Trends in Applied Spectroscopy
Vol. 9, 2012

Original Communication

Independent component analysis (ICA) and singular spectrum analysis (SSA) for solvent artifact suppression in one-dimensional NMR spectroscopy: A comparative analysis

Silvia De Sanctis, Wilhelm M. Malloni, Werner Kremer, Elmar W. Lang and Hans R. Kalbitzer*
Institute of Biophysics and Physical Biochemistry, University of Regensburg, D-93040, Germany

ABSTRACT

NMR spectroscopy is generally performed in aqueous solutions, thus a frequently occurring problem is the recovering of weak resonances superimposed by an intense solvent signal. Recently, the fully automated approach AUREMOL-SSA/ALS has been successfully applied to different types of n-dimensional NMR spectra; it uses singular spectrum analysis (SSA) for solvent artifact removal combined with an automated linear spline (ALS) for baseline correction. Here, independent component analysis (ICA) is introduced as a new method for the suppression of the solvent signal and compared with SSA. In principle, ICA can overcome some limitations of the SSA but requires a suitable experimental acquisition protocol for its application to one-dimensional NMR spectra. SSA is usually applied to the time domain signal (FID) and requires as input a single FID. In contrast, ICA is optimally applied to frequency domain signals and requires as input at least two different spectra. Here, different acquisition schemes tailored for ICA have been applied to one-dimensional synthetic and experimental datasets of the small globular protein HPr as well as on urine spectra. Excellent results have been obtained by the ICA especially under conditions where SSA fails.

*Corresponding author
hans-robert.kalbitzer@biologie.uni-regensburg.de
This work was presented at the EUROMAR 2011, Frankfurt am Main, Germany.

KEYWORDS: AUREMOL-SSA/ALS, singular spectrum analysis, independent component analysis, NMR NOESY spectra, pulse sequence, ICA-tailored experiments

ABBREVIATIONS

ALS, automated linear spline; FID, free induction decay; FIR, finite impulse response filter; ICA, independent component analysis; HPr, histidine containing phosphocarrier protein; SSA,Singular spectrum analysis

INTRODUCTION

NMR spectroscopy has become a powerful analytical method with many applications in various fields of biochemical and biophysical research. It is a suitable technique for identifying and quantifying small molecules in solutions and for determining the chemical structure of unknown substances. Quantification of small compounds in biofluids such as blood plasma and urine is a key feature in metabolomics (see e.g. [1, 2]). In addition, NMR can also provide information on the spatial arrangement of atoms and their dynamics in biomacromolecules such as proteins and nucleic acids and their macromolecular complexes (see e. g. [3, 4]). A common problem of NMR spectroscopy is due to artifacts caused by the strong solvent signal (see e.g. [5] and references herein) that is in biological samples usually water.

A simple way to reduce the water ^{1}H signal is by replacing normal water with heavy water.

However, in general this is not a solution of the problem. In metabolomics, the aqueous samples of body fluids should be analyzed as they are. In protein NMR spectroscopy, the signals of amide protons contain valuable structural information. Replacement of normal water by heavy water leads to an exchange of protons by deuterons in the amide groups and thus it leads to the disappearance of the correspondent proton signals.

For a protein dissolved in 90% H_2O/10% D_2O the concentration of solvent protons is more than five orders of magnitude greater than the typical concentration of the protein protons in the solution. Correspondingly, many experimental methods [6-11], have been developed, that at least partly suppress the water signal. Neither conventional solvent signal suppression nor tailored excitation procedures are able to reveal the solute signals very close to the solvent that may be important for the interpretation of the data. Alternatively, many post-processing methods were proposed that attempt to deal with this problem [12-28].

A powerful program that is able to strongly reduce the solvent signal in NMR spectra is AUREMOL-SSA/ALS. It is based on singular spectrum analysis (SSA) [29] and can be applied to one-dimensional [28] as well as to multi-dimensional spectra [27]. However, AUREMOL-SSA causes processing artifact when the solvent artifact is not the dominant signal in the spectrum. Independent component analysis (ICA) represents a promising alternative for those spectra that cannot be properly managed by the SSA. ICA [30] belongs to the class of Blind Source Separation (BSS) methods [31]. It has been successfully applied on EEG data revealing brain activities [32] and for feature extraction purposes from image and audio signals [33, 34]. The principal component analysis (PCA) [35] extracts uncorrelated components by means of second order statistics (variance maximization), while ICA looks for independent sources using higher order statistics (non-Gaussianity maximization).

SSA is typically applied on the time domain data, whereas the ICA can be used to decompose the overlapping signals directly in the frequency domain. SSA embeds a single time domain signal (FID) in a multi-dimensional space, yielding a trajectory matrix containing shifted versions of the same FID. An inherent property of ICA is instead that it needs at least as many different spectra of the same sample as the number of source signals that should be separated. In addition, the source components have to be differently weighted in the spectra. As shown earlier [25, 26], for higher dimensional spectroscopy one can use several rows in the mixed time-frequency domain for this purpose. In one-dimensional NMR spectroscopy this is not possible since usually only one FID is available. The solution to that problem lies in the creation of a set of one-dimensional spectra tailored for the application of ICA [36, 37]. In the following we will show the advantages and limits of the application of ICA for solvent removal as compared to SSA.

MATERIALS AND METHODS

Synthetic datasets and simulations

A synthetic one-dimensional spectrum has been calculated with the AUREMOL routine RELAX-*JT*2 [38] starting from the three-dimensional structure of a mutant of histidine-containing phosphocarrier protein HPr (H15A) [39] from *Staphylococcus aureus* and using the corresponding experimental chemical shifts. RELAX-*JT2* simulates multiplet structures as well as line widths. A one-dimensional 600 MHz NOESY spectrum with a mixing time of 0.15 s, a relaxation delay of 1.5 s, a spectral width of 12.65 ppm and 2048 complex time domain data points was simulated. The resulting time domain data was filtered by exponential multiplication with a line broadening of 3 Hz. The water artifact was produced by measuring a one-dimensional spectrum of 90% H_2O/10% D_2O with solvent pre-saturation at 600 MHz, having the same acquisition parameters as those ones used for the HPr spectrum simulation. Before the Fourier transformation the water artifact signal was added to the synthetic time domain signal of the protein. The phase and the amplitude of the added water signal were varied by performing respectively a phase correction and an intensity variation in the frequency domain followed by an inverse Fourier transformation. The same routine [38] has been used to generate a

600 MHz NOESY synthetic one-dimensional spectrum from the three-dimensional structure of the histidine-containing phosphocarrier protein HPr from *Staphylococcus carnosus* [40]. It has been back-calculated with a mixing time of 0.01 s, a relaxation delay of 1 s, a spectral width of 14.98 ppm and 32768 complex time domain data points.

Experimental datasets

All NMR spectra were recorded with a Bruker Avance-600 spectrometer operating at 600 MHz equipped with a cryoprobe. They have been recorded at 298 K.

Urine spectra

One-dimensional NOESY-type spectra of human urine have been recorded using oversampling [41] and digital filtering (Bruker DQD mode). The urine was buffered by 133 mM sodium phosphate, pH 7.4, 5% D_2O and 100 µM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) was added as internal reference. The water signal was reduced by a selective pre-saturation pulse of 5 s. 131,072 complex time domain points were recorded in the digital mode (group delay of 144 data points) with a spectral width of 20.03 ppm. A set of four experiments has been measured with different NOE mixing times of 10, 20, 800, and 1000 ms (for details see Fig. 1a). These two urine datasets (10, 20 ms and 800, 1000 ms) have been used separately as input to the ICA.

