Journal of Biomolecular NMR (2005) 33: 139–152
DOI 10.1007/s10858-005-3093-1


*Article*

# Automated protein NMR structure determination using wavelet de-noised NOESY spectra

Felician Dancea[a] & Ulrich Günther[b,*]
[a]*Center for Biomolecular Magnetic Resonance (BMRZ), Institute of Biophysical Chemistry, J. W. Goethe-University of Frankfurt, Frankfurt am Main, Germany;* [b]*Henry Wellcome Building for Biomolecular NMR Spectroscopy (HWB-NMR), CR UK Institute for Cancer Studies, University of Birmingham, Vincent Drive, Edgbaston, Birmingham, B15 2TT, United Kingdom*

Received 5 May 2005; Accepted 9 September 2005

*Key words:* NMR, NOESY, peak picking, protein structure determination, wavelet de-noising

**Abstract**

A major time-consuming step of protein NMR structure determination is the generation of reliable NOESY cross peak lists which usually requires a significant amount of manual interaction. Here we present a new algorithm for automated peak picking involving wavelet de-noised NOESY spectra in a process where the identification of peaks is coupled to automated structure determination. The core of this method is the generation of incremental peak lists by applying different wavelet de-noising procedures which yield peak lists of a different noise content. In combination with additional filters which probe the consistency of the peak lists, good convergence of the NOESY-based automated structure determination could be achieved. These algorithms were implemented in the context of the ARIA software for automated NOE assignment and structure determination and were validated for a polysulfide-sulfur transferase protein of known structure. The procedures presented here should be commonly applicable for efficient protein NMR structure determination and automated NMR peak picking.

*Abbreviations:* DWT – discrete wavelet transforms; NMR – nuclear magnetic resonance; NOE – nuclear Overhauser enhancement; NOESY – nuclear Overhauser enhancement spectroscopy; Sud – the polysulfide-sulfur transferase protein from *Wolinella succinogenes*.

## Introduction

Recent advances in automation of protein NMR structure determination were the product of a series of computational algorithms which link the iterative assignment of NOESY spectra with structure calculations (Mumenthaler and Braun, 1995; Mumenthaler et al., 1997; Nilges et al., 1997; Savarin et al., 2001; Herrmann et al., 2002a; Huang et al., 2003). While new types of constraints such as residual dipolar couplings (Tjandra and Bax, 1997), orientational information from heteronuclear relaxation in anisotropically tumbling molecules (Tjandra et al., 1997) or restraints obtained in the presence of paramagnetic centers in a protein (Banci et al., 1997) have facilitated protein structure determination, distance information from NOESY spectra remains an important basis for NMR structure elucidation. Peak picking in NOESY spectra has been a time consuming process, mainly due to spectral

*To whom correspondence should be addressed. Email: U.L. Gunther@bham.ac.uk.

overlap and because NOESY spectra are often obscured by noise and spectral artifacts. Therefore automation of the peak picking process requires reliable filters to select the relevant signals. A program which combines peak picking with automated structure determination by using intermediate protein structures as a guide for the interpretation of the NOESY spectra has been described previously (Herrmann et al., 2002b). Here we present a different approach to automated peak picking employing wavelet transforms to de-noise spectra prior to automated structure determination.

Discrete wavelet transforms (DWT) are commonly used for noise suppression and data compression. Recent applications of wavelet transforms to NMR show potential applications in NMR processing, in particular for the suppression of the water signal, for data compression and for de-noising (Hoch and Stern, 1996; Günther et al., 2002; Cancino-DeGreiff et al., 2002; Trbovic et al., 2005). Compared to other algorithms used to reduce spectral noise (such as Fourier and Savitzky-Golay filtering methods), wavelet de-noising is exceptionally stable and computationally efficient (Mittermayr et al., 1996; Shao et al., 2003). For optimal de-noising, noise reduction must be achieved while preserving the fine structure of the signals. The result depends predominately on three variables: the wavelet base function (e.g. Symmlet, Daubechies, Coiflet), the wavelet transform (e.g. periodic orthogonal, translation invariant) and the thresholding procedure (e.g. soft, hard). In this work the important de-noising variables were optimized for automated peak picking of protein NOESY spectra. As an additional filter we employ a consistency verification of the NOESY cross peak lists generated by the automated peak picking which is based on symmetries in, and between heteronuclear-edited NOESY spectra and on the fact that the NOE signals are usually part of a network of connectivities between adjacent spin systems. These algorithms were implemented in the context of the ARIA software (Linge et al., 2003) using routines from NMRLab (Günther et al., 2000), and were validated for a recently published polysulfide-sulfur transferase (Sud) protein structure (Lin et al., 2004).

## Theory

### *Multiresolution analysis (MRA) and wavelet series expansion*

MRA as introduced by Mallat (1989a) provides a general framework to construct a wavelet basis suitable to describe functions at different resolution levels. Starting from a father wavelet (scaling function) an orthonormal mother wavelet $\psi$ is obtained. Dyadic dilatations ($2^j$) yield nested subspaces which form a MRA. The base functions $\psi_{jk}$ are derived by additional translations ($k$):

$$\psi_{jk}(x) = 2^{j/2} \cdot \psi(2^j \cdot x - k). \tag{1}$$

The wavelet base functions have compact support, i.e. the wavelet is zero outside a finite interval $[k \cdot 2^{-j}, (k + 1) \cdot 2^{-j})$ and form an orthonormal basis for $\mathscr{L}^2(\mathrm{R})$ (the space of square integrable functions). Therefore any square integrable function $f(x)$ can be represented as a series of $\psi_{jk}$ with the corresponding scaling function $\phi_{j_0 k}$:

$$f(x) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k} \phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} \beta_{jk} \psi_{jk}(x), \quad j_0 \geq 0, \tag{2}$$

where the scaling $\alpha_{j_0 k}$ and the wavelet $\beta_{jk}$ coefficients are defined by:

$$\alpha_{j_0 k} = \int_0^1 f(x) \phi_{j_0 k}(x) dx, \quad \beta_{jk} = \int_0^1 f(x) \psi_{jk}(x) dx. \tag{3}$$

This representation of $f$ provides a location in both frequency (determined by $j$) and time (determined by $k$). The larger the value of $j$ the higher the frequency related to $\psi_{jk}$ and consequently the resolution.

### *Wavelet bases*

It is an essential feature of wavelets with good de-noising properties that smooth functions can be represented with a minimal set of coefficients. Historically, the first wavelet basis was the *Haar wavelet* which is defined by a simple step function: $\psi(x) = -1, 1(0 \leq x < 1/2, 1/2 \leq x < 1)$. Due to its discontinuity, the Haar basis is not suitable to

represent smooth functions with a sparse set of coefficients. More useful wavelet bases can represent high order polynomials with many wavelet coefficients of zero value. This feature is described by the number of vanishing moments of the wavelet basis. For a polynomial function $f(x)$, the coefficients $\alpha_{j_0k}$ and $\beta_{jk}$ are linear combinations of different order moments[1] of the wavelet base functions (see Equation 3). If the wavelet base function $\psi$ is chosen in a way that the order moments are zero: $\int x^k\psi(x)dx = 0$ for $k = \{0, \ldots, N\}$, the mother wavelet has $N$ vanishing moments.

