Application of Singular Spectrum Analysis (SSA), Independent Component Analysis (ICA) and Empirical Mode Decomposition (EMD) for automated solvent suppression and automated baseline and phase correction from multi-dimensional NMR spectra

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER NATURWISSENSCHAFTLICHEN FAKULTÄT III
BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

vorgelegt von

Silvia De Sanctis

aus Fermo, Italy

im Jahr 2011

Promotionsgesuch eingereicht am: 22.12.2011

Die Arbeit wurde angeleitet von: Prof. Dr. Dr. Hans Robert Kalbitzer

Prüfungsausschuß

Vorsitzender: Prof. Dr. Thomas Dresselhaus
Erstgutachter: Prof. Dr. Dr. Hans Robert Kalbitzer
Zweitgutachter: Prof. Dr. Elmar Wolfgang Lang
Drittprüfer: Prof. Dr. Rainer Merkl

Abstract

A common problem on protein structure determination by NMR spectroscopy is due to the solvent artifact. Typically, a deuterated solvent is used instead of normal water. However, several experimental methods have been developed to suppress the solvent signal in the case that one has to use a protonated solvent or if the signals of the remaining protons even in a highly deuterated sample are still too strong. For a protein dissolved in 90% H_2O / 10% D_2O, the concentration of solvent protons is about five orders of magnitude greater than the concentration of the protons of interest in the solute. Therefore, the evaluation of multi-dimensional NMR spectra may be incomplete since certain resonances of interest (e.g. H^{α} proton resonances) are hidden by the solvent signal and since signal parts of the solvent may be misinterpreted as cross peaks originating from the protein. The experimental solvent suppression procedures typically are not able to recover these significant protein signals. Many post-processing methods have been designed in order to overcome this problem.

In this work, several algorithms for the suppression of the water signal have been developed and compared. In particular, it has been shown that the Singular Spectrum Analysis (SSA) can be applied advantageously to remove the solvent artifact from NMR spectra of any dimensionality both digitally and analogically acquired. In particular, the investigated time domain signals (FIDs) are decomposed into water and protein related components by means of an initial embedding of the data in the space of time-delayed coordinates. Eigenvalue decomposition is applied on these data and the component with the highest variance (typically represented by the dominant solvent signal) is neglected before reverting the embedding. Pre-processing (group delay management and signal normalization) and post-processing (inverse normalization, Fourier transformation and phase and baseline corrections) of the NMR data is mandatory in order to obtain a better performance of the suppression. The optimal embedding dimension has been empirically determined in accordance to a specific qualitative and quantitative analysis of the extracted components applied on a back-calculated two-dimensional spectrum of HPr protein from *Staphylococcus aureus*.

Moreover, the investigation of experimental data (three-dimensional $^{1}H^{13}C$ HCCH-TOCSY spectrum of Trx protein from *Plasmodium falciparum* and two-dimensional NOESY and TOCSY spectra of HPr protein from *Staphylococcus aureus*) has revealed the ability of the algorithm to recover resonances hidden underneath the water signal.

Pathological diseases and the effects of drugs and lifestyle can be detected from NMR spectroscopy applied on samples containing biofluids (e.g. urine, blood, saliva). The detection of signals of interest in such spectra can be hampered by the solvent as well. The SSA has also been successfully applied to one-dimensional urine, blood and cell spectra.

The algorithm for automated solvent suppression has been introduced in the AUREMOL software package (AUREMOL_SSA). It is optionally followed by an automated baseline correction in the frequency domain (AUREMOL_ALS) that can be also used out the former algorithm. The automated recognition of baseline points is differently performed in dependence on the dimensionality of the data.

In order to investigate the limitations of the SSA, it has been applied to spectra whose dominant signal is not the solvent (as in case of watergate solvent suppression and in case of back-calculated data not including any experimental water signal) determining the optimal solvent-to-solute ratio.

The Independent Component Analysis (ICA) represents a valid alternative for water suppression when the solvent signal is not the dominant one in the spectra (when it is smaller than the half of the strongest solute resonance). In particular, two components are obtained: the solvent and the solute. The ICA needs as input at least as many different spectra (mixtures) as the number of components (source signals), thus the definition of a suitable protocol for generating a dataset of one-dimensional ICA-tailored inputs is straightforward.

The ICA has revealed to overcome the SSA limitations and to be able to recover resonances of interest that cannot be detected applying the SSA. The ICA avoids all the pre- and post-processing steps, since it is directly applied in the frequency domain. On the other hand, the selection of the component to be removed is automatically detected in the SSA case (having the highest variance). In the ICA, a visual inspection of the extracted components is still required considering that the output is permutable and scale and sign ambiguities may occur.

The Empirical Mode Decomposition (EMD) has revealed to be more suitable for automated phase correction than for solvent suppression purposes. It decomposes the FID into several intrinsic mode functions (IMFs) whose frequency of oscillation decreases from the first to the last ones (that identifies the solvent signal). The automatically identified non-baseline regions in the Fourier transform of the sum of the first IMFs are separately evaluated and genetic algorithms are applied in order to determine the zero- and first-order terms suitable for an optimal phase correction.

The SSA and the ALS algorithms have been applied before assigning the two-dimensional NOESY spectrum (with the program KNOWNOE) of the PSCD4-domain of the pleuralin protein in order to increase the number of already existing

distance restraints. A new routine to derive $^{3}J_{H^N}{}_{H^\alpha}$ couplings from torsion angles (Karplus relation) and vice versa, has been introduced in the AUREMOL software. Using the newly developed tools a refined three-dimensional structure of the PSCD4-domain could be obtained.

Acknowledgments

First and foremost I wish to express my sincerest gratitude to my Ph.D. supervisor Prof. Dr. Dr. Hans Robert Kalbitzer who supported me with his patience and vast knowledge.

I wish especially to thank the Prof. Dr. Elmar W. Lang for his many useful suggestions, his friendly company and complete disposal to encourage me to overcome both scientific and personal problems.

Prof. Dr. Werner Kremer, Prof. Dr. Claudia E. Munte and Prof. Dr. Wolfram Gronwald have helped me several times with the intent to answer to my scientific questions in an exhaustive way.

My deepest thanks are addressed to my family and in particular to my father that since the beginning of this project has fought with a bad disease and has never thought to put his discomfort over my freedom of living abroad far from home.

My husband has always encouraged me to work on this project and he is the reason why I dare to live.

De Sanctis Silvia

List of Tables

List of Figures

List of Abbreviations

AUREMOL-SSA/ALS	**AUREMOL S**ingular **S**pectrum **A**nalysis / **A**utomatic **L**inear **S**pline
NMR	**N**uclear **M**agnetic **R**esonance
FID	**F**ree **I**nduction **D**ecay
SSA	**S**ingular **S**pectrum **A**nalysis
PCA	**P**rincipal **C**omponent **A**nalysis
ICA	**I**ndependent **C**omponent **A**nalysis
SVD	**S**ingular **V**alue **D**ecomposition
KLT	**K**arhunen-**L**oeve **T**rasformation
EMD	**E**mpirical **M**ode **D**ecomposition
GA	**G**enetic **A**lgorithm
ALS	**A**utomatic **L**inear **S**pline
FIR	**F**inite **I**mpulse **R**esponse
FFT	**F**ast **F**ourier **T**ransformation
DECIM	**Decim**ation Factor
DSPFVS	**D**igital **Sp**ectrometer **V**ers**i**on
DW	**D**well **T**ime
DWOV	**O**versampling **D**well **T**ime
PKNL	Nonlinear phase correction
FnMODE	Acquisition mode for 2D, 3D, etc.
TD	Size of the FID
SW	**S**pectral **W**idth
PHC0	0^{th} order phase correction

PHC1	1st order phase correction
MC2	Acquisition mode for 2D, 3D, etc.
SVM	**S**upport **V**ector **M**achine
GRPDLY	**Gr**ou**p De**l**ay**
BC_mode	**B**aseline **C**orrection **mode**
RMSD	**R**oot **M**ean **S**quare **D**eviation

1 INTRODUCTION

1.1 NMR spectroscopy

1.1.1 PRINCIPLES OF NUCLEAR MAGNETIC RESONANCE

Nuclei having a nonzero nuclear spin quantum number $\vec{I}$ act like rotating charges whose nuclear spin angular momentum $\vec{J}$ generates a small nuclear magnetic moment $\vec{\mu}$. The isotopes ^{1}H, ^{13}C, ^{15}N and ^{2}H are the most used in NMR spectroscopy. If a static external magnetic field $\overrightarrow{B_0}$ is applied they align themselves in discrete states, namely they are positioned either with (lowest energy) or against it (highest energy) in accordance to the magnetic quantum number m_I. They precess around the magnetic field with an angular frequency ω_0 (Larmor frequency) which is proportional to $\overrightarrow{B_0}$ and to γ, the gyromagnetic ratio (see eq. 1.1). This latter defines the strength of the nuclear magnet field and it is different for each isotope.

$$\omega_0 = \gamma B_0 \tag{1.1}$$

At equilibrium the lower energy orientation of $\vec{\mu}$ (parallel to $\overrightarrow{B_0}$) is the more probable. Typically the spin population difference between the states is very small implying a very weak sensitivity of such technique. For instance, a suitable NMR sample must contain pure material in a larger order with respect to other methods as the mass spectrometry. The material concentration is directly proportional to the intensity of the observed signals. Considering that the energy gap is proportional to $\overrightarrow{B_0}$ (in accordance to eq. 1.2), the use of spectrometers operating at higher magnetic fields increases the population difference and more intense signals can be obtained with a consequent higher general sensitivity.

$$\Delta E = h\omega_0 \tag{1.2}$$

Applying an oscillating magnetic field $\overrightarrow{B_1}$ perpendicular to $\overrightarrow{B_0}$ as a 90° pulse of some microseconds produces a transverse magnetization.

Every perturbed spin tends to regain the equilibrium state relaxing back to the original condition. The population difference is exponentially restored through a longitudinal (spin-lattice) relaxation in a time T_1, where the spins re-align themselves along B_0. They also lose the precession coherence on the transverse magnetization plane (spin-spin relaxation) in a time T_2, where generally T_2 is shorter (equal) than T_1. This latter governs the acquisition rate of the signal where a shorter T_1 corresponds to faster acquisitions.

The transverse relaxation generates an exponentially decaying time domain signal known as FID (free induction decay). It must be Fourier transformed in order to obtain the NMR spectrum. The relaxation for population equilibrium restoring determines the necessary waiting time before repeating the experiment, the decay of the FID and consequently the total line width $\Delta\nu_{\frac{1}{2}}$ of the signals at half height in the frequency domain, according to eq. 1.3:

$$\Delta\nu_{\frac{1}{2}} = \frac{1}{\pi T_2} \tag{1.3}$$

Therefore, shorter T_2 correspond to broader line widths.

Both relaxation times (T_1 and T_2) are strictly related to the translational and to the rotational correlation times τ_c (the time to diffuse one diameter) and τ_r (the time to rotate one radian). The former, as described in eq. 1.4, is strictly related to the diffusion D. The latter (eq. 1.5) depends on the molecular mass (r is the molecular radius, η defines the viscosity of the solution, k_B is the Boltzmann constant and T represents the absolute temperature):

$$\tau_c = \frac{6\pi\eta D}{k_B T} \tag{1.4}$$

$$\tau_r = \frac{4\pi\eta r^3}{3k_B T} \tag{1.5}$$

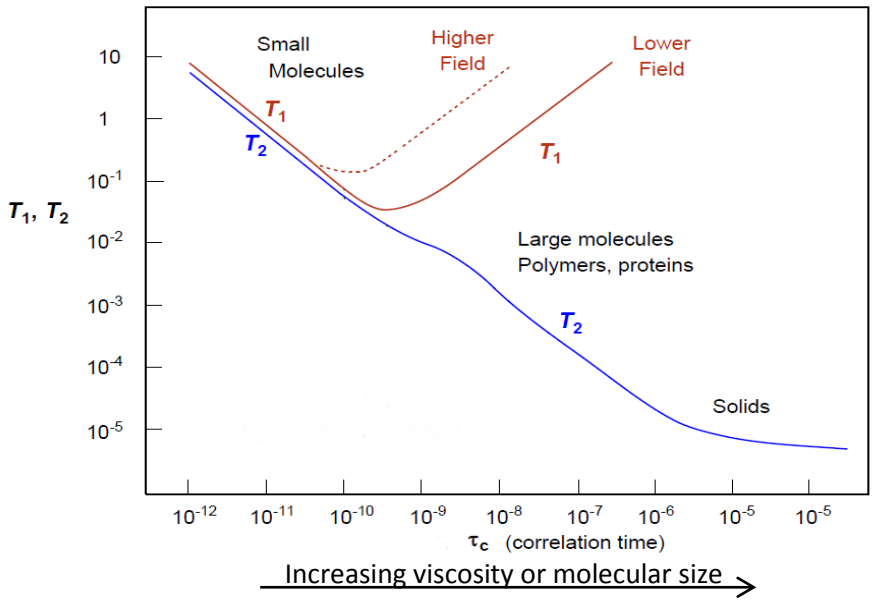


Figure 1.1 Longitudinal T_1 and transversal T_2 relaxation rates as functions of the correlation time τ_c: the correlation time increases with the size of the molecule. The longitudinal rate (T_1) decreases to reach a minimum, while the transversal relaxation time (T_2) continues to decrease [Reich, U.Wisc.Chem. 605].

Considering that the rotational correlation time is longer for large molecules, in absence of internal mobility it is possible to observe a specific behavior of T_1 and T_2 with respect to increasing τ_c, as reported in Fig. 1.1 [Reich, U.Wisc.Chem. 605].

1.1.2 CHEMICAL SHIFTS

Every spin precesses at a slightly different frequency due to the local chemical environment. This implies that they are observed at different positions (chemical shifts) in the NMR spectrum depending on their environment within the molecule. Motion of electrons around the nucleus shield or de-shield it decreasing or increasing the effective magnetic field B_{eff} that it feels.

For instance, electronegative atoms typically attract electrons de-shielding neighboring nuclei (in the bonding network) that are consequently more exposed to B_0 increasing their resonance frequencies. The effective field B_{eff} is obtained perturbing B_0 by a shielding constant σ, as described in the following equation:

$$B_{eff} = B_0(1-\sigma) = B_0 + B_{loc} \quad (1.6)$$

The eq. 1.1 needs to be consequently modified in a way that the resonance frequency becomes proportional to B_{eff}. In order to avoid such dependency, the chemical shift δ is normalized in part per million (ppm) relatively to the reference resonance (ω_{ref}) of a standard nucleus [Nowick et al, 2003]:

$$\delta = \frac{\omega - \omega_{ref}}{\omega_{ref}} \times 10^6 \quad (1.7)$$

The precession frequency difference among nuclei of the same type is very small, while it becomes larger for different nuclear isotopes. Some nuclei may have exactly the same chemical shift generating overlap ambiguities. The intensity of the signal is anyway proportional to the number of nuclei possessing the same chemical shift.

The distribution of the electronic charge is anisotropic, thus the intensity of the shielding effects produced by the electron clouds surrounding the nuclei depends on the molecular orientation with respect to B_0. Different chemical shifts are observable in dependence on such possible orientations (two parallel and one perpendicular to B_0). The typical measured chemical shift from NMR spectra (the isotropic one) represents the average value of the shielding constants of these three displacements:

$$\delta_{ISO} = \frac{(\sigma_{11} + \sigma_{22} + \sigma_{33})}{3} \quad (1.8)$$

The chemical shift anisotropy (CSA) is defined as the difference between the smallest position-related chemical shift and the average of the other two chemical shifts.

$$\Delta\sigma = \sigma_{11} - \frac{(\sigma_{22} + \sigma_{33})}{2} \quad (1.9)$$

It does not affect the chemical shifts of NMR in solution but it can contribute to the relaxation process.

1.1.3 COUPLINGS

The shape of the signals (splitting in a multiplet structure) in the NMR spectrum is affected by the presence of a neighboring nucleus (connected via bonds) and it is strictly dependent on the spin state of this latter. This interaction is also known as J or scalar coupling and can be spread only through bonds.

Two protons with different chemical shifts may be attached to two adjacent atoms (e.g. $H^{\alpha}C^{\alpha}$-$C^{\beta}H^{\beta}$). The nucleus of one proton can be aligned with or against B_0, decreasing or increasing respectively the magnetic field felt by its neighboring proton. If the effective field is decreased the neighboring proton resonates at lower frequency and vice versa.

Considering that two nuclei of the same type of proton attached to the same atom (e.g. $C^{\beta}2H^{\beta}$) possess four possible spin orientations (with two equivalent cases) with respect to B_0, the neighboring proton (H^{α}) results to be split in three signals (triplet) separated by a ${}^{N}J_{\alpha\beta}$ distance, expressed in Hz where N defines the number of covalent bonds (three in this case) between the nuclei α and β. Generally, N neighboring protons to a certain atom produce N+1 splitting signals of such atom with intensity ratio defined by the Pascal triangle.

