ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Magnetic Resonance 161 (2003) 174–182

JMR Journal of Magnetic Resonance

www.elsevier.com/locate/jmr

Restricted linear least square treatment processing of heteronuclear spectra of biomolecules using the ANAFOR strategy

Guy Lippens,[a,*] Philippe R. Bodart,[b] Françis Taulelle,[c] and Jean-Paul Amoureux[d]

[a] *Université de Lille2, CNRS-UMR 8525, Institut Pasteur de Lille, 1 rue du Professeur Calmette BP447, 59021 Lille Cedex, France*
[b] *Université de Lille1, Laboratoire de Catalyse de Lille, CNRS-UMR 8010, 59655 Villeneuve d'Ascq, France*
[c] *Université Louis Pasteur, CNRS-FRE 2423, RMN & Chimie du Solide, 4 Rue Blaise Pascal, 67070 Strasbourg, France*
[d] *Université de Lille1, LCPS CNRS-UMR 8012, 59655 Villeneuve d'Ascq, France*

Received 15 July 2002; revised 24 December 2002

Abstract

We explore the use of a processing procedure based on restricted least square minimization as a tool for reducing the time versus resolution dilemma often encountered for biomolecular multidimensional spectra. Using a 2D spectrum as a reference, we obtain the necessary input of frequency components and linewidths. Combined even with a limited time evolution in the indirect dimension, the amplitudes of the correlation peaks in all planes of the 3D spectra can be extracted, and can be used to reconstruct the interferograms in the third dimension. Parameters such as number of lines, threshold choice, resolution, lineshape, number of experimental data points and finally signal to noise ratio of the spectrum are examined starting from a triple-resonance HNCA spectrum of ubiquitin.

Keywords: Data processing; Resolution enhancement; Triple-resonance spectra

1. Introduction

Structural genomics, defined as the rapid throughput of protein structure determination, is promising to complement from both a functional and structural point of view the available data from the various genome projects [1]. Whereas X-ray crystallography remains the method of choice for large macromolecules, the modular nature of many proteins with an average domain length of 150–200 amino acids makes Nuclear Magnetic Resonance (NMR) a valid alternative, and recent developments have significantly extended the size accessible for NMR [2,3]. Double or triple labeling [4,5] combined with advanced triple-resonance techniques [6] and automated assignment software [7] have substantially cut down on the time to assign the various resonances, whereas improved analysis of the NOE contacts together with absolute constraints from partially oriented samples [8] lead to improved efficiency of the structure generation procedures. The inherently low sensitivity of NMR will benefit of low temperature reception coils such as recently implemented in the cryoprobes but, even with the best hardware, the data acquisition of a complete set of NMR data still requires several weeks to months of machine time, and therefore puts a limit to the throughput.

The time requirement of the triple-resonance spectra is intimately linked to the resolution requirement. Indeed, in order to obtain a reasonable resolution after Fourier transform of the time domain signal, one needs to acquire sufficient data points. Especially in the indirect dimensions (^{15}N or ^{13}C) of a typical 3D triple-resonance experiment, where the acquisition of each individual complex point is equivalent to recording two 2D experiments, time versus resolution is a real dilemma. Various approaches were tried to improve the resolution of the powerful Discrete Fourier Transform (DFT) computed by the Fast Fourier Transform (FFT) algorithm, such as Linear Prediction (LP) [9,10],

[*] Corresponding author. Fax: +00-33-3-20-87-12-33.
E-mail address: guy.lippens@pasteur-lille.fr (G. Lippens).

1090-7807/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S1090-7807(03)00005-3

Maximum Entropy Method (MEM), maximum likelihood, and Bayesian methods, reviewed in [11]. The recently described Filter Diagonalization Method (FDM), initially used to identify and quantify the number of NMR transitions in a given NMR spectrum, was developed to accomplish the ultimate data compressing [12]. FDM extracts parameters directly from the time domain signal by fitting the data to a sum of exponentially damped sinusoids, and has successfully been applied up to 4D experimental data [13]. However, it should be stressed that all these procedures work well and give comparable results if the signal to noise ratio (S/N) is high, but they may perform differently and produce artifacts or ghost signals when S/N is too low.

A very simple, reliable and fast procedure, based on a restricted least square minimization, has been applied in the field of Magnetic Resonance Imaging (MRI) to overcome strong truncation effects and improve the accuracy of any 1D or *n*D experiment provided the frequency components and line shapes are known [14]. In solid state NMR, provided the summation onto one of the indirect dimensions is isotropic and known in advance, a similar fitting procedure was developed under the name of TIGER [15], and was extended to the case of general line shapes by the ANAFOR strategy [16]. The latter procedure applies when everything but the complex amplitudes of the different lines is known, at least in the dimension where one wants to apply ANAFOR. The ANAFOR acronym has been suggested by Taulelle [16] as "ANAlysis of FOuRier," and recalls "anaphora" a literature term well known in literature, that consists of repeating the same word or phrase—the prior knowledge—at the beginning of successive clauses or sentences to reinforce their meaning. However restrictive this notion of prior knowledge might appear, we show here that this assumption is essentially fulfilled in all heteronuclear 3D spectra of biomolecules. Indeed, all these spectra are collections of a basic 2D experiment (often a ^{1}H–^{15}N HSQC spectrum), that is amplitude modulated to give the frequency components in the third (or higher) dimension. The aim of the present paper is to show that efficient use of this a priori information allows to overcome to a large extent the above described dilemma of resolution versus time.