*Daubechies wavelets* were defined as trigonometric polynomials which maximize the number of vanishing moments of the mother wavelet for a minimal compact support. If the length of the support of the base function is $2N$, the number of vanishing moments will be $N - 1$. Practically this means, that a polynomial of order $N - 1$ can be represented with zero value coefficients for the mother wavelet. They are consequently well suited to represent smooth signals with a sparse set of coefficients. The price for this improvement over Haar wavelets is the loss of symmetry of the base function, i.e. the wavelet transform of a mirror image of a function is not equivalent to the mirror image of the wavelet transform of the function. It has been shown that, except for the Haar system, wavelets cannot be at the same time compactly supported and symmetric (Daubechies, 1992). *Symmlet wavelets* were designed to be as close as possible to symmetry. As for Daubechies wavelets, for a width of the compact support of $2N$, the number of vanishing moments of the mother wavelet will be $N - 1$. *Coiflet wavelets* were derived from Daubechies wavelets and have in addition vanishing moments for the scaling function. For $N - 1$ vanishing moments the width of the compact support increases to $3N$.

### *Discrete wavelet transform*

A computationaly efficient implementation of the wavelet transform for digital signals is Mallat's fast discrete wavelet transform algorithm (Mallat, 1989b). The discrete wavelet transform can represented in a matrix form as:

$$d = \mathbf{W}f, \tag{4}$$

[1]The $k^{\text{th}}$ order moment of $\psi$ is defined as $\int x^k\psi(x)dx$.

were $f = \{f_1, f_2, \ldots, f_N\}'$ is the original signal represented as a column vector of $N = 2^n$ discrete data points, $d$ is a $N\times 1$ vector comprising both the discrete scaling coefficients $\alpha_{j_0k}$ and the discrete wavelet coefficients $\beta_{jk}$. $\mathbf{W}$ is a $N \times N$ orthogonal transformation matrix defined by the chosen orthonormal wavelet basis.

The connection between the discrete wavelet transform and MRA can be described by the operator representation of the quadrature mirror filters, known as the low band ($L$) and the high band ($H$) filters, which are specifically defined by the chosen orthonormal wavelet basis. If $f^{(n)}$ is the original signal (of $2^n$ data points), at each stage the wavelet decomposition moves to a coarser approximation, i.e $f^{(n-1)} = Lf^{(n)}$ and $d^{(n-1)} = Hf^{(n)}$, where $d^{(n-1)}$ is the detail lost by approximating $f^{(n)}$ by the averaged $f^{(n-1)}$. In this way the discrete wavelet decomposition of $f^{(n)}$ is represented as another sequence of length $2^n$, where the coarser approximation $f^{(n-1)}$ has only half of the original signal length. This procedure can be continued until one approximation coefficient remains. Thus the DWT (the equivalent of Equation 4) can be summarized as:

$$\begin{aligned} f \rightarrow & (Hf, HLf, HL^2f, \ldots, HL^jf, \ldots, HL^{n-1}f, H^nf) \\ & = (d^{(n-1)}, d^{(n-2)}, \ldots, d^j, \ldots, d^1, d^0, f^0), \end{aligned} \tag{5}$$

where the 'detail' sequences $d^j$ contain the wavelet coefficients $\beta_{jk}$.

*Wavelet de-noising* is based on the property of wavelets to represent signals with a set of coefficients which have desirable statistical properties in the suppression of noise (Daubechies, 1992). A substantial reduction of the noise level is achieved by applying a wavelet transform followed by a suppression of noise-related wavelet coefficients and backward wavelet transform (Figure 1). The most widely used methods to suppress noise-related coefficients are global hard- and soft-thresholding of the wavelet coefficients (Donoho and Johnstone, 1994, 1995). In the hard-thresholding procedure all coefficients below a threshold $\lambda$ are zeroed (keep or kill), while in soft-thresholding, in addition, all the other coefficients are also shrunk towards zero by subtracting $\lambda$ (shrink or kill):

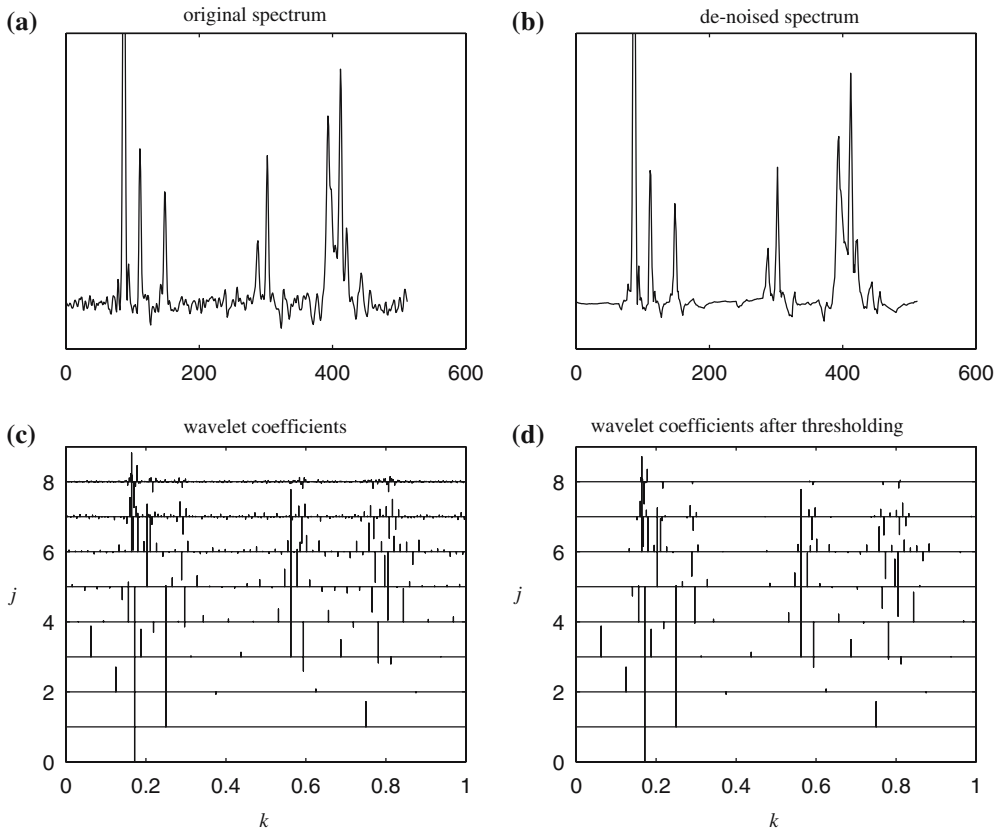


*Figure 1.* Schematic representation of wavelet de-noising: the DWT decomposes the original signal in wavelet coefficients ($k$) at different dyadic levels ($j$). Noise-related coefficients are eliminated by thresholding and the spectrum is reconstructed by an inverse wavelet transform.

$$\beta_{jk,\text{hard}} = \begin{cases} \beta_{jk} & \text{if} \quad |\beta_{jk}| > \lambda \\ 0 & \text{if} \quad |\beta_{jk}| \leq \lambda, \end{cases}$$

$$\beta_{jk,\text{soft}} = \begin{cases} \beta_{jk} + \lambda & \text{if} \quad \beta_{jk} < -\lambda \\ \beta_{jk} - \lambda & \text{if} \quad \beta_{jk} > \lambda \\ 0 & \text{if} \quad |\beta_{jk}| \leq \lambda. \end{cases} \qquad (6)$$

$\lambda$ is determined using the 'universal threshold' estimator: $\lambda = \sigma\sqrt{2\log N}$, where $\sigma$ represents the median absolute deviation of the wavelet coefficients obtained after the first wavelet decomposition step divided by an empirical factor of 0.6745 and $N$ is the total number of data points. This represents a robust procedure to estimate the noise level because the wavelet coefficients at the finest resolution level represent predominantly spectral noise. A large number of methods to estimate the wavelet coefficient threshold were compared in a review article (Antoniadis et al., 2001) some of which were also tested in this work.