The J coupling is defined vicinal if the protons are located three bonds apart, while it is known as geminal if two bonds are involved. Its magnitude decreases with the increasing number of bonds separating the atoms. The splitting effect can be avoided by means of a decoupling process where all the protons are irradiated contemporarily causing a very rapid transit between the states. This irradiation is applied during the acquisition of the FID.

If the scalar coupling is established between protons with a very different chemical shift, the splitting signal reveals a simple multiplet pattern and it is known as weak or first order J coupling. In case they have similar chemical shifts, distortions and complications on the splitting patterns are commonly observed (strong J coupling). The former is guaranteed if the chemical shift difference Δδ (in Hz) is much greater than the J value:

$$\frac{\Delta\delta}{J} > 5 \qquad (1.10)$$

1.1.4 NUCLEAR OVERHAUSER EFFECT

The equalization of the population distribution (saturation) of a specific nucleus (e.g. H^{α}) is obtained irradiating it continuously during the relaxation delay (before the 90° pulse). This effect propagates enhancing the population difference of another nucleus (e.g. H^{β}) that is close to the former in the space (5 Å). The observed relaxation of the latter is thus strictly dependent on the distance *r* between the considered nuclei.

The time for NOE building up is called mixing time (τ_m) where a combination of pulses and delay periods is applied to induce the magnetization transfer between the nuclei. In case of a long mixing time, a spin diffusion effect takes place where the perturbation of a certain nucleus propagates to a second one that in turn affects a third nucleus. The NOE is due to dipole-dipole interactions between two nuclei. It can occur in two manners: zero- and double-quantum relaxations that dominate the relaxation of large and small molecules respectively.

1.1.5 MULTI-DIMENSIONAL NMR

The one-dimensional experiment [e.g. Purcell et al, 1946] generally yields a spectrum of only one type of isotope, while the multi-dimensional one allows the observation of several isotopes simultaneously at different ppm frequency ranges and consents to overcome the overlapping problem. Dealing with one-dimensional NMR spectra furnishes information about the resonance frequencies and intensities and about the through-bond and space interactions. The signal overlaps limit the potential of such experiment.

In the two-dimensional NMR experiment [Aue et al, 1976] the direct direction is defined by the Fourier transform of the FID in a t_2 detection time, while the indirect one represents the Fourier transform of incrementally delayed FID at t_1 (evolution time) steps. In the multi-dimensional NMR the nuclei are identified by their chemical shifts in both directions and the existing correlation between them (J coupling or NOE) appears as a cross peak.

Each type of experiment consents to extract different information in accordance to the mixing sequence. In particular, there are two main categories of magnetization transfers that can be inferred by different mixing pulse sequences: those ones based on J-coupling (COSY, TOCSY, HSQC, HMBC and HMQC spectra) and the others based on NOE interactions (NOESY and ROESY spectra). The range of the J coupling (the amount of bonds where the magnetization is spread) can be directly selected in the mixing sequence. The TOCSY experiment is based on a hopping vicinal J coupling that is restrained to the atoms belonging to the same residue. The ROESY experiment is an

alternative to the NOESY case that can be particularly useful for small molecules as peptides.

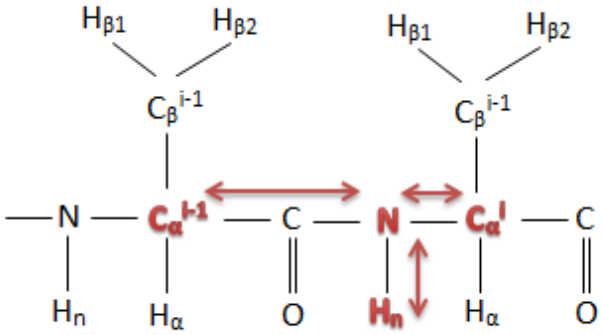
The multi-dimensional experiments may be classified in homonulcear (with a magnetization transfer between nuclei of the same type, as two protons in NOESY, ROESY, TOCSY and COSY spectra) and heteronuclear (with a magnetization transfer between two different types of nucleus, as proton and nitrogen in HSQC, HMQC and HMBC spectra). The latter involves the use of isotopically labeled proteins (^{13}C and ^{15}N) [Gardner et al, 1998]. The homonulcear spectra possess diagonal and symmetrical couples of cross peaks around it that arise from both possible magnetization patterns (i.e. magnetization from H^{α} to H^{β} and vice versa). These properties do not characterize the heteronuclear spectra.

The dimensionality of the measured spectra can be increased obtaining higher dimensional NMR data (as three-dimensional spectra) with less overlapping problems [Cavanagh et al, 1996; Oschkinat et al, 1988; Marion et al, 1989]. In order to obtain a third dimension adjunctive evolution and mixing steps must be intercalated combining the mixing sequences of the required two-dimensional experiments (i.e. two-dimensional ^{1}H-NOESY and two-dimensional ^{15}N-HSQC experiments as a three-dimensional $^{1}H^{15}N$-HSQC-NOESY). In order to observe three-dimensional experiments, the two-dimensional planes are separately evaluated in every direction of interest. Some of the most used three-dimensional experiments are HCCH-TOCSY, $^{1}H^{15}N$-HSQC-TOCSY and $^{1}H^{15}N$-HSQC-NOESY. A vast part of NMR experiments belong to the triple resonance class where three different types of nuclei (e.g. proton, nitrogen and carbon) can be observed simultaneously. Some examples are the HNCA, the HNCO, the CBCACONH, CBCANH, the HBHA(CBCACO)NH and the HBHA(CBCA)NH, whose magnetization transfer is represented in Fig. 1.2 In the former case (*a*) the scalar coupling is transferred from the amide proton to the nitrogen and further to the central carbon of its own amino acid and to that one of the previous residue. In such experiment both α-carbons can be contemporarily observed. The HNCO experiment reveals instead the carbon of the previous residue (*b*). In the CBCACONH spectrum the α- and β-carbons of the previous residue are revealed (*c*), while the CBCANH contains both α- and β-carbons of the considered amino acid and those ones of the previous residue (*d*). The latter couple of triple resonance experiments listed above allows the observation of the α- and β-protons only of the previous residue (*e*) and the detection of such atoms also in the considered amino acid (*f*) respectively.

In Fig. 1.3 is represented the three-dimensional data matrix of a HNCA experiment and the stripes along some different nitrogen planes are extrapolated including H^{N}-C^{α} cross peaks.

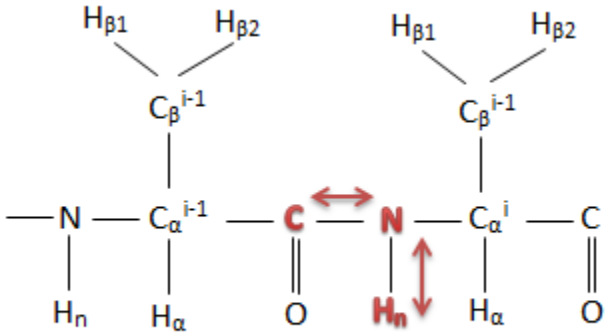
The spectral parameters extracted from such spectra are typically used to determine structural information of molecules.

HNCA



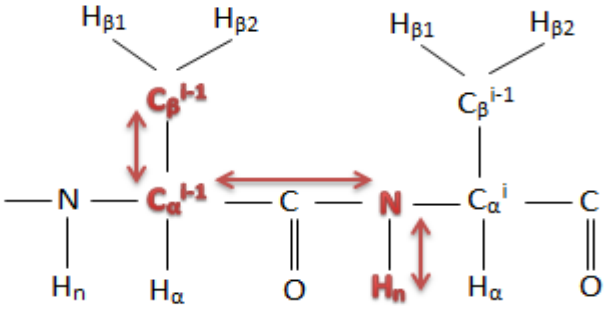
(a)

HNCO



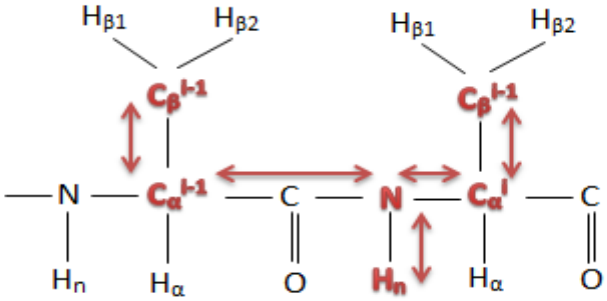
(b)

CBCACONH



(c)

CBCANH



(d)

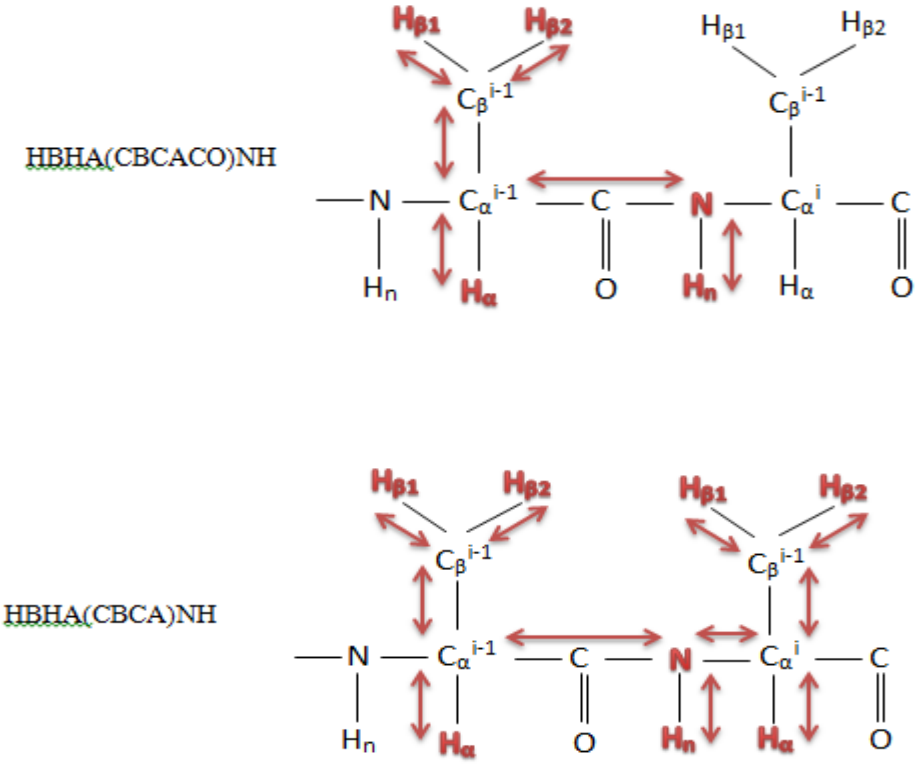


Figure 1.2 Schematic representation of the triple resonances experiments: HNCA (a), HNCO (b), CBCACONH (c), CBCANH (d), HBHA(CBCACO)NH (e) and HBHA(CBCA)NH (f). The atoms observed in each direction are highlighted in red in every experiment. The magnetization transfer is described by the red arrows.

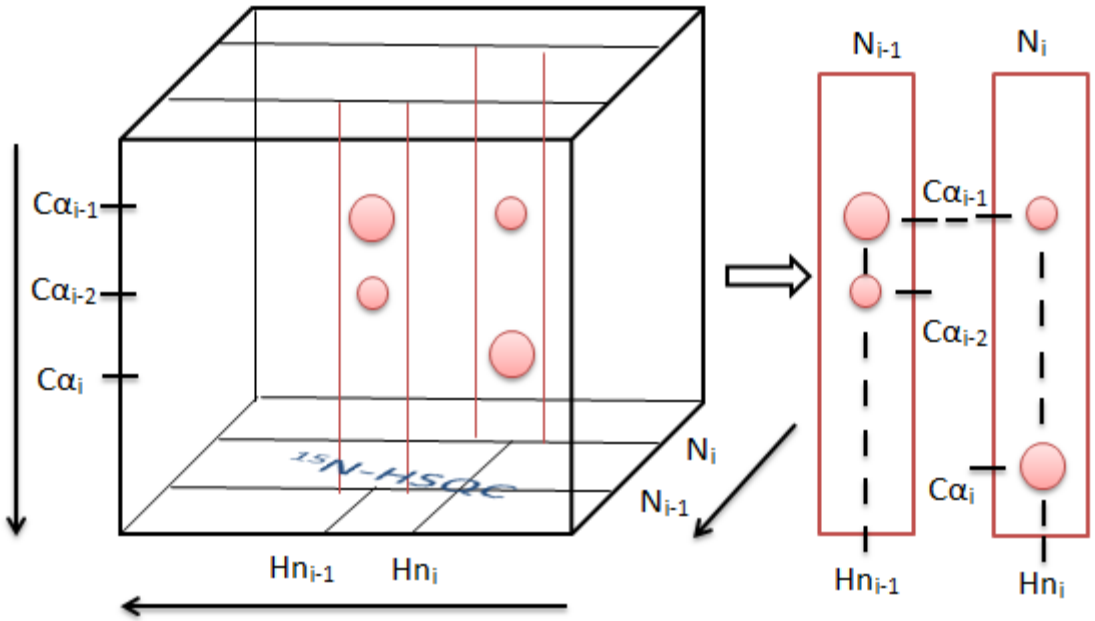


Figure 1.3 Three-dimensional HNCA experiment: planes extraction (N_i and N_{i-1}).

1.2 Data processing

Before Fourier transforming the time-domain acquired data, they need to be accurately processed. They can be multiplied by a weighting function (exponential, Gaussian, sine-bell, etc.) and they can be enlarged by adding zeros at the end of the FID. Moreover, after Fourier transforming additional correction may be applied as baseline flattening and phase adjusting. Solvent artifact removal is performed both during the acquisition of the data and by post-processing methods that can be applied either in the time or in the frequency domain.

1.2.1 BASELINE CORRECTION

The baseline in the spectrum should be ideally flat, but in reality it contains rolls and wiggles due to several reasons [Tang, 1994; Marion and Bax, 1988] including electronics errors. Especially, corruption of the early part of the FID [Hoult et al, 1983; Otting et al, 1986; Bax et al, 1991] due to the transient response of the spectrometer is critical for baseline distortions and it can be alleviated by using the Hahn echo pulse sequence [Bax and Pochapsky, 1992; Kay et al, 1992], by backward linear prediction [Marion and Bax, 1989; Güntert et al, 1992] of the corrupted points, by baseline cosine fitting [Heuer and Haeberlen, 1989] or by spectra oversampling [Wider, 1990; Halamek et al, 1994; Moskau, 2002].

Baseline distortions may also originate from the tails of the solvent spreading all over the spectrum [Bartels et al, 1995]. If it is not properly corrected it may lead to inaccurate signal integration (measurement of the resonance intensity). In multi-dimensional spectra (NOESY case), if the baseline is not properly corrected the volume calculation and consequently distance determination would be erroneous. Moreover, the weak resonances of interest may completely disappear under the noise threshold. *Bartels* has proposed an iterative flattening method (IFLAT) for baseline correction in multi-dimensional NMR spectra [Bartels et al, 1995] with strong solvent signals. It relies on a probabilistic spectral investigation for determining whether to attribute the considered data points to a baseline region or to a true resonance of interest.

Post-processing techniques for baseline correction have notably increased in the last decades. Some of them rely on the manual determination of the real signals and of the baseline regions followed by automated fitting of those baseline points to a polynomial function (up to 5^{th} order) that is then subtracted from the data [Barsukov and Arseniev, 1987; Dietrich et al, 1991]. The manual baseline point identification becomes extremely demanding increasing the dimensionality of the data, where the

baseline distortions appear like stripes extending in several directions from a true resonance.

Remarkable efforts have been done in order to automate the baseline point recognition and to fit the data with other functions, as the linear and the cubic spline interpolation [Saffrich et al, 1992; Zolnai et al, 1989]. The most commonly used approaches found in literature apply the automated baseline correction in the frequency domain [Pearson, 1977; Güntert and Wüthrich, 1991; Dietrich et al, 1991; Chylla and Markley, 1993; Rouh et al, 1993; Brown, 1995; Golotvin and Williams, 2000; Schulze et al, 2005; Cobas et al, 2006].

The FLATT algorithm proposed by *Güntert* and *Wüthrich* [Güntert and Wüthrich, 1991] is particularly effective since it automatically detects entire pieces of rows or columns (larger than the line width of a true signal) that can be fitted by a straight line. The average square deviation for a best fit of a straight line is computed on a stretch of 2n+1 data points with n chosen in such a way that 2n+1 corresponds to 75 Hz. *Dietrich* proposed [Dietrich et al, 1991] instead an automated recognition based on the computation of the first derivative of the average spectrum obtained with a moving filter with a width of 2 points sliding along the rows. The power spectrum is then generated where an iterative thresholding algorithm is applied. *Brown* in 1995 described Bernstein polynomial fitting functions of baseline regions [Brown, 1995]. Five years later *Golotvin* improved the baseline recognition [Golotvin and Williams, 2000] computing the maximal and the minimal values contained in each stretch of N points and evaluating if their difference exceed the noise standard deviation. *Cobas* in 2006 proposed instead the baseline recognition [Cobas et al, 2006] based on a continuous wavelet derivative transformation (CWT) followed by iterative threshold detection in the power spectrum [Dietrich et al, 1991] and by baseline interpolation with the Whittaker smoother algorithm [Whittaker, 1923; Eiler, 2003].