2. Restricted least square procedure

ANAFOR uses a restricted linear least square procedure in order to calculate in the time domain the complex amplitudes (A_k) of the components of the Free Induction Decay (FID), knowing in advance the number of resonances and their line features: frequency and lineshape. This prior knowledge limits its application but makes the procedure general as not only exponentially damped sinusoids are tolerated. Similarly to LP, a merit function χ^2 is minimized:

$$\chi^2 = \sum_{i}^{N_{\mathrm{f}}} \left(\frac{y_{\exp}(i) - y_{\mathrm{cal}}(i)}{\sigma_i} \right)^2, \tag{1}$$

where N_{f} is the number of experimental points, $y_{\exp}(i)$ and $y_{\mathrm{cal}}(i)$ are, respectively, the ith sampled experimental and calculated FID points, and σ_i is the standard deviation of $y_{\exp}(i)$. The summation over i can be performed for linearly or non-linearly sampled data points. This not only eliminates the need for the time signal to start at time zero, but also allows an easy tailoring of the recorded time series in order to optimize the S/N ratio and/or experimental time [17].

The general form of $y_{\mathrm{cal}}(i)$ is

$$y_{\mathrm{cal}}(i) = \sum_{k=1}^{M} A_k f_k(i), \tag{2}$$

where $M \leqslant N_{\mathrm{f}}$ is the number of resonances. The function $f_k(i)$ corresponds to the kth component of the FID (known in advance) and A_k is its amplitude that has to be calculated. This prior knowledge of the frequencies is the essential difference with the linear prediction algorithm, where the frequency components are a priori unknown variables that have to be calculated from the time signal. Moreover, the shape of the basis function $f_k(i)$ can be arbitrary, and is definitely not limited to the Lorentzian line usually assumed in LP algorithms. The minimum of Eq. (1) occurs where the derivatives of χ^2 with respect to all parameters A_k vanish, yielding the M normal equations:

$$0 = \sum_{i}^{N_{\mathrm{f}}} \frac{1}{\sigma_i^2} \left(y_{\exp}(i) - \sum_{k=1}^{M} A_k f_k(i) \right) f_k(i); \quad k = 1, \ldots, M. \tag{3}$$

Because the algorithmic used to solve this set of equations has been described in detail in the framework of its application to solid state NMR spectra [16], we give a brief analytical description of the concept for two simplified cases of a cosine transform of one single and two resonance lines on a given trace, neglecting any damping or noise contribution.

When we know a priori that there is only one frequency component in the time signal with known frequency ω_1, then the general Eq. (2) reduces to

$$y_{\mathrm{cal}}(i) = A_1 \cos(\omega_1 t_i). \tag{4}$$

Contrary to the well-known linear prediction algorithms, the a priori knowledge of the frequency value ω_1 will be used in ANAFOR, whereas LP first has to derive the number of frequency components and their frequencies, and only then further proceeds to derive their amplitudes. If the N_{f} first experimental points are considered, the sum to be minimized with respect to the unknown amplitude A_1 is

$$\sum_{i=1}^{N_f} \left(y_{\exp}(i) - A_1 \cos(\omega_1 t_i)\right)^2, \tag{5}$$

which leads to the following solution:

$$A_1 = \frac{\sum_{i=1}^{N_f} y_{\exp}(i) \cos(\omega_1 t_i)}{\sum_{i=1}^{N_f} \cos^2(\omega_1 t_i)}. \tag{6}$$

This is identical to the result that would be given by the cosine transform for the same time signal. If moreover $y_{\exp}(i) = A\cos(\omega_1 t_i)$, Eq. (6) reduces to $A_1 = A$.

The situation is somewhat more complex for a time signal that contains two frequency components, with frequencies ω_1 and ω_2, but unknown amplitudes A_1 and A_2. Similarly as above, we define the calculated signal

$$y_{\mathrm{cal}}(i) = A_1 \cos(\omega_1 t_i) + A_2 \cos(\omega_2 t_i) \tag{7}$$

and, derive the sum to be minimized with respect to the two independent parameters A_1 and A_2 as

$$\sum_{i=1}^{N_f} \left(y_{\exp}(i) - A_1 \cos(\omega_1 t_i) - A_2 \cos(\omega_2 t_i)\right)^2. \tag{8}$$

Partial deriving this sum with respect to both unknown amplitudes leads to a set of two linear equations, that can be easily solved to give

$$A_1 = \frac{\sum_{i=1}^{N_f} y_{\exp}(i)\cos(\omega_1 t_i) \sum_{i=1}^{N_f} \cos^2(\omega_2 t_i) - \sum_{i=1}^{N_f} y_{\exp}(i)\cos(\omega_2 t_i) \sum_{i=1}^{N_f} \cos(\omega_1 t_i)\cos(\omega_2 t_i)}{\sum_{i=1}^{N_f} \cos^2(\omega_1 t_i) \sum_{i=1}^{N_f} \cos^2(\omega_2 t_i) - \left(\sum_{i=1}^{N_f} \cos(\omega_1 t_i)\cos(\omega_2 t_i)\right)^2}. \tag{9}$$

When we fill out the explicit form of $y_{\exp}(i) = A\cos(\omega_1 t_i) + B\cos(\omega_2 t_i)$, we find as before $A_1 = A$ and $A_2 = B$. It is thus clear that two experimental points are sufficient to give the correct answer if, as stated before, noise is neglected.

It is instructive to compare this result with those of the regular cosine transform. Indeed, real Fourier transformation yields an amplitude of the frequency component at ω_1 given by

$$A(\omega_1) \approx \sum_{i=1}^{N_f} y_{\exp}(i) \cos(\omega_1 t_i), \tag{10}$$

which, involves both amplitudes A and B when the experimental signal is the sum of two frequency components

$$A(\omega_1) \approx A \sum_{i=1}^{N_f} \cos^2(\omega_1 t_i) + B \sum_{i=1}^{N_f} \cos(\omega_1 t_i)\cos(\omega_2 t_i). \tag{11}$$

The non-orthogonal nature of the finite summation of the two cosine components leads to an effective mixing, that will be all the more effective when the number of experimental points is smaller and the frequency separation lesser. This illustrates well the main advantage of ANAFOR over the DFT algorithm, as its use of the prior frequency knowledge will lead to a better separation of the different frequency components.

