boldsymbol{F2}^{\beta}$ and $\boldsymbol{F3}^{\beta}$ correspond respectively to the 1D carbon and proton time domain profiles (t_3 and t_4). Notably, only one pair of shapes for these two dimensions is adjusted for all signals described by a particular vector $\boldsymbol{F1}^{\beta}$. This explains why week NOE signals are reliably described by the 3DD, but may be missing if the 4DD is used in which case all shapes have to be adjusted for every individual peak.

The Equations 1 and 5 have the same form both in frequency and time domains or mixture of those for different dimensions. In practice it is convenient to perform Fourier transform in the non-sparsed dimensions. This speedups the calculations and allows removal of regions with artefacts, e.g. at the edges of the spectrum or near the water signal. Note that the data along the directly acquired dimension, i.e. t_4 for the 4D spectrum, is always complete and thus can be processed by regular Fourier transform prior to the MDD calculations.

Materials and methods

Reference 4D spectrum

A sample with uniformly $^{15}N/^{13}C$ labelled 106-residue protein PA1123 was generated by the Northeast Structural Genomics Consortium. A complete 4D HCCH-correlation NOESY spectrum (Vuister et al., 1993) was recorded at 25 °C on a Varian INOVA 600-MHz spectrometer. An acquisition time of 64 ms was used in the direct ^{1}H dimension, along with a 150 ms NOE mixing time and a relaxation delay of 1 s between the transients. The recorded spectrum contains 448, 50, 18 and 18 complex points along t_4 (^{1}H), t_1 ($^{1}H^{ind}$), t_2 (^{13}C) and t_3 (^{13}C) dimensions, respectively. The corresponding spectral widths are 11.6, 5.2, 19.9 and 19.9 ppm. With the four step phase cycle the total measurement time was 7.5 days.

The spectrum was processed using the nmrPipe software package (Delaglio et al., 1995). Signals in the time domain in ^{1}H and $^{1}H^{ind}$ dimensions were multiplied by a square sine and a sine weighting function, respectively, then zero filled to 512 and 64 points prior to Fourier transformation. In the t_2 (^{13}C) dimension signals were multiplied by a sine weighting function, zero filled and Fourier transformed. Subsequently, the time domain signals in t_3 (^{13}C) dimension were extended by 25% to 22 points by linear prediction (LP) using four coefficients, multiplied by a square sine function, zero filled to 32 points and Fourier transformed to frequency domain. After that, Hilbert transform was applied to the signals in t_2 (^{13}C) dimension, and the frequency domain data in this dimension were converted back into original time domain using the exact inverse of previous processing procedures. The final spectrum was obtained by processing the time domain signals in t_2 (^{13}C) dimension as following: linear predicted 4 points with eight coefficients, multiplied by a sine function, zero filled to the same size as in t_3 (^{13}C) dimension and Fourier transformed.

The processed spectrum in frequency domain will be referred to as ''reference'' spectrum. The standard deviation of the noise in this spectrum, which was defined by averaging the noise in ten arbitrarily chosen planes, is denoted as σ_{ref}. The value (equals to 2.42×10^4 arbitrary units of intensity) will be used as a unit of spectral intensity throughout the paper.

MDD reconstruction of the 4D spectrum

Processing of a spectrum recorded in the sparse mode with the MDD consists of five major steps, which are depicted in Figure 1 and described in the following paragraphs in more details: (i) Fourier transform in the acquisition dimension (t_4) and splitting of the spectrum into several regions; (ii) estimation of number of components (N_c) for every region; (iii) MDD calculations, i.e.

