2010-06

Automated solvent artifact removal and base plane correction of multidimensional NMR protein spectra by AUREMOL-SSA

Journal of Biomolecular NMR,Dordrecht : Springer Netherlands,v. 47, n. 2, p. 101-111, June 2010
http://www.producao.usp.br/handle/BDPI/49698

J Biomol NMR (2010) 47:101–111
DOI 10.1007/s10858-010-9414-z

ARTICLE

Automated solvent artifact removal and base plane correction of multidimensional NMR protein spectra by AUREMOL-SSA

Wilhelm M. Malloni · Silvia De Sanctis · Ana M. Tomé · Elmar W. Lang · Claudia E. Munte · Klaus Peter Neidig · Hans Robert Kalbitzer

Received: 22 February 2010 / Accepted: 22 March 2010 / Published online: 23 April 2010

Abstract Strong solvent signals lead to a disappearance of weak protein signals close to the solvent resonance frequency and to base plane variations all over the spectrum. AUREMOL-SSA provides an automated approach for solvent artifact removal from multidimensional NMR protein spectra. Its core algorithm is based on singular spectrum analysis (SSA) in the time domain and is combined with an automated base plane correction in the frequency domain. The performance of the method has been tested on synthetic and experimental spectra including two-dimensional NOESY and TOCSY spectra and a three-dimensional $^{1}H,^{13}C$-HCCH-TOCSY spectrum. It can also be applied to frequency domain spectra since an optional inverse Fourier transformation is included in the algorithm.

Keywords AUREMOL-SSA · Multidimensional NMR · Singular spectrum analysis · Solvent suppression · Base plane correction

Abbreviations

ALS	Automated linear spline
COSY	Correlation spectroscopy
FID	Free induction decay
FIR	Finite impulse response filter
HPr	Histidine containing phosphocarrier protein
ICA	Independent component analysis
KLT	Karhunen–Loeve transformation
PCA	Principal component analysis
SSA	Singular spectrum analysis
SVD	Singular value decomposition
Trx	Thioredoxin

W. M. Malloni · S. De Sanctis · E. W. Lang · C. E. Munte · H. R. Kalbitzer (✉)
Institute of Biophysics and Physical Biochemistry, University of Regensburg, 93040 Regensburg, Germany
e-mail: hans-robert.kalbitzer@biologie.uni-regensburg.de

A. M. Tomé
Department Electrical Engineering, Telecommunications and Informatics, University of Aveiro, 3100 Aveiro, Portugal

K. P. Neidig
BioSpin GmbH, Software Department, Silberstreifen 4, 76287 Rheinstetten, Germany

Introduction

NMR investigations of biomolecules are generally performed in aqueous solutions. Thus when studying the proton resonances, the dominant signal stems from the solvent and is often many orders of magnitude larger than the resonances of the molecules under consideration. The dominant solvent artifact is severely affecting the process of data evaluation, especially when automation is required. Suppressing this artifact signal, either by experimental means through proper pulse sequences or by post-processing methods using signal processing techniques, is thus a key issue in proton NMR [for a review see (Gronwald and Kalbitzer 2004)]. Early approaches make use of the fact that the water resonance is usually positioned at the center of the spectrum (i.e. at $\omega = 0$). As a consequence, its time domain signal can be described as a non-modulated exponential. This led Kuroda et al. (1989) to propose the computation of second derivatives of the FIDs and the application of a Fourier transformation to the latter, which suppresses signals at $\omega = 0$, such as the water artifact. An improvement to this filtering technique was later proposed by (Marion et al. 1989) applying a low-pass finite impulse

response filter (FIR). The contribution of the water signal can then be obtained by filtering out the oscillatory parts of the FIDs and then subtracting those parts from the original time domain data (Mitschang et al. 1990). This method is similar to the diagonal peak suppression method for phase-sensitive COSY spectra (Friedrichs et al. 1991). However, such linear phase filters always introduce severe signal distortions due to the damped nature of the time domain signals. For a critical review of such filtering approaches see (Coron et al. 2001). The authors compare several methods and conclude that the maximum phase FIR filter (Sundin et al. 1999) is the most accurate and efficient of these filtering techniques. The main disadvantage of FIR filtering techniques, however, is their limited use with unsuppressed water signals which cover a huge dynamic range and need attenuations up to −100 dB which conflicts with the practically available length of the filters beneath others.

Also wavelet transformation can be used for solvent suppression discarding the components corresponding to low frequencies before data reconstruction (Barache et al. 1997; Antoine et al. 2000; Guenther et al. 2002).

In the frequency domain, the dispersive tails of the water resonance can be largely attenuated by fitting these tails to a hyperbolic function which is then subtracted from the spectra (Adler and Wagner 1991). The dispersive tails of the water resonance can also be suppressed by phasing the water signal in absorption mode, zeroing the relatively small absorption signal in the frequency domain data, discarding the imaginary part and regenerating the signal from the processed real part via a Hilbert transformation (Tsang et al. 1990).

Matrix decomposition techniques like principal component analysis (PCA), singular value decomposition (SVD) and independent component analysis (ICA), follow similar ideas and form an important class of solvent suppression postprocessing methods. Application of the Karhunen–Loeve transformation (KLT) to multidimensional data removes undesired water artifacts based on their large intensity (Mitschang et al. 1991). Grahn et al. (1988) described the use of PCA for pattern recognition in two-dimensional NMR spectra by approximating the multivariate data matrix of the spectrum. The PCA method was afterwards applied by Hardy and Rinaldi (1990) for artifact reduction in COSY spectra using intensity matrixes. A related unsupervised approach is to apply singular value decomposition for large artifact removal and noise reduction on 2D NMR spectra as discussed by Brown and Campbell (1990) and by Pijnapple et al. (1992).

Related to PCA or SVD techniques are matrix pencil techniques which determine the eigenvectors and eigenvalues of a pair of time delayed correlation matrices (Lin et al. 1997). Recently these methods have been reconsidered using time-embedding techniques and simultaneous or joint diagonalization of a set of Toeplitz trajectory matrices (Parra and Sajda 2003). They belong to the class of unsupervised projective subspace techniques which decompose the signal into underlying component signals, some of which are related with the water resonance and are deliberately neglected during reconstruction. Note that during the PCA step of these methods, signal de-noising can be achieved by neglecting the eigenvectors related with the smallest eigenvalues (Gruber et al. 2006). These blind source separation techniques have been applied to 2D NOESY proton NMR spectra of proteins to remove the water resonance and any related artifacts (Stadlthanner et al. 2006). The tedious task of assigning components to the water signal has been fully automated (Boehm et al. 2006; Stadlthanner 2007).