3. Application to heteronuclear biomolecular spectra

Fig. 1 shows the annotated ^{1}H–^{15}N HSQC spectrum of ubiquitin. As stated above, because of the hyphenated nature of modern triple-resonance spectroscopy, many 3D spectra can be written down as a stack of this same 2D spectrum, where the amplitudes of the correlation peaks represent a wave modulation in the third dimension. For example, in a HNCA spectrum [6], the modulation corresponds to the carbon frequencies of one or more Cα nuclei, but it could well be proton frequencies as is the case in the NOESY-HSQC spectrum.

The strategy to exploit this prior information becomes clearer when we consider the ^{15}N trace through the isolated correlation peak 1. Because only one unique nitrogen frequency corresponds to this peak (Fig. 1), we do not need more than one complex point in the second dimension for all subsequent corresponding traces in the 3D experiment. Indeed, prior knowledge of the uniqueness of its line position and lineshape allows

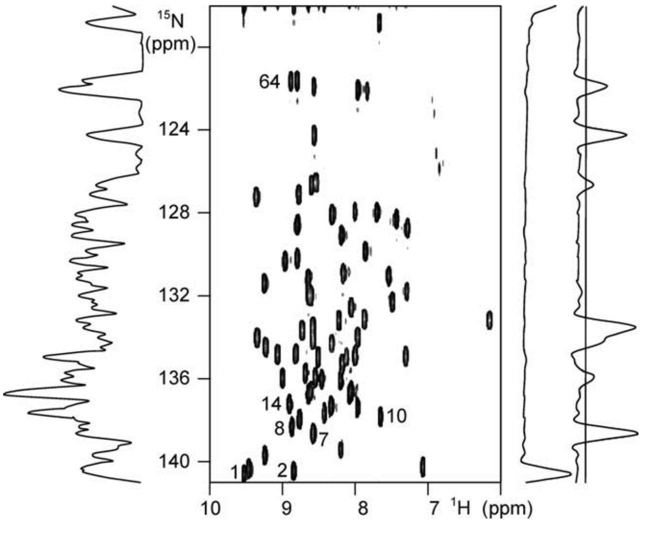

Fig. 1. ^{1}H–^{15}N correlation spectrum of ubiquitin corresponding to the first plane of the 3D HNCA experiment. The spectrum was recorded with 34 complex points in the ^{15}N direction, and 64 scans per increment. Peak picking was done automatically within the SNARF program (F. van Hoesel, Groningen, the Netherlands). Only those peaks that are discussed in the text have been annotated. The projection of the spectrum onto the nitrogen dimension is reported on the left. On the right, two individual ^{15}N traces through peak 1 and peak 7 are shown. In the latter trace, choosing the threshold as indicated by the line leads to 7 frequency components.

direct evaluation of its complex amplitude from the first data point, as Fourier transform tells us that the first point of the time domain signal is equivalent to the integral of the frequency spectrum. The presence of a unique signal in the ^{15}N dimension makes evidently any spreading out in this dimension without any consequence. For other cross peaks, amide proton chemical shift degeneracies can occur, and they will be even more severe as the protein is larger. For example, the ^{15}N trace corresponding to peak 7 (Fig. 1), contains at least seven peaks. Here again, however, a standard peak picking routine as implemented in most NMR packages can yield the frequency position of everyone of these peaks, and these will be invariant over further planes of the 3D experiment. The problem of linewidth estimation is less trivial, but can equally be overcome (see below). If only we give sufficient data points to assure that the least square problem of Eq. (1) is not underdetermined, we can hopefully extract the correct amplitudes for all further planes, and use these to reconstruct the time signal in the third dimension.

It is clear from these few examples that the procedure will require a reference 2D spectrum, in which both resolution and signal to noise ratio should allow a good determination of all correlation peaks that one wants to further consider. The requirement of a 2D reference spectrum distinguishes our procedure from the previously described TIGER algorithm [15] or even from the previously described application of the restricted linear least square procedure to 2D solid state spectra [16], that both required one single 1D spectrum for the indirect dimension. In the case of biomolecular spectra, this would be equivalent to the ^{15}N projection of the ^{1}H–^{15}N HSQC spectrum (Fig. 1, left), but it is intuitively clear that use of the individual traces rather than the projection greatly simplifies the problem.

In the next paragraph, we will first describe the concrete implementation of the algorithm as applied to a 3D HNCA spectrum of ubiquitin, before going deeper into such considerations as resolution and sensitivity enhancement.

4. Implementation for the 3D HNCA spectrum of ubiquitin

The 2D ^{1}H–^{15}N reference spectrum (here taken as the first plane of the HNCA, with a large number of complex points in the indirect ^{15}N dimension) was transformed by DFT in both dimensions, and peaks were picked automatically. The result is a number of coordinates for both proton and nitrogen frequencies, that are stored in two separate variables. Lineshapes can equally be extracted in a straightforward fashion when sufficient points are available in the indirect dimension, but we will see further that this is not even a stringent criterion. The 2D ^{1}H–^{15}N planes of the HNCA spectrum are then processed individually using these coordinates. For each plane, we first apply the DFT algorithm to the direct proton dimension, and then derive the 1D ^{15}N interferograms with N_f complex points for those lines that correspond to a relevant proton frequency according to the reference spectrum. Working with traces rather than strips simplifies the procedure, but the latter can be easily implemented by taking the relevant proton frequencies surrounded by a predefined number of points. The information for the least square processing of the interferogram comes from the equivalent ^{15}N trace in the reference spectrum, where the peak picking with a predefined threshold defines all peaks to be considered on that line. With this information of line positions and lineshapes, the set of base functions is constructed, and the least square problem of Eq. (1) is solved. The complex amplitudes can be used directly to construct the interferogram in the third dimension, or serve to reconstruct a ^{15}N interferogram with the desired number of points, that then is transformed by the regular DFT algorithm.