The first dyadic levels ($j = \{1, 2, 3, \ldots\}$ in Equation 5) represent the low frequency components of the signal in the wavelet representation, i.e. baseline and peak shape features. Therefore the suppression of these wavelet coefficients is not desirable and a 'low-frequency cutoff' $J$ which preserves the first dyadic levels is usually applied.

### *Translation invariant (TI) wavelet transform*

Wavelet suppression using hard- or soft-thresholding can cause truncation artifacts in the vicinity of the discontinuities introduced by suppressing individual coefficients which are often called Gibbs artifacts. These artifacts can be attributed to the lack of translation invariance of

the wavelet base. A simple method to average the translation dependence is 'cycle spinning' where data is shifted, de-noised and un-shifted. Subsequently the results for different shifts are averaged. A translation invariant transformation algorithm (Coifman and Donoho, 1995) was designed for fast cycle-spinning over all $N$ points of the spectrum. In conjunction with de-noising, the TI wavelet transform has significant advantages, particularly when sharp signals in an NMR spectrum cause pronounced Gibbs artifacts.

## Materials and methods

### *Experimental NMR data*

A $^{15}$N-edited NOESY spectrum of the Sud protein from *Wolinella succinogenes* (30 kDa homodimer, 137 residues in the monomer unit) was employed to test and calibrate different wavelet de-noising protocols. A reference NOESY peak list comprising 1410 NOE signals was collected manually and was verified by interactive structure calculation and NOE assignment using the ARIA software (Lin et al., 2004). The test $^{15}$N-NOESY spectrum was recorded with 48 increments in the $^{15}$N and 180 increments in the $^{1}$H dimension. The spectrum was processed in NMRLab using sine bell window functions and zero filling to 512 points in the incremented proton and 128 points in the incremented nitrogen dimension. For the automated NOE assignment and structure calculation two additional NOESY spectra were included: a $^{13}$C-separated NOESY and a $^{13}$C-separated NOESY with the carrier frequency on the methyl protons. In addition to the NOESY data, 11 hydrogen bonds and 136 dihedral angle constraints obtained with TALOS (Cornilescu et al., 1999) were utilized during the structure calculations. The backbone chemical shift assignment was 95% complete (Lin et al., 2000) with few flexible regions missing and the side chain resonance assignment was 74% complete.

*Wavelet de-noising* is achieved by performing a DWT and applying a threshold to the wavelet coefficients. In the simplest approach, a one-dimensional (1D) DWT has been applied to each 1D strip of the multidimensional NMR spectra. Alternatively, two-dimensional (2D) DWT has been used to de-noise 2D slices of the NMR spectra. The wavelet de-noising routines used in this work were based on the WAVELAB8.02 wavelet toolbox (Buckheit and Donoho, 1995).

### *Quantification of signal-to-noise and resolution*

To evaluate the effect of wavelet de-noising on the noise level in the spectrum, on peak intensities and on automatically generated peak lists we have used four different criteria: a statistical measure of the noise level in spectra and three scores which compare the peaks picked after de-noising with the reference peak list.

(1) For each $^{1}$H–$^{1}$H slice of the NOESY spectrum a noise standard deviation $\sigma$ was estimated by taking the minimum of the standard deviations of 256 adjacent square sections of the slice and a statistical *de-noising factor* was calculated as $dfactor = \sigma^{raw}/\sigma^{wav}$.

(2) The effect of the wavelet shrinkage on the fine structure of the NMR signals was quantified by a *fine structure score* which compares the reference peak volumes ($V_{ref}$) with the corresponding volumes after wavelet de-noising ($V_{wav}$):

$$fscore = 1 - \text{mean}\left(\frac{|V_{ref} - V_{wav}|}{V_{ref}}\right). \qquad (7)$$

The peaks volumes were obtained using the numerical integration algorithm described in the Appendix.

(3) To identify signals which fall below the peak picking threshold as a consequence of the smoothing effect of the wavelet de-noising, a *peak picking score* was defined as $pscore = N_{wav}/N_{ref}$, where $N_{wav}$ is the number of real peaks automatically picked on the wavelet de-noised spectrum and $N_{ref}$ the number of peaks in the reference list. This score measures the relative amount of small signals or signal shoulders which were lost.

(4) Because the noise standard deviation $\sigma$ did not always provide a useful measure for noise suppression in the peak list, an additional *de-nosing score* which calculates the ratio of the noise-related peaks obtained before ($N_{raw}^{noise}$) and after de-noising ($N_{wav}^{noise}$) was introduced: $dscore = 1 - N_{wav}^{noise}/N_{raw}^{noise}$.

With the exception of the de-noising factor, *dfactor*, which is always larger than one, these

scores have values between zero and one where a value of one represents the ideal case of a noise-free peak list without any missing signals. A negative value of *dscore* indicates truncation artifacts (causing additional local extrema) introduced by the wavelet transform.

A *consistency check of the NOESY peak lists* was introduced to validate and partially assign NOESY cross peaks after automated peak picking using sequence-specific resonance assignments. The checking procedure is based on the following principles: (1) a NOESY cross peak is usually part of a network of connections between pairs of spin systems (network anchoring) and (2) NOESY spectra have an intrinsic symmetry (symmetry mapping). In $^{15}$N-edited NOESY spectra symmetry mapping selects pairs of NH–NH signals, whereas between $^{15}$N- and $^{13}$C-edited NOESY spectra HN–HC pairs are identified. A similar scheme was originally introduced to discriminate between multiple NOE assignments (Herrmann et al., 2002a) and later used for NOESY cross peak validation (Herrmann et al., 2002b). Here we use a combination of network anchoring and symmetry mapping for peak validation and alternatively, as a method to select NOE assignments. The individual assignment possibilities allowed by the frequency tolerance are subject to a two-pass filtering which yields a zero-or-one scoring as follows: (1) the network anchoring score is positive if at least a second non-diagonal NOESY peak between the same pair of residues is found and (2) the symmetry mapping score is positive if a symmetric partner exists, which is also anchored in its own network of NOE contacts. The last condition was introduced to minimize the amount of erroneous symmetry partners owing to the residual noise or missing chemical shifts. 'Lonely' NOESY cross peaks which do not belong to any possible network of NOE contacts and do not have any symmetry partner are rejected. To discriminate between the different assignment possibilities of a NOESY cross peak the conditions are more restrictive: an assignment is made only if it anchors the peak in a network of contacts and if it allows a symmetry related partner.

*The iterative NOE assignment and structure calculations* were carried out with ARIA (*a*mbiguous *r*estraints for *i*terative *a*ssignment) program (Linge et al., 2003). The consistency filters for the NOESY peak lists were embedded into the ARIA interface.

## Results

### *Optimal wavelet-based de-noising scheme*

Different schemes for wavelet de-noising were evaluated and compared. These included 1D and 2D DWT, where each was evaluated for several mother wavelets (Symmlet wavelet with 5, 8 and 10 vanishing moments: S5, S8 and S10; Daubechies wavelet with 2 and 10 vanishing moments: D4 and D20; Coiflet wavelet with 2 and 10 vanishing moments: C1 and C5; and Haar wavelet), different de-noising schemes (hard-, soft-, TI hard- and TI soft-thresholding) and various low-frequency cutoffs ($J = 2-5$). De-noising was always applied to the $^{1}$HN–$^{1}$H planes of the NOESY test spectrum and in the case of the 1D DWT we examined the effect of the order in which the two dimensions were de-noised ($^{1}$H/$^{1}$HN or $^{1}$HN/$^{1}$H).