In summary, many experimental and numerical techniques achieving water suppression have been developed so far (Hore 1989), but most of them cannot recover the solute resonances hidden underneath the water artifact. Furthermore automation requires that no additional parameters have to be set by the user. In this paper, a method based on singular spectrum analysis (SSA) (Ghil et al. 2002) for the removal of the solvent artifact and the recovery of hidden solute resonances is described. This technique is an extension of the PCA applied to a time-lagged data set. Whereas the application of the Karhunen-Loeve transformation reduces to the creation of an autocorrelation matrix by time averaging over a sample of free induction decays, the SSA embeds each FID separately in an M-dimensional vector space considering a matrix of M lagged copies of the single time series. A related method was previously reported by Zhu et al. (1997) in which a singular value decomposition was applied to the trajectory matrix derived from the data. The interrelationships among SVD, PCA and KLT have been discussed by Gerbrands (1981) pointing out that they can differ significantly.

Materials and methods

Test data sets

The two-dimensional experimental NOESY and TOCSY spectra were recorded from a sample containing 2.7 mM uniformly ^{15}N-enriched HPr protein from *Staphylococcus aureus* in 500 μL 95% H_2O/5% D_2O, pH 7.0. The three-dimensional experimental ^{1}H,^{13}C HCCH-TOCSY spectrum has been recorded from a sample containing Thioredoxin protein (Trx) from *Plasmodium falciparum* in D_2O 99.5%, pH 7.0. Trx is a medium size protein with 104 residues and formed by four α-helices and five stranded β-sheets (Munte

et al. 2009). The NMR spectra were recorded on Bruker Avance-800 and Avance-600 spectrometers operating at 800 MHz and 600 MHz, respectively, employing a mixing time of 100 ms in both two-dimensional spectra and a mixing time of 12 ms in the three-dimensional case. The water signal was reduced by selective pre-saturation in the NOESY and in the TOCSY spectra, whereas no pre-saturation was needed for the three-dimensional spectrum since measured in D_2O. The two-dimensional TOSCY and NOESY spectra have been recorded using relaxation delays of 1 and 2 s each and with 1,024 × 2,048 and 512 × 512 complex time domain points, respectively. The three-dimensional HCCH-TOCSY spectrum has been recorded with a relaxation delay of 1 s and 2,048 × 96 × 128 time domain points. The spectral widths in the two dimensions were 13.9486 ppm in the two-dimensional TOCSY spectrum, whereas they were 13.961 ppm in the two-dimensional NOESY case. The three-dimensional spectrum has a spectral width of 6.9945 ppm in the direct direction, 70.0 ppm for the first indirect and 6.9945 ppm for the second indirect direction, being acquired with the 3-1-2 order and having the first indirect direction related with the ^{13}C.

All spectra were measured at 303 K. The NMR data were acquired with the program TOPSPIN (Bruker, Karlsruhe).

A synthetic two-dimensional NOESY spectrum was calculated with the AUREMOL module RELAX-*JT*2 (Ried et al. 2004) from the three-dimensional structure of HPr (H15A) from *Staphylococcus aureus* and from the corresponding experimental chemical shifts. The simulation is based on a full relaxation matrix approach and includes also T_2-calculations and J-couplings. Gaussian noise was added corresponding to a signal to noise ratio of approximately 2σ for a proton-proton pair in a distance of 0.5 nm as described by Baskaran et al. (2009). The resulting time domain data were filtered by exponential multiplication with a line broadening in the two dimensions of 3 Hz and finally the data have been Fourier transformed. The water artifact was produced by measuring a 2D-NOESY spectrum of 90% H_2O/10% D_2O with solvent pre-saturation, having the same acquisition parameters used for spectra simulation. After Fourier transformation this spectrum was added to the synthetic spectrum scaled in such a way that the maximum of the water was about 5,000 times stronger than a typical amide signal.

Availability of the program

All the developed routines have been integrated in the program AUREMOL (Gronwald and Kalbitzer 2004) and can be downloaded from www.auremol.de.

Theory

Projective subspace techniques

Time series analysis techniques often rely on embedding one-dimensional sensor signals, the FIDs, in the space of their time-delayed coordinates. Embedding can be regarded as a mapping that transforms a one-dimensional time series $x^i = (x^i[0], x^i[1], \ldots, \ldots, x^i[N-1])^T$ into a sequence of M time-lagged vectors. A multidimensional NMR spectrum then consists of Q FIDs $x^i, (i = 1, \ldots, Q)$. Each FID x^i of length N is embedded in its delayed coordinates with an $(N - M + 1)$ window size, to form a trajectory matrix X with its characteristic Toeplitz structure.

$$X = \begin{bmatrix} x[M-1] & x[M] & .... & ... & x[N-1] \\ x[M-2] & x[M-1] & .... & ... & x[N-2] \\ x[M-3] & x[M-2] & x[M-1] & ... & x[N-3] \\ .... & .... & .... & & .... \\ .... & .... & .... & & .... \\ x[1] & x[2] & x[3] & .... & x[N-M+1] \\ x[0] & x[1] & x[2] & .... & x[N-M] \end{bmatrix} \quad (1)$$

The embedding dimension can be estimated using model order selection techniques (Liavas and Regalia 2001). Note also that the Toeplitz matrix X of $(M \times (N - M + 1))$ dimensions has identical entries along its (top left to bottom right) diagonals. Any multidimensional signal vector x_k that constitutes the columns of X, is projected onto the directions (eigenvectors) related with the largest eigenvalues of the covariance matrix. Dimension reduction can be achieved by deliberately projecting the data vectors x_k only onto the $L < N$ directions corresponding to the L eigenvectors related with the L largest eigenvalues. In this case it corresponds to a denoising procedure as the eigenvectors related to the smallest eigenvalues encompass just noise.

The reconstruction after zeroing the eigenvector related to the largest eigenvalue that corresponds to the dominant solvent signal, leads to a new set of vectors $x_k^{'}$ forming the estimated trajectory matrix $X^{'}$. Note that in general elements along each descending diagonal of $X^{'}$ will not be identical like in case of the original trajectory matrix X. This can be cured, however, by replacing the entries in each diagonal by their average, obtaining again a Toeplitz matrix X_r. This procedure assures that the Frobenius norm of the difference $\left(X_r - X^{'}\right)$ attains its minimum value among all possible solutions to get a matrix with all diagonals equal (Golyandina et al. 2001; Teixeira et al. 2008). The one-dimensional signal $x^i[n]$ is thus obtained by reverting the embedding, forming the signal with the mean of the values along each diagonal of $X^{'}$.

The Singular Spectrum Analysis (SSA) is essentially a principal component analysis applied to the covariance matrix C formed with the centered trajectory matrix of each FID. The following steps (1–7) need to be repeated for every FID.