In order to make our strategy more tangible, we show the different steps in Fig. 2. The ^{15}N trace of peak 10 in the reference spectrum contains 3 peaks at the frequencies of −666, −108, and 494 Hz with respect to the offset, with an estimated natural linewidth of 1 Hz (Fig. 2a). With this information and the experimental points of the interferogram corresponding to the first plane of a

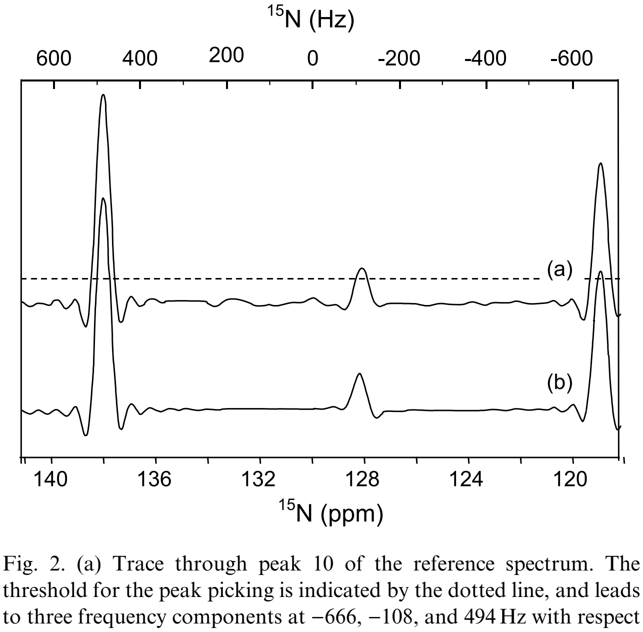

Fig. 2. (a) Trace through peak 10 of the reference spectrum. The threshold for the peak picking is indicated by the dotted line, and leads to three frequency components at −666, −108, and 494 Hz with respect to the offset. This information together with the corresponding interferogram from first plane of the HNCA (recorded with eight complex points in the ^{15}N direction and 4 scans per increment) is used as input for the restricted least square minimization algorithm, that determines the three amplitudes. (b) The interferogram reconstructed to 34 complex points is transformed by FFT.

3D HNCA spectrum recorded with only eight complex points in the ^{15}N dimension, the complex amplitudes of the 3 components are calculated, and these are used to reconstruct an interferogram of 34 complex points. This latter is then transformed by DFT (Fig. 2b), and written into the 2D matrix, before going to the next ^{15}N interferogram corresponding to the following cross peak. All consecutive 2D planes are transformed in a similar way, building up the full 3D matrix. Whereas we at this moment could consider a similar restricted least square procedure for the third dimension, this would require a different ^{13}C reference spectrum for the different HNCA, HNCO, CBCACONH,... triple-resonance experiments. In order to avoid this lengthy recording of all reference spectra, we limit here the restricted least square procedure to the ^{15}N dimension. We therefore recorded the standard number of complex points in the carbon dimension, and transformed the spectra in this third dimension by the regular DFT algorithm.

In the following paragraphs, we will discuss in further detail the importance of such parameters as number of lines and threshold choice, resolution, lineshape, number of experimental data points and finally S/N ratio of the spectrum.

5. Number of lines

If we imagine a ^{15}N trace of the reference 2D spectrum that contains two lines, it is clear that one complex point is not sufficient to calculate the independent intensity of both lines. Indeed, the mathematical minimization problem of Eq. (1) is underdefined, or, alternatively, with one single point, the amplitude can be freely distributed between both lines. If we give a second experimental point, and the data are noiseless, the problem is in principle totally determined. However, the data are subject to noise, and the amplitude distribution between both lines will be more reliable with an increasing number of experimental points and with a growing frequency difference between both lines. For the ubiquitin spectrum, the S/N ratio is very good, and as shown in Fig. 3, even with a separation of 40 Hz, both lines can be distinguished with reasonable accuracy when we use as little as four complex points.

This analysis, however, does not deal with the general problem of underdetermination: what happens when the number of lines on a trace is larger than the number of experimental points? This problem that, especially for the central region of the ^{1}H–^{15}N spectrum of a larger protein, seems to limit the usefulness of the method, has been investigated previously by Knijn et al. [18] in the framework of frequency-selective quantification in the time domain, and relates to the non-orthogonality of the cosine functions whenever they are sampled over a finite number of points. Knijn et al. [18] concluded that the contribution of the ignored sinusoid relates to the height of the corresponding frequency-domain spectral peak at the position of the wanted peak. This result is illustrated by the simulations of a spectrum composed of two lines of intensities 1 and 0.5, processed with the assumption of a single line (Fig. 4). If the two frequencies are very close, the algorithm assigns both intensities to the same line. As the separation between the lines increases, omitting the second line has less and less influence on the calculated intensity of the first one. When normalized to the initial intensity of 1.5, the intensity approaches rapidly its true value of 1 as the frequencies start to diverge (Fig. 4). That the convergence speed

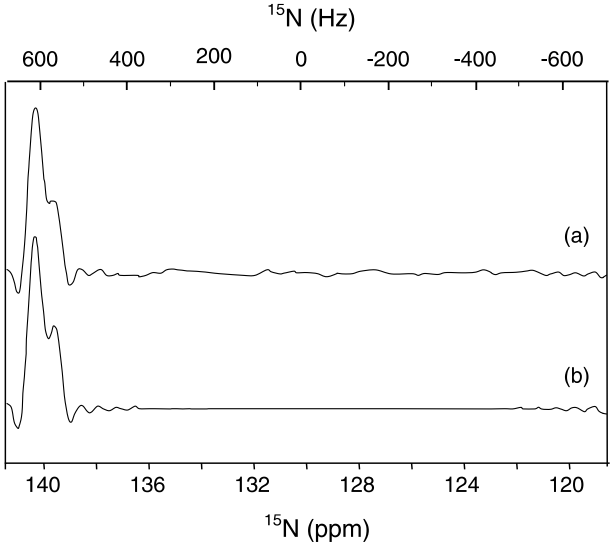

Fig. 3. Reconstruction of a trace containing two lines separated by 40 Hz. In (a) the trace through peak 1 of the reference spectrum was added to itself after scaling by a factor of 0.5 and a frequency shift of 40 Hz. (b) Reconstructed trace using four complex points of the interferogram corresponding to the spectrum in (a), and the information on the two frequency components.

