Journal of Biomolecular NMR **27:** 165–173, 2003.

Optimizing resolution in multidimensional NMR by three-way decomposition

Vladislav Yu. Orekhov[a,*], Ilghiz Ibraghimov[b] & Martin Billeter[c]
[a]*Swedish NMR Centre at Göteborg University, Box 465, 40530 Göteborg, Sweden;* [b]*Saarbrücken University, Mathematical Department, 66041 Saarbrücken, Germany,* [c]*Biochemistry and Biophysics, Göteborg University, Box 462, 40530 Göteborg, Sweden*

Received 11 March 2003; Accepted 15 May 2003

Key words: azurin, maximum entropy, MUNIN, NOESY, non-uniform sampling, PARAFAC

Abstract

Resolution depends on the number of points sampled in a FID; in indirectly detected dimensions it is an important determinant of the total experiment time. Based on the high redundancy present in NMR data, we propose the following timesaving scheme for three-dimensional spectra. An extensive grid of discrete $t1$- and $t2$-values is used, which increases resolution while preserving the spectral width. Total experiment time is reduced by avoiding the recording of $t3$-FIDs for selected pairs of $t1$ and $t2$; typically the recording is omitted for about 75% of the ($t1$,$t2$) combinations. These data sets are referred to as *sparse*, and post-experimental processing making optimal use of spectral redundancy provides the missing, non-recorded data. We have previously shown that three-way decomposition (TWD) within the MUNIN approach provides a practical way to process *dense* NMR data sets. Here, a novel TWD algorithm [Ibraghimov, (2002) *Numer. Linear Algebra Appl.* **9**, 551–565] is used to complement a *sparsely* recorded time-domain data set by providing the missing FIDs for all ($t1$,$t2$) combinations omitted in the experiment. A necessary condition is that for each $t1$-value at least a few FIDs are recorded, and similar for each $t2$-value. The method is demonstrated on non-uniformly sampled ^{15}N-NOESY-HSQC data sets recorded for the 14 kD protein azurin. The spectra obtained by TWD, reconstruction and ordinary transform to frequency-domain are, in spite of the large number of signals and the high dynamic range typical for NOESYs, highly similar to a corresponding reference spectrum, for which all ($t1$,$t2$) combinations were recorded.

Abbreviations: DFT – discrete Fourier transform; ME – maximum entropy; MUNIN – multidimensional NMR spectra interpretation; TWD – three-way decomposition; 1D – one-dimensional; 3D – three-dimensional.

Introduction

The application of NMR to challenging problems in structural biology often requires spectroscopy at the very limits of resolution and sensitivity. In multidimensional NMR, both characteristics come at significant cost of experimental time, since every new value in the indirectly detected dimensions has to await prior equilibration of the spin system. The traditional method of spectra processing, discrete Fourier transform (DFT), requires data sampling with regular time intervals and therefore dictates a straightforward relation between spectral width, resolution and experiment time. Variations of this simple processing scheme have been introduced to improve on resolution or sensitivity, examples being linear prediction or maximum entropy reconstruction (reviewed by Koehl, 1999; Hoch and Stern, 2001; Stern et al., 2002).

In the present context, we consider a three-dimensional (3D) NMR time-domain data set to be defined on a 3D grid with the evolution times $t1$, $t2$ and $t3$ forming the axes. For normal processing by DFT, measured data exists for all grid points, and the latter are regularly spaced. This spacing interval is

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

dictated by the sampling bandwidth (spectral width), and the number of grid points defines evolution time and delimits the resolution. The limitation imposed on resolution by available experiment time may be overcome in various ways. Experimental data may be extended towards larger evolution times by linear prediction prior to DFT. The assumption is that all information needed for this extension is already present in the data for short evolution times, and a fit of the existing data with analytical expressions will provide reasonable values for longer evolution times (Koehl, 1999). A different idea is to drop the requirement for regularly spaced data (uniform sampling), allowing to cover during the same experiment time a wider range of evolution times while preserving in principle a sufficiently short spacing (Barna et al., 1987; Schmieder et al., 1993). A bonus of this non-uniform sampling is that a sampling biased towards data points with higher signal-to-noise should yield an improvement in sensitivity (Schmieder et al., 1993; Stern et al., 2002). For many NMR spectra this could simply consist of a denser sampling of shorter evolution times. In the context of our regular grid in $t1$, $t2$ and $t3$, a non-uniformly sampled data set would contain holes, i.e. grid points for which no data was collected. DFT can no longer be applied to such data, and more sophisticated methods such as maximum entropy reconstruction need to be applied (Hoch and Stern, 2001). For the latter, the ambiguity of incomplete experimental data is removed by defining an entropy measure that subsequently is used to identify an optimal solution. This entropy measure is defined in frequency space and its function is to suppress unnecessary spectral features deviating from a flat baseline.