HPr spectra

One-dimensional NMR spectra have been measured with a sample containing 1 mM uniformly ^{15}N-enriched HPr protein from *Staphylococcus carnosus* [40] in 95% H_2O/5% D_2O, pH 7. Two types of spectra were recorded: a set of 1D-NOESY spectra with different phase cycles (see Fig. 1a) and a set of diffusion weighted 1D-NOESY spectra (see Fig. 1b). The spectral width was 14.98 ppm and the mixing time was 10ms. 32,768 complex time domain points (including 140 points of the group delay) have been recorded. The water signal was reduced by selective pre-saturation of 1 s.

Software

All the NMR data were acquired with the program TOPSPIN (Bruker, Karlsruhe). AUREMOL-SSA/ALS and AUREMOL-ICA are parts of the program AUREMOL [5].

Theoretical considerations and implementation

Singular spectrum analysis of NMR data

The application of SSA to n-dimensional NMR spectroscopy ($n \geq 1$) has been described in detail earlier [27, 28]. It is applied to time domain NMR data independently on the dimensionality since it manages separately each (complex) FID. SSA is an extension of the PCA (principal component analysis) and it is a nonparametric method that allows decomposing a time series into a sum of M interpretable components (Fig. 2). The number M has to be adapted to the complexity and digital resolution of the spectra. In typical multi-dimensional NMR spectra of proteins with relatively low resolution (typically 2 K complex time domain data points in the direct dimension) the embedding in 20 dimensions (M = 20) is sufficient. In one-dimensional spectra (e. g. of body fluids) with much higher digital resolution (typically 16 K complex time domain data points) the embedding with M = 40 showed to be appropriate [28]. For solvent artifact removal, the component with the largest eigenvalue is nullified before reconstructing the signal.

Consider a one-dimensional time domain signal (FID) x_i of length N. When a finite impulse response filter (FIR) is used for oversampling [41], the initial portion of the signal does not contain useful information, thus it is excluded from the calculation and stored for a successive regeneration of the original dataset. The corresponding group delay (GRPDLY) severely affects the extracted components if it is not properly managed, being responsible of wiggles in the Fourier transformed spectrum. The SSA procedure needs only one time domain signal (FID) as input. The centered and normalized FID of length (1 x N_g) with N_g = (N-GRPDLY) is embedded in its delayed coordinates with an (N_g-M+1) window size. A phase correction is automatically applied by the developed algorithm in accordance to the time shift due to the initial group delay. An automated baseline correction is performed as well in the frequency domain in order to obtain the final spectrum. Thus, the already developed AUREMOL-SSA/ALS solvent

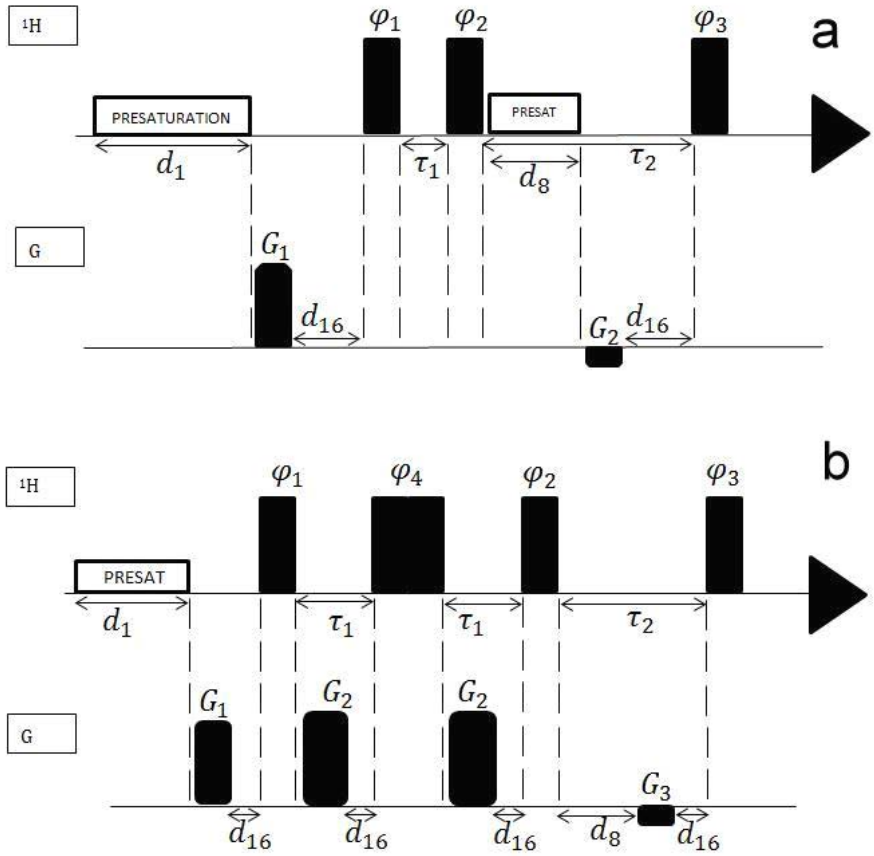

Fig. 1. ICA-tailored NMR experiments. The thin and the thick solid bars represent 90° and 180° hard RF-pulses. Gradient pulses applied along the z-axis are represented by black boxes.
(a) NOESY-type pulse sequence for editing the sign or intensity of the water signal. Here, two spectra with different weighting of the water and the protein signals can be obtained by variation of the mixing time (T_1-weighting). Alternatively, the sign of the water signal can be varied in the two spectra by a modified phase cycle that makes use of the fast radiation damping of the water signal.
Two separate spectra with a different sign of the water signal can be obtained by selecting in the first spectrum φ_1 = (x) (-x), φ_2 = 4(x,-x), φ_3 = 2(x)2(-x),2(y), 2(-y), φ_{rec} = 2(x) 2(-x),2(y)2(-y), and in the second spectrum φ_1 = (x) (-x), φ_2 = 4(-x,x), φ_3 = 2(x)2(-x),2(y), 2(-y), φ_{rec} = 2(-x) 2(x),2(-y)2(y). Typical settings are: relaxation delay, 5 s; presaturation, 1.5 s; power level, 60 dB: gradient pulses G_1, 1 ms duration with 50 G cm^{-1}, G_2, 1 ms and -10 G cm^{-1}; presaturation during mixing time d_8; sine gradient shape; delay for gradient recovery d_{16}, 8 ms. Mixing times τ_2 are 10 ms, 20 ms, 0.8 s, 1 s.
(b) Diffusion weighted NOESY-type spectrum for editing the intensity of the water signal. The phase cycle is: φ_1 = (x)(-x), φ_4 = (y)(-y), φ_2 = 8(x)8(-x), φ_3 = 2(x)2(-x)2(y)2(-y), and φ_{rec} = (x)2(-x)(x)(y)2(-y)(y)(-x)2(x)(-x)(-y)2(y)(-y). Relaxation delay, 1 s; gradient pulses G_1, 1 ms, 50 G cm^{-1}, G_2, 4 ms, 80 G cm^{-1} (first experiment) and 50 G cm^{-1} (second experiment), G_3, 1 ms, -10 G cm^{-1}; mixing time, 10 ms; delay for gradient recovery, 0.5 ms; gradient shape, sine.

removal method involves an adequate pre-processing step, including digitally filtered data managing and time domain signal normalization. It also encompasses a consecutive post-processing step concerning baseline and phase correction in the frequency domain. Moreover, from the theory it is evident that if the solvent signal is not the dominant one, nullifying the largest eigenvalue leads to an undesired removal of the strongest solute resonances.