1.2.2 PHASE CORRECTION

After Fourier transforming not all signals in the spectrum have an absorptive line shape since phase distortions due to instrumental errors may arise. Ideally the real spectrum should be in pure absorptive mode and the imaginary should possess both absorptive and dispersive signals. In order to correct deviations from such ideal case a phase rotation angle ϕ of both spectra must be accurately defined, taking into account that it is a linear function of the chemical shift δ (see eq. 1.11).

$$\phi_i = (phc1 \times \omega) + phc0 = \left(phc1 \times \frac{i}{n}\right) + phc0 \quad (1.11)$$

in which the zero-order (phc0) and the first-order (phc1) phase corrections must be determined for each *i*th data point in the real and imaginary spectra over a total of *n* data points.

Typically, the zero-order phase distortion comes from differences between the reference phase and the receiver detector phase and it is frequency independent. The first-order phase distortion is due to several factors as time delay between excitation and detection, flip-angle variations and filtering procedures [Craig and Marshall, 1988; Neff et al, 1977; Daubenfeld et al, 1985]. Unlike the zero-order case, it is frequency dependent.

The phase correction is applied in the real and imaginary spectra accordingly to eq. 1.12 and eq. 1.13:

$$R_i^{ac} = R_i^{bc}\cos(\phi_i) + I_i^{bc}\sin(\phi_i) \tag{1.12}$$

$$I_i^{ac} = I_i^{bc}\cos(\phi_i) - R_i^{bc}\sin(\phi_i) \tag{1.13}$$

where R_i^{bc}, I_i^{bc} and R_i^{ac}, I_i^{ac} represent the *i*th data points of the real and the imaginary parts before and after phase correction respectively.

Using the modern software (e.g. TOPSPIN, XWINNMR), the phase correction is typically performed in the frequency domain with a manual evaluation of the signals in the spectrum. Dealing with multi-dimensional data implies the simultaneous correction on more than one row and more than one column. Many efforts have been done in order to automate this task.

Automated phase correction of 1D NMR spectra has been deeply analyzed [Chen et al, 2002; Koehl et al, 1995; Balacco, 1994; Ernst, 1969; Heuer, 1991; Brown et al, 1989; Craig and Marshall, 1988]. *Ernst* in 1969 [Ernst, 1969] firstly proposed to use the Hilbert transform for finding dispersion and absorption signals with null and maximal integral respectively. *Chen* in 2002 developed an automated phase correction algorithm for one-dimensional spectra based on entropy minimization. This method [Chen et al, 2002] overcame the limitations due to signal-to-noise ratio and signal overlapping affecting previous techniques [Heuer, 1991; Brown et al,

1989; Craig and Marshall, 1988] based on symmetrizing lines, baseline optimization and DISPA (dispersion versus absorption plots) respectively.

The multi-dimensional problem is actually a challenging task [Cieslar et al, 1988; Hoffman et al, 1992; Dzakula, 2000; Balacco and Cobas, 2009]. *Cieslar* has been the first to propose an automated multi-dimensional phase correction algorithm [Cieslar et al, 1988] based on the maximization of signal asymmetry and of signal height in the diagonal. It was limited to homonuclear spectra. *Hoffman* exploited the DISPA method [Hoffman et al, 1992] for correcting the phase in multi-dimensional spectra. *Dzakula* in 2000 [Dzakula, 2000] developed the PAMPAS algorithm (Phase Angle Measurement from Peak Areas) and as last proposal *Balacco* and *Cobas* used a whitening algorithm in 2009 [Balacco and Cobas, 2009].

1.2.3 SOLVENT SUPPRESSION

The one-dimensional FID is represented by the sum of K exponentially damped sinusoids (see eq. 1.14) with a_k amplitude, φ_k phase, d_k damping factor and f_k frequency. The different precession rates of the spins in the transverse plane are observable as different oscillation rates f_k. After Fourier transforming, each signal is translated in to a Lorentzian line disposed at a specific position (chemical shift) in the frequency domain. In particular, n is related to the n^{th} data point and Δt defines the sampling interval:

$$y_n = \sum_{k=1}^{K} a_k \, e^{i\varphi_k} e^{(-d_k + i2\pi f_k)n\Delta t} \tag{1.14}$$

The solvent signal is responsible for the low-frequency component of the FID signal that can be eventually subtracted from the original dataset before Fourier transforming. Typically, in the frequency domain it is positioned in the middle of the spectrum. Temperature changes induce phase and amplitude variations of the signals (i.e. the solvent) in the time domain that correspond to a shift of the resonance positions in the frequency domain.

1.2.3.1 DEUTERATED SAMPLES

The use of liquid-state NMR data involves the dissolution of the samples in a solvent. Since in proton ^1H NMR the solvent resonance must not dominate the spectrum, the

hydrogen atoms of the solvent molecule can be replaced with atoms of deuterium (^2H).

The deuterium is also necessary to lock the strength of the magnetic field B_0 that must be unchanged during the experiment. Since there are several scans for each FID (the same experiment is acquired several times and added up), a field strength variation, thus a frequency change, would not allow a correct sum of the same peak through the various acquisitions. The lock channel monitors the drift of the magnetic field detecting continuously the chemical shift position of the deuterium signal. Any shift on the deuterium resonance is detected and B_0 is adjusted to keep it constant. The resonance of the deuterium becomes sharper enhancing the shimming (correcting the inhomogeneity of the magnetic field), thus it is easily detectable keeping constant the amount of deuterium in the sample.

The D_2O (heavy water) is the ideal solvent when dealing with water soluble molecules.

1.2.3.2 SOLVENT SUPPRESSION: EXPERIMENTAL METHODS

The solvent-to-solute ratio concentration is often in the order of 10^5, thus recording the spectrum in completely non-deuterated solutions would lead to an almost complete disappearing of the solute resonance in the spectrum. In order to detect amide hydrogen atoms, the spectrum is generally acquired in a solution of 90% H_2O / 10% D_2O. In samples containing small amount of deuterated solvent the water protons may still be too strong and hide resonances of interest. In order to avoid that problem, instrumental techniques for solvent suppression must be always applied.

Among all the existing experimental methods, one of the most used is the presaturation [Hoult, 1976] where a long low-power irradiation at the solvent frequency is applied in order to saturate the solvent protons during the relaxation time. It is immediately followed by a normal pulse exciting them. Typically, partial saturation of H^N spins due to rapid chemical exchange between amide protons and saturated water protons affects spectra with presaturation. This effect and the presence of protons resonating close to the water position (H^α) cause a reduction of the intensity of the amide proton resonances. Alternative methods without water saturation have been designed with the aim to overcome these problems.

The WATERGATE (WATER suppression by GrAdient Tailored Excitation) solvent suppression method encompasses gradients pulses (PFG) de-phasing the magnetization of the water and of the solute with a successive refocusing of the solute signals [Piotto et al, 1992; Sklenar et al, 1993; Saudek et al, 1994; Liu et al, 1998]. The water signal is much more strongly suppressed compared to presaturation. Intensity reduction of the resonances of interest is still observable. The water signal needs a

longer T_1 relaxation time to regain its equilibrium state with respect to the protein resonances, thus it cannot fully relax back and it is partially saturated.

Water-flip-back [Grzesiek and Bax, 1993a,b; Lippens et al, 1995] with an adjunctive selective pulse at the water resonance before the standard watergate sequence reducing the saturation transfer.

Alternative method to presaturation and selective excitations is the use of spin-lock pulses [Messerle et al, 1989], the jump-return suppression method [Plateau and Gueron, 1982] and the use of binomial sequences [Hore, 1983].

Using modern NMR spectrometers an excellent water suppression can be achieved with selective excitation techniques such as WET (Water suppression Enhanced through T_1 effects) [Smallcombe et al, 1995], excitation sculpting [Hwang et al, 1995] or by applying more complicated selective presaturation sequences such as PURGE [Simpson and Brown, 2005].

The above mentioned experimental techniques are typically complemented with post-processing solvent suppressing methods.

1.2.3.3 SOLVENT SUPPRESSION: POST-PROCESSING METHODS

The experimental suppression alone is not sufficient for resonance recovering after water suppression, thus several post-processing methods have been developed.

Some of the first approaches exploited the fact that the water resonance is usually positioned at the center of the spectrum. *Kuroda* in 1989 proposed the Fourier transform of second derivatives of the FID for solvent suppression [Kuroda et al, 1989]. *Marion* in the same year used convolution filters with the same purposes. In this case the FID was typically filtered by a low pass finite digital filter with a specific bandwidth positioned at the resonance frequency of the water [Marion et al, 1989]. The convolution of the FID with a Gaussian or sine bell window is then subtracted for water suppression. A modified low-frequency deconvolution filter on COSY spectra has been proposed by *Friedrichs et al*, in 1991. The Karhunen-Loeve transformation has been applied for filtering out low-frequency contributions in the time domain signal [Mitschang et al, 1990]. The continuous wavelet transform (CWT) and the Gabor transform furnish a time-frequency representation of the signal and they can also be used for suppressing large unwanted spectral resonances as the solvent [Barache et al, 1997; Antoine et al, 2000]. *Günther* applied the dyadic discrete wavelet transform (DWT) as alternative method [Günther et al, 2002]. For a critical review of such filtering approaches see *Coron et al*, in 2001, where several methods are compared concluding that the finite impulse response FIR filter [Sundin et al, 1999] is the most efficient one.

The dispersive tails of the water resonance can be eliminated applying suitable baseline correction [Adler and Wagner, 1991]. They are largely attenuated by fitting these tails to a hyperbolic function which is then subtracted from the spectra. The dispersive tails of the water resonance can also be suppressed by phasing the water signal in absorption mode, eliminating such signal from the real part of the spectrum, discarding the imaginary part and regenerating the correct imaginary data from the processed real part via a Hilbert transform [Tsang et al, 1990].

Signal decomposition techniques as principal component analysis (PCA) and singular value decomposition (SVD) have been also applied for solvent suppression purposes. *Grahn* in 1988 described the use of PCA for pattern recognition in two-dimensional NMR spectra [Grahn et al, 1998]. The PCA was previously used for artifact reduction in COSY spectra [Hardy and Rinaldi, 1990]. Singular value decomposition (SVD) on a Hankel-type matrix of the FID was also employed for large artifact removal [Brown and Campbell, 1990; Pijnappel et al, 1992].

The SSA (singular spectrum analysis) [Ghil et al. 2002] is an extension of the PCA. The latter creates an autocorrelation matrix by time averaging over a sample of free induction decays, while the former embeds each FID separately in an M-dimensional vector space [Zhu et al, 1997]. The interrelationships among SVD, PCA and KLT have been discussed by *Gerbrands* [Gerbrands, 1981].

The matrix pencil techniques are related to PCA and SVD determining the eigenvectors and eigenvalues of a pair of time delayed correlation matrices [Lin et al, 1997]. Time-embedding techniques and simultaneous or joint diagonalization of a set of Toeplitz trajectory matrices recently led to reconsider those methods [Parra and Sajda, 2003]. These blind source separation (BSS) techniques based on a GEVD (generalized eigenvalue decomposition) of a matrix pencil have been applied to 2D NOESY proton NMR spectra of proteins to remove the water resonance [Stadlthanner et al, 2006]. The extracted components are automatically identified using the simulated annealing [Boehm et al, 2006].

The ideal procedure must not only remove the solvent signal, but it also must not distort the rest of the spectrum, reveal the hidden resonances of interest and require no user intervention.

1.3 Molecules

Proteins are biochemical essential compounds of the organism, having diverse biological functions: oxygen transport (hemoglobin), hormone transport (albumin), cell signaling (transduction proteins), antibody (immunoglobulin), antigen (bacterial and viral proteins), hormone effector (insulin), mobility (myosin, actin), receptor,

repressor, storage (ferritin), catalyst (enzymes) and structure (keratin, collagen). Generally, they build up complexes connecting to other biomolecules such as lipids (lipoproteins), carbohydrates (glycoproteins), phosphate groups, nucleic acids and prosthetic groups.

The proteins are constituted by different compositions of 20 canonical amino acids forming a polypeptide chain. The peptide covalent bond is formed between the carboxyl and the amide groups of adjacent amino acids losing a water molecule. When they are joined together they are called residues. The function of each protein is strictly related to the properties of its constituting amino acids. As described in Fig. 1.4 they possess common basic structure (i.e. backbone or main chain atoms) including an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom (H^{α}) and a carbon (C^{α}). A variable side chain (R) is bonded to this latter.

The side chains of the various amino acids reveal different properties as the hydrophobicity. Glycine is the simplest one exhibiting a certain conformational flexibility. Alanine, valine, leucine and isoleucine possess aliphatic side chains mostly involved in hydrophobic interactions.

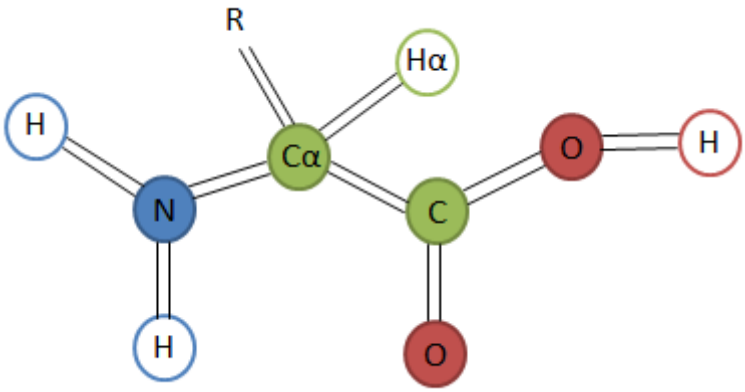


Figure 1.4 Chemical structure of the backbone (main chain) of an amino acid: the backbone is the same for all amino acids that are distinguished only by different side chains (R).

The hydrophobicity increases with the number of aliphatic C atoms in the side chain. Aspartic and glutamic acids are polar and behave as organic acids. Their amides (asparagine and glutamine) are very hydrophilic. Tyrosine, tryptophan and phenylalanine are aromatic amino acids absorbing ultraviolet light. The basic ones as histidine, lysine and arginine are hydrophilic. Proline is the only cyclic one exhibiting similar behavior of the aliphatic group. Serine and threonine are hydroxyl hydrophilic amino acids. Cysteine and methionine contain sulfur atoms and are considered hydrophobic.

Alanine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, proline, serine and tyrosine can be directly produced by the human organism, while the other 10 amino acids must be provided by the food.

The extended polypeptide chain that constitutes the protein is stabilized by the planarity of the non-hydrogen atoms involved in the peptide bonds and by the limited rotation about this bond. The peptide bond angle ω can therefore assume only two values: 0° (*cis* conformation) and 180° (*trans* conformation). There are two other bonds in the polypeptide backbone, between N and Cα and between Cα and C. The former is known as phi (ϕ) torsion angle and the latter is called psi (ψ) torsion angle. Since relative free rotation about them is permitted, the number of possible conformations of the polypeptide chain is restricted by the combination of rotatable bonds with rigid planar regions as described in Fig. 1.5.

The values of the torsion angles are anyway sterically constrained by unfavorable contacts between atoms. The Φ and ψ allowed values are described in the Ramachandran plot, reported in Fig. 1.6. Glycine is the only residue that possesses a different Ramachandran plot since the allowed regions are typically larger for residues with a very short side chain.

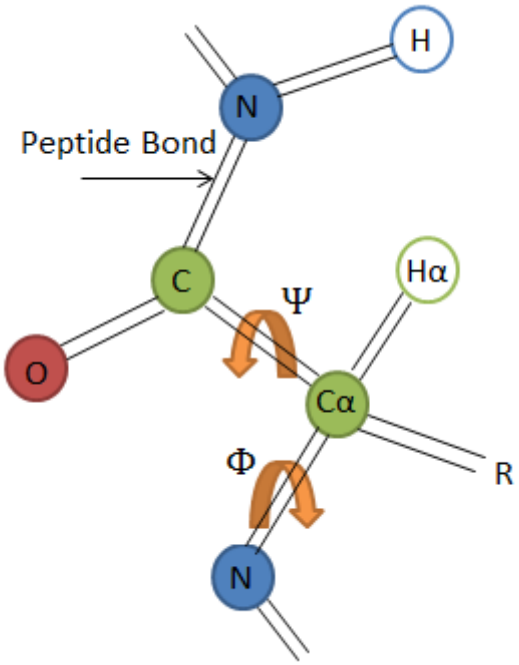


Figure 1.5 Polypeptide chain with backbone bonds: the peptide bond connects two amino acids. Phi and psi torsion angles limit the allowed conformations of the polypeptide chain.