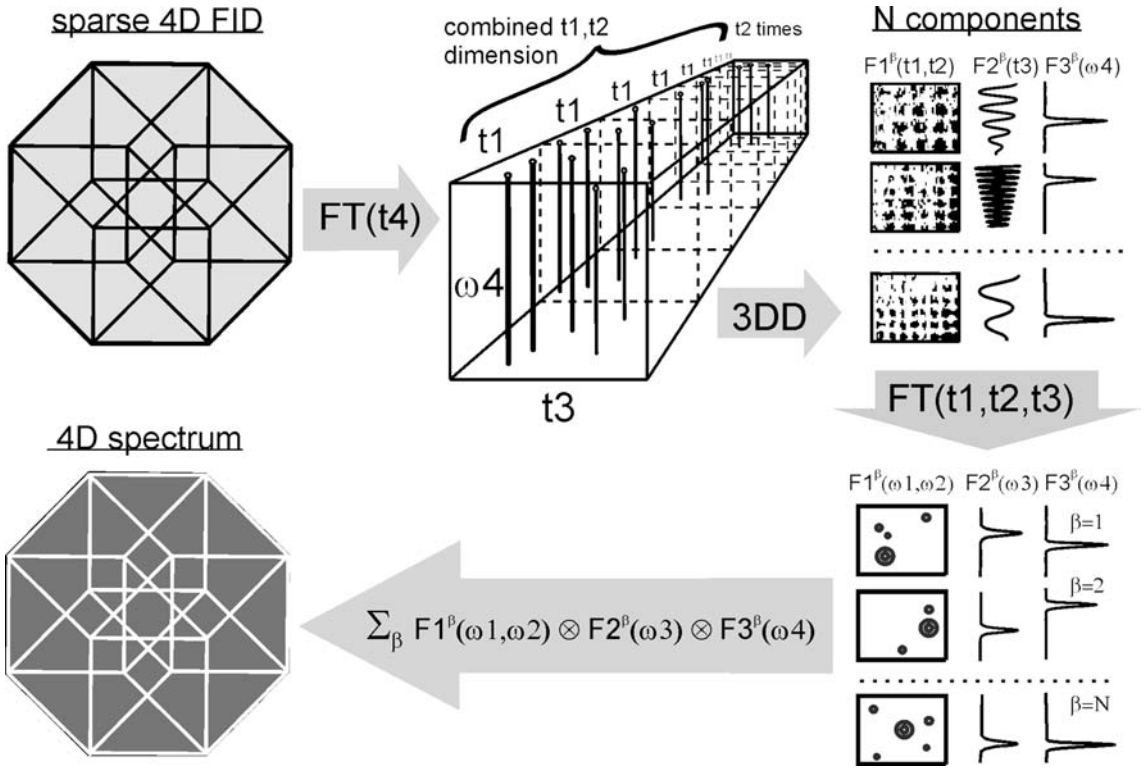

Figure 1. Schematic of the procedure by which the frequency domain spectrum of a sparsely sampled multi-dimensional data set is reconstructed. See text for details.

minimization of the Equation 1; (iv) processing of individual $\boldsymbol{F1}^{\beta}$ ($\boldsymbol{F2}^{\beta}$) shapes as regular 2D (1D) time series; (v) reconstruction of the spectrum for every region using the resulting set of shapes in the frequency domain and combining all regions into final 4D spectrum. In addition specifically in this work, sparse data is extracted from a complete spectrum prior to step (ii) and the resulting reconstructed 4D spectrum is compared with the reference for the evaluation of the quality of the reconstruction. Note that normally the sparse FID is obtained directly from experiment. Steps (i) and (iv) are performed using a standard processing tool nmrPipe (Delaglio et al., 1995). For the remaining steps different routines from the home-built software mddNMR are used. The mddNMR package is to be released soon.

Region selection in ω_4

First the acquisition t_4-dimension of the spectrum is processed in the same way as for the reference spectrum (see above). Then the 10 overlapping aliphatic regions in the range −0.5 to 4.5 ppm in the ω_4 dimension were extracted. Each of the eight central regions covered 0.7 ppm, and two flanking regions covered 0.6 ppm. (See first column in Table 1 for spectral range of each region.) All regions had a 0.2 ppm overlap with their immediate neighbours in order to eliminate boundary effects in the reconstructed spectrum. For methodological purposes, variations of several parameters were performed on one of the regions, which in the following is referred to as a *representative*. This region covers range of chemical shifts from 0.4 to 1.1 ppm in the acquisition dimension ω_4. The division of the spectrum into a set of regions is beneficial only for reduction of MDD computation time.

Determination of number of components and Tikhonov regularization factor λ

The number of components (N_c) is an important parameter for MDD. It should be sufficiently high for describing all signals in the spectrum. On the other hand N_c values should be kept low for

Table 1. The results of the MDD reconstruction in the 10 regions of 4D HCCH-correlation NOESY spectrum of the protein PA1123

ω_4 region[a] (ppm)	Number of components N_c used for 3DD	σ_{dif}[b]
−0.5–0.1	4	1.08
−0.1–0.6	16	1.15
0.4–1.1	50	1.19
0.9–1.6	36	1.18
1.4–2.1	50	1.21
1.9–2.6	42	1.18
2.4–3.1	26	1.15
2.9–3.6	19	1.18
3.4–4.1	43	1.28
3.9–4.5	29	1.63

[a]The ranges of the extracted regions in ppm along ^{1}H-direct dimension.
[b]The value of σ_{dif} is normalized as defined in *Materials and methods* section.

reducing the number of adjustable parameters and avoiding over-parameterization in the minimization of Equation 1. As it follows from the theory presented above, the 4D ^{1}H–^{13}C HSQC-NOESY-HSQC spectrum can be adequately described by the 3DD with approximately so many components as the number of peaks in the corresponding 2D ^{1}H–^{13}C HSQC spectrum. Thus, the most straightforward and simple method of estimating the N_c is counting the number of peaks in the corresponding region of the HSQC spectrum. This method has been used in our previous publications (Orekhov et al., 2003; Tugarinov et al., 2005). As an equal alternative, in this study N_c is estimated from the same input 4D sparse data. N_c is assumed to correspond to the number of strong diagonal signals observed in the reconstruction of the 4D spectrum, which is obtained using a short run of the 4DD. The 4DD is fast converging as for given input data and number of components, it uses much fewer adjustable parameters than 3DD, i.e. $N_c \times (P+Q+M+N-3)$ for 4DD versus $N_c \times (P \times Q+M+N-2)$ for 3DD. Consequently, 4DD is more robust in describing bright spectral features. The protein used in this study has approximately 300 proton–carbon pairs with unique H, C chemical shifts. Thus, for each of the 10 regions, the average number of components N_c is 30. In the 4DD two times higher N_c values were used for each region to ensure sufficient number of components for reconstruction of all diagonal peaks. For these calculations the value of λ was set to 0.001, and each run had 200 iterations. The number of diagonal peaks was determined by peak-picking script *pkFindROI* from the nmrPipe package (Delaglio et al., 1995) using the threshold of three noise levels with other parameters set at their default values. Finally, for every region the number of observed strong diagonal peaks was increased by 30% to describe small intensity features and used as the N_c (see second column of Table 1) for the further 3DD calculations.