Independent component analysis of NMR data

ICA represents the solution of the cocktail party problem [30], where the signal detected by a

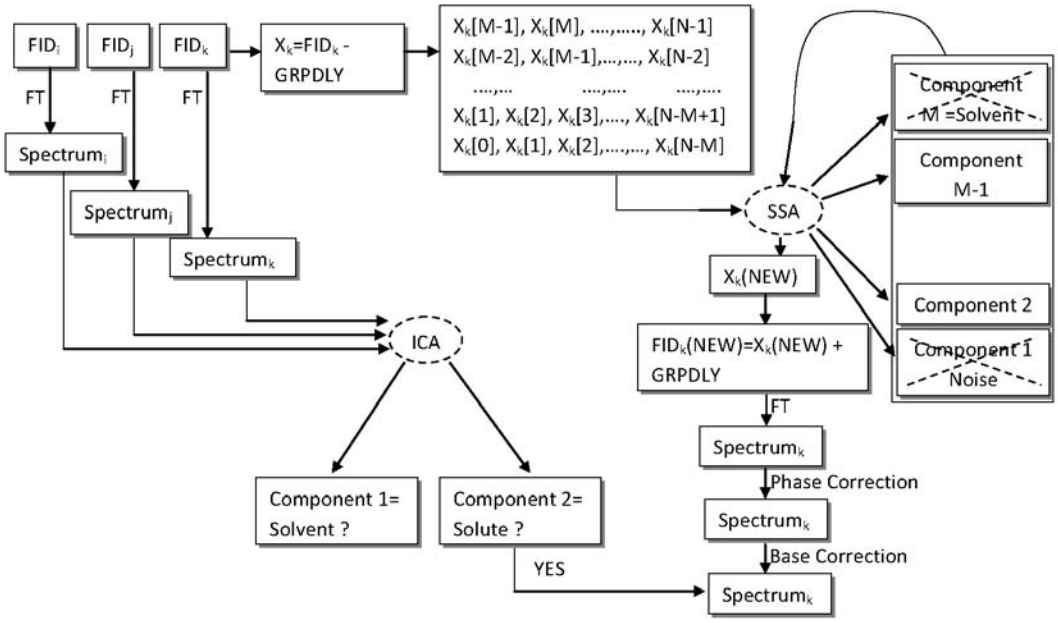

Fig. 2. Schematic description of ICA and SSA applied to NMR data. Application of ICA requires a set of n one-dimensional spectra with different contributions of the solvent signal. It performs the separation of the frequency domain data in two main components, the solvent and the solute signals that need to be (automatically) recognized. SSA uses only one of the FIDs as input. The points belonging to the group delay are excluded and afterwards a trajectory matrix containing M shifted versions of the FID is generated. The algorithm extracts the M components and nullifies that one corresponding to the signal with the highest variance (the solvent). An inverse reconstruction process is then applied by the SSA and a new FID is built. The points belonging to the group delay are re-appended at the beginning of the FID, then it is Fourier transformed, phase corrected in accordance with time shift due to the group delay and baseline corrected. The ICA avoids all the previously described procedural steps.

microphone is a linear superposition of the signals coming from many different sources (talking persons attending the party). These signals can be decomposed when not only one microphone is present in the room but there are many microphones at different locations. The ICA works by finding a transformation of the measured signals (mixtures) that produces independent components (sources), assuming that each of these independent signals is associated with a different physical process. The general scheme is depicted in Fig. 2.

Considering the signals of two different sources, denoted by s_1 and s_2, the linear superposition $x_i(t)$ detected at microphone i can be expressed as follows:

$$x_i(t) = a_{i1}s_1(t) + a_{i2}s_2(t) \quad (1)$$

The number of detected signals x_i must be at least equal to the number of sources to be separated. In our application, we would need at least two different spectra where the signals of the solute and the solvent are weighted differently.

The set of mixtures (NMR spectra) can be represented as a matrix $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2})^T$

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^N \\ x_2^1 & x_2^2 & \cdots & x_2^N \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} s_1^1 & s_1^2 & \cdots & s_1^N \\ s_2^1 & s_2^2 & \cdots & s_2^N \end{pmatrix} = \mathbf{AS} \quad (2)$$

where **X** is the (2 x N) matrix of observations (i. e. the two one-dimensional NMR spectra) at each value t = 1,…,N, while **A** is the (2 x m) mixing matrix and **S** is the (m x N) matrix containing the m independent source signals (the solute and the solvent) at each time point t.

In principle, ICA tries to find the inverse matrix $\mathbf{A}^{-1}$ that solves the problem. However, in general, mathematically a unique solution of eq. 2 does not exist and additional conditions have to be fulfilled

by the data (and regarded by the experimental scheme) to assure the identification of the underlying components.

In order to estimate the coefficients a_{ij} it must be assumed that the source signals $\mathbf{s}_1$ and $\mathbf{s}_2$ are statistically independent, hence de-correlated, for each value of the parameter t. Furthermore, only one of the components can have a Gaussian distribution which represents the Gaussian noise that is assumed superimposed onto all signal channels. In addition, it is advantageous but not required when all rows of **X** are linearly independent i. e. at least the intensity and/or the phase (in complex data) of one of the source signals differ in all spectra.

For finding a unique mathematical solution of the problem, additional conditions have to be imposed that do not restrict the experimental data structure. Since multiples of the independent source vectors also represent valid solutions of eq. 2, the identified components should be whitened. This is achieved by two operations: the constant offset of the mixed signals is removed by subtracting the row-wise mean of **X** from the observation matrix (zeroing) and the variances of the individual source vectors (rows of **S**) of the solution are normalized to 1. Due to the whitening process, a new orthogonal mixing matrix $\tilde{\mathbf{A}}$ is obtained whose free parameters to be estimated are notably reduced from N^2 to N(N-1)/2. Instead of estimating any arbitrary full-rank matrix **A**, the orthogonal matrix $\tilde{\mathbf{A}}$ can be estimated more easily. However, note that the total transformation, including the whitening, is in general represented by a non-orthogonal matrix. The above restrictions help to solve the mathematical problem but do not solve the inherent ambiguity of the problem that the "true" complex amplitude of the source signals is not defined. However, under special conditions this amplitude factor can be calculated for a certain component. In NMR spectroscopic applications the amplitude can be obtained by fitting the calculated ICA component of interest to a spectrum whose intensity is unperturbed.

After calculating the mixing matrix $\tilde{\mathbf{A}}$, its inverse **U** is computed and the sources are obtained as follows:

$$\mathbf{S} = \begin{pmatrix} s_1^1 & s_1^2 & \cdots & s_1^N \\ s_2^1 & s_2^2 & \cdots & s_2^N \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix} \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^N \\ x_2^1 & x_2^2 & \cdots & x_2^N \end{pmatrix} = \mathbf{UX} \quad (3)$$

The de-mixing matrix **U** is usually found by optimizing cost functions that measure the non-Gaussianity (i. e. kurtosis and negentropy), the independence or the maximum likelihood of the extracted components.

Since there are several cost functions, different ICA methods have been developed such as the FastICA [42], the InfoMax [43] and the JADE [44] algorithms. In particular, the former has been used in this work since it allows a faster and more reliable learning procedure with a cubic convergence. In our experience, the FastICA algorithm produces the optimal decomposition within a smaller computational time than the other algorithms applied on the same dataset.

ICA gives more stable results when at least part of the component signals are separated somewhere in the dataset. This is the reason why generally the frequency domain NMR signal is preferred to the time-domain NMR signal. In the applications discussed in the following, the solvent signal needs to be separated from the solute signal. This reduces the number of sources m to be estimated to two. This condition only holds when proper experimental conditions (pulse sequences) are used. Otherwise, experimental artifacts have to be (at least partly) corrected by assuming more components. In SSA, usually the component with the largest eigenvalue is removed for solvent suppression, while in ICA the component containing the water signal in the center of the spectrum must be removed after a direct inspection of the data. In particular, ICA produces a permutable output with scaling and sign ambiguities, which must be evaluated directly by the user or by an adjunctive method for the automated recognition of the components. In the practical problems discussed below, the solvent signal is centered in the spectrum, whereas the solute signal is mainly located in other regions of the spectrum. This feature can be easily used for

the selection of the proper component and the calculation of the correct scaling.