Initially, 384 different de-noising protocols were applied to a 2D $^{1}$H–$^{1}$H cross section of the 3D $^{15}$N-edited NOESY spectrum (Figure 3, panel (e)) and evaluated using the scores described in Materials and methods (Figure 2). As expected, the Haar wavelet scored low regardless of the shrinkage scheme (methods: 8, 16, 24 and 32 in Figure 2) because the Haar wavelet basis is not continuous and therefore less suitable to represent smooth functions. Wavelets with good smoothing properties which were designed to minimize the wavelet coefficients for smooth functions, such as Symmlet and Daubechies wavelets, represent a good compromise between noise reduction and the preservation of the fine structure (methods: 1–5, 9–13, 17–21 and 25–29 in Figure 2). Compared to the 1D DWT the 2D decomposition is computationally more efficient, however the overall scores are inferior (Figure 2: blue spots). The decomposition order for the 1D DWT within the 2D data matrix has little influence although slightly better scores were obtained when the incremented proton dimension was de-noised first ($^{1}$H/$^{1}$HN, Figure 2: red symbols). Soft-thresholding yields the best possible noise suppression (large *dfactor* and *dscore*) at the expense of fine structure (low *fscore*) and completeness of the peak list (low *pscore*). In contrast, hard-thresholding preserves the fine structure at a modest gain of signal-to-noise. TI de-noising proved superior in all scores because

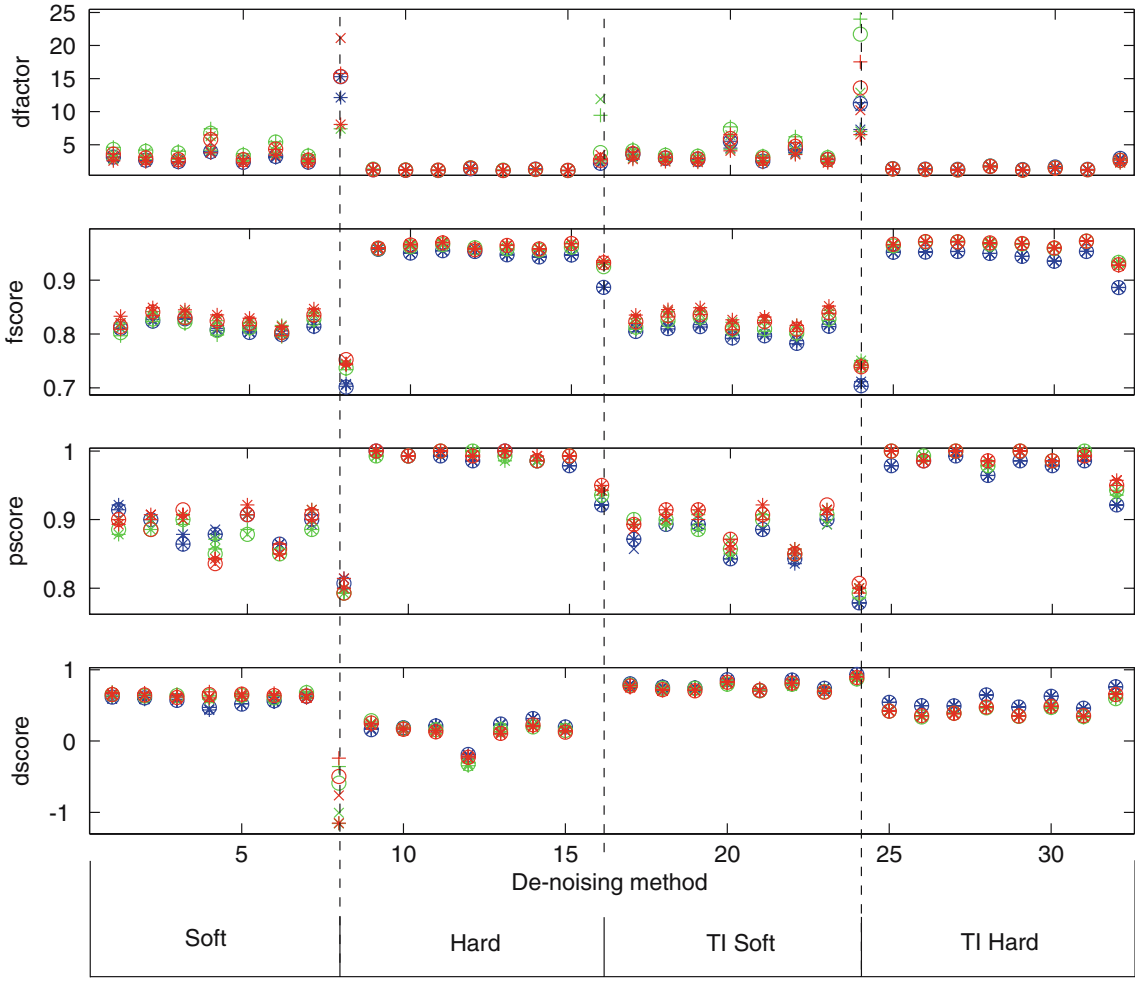


*Figure 2.* Scores for 384 de-noising protocols using a test plane of a 3D $^{15}$N-edited NOESY spectrum of the Sud protein. The four sections separated by dashed lines correspond to soft-, hard-, TI soft- and TI hard-thresholding (methods: 1–8, 9–16, 17–24 and 25–32). For each section the following wavelet bases were used: S5, S8, S10, D2, D20, C1, C5 and Haar (in the specified order). Red and green colors represents the 1D DWT de-noising order $^{1}$H/$^{1}$HN and$^{1}$HN/$^{1}$H, respectively whereas blue represents the 2D DWT. The low-frequency cutoffs are represented by the symbols + ($J=2$); ○ ($J=3$); × ($J=4$); and * ($J=5$).

it eliminates truncation artifacts and averages residual noise (de-noising methods 16–32 in Figure 2). The low-frequency cutoff ($J$) had little influence in combination with hard-thresholding schemes, presumably owing to the large wavelet coefficients of the peaks (intense singularities) compared to the baseline areas. For soft-thresholding, low values of $J$ caused smoothing because all wavelet coefficients were shrunk regardless of their absolute value. As a general result, for signals shorter than 2500 digital points, a low-frequency cutoff of three ($J=3$) represents a good compromise between signal-to-noise and resolution.

In a second step we applied the S5 (Symmlet 5) and D4 (Daubechies 4) wavelet de-nosing protocols to the full 3D spectrum (Table 1). The analysis confirmed the previous result of two possible de-noising strategies which yield either strong de-noising or high preservation of fine structure, respectively. Soft-thresholding yielded a high signal-to-noise ratio (*dfactor* = 2.7–5.4) but suppressed the low intensity signals (*pscore* = 0.77–0.83), whereas hard-thresholding preserved the fine structure (*pscore* = 0.92–0.94) on the expense of the signal-to-noise gain (*d factor* = 1.2–1.8). Best results were obtained when the 1D DWT was used in combination with the TI de-noising.