Apart from the number of components the Tikhonov regularization factor, denoted as λ in Equation 1, is the only parameter of the MDD calculations. Whereas for complete data, the value of λ mainly determines how fast convergence of the algorithm is achieved, for sparse data usage of regularization also improves quality of the solution. In our previous studies on sparse 3D and 4D NOESY spectra (Orekhov et al., 2003; Tugarinov et al., 2005) the optimum of the λ-values was determined as the broad range: 10^{-1}–10^{-3}. In this work, if not specified otherwise, $\lambda = 0.01$ was used for all runs. In addition, to check how optimal this value is, we optimized the λ-value for the representative region with the sparse level of 0.3, i.e. 30% of the data. For this purpose, quality of spectral reconstruction was evaluated for a series of 3DD runs with systematic variation of the λ value.

MDD calculations and computational costs

For a given input spectrum and defined values of N_c and λ, the MDD calculations (3DD or 4DD) that is minimization of Equation 1 is a batch job performed by the mddNMR software. On output a file is produced, which contains amplitudes and 1D vectors of shapes for all components. The amount of computation depends on specified number of iterations in the minimization. Convergence is usually achieved within the first 50–100 iterations out of total 1000 used in this study. The 3DD calculations with 1000, 2000 and 5000 iterations gave practically identical reconstructed 4D spectra of the representative region. A typical calculation with 1000 iterations at a Linux cluster takes 2 h using 16 Intel Xeon 2.2 GHz processors.

Processing of the 1D shapes and reconstruction of 4D spectrum

The one-dimensional shapes produced by the MDD for individual components are either ready frequency line shapes (i.e. for ω_4) or regular complex FID time series (Figure 1). Using nmrPipe processing scripts the later are converted into 1D frequency line shapes. In the special case of 3DD applied to the 4D spectrum, the first shape (***F1***) corresponds to a 2D hyper-complex FID and, consequently, is processed as a 2D spectrum. Subsequently, for each component a 4D matrix is produced by the tensor product of all three (or four for 4DD) frequency shapes. Sum of all matrices for individual components gives the final 4D reconstruction of the spectrum for the region. It is possible to reconstruct first the complete 4D spectrum in the time domain and then process it as usual. However, while the final result is almost identical, the processing of shapes gives significant practical advantage since 1D Fourier processing is much faster, it does not require Hilbert transform and inverse processing in one of the carbon dimensions. It is also worth noting that the former approach allows almost immediate zooming into any region of the frequency domain spectrum without having to process the whole spectrum beforehand.

The processing parameters for the shapes, i.e. weighting functions, zero-filling, phasing and linear prediction, were identical to those used for the processing of the corresponding dimensions of the complete reference spectrum. A short linear prediction was used for avoiding well-known artifacts that might occur with a more extensive use of the LP, i.e. t_1 noise ridges and uncontrollable distortions of line shapes. These artifacts, which would have nothing to do with the MDD processing, might interfere with quantitative comparison of the reference and the reconstructed spectra. Note that both LP and zero filling were applied to the output of the MDD calculations, i.e. to the shapes, and do not effect MDD calculations *per se*.

Comparison with the reference spectrum

All sparse data sets in this work were produced as subsets of the complete reference 4D spectrum described above. In each case a specified amount of complete 1D FIDs (t_4) were extracted from this spectrum. The sparse level, which varies from 0 (no data) to 1 (full data), is defined as a fraction of the selected to the total number of FIDs. The FIDs were selected in accordance with a random distribution, which was exponentially biased to match the transverse relaxation times of 50, 10 and 10 ms in dimensions t_1, t_2 and t_3, respectively (Tugarinov et al., 2005).

Both reconstructed and reference spectra are organized in the frequency domain as sets of 2D spectral planes (64×128 data points) along the proton (ω_1) and carbon (ω_2) dimensions of the origin aliphatic group. In total, a typical region of the 4D spectrum contains $64 \times 63 = 4032$ of these planes, which corresponds to the number of points in the two remaining dimensions ω_3 and ω_4. Quality of the reconstruction was assessed using two quantitative measures: the normalized (i.e. divided by σ_{ref}) standard deviation of intensities in the difference spectrum σ_{dif} and the noise level in the reconstructed spectrum σ_{rec}. Note that while the noise level is uniform throughout most of the reference spectrum, it differs among the 2D planes of the reconstruction. Namely, significant denoising occurs in the planes without strong signals that are described as MDD components. To obtain a more meaningful noise estimate for planes with signals, σ_{rec} values were averaged only at the planes containing the diagonal signals using subroutine *estNoise* from nmrPipe software. When applied to the same planes of the reference spectrum this procedure gave the noise estimate very close to σ_{ref}.