We propose a different approach that combines elements of the above ideas with a recently introduced processing scheme. Time-domain spectra with holes, which we call *sparse* data sets, form the input, allowing both for non-uniform, optimized sampling and an extension of the grid to long evolution times without changing the fundamental grid spacing or increasing measurement time. Rather than directly transforming these data sets using an additional principle such as entropy, the missing data points are predicted in time domain, and a full data set defined on a large grid with narrow, regular spacing is reconstructed. This prediction is not an extension from only short evolution times, but it uses all available experimental data. The resulting large and dense data set may then be processed as if all of it had been recorded, i.e. linear prediction to further extend evolution times may be performed if desired, and normal DFT may be applied.

Three-way decomposition (TWD) as a general analysis tool has been introduced and discussed about 30 years ago (Caroll and Chang, 1970; Harshman, 1970). It relies on two principles: signals in a multidimensional data set can be described by direct products of 1D vectors, and the resulting decomposition is unique for data sets with at least three dimensions (Kruskal, 1977). It thus differs from other types of decompositions, which were for example presented for 2D NMR spectra (Havel et al., 1994). Recently, we showed with the MUNIN approach that TWD is a valid method to process experimental NMR spectra and to faithfully reproduce signal positions and intensities in for example a ^{15}N-NOESY-HSQC with its large dynamic range and its high density of signals (Orekhov et al., 2001). Subsequently, the correctness and completeness of the structural data extracted from this NOESY with the help of TWD was demonstrated (Gutmanas et al., 2002). Other applications of MUNIN to NMR data sets included series of ^{15}N-HSQC spectra recorded either for the determination of relaxation parameters (Korzhnev et al., 2001) or for the efficient identification of the binding of small molecules from a large library to a target protein (‘drug discovery’; Damberg et al., 2002). However, it is important to note that the present approach differs fundamentally from our earlier applications of TWD to non-uniformly sampled data (Orekhov et al., 2001). In the previous work, all data points for a given *t1*-value were missing, i.e. in the grid of the input spectrum an entire plane was removed. Consequently, after dropping this plane, a *dense* data set was again obtained, which could then be processed in the same way as a full, uniformly sampled data set. However, reconstruction of the missing plane was impossible, and in all output obtained any data for this $t1$-value was also missing. In contrast, the present approach allows for reconstruction of missing data, but requires that no complete plane with experimental data be removed.

Before demonstrating this method on examples where 75% and more of the data points are missing in the input and thus reconstructed from the remaining 25% of experimentally obtained data, a few remarks are in place. Reconstruction is only possible because NMR data contain a significant amount of redundancy. 3D spectra often consist of several million data points; they contain however only a few thousand signals (peaks), which can be characterized by a few dozen numbers each, even if line shapes are to be described.

Another issue concerns the design of sparse data sets for efficient saving of experimental time. Evolution in *t3* comes at no cost since equilibrium has to be awaited anyway before the next execution of the pulse sequence. Therefore, when recording a sparse data set one collects for any given ($t1$,$t2$) combination either the entire FID in $t3$, or the pulse sequence is not run at all for this ($t1$,$t2$) combination. A final remark illustrates the optimal use by TWD of experimental data when reconstructing missing points. For ease of description, an example of a frequency domain spectrum is considered (which would indeed be processed by TWD in the way described here). Assume that the data point corresponding to the maximum of a cross peak is missing. Averaging the intensity of the immediate neighbors would use only limited data, and it would significantly reduce the maximum and thus falsify the peak intensity. However, using the line shapes of all other peaks that have one frequency in common allows TWD to perform in this situation an almost perfect reconstruction. This example also shows the need for at least a few measured data points for each value of $t1$, respectively $t2$; thus no complete ($t1$,$t3$) or ($t2$,$t3$) plane may be missing.

In the following, we shortly define the basic model used by TWD to describe data sets, and then introduce the novel features to the algorithm that allow treatment of sparse data sets. Results are shown using real data, namely from a ^{15}N-NOESY-HSQC, and the influence of various parameters such as the ratio of experimental vs. predicted data is illustrated. The discussion includes comparison to other processing tools, in particular to maximum entropy reconstruction (Hoch and Stern, 2001).

Methods

The fundamental idea of TWD applied to NMR data is that a 3D input data set **S** (in our case an experimental time-domain spectrum) is optimally approximated by a given number of 3D components, which in turn are described by the tensor product of 1D shapes **F1**, **F2** and **F3** along the three dimensions (Orekhov et al., 2001). Note that both the input **S** as well as the output **F1**, **F2** and **F3** are entities that consist of discrete points, and thus form a 3D matrix and 1D vectors, respectively (usually, all entities are defined on a regular grid of $t1$-, $t2$- and $t3$-values). Finding the optimal fit consists of minimizing the following penalty function:

$$\min \sum_{ijk} \left| S_{ijk} - \sum_{m=1}^{M} a^m \cdot F1_i^m \cdot F2_j^m \cdot F3_k^m \right|^2 + \lambda \sum_{m=1}^{M} (a^m)^2 \quad (1)$$

by optimizing the components, i.e. varying the values a^m, $F1_i^m$, $F2_j^m$ and $F3_k^m$. Here, the first sum over m enumerates the M components, a^m is the amplitude of a component, $F1_i^m$, $F2_j^m$ and $F3_k^m$ are numbers representing elements of the normalized shapes for a component, and the indices *i, j* and *k* identify grid points along the three dimensions. The purpose of the second sum over m is to suppress large differences in the size of components (Ibraghimov, 2002). It is controlled by the size of Tikhonov's regularization factor λ (Tikhonov and Samarskij, 1990). For an input matrix S of size I^*J^*K, the number of parameters for optimization is $M^*(I + J + K - 2)$; thus, for a typical size of S the number of input measurements far exceeds the number of fit parameters. So far, this corresponds to the earlier presented MUNIN approach of applying TWD to NMR data (Orekhov et al., 2001).