*Table 1.* Scores for different de-noising procedures[a] applied on a $^{15}$N-edited NOESY

| De-noising method | 1D DWT | | | | | 2D DWT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *dfactor* | *fscore* | *pscore* | *dscore* | CPU[b] time (s) | *dfactor* | *fscore* | *pscore* | *dscore* | CPU[b] time (s) |
| S5 soft | 3.254 ± 0.298 | 0.850 | 0.812 | 0.522 | 1437.3 | 2.673 ± 0.260 | 0.829 | 0.813 | 0.517 | 517.1 |
| S5 TI soft | 2.850 ± 0.262 | 0.857 | 0.828 | 0.576 | 5711.8 | 2.943 ± 0.298 | 0.829 | 0.806 | 0.609 | 8952.0 |
| D4 soft | 5.398 ± 0.651 | 0.848 | 0.805 | 0.525 | 1401.2 | 4.045 ± 0.556 | 0.832 | 0.798 | 0.532 | 519.2 |
| D4 TI soft | 4.277 ± 0.491 | 0.849 | 0.799 | 0.611 | 5723.1 | 4.378 ± 0.533 | 0.823 | 0.774 | 0.655 | 8802.8 |
| S5 hard | 1.242 ± 0.056 | 0.973 | 0.945 | 0.134 | 1410.7 | 1.232 ± 0.056 | 0.960 | 0.916 | 0.183 | 505.2 |
| S5 TI hard | 1.335 ± 0.068 | 0.975 | 0.943 | 0.264 | 5663.2 | 1.400 ± 0.083 | 0.957 | 0.923 | 0.354 | 8880.3 |
| D4 hard | 1.621 ± 0.180 | 0.968 | 0.942 | −0.401 | 1412.5 | 1.473 ± 0.130 | 0.960 | 0.932 | −0.447 | 509.4 |
| D4 TI hard | 1.800 ± 0.194 | 0.978 | 0.933 | 0.232 | 5602.0 | 1.807 ± 0.155 | 0.964 | 0.903 | 0.426 | 8717.2 |

[a]The low-frequency cutoff $J$ was set to a value of 3 for all de-noising procedures.
[b]CPU time required on a 1.5 GHz AMD processor for wavelet de-noising. For 1D DWT the incremented proton dimension was de-noised first.

*Table 2.* Quality scores after NOESY peak list validation using the network anchoring and symmetry mapping filters

| De-noising method | 1D DWT | | | 2D DWT | | |
|---|---|---|---|---|---|---|
| | *fscore* | *pscore* | *dscore* | *fscore* | *pscore* | *dscore* |
| None[a] | 1 | 0.961 | 0.709 | – | – | – |
| S5 soft | 0.851 | 0.791 | 0.845 | 0.831 | 0.731 | 0.864 |
| S5 TI soft | 0.858 | 0.806 | 0.868 | 0.830 | 0.783 | 0.874 |
| D4 soft | 0.848 | 0.774 | 0.852 | 0.832 | 0.722 | 0.881 |
| D4 TI soft | 0.850 | 0.777 | 0.891 | 0.829 | 0.752 | 0.901 |
| S5 hard | 0.973 | 0.920 | 0.727 | 0.960 | 0.894 | 0.737 |
| S5 TI hard | 0.975 | 0.918 | 0.771 | 0.957 | 0.899 | 0.799 |
| D4 hard | 0.969 | 0.913 | 0.586 | 0.961 | 0.908 | 0.559 |
| D4 TI hard | 0.978 | 0.903 | 0.775 | 0.964 | 0.878 | 0.819 |

[a]Automated picked peaks using the original spectrum.

For soft-thresholding best scores were obtained with Daubechies wavelets while the Symmlet basis scored better with hard-thresholding. No further improvement could be found with more sophisticated thresholding schemes (data not shown).

### *NOESY peak list validation*

By incorporating the validation filters based on network anchoring and symmetry mapping all de-noising scores were further improved with a minimal loss of peaks. This is reflected by a larger de-noising score (*dscore*) and minimally smaller peak picking scores (*pscore*) (see Table 2). Limitations for this validation scheme are excessively noisy peak lists, incomplete assignment tables, shifted peaks or tight frequency tolerances. Furthermore, unique contacts between amino acids of different structural elements of proteins with high information content may be eliminated. When the validation filters were applied without prior wavelet de-noising the quality scores indicate that 4% of the real peaks were eliminated while 70% of the noisy entries were removed. However, by combining wavelet de-noising and validation filters up to 90% of the residual noise was removed while only 2% additional real peaks were eliminated.

### *Iterative NOE assignment and structure calculation using wavelet de-noised spectra*

The two de-noising strategies derived in this analysis have complementary features for automated NOE assignment strategies. The first de-noising scheme employing soft-thresholding

(1D-DWT-D4-TI-Soft) yields a peak lists which is approximately 80% complete and 60% de-noised (list *(i)*). The second de-noising scheme which uses hard-thresholding of the wavelet coefficients (1D-DWT-S5-TI-Hard) provides a peak list which is 95% complete and 25% de-noised (list *(ii)*). Combined with NOESY peak list validation the peak lists were 75% complete and 90% de-noised *(i)* or 90% complete and 75% de-noised *(ii)*, respectively. Automated iterative NOE assignment and structure calculation can take advantage of the complementary features of the two schemes if the two peak lists are employed incrementally. In a first stage only the best and most reliable peak list *(i)* is used while peak list *(ii)* with modest noise suppression and a large number of signals can be introduced in a later stage when a structural model is already available.

This strategy was validated using the experimental NOESY data of the Sud dimer for which a high resolution solution structure was previously reported (Lin et al., 2004). To simplify the assignment procedure the NOE assignment and structure calculations were carried out only for the monomer unit (residues 20–130). The N-terminal $\alpha$-helix was not considered since its positioning is essentially determined by the dimer fold. The monomer reference structure was recalculated using the intra-monomer distance constraints originating from the $^{15}N$,$^{13}C$ and methyl-$^{13}C$ edited-NOESY spectra.

### *Three stage NOE assignment and structure calculation protocol*

The *first stage* of iterative NOE assignment and structure calculation started with the 'cleanest' NOESY peak list *(i)* and five iterations in ARIA. In this stage 2117 NOEs were collected from the three heteronuclear NOESY spectra. Besides validation of NOESY peaks, the network anchoring and symmetry mapping filters allowed 562 unambiguous NOE assignments. The coupled NOE assignment and structure calculation protocol followed the standard ARIA scheme (Linge et al., 2001) of the first five iterations. To take advantage of the clean but incomplete peak list *(i)* and to minimize the amount of peaks that may be incompatible with the transient 3D models owing to underestimated upper limits, the *qmove* flag of the violation analysis module in ARIA was used throughout these initial five iterations[2]. In each iteration 30 structures were calculated and the ten models with lowest energy were used to interpret the spectra in the following cycle. The ambiguity cutoff in ARIA[3] was gradually decreased from 1 to 0.98. At this stage a bundle of conformers with a mean backbone RMSD of $4.68 \pm 1.08$ Å between the best ten models was obtained. The RMSD between the average structure and the reference model was 2.64 Å (Figure 3, panel (b)).

In the *second stage* these models were used as a starting point for a new cycle of four ARIA iterations using the peak list *(ii)* and after the anchoring/symmetry based validation (2615 NOEs). The protocol was identical with the one employed in the first part of the run but no initial assignments were imposed. In this way all assignment possibilities were reassessed based on the previously calculated structural models. After four iterations a bundle of conformers with a mean backbone RMSD of $2.00 \pm 0.36$ Å and a deviation between the average and the reference structure of 1.72 Å was achieved (Figure 3, panel (d)). Despite a high ambiguity cutoff for the NOE assignments (0.98) which allows for a large number of ambiguous distance restraints, the calculation converged to a well defined model. The sparseness of the cross peak list in this stage does not represent a drastic limitation because NOESY-based structure calculations are tolerant with respect to the data incompleteness (Jee and Güntert, 2003).