Cross-peaks in the *representative region* of reference and reconstructed spectra were identified and characterized using script *pkFindROI* from the nmrPipe package (Delaglio et al., 1995). Peaks are picked using *Reject Noise Peaks* mode at the Chi2 level of 0.01, with the intensity threshold of $2\sigma_{ref}$ and the noise level σ_{ref}. The cross-peaks found in the planes without diagonal signals were excluded from the resulting peak list. False entries in the peak list obtained from the reconstructed spectrum were identified using two thresholds, i.e. as those with peak intensities higher than $1.3*(I_{ref}+3)$ and $1.3*(I_{ref}+5)$, where I_{ref} is intensity at the corresponding position in the reference spectrum. Missing peaks are defined as peaks present in the reference spectrum but missing in the reconstruction. Quantitatively, a peak with amplitude A_{ref} from the reference peak list was defined as missing if the intensity at the corresponding position in

the reconstructed spectrum was lower than $0.7*(A_{ref} - 3)$. To check how many peaks are expected to be lost at different sparse levels because of the increased noise, the intensities of the peak list entries were compared against thresholds increasing as 2*sqrt(1/ < sparse level >).

Accuracy of the line width reconstruction was assessed by comparison of the corresponding values of the cross-peaks matched by all frequencies in the peak lists obtained for the reconstructed and reference spectra. The match frequency tolerances were 0.02 and 0.24 ppm for the proton and carbon dimensions, respectively.

Results and discussion

Applicability of the MDD for processing of multidimensional spectra is determined by its performance in reproducing all above-noise spectral features without introducing signal distortions and false peaks that could hamper the spectra interpretation. In this study we compare the complete reference spectrum and a number of 4D spectral reconstructions, which were obtained using the 3DD applied to several sparse data sets. This comparison addresses the following questions: (i) are all of the signal intensities correctly reproduced in the full dynamic range, and are there false peaks that significantly exceed the level of the baseline noise; (ii) how does the reconstruction quality depend on the sparse level; (iii) is there a variation in the quality of reconstruction from one region of the spectrum to another, e.g. due to different signal density or presence of spectral artefacts; (iv) how sensitive are the results to the settings of the MDD parameters, i.e. the number of components and the λ-value?

4D reconstruction results

Accuracy of NOE intensity reconstruction

The most important aspect of quality of spectra obtained from sparsed data is accuracy of reconstruction of NOE peaks. We applied a commonly used procedure that is to detect peaks in both reference and reconstruction spectra using a peak-picking procedure, match them and to compare the peak parameters obtained. However, as usage of any particular peak-picking program makes a quality estimation to be dependent on and specific for that program, the analysis was complemented by a more direct measure of reconstruction quality. Namely, we perform data-point to data-point comparison of spectral intensities between the reconstructed and reference spectra for all 230686720 points in the region from −0.5 to 4.5 ppm (in ω4) of the 4D spectra. The correlation factor $R = 0.997$ was calculated for the range of amplitudes starting from just about the noise level ($2.2\sigma_{rec}$) till highest diagonal peaks ($1300\sigma_{rec}$) in either reference or reconstructed spectrum. The subset of points with relatively low intensities in the range $2.2–110\sigma_{rec}$ correspond to NOE cross-peaks. The correlation factor R for this range is 0.984.

Correspondence of intensities for all data points in the two data sets indicates the correctness of reconstruction for all peak parameters: intensities, line-shapes and peak positions. In the following this is further verified by comparing parameters of the cross-peaks identified in the representative region of the reference and reconstruction spectra. Figure 2 presents the reconstruction–reference correlation plot between the NOE peak intensities for 30% sparse data. A dot significantly above the diagonal line would indicate a false peak, i.e. having significantly weaker intensity in the reference spectrum at position of the peak in the reconstruction. The distance between a dot and the diagonal reflects accuracy of the intensity reconstruction. Figure 2 and Table 2 illustrate a great advantage of the MDD method that almost no or very small number of weak false signals are generated in the reconstructed spectrum for all used sparse levels.

Equally important question is how many peaks that are found in the reference spectrum are missing in the reconstruction. Out of total 219 cross-peaks found in the representative region 10 correlations were missing already for the sparse level 100%, i.e. for the MDD reconstruction obtained using the complete data set. These peaks were strongly attenuated in the reconstruction together with seven corresponding weak diagonal signals. These diagonals were the weakest of all, with intensities in the reference spectrum below 63 (the level is five times below the average diagonal intensity), and apparently correspond to the signals from minor conformations, small fraction of degraded protein or impurities in the sample. Although these peaks can be recovered when an analysis of minor signals is necessary by using

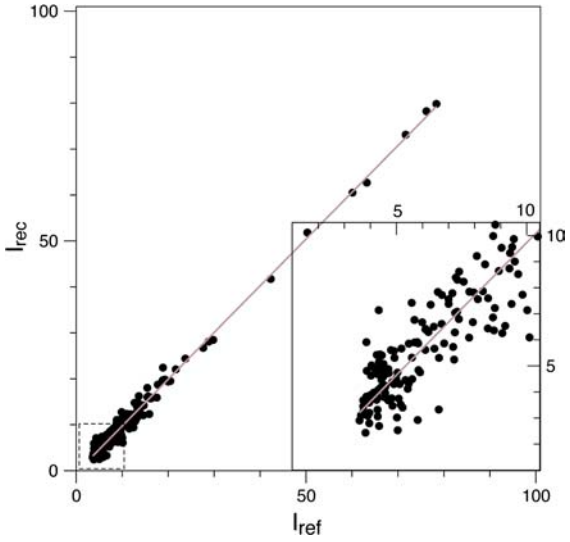

Figure 2. Accuracy of the peak intensities in the 30% sparse reconstruction. The intensities are correlated at the cross-peak positions in the representative region of the reference and reconstructed spectra. The intensities are measured in units of standard deviation of noise in the reference spectrum, σ_{ref}. Inset: magnification of the intensity range from 1 to 10.5. The cross-peaks were obtained by the peak-picking in the reconstructed spectrum (see *Materials and methods* for the procedure parameters).

more components for the region (data not shown), we decided to exclude them from the further consideration. Question regarding selection of a number of components is considered below.