1. The data matrix X need to be centered to render it zero mean.
2. An $(M \times M)$-dimensional correlation matrix C is computed via

$$C = \frac{XX^T}{(N - M + 1)} \tag{2}$$

3. The eigenvalues decomposition of the covariance matrix is computed, yielding the eigen-representation of the correlation matrix with the U matrix containing the M eigenvectors in its columns and D is the diagonal matrix of the eigenvalues

$$C = \frac{1}{(N - M + 1)} XX^T = UDU^T \tag{3}$$

After this step denoising can be achieved by projecting the multidimensional signal into the subspace spanned by the eigenvectors corresponding to the $L < M$ largest eigenvalues. The M components of each FID are extracted by projecting the trajectory matrix along the directions given by the eigenvectors. The components are contain in the subspace matrix S calculated by

$$S = U^T X \tag{4}$$

4. During the reconstruction process the eigenvector related to the largest eigenvalue is nullified, yielding the new trajectory matrix in the following manner

$$X' = U_{\text{null}} U_{\text{null}}^T X \tag{5}$$

5. The mean value is added to the new data matrix X'.
6. The reconstructed data matrix X' does not possess a Toeplitz structure anymore. However, it is easily reconstituted by diagonal averaging and replacing every element along a diagonal by its averaged element.
7. The reconstructed one-dimensional signal x', i.e. the reconstructed FID is finally obtained reverting the embedding process.

The reconstructed total data matrix $X^{\text{tot}} = [x'^{(1)}, x'^{(2)}, \ldots, x'^{(q)}]$ is obtained using the extracted one-dimensional signals $x'^{(i)}$ with $i = 1, \ldots, q$, according to the number of recorded FIDs.

Note that the reconstruction process can proceed in two different ways which should be equivalent in principle. The reconstructed FID is obtained by nullifying the projection onto the eigenvector corresponding to the largest eigenvalue, i.e. zeroing this eigenvector in the eigenvector matrix U_{null}. An equivalent approach would be by reconstructing the solvent signal using only the largest eigenvector. Then the FID related to the protein resonances can be obtained subtracting it from the original data, i.e. $y[n] = x[n] - x'[n]$.

Automated base plane correction in the frequency domain

After removing the strong solvent signal, the base plane usually needs to be corrected in the frequency domain. Several methods for baseline correction have been developed. The most efficient and robust one is probably the cubic spline interpolation (Zolnai et al. 1989). The latter, however, induces new artifacts in areas where only few baseline points can be defined. In the present work, the linear spline interpolation (Saffrich et al. 1993) of the base points has been used as a valid alternative, being even more efficient and simpler. Traditionally, the base points where the base plane should be zero and where no relevant peaks are expected are defined interactively by the user. In automation this is not acceptable; hence the points have to be identified by the program. A method similar to the one proposed by Guentert and Wuethrich (1992) is used to automatically recognize the baseline regions in the spectrum. It is based on the observation that a contiguous piece of a row or a column of the data matrix can be well fitted by a straight line only if it lies in a pure baseline region. The most important parameter here is the size W of the window examined around a data point k that must be clearly larger than the expected line width of a protein resonance peak. Therefore the window size W must automatically change depending on the investigated spectrum. In general the default value of 75 Hz described by Guentert and Wuethrich (1992) is suitable for homonuclear proton NMR spectra. In order to allow the use of an appropriate adaptable window size W dependent on the type of experiment, the algorithm has been modified. The spectrum is firstly evaluated peak by peak fitting a Lorentzian function to the datasets optimized by the nonlinear least-squares algorithm of Levenberg–Marquardt (Levenberg 1994; Marquardt 1963). Only peaks are considered having intensities larger than 3-times the noise level σ_N. The maximum values LW of the line widths lw of all peaks are then computed separately for each dimension in the following manner:

$$\begin{aligned} &\text{LW}_1 = \max(\max(\text{lw}_r)) \text{ with} \\ &\quad r = 1, 2, \ldots, \text{number of rows} \\ &\text{LW}_2 = \max(\max(\text{lw}_c)) \text{ with} \\ &\quad c = 1, 2, \ldots, \text{number of columns} \end{aligned}$$

In a two-dimensional case, two line widths histograms are generated representing the line width distributions

within the frequency range (0, LW_1) and (0, LW_2), respectively. The two histograms contain the maximum line widths values of each row and column, respectively. The most frequently occurring maximum line width is used as reference to fix the window size W on the considered dimension. The actual window size is thus chosen as twice the line width at the maximum of the line width histogram.

The window slides row-wise and column-wise. Within each sliding window centered at data point k, the measured data points are approximated by a straight line and the mean square deviation χ_k^2 of the data points from the best fitting straight line is determined. Finally a threshold is defined as $\text{th} = \tau\chi^2_{\min}$ and those regions in the spectrum where $\chi^2 < \text{th}$ are identified as pure baseline regions. Typically, τ was set to 10.

The total set of pure baseline points is again approximated by a straight line using the linear spline interpolation method. Afterwards those approximated rows and columns are subtracted from the original dataset. However, long stretches of interpolated baseline regions yield to straight lines of zeros in the baseline corrected spectrum. In order to avoid this problem, the interpolation is not applied to all the consecutive recognized baseline points. In a stretch of five consecutive baseline points, only one is chosen to be interpolated (i.e. the middle one). Moreover, the intensity value of each chosen baseline point is not directly linearly interpolated, but it is substituted by the mean value between its own intensity and the intensities of the two adjacent points. As last step of baseline correction, the same procedure described above is applied column-wise excluding the points already corrected along the rows from the search of baseline points.

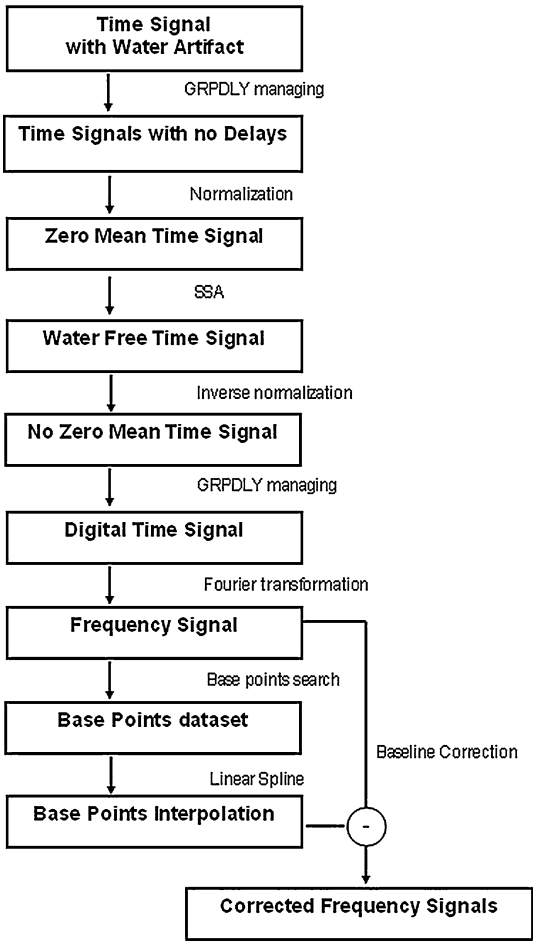

Fig. 1 Schematic overview of the automated water artifact removal procedure by means of SSA from digitally filtered data and subsequent automated baseline correction by linear spline interpolation