In the *third stage* the previously calculated models were used to interpret the peak lists obtained by automated peak picking performed on the original data (approximately 3500 assignable peaks). Four cycles of ARIA (iterations 5–8) were carried out imposing strict violation tolerances (1.0–0.1 Å) and spin diffusion correction. The ambiguity cutoff was gradually decreased from 0.96 to 0.8. It is important to use the original spectra for the final NOE assignment and structure calculation because the most informative long-distance NOE signals may have very low intensities and can be suppressed even with the most conservative de-noising schemes. After a final ARIA

[2]The *qmove* feature moves the upper limit for each systematically violated restraint to 6 Å, repeats the violation analysis and rejects only the remaining violated restraints.

[3]The number of assignment possibilities which are ranked and taken into account based on the previously calculated structures.

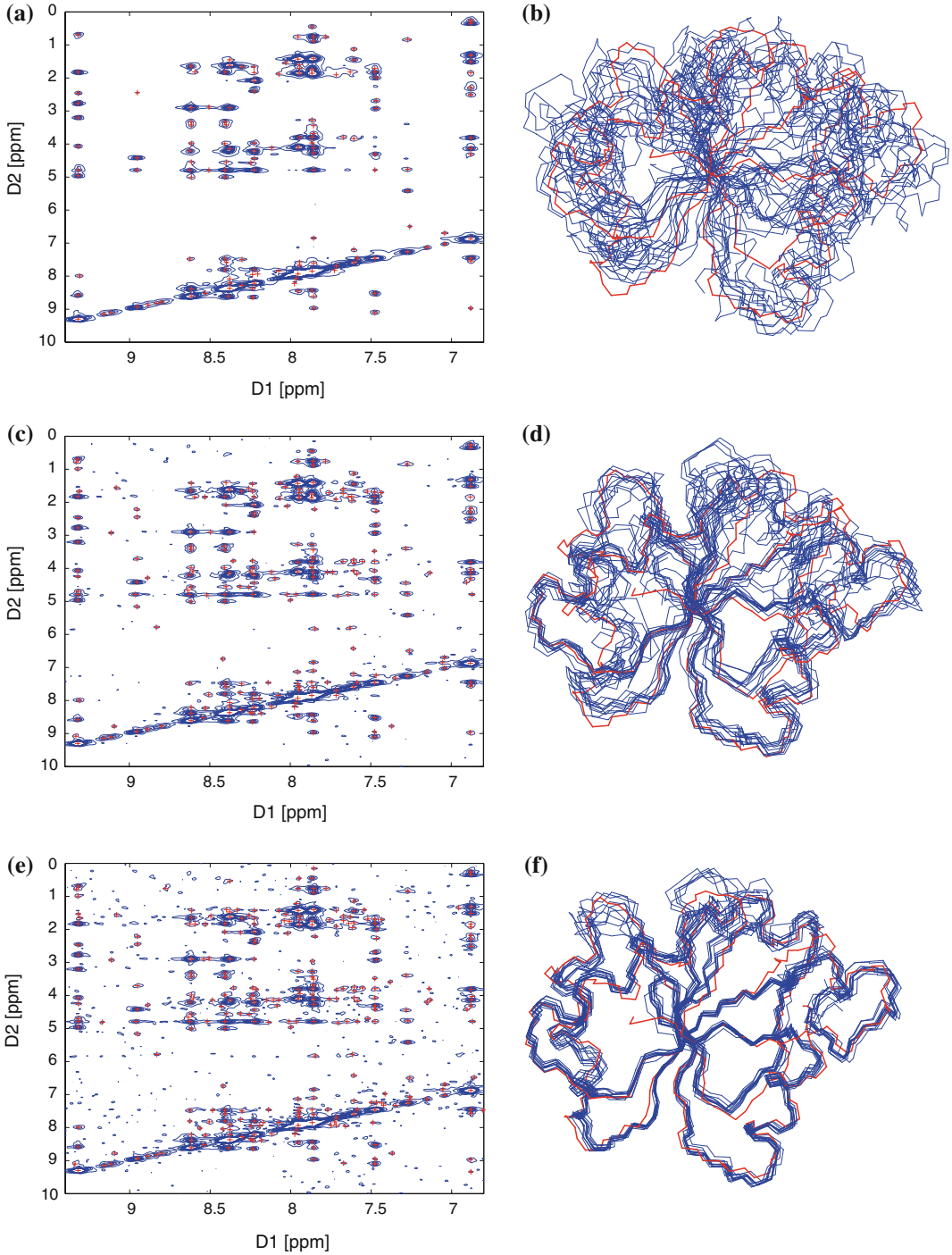


*Figure 3.* (a), (c) and (e) represent a 2D slice of the 3D $^{15}$N-edited NOESY spectrum of the Sud protein from *Wolinella succinogens*: (a) after 1D-DWT-D4-TI-Soft de-nosing, (c) after 1D-DWT-S5-TI-Hard de-noising and (e) the original cross section. Red crosses depict the automatically picked peaks for each spectrum. (b), (d) and (f) show backbone plots of the reference structure (red) together with the ten best conformers (blue) obtained in subsequent stages of automated NOE assignment and structure calculation using NOESY spectra (a), (c) and (e), respectively.

iterative structure calculation 1923 non-redundant NOEs were assigned leading to a bundle of the ten best conformers with a mean backbone RMSD of $0.85 \pm 0.2$ Å (Figure 3, panel (f)).

An identical structure calculation protocol was applied using the distance restraints previously obtained by manual peak picking. The automated and manual schemes gave similar target functions and almost identical RMSD values. The backbone RMSD between the mean structures of the two bundles (automatic versus manual) was 1.06 Å. Additionally, a control run of iterative NOE assignment and structure calculations with ARIA using the raw peak lists obtained by automated peak picking on the original spectra (same input data as for *stage 3*) was performed. In this case, both the target function value and the RMSD of the NMR ensemble were considerably higher indicating a poorer convergence. The accuracy given by the RMSD deviation from the reference structure increased to 1.59 Å. Table 3 presents the structural statistics summary of the three-stage automated NOE assignment and structure calculation compared to the corresponding values for the interactive manual approach and for the standard ARIA protocol using peak lists generated by automated peak picking without wavelet de-noising.

## Discussion

The difficulty of *de novo* protein structure calculation using iterative NOE assignment strategies is to distinguish between multiple assignment possibilities of the NOESY cross peaks in the presence of different types of noise. The most direct type of noise is spectral noise arising from the NMR hardware. Although this has been substantially reduced by the introduction of cryogenically cooled probes there is always remaining noise, especially as NMR spectroscopists use proteins at lower concentrations. In addition, there is noise in the peak lists after peak picking, typically arising from artifacts or chemical shift ambiguities in the spectrum. The method described in this work takes advantage of direct spectral noise to determine de-noised peak lists at different levels of reliability. Clearly, this method is limited to noise present in the data and will fail for perfect spectra.

The analysis of many different wavelet de-noising schemes applied to a sample NOESY spectrum showed that no single wavelet de-noising strategy yields a perfect peak list. High levels of de-noising are usually associated with smoothing effects causing suppression of low intensity signals and signal shoulders. However, the special features of different de-noised peak lists provide complementary information which facilitate a combination of automated peak picking, NOE assignment and structure calculations employing the ambiguous distance restraints (ADR) concept (Nilges, 1995) in ARIA.