As sparse level gets lower, the measurement time decreases, noise in the reconstructions becomes higher and the weakest peaks may disappear in the noise. As it is shown in Table 2 the small percentages of missing peaks in the reconstructions correspond very well to the losses anticipated due to the shortening of the measurement time. Thus, four peaks (1.9% of 209) are missing for 30% sparse while the same number of peaks are not found in the reference spectrum if the sqrt(1/0.3) = 1.82 times higher threshold is used for the peak-picking. Note also that 1–2% values are on the order of the errors in the peak-picking procedure.

Values of NOE peak intensities or integrals carry primary information about interproton distances in protein structure. It is thus important that NOE intensities are well reproduced in full range of peak amplitudes. Correlation coefficients were calculated between intensities of the NOE peaks in the reference spectrum and intensities at

Table 2. The quality of the MDD reconstruction for the representative region (0.4–1.1 ppm in ω_4) at different sparse levels quantified as percentage of missing and false cross-peaks

Sparse level	Percentage[a] of missing and false peaks[b]			
	Missing observed	Missing expected[c]	False ($>5\sigma_{ref}$)	False ($>3\sigma_{ref}$)
1.00	0.0	0.0	0	1.9
0.70	1.9	1.4	0	1.9
0.50	4.3	1.4	0	2.4
0.30	1.9	1.9	0	1.0
0.20	3.8	3.3	0	4.3
0.15	9.6	10.5	0	10.0

[a]The values are normalized to 209 – the total number of cross-peaks detected for the same region in reference spectrum (see text).
[b]See *Materials and methods* for the criteria for attributing peaks to different groups.
[c]Estimate of number of peaks lost due to shorter measurement time as described in *Materials and methods*.

the corresponding positions in the 30% sparse reconstruction. Out of total 209 peaks in the representative region of the reference spectrum 131 peaks with intensities I_{pk} in the range 3–10 were grouped as weak; 72 peaks ($10 < I_{pk} < 100$) were grouped as medium; 6 peaks ($I_{pk} > 100$) were strong. For the groups of weak, medium and strong peaks the correlation coefficients are 0.83, 0.9965, and 0.9997, respectively. Considering that the peak amplitudes are normalized to the noise level in the reference spectrum while the noise in the 30% reconstruction is 1.82 times larger, the correlation is good even for very weak peaks.

Accuracy of the peak line widths is another important aspect to be considered when judging quality of the reconstructed spectrum. Average relative errors in the line width of 0.1($^1H^{ind}$, ω_1), 0.1(^{13}C, ω_2), 0.1(^{13}C, ω_3), and 0.15(^1H, ω_4) were calculated for the representative region (see *Materials and methods*). These error values closely correspond to the digital resolution in the spectral dimensions and are almost independent of the peak amplitudes. Furthermore, noise level intensities in the difference spectrum (Figure 5) indirectly indicate good correspondence both for the peak intensities and line widths. A typical example of a nearly empty plane from the difference spectrum is shown in Figure 3a. Figure 3c illustrate that the cross-peak line shapes are reproduced in the wide range of

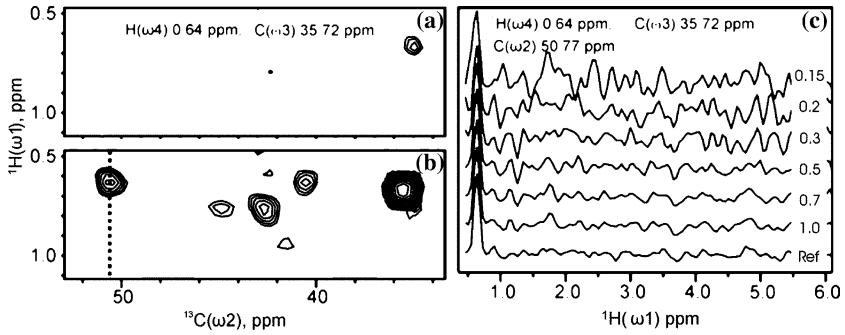

Figure 3. An example of a 2D strip from the 30% sparse 4D spectrum (b), the difference spectrum (a), and 1D profiles with a cross-peak for the reference spectrum and the reconstructions for various sparse levels 0.15–1.00 (c). The strip in a, b is at ω_4 = 0.64 ppm and ω_3 = 35.72 ppm, extracted from 0.47 to 1.1 ppm along the ω_1 and had full size in ω_2 dimension. The vertical line at ω_2 = 50.77 ppm indicates the location of the ω_1 cross-sections shown in c. To show noise the size of the 1D cross sections is full, unlike that of the strip. Data for different sparse levels, as indicated by number on the right, are shown with a constant vertical offset. The first contour (in a, b) was at three standard deviations of noise in the reference spectrum ($3\sigma_{ref}$), while contour multiplication factor was 1.3.

sparse levels. Finally, it is worth noting that since the MDD procedure makes no assumptions about the line-shapes, it is very unlikely to result in biased line widths.