The central new feature, requiring a new implementation of the procedure, is a modified penalty function required to analyze *sparse* data sets (Ibraghimov, 2002):

$$\min \sum_{ijk} G_{ijk} \cdot \left| S_{ijk} - \sum_{m=1}^{M} a^m \cdot F1_i^m \cdot F2_j^m \cdot F3_k^m \right|^2 + \lambda \sum_{m=1}^{M} (a^m)^2, \quad (2)$$

where the elements of the matrix G are $G_{ijk} = 1$ for a recorded and $G_{ijk} = 0$ for an omitted spectral data point S_{ijk} in the NMR experiment. Thus, not recorded data will not contribute to the penalty function. The key feature about the second expression is that while the input matrix S is *sparse*, the output shapes $\mathbf{F1}^m$, $\mathbf{F2}^m$ and $\mathbf{F3}^m$ are complete, i.e. vectors with *all* elements. This can only be achieved if S never misses an entire plane; since FIDs along $t3$ are either fully sampled or completely omitted, this means that no row or column in the ($t1$,$t2$) plane is empty. The importance of this new variant of decomposition lies in the following: By multiplying the output shapes $\mathbf{F1}^m$, $\mathbf{F2}^m$ and $\mathbf{F3}^m$ and the amplitudes a^m as in the right part of the above expressions, one can reconstruct a full spectrum S^*. The latter represents an

optimal approximation of the spectrum S for the elements that have been experimentally obtained. The full spectrum S^* can be transformed and analyzed in the usual way (in our case of a time-domain data set by linear prediction and Fourier transform). Details regarding the implementation for NMR purposes of the least-squares minimization, the processing of sparse data and the Tikhonov regularization can be found in Orkehov et al. (2001) and Ibraghimov (2002).

All examples in the Results section are based on a 3D ^{15}N-NOESY-HSQC recorded for the 128 residue long protein azurin (Karlsson et al., 1989); this spectrum has been described earlier (Orekhov et al., 2001). Rather than recording FIDs for only selected pairs of evolution times ($t1$,$t2$), a full spectrum was collected that serves as reference to assess the quality of the reconstructions. For the illustrations, the FIDs in the HN dimension were zero-filled doubling their size and Fourier transformed, and the region 8.67–8.90 ppm was extracted. Subsets of FIDs were formed as input to TWD with various amounts of ($t1$,$t2$) combinations. The degree to which such a subset is sparse is defined by the ratio between the number of ($t1$,$t2$) combinations used and those in the full reference data set; this ratio is subsequently denoted by R. The selections were based on random numbers, but exponentially biased towards shorter values of the evolution times using an exponential decay time of *20 ms* in both dimensions. For every subset, the last few percent of ($t1$,$t2$) combinations were used to ensure that all rows along $t1$ or $t2$ contain a few entries. The input for TWD thus consists of a sparse matrix with various extent of missing elements. The size of the corresponding full reference data set was 160 and 44 complex time-domain points along the H_{NOE} and ^{15}N dimensions, respectively, and real 24 frequency domain points in the HN dimension. Other parameters of the TWD runs that were varied systematically are the number of components M, the seed for the random selection of ($t1$,$t2$) combinations, and Tikhonov's regularization factor λ. The reconstructed spectra as well as the full reference spectrum were then processed as follows: DFT was directly applied to the H_{NOE} dimension. In the ^{15}N dimension the time domain signal was extended by 50% to 66 points by linear prediction using twelve coefficients prior to DFT. For all dimensions, the size of the time domain data was doubled by adding zeros prior to DFT, and the signals were multiplied by a square sine weighting function prior to DFT. For another comparison, an additional subset with 25% $t3$-FIDs of the full spectrum was constructed for processing without TWD. This requires uniform sampling, and the most reasonable way (if this is at all possible) is to truncate the data along ^{15}N to 11 points. This latter spectrum was then also extended in the ^{15}N dimension to 66 points by linear prediction using five coefficients and transformed like all other spectra. Execution time for a decomposition on a single 2.4 GHz processors (running Linux) was 85 minutes on a single CPU, and it was shown that the algorithm can be efficiently parallelized.

Results

For a systematic analysis of TWD, a region from a 3D ^{15}N-NOESY-HSQC as described in Methods was used. The flow chart of Fig. 1 summarizes the procedure, where the bold part describes a normal application of TWD to a sparse data set (left side of Figure 1). The other parts of this figure concern the use of a reference data set, which was experimentally recorded to 100%. A first purpose of this reference data set is to derive different sparse data sets by eliminating FIDs along the third dimension; the use of a complete reference avoids the recording of every new sparse data set when varying selected parameters such as the ratio R of missing and total data. Secondly, it provides an absolute reference for the evaluation of the reconstructed data sets obtained by TWD. Note that for simplicity, the reference data set (upper right corner of Figure 1) was defined after application of Fourier transform in the third dimension.