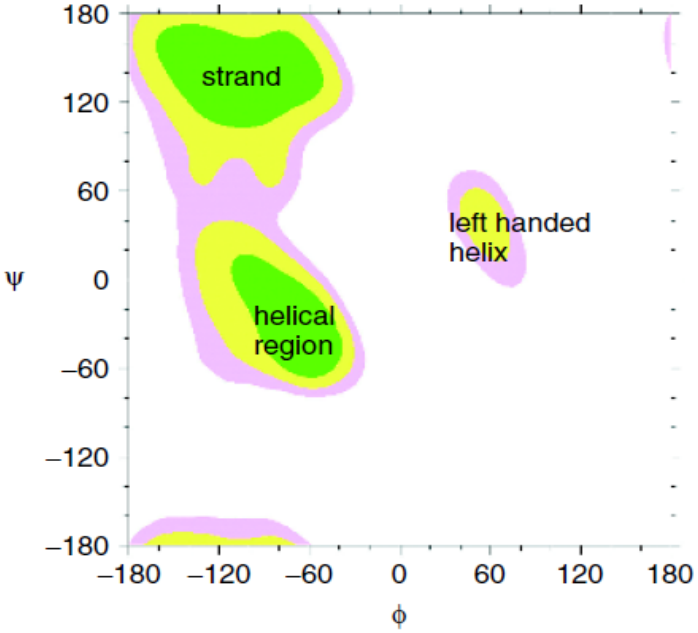


Figure 1.6 Ramachandran plot: most favorable regions highlighted in green with phi and psi angle values giving rise to well-defined structural motifs.

Additional covalent bonds may be present in some proteins in the form of disulfide bridges built up between the sulfur atoms in the side chains of the cysteine residues. They can be broken at high temperatures and acidic pH when a denaturing process is imposed.

1.3.1 PROTEIN STRUCTURE

The primary structure is the linear sequence of the residues along the polypeptide chain that is unique for each protein. Spatial relationships between close residues are responsible of local conformations (secondary structures) of the polypeptide chain. Mainly, they are represented by α helix, β strand and turns. A typical alpha helix contains 3.6 residues per turn (as described in Fig. 1.7). Its existence is strictly related to the values of the torsion angles allowing the building up of hydrogen bonds between the backbone carbonyl oxygen of one residue and the amide hydrogen of the residue located four positions ahead in the polypeptide chain. Proline cannot participate in helical structures due to the lack of an amide proton. The 3_{10} helix is another type of helix whose hydrogen bonds are formed between the residues i and i+3 with 3 residues per turn, while the π helix has hydrogen bonds between i and i+5 residues with 4.4 residues per turn.

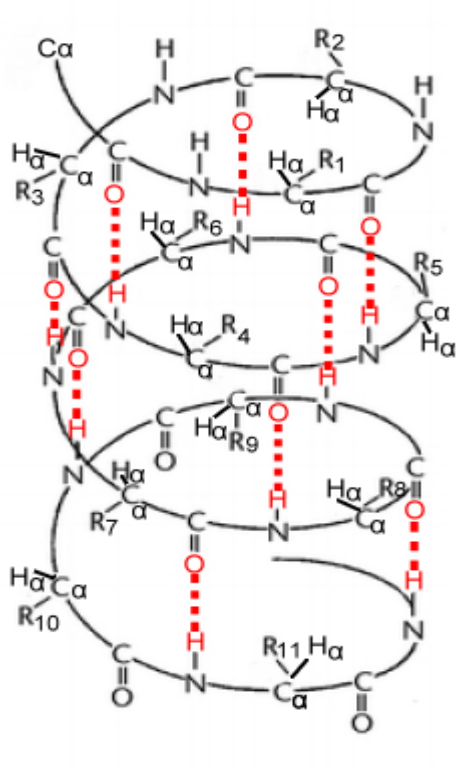


Figure 1.7 Alpha helix secondary structure: hydrogen bonds highlighted in red.

The β sheet is formed by two or more β strands stabilized by hydrogen bonds between the residues i and i+3, with a reversion of the chain direction. Adjacent strands can be parallel or antiparallel (more common) arranged as shown in Fig. 1.8.

The turns often join together such secondary structures and can be distinguished in γ turns, containing three residues and β turns with four residues. The amino acids are commonly found in certain secondary structure. For instance, those ones having a long side chain (leucine, methionine, glutamine and glutamic acid) are typically disposed in α helices, while valine, isoleucine and phenylalanine prevalently form β sheets. Proline and glycine often appear in turn structures. Exceptions must be considered anyway.

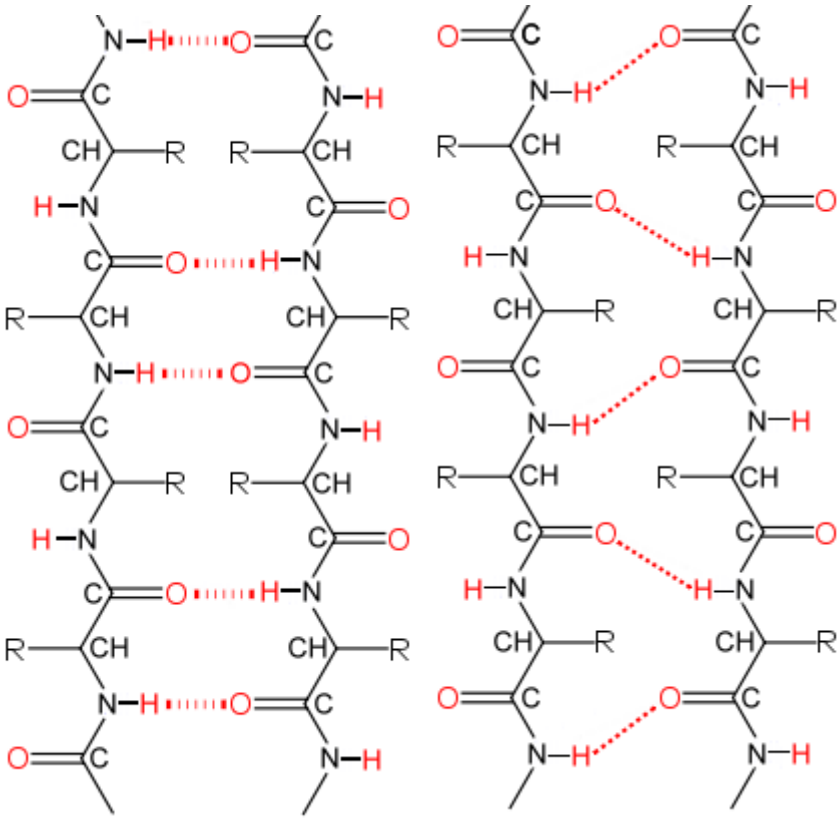


Figure 1.8 Beta Sheet secondary structure: antiparallel (left) and parallel (right) beta sheets (hydrogen bonds highlighted in red).

The folding or the linking of those secondary structures represents the tertiary structure of the protein, where amino acids located far apart in the primary sequence are spatially closely arranged. This structure derives from the interactions arising among secondary elements by means of hydrogen bonds, disulfide bridges, van der Waals interactions, hydrophobic contacts and electrostatic interactions. The polypeptide chain must fold into a correct tertiary structure (native state) in order to become a functional protein. There is a tide relation between primary sequence and folded structure that in turn determines the protein function.

1.3.2 PSCD4-DOMAIN OF PLEURALIN PROTEIN

The pleuralin cell wall protein is obtained from the diatom *Cylindrotheca fusiformis* organism [Kröger et al, 1997; Wenzler et al, 2001]. Diatoms are unicellular organisms belonging to the class of the algae [Van den Hoek et al, 1993]. The scaffold of the diatom cell walls is made up of silica that is associated with proteins [Volcani, 1981]. Within the cell wall there are proteins resistant to EDTA and SDS treatment, thus they are only extractable after complete dissolution of the cell wall in anhydrous hydrogen

fluoride (HF) [Mort and Lamport, 1977]. The pleuralin HEP200 [Kröger et al, 1997] belongs to this class of proteins.

The pleuralin is linked to the silanol groups (SiOH) in the silica surface of the cell wall through covalent bonds (glycoside bonds, phosphodiester-bonds and OH-groups linkages [Hecky et al, 1973; Lobel et al, 1996]). In all the described cases, the pleuralin shows the same modular construction with an N-terminal pre-sequence (amino acids: 1-4), a proline-rich (65%) domain (amino acids: 46-82), 3 to 5 proline-rich (22%) conserved PSCD-domains (with 87 or 89 amino acids for each PSCD domain) and a different C-terminal (amino acids: 576-946) for each pleuralin.

PSCD1	NPSSQPSECA DVLEECPIDE CFLPYSDASR PPSCL—SFG RPDCDVLPTP QNINCPRCCA TECRPDNPMF TPSPDGSPPI CSPTMLPTN
PSCD2	EPSSAPSDCG EVIEECPLDT CFLPTSDPAR PPDCT—AVG RPDCDVLPFP NNLGCPACCP FECSPDNPMF TPSPDGSPPN CSPTMLPTP
PSCD3	APSSQPSQCA EVIEQCPIDE CFLPYGDSSR PLDCTDPAVN RPDCDVLPTP QNINCPACCA FECRPDNPMF TPSPDGSPPI CSPTMMPSP
PSCD4	EPSSQPSDCG EVIEECPIDA CFLPKSDSAR PPDCT—AVG RPDCNVLPFP NNIGCPSCCP FECSPDNPMF TPSPDGSPPN CSPTMLPSP
PSCD5	QPSSQPSECA DVLELCPYDT CFLPFDDSSR PPDCTDPSVN RPDCDKLSTA IDFTCPTCCP TQCRPDNPMF SPSPDGSPPV CSPTMMPSP

Table 2.1 Sequence comparison of the five PSCD-domains of the pleuralin HEP200: PSCD4-domain is the investigated one. Red letters of amino acids identify the parts of the sequences that are identical to the PSCD4-domain. The ten residues of cysteine have conserved the position in all the five domains (green highlighted) being responsible of establishment of disulfide bridges.

The PSCD domains share the 73-91% of the sequence and they contain 10 cysteine residues at exactly the same positions. The PSCD4-domain (amino acids: 372-458) of HEP200 is shown in Table 2.1 and it is compared with the other four domains. Only the third and the fourth domains are directly connected while the others are separated by short sequences of amino acids. The recombinant His_6PSCD4 [Wenzler et al, 2001] is one hundred-twelve amino acids long and contains the PSCD4-domain (amino acids: 16-102) as reported in Appendix A.

1.4 Protein Structure Determination

The three-dimensional structure of the protein is essential information that can be exploited in order to understand the biological function of the protein. There are two major methods for protein structure determination: X-ray crystallography and NMR spectroscopy. They are often complemented by computational approaches based on the simulation of the molecular dynamics.

1.4.1 PROTEIN STRUCTURE DETERMINATION: X-RAY CRYSTALLOGRAPHY

The diffraction of X-ray by sodium chloride crystals has been discovered by *William* and *Lawrence Bragg* in 1907 [Bragg, 1907] who determined the direct relationship between the diffraction pattern and the crystal structure (known as Bragg's law). *James Sumner* has crystallized the first protein in 1926 (enzyme urease). The first protein structure has been determined in 1958 by *John Kendrew* [Kendrew et al, 1958]. Patterns of diffracted X-rays have been used to obtain information about the orientation of the atoms in the molecule.

Solid three-dimensional crystals of the molecule have to be grown for diffraction and it is a time-consuming process, thus this method would be more appropriate for small molecules. The crystal acts as an amplifier since the X-ray irradiation of a single molecule would be too weak, while the crystal contains numerous molecules oriented in the same direction. The crystallized protein is positioned in a tube and it is irradiated by an X-ray beam. The wavelength of the electromagnetic radiation must equalize the distance between the atoms in order to observe such diffraction. The intensity and the positions of the diffraction spots are collected, the crystal is rotated and the measurement is repeated. An electron density map (a map of the distribution of the electrons in the molecule) is obtained which being a good approximation of the atomic position provides information about the relative distance between the atoms. Some fitting models must be computationally determined and refined. The R-factor is the measure of agreement with the measured diffraction data.

The majority of the protein structures that have been collected up today have been determined by X-ray crystallography. This technique can be applied also to very large molecule while NMR is actually limited to 80 kDa and it is recently increased by the use of TROSY (Transverse Relaxation-Optimized Spectroscopy) experiments [Pervushin et al, 1997]. The crystallization process is extremely difficult for non-globular proteins, thus X-ray is not always applicable. NMR spectroscopy allows a more realistic determination of the structure, since it is performed in solution. The different flexibility of the various domains of the molecule is easily detected by NMR.

1.4.2 PROTEIN STRUCTURE DETERMINATION: NMR SPECTROSCOPY

The sample containing the molecule in solution must be located in a spectrometer in order to collect a set of NMR spectra. At this point, the spectroscopist must investigate such data assigning each resonance with a specific nucleus in the molecule. Such task is performed with different strategies depending on the type of measured spectra. In particular, the TOCSY and the COSY spectra allow the

identification of resonance positions in each amino acid separately, due to the fact that the scalar coupling between different residues is four bonds apart and it is negligible. The NOESY experiment is used to perform a sequential assignment building a sort of path between H^{N}_{i}-H^{α}_{i} cross peaks connected through intense sequential resonances of the type H^{N}_{i}-H^{α}_{i-1}. The assignment task can be slighted by the use of adjunctive triple resonance experiments.

After completing the assignment, some restraints must be derived from the spectra. They are necessary to infer accurately the molecular three-dimensional structure from the NMR spectra. The main constraint is related to measurements of the NOEs [Macura and Ernst, 1980; Neuhaus and Williamson, 1989] that define the distances between hydrogen atoms of residues that are close in the space but can be even far away in the primary sequence. This information is extracted from cross peaks volumes of the NOESY spectra, considering that for small mixing time τ_m, the intensity of the peak is inversely proportional to the sixth power of the distance r separating two considered atoms [Kumar et al , 1980], as expressed in eq. 1.15:

$$V_{ij} \propto \frac{1}{r_{ij}^{6}} \tag{1.15}$$

The distance between atoms can be computed as described in eq. 1.16, given a well-known reference distance r_{ref} that arises either from cross peaks of methylene groups (i.e. H^{β}_{1} and H^{β}_{2} of aspartic acid) and it is typically 1.78 $\dot{A}$ or from cross peaks of aromatic rings (as in phenylalanine, tyrosine and tryptophan) that corresponds to 2.42 $\dot{A}$.

$$r_i = \left(\frac{I_{ref}}{I_i}\right)^{\frac{1}{6}} r_{ref} \tag{1.16}$$

in which I_{ref} and I_i represent the cross peak intensities.

The mixing time used to detect the NOEs must be neither too long to avoid spin diffusion [Kalk and Berendsen, 1976; Mertz et al, 1991] effects nor too short in order to make it observable. The volume of the cross peaks must be measured by means of integration tools but a very precise determination is not thinkable. The NOEs must be considered as upper bounds on such distances and can be classified as strong (in the range 1.9 -2.7 Å), medium (between 2.7-3.3 Å) and weak (up to 5 Å).

NMR experiments provide other information as hydrogen bonds developed through scalar couplings. In particular, the long-range HNCO experiment is typically used to find such type of bonds [Cordier et al, 2008; Cordier and Grzesiek, 1999]. In particular, the observed hydrogen bonds are translated in distance constraints as described in Fig. 1.9.

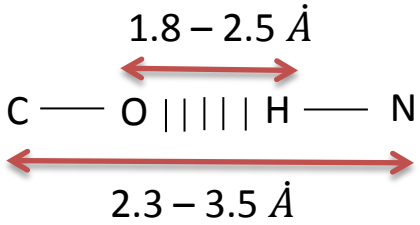


Figure 1.9 Hydrogen bonds developed via scalar coupling: distance restraints obtained from observed hydrogen bonds.

Such restraints are strictly involved in the secondary structure (α-helix and β-sheets) definition.

The chemical shifts are also valuable parameters that can be used in order to infer some structural information. In particular, some studies have been conducted to reveal recurrent chemical shift values for backbone atoms of specific amino acids in well-defined secondary structures [Wishart et al, 1995; Wang et al, 2001]. The TALOS program [Corneliescu et al, 1999] has been developed with the aim to compare the observed chemical shifts to an existing database of chemical shift values in order to predict which residues are involved in secondary structure motifs that in turn define the allowed values of dihedral angles along the main chain of such amino acids.

The torsion angle restraints can be derived not only from chemical shifts, but they are mainly obtained through vicinal scalar coupling constant ^{3}J, whose value can be observed from coupled COSY spectra or from triple resonance experiments as the HNCA E.COSY (Exclusive Correlation Spectroscopy) [Griesinger et al, 1987].