Noise in the reconstruction and the sparse level

To analyse the effect of missing data on the MDD results, a series of 3DD runs were made for the representative region (0.4–1.1 ppm) with varying sparse level (0.15, 0.20, 0.30, 0.50, 0.70 and 1.00). In these calculations N_c was 50 and λ was 0.01.

To illustrate the effect of different sparse ratios, a strip of 2D plane in the reconstructed spectrum with 0.64 ppm and 35.72 ppm in ^{1}H (ω_4) and ^{13}C (ω_3) dimensions was extracted (Figure 3b). The cross-sections of the strip at ω_2 = 50.77 ppm in the plane for various fractions of sparse data were taken and displayed in Figure 3c. The cross-sections show that with the sparse level decreasing, the baseline noise in the corresponding spectrum gradually increased. Note that while very weak peaks may disappear as they submerge into the noise, the intensity and linewidth of the cross-peak is preserved for all sparse ratios. The Figure 3a shows the difference spectrum. The difference of 3 contours is visible for the strong diagonal signal.

The Figure 4 shows observed noise value σ_{rec} in the reconstructed spectrum as a function of the sparse level. When the sparse level decreases so does the data measuring time. It is therefore not surprising that the σ_{rec} is higher for low sparse levels. Indeed, it is well known that for the regular spectra the noise is proportional to the inverse of square root of the total measurement time. As a benchmark a dashed line in Figure 4 was calculated according to such relationship. This represents an anticipated noise increase in the reference spectrum recorded in shorter time, e.g. with fewer transients. The dependences for the measured σ_{rec} and anticipated noise on measurement time show similar trend and largely overlap. This allows reducing total time of the experiment in the so called *sample limiting* situation without sacrificing resolution. In addition sensitivity per unit time is conserved. Thus, total measurement time of a sparse 4D experiment is defined only by desired

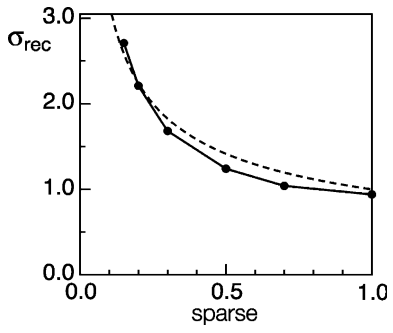

Figure 4. Noise as a function of sparse level or fraction of total measurement time: (thick line, closed circles) average noise in the reconstruction spectra σ_{rec}; (thin dash line) noise extrapolated from the reference spectrum. The noise extrapolation was performed using inverse square root dependence on the measurement time, which is proportional to the sparse level. The values are given in units of standard deviation of noise in the reference spectrum σ_{ref}. Note that the noise estimates are conservative, since only planes containing the diagonal peaks were used for calculating σ_{rec} while the empty denoised planes were ignored (see text).

sensitivity rather than by necessity to sample many data points in order to achieve target resolution. Small differences between the curves in Figure 4 can be easily rationalized on the qualitative level. For the high sparse levels (right side of the plot) noise in the reconstructed spectrum is somewhat lower. Here, the bias in the sampling density ensures that only data points with the lowest signal are missing. Omission of these points hardly reduces the quality of the reconstruction. As the sparse level goes down the gain provided by the matched acquisition levels off while losses introduced by the MDD increase since the number of measured data points approaches the number of adjustable parameters of the MDD model. This causes elevated noise values on the left side of the plot. In other words, slightly lower noise values for sparse levels 30–70% can be attributed to the effect of sensitivity optimized matched acquisition. This effect is much more significant when acquisition is longer than transverse relaxation time. The acquisition times in our spectrum were short (16 ms for 1H and 6 ms for 13C) in comparison with exponential decay times (50 and 10 ms, respectively) used for generation of the sampling matrix G in Equation 1. Therefore the sampling density in the time domain did not differ too much and the sampling effect on S/N was expected to be small.

As it follows from the results of simulations presented here, a reasonable compromise for the sparse level for the 4D NOESY's is in the range 0.25–0.4. This is in line with the value used in our previous practical presentation of the MDD (Tugarinov et al., 2005).

Reconstruction of the 10 regions of the spectrum

For each of the ten regions (see Table 1), a data set with sparse level of 30% was prepared as input for the 3DD calculations. The number of components was determined as described in *Materials and methods* and shown as the second column of the Table 1. The third column in Table 1 shows normalized standard deviation in the difference spectrum σ_{dif}. These values for the first nine regions (−0.5 to 3.6 ppm) fall in a narrow range between 1.08 and 1.21, which implies similar accuracy of the reconstructions for these regions. Relatively high value of 1.63 was obtained for the leftmost region of the spectrum (3.4–4.5 ppm). The elevated value of σ_{dif} in this case is due to destructive influence of not fully suppressed water ridges in this region.