For a first application, a data set with 25% of the FIDs along $\omega3$, i.e., $R = 0.25$, was formed by random selection of ($t1$,$t2$) combinations. Instead of purely random sampling we used an exponentially biased sampling with a denser distribution of data points at the beginning of the decaying signals (Figure 2). This scheme was proven to be superior both with respect to resolution and sensitivity in comparison with completely random sampling (Stern et al., 2002). Other parameters for this decomposition are a Tikhonov regularization factor $\lambda = 0.005$ and a maximal number of components $M = 30$ (see expressions 1 and 2). The latter choice was based on knowledge from a ^{15}N-HSQC of the number of HN-H groups in the spectral region considered. Figure 3 displays planes with $\omega_{HN} = 8.82$ ppm of the transformed 3D spectra for comparison of the output of TWD after reconstruction and the conventionally processed full reference data set. The sparse data set with $R = 0.25$ (Figure 3A)

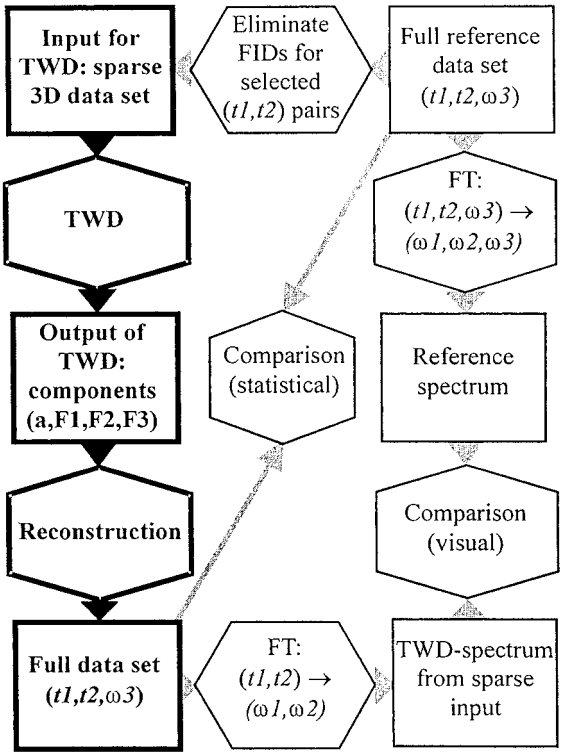

Figure 1. Flow-chart summarizing the application of TWD to various data sets derived from a ^{15}N-NOESY-HSQC. The left side (bold lines) describes the normal use where only a sparse data set is recorded. The other parts (thin lines) indicate the double use of a full reference data set (upper right corner). It allows derivations of various sparse data sets for systematic tests of TWD (top of figure), and it serves as a reference for comparison of the results from TWD, both in time-domain (center of figure) and in frequency domain (lower right corner). Note that for simplicity the initial reference data set is already transformed in the third dimension; in this dimension all data points are available (see text). Six-cornered boxes indicate processing steps and rectangular boxes describe input or output data.

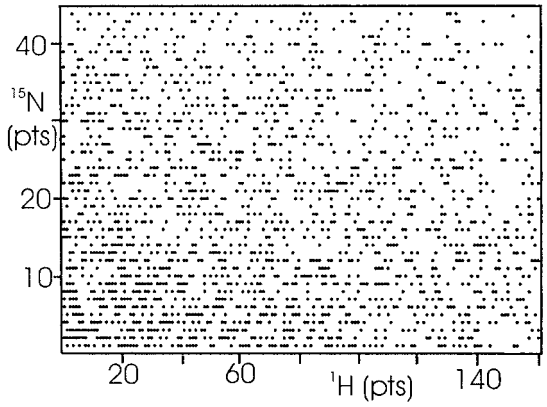

Figure 2. Illustration of a sparse data set for input to TWD (see Figure 1, upper left corner). Shown is the distribution of ($t1$,$t2$) combinations used in the TWD application yielding the result of Figure 3A. The dots indicate the 25% of data chosen as input for this decomposition using an exponentially biased random sampling with a preference for short $t1$- and $t2$-values. Whenever a ($t1$,$t2$) combination is selected for measurement, a complete FID along $t3$ is recorded, allowing immediate Fourier transform in this direction.

provides an excellent reproduction of both strong and weak signals in the reference spectrum (Figure 3B). A difference spectrum (not shown) would contain only few features at the level of the two lowest contours in Figure 3. The sparse experimental input does not introduce any significant artifacts in the resulting spectrum. Furthermore, processing with TWD does not suffer from the high dynamic range and the large number of signals found in NOESY-type spectra. Peaks at ω_N = 117.4 ppm result from folding and are thus negative; this has no consequence on their correct reproduction in Figure 3A. Figure 3C shows another spectrum processed with TWD, but this time only 18.75% of the data from the reference spectrum were selected by exponentially biased random sampling. Although some differences to the reference spectrum start to show up, most features of the latter spectrum are faithfully reproduced. For comparison, another subset with 25% of the FIDs from the reference spectrum was chosen, using this time uniformly sampled data. The acquisition time in the nitrogen dimension had to be reduced four times, and the signal size had to be extended six times in this dimension by linear prediction to achieve the same final data size prior to DFT (see Methods). The resulting spectrum, shown in Fig. 3D, clearly contrasts in quality when compared to the other spectra of this figure.