In Fig. 2 the one-dimensional NMR data separation by ICA is compared with the removal of the solvent artifact by SSA. The ICA simply does not require all the previously described pre- and post-processing procedural steps typically used by the SSA algorithm, since it is applied directly in the frequency domain. For instance, the group delay points at the beginning of the FIDs must not be removed before the decomposition and the trajectory matrix containing shifted versions of the FID is not built. However, the exact number of those points belonging to the delay is implicitly calculated by the software TOPSPIN during the experimental acquisition, in order to apply a first order phase correction into the spectrum coherently with the group delay information without any user intervention. In such way, the undesired effect of arising wiggles in the frequency domain is avoided and the spectrum is not distorted. All the steps concerning the group delay management need to be taken in account only if dealing with time domain data, as in the SSA case. It implies that ICA is faster than SSA avoiding those pre-processing steps. The zero meaning and the whitening of the Fourier transformed signals are automatically applied by the FastICA algorithm. Moreover, it is assumed that the spectra have been already baseline and phase corrected before applying the ICA removal procedure, thus they do not need any further automated correction.

RESULTS AND DISCUSSION

Application of ICA to synthetic phase and intensity modulated spectra.

Since ICA needs as input at least as many experiments as the number of sources to be separated, one has to produce at least two 1D datasets where the two components (solute and solvent) have different weights (i. e. intensities and/or phases) in case of complex NMR data. Fig. 3a and Fig. 3b (zoom of the solvent region of Fig. 3a) show a synthetic dataset where a water artifact with different phases and/or intensities was mixed to a simulated spectrum of HPr protein from *Staphylococcus aureus* containing in addition artificial white noise (b traces). The water signal at 4.7 ppm is surrounded by dashed lines in Fig. 3.

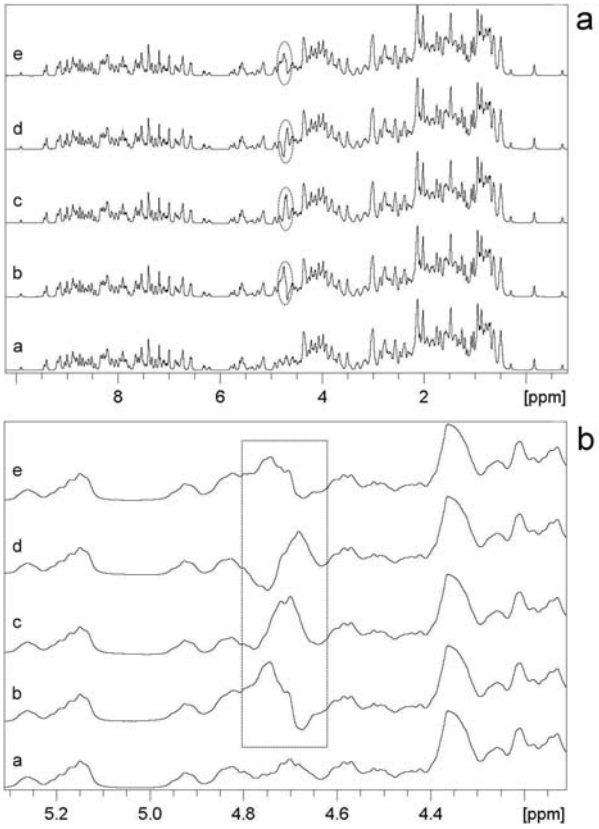

Fig. 3. Synthetic protein test spectra for application of SSA and ICA. Synthetic spectra were calculated by adding to a simulated spectrum of HPr (H15A) from *S. aureus* a weak experimental water signal (highlighted in the center of the spectra) with different phases and/or intensities. (a) complete spectra, (b) zoom of the spectra in the range of the solvent artifact. Description of the traces: (a) back-calculated spectrum without any solvent signal; (b) addition of an experimental water artifact whose intensity is approximately 50% of the strongest resonance of the protein; (c) as (b) but with an additional phase shift of the water signal of 90 degrees; (d) as (b) but with an additional phase shift of the water signal of 180 degrees; (e) as (b) where both the phase has been slightly modified by 2.6 degrees and the intensity has been additionally reduced by the 10%. Mixing time 0.15 s; spectral width 12.65 ppm, 2048 time and frequency domain data points, exponential window multiplication with a line broadening of 3 Hz.

The intensity of the water artifact has been set to 50% of the strongest spectral component, a condition where SSA starts producing spectral artifacts. As discussed above, the ICA requires at least the same number of different spectra as the number of independent components in the spectra.

In the simplest case, only two components have to be separated, the water artifact and the signal of the compound (e. g. a protein) to be analyzed. However, for this goal NMR experiments have to be used that influence the intensities and/or phases of these two components separately (see Fig. 1) in order to obtain the desired independent components. If the used experiments also influence the signals of individual atoms or groups of the compound(s) contained in the samples in a non-uniform way, the number of spectra needed for ICA must be increased correspondingly.

ICA assuming two components has been used for recovering the original HPr spectrum depicted in Fig. 3 (a traces) and in Fig. 4 (a traces). The complete simulated dataset as well as a zoom of the solvent signal region is shown in Fig. 3. An experimental water signal has been added to the simulated protein spectrum, as shown in Fig. 3 (b traces). The phase of the additional water signal has been changed by 90° (Fig. 3, c traces) and by 180° (Fig. 3, d traces) before mixing it with the artificial protein spectrum. In Fig. 3, e traces, both the intensity and the phase of the solvent artifact have been changed. The results of an application of ICA or SSA to the dataset are shown in Fig. 4a and their zoom in the artifact region is reported in Fig. 4b. The Fig. 4 (b traces) shows the result of ICA when a 90° phase shift is applied to the second spectrum (ICA applied to the spectra shown in Fig. 3, b and c traces). Fig. 4 (c traces) shows the result of ICA when a large phase shift of 180° is produced (ICA applied to the spectra shown in Fig. 3, b and d traces). In Fig. 4 (d traces) is reported the result of ICA when both intensity (reduced by the 10%) and phase (2.6°) are modified by the pulse sequence (ICA applied to the spectra shown in Fig. 3, b and e traces). Finally, in Fig. 4 (e traces) the result produced by the SSA applied to the spectrum described in Fig. 3 (b traces) is reported. The investigated cases clearly demonstrate that ICA can recover almost perfectly the unperturbed signal when the two original spectra contain a relatively weak solvent signal (that is unfavorable for SSA). The optimal recovering was obtained by 180° shift of the solvent signal in the second spectrum or by a moderate intensity variation. SSA gives a clearly inferior recovery of the signals superposed by the solvent signal.