The Karplus relation [Karplus, 1963] connects the three-bond scalar coupling ^{3}J with the torsion angle θ as follows:

$$^3\mathrm{J}(\theta) = A\cos^2\theta + B\cos\theta + C \qquad (1.17)$$

in which A, B and C are empirically computed constants [Habeck et al, 2005]. In particular, the φ dihedral angle restraints are derived from $^3J_{H^N H^\alpha}$ coupling constant, while the χ_1 is obtained from the constant $^3J_{H^\alpha H^\beta}$ [Pardi et al, 1984].

A new class of restraints is related to the measurement of residual dipolar couplings (RDCs) in partially aligned molecules. It provides distance restraints between two atoms (N-H^N) and orientational information about the angle formed between the vector connecting them and the magnetic susceptibility tensor [Tjandra et al, 1997]. They define the bond orientations with respect to the tensor, a global defined axis in the molecule.

The RDC arises from dipolar couplings that have been not averaged out, as in the case of anisotropic conditions. In solution NMR the RDCs can be obtained adding a co-solute inducting a partial alignment (i.e. liquid crystals [Saupe et al, 1964; Saupe et al, 1968; Tjandra and Bax, 1997], bicelle [Sanders et al, 1994; Sanders and Schwonek, 1992; Cavagnero et al, 1999] and other media [Clore et al, 1998a; Hansen et al, 1998; Sass et al 1999; Sass et al, 2000]). The degree of alignment is adjusted varying the concentration of co-solute in the sample [Bryce and Bax, 2004]. They behave isotropically at room temperature, while at high temperatures they cause a signal splitting [Ottiger and Bax, 1998].

The RDC is an additional contribution to the J coupling splitting that is observable in coupled aligned HSQC spectra. The dipolar coupling $D_{AB}(\theta,\Phi)$ is described by eq. 1.18, where A_a and R are the axial and rhombic components of the alignment tensor, while θ and Φ are the polar and azimuthal angles describing the orientation of the internuclear vector with respect to the alignment frame. Both the rhombicity and the axiality parameters must be determined for structure calculation. The magnitude of the molecular alignment tensor is described by three components $|A_{zz}| \geq |A_{yy}| \geq |A_{xx}|$ or equivalently by $A_a = \frac{3}{2}A_{zz}$ and $A_r = A_{xx} - A_{yy}$. The orientational restraints are measured building a histogram [Clore et al, 1998b, Bryce and Bax, 2004; Wei and Werner, 2006] of the distribution of normalized RDCs values where the extremes (A_{yy} and A_{zz}) and the most likely value (A_{xx}) are determined as reported in Fig. 1.10. The accuracy of the histogram is enhanced increasing the amount of available RDCs.

$$D_{AB}(\theta,\Phi) = D_a^{AB}\left\{(3\,cos^2\theta - 1) + \frac{3}{2}Rsin^2\theta cos2\Phi\right\} \quad (1.18)$$

where $$D_a^{AB} = -\frac{\mu_0\gamma_a\gamma_B h}{8\pi^3 r_{AB}^3}\left(\frac{A_a}{2}\right) \quad (1.19)$$

with μ_0 defining the magnetic moment, h is the Planck's constant, γ_i are the gyromagnetic ratios and r is the distance between the nuclei A and B. The rhombicity $R = \frac{A_r}{A_a}$ is obtained as follows:

$$R = \frac{2}{3}\frac{A_{xx}-A_{yy}}{A_{zz}} \tag{1.20}$$

while the axiality is computed in the following manner:

$$A_a = -\frac{A_{zz}}{2} \tag{1.21}$$

The RDC tensor components fulfill the relation:

$$A_{zz} + A_{yy} + A_{xx} = 0 \tag{1.22}$$

The histogram pattern provides the magnitude of the alignment tensor, while the orientation is optimized during the structural calculation.

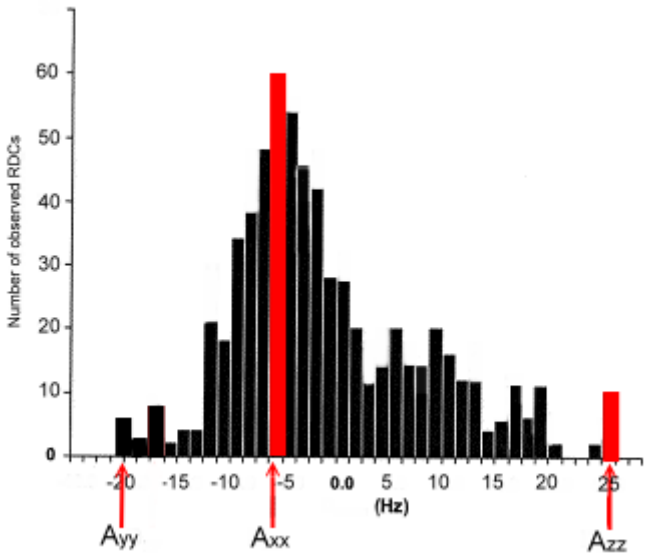


Figure 1.10 Residual dipolar coupling restraints: histogram of the RDCs distribution with extremes highlighted in red.

In order to derive a high quality structure from NMR data at least 20 restraints per residue must be obtained. The calculation of a three-dimensional molecular structure is a minimization problem of a potential energy function describing the agreement between an initial conformation and a given set of restraints. Generally, the folded protein possesses the lowest potential energy, thus a simulated annealing algorithm [Kirkpatrick et al, 1983; Nilges et al, 1988; Bruenger and Nilges, 1993] must be performed for avoiding that the energy is trapped in a local minima. The initial structure is generated by metric matrix distance geometry [Nilges et al, 1988; Crippen and Havel, 1988; Havel, 1991] or by the variable function method [Braun and Go, 1985; Güntert et al, 1991]. It is also possible to start from an extended structure [Brooks et al, 1983] at the expense of the computational time.

The potential energy function that must be minimized is described in eq. 1.23:

$$E_{pot} = E_{bond} + E_{angle} + E_{dihedral} + E_{electrostatic} + E_{vanderWaals} \qquad (1.23)$$

where the first three pseudo-energy terms correspond to three different types of atom movement involving bond length stretching, bond angles stretching and bond rotations. The last two terms represent the contribution of non-bounded interactions.

Some of the existing molecular dynamics program widely used are: Amber [Cornell et al, 1995], Xplor [Bruenger, 1992] and CNS [Bruenger et al, 1998].

In particular, an ensemble of several conformations having low energy is generated, which reflect the flexibility of the molecule in solution. The RMSD (root mean square deviation) is computed (in accordance to eq. 1.24) as the difference between the average structure and the other components of the bundle [Renugopalakrishnan et al, 1991, Hyberts et al, 1992]. The lowest is the RMSD the more similar are the conformations in the ensemble.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_{mean} - s_i)} \qquad (1.24)$$

1.4.2.1 NMR in METABOLOMICS

The metabolome represents the complete set of metabolites (low-molecular-weight compounds as amino acids, carbohydrates, vitamins, lipids, etc.) in cell, tissue, organ and biological fluid (i.e. urine, plasma, etc.). The metabolomics provides an exhaustive analysis of the metabolites contained in a sample furnishing information about physiological status of an organism, pathological diseases, and effects of nutrients, drugs, lifestyle, environment and toxic agents.

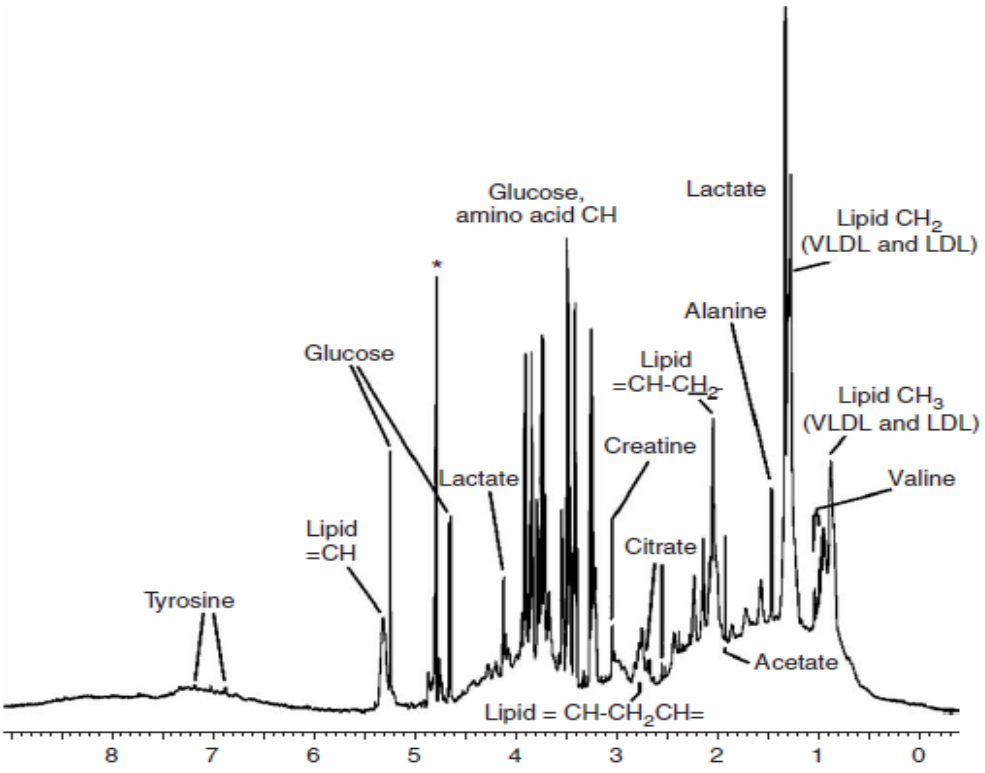


Figure 1.11 NMR-metabolomics: ^{1}H NMR spectrum of blood serum [Beckonert et al, 2007]

NMR spectroscopy and mass spectrometry have been typically used in metabolomics. The former has revealed to be particularly significative in this field [Lindon et al, 1999; Aranibar et al, 2011]. The first applications of NMR to metabolism studies (metabonomics) have been conducted since the 1970s by Brown [Brown et al, 1977]. Successive toxicity metabolism studies have been developed by Nicholson's group [Wilson et al, 1987; Nicholson et al, 1999]. They have designed a method based on PCA in order to ease the urine spectra assignment and to discover new signals [Holmes et al, 1992].

Biofluids are typically used in NMR-metabolomics since they are easily obtainable (i.e. urine, plasma and saliva). This noninvasive technique is also well-suited for metabolic analysis of *in vitro* cell systems as yeast and tumor cells.

The drug and metabolite concentration levels in blood plasma are especially important to determine the precise dosage of a medicine. As in the urine case the signal overlapping is very common and is mainly due to the presence of broad resonances caused by macromolecules as lipids and proteins (albumin and immunoglobulins) covering some minor metabolites of drugs. A typical ^{1}H NMR spectrum of urine reveals thousands of overlapping sharp signals from low molecular weight metabolites (as creatinine and hippurate). The typical spectrum of blood plasma contains instead also the high molecular weight components (as lipids) resulting in a larger line width of the signals. An example of one-dimensional NMR spectrum of blood serum is reported in Fig. 1.11 [Beckonert et al, 2007].

Accuracy in signal assignment and intensity quantification is particularly critical in NMR-metabolomics since many small but significant resonances may be neglected cause of baseline distortions. This problem would lead to incorrect detection of important metabolites and biomarkers. Time and frequency domain baseline correction methods are typically applied on NMR data (see par. 1.2.1). They usually need a robust identification of noise regions [Brown, 1995; Saffrich et al, 1992; Golotvin and Williams, 2000] that can represent a challenging task in metabolomics. *Xi* has proposed a baseline correction of NMR metabolomics spectra based on a penalized smoothing model. Neither an explicit identification of noise data points nor fixed baseline fitting curves are required [Xi and Rocke, 2008]. *Chang* pointed out that many existing algorithms are aggressive [Pearson, 1977; Bartels et al, 1995] proposing a baseline correction method based on sliding window over high pass filtered signal dense spectra [Chang et al, 2007].

1.5 Automated structure determination

In comparison to the large number of protein sequences available, only relatively few 3D protein structures have been solved so far. The gap between the number of experimentally solved structures and the number of known protein sequences is huge and will most probably continue to widen in the future. As long as computational methods are not sufficiently well developed for accurately predicting a protein's structure based on its amino acid sequence alone [Chothia et al, 1986], experimental methods for structure determination at atomic resolution will continue to play a dominant role in structural biology. NMR-based structure determination of small well-behaved proteins (highly soluble, globular and uniquely folded) is nowadays a manageable scientific problem which generally leads to a trustworthy solution. However, an expert must be involved and may need several months for a complete, self-consistent analysis. This situation is, in general, not acceptable in proteomics research, where a large number of structures need to be solved in as short time as possible. Considerable efforts have been made for a complete automated structure determination.

1.5.1 AUREMOL

The twelve years old AUREMOL software [Gronwald and Kalbitzer, 2004] based on AURELIA [Neidig et al, 1995] has been successfully used for the determination of three dimensional protein structures. The goal of AUREMOL is to provide routines for a reliable and automatic protein structure determination from a minimum of experimental NMR data with no or a minor user intervention. AUREMOL relies on a molecule-centred top-down strategy, which means that starting from an initial structure it is iteratively refined until it fits the experimental data.

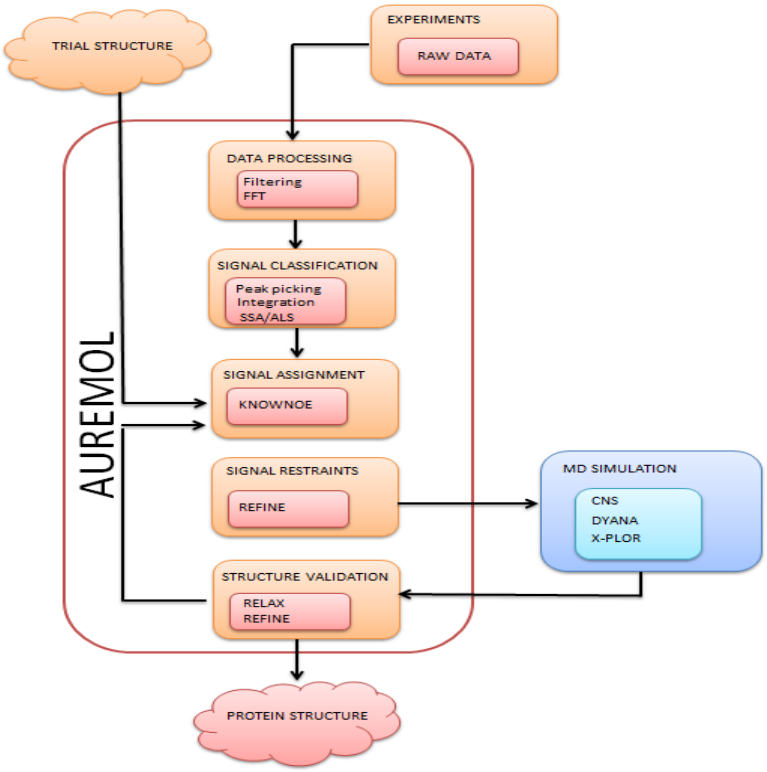


Figure 1.12 AUREMOL top-down strategy: the initial trial structure is iteratively refined in order to optimally fit the experimental data. External programs for molecular modelling are used to validate the structure.

As shown in Fig. 1.12 it is evident that starting from some filtered and Fourier transformed measured spectra, the AUREMOL Bayesian peak picking [Antz et al, 1995] can be applied in order to find all the true resonances of interest and automatically recognize artefacts and noise signals. All the true signals are automatically segmented [Geyer et al, 1995] to obtain volume information. They are automatically assigned by the KNOWNOE module [Gronwald et al, 2002] with a successive optimization step for chemical shift assignment [Baskaran et al, 2009; Baskaran et al, 2010]. The REFINE routine [Trenner, 2006] is used to automatically reproduce NOEs distance restraints and to estimate errors that together with some adjunctive dihedral angles, J couplings and H-bond restraints are required for an external protein molecular dynamics (MD) simulation. Once the MD has been performed, the RFAC module [Gronwald et al, 2000] is applied to validate the molecular structures back calculated through the RELAX module [Görler and Kalbitzer, 1997; Ried et al, 2004; Görler et al, 1999] comparing it with an initial trial structure. A homology modelling routine [Chothia et al, 1986] based on simulated annealing allows one to create suitable starting structures. AUREMOL contains a sequential assignment tool for 3D heteronuclear NMR data and the results can be used to simulate 2D and 3D NOESY spectra based on the complete relaxation matrix formalism using different motional models, including simulation of individual T_2 relaxation times. AUREMOL currently relies on external software only for raw data processing and restrained molecular dynamics calculations. Accurate constraint

information that may be used as input for structure calculation programs such as CNS, X-PLOR, or DYANA, for example, is automatically generated from 2D and 3D NOESY spectra using iterative relaxation matrix calculations. AUREMOL provides routines and utilities for the data analysis and structure validation steps in one NMR program with an emphasis on NOESY evaluation.