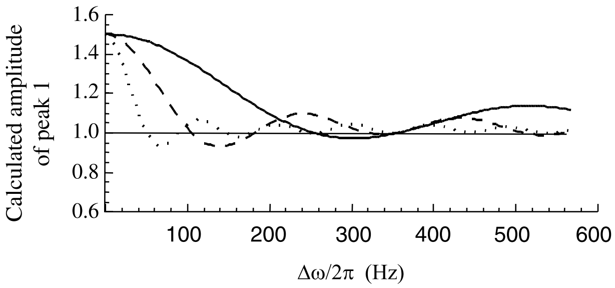

Fig. 4. Effect of omitting a frequency component. Synthetic time signals composed of two frequency components shifted by variable $\Delta\omega$ (generated as described in legend of Fig. 3), have been processed assuming the presence of a single component. The calculated amplitude is reported as a function of the separation between the two lines. When both lines coincide, the full intensity (1.5) is associated with the unique line, but its intensity correctly converges to one as the separation between the lines increases. The amplitudes were calculated with 4 (solid line), 8 (hyphened) or 16 (dotted line) complex points.

increases with an increasing number of experimental points is again intuitively clear. With more points, the component in the time signal that is not at the given frequency will contribute less and less effectively to the calculated intensity of the first line. We found that the influence of omitting a frequency component that is actually present in the FID becomes negligible when $\Delta \nu N_{\mathrm{f}} \tau > 0.6$, where $\Delta \nu$ is the frequency difference and τ is the dwell time. With eight complex points recorded, and a typical dwell time of 500 μs, this result implies that if there is more than eight frequency components on a given trace, the base functions with frequency values outside the range of ±150 Hz from the line considered can safely be neglected.

6. Influence of the threshold

The threshold chosen to pick an individual trace in the reference spectrum immediately influences the number of basis functions that one will impose in the least square treatment of the corresponding traces in the 3D spectrum. This paragraph therefore serves as a concrete example of the above described discussion on the number of lines. If we look at the ^{15}N trace of the reference spectrum through peak 8 (Fig. 5a), we find, close to peak 8 at 524 Hz, three other components that correspond to other amide groups with almost the same ^{1}H frequency (peak 2 at 656 Hz, peak 14 at 458 Hz and peak 64 at −492 Hz). The weaker intensity of these lines reflects the fact that the trace through peak 8 does not intersect the center of the other lines but rather their wings. Depending on the threshold that we will choose to pick this trace, the number of peaks can vary between one and four. If we choose the threshold above the maximum of the three minor peaks (solid line in Fig. 5a), the algorithm will assign all intensity to peak 8. Peaks 2 and 64 (at 656 and −492 Hz, respectively) will not influence significantly its intensity if we use eight complex points, as $\Delta \nu N_{\mathrm{f}}\ \tau > 0.6$. However, peak 14 separated by only 66 Hz from peak 8 (and thus within the range defined by $\Delta \nu N_{\mathrm{f}}\ \tau < 0.6$) will contribute to the amplitude of the latter. As moreover the intensity of peak 14 is modulated by its own ^{13}C frequency, its contribution to peak 8 is equally modulated. A Fourier transform in the third dimension (Fig. 5e) therefore shows, next to the ^{13}Cα frequencies corresponding to peak 8, ^{13}Cα signals that correspond to peak 14 (Fig. 5f). Decreasing the threshold a first time (hyphened line in Fig. 5a) introduces two additional frequency components to the fit, but not the one corresponding to peak 14 (dot line in Fig. 5a), and the spurious ^{13}Cα signals remain (Fig. 5d). Only when we decrease the threshold such as to take explicitly into account peak 14, its amplitude modulation does not contribute anymore to the calculated intensity for peak 8. As a result, the spurious modulation at the ^{13}Cα frequency of peak 14 disappears from the ^{13}Cα line of peak 8 (Fig. 5c). Whereas one could object that the threshold choice has such an important influence that it should almost be performed manually for every peak, putting an impossible burden on the operator, we found that a threshold at three times the noise level of the spectrum does efficiently eliminate the above described problems, without introducing any artifacts.