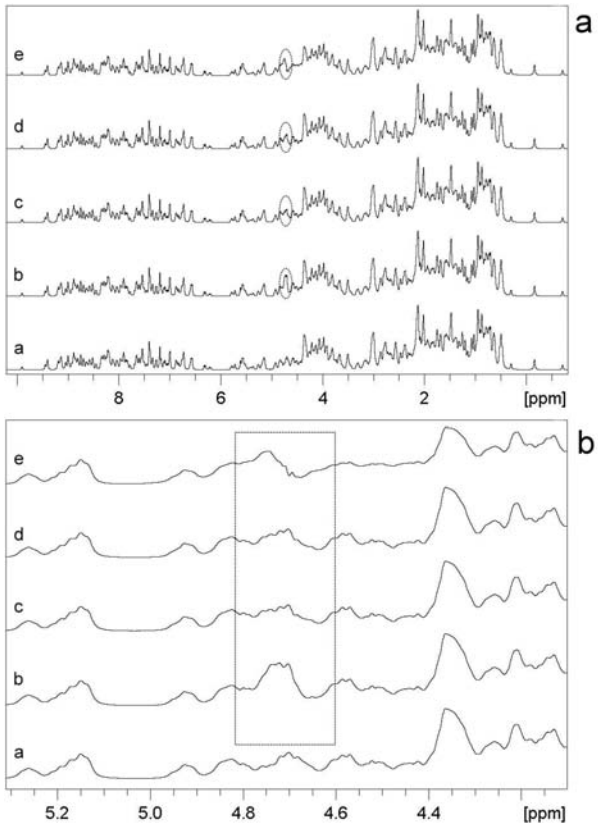

Fig. 4. Application of SSA and ICA to a synthetic one-dimensional dataset of the HPr protein from *S. aureus* (H15A). ICA and SSA have been applied to the spectra shown in Fig. 3. (a) complete spectra, (b) zoom of the solvent region. Description of the traces: (a) original back-calculated spectrum without any solvent signal; (b) ICA assuming two sources applied to the two spectra shown in Fig. 3, b and c traces; (c) ICA applied to the two spectra shown in Fig. 3, b and d traces; (d) ICA used on the two spectra shown in Fig. 3, b and e traces; (e) SSA with an embedding of 20 components applied to the spectrum shown in Fig. 3, b trace. The residual water signal is highlighted.

When the differences (of the phases of the solvent signals) between the two spectra used are only very small, the results are not as good as shown in Fig. 4. Here, the use of more than two datasets as input can help. For example, pure protein spectra with a quality of those shown in Fig. 4 can be obtained when three spectra with small angle variations of the water resonance phase of 0, 1.2 and 1.6 degrees are used for ICA (data not shown). When instead the solvent signals of the two spectra differ only by the intensities, even a

small variation produces an almost optimal recovery of the resonances of interest. In general, experimentally either large phase variations or small intensity change of the components to be separated should be generated by the pulse sequence in question whenever possible. Only when this is not suitable, more than two spectra should be used to separate the components.

The result is different when a complex SSA is applied to these spectra. SSA needs as input only one time domain signal. As an example the time domain signal corresponding to the spectrum depicted in Fig. 3 (b traces) was used with a SSA embedded in a 20 dimensional space. Although the water signal is almost completely removed, as demonstrated in Fig. 4 (e traces), the resonances under the water signal are not optimally recovered and some of the neighboring signals are distorted. Often SSA produces also artifacts in other parts of the spectrum when the water signal is not the strongest signal in the spectrum [28]. Such an effect has not been observed in our example, whose water intensity was the half of the strongest protein resonance in the spectrum. A major reduction of the intensity of the solvent signal (as shown in Fig. 3, e traces) would produce such artifacts, destroying the strongest parts of the protein spectrum (data not shown). In all other cases (c and d traces of Fig. 3), the SSA produced similar results to those ones shown here in Fig. 4 (e traces).

SSA is a powerful method when a strong water signal is present [27, 28], but also in such favorable cases it can result in a too strong solvent suppression and spectral distortions in the vicinity of the water signal. This can happen when an inadequate number of components are extracted and are successively nullified. We tested that in a spectrum corresponding to that one shown in Fig. 5 (c trace) where the water signal was 100-times stronger than the strongest protein signal in the spectrum. Fig. 5 shows again that SSA (a trace) as well as ICA (b trace) produce a strong reduction of the water signal around 4.7 ppm. For the application of ICA a second spectrum was generated where the intensity of the water signal was reduced by 30% but was still 70-times stronger than the strongest resonance of the protein. SSA with a default embedding of 20 components leads

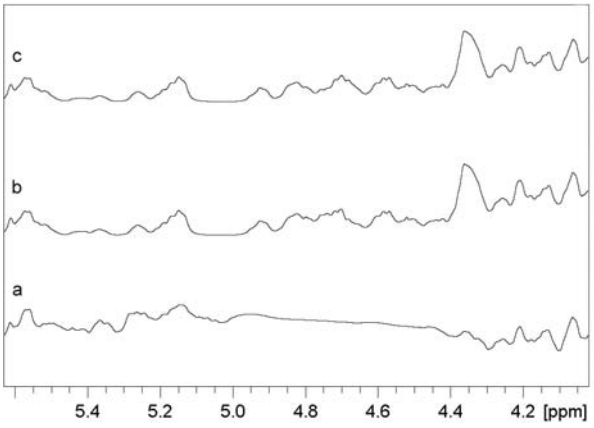

Fig. 5. Application of SSA and ICA to a spectrum containing a strong solvent artifact. Part of a one-dimensional NOESY-type NMR spectrum of the HPr protein from *S. aureus* (H15A) back-calculated from the 3D-structure and the known chemical shifts with RELAX-*JT*2. A water signal that was 100-times more intense than the strongest protein signal and Gaussian noise were added to the synthetic data before Fourier transformation. In a second spectrum a water signal that was reduced by 30% was added. (a) SSA applied to the first spectrum, (b) ICA applied to the two spectra, (c) the back-calculated spectrum without any solvent addition.

to a complete removal of the solvent signal at 4.7 ppm (Fig. 5a) and influences also the resonances in the vicinity. They are either completely suppressed or distorted. An example is the protein resonance at 4.3 ppm (Fig. 5c) that is completely suppressed. Increasing the number of extracted and nullified components could improve the performance of SSA. As demonstrated recently [28], the optimal removal of the solvent and recovery of signals close to the solvent resonance by SSA is reached when the solvent artifact is approximately twice as strong as the most intense solute signal. In contrast, ICA applied to the corresponding set of spectra results in a perfect recovery of all protein resonances (Fig. 5b) that were completely superposed by the strong water resonance.

The performance of ICA was tested systematically for the synthetic dataset described above by variation of the intensity and/or phase of the solvent signal added to the protein spectrum (Fig. 6). Firstly, a synthetic spectrum of the HPr protein was added to an experimental solvent signal

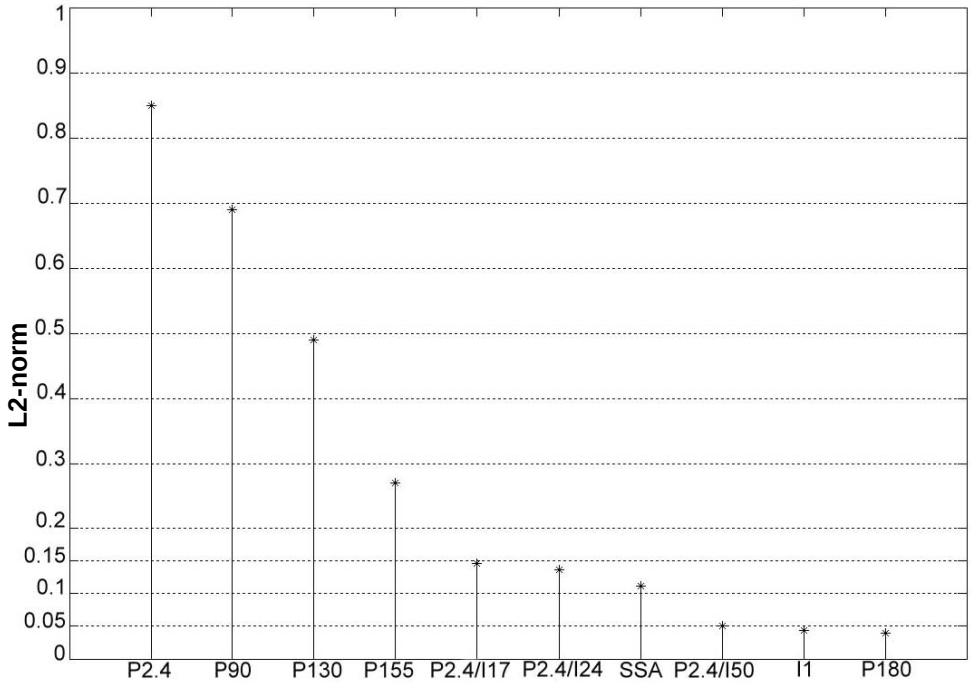

Fig. 6. Dependence of the performance of ICA and SSA on the type of inputs. Synthetic datasets were created as described in Fig. 3 but phases and intensities were varied over a wide range of conditions. As a measure for the performance the L^2-norm was calculated for the difference between the pure protein spectrum and the spectrum obtained after the addition of a solvent signal that was twice intense than the strongest protein signal (first spectrum). The L^2-norm of the point-wise differences of these two spectra was arbitrarily set to 1. The result after application of SSA is labeled as (SSA). ICA has been applied to a dataset consisting of the first spectrum and additional spectra where the phase (P) and/or the intensity (I) were varied. (P*x*) means that the phase of the water signal was shifted relatively to the first spectrum by *x* degrees. (I*x*) means that the intensity of the water signal was reduced by *x*% of the intensity in the first spectrum.