Implementation

The strategy described in Fig. 1 has been developed and integrated into the AUREMOL software package for automated water artifact removal from multidimensional NMR spectra and automated baseline correction. Starting from a multidimensional time domain signal, the singular spectrum analysis is applied to each experimental FID (to the rows of the data matrix) separately, performing steps 1–7 as explained in the theory section. This procedure generates an ensemble of artifact-free signals. Using a modern Bruker spectrometer the experimental data are typically oversampled and digitally frequency filtered (DQD-mode) (Moskau 2002). In this case, the data have to be preprocessed before applying SSA, since they do not correspond to a classical FID (damped cosine function). Here, the first few data points of each FID represent the group delay due to the digital frequency filtering and do not contain specific spectral information. Before applying the SSA, they have to be removed from the FID and the remaining data have to be left-shifted correpondingly. Moreover, during the embedding step the data are transformed to zero mean and normalized to unit norm (z-transformation). Each FID, or equivalently each row of a two-dimensional data matrix, is embedded into a feature space of dimension $K = 20$, whereas a fixed shift of one data point is used (see eq. 1). The embedding dimension has been chosen empirically. In Fig. 2, the resulting 20 estimated components, obtained from a two-dimensional NOESY spectrum, demonstrate a clear separation between the water artifact signal and the rest of the signal in the time domain. Typically, as described in Fig. 3 (showing the corresponding components in the frequency domain), one of the estimated components represents the water artifact almost perfectly, and about ten components are needed to reconstruct the protein signals. All the remaining components only contain pure noise. Therefore, the restriction to a smaller number of dimensions during

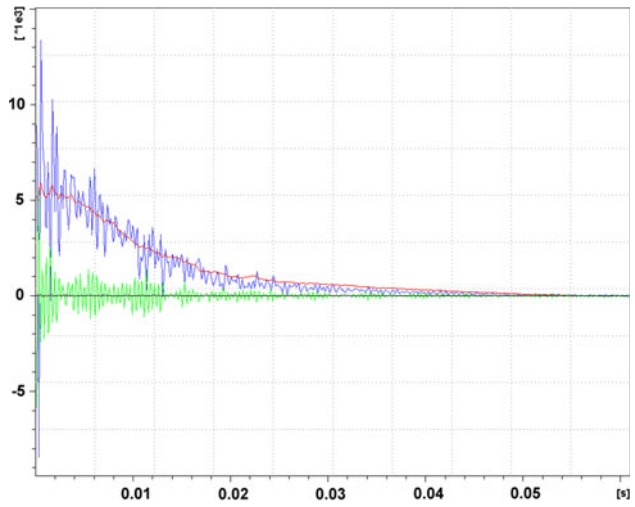

Fig. 2 Representation of some components of the subspace matrix S (eq. 4) calculated from the first FID of the two-dimensional NOESY experimental spectrum of HPr. Time domain data, 1 × 512 complex data points; embedding dimensions of the trajectory matrix 20 × 492, number of extracted components 20. Superimposition in the time domain of the first component (*red*) related to the solvent signal, the second component (*green*) representing a portion of the protein signal, the last component (*black*) containing only noise and the first original FID before the decomposition (*blue*)

the embedding step would be possible, but since computations are fast and efficient, the proposed embedding dimension $K = 20$ has been set as default.

The resulting trajectory matrix from each FID is fed into the PCA algorithm to determine the eigenvectors and eigenvalues of the embedded data. Next, the eigenvector belonging to the largest eigenvalue extracted is set to zero. Before reconstructing the original signal, data along the diagonals of the estimated trajectory matrix need to be replaced by their average to preserve the Toeplitz structure of the trajectory matrix. An inversion of the zero mean and data normalization steps is applied at the output of the SSA, rendering the water removal procedure more effective and avoiding scaling problems on the data. The previously stored group delay points (when existing) are then re-appended to the corrected FIDs. This particular treatment of the digitally filtered data for water removal avoids the generation of undesired artifacts.

The procedure is iteratively repeated for all the trajectory matrixes of all FIDs, independent of the dimensions of the data set. Once the water artifact is removed from the whole multi-dimensional time domain data set, the signal is automatically Fourier transformed to the frequency domain and a phase correction is applied coherently with the group delay time shift introduced by the digital filter. The baseline points are then selected from the multi-dimensional artifact corrected spectrum as described above and then linearly interpolated row by row and column by column. The interpolation is then subtracted from the artifact-free spectrum row-wise and column-wise.

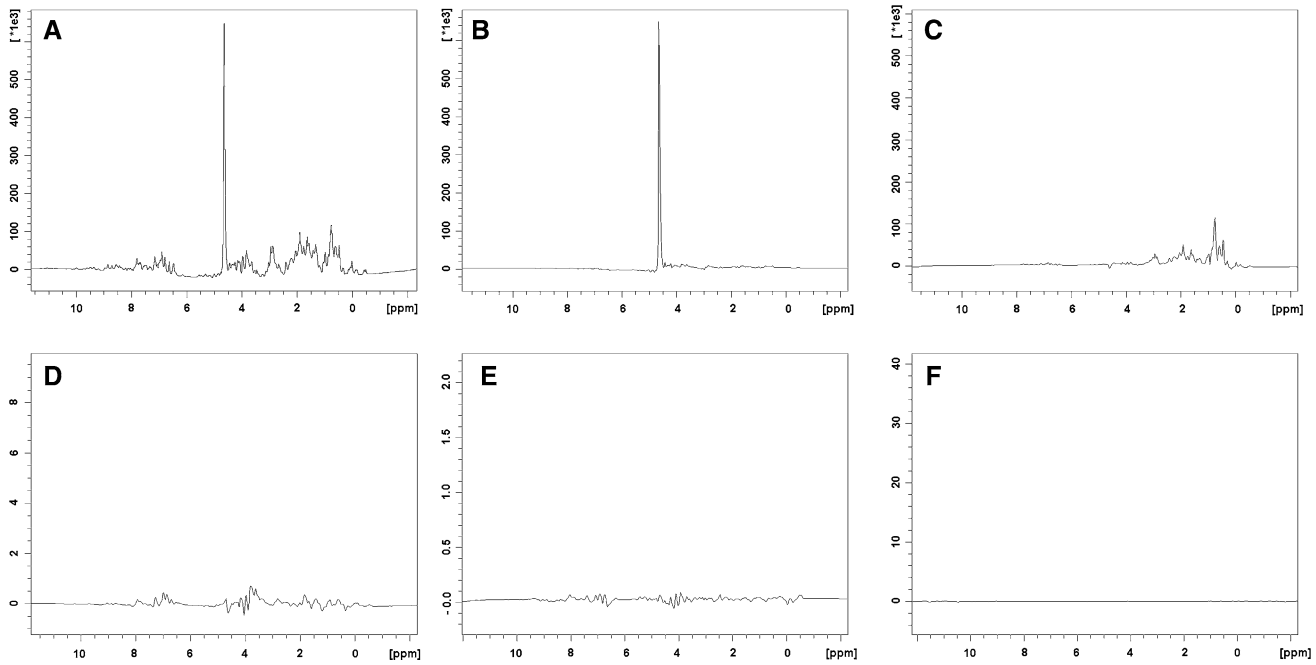

Fig. 3 Representation in the frequency domain of some of the extracted components from the trajectory matrix of the first FID of the two-dimensional NOESY experimental spectrum of HPr protein. Time domain data, 1 × 512 complex data points, embedding dimensions of the trajectory matrix 20 × 492, number of extracted components 20, size of the real data after Fourier transformation 20 × 492. A representation of the original data after Fourier transformation of the first FID (**A**), the first estimated component (**B**), the second component (**C**), the fifth component (**D**), the tenth component (**E**) and the last component (**F**) in the frequency domain