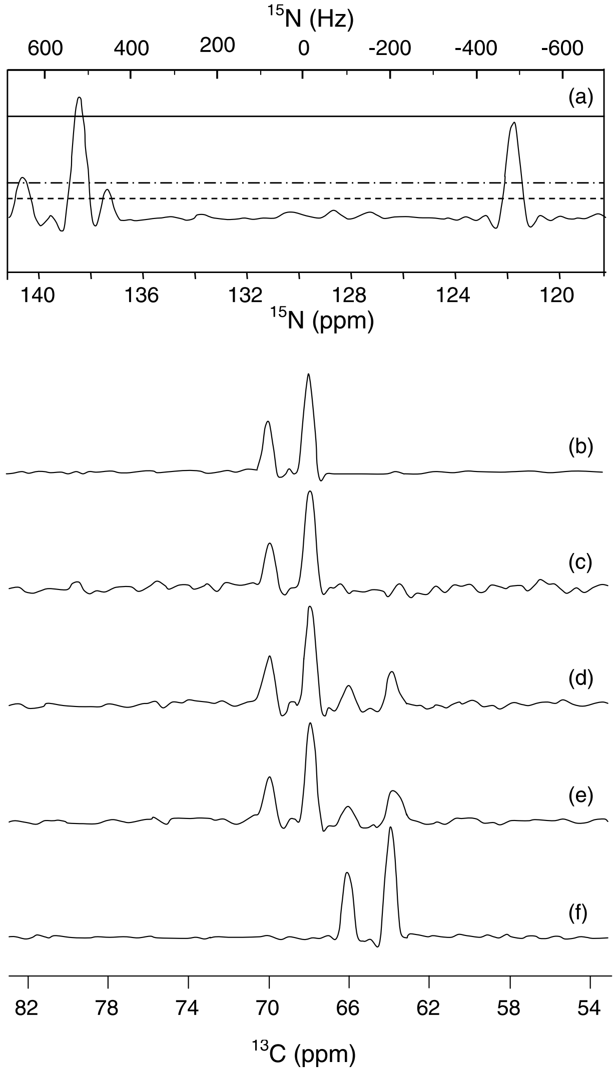

Fig. 5. (a) Trace of the reference spectrum through peak 8 with four frequency components at 656, 524, 458, and −492 Hz with respect to the offset. The choice of the threshold (in arbitrary units, 6 for the full line, 2 for the hyphened line and 1 for the dotted line) determines whether 1, 3, or all 4 peaks are taken into account for the subsequent minimization. Spectrum (b) is the ^{13}C trace of peak 8 from the HNCA acquired with 34 complex points in the ^{15}N dimension and transformed by FFT. From (c) to (e), the resulting ^{13}C spectra in the third dimension coming from a restricted least square minimization of the HNCA planes with eight complex points in the ^{15}N dimension, and taking the threshold at 1 (c), 2 (d), or 6 (e). Spectrum (f) is the ^{13}C modulation corresponding to peak 14, obtained from the full FFT transformed HNCA spectrum, that leaks into the ^{13}C interferogram of peak 8 when not taken into account.

The above results illustrate clearly that taking into account all frequency components in the immediate neighborhood of the peak under consideration is an important criterion. Based on the requirement of a number of experimental points larger than the number of frequency components within the range of $\Delta\nu N_f\ \tau < 0.6$ from the peak under consideration, taking eight experimental complex points, limits to eight the number of ^{15}N frequency components that we can consider within a range of 150 Hz of the central line, which is totally acceptable for the ^{15}N resolution on a 14.1T spectrometer.

7. Influence of the linewidth

Although applicable to any line shape [16], our initial assumption of Lorentzian lines for the ubiquitin ^{15}N resonances makes the algorithm fit the experimental FID to a sum of damped exponential sinusoids. We therefore need not only the frequencies but also the damping constants of every component. These damping components can easily be extracted from the reference HSQC, providing a large number of ^{15}N increments are recorded. Since this experiment has to be performed only once, it is not very time-consuming to record it with a ^{15}N evolution time long enough (3 times the T_2 or more). Because in the case of ubiquitin lines are extremely narrow, we have decided to treat the influence of the linewidth on a different protein, the PA92 mutant of the 13 kD Cyclin dependent Kinase Subunit (CKS) of *S. pombe,* which shows a faster relaxation than ubiquitin [19]. A ^{1}H–^{15}N HSQC recorded with 512 complex points in the ^{15}N evolution was transformed by DFT without application of any filter. Whereas this gave immediately a good estimate of the linewidth of every peak, the fitting algorithm itself can be used to estimate the widths when few lines are present. Starting from a trace of the reference ^{1}H–^{15}N HSQC spectrum of this protein recorded with only 34 points in the ^{15}N dimension, we performed the reconstruction by introducing different linewidths (Fig. 6a). When the linewidth is underestimated (3 Hz instead of the actual 11 Hz), the initial amplitude of the reconstructed FID decreases as the number of experimental points taken into account for the reconstruction grows. Inversely, when we overestimate the linewidth (30 Hz instead of 11 Hz), giving more experimental points leads to an increase in initial signal intensity, as the algorithm tends to compensate the too rapid signal decrease by an initial higher intensity (Fig. 6a). This indicates not only the error introduced by an erroneous estimation of the linewidth, but it also gives a method to optimize the estimation. Only for the correct linewidth, the reconstructed FID follows the experimental signal, irrespective of the number of experimental points taken into account. For a single line, we therefore can extract the linewidth by reconstruction of the corresponding trace of the reference spectrum with a varying number of experimental points, and by requiring that at the correct linewidth the initial amplitude of the reconstructed FID does not depend on the number of experimental points that we take into account (Fig. 6b). Whereas this method leads to a precise estimate of the true linewidth, but at the same time shows that the input T_2 values are not a very critical parameter, it is not well suited for the simultaneous linewidth estimation of several frequency components, as compensation effects can occur. Therefore, especially in the case of inhomogeneous line broadening of the different correlation peaks, we recommend using the reference spectrum to extract the linewidths.