whose maximum intensity was twice larger than that one of the strongest protein signal which represents the optimal case for solvent removal by SSA [28]. As a measure of the performance, the L^2-norm of the point-wise difference between that protein spectrum obtained before the solvent suppression and the original simulated spectrum (without any experimental water) was used. The norm of this spectrum was set to 1. As a bench mark SSA is able to reduce the L^2-norm by approximately 89% (SSA in Fig. 6). Quantitatively, similar results can be obtained when a second spectrum is used by the ICA where either the phase of the water signal is strongly shifted (in the range between 160 and 180 degrees) or a small phase change (i. e. 2.4 degrees) of the solvent signal is accompanied by a remarkable intensity reduction (in the range of about 24% and 50% of the initial solvent signal (P2.4-I24 and P2.4-I50 in Fig. 6) that correspond respectively to a ratio of 1.64 and 1 between the solvent and the strongest solute signals). However, in the ICA cases, close to the resonance frequency of the solvent, the protein spectrum is much better reconstructed, analogously to the data shown above. A notably better result is obtained when the water phase in the second spectrum is shifted by 180 degrees (P180 in Fig. 6). Here a reduction by 96% is reached. Similar results are obtained with different scenarios, e. g. when only the intensity of the water signal in the second spectrum is slightly reduced, even just by 1% of the initial solvent signal (I1 in Fig. 6).

Inspecting the results obtained from the simulated dataset, one can conclude that ICA is a powerful method to separate the water artifact from the true signal when ideal conditions are met, namely that the spectra used are composed of two components

with mutual phase and/or intensity modulation but no solute signal variations. If the solvent signal is rather weak, the wanted signal(s) can be virtually completely recovered (Fig. 4, c and d traces). In contrast, in case of weak solvent signals SSA leads to spectral distortions.

Application of ICA to experimental spectra with phase and intensity modulation

In experimental spectra the ideal conditions usually cannot be established perfectly and the result is expected to depend on the sample as well as on the pulse sequence used. The pulse sequence has to manipulate the components to be separated in a different way. This can only be done when the corresponding spin systems are characterized by different NMR parameters. One possibility is the use of differences in longitudinal relaxation times T_1 between solvent and solute. A simple way to use this property is the application of the typical NOESY-type pulse sequence that is generally used in metabolomic applications of NMR (Fig. 1), with different mixing times. At high Q-values of the receiver coils the water signal relaxes much faster than the solute signal because of radiation damping and can therefore be separated from the solute signal. This effect should be especially prominent for short mixing times. Fig. 7 shows the results of SSA and ICA applied to experimental NOESY-type 1D spectra of small molecules recorded with different mixing times. One has to mention that the urine spectra are used here only as an example for testing the performance and side effects of the two methods. In practice, in urine spectra the signals of the metabolites are rather strong compared to the water signal suppressed by the NOESY-type sequence. Additional data post-processing usually is not required here.

The phase and intensity of the water signal strongly depend on the mixing time. ICA was applied to the pair of spectra with short mixing times and to the pair with long mixing times. In addition, SSA has been applied separately on the experiments with a mixing time of 10 ms and of 1000 ms, with M= 40. In the experimental dataset with the short mixing times, the water signal is still much stronger than the remaining signals. Here, SSA provides a strong suppression of the solvent signal (Fig. 7a). The most visible disadvantage of SSA is again that the signals very close to the water resonance are remarkably attenuated. In the dataset with the longer mixing times (Fig. 7b) the water signal in the original spectrum with a mixing time of 1 s is not the strongest signal anymore. Therefore, SSA erroneously removes the strongest solute signals. In both cases, ICA leads to a strong reduction of the solvent signal without any distortion or attenuation of the wanted signal and clearly represents the superior method. When ICA is used to a pair of spectra with a short mixing time and a very long mixing time, no satisfactory results are obtained. This is probably due to the fact that the long selective irradiation during the mixing time strongly modifies the shape of the solvent signal compared to the shape obtained at the short mixing time and thus destroys the precondition for a successful application of ICA. Instead of varying the mixing time it is also possible to create two spectra with solvent signals with different signs by using different phase cycles (see Fig. 1a). Theoretically, this method creates ideal conditions for ICA but field inhomogeneities create regions where radiation damping is small. In these parts, the required inversion of signs does not work. The obtained results are comparable to those shown in Fig. 7.

For many applications the most important property of ICA is that it can recover resonances that are hidden below the water artifact when a proper pulse sequence is used. We have already shown that ICA spectra work almost perfectly for ideal synthetic datasets (Fig. 5). For obtaining similar conditions, in protein NMR spectroscopy, diffusion weighting (Fig. 1b) turned out to be the optimal method to obtain spectra with differential attenuation of the remaining water signal. As an example the experimental diffusion weighted NOESY-type spectra of HPr from *S. carnosus* were recorded (Fig. 8). The insert in Fig. 8a shows the water signal in the two spectra used for ICA that is differently attenuated. SSA strongly reduces the water signal and leads to a clear improvement of the spectral quality. However, protein resonances close to the water resonance are not visible. In contrast, the protein signals superposed by the water resonance can be recovered almost completely by ICA (Fig. 8b). A simulation of the spectrum from the known