ADR-based structure calculations suffer from additional local minima introduced in the NOE hybrid energy function by incorrect assignment possibilities which lead to a more demanding minimization problem. To simplify the landscape of the NOE potential surface and to reduce the effect of spectral artifacts additional filters based on the chemical shift assignments and the intrinsic properties of the NOESY spectra (network anchoring, symmetry mapping, restraint combination and Gaussian frequency windows) were previously introduced (Herrmann et al., 2002a, b). However, for these sophisticated filtering strategies almost complete chemical shift assignments and clean NOESY cross peak lists are required (Günter, 2003; Jee and Güntert, 2003).

*Table 3.* Structural statistics for the three stages of automated NOE assignment and structure calculation; comparison with the results of the interactive manual approach and standard ARIA protocol

| | Stage 1 | Stage 2 | Stage 3 | Manual | Standard ARIA |
|---|---|---|---|---|---|
| NOE cross peaks | 2117 | 2615 | 3507 | 2700 | 3507 |
| NOE distance restraints[a] | 1615 | 1965 | 1923 | 1896 | 1901 |
| Target function (kcal/mol) | $2215.1 \pm 417.3$ | $944.3 \pm 309.1$ | $132.9 \pm 7.0$ | $110.6 \pm 3.4$ | $215.1 \pm 16.2$ |
| Backbone RMSD (Å)[b] | $4.68 \pm 1.08$ | $2.00 \pm 0.36$ | $0.85 \pm 0.20$ | $0.84 \pm 0.10$ | $1.16 \pm 0.21$ |
| | 2.64 | 1.72 | 1.06 | | 1.59 |

[a]Unambiguous and ambiguous (ADR) distance restraints.
[b]First row denotes the mean backbone RMSD of energetically best ten models, the second row the RMSD between the ensemble average structure and the reference model. For all RMSD calculations residues 21–89 and 95–129 were used.

The strategy presented here combines filters which use the intrinsic logic of the peak list (symmetry mapping and network anchoring) with wavelet de-nosing which reduces the spectral noise independent of any specific features of the peak list. Different stages of de-noising complement the requirements of the ADR algorithm by providing a highly reliable but incomplete peak list in a first stage followed by a less stringently de-noised but almost complete peak list in a second stage of combined assignment and structure calculation. This strategy is less prone to move into local minima than other concepts which emphasize filters relying on the internal logics of the peak list.

The advantages of the de-noising strategy will be most significant for somewhat noisy NOESY spectra. Wavelet de-noising is computationally efficient, in fact commonly used DWT algorithms are faster than the Fast Fourier Transformation. Therefore de-noising and peak picking require little additional computational time to obtain peak lists for different stages of the procedure. The combined software tools provide wavelet de-noising, peak picking and integration with export modules to different file formats. Therefore this software will be commonly applicable with different programs for combined NOESY assignment and structure calculation. The symmetry and network anchoring filters were directly incorporated into the ARIA software. An initial version of the software is available from the authors.

## Acknowledgements

This work was supported by the Large Scale Facility Frankfurt (UNIFRANMR) and by the RTD project FIND from the European Community.

**Electronic supplementary material** is available at http://dx.doi.org/10.1007/s10858-004-3093-1.

## Appendix

*Peak picking and peak integration*

A robust numerical procedure for automated peak picking and peak integration of the multidimensional NMR spectra was developed as an integrated tool in this project. This peak picking procedure consists of four distinct steps which will be described for a paradigmatic 2D dataset.

(1) To overcome distortions from non-uniform noise distributions and noise bands (water line, diagonal and $T_1$-noise bands) the spectral local background noise levels were determined as described previously (Koradi et al., 1998). For each one-dimensional strip of the spectrum a noise standard deviation $\sigma$ was calculated by taking the minimum of the standard deviations for 16 consecutive sections of the strip. The local background noise level of a point $P$ of coordinates ($i_1$, $i_2$, ..., $i_n$), belonging to a $n$-dimensional NMR spectrum, is calculated according to:

$$bnoise(P_i) = F \cdot \sqrt{\sum_{\mathrm{dim}=1}^{n} \sigma^2_{\mathrm{dim},i_{\mathrm{dim}}} - (n-1) \cdot \min_{\mathrm{dim},i}(\sigma_{\mathrm{dim},i})^2}, \quad \text{(A1)}$$

where $F$ is an empirical user-adjustable factor (between 2 and 5).

(2) In a second step, the spectrum was segmented into regions of points with the absolute value of the intensity larger than the local noise levels $bnoise(P)$ (Figure A1: blue crosses). Because the standard deviation of the signal after de-noising is not a suitable descriptor for noise levels, the $bnoise(P)$ values obtained for the native spectra were also used for the segmentation of the de-noised spectra.

(3) The local extrema (maxima or minima, depending on the peak signs) were determined by a grid search using the sparse matrix obtained after segmentation. In our implementation the width of the grid cell can be adjusted by the user according to the digital resolution of the dataset; in this work the smallest possible grid cell size of $3\times3$ points was used. A peak list containing the coordinates of all the local extrema above the local noise levels is obtained.

(4) An algorithm for *digital peak integration* which can separate overlapping signals (even if those have very different sizes) was designed. This algorithm first defines an initial integration box around each local maximum[4] (Figure A1:

[4]For negative signals a positive mirror image of the initial integration box is computed prior to integration.

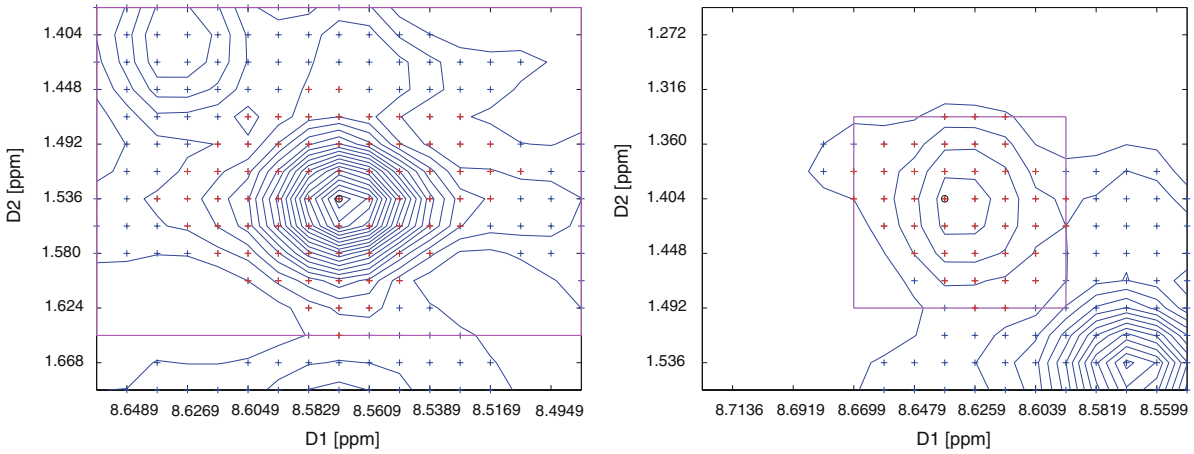


*Figure A1.* Two examples for peak integration in the presence of the spectral overlap. The plots represent the initial integration boxes, blue crosses depicts the digits with an intensity larger than the estimated local noise levels, magenta squares are the refined integration boxes and the red crosses the actual points which are found to be part of the peak subject to integration.