Expanding on the description of the effect of different degrees of sparse data sets in Figures 3A–C, a systematic variation of various parameters was performed. The parameters investigated for their influence on the result include the ratio R between experimentally observed and full data (as in Figures 3A and 3C) and the number of components M. Two measures were used to determine the success of each run: the residual of the decomposition, i.e. the root mean square difference between the reconstructed and the references data sets, and the corresponding kurtosis (Press et al., 1992). For these statistical analyses all data points, i.e., the full reference data set, were used; thus all values for residual and kurtosis are based on the same number of points. The residual should be minimal, but mainly due to spectral noise it will not reach zero. A small kurtosis indicates that the differences between the input and the output to TWD

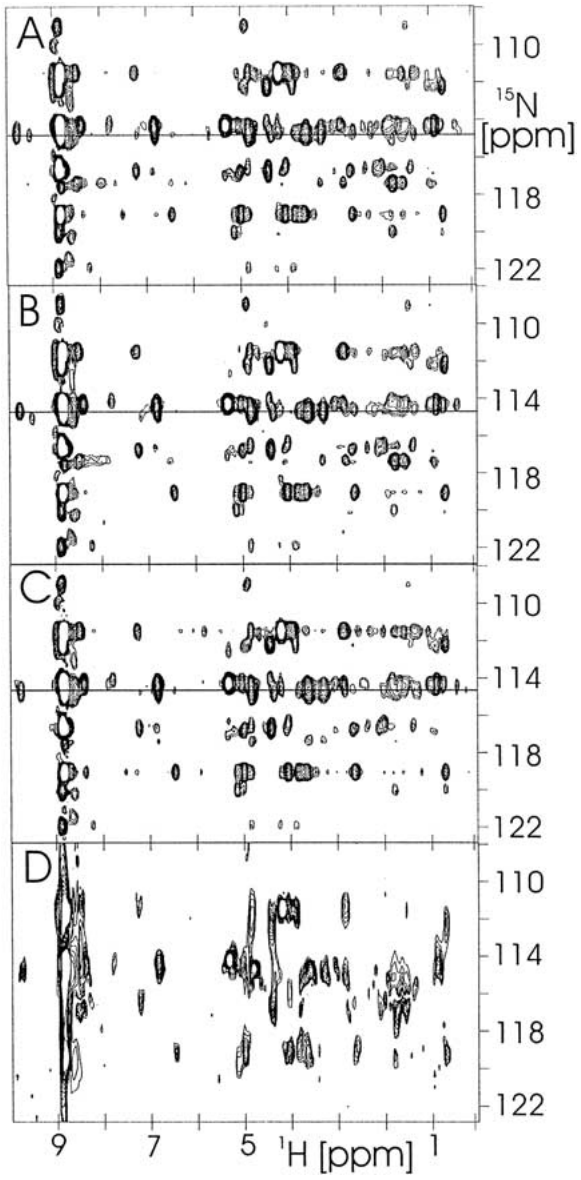

Figure 3. Selected region of the 3D ^{15}N-NOESY-HSQC spectrum recorded for the 14 kD protein azurin. Planes along the frequencies of H_{NOE} and ^{15}N with $\omega_{HN} = 8.82$ ppm are shown. (A) Result from a TWD-calculation with $R = 0.25$, $M = 30$ and $\lambda = 0.005$. (B) Corresponding region from the full reference spectrum. (C) Result from a TWD-calculation with $R = 0.1875$, $M = 30$ and $\lambda = 0.005$. The reconstructed time-domain data used for (A) and (C) were transformed in the same manner as the full reference time-domain data yielding (B). (D) Result from a data set truncated to 11 complex points in the ^{15}N dimension followed by conventional linear prediction and DFT (see text). The horizontal lines at $\omega_N = 114.5$ ppm in panels A–C indicate the location of the cross sections shown in Figure 5.

corresponds to Gaussian noise. Figure 4 displays the residual and kurtosis as a function of the fraction R of sparse data for different numbers of components M. The knee of all curves is near $R = 0.20$. Therefore, decompositions using 25% of the total data as in Figure 3A, or also 18.75% as in Figure 3C, come at little cost in terms of spectral differences compared to using a full data set. The number of components plays a minor role as long as it is not chosen too small to describe the signals present as for $M \leq 20$. Another parameter whose influence on the outcome of a TWD application was checked is Tikhonov's regularization factor λ. On all previously discussed runs it was set to 0.005. When varying it from 0.0005 to 0.05 for the decompositions with $R = 0.25$ and $M = 30$, the maximally observed increase of the residual and the kurtosis was 0.03 and 0.23, respectively. These two values are smaller than the corresponding variations when changing the number of components M in the range 25–50 (see Figure 4). Therefore, the regularization factor λ does not represent a critical parameter. Also, several decompositions were performed with the same parameters as for the result of Figure 3A, but choosing different seeds for the random selection of $(t1,t2)$ combinations; no significant variations of the results were observed.