2 MATERIALS and METHODS

2.1 Materials

2.1.1 BACK-CALCULATED DATASET: HPr PROTEIN

The back-calculated dataset is made up of a two-dimensional NOESY spectrum of HPr protein from *Staphylococcus aureus* (mutant H15A). SSA has been applied on the time domain both before and after adding the experimental solvent signal (also not positioned in the middle of the spectrum) and in the mixed domain (t1, ω2) including the water experimental artifact. The PCA has been performed directly on the spectrum without any time embedding of the data and it has been compared with the SSA. The two-dimensional dataset with additional experimental water has been also used to perform baseline correction (ALS) without any prior solvent removal.

The one-dimensional simulated spectrum of the HPr protein from *Staphylococcus aureus* (mutant H15A) has been extracted (first FID) from the two-dimensional one. SSA has been applied in the time domain, in order to investigate quantitatively the number of extracted components (embedding dimension of SSA) and to determine specific limitations of the algorithm in dependence on the magnitude of the added experimental solvent signal, and in the mixed domain (t1, ω2). ICA has been applied in the frequency domain and it has been compared with SSA.

2.1.1.1 SSA AND PCA OF TWO-DIMENSIONAL SPECTRA

A synthetic two-dimensional NOESY spectrum has been back-calculated with the AUREMOL module RELAX-*JT*2 [Ried et al, 2004] using the three dimensional bundle of structures of histidine-containing phosphocarrier protein HPr (H15A) from *Staphylococcus aureus* made up of 20 conformations and the corresponding experimental chemical shifts. The HPr is made up of eighty-eight residues and its structure consists of three α-helices and four stranded anti-parallel β-sheets [Maurer et al, 2004]. The mutation H15A involves the substitution of the histidine residue number fifteen with an alanine residue.

The HPr spectrum has been simulated at a temperature of 303 K with a τ_{rot} of 3.65 ns. The order parameters of the backbone-backbone, backbone-sidechain and sidechain-sidechain correspond to 0.85, 0.80 and 0.65 respectively. The H_2O content is 90%. The following parameters have been used: mixing time, 150 ms; cut off distance, 0.5 nm; relaxation delay, 1.3 s; time-domain data, 1024x2048; size of the

real data after Fourier transformation, 512x1024; spectrometer frequency, 600.13 MHz; spectral width, 13.9791 pm in both directions; acquisition modes, qsim and States in w_2 and $w_1$respectively. Another synthetic two-dimensional spectrum has been simulated with the same parameters described above, with 512x16384 time domain data points in order to perform a quantitative investigation of the embedding dimension of the SSA.

The Lorentzian line shape of the peaks has been used in order to compute the artificial dataset with a simultaneous phase increment on both directions. Gaussian noise has been added to the spectrum by means of an existing AUREMOL routine. It has been added with a signal to noise ratio of approximately 2σ for a proton-proton pair in a distance of 0.5 nm [Baskaran et al, 2009]. The resulting time domain data has been filtered by exponential multiplication with a line broadening in the two dimensions of 3Hz and zero filled.

The water artifact was produced by measuring a two-dimensional NOESY spectrum of 90 % H_2O/10 % D_2O with solvent presaturation [Hoult, 1976] at 600.13 MHz, having the same acquisition parameters used for the spectrum simulation. The time domain signal of the synthetic HPr (obtained using the Cfid routine of AUREMOL) has been added to the time-domain signal of the experimental water (inverse Fourier transformed) scaled in such a way that the maximum of the water was about 500 times stronger than the protein signals.

2.1.1.2 SSA AND ICA OF ONE-DIMENSIONAL SPECTRA

The synthetic one-dimensional spectrum of histidine-containing phosphocarrier protein HPr (H15A) from *Staphylococcus aureus* has been obtained from the two-dimensional (with 1024x2048 time domain data points) back-calculated one [Ried et al, 2004] whose parameters have been described in par. 2.1.1.1. In particular, the first FID has been extracted from the two-dimensional raw data (Cfid routine of AUREMOL), while Gaussian noise and experimental water have been added to it before Fourier transforming. Moreover, several one-dimensional spectra have been obtained summing different water artifact signals to the simulated one-dimensional spectrum. The solvent signal has experienced several modifications as scaling and phase change as described in details in chapter four. The experimental water artifact has been measured with the same acquisition parameters of the simulated one as described in par. 2.1.1.1. Moreover, the first FID of the second two-dimensional spectrum (with 512x16384 time domain data points) has been used to investigate the embedding dimension of the SSA in dependence on the data dimensionality.

2.1.2 EXPERIMENTAL DATASET

All the experimental spectra successively described (except the data of the PSCD4-domain of the pleuralin protein) have been acquired and processed with the TOPSPIN software. SSA has been applied on the three-dimensional $^{1}H^{13}C$ HCCH-TOCSY spectrum of Trx protein from *Plasmodium falciparum*, while baseline correction (ALS) has been performed on the same spectrum before SSA. This latter has been applied on the two-dimensional NOESY spectra of HPr protein from *Staphylococcus aureus* (mutant H15A) both digitally (used also for a qualitative analysis of the extracted components) and analogy acquired to define the group delay data points management, on the two-dimensional TOCSY spectrum of HPr protein from *Staphylococcus aureus* (mutant H15A) and on the two-dimensional NOESY spectrum of the PSCD4-domain of the pleuralin protein (successively baseline corrected by means of ALS). It has also been applied on the one-dimensional spectrum of HPr protein from *Staphylococcus aureus* (mutant H15A) measured with watergate solvent suppression (to verify the SSA and the ALS performances when the solvent is not the dominant signal) and on metabolomics data as the one-dimensional spectra of blood plasma, human urine and COS7 cell. Moreover, the urine spectrum has been intentionally baseline distorted before applying the ALS in cascade after the SSA. The ICA has been performed in the frequency domain of two datasets of the one-dimensional human urine spectrum and it has been compared with the SSA. The first dataset is composed of two experiments acquired with mixing times of 10 and 20 ms, while the second one is made up of two experiments measured with a mixing time of 1500 and 2000 ms. Two datasets of ICA-tailored one-dimensional spectra of HPr protein from *Staphylococcus carnosus* have been obtained with specific pulse sequences in order to improve the solvent removal performance. In particular, the first dataset contains two experiments acquired with a different phase cycling, while the second one is composed by two experiments measured with different diffusion times. EMD of the time domain signal of the one-dimensional spectrum measured from a sample containing a mixture of five amino acids has been investigated in details in order to determine the solvent removal performance. It has been also applied to the time domain of the one-dimensional spectrum of a mixture of twenty amino acids with phase correction purposes.

2.1.2.1 SSA OF THREE-DIMENSIONAL SPECTRA: Trx PROTEIN

A three-dimensional experimental $^{1}H^{13}C$ HCCH-TOCSY spectrum has been recorded from a sample containing Thioredoxine protein (Trx) from *Plasmodium falciparum* in D_2O 99.5%, pH 7.0. It is a medium size protein with one hundred-and-four residues and formed by four α-helices and a five stranded β-sheet [Munte et al, 2009]. The spectrum has been recorded on a Bruker DMX-600 spectrometer operating at 600.13

MHz, employing a mixing time of 12 ms, a relaxation delay of 1 s, 2048x96x128 time domain points and 1024x64x64 real frequency domain points. The water signal has not been experimentally reduced since the spectrum has been measured in deuterium. The spectral widths were 6.9945 ppm in the direct direction (ω_3), 70.0 ppm for the first indirect (ω_1) and 6.9945 ppm for the second indirect direction (ω_2). It has been acquired with the 3-1-2 order with the first indirect direction related with the ^{13}C. The acquisition modes used are DQD in the direct direction and States-TPPI in both indirect directions. The spectrum has been filtered with a Gaussian function with a line broadening of -6.0 Hz in the direct direction and -8.0 Hz in both the indirect directions and positioned at 0.12 and 0.1 respectively. It has been acquired at 303 K.

2.1.2.2 SSA OF TWO-DIMENSIONAL SPECTRA

2.1.2.2.1 NOESY SPECTRA OF HPr PROTEIN

A two-dimensional NOESY spectrum has been recorded from a sample with 2.7 mM uniformly ^{15}N-enriched histidine-containing phosphocarrier (HPr) protein from *Staphylococcus aureus* (H15A) in 500 µL 95% H_2O/5% D_2O, pH 7.0. It was recorded on a Bruker Avance-800 spectrometer operating at 800.13 MHz with a mixing time of 100 ms. The two-dimensional NOESY spectrum has been recorded using a relaxation delay of 2 s, with 512x1024 time domain points and with 512x512 real frequency domain points. The spectral widths in the two dimensions were 13.970 ppm. It has been acquired at 303 K. The acquisition modes were DQD and States in the direct and indirect direction respectively. Solvent presaturation [Hoult, 1976] of 2.1 s has been used. Digital filtering [Moskau, 2002] have been applied as well.

Another NOESY spectrum of histidine-containing phosphocarrier HPr protein from *Staphylococcus aureus* has been obtained from the same sample recorded without digital filtering (analog mode). It has been recorded at 600.13 MHz (Bruker Avance-600.13) employing a mixing time of 100 ms, a relaxation delay of 2 s and with 512x1024 complex time domain points. The spectral widths were 13.9790 in both directions. The acquisition modes were qsim and States in the direct and indirect direction respectively. It has been acquired at 303 K. Solvent presaturation of 2.1 s [Hoult, 1976] has been applied.

2.1.2.2.2 TOCSY SPECTRUM OF HPr PROTEIN

A two-dimensional TOCSY spectrum has been recorded from a sample with 2.7 mM uniformly ^{15}N-enriched histidine-containing phosphocarrier (HPr) protein from

Staphylococcus aureus (H15A) in 500 µL 95% H_2O/5% D_2O, pH 7.0. It was recorded on a Bruker Avance-800 spectrometer operating at 800.13 MHz with a mixing time of 100 ms. The two-dimensional TOCSY spectrum has been recorded using a relaxation delay of 1 s, with 1024x4096 time domain points and with 1024x2048 real frequency domain points. The spectral widths in the two dimensions were 13.9486 ppm. It has been acquired at 303 K. The acquisition modes were qsim in w_2 and TPPI in w_1. Solvent presaturation [Hoult, 1976] of 1 s has been used. Digital filtering [Moskau, 2002] has been applied as well.

2.1.2.2.3 NOESY SPECTRUM OF PSCD4-DOMAIN OF THE PLEURALIN PROTEIN

A two-dimensional NOESY spectrum has been acquired at 298 K on a DRX-800 spectrometer operating at 800.13. The sample contained 10 mM sodium-phosphate buffer with a pleuralin protein concentration of 10 mg/ml and 10% D_2O. In addition, 0.1 mM EDTA, 1 mM NaN_3, 1 µM Leupeptin, 1 µM Pepstatin, 1 µM BPTI and 0.1 mM DSS as internal reference have been added in the solution. The NOESY spectrum has been obtained with the following acquisition and processing parameters: mixing time, 100 ms; relaxation delay, 1.6 s; acquisition modes, DQD in w_2 and TPPI in w_1; time-domain data points, 512x4096; real points after Fourier transforming, 256x2048; spectral width, 13.8858 ppm in both directions; window filter, Gaussian; line broadening, -6 Hz and –8 Hz in the direct and indirect dimensions respectively. Solvent presaturation [Hoult, 1986] of 1.7 s and further baseline correction have been applied on the two-dimensional NOESY spectrum. The data have been acquired and processed with XWINNMR 2.6.

2.1.2.3 SSA OF ONE-DIMENSIONAL SPECTRA

2.1.2.3.1 HPr PROTEIN SPECTRUM WITH WATERGATE SOLVENT SUPPRESSION

The one-dimensional spectrum of histidine-containing phosphocarrier protein has been acquired from a sample of HPr protein from *Staphylococcus aureus* (H15A) in a 1.5 mM buffer, pH 7.0 in 95% H_2O/5% D_2O. For the reference, 0.05 mM DSS has been added to the sample. To prevent degradation 0.5 mM EDTA and 1 mM NaN_3 have been used. The spectrum was recorded at 298 K and 800.13 MHz (Bruker Avance-800) using a DQD acquisition mode, a mixing time of 60 ms and a relaxation delay of 2

s. The spectral width was 14 ppm, the data were recorded using oversampling and digital quadrature detection. 64 K complex time domain points were obtained. Watergate solvent suppression has been applied [Piotto et al, 1992; Sklenar et al, 1993; Saudek et al, 1994; Liu et al, 1998].

2.1.2.3.2 BLOOD PLASMA SPECTRA WITH SOLVENT PRESATURATION

A sample of human blood plasma with EDTA added for anticoagulation has been used to record a one-dimensional spectrum at 310 K and 600.13 MHz (Bruker AvanceII-600) with a DQD acquisition mode, a mixing time of 8 ms and a relaxation delay of 2 s. The spectral width was 20.0 ppm, the data were recorded using oversampling [Moskau, 2002] and digital quadrature detection. Water presaturation [Hoult, 1976] of 2.8 s has been applied on the spectra. 32 K time domain points were collected. Three different groups of spectra have been acquired: 1) from fasting patient; 2) ninety minutes after eating; 3) one-hundred-fifty minutes after eating.

2.1.2.3.3 CELL SPECTRUM WITH PRESATURATION AND WATERGATE SUPRRESSION

A one-dimensional spectrum has been measured from a sample containing african green monkey fibroblast cells [Couillard-Depres et al, 2004]. COS7 cells were grown in Dulbecco's modified essential medium (DMEM) containing 10% fetal bovine serum, 4 mM glutamine, 1.5 mg/ml sodium bicarbonate, 4.5 mg/ml glucose, 1mM sodium pyruvate, 100 U/ml penicillin and 100 μg/ml streptomycin. 1-10 millions of cells per sample were washed twice in phosphate-buffered saline (PBS) and embedded in ultralow gelling point agarose (Sigma Aldrich; 1 % agarose in PBS solution containing 10 % D_2O and 40 μM DSS) to avoid inhomogeneous distributions and sedimentations inside the 5 mm NMR tubes. Samples were cooled to 5°C and NMR measurement started within 15 min thereafter. During measurement, the temperature was kept at 5°C [Ramm et al, 2009]. The spectrum was recorded at 278 K and 800.13 MHz using a DQD acquisition mode, a mixing time of 10 ms and a relaxation delay of 1 s. The spectral width was 12 kHz, the data were recorded using oversampling [Moskau, 2002] and digital quadrature detection. 64 K time domain points were gathered. Solvent presaturation [Hoult, 1976] and watergate [Piotto et al, 1992; Sklenar et al, 1993; Saudek et al, 1994; Liu et al, 1998] solvent suppression have been applied on the spectrum.

2.1.2.4 SSA AND ICA OF ONE-DIMENSIONAL SPECTRA

2.1.2.4.1 URINE SPECTRA WITH SOLVENT PRESATURATION

The one-dimensional human urine spectrum was recorded with a Bruker Avance-600 spectrometer operating at a ^{1}H frequency of 600.13 MHz. It was acquired using oversampling [Moskau, 2002] and digital filtering (Bruker DQD mode). A NOESY-type 1D pulse sequence was used for the sample, including a selective presaturation of the solvent resonance [Hoult, 1976] of 5 s and a spoiler z-gradient pulse applied during the mixing time. For obtaining standardized conditions 133 mM sodium phosphate buffer, pH 7.4, 5% D_2O, and 0.1 mM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) was added. The spectrum was recorded at 298 K using a DQD acquisition mode, a mixing time of 10 ms and a relaxation delay of 5 s. The spectral width was 20.0276 ppm, the data was recorded using oversampling [Moskau, 2002] and digital quadrature detection. 128 K time domain points were sampled. Two datasets of experiments have been measured with a different NOE mixing times of 10 and 20 ms in the first dataset and of 1500 and 2000 ms in the second case.

2.1.2.5 ICA-TAILORED ONE-DIMENSIONAL SPECTRA

The one-dimensional experimental NMR spectra have been measured from a sample containing 1mM of uniformly ^{15}N-enriched HPr protein from *Staphylococcus carnosus* 95% H_2O/5% D_2O, pH 7. The NMR spectra were recorded on Bruker Avance-600 operating at 600 MHz employing a mixing time of 10 ms, a relaxation delay of 1 s, a spectral width of 14.9872 ppm and with 32 K time domain points (including 140 points of the group delay). The water signal was reduced by selective pre-saturation. All spectra were measured at 298 K. In particular, two datasets made up each one of two experiments have been obtained with specific pulse sequences. The first dataset has been generated measuring two experiments with a different phase cycling, while the second one is made up of two experiments acquired with different diffusion times (gradient weights of 80 G/cm and 50 G/cm for each case).