Results and discussion

The performances of the different methods have been validated by comparing the results with experimental as well as synthetic data. These latter have the advantage that the pure artifact-free spectrum is available and can be used as the "gold" standard for the obtained results. These synthetic spectra were calculated for a medium sized protein, namely the histidine-containing phosphocarrier protein (HPr) from *Staphylococcus aureus.* It is 88 residues in size and its structure consists of three α-helices and a four stranded anti-parallel β-sheet (Maurer et al. 2004). First a noiseless 2D NOESY spectrum was calculated from the 3D structure of the protein with RELAX-JT2 implemented in AUREMOL with inclusion of J-couplings and T_2-relaxation terms and the addition of Gaussian noise (see "Materials and methods"). Since only base line artifacts are removed by the methods described here, this spectrum is used as reference spectrum. Finally, a strong water-artifact signal was added to obtain a test spectrum. For the used experimental spectra, reference spectra do not exist, thus the performance of the routines cannot be quantified absolutely but only a visual inspection of the data can be applied for quality assessment.

Performance of the automated linear spline

The performance of spline-like base plane corrections depends critically on the selection of the baseline points assumed to be part of optimal base plane and to do not contain valid signals. With an interactive selection of these points the results are rather satisfying (Zolnai et al. 1989; Saffrich et al. 1993); however, for automation this manual selection is not acceptable. Therefore we adapted and generalized a method for base point selection proposed by Guentert and Wuethrich (1992) for the use of a linear spline base plane correction. Here, the inherent criteria for a base plane point are (1) that it is part of a region that can be approximated by a straight line but (2) is not part of a peak itself. The second criterion is confirmed by selecting a window that is clearly larger than line width of true peaks. If the window would contain the peak maximum, criterion 1 would not be fulfilled. In its original paper Guentert and Wuethrich selected a fixed number of data points for the definition of the sliding window, a selection that is appropriate if only one type of spectra (e. g. homonuclear 2D-spectra) is used. Since a peak width recognition procedure that is applicable to all kinds of NMR spectra was required, we had to devise a method that determines the optimal size of the windows in all dimensions in a given spectrum. For that the distribution of line widths of the peaks in the spectrum under consideration was determined and the window size was set as a multiple of the most frequent line width. When baseline points are direct neighbors, the linear spline creates regions with a noiseless baseline, that visually looks nice but is not acceptable by some processing methods. Therefore, a special selection method had to be devised (see "Theory").

An example for the application of ALS is shown in Fig. 4 where the synthetic NOESY spectrum is shown simulated without a water signal and any base plane deviations. The same NOESY spectrum containing the

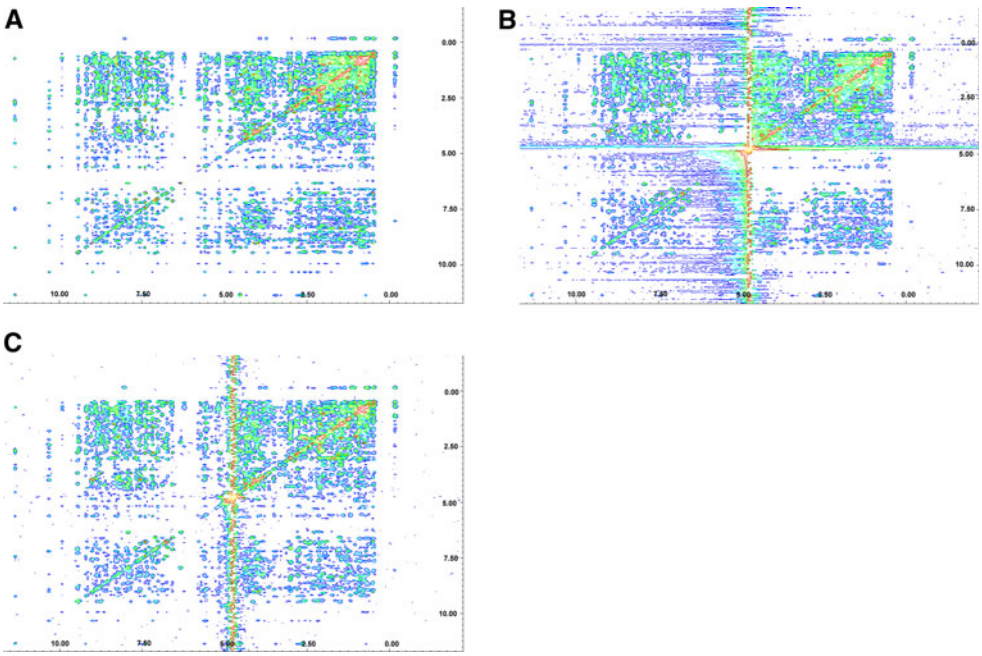

Fig. 4 ALS baseline correction. The NOESY spectrum of the histidine-containing phosphocarrier protein (HPr) from *Staphylococcus aureus* was back calculated using the AUREMOL module RELAX-JT2. Mixing time 100 ms, time domain data, 512 × 1,024 complex data points, size of the real data after Fourier transformation 51 × 1,024, zero filling before transformation. **A** Synthetic two-dimensional NOESY spectrum of HPr, **B** the simulated NOESY spectrum with a water artifact added, and **C** the spectrum obtained after baseline correction of spectrum **B**

solvent signal and severe base plane deviations is compared with the base plane corrected spectrum in Fig. 4. It shows that the method is working outside the region around the water signal but of course the method cannot remove the water artifact itself. This should be the domain of the SSA-module. However, the application of the ALS after the application of SSA leads to a significant improvement of the results.

Application of singular spectral analysis to multi-dimensional NMR spectra

The SSA-algorithm was first tested on the same synthetic spectrum (Fig. 4) already used for the application of the ALS algorithm. Figure 5 shows the result; the water artifact was almost completely removed and the obtained spectrum looked almost as the unperturbed spectrum shown in Fig. 4. It is evident that water and baseline artifacts are strongly suppressed, whereas hidden protein resonances are recovered. A problem occurring with many processing methods is a change of peak intensities that cannot be accepted for a quantitative analysis of the data. The residual calculated as difference between the processed spectrum and the original artifact-free spectrum is almost zero (within the limits of pure noise) for the protein cross peaks as shown in Fig. 5.