A more comprehensive view of the influence of parameters is afforded by Figure 5, where cross-sections through the planes shown in Figure 3 are assembled (see thin lines in Figure 3). Starting again from the parameter choice of the run of Figure 3A, i.e., $R = 0.25$, $M = 30$ and $\lambda = 0.005$, two sets of runs are compared to the full reference spectrum. In Figure 5A, the ratio R is varied from 0.0625 to 0.75; in Figure 5B, the number of components M adopts values between 15 and 50. Lowering the percentage R of FIDs used yields an increase in noise, which starts to hide relevant peaks with $R < 18\%$ (Figure 5A). This gradual growing of noise without the arising of sizeable signal artifacts demonstrates the robustness of the approach: even at a fraction of 12.5% the peaks that are still detectable will not be confused with artifacts. A reduction in experimental data may be considered as an increase of the noise level. Figure 5A indicates that the algorithm responds in a very stable way to added noise. The influence of the choice of the number of components M is hardly visible in Figure 5B. However, too small values of this parameter may yield localized distortions when different components are forced to merge (Gutmanas et al., 2002).

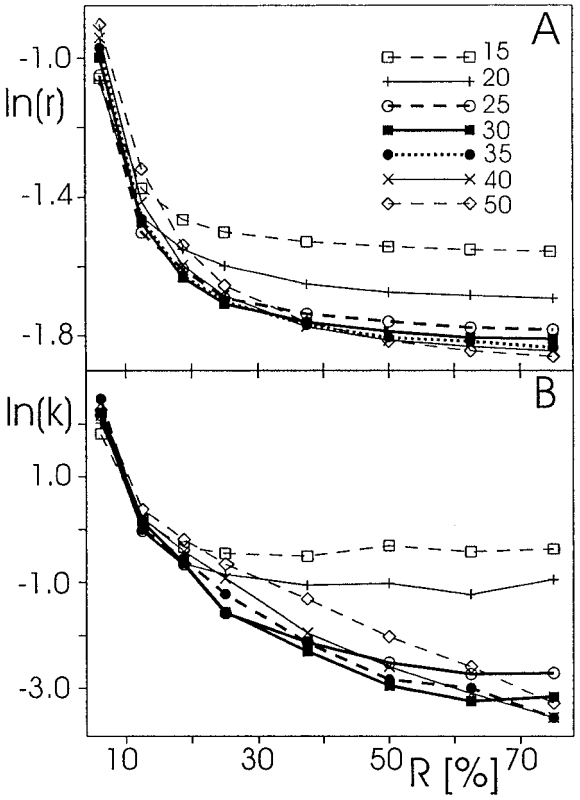

Figure 4. Plots of (A) residual and (B) kurtosis for different fractions of sparse data R. Seven curves are shown for different number of components $M = 15$, 20, 25, 30, 35, 40 and 50 (line types and symbols for the different choices of M are indicated in panel A). In all runs λ was set to 0.005. For residual and kurtosis values, natural logarithms were calculated and indicated on the vertical axes as $\ln(r)$ and $\ln(k)$, respectively.

Discussion

TWD has been shown earlier to present a very general tool for the processing of multidimensional NMR data (Orekhov et al., 2001). One may also add that an inherent suitability of TWD for processing of multidimensional NMR data is indicated by the fact that the TWD model of expression 1 can be directly derived from a general formulation of NMR pulse sequences. Besides the demonstration of its applicability to sparse, non-uniformly sampled data sets, the main conclusions of the present study regards the robustness and reliability of TWD. A first issue to mention is the small number of parameters that are required. The only sensitive and application dependent parameters are the percentage R of recorded data and the number of components M. Reducing the amount of sampled data obviously leads at some point to failure of a proper reconstruction.