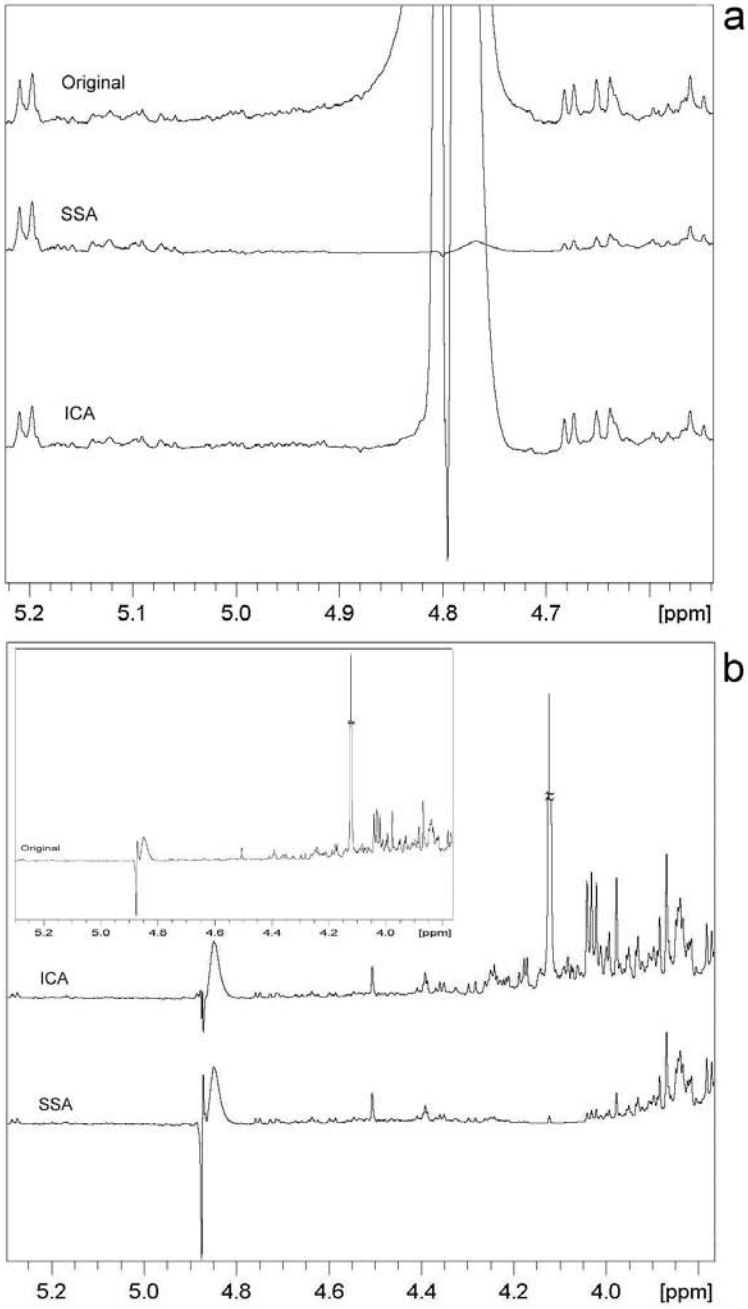

Fig. 7. Separation of solute and solvent signal by radiation damping and selective saturation. One-dimensional human urine spectra recorded with a NOESY-type pulse sequence (see Fig. 1a) with different mixing times of 10 ms, 20 ms, 0.8 s, and 1 s. Relaxation delay, 5 s; 128 K complex time domain data points; 64 K frequency domain data points; DQD acquisition mode; 144 time domain group delay data points; spectral width, 20.03 ppm; 128 scans. The water signal was saturated by a selective presaturation pulse of 5 s and an additional short saturation pulse during the mixing time. (a) ICA applied to the pair of spectra with short mixing times (10 ms and 20 ms) and SSA (embedding M = 40) on the experiment with a mixing time of 10 ms. (b) ICA applied to the pair with long mixing times (0.8 s and 1 s) and SSA used on the spectrum with a mixing time of 1000 ms, with M= 40. The insert shows part of the original spectrum measured with a mixing time of 1 s. During this time the water signal has been saturated by selective irradiation.

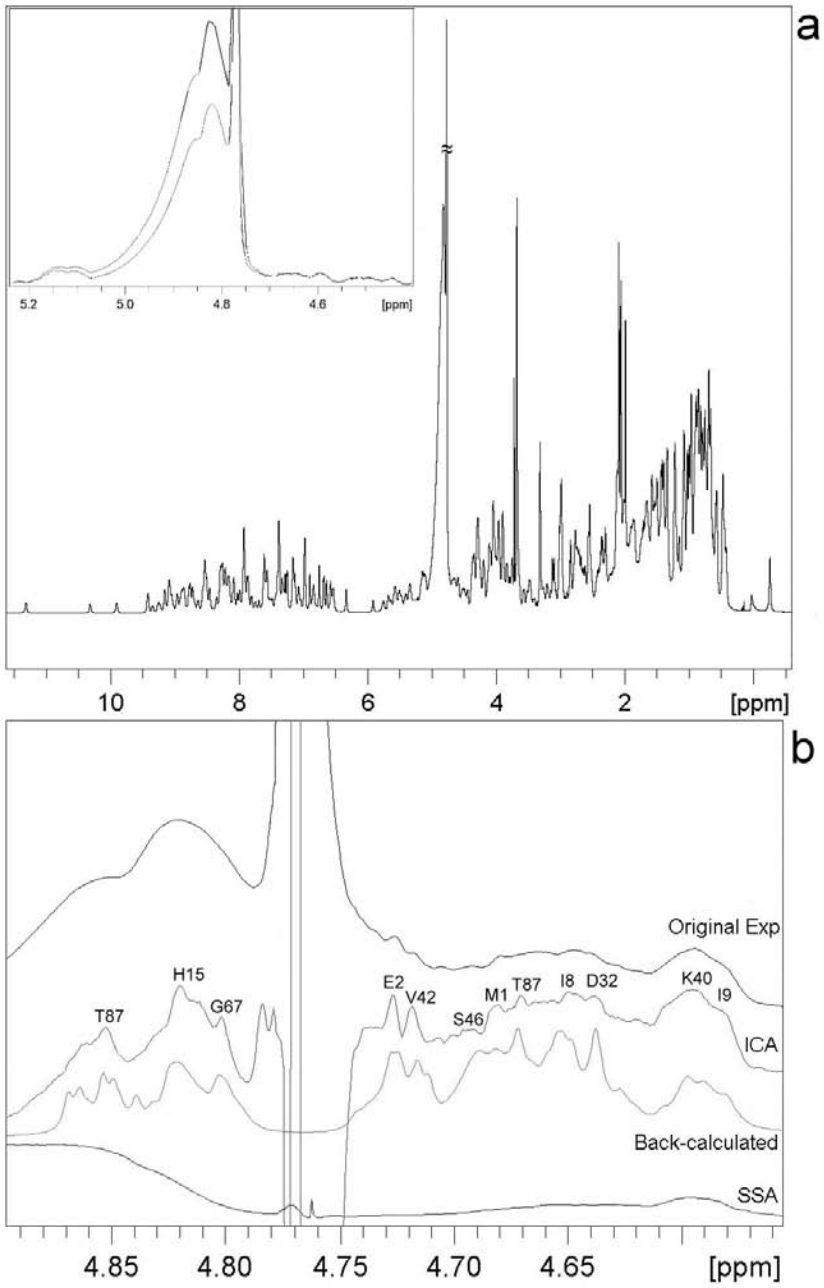

Fig. 8. ICA and SSA application on the ICA-tailored one-dimensional experimental spectra of HPr protein. The sample contained 1 mM HPr protein from *Staphylococcus carnosus* in 95% H_2O/5% D_2O, pH 7. Spectra were measured with the pulse sequence shown in Fig. 1b. (a) Original spectra. Gradient weights G_2: 80 G cm^{-1}, (the solvent artifact shown in lower trace of the insert) and 50 G cm^{-1}, (the solvent artifact shown in the upper trace of the box); spectral width, 14.9 ppm; 32,768 complex time domain data points. (b) Spectral zoom of the artifact region before (original) and after application of SSA/ALS (SSA) to the first spectrum (gradient strength 80 G cm^{-1}) and after ICA applied to both spectra. Before application of ICA, the time domain data were filtered by an exponential multiplication corresponding to an additional line width of 0.3 Hz, a phase correction was applied and the baseline was corrected by ALS. The back calculated spectrum was simulated with the experimental parameters by AUREMOL-RELAX-*JT2* [38] from the published assignments and the three-dimensional structure of the HPr protein [40].

structure and chemical shifts by AUREMOL-RELAX-*JT2* [38] is almost identical to that obtained after application of ICA. The assignment of the recovered resonances to the known resonance frequencies shows that in fact a large part of the H^{α}-resonances can be detected. The quality of the spectrum after application of ICA corresponds to that one recorded in D_2O, but here no exchange of normal water by heavy water was necessary and the resonances of the amide protons are still visible. These latter would disappear in D_2O.