Figure 6 shows as an example the application of SSA and ALS to an experimental two-dimensional TOCSY spectrum of the HPr protein. It is clearly seen that the solvent artifact and the base plane variations are largely suppressed and hidden protein resonances lying underneath the water are recovered. The boxed part of the spectrum close to the water resonance is shown in larger

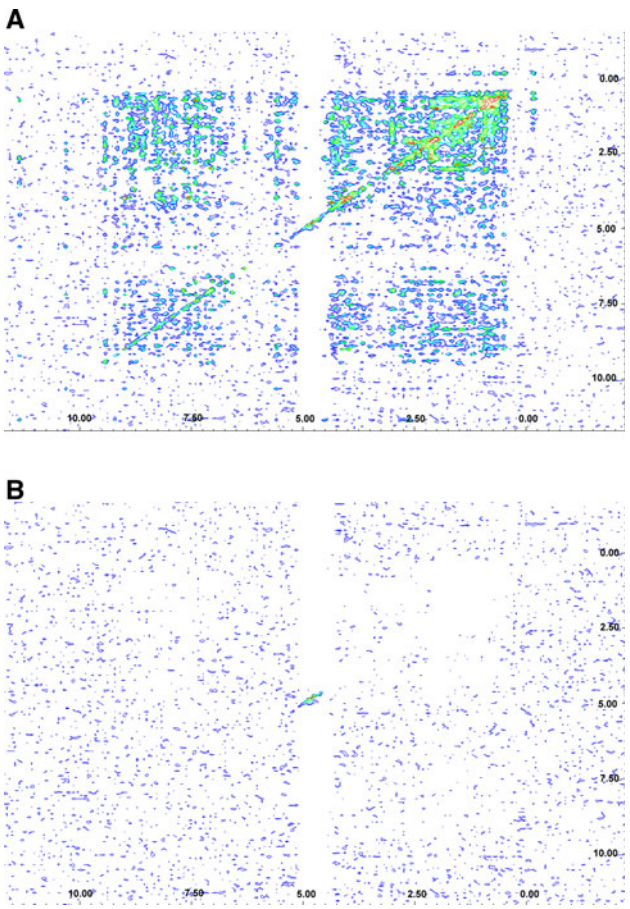

Fig. 5 Demonstration of artifact removal by SSA on a synthetic spectrum. The same synthetic two-dimensional NOESY spectrum of the histidine-containing phosphocarrier protein (HPr) from *Staphylococcus aureus* was used as in Fig. 4B. **A** Synthetic spectrum after solvent removal by SSA and ALS baseplane correction in two dimensions. **B** The residual obtained by substracting the original, artifact-free spectrum (Fig. 4A) from spectrum (**A**)

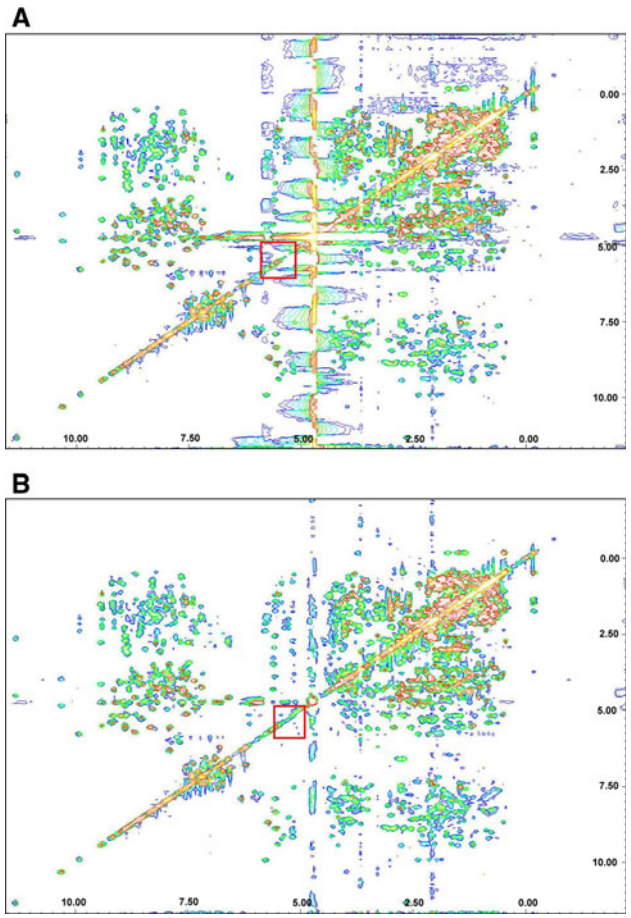

Fig. 6 Demonstration of artifact removal by SSA on an oversampled experimental two-dimensional spectrum. A TOCSY spectrum of the histidine-containing phosphocarrier protein (HPr) from *Staphylococcus aureus* in 500 µl of 95% H_2O/5% D_2O was measured on a Bruker Avance-800 spectrometer. Mixing time, 100 ms, relaxation delay 1 s, time-domain data matrix 1,024 × 2,048 complex data points, zero filling before Fourier transformation, size of the real data matrix after Fourier transformation 1,024 × 2,048, linear interpolation for baseline correction after transformation. **A** Original spectrum and **B** spectrum after application of SSA and ALS

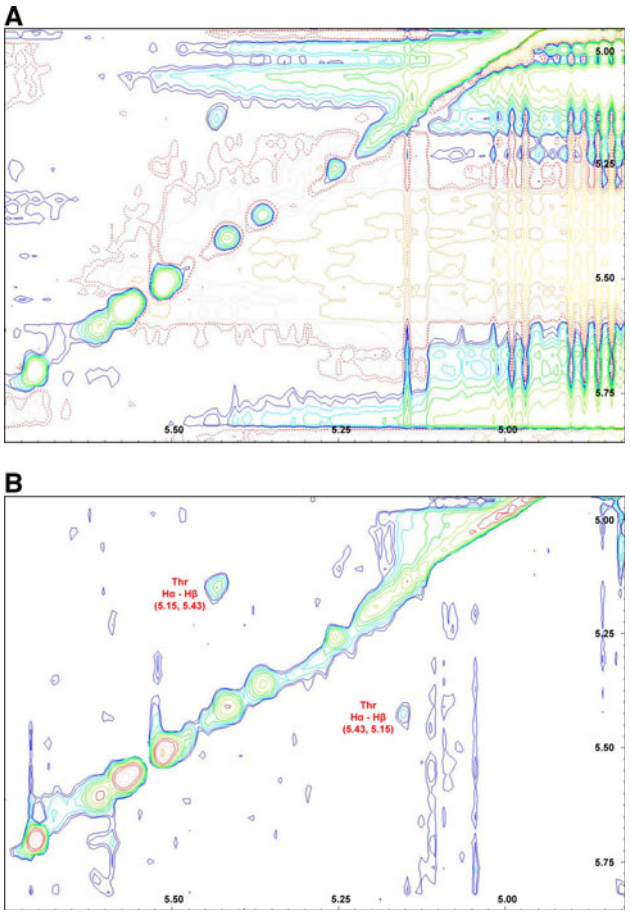

Fig. 7 Recovery of signals close to the solvent line. Enlargement of the artifact area with $\delta_2 \in [5.75,4.85]$ ppm and $\delta_1 \in [5.80,4.90]$ ppm of the experimental two-dimensional TOCSY spectrum (*red box* depicted in Fig. 6). **A** Original spectrum and **B** the spectrum after solvent removal procedure. A threonine H^{α}-H^{β} cross peak is marked