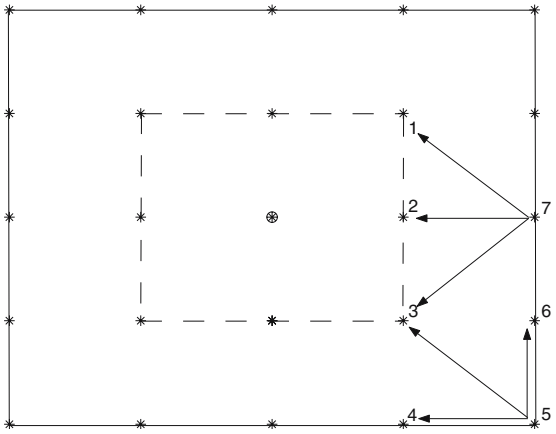


*Figure A2.* Object-related growing algorithm used for the peak integration: the dashed and solid lines represent the first and second shells which define the peak limits around the central local maximum, respectively. Point 7 is considered to be a part of the peak if the first order neighbor (point 2) has a higher intensity and if the two second order neighbors (points 1 and 2) have intensities above the noise threshold. For corners a variation of this definition is used: the first order neighbor of point 5 is point 3 and the second order neighbors are points 6 and 4.

full boxes). Its rational size is determined by a rectangular local minimum search starting from the central maximum along the Cartesian dimensions of the spectrum which stops either if the background noise level is reached or if a local minimum is encountered (Figure A1: magenta boxes). Within the refined rectangular integration box the peak shape is resolved by an object-related growing algorithm around the local maximum (Figure A1: red crosses) which iteratively adds one square shell centered on the central maximum (Figure A2: continuous line) until the end of the integration box is reached in each dimension. A point of the new shell (Figure A2: point 7) is added to the peak if the first order neighbor (Figure A2: point 2) has a higher intensity in the previous layer and if the second order neighbors have intensities above the local noise levels (points 1 and 3 in Figure A2). For corners (Figure A2: point 5) the condition is slightly different, the first order neighbor is considered to be the edge of the previous shell (Figure A2: point 3) and the second order neighbors are located within the same layer (points 4 and 6 in Figure A2).

Using this algorithm all data points which are part of a given peak can be determined, even in the presence of strong chemical shift degeneracies without any *a priori* assumptions about the shape of the signals. The peak integrals are calculated by adding the data points determined in (1)–(4). In our implementation this integrator also provides the matrices describing the peak shapes for further statistical multivariate or Bayesian analysis (Grahn et al., 1989; Schulte et al., 1997).

## References

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001) *J. Stat. Software*, **6**, 1–83.

Banci, L., Bertini, I., Savellini, G., Romagnoli, A., Turano, P., Cremonini, M., Luchinat, C. and Gray, H. (1997) *Proteins*, **29**, 68–76.

Buckheit, J. and Donoho, D. (1995) *Wavelet and Statistics*, chapter Wavelab and Reproducible Research. Springer, Berlin, pp. 53–81.

Cancino-De-Greiff, H.F., Ramos-Garcia, R. and Lorenzo-Ginori, J.V. (2002) *Concepts Magn. Reson.*, **14**, 388–401.

Coifman, R.R. and Donoho, D.L. (1995) *Wavelet and Statistics*, chapter Translation-Invariant De-noising. Springer, Berlin, pp. 103–125.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.

Daubechies, I. (1992) *Ten Lectures on Wavelets* SIAM, Philadelphia.

Donoho, D. and Johnstone, I. (1994) *Biometrika*, **81**, 425–455.

Donoho, D. and Johstone, I. (1995) *J. Am. Stat. Assoc.*, **90**, 1200–1224.

Grahn, H., Edlund, U., van den Hoogen, Y., Altona, C., Delaglio, F., Roggenbuck, M. and Borer, P. (1989) *J. Biomol. Struct. Dyn.*, **6**, 1135–1150.

Günter, P. (2003) *Prog. Nucl. Magn. Res. Spectrosc.*, **43**, 105–125.

Günther, U., Ludwig, C. and Rüterjans, H. (2000) *J. Magn. Reson.*, **145**, 201–208.

Günther, U., Ludwig, C. and Rüterjans, H. (2002) *J. Magn. Reson.*, **156**, 19–25.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002a) *J. Mol. Biol.*, **319**, 209–227.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002b) *J. Biomol. NMR*, **24**, 171–189.

Hoch, J.C. and Stern, A.S. (1996) *NMR Data Processing*, chapter Emerging Methods. Wiley-Liss, New York, pp. 136–144.

Huang, Y., Swapna, G., Rajan, P., Ke, H., Xia, B., Shukla, K., Inouye, M. and Montelione, G. (2003) *J. Mol. Biol.*, **327**, 521–536.

Jee, J. and Güntert, P. (2003) *J. Struct. Funct. Genomics*, **4**, 179–189.

Koradi, R., Billeter, M., Engeli, M., Güntert, P. and Wüthrich, K. (1998) *J. Magn. Reson.*, **135**, 288–297.

Lin, Y., Dancea, F., Löhr, F., Klimmek, O., Pfeiffer-Marek, S., Nilges, M., Wienk, H., Kröger, A. and Rüterjans, H. (2004) *Biochemistry*, **43**, 1418–1424.

Lin, Y., Pfeiffer, S., Löhr, F., Klimmek, O. and Rüterjans, H. (2000) *J. Biomol. NMR*, **18**, 285–286.

Linge, J., Habeck, M., Rieping, W. and Nilges, M. (2003) *Bioinformatics*, **19**, 315–316.

Linge, J., O'Donoghue, S. and Nilges, M. (2001) *Methods Enzymol.*, **339**, 71–90.

Mallat, S. (1989a) *Trans. Am. Math. Soc.*, **315**, 69–87.

Mallat, S. (1989b) *IEEE Trans. Pattern Anal. Machine Intell.*, **11**, 674–693.

Mittermayr, C., Nikolov, S., Hutter, H. and Grasserbauer, M. (1996) *Chemomet. Intell. Lab. Syst.*, **34**, 187–202.

Mumenthaler, C. and Braun, W. (1995) *J. Mol. Biol.*, **254**, 465–480.

Mumenthaler, C., Güntert, P., Braun, W. and Wüthrich, K. (1997) *J. Biomol. NMR*, **10**, 351–362.

Nilges, M. (1995) *J. Mol. Biol.*, **245**, 645–660.

Nilges, M., Macias, M., O'Donoghue, S. and Oschkinat, H. (1997) *J. Mol. Biol.*, **269**, 408–422.

Savarin, P., Zinn-Justin, S. and Gilquin, B. (2001) *J. Biomol. NMR*, **19**, 49–62.

Schulte, A., Gorler, A., Antz, C., Neidig, K. and Kalbitzer, H. (1997) *J. Magn. Reson.*, **129**, 165–172.

Shao, X.-G., Kai-Man Leung, A. and Chau, F.-T. (2003) *Acc. Chem. Res.*, **36**, 276–283.

Tjandra, N. and Bax, A. (1997) *Science*, **278**, 1111–1114.

Tjandra, N., Garrett, D., Gronenborn, A., Bax, A. and Clore, G. (1997) *Nat. Struct. Biol.*, **4**, 443–449.

Trbovic, N., Dancea, F., Langer, T. and Günther, U. (2005) *J. Magn. Reson.*, **173**, 280–287.