Variation of the MDD parameters

The Tikhonov regularization factor λ

Tikhonov regularization improves the convergence of the MDD minimization (Ibraghimov, 2002) and quality of the reconstruction (Orekhov et al., 2003). When applied to a complete data set, result of the decomposition is not very sensitive to the value of λ. However, for a large fraction of missing data, usage of regularization becomes necessary for improving the outcome of the decomposition. Up to certain λ-value the regularization can be thought as a mild additional constrain, which minimizes energy of the output components calculated throughout whole spectrum including missing points. A reasonable λ-value allows avoiding unjustifiably high amplitudes for the missing data points, which would result in additional noise in the reconstruction. On the other hand a too big value might cause distortions in the solution. Thus, both too low and too high λ-values can result in a large variance in the difference spectrum. This is illustrated by the results obtained for the representative 4D data region for a set of λ-values (1.0, 0.1, 0.01, 0.001, 10^{-5}). Figure 5a shows a plot of the σ_{dif} values as a function of λ. A shallow minimum, which spawns almost two orders of magnitude, is located near 0.01–0.1; it is in line with the optimal values previously reported for different 3D and 4D data sets (Orekhov et al., 2003; Tugarinov et al., 2005). Therefore, it once again confirms that the regularization factor λ does not represent a critical parameter for the MDD calculations and one can safely use the default value, e.g. $\lambda = 0.01$.

The number of components

Figure 5b shows σ_{dif} values for the representative region as a function of the number of components N_c. The plot in Figure 5b shows an almost flat curve with a minimum at 50, which matches the value predicted for this region (Table 1, strip 0.4–1.1 ppm) by the procedure described in the *Materials and methods* section. This shows that overall reconstruction works well for the wide range of N_c values. Nonetheless, N_c values much smaller than the optimum should be avoided. It was shown earlier (Gutmanas et al., 2002) that

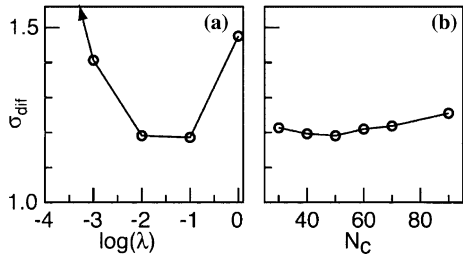

Figure 5. Average variance σ_{dif} in the reference-reconstruction difference spectra as a function of the MDD parameters: λ (a) and N_c (b). The average variance values shown as open circles connected by a line are normalized to σ_{ref}. The arrow on the (a) is directed towards point at $log(\lambda) = -5$, $\sigma_{dif} = 3.72$, which is outside the plot. The vertical scale and limits are the same for the two graphs. The calculations were performed with sparse level of 0.3 for the region 0.4–1.1 ppm. The fixed parameters were: $N_c = 50$ (a), $\lambda = 0.01$ (b).

when choosing values that are clearly too small, the weakest signals can be lost, or signals with high overlap may be merged into a single component. On the other limit, when N_c is too large, the somewhat higher values of σ_{dif} are caused by over-parameterization of the model of expression (1). In both cases the consequences of using improper number of components remain localized and small. Overall, the MDD procedure appears to be tolerant to both overestimation and underestimation of N_c. The calculation with $N_c = 90$, approximately twice overestimate of the optimal number, gives mostly the same result, except for slight noise increase. More specifically, the tolerance of the MDD to overestimation of N_c is related to the degree of the redundancy of the data, which is defined as the ratio of the number of experimental measurements to the number of model parameters fitted in the MDD. The latter number corresponds to the number of elements in all shapes of all selected components. Clearly, to ensure an unambiguous solution in the minimization of Equation 1, the redundancy of the data must be significantly larger than one. In other words the number of adjustable parameters must be smaller than the number of measurements. The consideration sets an upper limit on the number of components to select. For the region of the reference spectrum used here, the number of experimental measurements is the number of points in all dimensions $100 \times 36 \times 36 \times 63 = 8164800$. This number becomes 2449440 for the sparse level of 0.3. For the 3DD and 4DD calculations applied to the same sparse data set, the number of parameters fitted inside the models is significantly different. In 4DD (t_1,t_2,t_3, ω_4) calculation the number of model parameters is $N_c \times (100 + 36 + 36 + 63 - 3) = N_c \times 232$ and the corresponding data redundancy can be calculated as $2449440/(N_c \times 232) = 10558/N_c$. In the 3DD ([$t_1$,$t_2$],$t_3$,$\omega_4$) the number of model parameters $N_c \times (36*100 + 36 + 63 - 2) = N_c \times 3697$ is larger, since the first and second dimensions are merged into one long dimension, as described in the theoretical part. The redundancy of the data is $2449440/(N_c \times 3697) = 662/N_c$. For example, the representative region with $N_c = 50$ the 3DD (4DD) has redundancy of 13.2 (200). Thus, to achieve the same level of redundancy, 4DD allows approximately 16 times larger number of components compared to 3DD. It is for this reason 4DD calculation is more tolerant to overestimation in N_c and can be used for a robust spectral component estimation. Note, however, that the actual local redundancy levels can differ from the average values estimated above for the entire region.