magnification in Fig. 7. It shows that a threonine $H^{\alpha} - H^{\beta}$ cross peak at (5.430, 5.150 ppm) superposed by the water resonance is clearly observable after the application of SSA. Finally, Fig. 8 shows the results of the artifact removal procedure to an oversampled three-dimensional HCCH-TOCSY spectrum. SSA was applied in the direct (t_3)-dimension to all rows (FIDs) of the 3D-time domain data. After performing the solvent removal, the data was Fourier transformed and ALS was applied in all (ω_2, ω_3)-planes. In general, the baseline correction could be performed not plane-by-plane wise but direction-wise, an option not implemented in the actual version of AUREMOL, that could give a slightly improved performance of the algorithm. The water resonance and its tails were almost completely removed. The recovery of the peak lying under the water is demonstrated in Fig. 9, showing the projection on the F1–F3 plane of the three-dimensional HCCH-TOCSY spectrum before and after the solvent removal.

The singular spectrum analysis has been tested both on pure time domain signals and on mixed time–frequency domain signals. In a two-dimensional spectrum, the mixed domain is generated by Fourier transforming only along the columns of the data matrix and consequently using each row separately as input to the SSA, since the rows can still be considered as time domain signals. In both cases the performance of the method is almost equally good but in particular starting from the time domain is preferable for an easier data managing. The NMR time domain signals are usually recorded as complex data. Practical tests show that the SSA method works better when the data are treated as complex numbers instead of working on the a priori

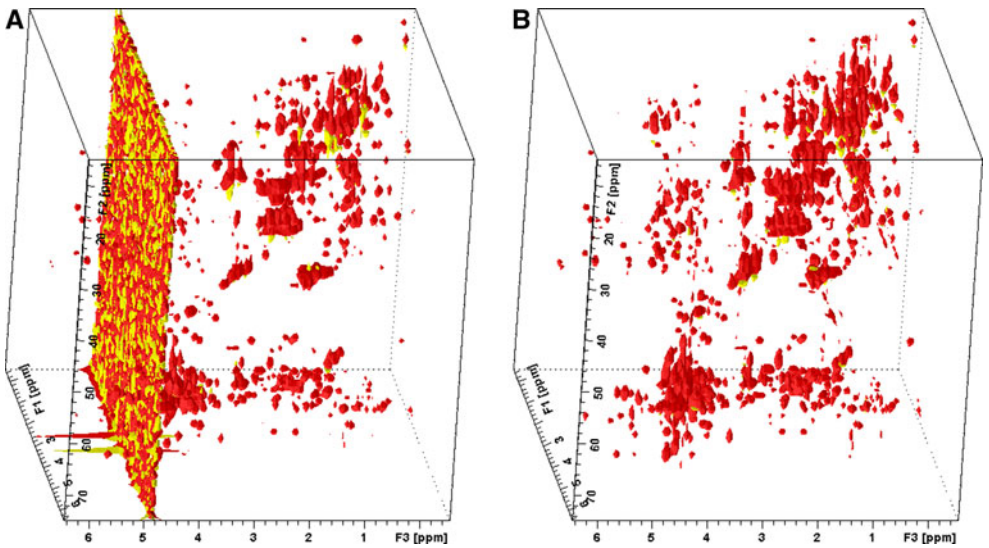

Fig. 8 Application of the singular spectrum analysis for water artifact removal to a three-dimensional spectrum. A subcube of a three-dimensional $^{1}H,^{13}C$ HCCH-TOCSY spectrum of the thioredoxin protein (Trx) from *Plasmodium falciparum* is shown prior (**A**) and after (**B**) artifact removal by SSA and ALS. Size of the subcube 2,048 × 96 × 128 real data points

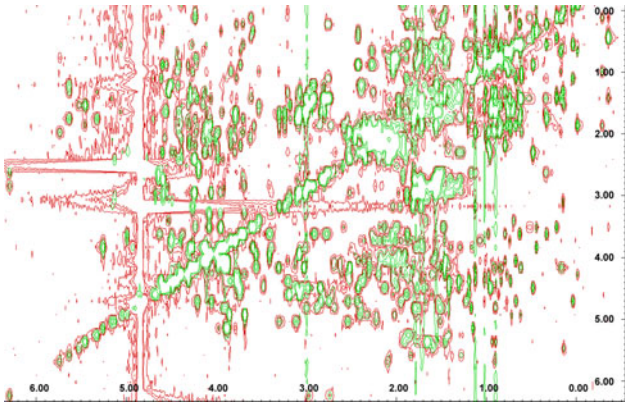

Fig. 9 Projection of a three-dimensional spectrum before and after application of SSA and ALS. The $^{1}H,^{13}C$ HCCH-TOCSY spectrum depicted in Fig. 8 was projected onto the F1-F3 plane. A portion of this projection is shown before (*red*) and after (*green*) the water artifact removal procedure

separated real and imaginary parts. When performing SSA with oversampled data, the reconstruction algorithm produces at the beginning of an FID an increasing signal (called the group delay) before the real free induction signal starts. If SSA is directly applied to that FID, no satisfactory result can be obtained. Therefore it is mandatory to perform the removal of these data points from each FID before starting the SSA procedure. The performance of the embedding step used before the PCA algorithm can be highlighted if compared with the same procedure applied on the whole set of FIDs without generating any trajectory matrix. The number of projections onto the directions related with the largest eigenvalues of the covariance matrix is not more related to the embedding dimensions but it is equal to the number of measured FIDs or rows of a multi-dimensional spectrum. Dealing with a larger number of estimated components and with many different time signals simultaneously, increases the computational time and generates the problem of the components identification and assignment, since more than one projection can be related to the water artifact.

Conclusions

The application of singular spectrum analysis, followed by an automated linear spline, is mathematically rather simple and straightforward and gives at least as good results as do more complicated methods. An advantage for practical applications is the complete automation. In the AUREMOL software package it is used in a fully automated way that includes application of SSA to the time domain data set, followed by filtering, Fourier transformation, phase correction related to the group delay management and application of ALS without any user intervention. The only step that still must be performed interactively is the determination of the parameters used for phase correction since here still no stable method exists. We believe that, considering the simplicity and the efficiency of the automated implementation and the ease with which the spectroscopist can insert it to the own processing strategy, singular spectrum analysis for water artifact removal, as presented in this work, should become an useful tool for the treatment of multidimensional NMR spectra.

Acknowledgments This work was supported by the Bundesministerium für Forschung (BMBF), the Deutsche Forschungsgemeinschaft (DFG), the European Union, and the Fonds der Chemischen Industrie (FCI).