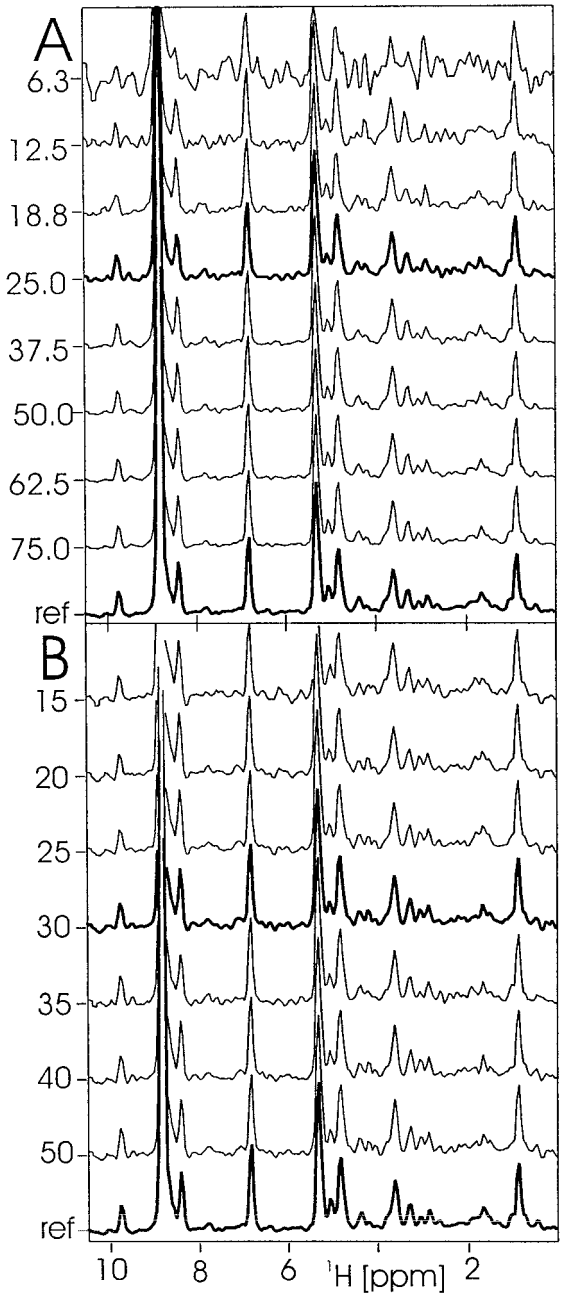

Figure 5. Cross sections for various decompositions through corresponding planes of the ^{15}N-HSQC-NOESY (see horizontal lines at $\omega_N = 114.5$ ppm in Figure 3). The thick lines are cross sections from the application of Figure 3A and from the reference spectrum of Figure 3B. The thin lines stem from TWD calculations with systematic parameter variation. For visibility, lines from different parameter choices are shown with a constant offset. In (A), R is varied from 0.0625 to 0.75; in (B), M is varied from 15 to 50. For all runs λ was set to 0.005, and 3000 iterations were performed. The vertical direction provides intensities; however, due to the offsets used, the labeling along this axis indicates the value of the parameter that is varied, i.e. of R in panel A and of M in panel B.

Figure 5 shows, however, that this failure occurs by a gradual increase of the noise rather than by a sudden appearance of artifacts, making TWD a reliable method. With respect to the choice of number of components M, the similarity of the curves in Figure 4 indicates only marginal sensitivity, once M is sufficiently large. The number of necessary components M can usually be well estimated; in the application presented here a simple counting of peaks in a corresponding ^{15}N-HSQC is sufficient. The consequences when choosing values that are clearly too small such as $M = 15$ or 20 remain localized. It was shown earlier that in this case components with high overlap may be merged into a single component (Gutmanas et al., 2002). The somewhat different behavior for $M = 50$ (Figure 4B) seems to be caused by overparameterization of the model of expression 2. The choice of the regularization factor λ is not critical, but some regularization is advisable to avoid pairs of large components that nearly cancel each other (Damberg et al., 2002; Ibraghimov, 2002). This parameter appears largely independent of the type of application as its use in another context showed (Damberg et al., 2002); thus our present choice of $\lambda = 0.005$ can be suggested quite generally for most applications. The choices of more 'technical' parameters, namely the seed for the sampling of FIDs used and the number of iterations in the optimization, resulted in no unexpected behavior.

The penalty function of expression 2 is, similar to the previously used expression 1, non-convex, and thus convergence can formally not be guaranteed. Our experience, which is based on probably over a thousand runs of MUNIN, indicates reliable convergence of the optimization procedure chosen, even when applying it to very different types of data sets (this study, Gutmanas et al., 2002; Damberg et al., 2002). From the hundreds of runs on sparse data sets performed in this study, only one showed a somewhat outstanding behavior. With a seed different than the one used for Figure 4, the decomposition for the parameter set $R = 0.625$ and $M = 25$ exhibited a kurtosis of similar size as when using $R = 0.25$ and $M = 30$, i.e. the parameter choice of Figure 3A. Speed of convergence represents a separate issue. It is known that the type of optimization used (see appendix of Orekhov et al., 2001) is intrinsically slow. Other, faster procedures may, however, cost a price in terms of the quality of convergence. While we always used 3000 iterations, the second half of the run was mainly used to make sure that no sudden drop of the penalty function occurred after it had leveled of. In fact, we did not observe any significant difference for corresponding runs stopped after 1000 iterations. The procedure lends itself very well to parallelization, where little overhead is observed for runs using ten CPUs (unpublished data).

As in any other least square fit, the number of model parameters must be smaller than the number of experimental measurements. This imposes a natural lower limit to the amount of data points in a sparse spectrum. TWD is a true three-dimensional analysis, and therefore the number of experimental measurements coincides with the spectral size; for the region of the reference spectrum used here (Figure 3B), this is $12 \times 88 \times 320 = 337920$. The number of model parameters in expression (2) corresponds to the number of elements in all shapes of all components: $30 \times (12 + 88 + 320 - 2) = 12540$. The ratio between the two numbers yields about a 27-fold redundancy of the data. Thus, one must have at least 3.7% of the data points to reconstruct the complete spectrum. In our calculations we found that the reconstructed spectrum still looks reasonable when using about 18% of the full data set. Such a sparse data set theoretically still provides a fivefold redundancy. The region presented here is one of the most crowded in the spectrum of azurin. It contains 25 HN-N groups and thus the choice of $M = 30$ components is appropriate (Gutmanas et al., 2002). Redundancy of the data and, consequently, the quality of the reconstruction would increase for less crowded regions, where a smaller number of components is needed, and vice versa.