CONCLUSIONS

In this work, two post-processing methods (SSA and ICA) for the removal/attenuation of the solvent signal in NMR spectra are described. Both give good results, but have different applications and advantages. The SSA is applied on the time domain signal (FID) but it needs some pre-processing steps (as digitally filtered data management) and some post-processing steps (as baseline and phase correction). It can be easily applied to spectra of any dimensionality in an automated procedure, as it is implemented in AUREMOL/SSA-ALS [27, 28]. It can be applied to any type of NMR spectra. The major limitation is that the solvent signal must be clearly more intense than any other resonance in the spectrum. If this condition is not fulfilled spectral artifacts are produced as distortions of the resonances of interest close to the solvent signal and suppression of the strongest resonances in the solute spectrum. Additionally, even when the solvent signal is more intense than any other solute signal, SSA results in a very strong solvent suppression leading to the attenuation of the protein resonances located in the artifact region. This effect could be reduced varying the embedding dimensions: increasing the number of extracted and nullified components. In this case the automation would get difficult. ICA has demonstrated to overcome those limitations. It requires as input more than one experiment with different solvent or solute signal intensities and/or phases. Optimal results are obtained when the line shapes of the solvent and the solute resonances are unperturbed but when their relative intensities are different in the analyzed spectra. Thus, the experimental scheme (pulse sequence) has to be optimized for an appropriate application of ICA. If this is the case, resonances below the water signals can be detected with high quality. The AUREMOL-ICA is actually an ongoing project that has been thought not only for automated solvent suppression, but also for other purposes as metabolites quantification in diffusion-weighted spectra of biofluids.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft and the Bayerische Forschungsstiftung. We are thankful to A. M. Tomé for her advices and valuable discussions.

REFERENCES

1. Aranibar, N., Borys, M., Mackin, N. A., Ly, V., Abu-Absi, N., Abu-Absi, S., Niemitz, M., Schilling, B., Jian Li, Z., Brock, B., Russell II, R. J., Tymiak, A. and Reily, M. D. 2011, J. Biomol. NMR, 49(3-4), 195.
2. Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C. and Nicholson, J. K. 2007, Nat. Prot., 2(11), 2692.
3. Wuethrich, K. 1986 NMR of Proteins and Nucleid Acids, Wiley.
4. Cavanagh, J., Fairbrother, W. J., Palmer III, A.G. and Skelton, N. J. 1996, Protein NMR Spectroscopy: Principles and Practice, Academic Press.
5. Gronwald, W. and Kalbitzer, H. R. 2004, Progr. NMR Spectr., 44, 33.
6. Plateau, P. and Gueron, M. 1982, J. Am. Chem. Soc., 104(25), 7310.
7. Smallcombe, S. H., Patt, S. L. and Keifer, P. A. 1995, J. Magn. Reson. Ser. A, 117, 295.
8. Piotto, M., Saudek, V. and Sklenar, V. 1992, J. Biomol. NMR 2(6), 661.
9. Sklenar, V., Piotto, M., Leppik, R. and Saudek, V. 1993, J. Magn. Reson. A, 102, 241.
10. Hwang, T. L. and Shaka, A. J. 1995, J. Magn. Reson. A, 112, 275.
11. Simpson, A. J. and Brown, S. A. 2005, J. Magn. Reson., 175(2), 340.
12. Kuroda, Y., Wada, A., Yamazaki, T. and Nagayama, K. 1989, J. Magn. Reson., 84, 604.
13. Marion, D., Ikura, M. and Bax, A. 1989, J. Magn. Reson., 84, 425.
14. Mitschang, L., Neidig, K. P. and Kalbitzer, H. R. 1990, J. Magn. Reson., 90, 359.

15. Barache, D., Antoine, J. P. and Dereppe, J. M. 1997, J. Magn. Reson., 128, 1.
16. Antoine, J. P., Coron, A. and Dereppe, J. M. 2000, J. Magn. Reson., 144, 189.
17. Guenther, U. L., Ludwig, C. and Rueterjans, H. 2002, J. Magn. Reson., 156, 19.
18. Adler, M. and Wagner, G. 1991, J. Magn. Reson., 91, 450.
19. Tsang, P., Wright, P. E. and Rance, M. 1990, J. Magn. Reson., 88, 210.
20. Mitschang, L., Cieslar, C., Holak, T. A. and Oschkinat, H. 1991, J. Magn. Reson., 92, 208.
21. Hardy, J. K. and Rinaldi, P. L. 1990, J. Magn. Reson., 88, 320.
22. Brown, D. E. and Campbell, T. W. 1990, J. Magn. Reson., 89, 255.
23. Pijnapple, W. W. F., Van Den Boogaart, A., De Beer, R. and Van Ormondt, D. 1992, J. Magn. Reson., 97, 122.
24. Zhu, G., Smith, D. and Hua, Y. 1997, J. Magn. Reson., 124, 286.
25. Stadlthanner, K., Tomé, A. M., Theis, F. J., Lang, E. W., Gronwald, W. and Kalbitzer, H. R. 2006, Neurocomp., 69, 497.
26. Boehm, M., Stadlthanner, K., Gruber, P., Theis, F. J., Lang, E. W., Tomé, A. M., Teixeira, A. R., Gronwald, W. and Kalbitzer, H. R. 2006, IEEE Trans. Biom. Eng., 53(5), 810.
27. Malloni, W. M., De Sanctis, S., Tomé, A. M., Lang, E. W., Munte, C. E., Stadlthanner, K., Neidig, K. P. and Kalbitzer, H. R. 2010, J. Biomol. NMR., 47(2), 110.
28. De Sanctis, S., Malloni, W. M., Kremer, W., Tomé, A. M., Lang, E. W., Neidig, K. P. and Kalbitzer, H. R. 2011, J. Magn. Res., 210(2), 177.
29. Ghil, M., Allen, M. R., Dettinger, M. D. and Die, K. 2002, Rev. Geoph., 40(1), 1003.
30. Comon, P. 1994, Sign. Proces., 36(3), 287.
31. Hyvaerinen, A., Karhunen, J. and Oja, E. 2001, Independent component analysis, Wiley.
32. Makeig, S., Jung, T. P., Bell, A. J. and Sejnowski, T. J. 1996, Adv. Neur. Inf. Proc. Sys., 8, 145.
33. Bell, A. J. and Sejnowski, T. J. 1997, Vis. Res., 37(23), 3327.
34. Bell, A. J. and Sejnowski, T. J. 1995, Neur. Comput. 7, 1129.
35. Jolliffe, I. T. 1986, Principal component analysis, Springer-Verlag.
36. De Sanctis, S. 2011, PhD thesis, University of Regensburg, Germany.
37. Harsch. T., Donaubauer, H., Malloni, W., De Sanctis, S., Neidig, K. P. and Kalbitzer, H. R. 2011, EUROMAR, Frankfurt am Main.
38. Ried, A., Gronwald, W., Trenner, J. M., Brunner, K., Neidig, K. P. and Kalbitzer, H. R. 2004, J. Biomol. NMR, 30(2), 121.
39. Maurer, T., Meier, S., Kachel, N., Munte, C. E., Hasenbein, S., Koch, B., Hengstenberg, W., Kalbitzer, H. R. 2004, J. Bacteriol., 186(17), 5906.
40. Kalbitzer, H. R., Görler, A., Li, H., Dubovski, P. V., Hengstenberg, W., Kowolik, C., Hiroaki, Y. and Akasaka, K. 2000, Prot. Sci., 9(3),693.
41. Moskau, D. 2002, Conc. Magn. Reson., 15(2), 164.
42. Hyvaerinen, A. and Oja, E. 1997, Neur. Comp., 9(7), 1483.
43. Lee, T. W. and Sejnowski, T. J. 1996, Proc. 4th Joint Symp. Neur. Comp., 7, 132.
44. Cardoso, J. F. 1999, Neur. Comp., 11(1), 157.