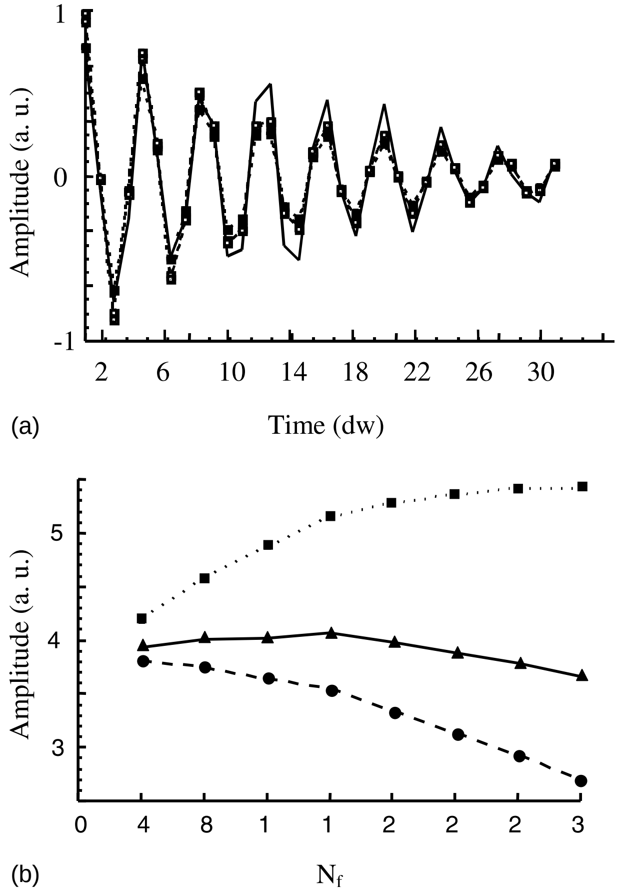

Fig. 6. (a) The interferogram of a trace through the ^{15}N HSQC of the CKS protein p13^{suc1} (solid line) was calculated using 4 (filled squares), 16 (open squares), or 32 (dotted squares) complex points, and an overestimated linewidth of 30 Hz. Enforcing a too rapid decay of the FID leads to an increase of the initial amplitude when we take more complex points into account. Amplitude axis is in arbitrary unit and time axis in unit of dwell time. (b) Initial amplitude of the same interferogram as a function of the number of complex data points used, with estimated linewidths of 3 Hz (circles), 11 Hz (triangles) or 30 Hz (squares). When the linewidth is correctly estimated, the calculated FID follows closely the experimental one, and the initial amplitude becomes independent of the number of data points used in the minimization procedure.

8. Uncertainty of the calculated amplitude as a function of N_f

For a trace with one single peak, such as the one corresponding to peak 1 in the 2D spectrum of ubiquitin (Fig. 1), a single complex point of the interferogram is sufficient to calculate the complex amplitude of the corresponding spectral line. The uncertainty of the calculated amplitude will, in this case, be that of the experimental data point. If we use more data points, and impose that the time signal corresponds to one peak with a given frequency and lineshape, one can readily understand that the error margin on the amplitude will decrease. In fact, it has been shown [16] that in the case of one single Lorentzian line, the incertitude of the calculated amplitude is given by

$$\sigma_A = \sqrt{\frac{1 - e^{-2/\rho}}{1 - e^{-2N_f/\rho}}} \cdot \sigma_n, \tag{12}$$

where ρ is a density of points, corresponding to the number of sampled points during one T_2^* interval, σ_n is the standard deviation of the noise and, σ_A the standard deviation of the real and imaginary parts and also of the modulus of the complex amplitude A. Eq. (12) has been confirmed in the particular case of solid state REDOR measurements but it is important to check its validity in our frame of biomolecular NMR. In the case of solid state spectra, T_2^* could be as short as the dwell time ($\rho < 10$) but in the case of workable protein spectra, it will be large compared to a typical dwell time of hundreds of μs ($\rho > 100$).

In the ideal case of an infinite density of points ($\rho \gg N_f$), the error estimate simplifies to

$$\sigma_A = \frac{\sigma_n}{\sqrt{N_f}}. \tag{13}$$

This expression signifies that every point of the interferogram contributes equivalently to the signal, and therefore an increase in the number of experimental points in the interferogram is identical to an increase in the number of accumulated scans. In Fig. 7, the estimated uncertainty on the calculated amplitude of the single frequency component corresponding to the trace through peak 1 of ubiquitin as a function of the number of experimental points fitted confirms the above derived expression.

In the case of extensive line broadening, due to the slow tumbling for larger proteins or due to non-uniform T_2 relaxation in the case of conformational heterogeneity, limiting the data points to short evolution times does lead to some sensitivity enhancement. Eq. (12) was then tested with a Monte Carlo simulation on a corpus of 100 synthetic signals. As a noise estimate, we extracted traces from the empty region of the ^{1}H–^{15}N HSQC, corresponding to proton frequencies between 0 and 2 ppm, and added a synthetic signal corresponding to the same frequency component of 16 Hz width (T_2 value of 20 ms). Because the restricted linear least square minimization is identical to DFT when one single component is present, we could use our procedure to calculate the amplitude of the single component both with eight or 32 complex points. The variance of the calculated amplitude when considering the 100 noise traces was 7% of the full signal amplitude when we used eight complex points from the HSQC recorded with 512 scans, but increased to 11% when we took 32 complex points from the HSQC recorded with 128 scans. In this example, where we have 29 sampled points in a T_2 interval, the general form [12] derived for the variance of the signal amplitude predicts a 60% increase for the variance, in agreement with the numbers we found in our simulations. In conclusion, Eq. (12) describing the standard deviation of the calculated amplitude as derived in [14] remains valid for biomolecular NMR. When linewidths are narrow, it can be simplified to Eq. (13). We may further add that the uncertainty is mathematically expected to be less with ANAFOR than with DFT but no major gain is to be expected when DFT is performed in optimal conditions without truncation and with the use of a matched filter [14].

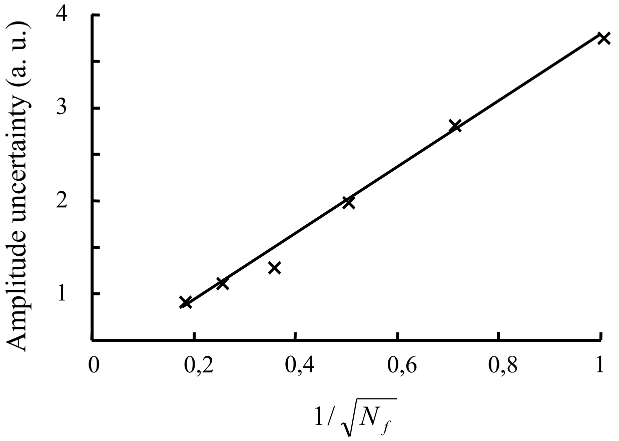

Fig. 7. Estimated uncertainty on the calculated amplitude as a function of the number N_f of fitted experimental points. Input data were the single frequency and linewidth of the trace through peak 1 of the reference spectrum, and the interferogram corresponding to the same peak 1 in the first plane of the HNCA experiment.