Comparison with other methods is restricted here to tools with a demonstrated capability to process sparse, non-uniformly sampled spectral data from multidimensional NMR applied to proteins. The most prominent and best-documented method in this context is maximum entropy (ME) reconstruction. Similar to TWD, it is proposed as a way to improve on resolution and sensitivity by using non-uniformly sampled data (Stern et al., 2002). ME reconstruction maximizes an entropy measure defined in frequency space; this aspect is sometimes described as choosing a spectrum with least information (Sibisi, 1983). In practical applications, noise levels need to be estimated *a priori* in order to define run-time parameters (Hoch and Stern, 2001). Noise below the chosen level is then efficiently suppressed, but the choice has direct implications on the final sensitivity achieved. In TWD, the only real parameter for processing of a given sparse data set is the number of components M. Once this parameter

is chosen large enough it has hardly any effect (Figures 4 and 5); the additional components will either describe signal irregularities or strong noise features. Sensitivity is thus little affected by any *a priori* defined run-time parameter. Ideally, all noise will end up in the residual, because the minimization of expression 1 or 2 will add as much intensity as possible to the limited number of components. These will include preferentially intensity from signals, since they are often stronger, and several can be collected into one component. It has indeed been demonstrated that the noise in the resulting shapes from TWD processing is smaller than the noise in a conventionally processed reference spectrum (Gutmanas et al., 2002).

When comparing the suitability of TWD and ME reconstruction for multidimensional NMR spectra, complementing preferences regarding the input are revealed. ME reconstruction is most reliable when the spectra contain a relatively small number of peaks, and when signal quantification is not very critical (Hoch and Stern, 2001). On the other hand, TWD in the current implementation is little affected by a high dynamic range, and component definition becomes more reliably with increasing number of peaks included in them. Therefore, NOESY-type spectra are very well suited for TWD, whereas ME reconstruction aims rather at triple-resonance spectra. While an extensive literature describes the power of ME reconstruction over conventional methods for 1D and 2D NMR data, only a few applications to 3D spectra are published (Schmieder et al., 1993; Hodgkinson et al., 1993), none of which includes a NOESY-type spectrum. A direct comparison between the two methods can therefore not be accomplished at this time. With the new formulation of the TWD model in expression 2, and with the corresponding new implementation, reconstruction of full time-domain data from a sparse input data set becomes possible. This novel feature provides new opportunities for unstable or low-concentration samples through sampling schemes optimized for resolution, sensitivity and/or experiment time. In addition, savings of 75% or more of instrument time open new avenues to high-throughput NMR studies in structural genomics.

Acknowledgements

This work was supported by grants 621-2001-3095 and 621-2001-3014 from the Swedish Research Council. The NMR experiment was performed at the Swedish NMR Center on ^{15}N labeled azurin sample kindly provided by Dr G. Karlsson. The authors would like to thank Prof L. Kay for fruitful discussions.

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.

Caroll, J. D. and Chang, J. (1970) *Psychometrica,* **35,** 283–319.

Damberg, C. S., Orekhov, V. Y. and Billeter, M. (2002) *J. Med. Chem.*, **45**, 5649–5654.

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

Harshman, R.A. (1970) *UCLA Working Papers in Phonetics,* **16,** 1–84.

Havel, T. F., Najfeld, I. and Yang, J.X. (1994), *Proc. Natl. Acad. Sci. USA*, **91**, 7962–7966.

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

Hodgkinson, P., Mott, H.R., Driscoll, P.C., Jones, J.A. and Hore, P.J. (1993) *J. Magn. Reson.*, **B101**, 218–222.

Ibraghimov, I. (2002) *Numer. Linear Algebra Appl.* **9**, 551–565.

Karlsson, B.G., Pascher, T., Nordling, M., Arvidsson, R.H. and Lundberg, L.G. (1989) *FEBS Lett.,* **246,** 211–217.

Koehl, P. (1999) *Prog. NMR Spectrosc.*, **34**, 257–299.

Korzhnev, D.M., Ibraghimov, I.V., Billeter, M. and Orekhov, V.Y. (2001) *J. Biomol. NMR,* **21,** 263–268.

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

Orekhov, V.Y., Ibraghimov, I.V. and Billeter, M. (2001) *J. Biomol. NMR,* **20,** 49–60.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical Recipes*, Cambridge University Press, Cambridge.

Schmieder, P, Stern, A.S., Wagner, G. and Hoch, J.C. (1993) *J. Biomol. NMR*, **3**, 569–576.

Sibisi, S. (1983) *Nature*, **301**, 134–136.

Stern, A.S., Li, K.-B. and Hoch, J.C. (2002) *J. Am. Chem. Soc*, **124**, 1982–1993.

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