References

Adler M, Wagner G (1991) Removal of dispersive baseline distortions caused by strong water signals. J Magn Reson 91:450–454

Antoine JP, Coron A, Dereppe JM (2000) Water peak suppression: time-frequency vs time-scale approach. J Magn Reson 144:189–194

Barache D, Antoine JP, Dereppe JM (1997) The continuous wavelet transform, an analysis tool for NMR spectroscopy. J Magn Reson 128:1–11

Baskaran K, Kirchhoefer R, Huber F, Trenner J, Brunner K, Gronwald W, Neidig KP, Kalbitzer HR (2009) Chemical shift optimization in multidimensional NMR spectra by AUREMOL-SHIFTOPT. J Biomol NMR 43(4):197–210

Boehm M, Stadlthanner K, Gruber P, Theis FJ, Lang Ew, Tome AM, Teixeira AR, Gronwald W, Kalbitzer HR (2006) On the use of simulated annealing to automatically assign decorrelated components in second-order blind source separation. IEEE Trans Biom Eng 53(5):810–820

Brown DE, Campbell TW (1990) Enhancement of 2D NMR spectra using singular value decomposition. J Magn Reson 89:255–264

Coron A, Vanhamme L, Antoine JP, Hecke PV, Van Huffel S (2001) The filtering approach to solvent peak suppression in MRS: a critical review. J Magn Reson 152:26–40

Friedrichs MS, Metzler WJ, Mueller L (1991) Removal of diagonal peaks in two-dimensional NMR spectra by means of digital filtering. J Magn Reson 95:178–183

Gerbrands JJ (1981) On the relationships between SVD, KLT and PCA. Pattern Recogn 14:375–381

Ghil M, Allen MR, Dettinger MD, Ide K (2002) Advanced spectral methods for climatic time series. Rev Geophys 40:3.1–3.41

Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of time series structure: SSA and related techniques. Chapman and HALL/CRC, London

Grahn H, Delaglio F, Delsuc MA, Levy GC (1988) Multivariate data analysis for pattern recognition in two-dimensional NMR. J Magn Reson 77:294–307

Gronwald W, Kalbitzer HR (2004) Automated structure determination of proteins by NMR spectroscopy. Progr NMR Spectr 44:33–96

Gruber P, Stadlthanner K, Boehm M, Theis FJ, Lang EW, Tome AM, Teixeira AR, Puntonet CG, Gorriz-Saez JM (2006) Denoising using local projective subspace methods. Neurocomp 69:1485–1501

Guentert P, Wuethrich K (1992) FLATT—a new procedure for high-quality baseline correction of multidimensional NMR spectra. J Magn Reson 96:403–407

Guenther UL, Ludwig C, Rueterjans H (2002) WAVEWAT—improved solvent suppression in NMR spectra employing wavelet transforms. J Magn Reson 156:19–25

Hardy JK, Rinaldi PL (1990) Principal component analysis for artifact reduction in COSY spectra. J Magn Reson 88:320–333

Hore PJ (1989) Methods in enzymology. In: Oppenheimer N, James TL (eds) vol 176, pp 64–77

Kuroda Y, Wada A, Yamazaki T, Nagayama K (1989) Postacquisition data processing method for suppression of the solvent signal. J Magn Reson 84:604–610

Levenberg K (1994) A method for the solution of certain problems in least squares. Quart Appl Math 2:164–168

Liavas AP, Regalia PA (2001) On the behavior of information theoretical criteria for model order selection. IEEE Trans Sign Proc 49:1689–1695

Lin YY, Hodgkinson P, Ernst M, Pines A (1997) A novel detection-estimation scheme for noisy NMR signals: applications to delayed acquisition data. J Magn Reson 128:30–41

Marion D, Ikura M, Bax A (1989) Improved solvent suppression in one- and two-dimensional NMR spectra by convolution of time-domain data. J Magn Reson 84:425–430

Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. J Appl Math 11:431–441

Maurer T, Meier S, Kachel N, Munte CE, Hasenbein S, Koch B, Hengstenberg W, Kalbitzer HR (2004) High-resolution structure of the histidine-containing phosphocarrier protein (HPr) from *Staphylococcus aureus* and characterization of its interaction with the bifunctional HPr kinase/phosphorylase. J Bacteriol 186(17):5906–5918

Mitschang L, Neidig KP, Kalbitzer HR (1990) Suppression of oscillatory artifacts in two-dimensional NMR spectra. J Magn Reson 90:359–362

Mitschang L, Cieslar C, Holak TA, Oschkinat H (1991) Application of the Karhunen-Loéve transformation to the suppression of undesired resonances in three-dimensional NMR. J Magn Reson 92:208–217

Moskau D (2002) Application of real time digital filters in NMR spectroscopy. Conc Magn Resonan 15:164–176

Munte CE, Becker K, Schirmer RH, Kalbitzer HR (2009) NMR assignments of oxidized thioredoxin from *Plasmodium falciparum*. Biomol NMR Assign 3:159–161

Parra L, Sajda P (2003) Blind source separation via generalized eigenvalue decomposition. J Mach Learn Res 4:1261–1269

Pijnapple WWF, Van Den Boogaart A, De Beer R, Van Ormondt D (1992) SVD-based quantification of magnetic resonance signals. J Magn Reson 97:122–134

Ried A, Gronwald W, Trenner JM, Brunner K, Neidig KP, Kalbitzer HR (2004) Improved simulation of NOESY spectra by RELAX-JT2 including effects of J-coupling, transverse relaxation and chemical shift anisotropy. J Biomol NMR 30(2):121–131

Saffrich R, Beneicke W, Neidig KP, Kalbitzer HR (1993) Baseline correction in *n*-dimensional NMR spectra by sectionally linear interpolation. J Magn Reson 101(B):304-308

Stadlthanner K (2007) Nonlinear ICA analysis of time series data. Logos Verlag, Berlin

Stadlthanner K, Tome AM, Theis FJ, Lang EW, Gronwald W, Kalbitzer HR (2006) Separation of water artifacts in 2D NOESY protein spectra using congruent matrix pencils. Neurocomp 69:497–522

Sundin T, Vanhamme L, Van Hecke P, Dologlou I, Van Huffel S (1999) Accurate quantification of 1H spectra: from finite impulse response filter design for solvent suppression to parameter estimation. J Magn Reson 139:189–204

Teixeira AR, Tome AM, Boehm M, Puntonet CG, Lang EW (2008) How to apply non-linear subspace techniques to univariate biomedical time series. IEEE Trans Instrum Meas

Tsang P, Wright PE, Rance M (1990) Signal suppression in the frequency domain to remove undesirable resonances with dispersive lineshapes. J Magn Reson 88:210–215

Zhu G, Smith D, Hua Y (1997) Post-acquisition solvent suppression by singular-value decomposition. J Magn Reson 124:286–289

Zolnai Z, Macura S, Markley JL (1989) Spline method for correcting baseplane distortions in two-dimensional NMR spectra. J Magn Reson 82:496–504