Conclusion

In this paper the MDD methodology was applied to processing of the exponentially sampled high-resolution 4D HCCH-correlation NOESY spectrum of a 106-residue protein produced in the pipeline of structural genomics. The results of the systematic analysis demonstrate the robustness of the MDD procedure and its applicability to reconstructing of non-uniformly sampled spectra with sparse levels from 15 to 100%. Throughout this range the resolution and the sensitivity per unit of measurement time are preserved in the reconstructed spectrum. In the example using 30% data we demonstrate that 4D NOE spectral intensities are correctly reproduced in the large dynamic range, as well as cross-peak intensities. Together with accurate cross-peak line widths and increased resolution delivered by the sparse acquisition these intensities should translate into a set of correct NOE constraints to be used for accurate spatial structure calculation. It was demonstrated that the method does not produce false peaks and the loss of peaks in the reconstructed planes relative to the reference spectrum corresponds to the anticipated increase in the noise level due to shorter measurement time of

the sparse spectrum. While giving prospects for better resolution and a significantly simplified peak assignment, the method does not sacrifice sensitivity. This makes it attractive not only in the field of high throughput structural proteomics but also for all types of molecular systems amenable for NMR structural studies. The benefits are especially appreciated in application to large proteins (Tugarinov et al., 2005). As such the MDD in combination with sparse sampling complements a toolbox of fast techniques that so far were demonstrated on spectra other than NOESY, e.g. triple-resonance experiments for signal assignment as reviewed for example in (Freeman and Kupče, 2003) (Hoch and Stern, 1996, 2001).

Acknowledgements

This work was supported by grants from the Swedish Foundation for Strategic Research (A3 04:160d), the Swedish National Allocation Committee (SNIC 3/04–44), the US. National Institutes of Health (PSI grant to the Northeast Structural Genomics Consortium), the Ontario Research and Development Challenge fund, and Genome Canada.

References

Barna, J.C.J, Laue, E.D., Mayger, M.R., Skilling, J. and Worrall, S.J.P. (1987) *J. Magn. Reson.*, **73**, 69–77.

Bax, A. (1994) *Curr. Opin. Struct. Biol.*, **4**, 738–744.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M. and Arrowsmith, C.H. (2000) *Nat. Struct. Biol.*, **7**, 903–909.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

Ernst, R.R. (1992) *Biosci. Rep.*, **12**, 143–187.

Fernandez, C. and Wider, G. (2003) *Curr. Opin. Struct. Biol.*, **13**, 570–580.

Freeman, R. and Kupče, E. (2003) *J. Biomol. NMR*, **27**, 101–113.

Gerstein, M., Edwards, A., Arrowsmith, C.H. and Montelione, G.T. (2003) *Science*, **299**, 1663–1663.

Goto, N.K. and Kay, L.E. (2000) *Curr. Opin. Struct. Biol.*, **10**, 585–592.

Gutmanas, A., Jarvoll, P., Orekhov, V.Y. and Billeter, M. (2002) *J. Biomol. NMR*, **24**, 191–201.

Hoch, J.C. and Stern, A.S. (1996) *NMR Data Processiong*, Wiley-Liss, New York.

Hoch, J.C. and Stern, A.S. (2001) *Methods Enzymol.*, **338**, 159–178.

Ibrahimov, I. (2002) *Numer. Linear Algebr. Appl.*, **9**, 551–565.

Kennedy, M.A., Montelione, G.T., Arrowsmith, C.H. and Markley, J.L. (2002) *J. Struct. Funct. Genomics*, **2**, 155–169.

Kruskal, J.B. (1977) *Linear Algebra Appl.*, **18**, 95–138.

Laue, E.D., Mayger, M.R., Skilling, J. and Staunton, J. (1986) *J. Magn. Reson.*, **68**, 14–29.

Orekhov, V.Y., Ibraghimov, I. and Billeter, M. (2003) *J. Biomol. NMR*, **27**, 165–173.

Pervushin, K., Riek, R., Wider, G. and Wüthrich, K. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 12366–12371.

Peti, W., Etezady-Esfajani, T., Herrmann, T., Klock, H.E., Lesley, S.A. and Wüthrich, K. (2004) *J. Struct. Funct. Genomics*, **5**, 205–215.

Rovnyak, D., Frueh, D.P., Sastry, M., Sun, Z.Y.J., Stern, A.S., Hoch, J.C. and Wagner, G. (2004) *J. Magn. Reson.*, **170**, 15–21.

Schmieder, P., Stern, A.S., Wagner, G. and Hoch, J.C. (1994) *J. Biomol. NMR*, **4**, 483–490.

Szyperski, T., Yeh, D.C., Sukumaran, D.K., Moseley, H.N.B. and Montelione, G.T. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 8009–8014.

Tikhonov, A.N. and Samarskij, A.A. (1990) *Equations of Mathematical Physics*, Dover, New York.

Tugarinov, V., Hwang, P.M. and Kay, L.E. (2004a) *Annu. Rev. Biochem.*, **73**, 107–146.

Tugarinov, V., Kay, L.E., Ibraghimov, I.V and Orekhov, V.Y. (2005) *J. Am. Chem. Soc.*, **127**, 2767–2775.

Tugarinov, V., Sprangers, R. and Kay, L.E. (2004b) *J. Am. Chem. Soc.*, **126**, 4921–4925.

Vuister, G.W., Clore, G.M., Gronenborn, A.M., Powers, R., Garrett, D.S., Tschudin, R. and Bax, A. (1993) *J. Magn. Reson. B*, **101**, 210–213.

Wüthrich, K. (2003) *J. Biomol. NMR*, **27**, 13–39.

Yamazaki, T., Lee, W., Revington, M., Mattiello, D.L., Dahlquist, F.W., Arrowsmith, C.H. and Kay, L.E. (1994) *J. Am. Chem. Soc.*, **116**, 6464–6465.

Yee, A., Pardee, K., Christendat, D., Savchenko, A., Edwards, A.M. and Arrowsmith, C.H. (2003) *Acc. Chem. Res.*, **36**, 183–189.