9. Conclusion

In the present paper, we have revisited the restricted least square minimization procedure as a tool for the processing of biomolecular multidimensional spectra. It was shown that high resolution 2D HSQC spectrum can advantageously be used as a reference spectrum to give the necessary input of frequency components and linewidths for data treatment of the 3D spectrum. Even combined with a limited time evolution in the indirect dimension, the amplitudes of the correlation peaks in all planes of the 3D spectra can be extracted, and can be used to reconstruct the interferograms in the third dimension. We have demonstrated this for a HNCA spectrum, where the use of only eight complex points leads to a fourfold gain in time, by far compensating the initial time spent on the 2D reference spectrum. Evidently, in other applications such as heteronuclear relaxation measurements, where in a similar way a series of 2D spectra has to be acquired, equivalent gains in time can be obtained. Because moreover one single ^{1}H–^{15}N HSQC can form the reference spectrum for the whole set of triple-resonance spectra commonly used to assign small to intermediate sized proteins, we believe that the proposed processing will show its use in the field of structural genomics, as it allows a simple procedure to work with truncated data sets.

Acknowledgments

We thank Dr. J.-M. Wieruszeski for recording the 3D spectra on ubiquitin and for expert help with the figures, Prof. B. Kusse (Cornell University, USA) for stimulating discussions on the DFT algorithm, Dr. S. Macura (Mayo Clinic, USA) for pointing out the MRI literature and Dr. J. Martins (University of Ghent, Belgium) for carefully reading the manuscript. Part of this work was accomplished in the framework of the Groupement Scientifique RMN Nord-Pas de Calais (France).

References

[1] S.K. Burley, An overview of structural genomics, Nature Struct. Biol., Structural Genomics Supplement, (2000) November issue.

[2] K. Pervushin, R. Riek, G. Wider, K. Wüthrich, Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution, Proc. Natl. Acad. Sci. USA 94 (1997) 12366–12371.

[3] J. Fiaux, E.B. Bertelsen, A.L. Horvich, K. Wüthrich, NMR analysis of a 900kD GroEL/GroES complex, Nature 418 (2002) 207–210.

[4] L.P. McIntosh, F.W. Dahlquist, Biosynthetic incorporation of ^{15}N and ^{13}C for assignment and interpretation of nuclear magnetic resonance spectra of proteins, Quart. Rev. Biophys. 23 (1990) 1–38.

[5] D.M. LeMaster, Deuterium labelling in NMR structural analysis of larger proteins, Quart. Rev. Biophys. 23 (1990) 113–174.

[6] M. Sattler, J. Schlecher, C. Griesinger, Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients, Prog. Nucl. Magn. Reson. Spec. 34 (1999) 93–158.

[7] H.N.B. Moseley, G.T. Montelione, Automated analysis of NMR assignments and structures for proteins, Curr. Opin. Struct. Biol. 9 (1999) 635–642.

[8] N. Tjandra, A. Bax, Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium, Science 278 (1997) 1111–1114.

[9] H. Barkhuijsen, R. De Beer, W.M. Bovée, D. van Ormondt, Retrieval of frequencies amplitudes damping factors and phases from time-domain signals using a linear least-squares procedure, J. Magn. Reson. 61 (1985) 465–471.

[10] P. Koehl, Linear prediction spectral analysis of NMR data, Prog. Nucl. Magn. Reson. Spec. 34 (1999) 257–299.

[11] R.E. Hoffman, G.C. Levy, Modern methods of NMR data processing and data evaluation, Prog. Nucl. Magn. Reson. Spec. 23 (1991) 211–258.

[12] V.A. Mandelshtam, The multidimensional filter diagonalization method, J. Magn. Reson. 144 (2000) 343–351.

[13] H. Hu, A.A. De Angelis, V.A. Mandelshtam, A.J. Shaka, The multidimensional filter diagonalization method, J. Magn. Reson. 144 (2000) 357–366.

[14] R. De Beer, D. van Ormondt, Analysis of NMR data using time domain fitting procedures, in: P. Diehl, et al. (Eds.), NMR Basic Principles and Progress, 26, Springer Verlag, Berlin, 1992, pp. 202–248.

[15] G. McGeorge, J.Z. Hu, C.L. Mayne, D.W. Alderman, R.J. Pugmire, D.M. Grant, Technique for importing greater evolution resolution in multidimensional NMR spectrum, J. Magn. Reson. 129 (1997) 134–144.

[16] (a) S. Steurnagel, A. Bailly, J.P. Amoureux, V. Munch, P. Koehl, F. Taulelle, Magic Angle Turning of Q8M8 processed by Anafor. Experimental NMR Conference, Orlando, 1999;
(b) F. Taulelle, C. Carlotti, V. Munch, G. Fink, P. Bodart, J.P. Amoureux, P. Koehl, Anafor processing of solid state experiments, Experimental NMR Conference, Asilomar, 2000;
(c) P. Bodart, J.P. Amoureux, G. Fink, F. Taulelle, Anafor in MQMAS-based methods, Experimental NMR Conference, Orlando, 2000;
(d) G. Fink, F. Taulelle, Processing spectra using Anafor, European Experimental NMR Conference, Prague, 2002;
(e) P.R. Bodart, J.-P. Amoureux, F. Taulelle, ANAFOR: application of a restricted linear least squares procedure to NMR data processing, Solid State NMR 21 (2002) 1–20.

[17] P. Schmieder, A.S. Stern, G. Wagner, J.C. Hoch, Improved resolution in triple-resonance spectra by nonlinear sampling in the constant-time domain, J. Biomol. NMR 4 (1994) 483–490.

[18] A. Knijn, R. De Beer, D. van Ormondt, Frequency selective quantification in the time domain, J. Magn. Reson. 97 (1992) 444–450.

[19] B. Odaert, I. Landrieu, K. Dijkstra, G. Schuurman-Wolters, P. Casteels, J.-M. Wieruszeski, D. Inzé, R. Scheek, G. Lippens, Solution NMR study of the monomeric form of p13^{suc1} protein sheds light on the hinge region determining the affinity for a phosphorylated substrate, J. Biol. Chem. 277 (2002) 12375–12381.