2.1.2.6 EMD OF ONE-DIMENSIONAL SPECTRA

2.1.2.6.1 SPECTRUM OF A METABOLITE MIXTURE OF FIVE AMINO ACIDS

The one-dimensional spectrum of a metabolite mixture [Snyder et al, 2008] made up of five different amino acids (S,G,V,T,L) with a concentration of 0.1 mg/ml has been recorded from a sample containing, in addition to the residues, 80% D_2O including 100 µM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid). It has been measured at 600.13 MHz (Bruker Avance-600) employing a qsim acquisition mode and a mixing time of 10 ms. The water signal was reduced by selective presaturation [Hoult, 1976] of 20.1 s. The spectrum has been recorded using a relaxation delay of 20 s and with 64 K time domain points. The spectral width of the spectrum is 20.0 ppm. It has been measured at 278 K.

2.1.2.6.2 SPECTRUM OF A METABOLITE MIXTURE OF TWENTY AMINO ACIDS

The one-dimensional spectrum of a sample made up of twenty amino acids [Snyder et al, 2008] with a concentration of 0.1 mg/ml has been recorded in 80% D_2O including 100 µM DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) on a Bruker Avance-600 spectrometer. The acquisition parameters are the following: acquisition mode, qsim; mixing time, 10 ms; relaxation delay, 20 s; time domain data, 64 K; spectral width, 20.0 ppm. It has been recorded at 278 K with water presaturation [Hoult, 1976] of 20.1 s.

2.1.3 PROTEIN STRUCTURE DETERMINATION: PSCD4-DOMAIN

The sequential assignment [see e.g. Sattler et al, 1999] has been obtained using the following two- and three-dimensional experiments [Wenzler, 2003]:

1. HNCA
2. HNCO
3. HACACO
4. HBHA(CO)NH
5. CBCA(CO)NH
6. 3D ^{1}H ^{15}N-TOCSY-HSQC
7. 3D ^{1}H ^{15}N-NOESY-HSQC
8. 3D ^{1}H ^{13}C-NOESY-HSQC
9. 3D HCCH-TOCSY
10. 3D HNCA-E.COSY
11. 2D ^{1}H ^{15}N-HSQC

12. 2D TOCSY
13. 2D NOESY

The first eleven spectra have been measured on a DRX-600 spectrometer operating at 600.13 MHz, while the last two has been acquired on a DRX-800 spectrometer operating at 800.13 MHz. All the spectra have been measured at 298 K. The sample used for the NMR measurements contained 10 mM sodium-phosphate buffer with a protein concentration of 10 mg/ml and 10% D_2O. In addition, 0.1 mM EDTA, 1 mM NaN_3, 1 μM Leupeptin, 1 μM Pepstatin, 1 μM BPTI and 0.1 mM DSS as internal reference have been added in the solution.

The sequential assignment of the atoms of the backbone has been obtained using the first eight spectra listed above. The side chain assignment has been performed using the experiments number 4, 5, 6, 8, 12 and 13. The assignment of the NOESY spectrum (number 13) has been automatically performed using the KNOWNOE routine [Gronwald et al, 2002].

The assignment furnished chemical shift restraints used to predict secondary structures. Statistical analysis has demonstrated the relationship existing between the chemical shift of certain atoms of the backbone and the presence of specific secondary motives [Wishart et al, 1995]. The TALOS+ [Corneliescu et al, 1999] program has been used to predict those secondary structures from the chemical shifts reported in Appendix A. It has produced dihedral angle restraints (φ angles) analyzing a database of homologous tri-peptides.

The $^3J_{H^N H^\alpha}$ coupling restraints have been observed (Appendix B) in the HNCA-E.COSY (number 10) spectrum [Griesinger et al, 1987]. Such restraints have been compared and added to those ones obtained with TALOS+ (Appendix C).

The distance restraints have been obtained from NOESY spectra (number 13) using the REFINE routine [Trenner, 2006]. Lower and upper limits on those automatically computed distances have been manually adjusted [Kalbitzer and Hengstenberg, 1992].

Hydrogen bonds restraints (Appendix D) have been obtained using an HNCO spectrum with a longer transfer delay (66 ms) respect with the experiment number 2 (16.5 ms) [Cordier et al, 2008; Cordier and Grzesiek, 1999].

Residual dipolar coupling restraints (Appendix E) have been obtained from the ^1H-^{15}N HSQC experiment that has not been decoupled in the ^{15}N-dimension (number 11). The anisotropic solution has been obtained adding bicelle as co-solute [Sanders et al, 1994; Sanders et al, 1992; Cavagnero et al, 1999].

All the spectra have been acquired and processed with XWINNMR 2.6. The three-dimensional structure has been determined using the CNS program [Bruenger et al, 1998].

2.2 Methods

2.2.1 SIGNAL DECOMPOSITION: PCA

Principal Components Analysis (PCA) [firstly developed by Pearson in 1901] is also known as Proper Orthogonal Decompositions (POD) since it is an orthogonal linear transformation of the data. It transforms the original signal **x** in a new coordinate system that is built projecting the data along the N directions spanned by the eigenvectors with the highest variances. Therefore, the orthogonal components can reveal the variance of the data in a space of reduced dimensionality. It is in fact often used to find a lower dimensional representation **y** of the original zero-mean data **x**. The first component y_1 can be defined as a linear combination of the K elements of the signal **x** weighted by scalar coefficients w_i that maximize the variance of y_1, as described in eq. 2.1

$$y_1 = \sum_{i=1}^{K} w_i x_i = \mathrm{w}_1^T \mathrm{x} \tag{2.1}$$

The unitary norm constraint $||\mathrm{w}|| = 1$ on the weighting vector **w** is additionally imposed in order to maximize the following criterion:

$$PCA\ (\mathrm{w}_1) = \mathrm{E}\{y_1^2\} = \mathrm{E}\{(\mathrm{w}_1^T \mathrm{x})^2\} =$$
$$= \mathrm{w}_1^T \mathrm{E}\{\mathrm{x}\mathrm{x}^{\mathrm{T}}\}\, \mathrm{w}_1 = \mathrm{w}_1^T \mathrm{R}\, \mathrm{w}_1 \tag{2.2}$$

in which E defines the expected value and R represents the (NxN) covariance matrix of **x** that is symmetric and positive definite. In particular, if **x** is a vector it corresponds to the variance of **x**. R may also represent the correlation matrix [Borgognone et al, 2001].

The PCA problem consists in finding the eigenvectors e_i and eigenvalues λ_i of R, which corresponds to the eigenvalue decomposition of R:

$$R = E\Lambda E^T \tag{2.3}$$

where **Λ** is a diagonal matrix containing the eigenvalues λ_i in a descent order ($\lambda_i \geq \lambda_2 \geq \cdots \geq \lambda_N$), while **E** represents the matrix of eigenvectors.

The scalar coefficient w_1 maximizing eq. 2.2 corresponds to the first eigenvector e_1 related to the eigenvalue λ_1 encompassing the highest variance of **x**. The projection of the original data along the direction spanned by the first eigenvector describes the first component (that one with the highest variance):

$$y_1 = e_1^T x \tag{2.4}$$

The successive components must fulfill the criterion reported in eq. 2.2 and they must be uncorrelated (orthogonal) with the previous ones:

$$E\{y_i y_j\} = E\{(w_i^T x)(w_j^T x)\} = w_i^T R\, w_j = 0 \tag{2.5}$$

Iteratively, it yields:

$$w_k = e_k \quad \text{and} \quad y_k = e_k^T x \tag{2.6}$$

Writing eq. 2.6 in a matrix form, we obtain:

$$Y = E^T X \tag{2.7}$$

As a consequence of eq. 2.6 the variances of the principal components are described by the eigenvalues of R in accordance to eq. 2.8:

$$\mathrm{E}\{y_k^2\} = \mathrm{E}\{e_k^T x x^T e_k\} = e_k^T R e_k = \lambda_k \quad (2.8)$$

Typically, the principal components related to the largest eigenvalues describe the dominant signals while those ones corresponding to the smallest eigenvalues represent the noise.

The principal components that eventually need to be discarded correspond to some of the projections of the signal along the directions spanned by the eigenvectors (i.e. the *k*th component is discarded as in eq. 2.9). They are nullified (see eq. 2.9) before transforming back to the original coordinate system (eq. 2.10):

$$y_k^{null} = \mathrm{e}_{k_{null}}^T \mathrm{x} = 0 \quad (2.9)$$

$$\mathrm{X}^{\mathrm{new}} = \mathrm{E}_{null}\ \mathrm{E}_{null}{}^{\mathrm{T}}\mathrm{X} = \mathrm{E}_{null} Y \quad (2.10)$$

The mean can be added on the reconstructed signal as last step.

2.2.2 SIGNAL DECOMPOSITION: SSA

The Singular Spectrum Analysis (SSA) was initially published by Broomhead in 1986 [Broomhead and King, 1986]. Actually, SSA is applied with several purposes as smoothing, extraction of periodicities or trends and many others. The theoretical aspects of SSA have been widely described (see Danilov and Zhigljavsky, 1997 and Golyandina et al, 2001).

SSA encompasses two different stages: decomposition and reconstruction. The former involves embedding of the signal and eigenvalue decomposition of the covariance matrix of the embedding as described in eq. 2.3. SSA is in fact an extension of PCA that allows the decomposition of an embedded one-dimensional time series into a sum of M components. The second stage consists in components selection and diagonal averaging of the reconstructed dataset with an implicit reverse embedding.

Time series analysis techniques often perform the embedding of one-dimensional signals, in the space of their time-delayed coordinates (e.g. Zhu et al, 1997). Embedding can be considered as a mapping that transforms a one-dimensional time series $x = (x[0], x[1], \dots, \dots, x[N-1])^T$ into a sequence of M time-lagged vectors $x = (x[0], x[1], \dots, \dots, x[N-M])^T$. The zero-mean time signal x of length N is embedded in its delayed coordinates with an (N-M) window size, to build up a trajectory matrix **X** (see eq. 2.11) whose rows constitute the lagged vectors. The embedding requires defining the window size that can vary in the range $2 \leq M \leq N$.

$$X = \begin{bmatrix} x[M-1] & x[M] & .... & ... & x[N-1] \\ x[M-2] & x[M-1] & .... & ... & x[N-2] \\ x[M-3] & x[M-2] & x[M-1] & ... & x[N-3] \\ .... & .... & .... & & .... \\ .... & .... & .... & & .... \\ x[1] & x[2] & x[3] & .... & x[N-M+1] \\ x[0] & x[1] & x[2] & .... & x[N-M] \end{bmatrix} \quad (2.11)$$

The embedding dimension can be estimated using model order selection techniques [Liavas and Regalia, 2001]. Generally a fixed shift of $\Delta n_t = 1$ is used between the lagged vectors. The trajectory matrix **X** thus contains as many time lagged copies of the signal **x** as the number of components, shifted by one data point for every row of the data matrix. Moreover, the trajectory matrix **X** of (Mx(N-M)) dimensions is a Töplitz matrix which means that it has identical entries along its (top left to bottom right) diagonals.

The eigenvalue decomposition of the trajectory matrix **X** is then applied in accordance to eq. 2.3. The (*i*, *j*) entry of the covariance matrix R of the trajectory matrix **X** is obtained as following:

$$R_{ij} = E\left[(X_{i,k} - \mu_i)(X_{j,k} - \mu_j)\right] \quad (2.12)$$

where $u_i = E(X_i)$ and $u_j = E(X_j)$ are the expected values of the *i*th and the *j*th (among the M rows) row vector (of dimension 1xN-M) of **X** and k =1,...,N-M.

Any signal x_k that constitutes the columns of **X**, is projected along the directions spanned by the eigenvectors related with the eigenvalues of the covariance matrix. Reduction of the dimension can be obtained at this point by projecting the data x_k

only onto the L< N directions defined by the L eigenvectors related with the L largest eigenvalues. This process represents a denoising procedure since the eigenvectors related to the smallest eigenvalues encompass just noise.

The reconstruction stage after nullifying some of the projections (see eq. 2.9) leads to a new set of vectors x'_k forming the estimated trajectory matrix $\mathbf{X}'$. In particular, the elements along each descending diagonal of $\mathbf{X}'$ are no more identical as in the original trajectory matrix **X**. This is repaired by replacing the entries in each diagonal by their average, obtaining again a Töplitz matrix $\mathbf{X_r}$. This procedure provides the best approximation so that the Frobenius norm of the difference ($\mathbf{X_r} - \mathbf{X}'$) is minimal [Golyandina et al, 2001; Teixeira et al, 2008]. Therefore, the one-dimensional signal $\mathbf{x}'$ is reproduced by reverting the embedding.

In summary, in order to perform SSA, the (Mx(N-M)) time domain signal *x* must be embedded into the trajectory matrix **X** (eq. 2.11) and the eigenvalue decomposition of the (MxM) covariance matrix **R** of **X** must be computed (eq. 2.12). The eigen-representation of **R** yields an (MXM) diagonal matrix **Λ** of eigenvalues and an (MXM) matrix **E** of eigenvectors. The projections along the eigenvectors related to certain eigenvalues (i.e. the smallest for denoising purposes) are nullified (eq. 2.9) and the new (Mx(N-M)) trajectory matrix $\mathbf{X}'$ is obtained transforming back to the original coordinate system (eq. 2.10). Diagonal averaging is applied on $\mathbf{X}'$ replacing every element along each descending diagonal by its averaged element. The reconstructed one-dimensional signal $\mathbf{x}'$ is finally obtained by reverting the embedding process. The mean can be added again to the reconstructed signal.

The reconstruction process can proceed in two different ways which should be equivalent in principle. Nullifying the projection related to the smallest eigenvalue (i.e. for denoising purposes) corresponds to reconstructing the signal defined by the last component (with the lowest variance) and to subtract it from the original data, i.e. $\mathrm{y}(i) = \mathrm{x}(i) - \mathrm{x}'(i)$.

2.2.3 SIGNAL DECOMPOSITION: ICA

The Independent Component Analysis (ICA) [Comon, 1994] belongs to the class of Blind Source Separation (BSS) methods [Hyvärinen et al, 2001]. It has been successfully applied on EEG data revealing brain activities [Makeig et al, 1996; Jung et al, 1998; Vigario et al, 2000; Krishnaveni et al, 2006] and for feature extraction purposes from image and audio signals [Bell and Sejnowski, 1995; Bell and Sejnowski, 1997]. The PCA [Joliffe, 1986] extracts uncorrelated components by means of second order statistics (variance maximization), while ICA looks for independent sources using higher order statistics (i.e. non-Gaussianity maximization). 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 *i*th observed signal ($\mathbf{x_i}$) can be represented as a weighted sum of several sources (*m* components), denoted by s_i. This can be expressed as follows:

$$x_i(t) = a_{i1}s_1(t) + a_{i2}s_2(t) + \ldots + \ldots a_{im}s_m(t) \quad (2.13)$$

The number of mixtures (i=1,….,N) must be at least equal to the number of underlying sources ($N \geq m$), as N microphones located at different positions in a room, recording *m* talking persons (sources) in the well-known Cocktail-party problem [Comon, 1994]. In this case the a_{ij} mixing coefficients are determined by the different distances of each person from every microphone.

In order to estimate a_{ij} it must be assumed that s_j and s_{j+1} are statistically independent at all times *t*. The set of mixtures can be represented as a matrix of vectors $\mathbf{x} = (x_1, x_2, \ldots, \ldots, x_N)^T$ of length M, with the following expansion of eq. 2.13:

$$\mathbf{x} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^M \\ x_2^1 & x_2^2 & \cdots & x_2^M \\ \ldots & \ldots & \cdots & \ldots \\ x_N^1 & x_N^2 & \cdots & x_N^M \end{pmatrix} = \quad (2.14)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{N1} & a_{N2} & \ldots & a_{Nm} \end{pmatrix} \begin{pmatrix} s_1^1 & s_1^2 & \cdots & s_1^M \\ s_2^1 & s_2^2 & \cdots & s_2^M \\ \ldots & \ldots & \cdots & \ldots \\ s_m^1 & s_m^2 & \cdots & s_m^M \end{pmatrix} = \mathbf{As}$$

where **x** is the (NxM) matrix of the observations at each time t = 1,…,M, while **A** represents the (Nxm) mixing matrix and **s** is the (mxM) matrix containing the *m* independent source signals at each time point t = 1,…,M.

The vector **x** must be centered by zero-mean normalization, involving that **s** is zero mean as well. The variance of the independent component cannot be determined since both **s** and **A** are unknown. As a consequence the complexity of the estimation

problem must be reduced by means of a whitening pre-processing step restricting the sources to have unit variance. Due to the whitening process a new orthogonal $\tilde{\mathbf{A}}$ mixing matrix is obtained whose free parameters to be estimated are notably reduced. Instead of estimating any arbitrary full-rank matrix **A**, the orthogonal matrix $\tilde{\mathbf{A}}$ can be easily found.

This method allows a faster approach to the decomposition problem but it obviously leads to ambiguities on the magnitude and on the sign of the independent components. Moreover, for the same reason the order of the terms in eq. 2.13 may be permutated conducting to indeterminacy on the order of the estimated independent components. As a consequence the independent component analysis requires a visual inspection of the extracted sources in order to identify the undesired ones (i.e. artifacts) and to remove them. A semi-automated ICA method has been proposed by Delorme [Delorme et al, 2001], while Joyce [Joyce et al, 2004] and Nicolaou [Nicolaou and Nasuto, 2004] have presented automated identification of the components via correlation metrics and support vector machines (SVM) respectively.

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

$$\hat{\mathbf{s}} = \begin{pmatrix} \hat{s}_1^1 & \hat{s}_1^2 & \cdots & \hat{s}_1^M \\ \hat{s}_2^1 & \hat{s}_2^2 & \cdots & \hat{s}_2^M \\ \cdots & \cdots & \cdots & \cdots \\ \hat{s}_m^1 & \hat{s}_m^2 & \cdots & \hat{s}_m^M \end{pmatrix} = \tag{2.15}$$

$$\begin{pmatrix} u_{11} & u_{21} & \cdots & u_{N1} \\ u_{12} & u_{22} & \cdots & u_{N2} \\ \cdots & \cdots & \cdots & \cdots \\ u_{1m} & u_{2m} & \cdots & u_{Nm} \end{pmatrix} \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^M \\ x_2^1 & x_2^2 & \cdots & x_2^M \\ \cdots & \cdots & \cdots & \cdots \\ x_N^1 & x_N^2 & \cdots & x_N^M \end{pmatrix} = \mathbf{Ux}$$

The statistically independent non-Gaussian components are obtained by means of high order moments. ICA can be defined as a method that finds a linear transformation which maximizes the non-Gaussianity or equivalently minimizes the mutual information of the sources [Kraskov et al, 2004]. Exploiting the central limit theorem it is straightforward that the distribution of a mixture of many variables tends to have a Gaussian behavior, thus minimizing it would lead to the optimal signal decomposition.

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

Computing the kurtosis *k* or the fourth-order cumulant of a random variable *y*, it is possible to measure the non-Gaussianity:

$$k\,(y) = E\{y^4\} - 3(E\{y^2\})^2 \tag{2.16}$$

In case of a Gaussian variable *y*, the fourth moment $E\{y^4\}$ corresponds exactly to $3(E\{s^2\})^2$ thus, the kurtosis is zero for a Gaussian variable. Sub-Gaussian are the random variables yielding a negative kurtosis, while super-Gaussian are those ones having positive kurtosis. Starting from a weighting vector **w**, the direction in which the kurtosis of $y = \boldsymbol{w^T x}$ increases (positive kurtosis) or decreases (negative kurtosis) more strongly is computed. At this point a gradient method is applied in order to find a new vector **w**.

Using the kurtosis as a measure of non-Gaussianity has some drawbacks [Huber, 1985] thus, typically the negentropy is alternatively used.

Considering the differential entropy *H* of a random vector **y** with density *f* (**y**):

$$\mathrm{H}(y) = -\int f(\mathrm{y}) \log f(\mathrm{y}) \mathrm{dy} \tag{2.17}$$

it is straightforward the definition of negentropy **J**:

$$\mathrm{J}\,(y) = \mathrm{H}\left(\mathrm{y}_{\mathrm{gauss}}\right) - \mathrm{H}\,(y) \tag{2.18}$$

where $\mathbf{y}_{\mathbf{gauss}}$ is a Gaussian random vector. A Gaussian variable has the largest entropy, thus the negentropy can be used to measure the non-Gaussianity of the extracted sources.

The negentropy **J** is typically approximated using higher-order moments [Jones and Sibson, 1987; Hyvärinen, 1998]:

$$J\,(y) \approx c\,[E\{G\,(y)\} - E\{G\,(\upsilon)\}]^2 \tag{2.19}$$

in which G is any non-quadratic function [Hyvärinen, 1998], c is a constant and υ is a Gaussian variable. G(y) can be arbitrarily chosen and if G(y) = y^4, a kurtosis-based approximation would be obtained.

The most robust estimators are defined in eq. 2.20 and 2.21:

$$G_1(y) = \frac{1}{a_1}\log\cosh a_1\, y \tag{2.20}$$

$$G_2(y) = -e^{\frac{-y^2}{2}} \tag{2.21}$$

where $1 \leq a_1 \leq 2$.

The mutual information **I** defines the dependence between *n* random variables and it can be expressed in terms of negentropy:

$$\mathrm{I}\,(y_1, y_2, \ldots, y_n) = \mathrm{J(y)} - \sum_{i=1}^{m} J\,(y_i) \tag{2.22}$$

The mutual information has to be minimized or equivalently the negentropy of the extracted sources needs to be maximized (it is zero only if **y** has a Gaussian distribution).

Since there are several cost functions, different ICA methods have been developed such as FastICA [Hyvärinen and Oja, 1997], InfoMax [Bell and Sejnowski, 1995; Lee and Sejnowski, 1996], JADE [Cardoso, 1999], SOBI [Beloucharani et al, 1997], MILCA [Kraskov et al, 2004] and other similar algorithms.

The FastICA uses a fixed-point iteration scheme to maximize the non-Gaussianity of the sources. In order to provide an approximation of negentropy, the algorithm used originally kurtosis as contrast function, but more recent versions apply the hyperbolic tangent, exponential and cubic functions. *Hyvärinen* highlighted that the ordering of the sources is influenced by the contrast function used. In particular, using kurtosis as contrast function, the super-Gaussian sources tend to be the first ones. Ordinary ICA algorithms (InfoMax) are based on gradient descent whose

convergence is slow (linear). The FastICA has allowed a faster and more reliable learning, using a fixed-point iteration algorithm with a cubic convergence, thus it has been used in the present work.

The InfoMax exploits an adaptive learning algorithm which maximizes the information passed through a neural network. The joint entropy H(*y*) of the outputs of the network is maximized which corresponds to a minimization of the mutual information among the outputs. *Bell and Seinowski* initially proposed an algorithm well-suited for separating signals with positive kurtosis (super-Gaussian distribution), while *Lee and Sejnowski* extended the original version of the InfoMax to make it able to handle both super- and sub-Gaussian distributions.

The JADE algorithm (Joint Approximation Diagonalization of Eigenmatrices) calculates the maximization by means of joint diagonalization. This method can be described in two steps:

1) diagonalization of the covariance matrix (eigenvalue decomposition) of the mixtures
2) diagonalization of the kurtosis matrices of the observations.

The SOBI (Second-Order Blind Identification) algorithm relies on joint-diagonalization of time delayed second order covariance matrices. MILCA (Mutual information based Least dependent Component Analysis) minimize the mutual information between the estimated sources.

2.2.4 SIGNAL DECOMPOSITION: EMD

The Empirical Mode Decomposition (EMD) method [Huang et al, 1998] can be seen as an exploratory data analysis technique. It is ideally suited to decompose any non-stationary time-domain signal into its oscillatory components. In general, it aims to decompose an arbitrary signal via a, usually small, number of IMFs (intrinsic mode functions) and the residual *r*. The IMFs represent zero-mean amplitude and frequency modulated components. The EMD relies on a process called sifting which allows the decomposition of the signal into a finite set of oscillatory components.

It has been successfully applied to solve many practical problems [Wu et al, 2001; Coughlin and Tung, 2004; Wang et al, 2008; Lo et al, 2008; Lo et al, 2009; Cong et al 2009].

On a signal *x(t)* the EMD performs the mapping:

$$x(t) = \sum_n x_n(t) + r(t) \tag{2.23}$$

where the $x_n(t)$ term, with n=1,....,N, denotes the set of IMFs and r(t) is the trend within the data (also known as the last IMF or residual). Any IMF is a function which is characterized by the following properties (criteria):

a) the upper and the lower envelopes of the signal have to be symmetric (the local average is zero).
b) the number of zero-crossings and the number of extremes are equal or they differ at most by one.

The IMF represents an oscillatory mode with a variable amplitude and frequency along the time axis.

In order to extract the IMFs from a signal, the sifting algorithm [Huang et al, 1998] is applied as described below:

(1) Identify all local maxima and minima of $x_k(t)$

(2) Connect all the local maxima of the signal by cubic spline yielding the upper envelope $U_k(t)$ of the signal

(3) Connect all the local minima of the signal by cubic spline yielding the lower envelope $L_k(t)$

(4) Subtract the mean envelope $m_k(t) = \frac{1}{2}(U_k(t) - L_k(t))$ from the signal yielding the most oscillating pattern

$$x_{k+1}(t) = x_k(t) - m_k(t) \tag{2.24}$$

(5) Verify all IMF criteria. If not satisfied repeat all previous steps on $x_{k+1}(t)$, otherwise set:

$$x_n(t) = x_{k+1}(t) \tag{2.25}$$

and

$$r_{n+1}(t) = r_n(t) - x_n(t) \tag{2.26}$$

(6) The algorithm is completed if the residual of Step 5 is a monotonous function. If not, the same procedure is applied on the residue signal in order to identify the next highest oscillating pattern.

Typically after extracting all IMFs, the components are further analyzed by applying the Hilbert-Huang transform or by processing them in any other suitable way [Quiroga et al, 2000; Quiroga et al, 2002]. The Hilbert spectral analysis of the IMFs is particularly suitable for non-stationary and non-linear data in order to obtain instantaneous frequency information evolving with time.

3 Singular Spectrum Analysis of NMR data

3.1 Solvent suppression and baseline correction

3.1.1 GENERAL CONSIDERATIONS

Time series embedding in the space of time-delayed coordinates (by means of a trajectory matrix $\mathbf{X}$) is applied on the measured FID of length N. This procedure is straightforward when a one-dimensional spectrum is analyzed, while it must be iteratively repeated over all the rows (representing the same FID acquired at a varying evolution time t_1) of a multi-dimensional spectrum. This latter in fact consists of Q FIDs x^i, $(i = 1, \dots \dots, q)$, each one of length N. As many trajectory matrices ($\mathbf{X^Q}$) as the number of acquired FIDs (that corresponds to the size of the time-domain t_1 in the indirect direction) must be constructed and the eigenvalue decomposition has to be separately performed. The embedding dimension M of the trajectory matrices has been determined empirically. The projections of the eigenvectors related to the eigenvalues of interest can be nullified and new trajectory matrices $\mathbf{X}'$ are obtained. In particular, the eigenvector related to the largest eigenvalue describing the component with the highest variance (thus corresponding to the dominant signal) has been rejected in order to obtain solvent suppression. Reverting the embedding leads to a modified FID (not containing water artifacts), one for each row of a multi-dimensional spectrum.

Pre-processing of the NMR spectra is mandatory and it includes:

(1) the computation of the number of time-domain points belonging to the group delay (GRPDLY in accordance to the Bruker parameters) due to the oversampling and to the digital filtering;

(2) the signal normalization to unit norm (z-transform).

Post–processing steps are instead related to the inverse normalization, to the Fourier transformation, to phase correction in accordance to the group delay points and to baseline correction in the frequency domain. A general schema of the SSA application on NMR data for water suppression is presented in Fig. 3.1.

The performance of the solvent suppression by means of the SSA has been evaluated applying this technique on several dataset in order to determine its ability to deal with analog or digital (par. 2.1.2.2.1) spectra (group delay management), with mixed (time-frequency) domain data (in multi-dimensional cases as described in par. 2.1.1.1) and with spectra whose dominant signal is not the solvent (as in case of watergate suppression [Piotto et al, 1992; Sklenar et al, 1993; Saudek et al, 1994; Liu et al, 1998]as in par. 2.1.2.3.1, or in case of back-calculated two-dimensional data [Ried et al, 2004] not including any experimental solvent spectrum as in par. 2.1.1.1).

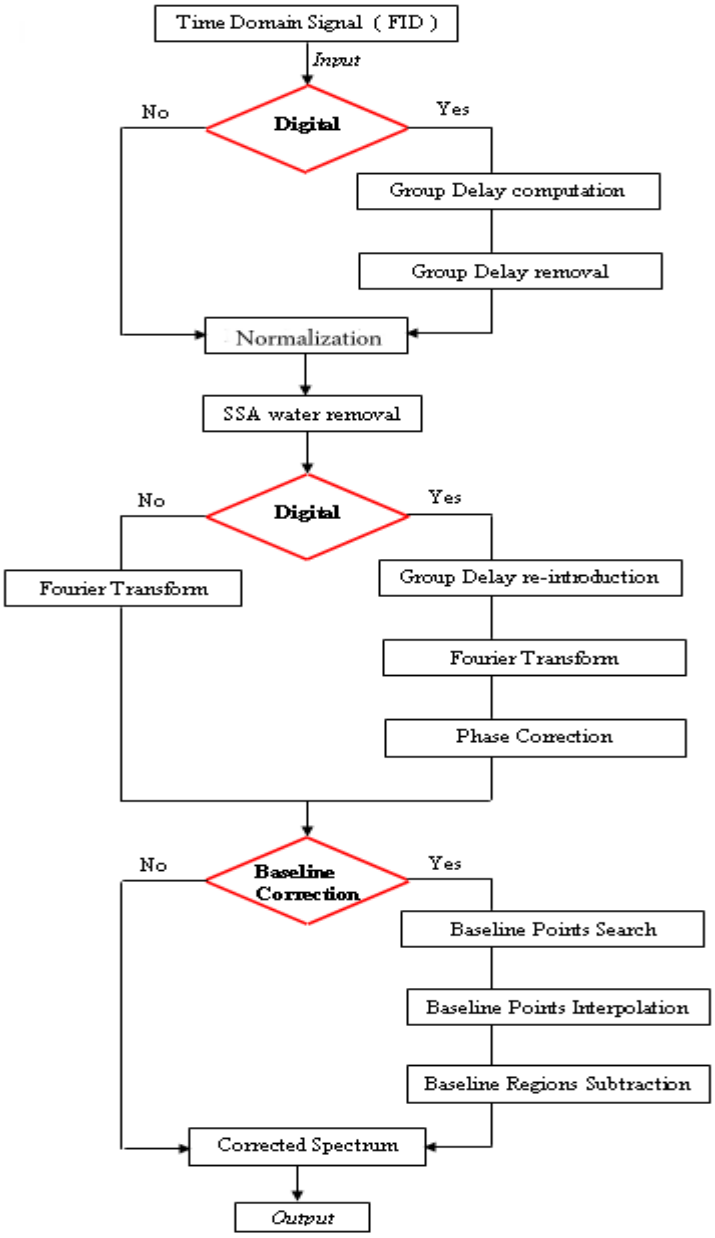


Figure 3.1 Flowchart describing the SSA application on NMR spectra: pre-processing encompasses group delay management and normalization in the time domain; post-processing is related to phase and baseline correction in the frequency domain [De Sanctis et al, 2010].

The most suitable embedding dimension M and the optimal solvent-to-solute ratio for a reliable solvent suppression have been investigated on the back-calculated one-dimensional HPr spectrum (par. 2.1.1.2).

The SSA has been applied on three- (par. 2.1.2.1), two- (par. 2.1.2.2) and one-dimensional (par. 2.1.2.3 and 2.1.2.4) experimental spectra and on two- (par.2.1.1.1) and one-dimensional (par.2.1.1.2) back-calculated data. Its application to one-dimensional cases has been compared with other techniques such as the EMD (in the time domain), the ICA and the PCA (in the frequency domain) as explained in chapter four.

3.1.1.1 PRE-PROCESSING:OVERSAMPLING AND DIGITAL FILTERING

Modern spectrometers use oversampling of the data followed by digital frequency filtering by finite impulse response filters (FIR) and a reduction of the stored spectral range [Moskau, 2002]. The Analog to Digital Converter allows data oversampling that has some advantages as mayor accuracy and increased dynamic range. As defined by the Bruker parameters, it means that they are not sampled according to the DW (dwell time) but to DWOV (oversampling dwell time). However, it yields more data points TD than those ones determined by the SW (spectral width) imposed from the Nyquist theorem, thus a decimation procedure must be applied in order to reduce them (in accordance to the Bruker parameters: DECIM= DW/DWOV).

The ideal digital filter would be a rectangular curve that matches perfectly the spectral window. A rectangular filter function in the frequency domain corresponds to a sinc function in the time domain that must be cut off at some point and whose consequent truncation effects are cured by some internal optimizing procedures. The filter function slides through the raw data with unknown manufacturer weights. FIR filtering leads to a delayed response where the first time points of the FID are corrupted. The dead time (group delay or GRPDLY) visible at the beginning of a digitally filtered FID corresponds to the time necessary for the filter to slide over the data. For example, using a sinc filtering function the highest intensity of such a filter lays in the center, thus the filtered FID starts to show a significant behavior only when this part of the filter reaches the beginning of the FID that may happen after N_g complex data points (e.g. 70 complex data points). The length of the GRPDLY reveals the steepness of the filter.

In a two-dimensional case the digital filter is always applied only on the direct dimension (t_2).

In Fig. 3.2 the first FIDs of each experimental two-dimensional NMR NOESY spectra (digital and analog acquisition mode) of HPr from *Staphylococcus aureus* (H15A) are depicted (par. 2.1.2.2.1). In particular, the signal in the upper part represents the

digitally filtered FID with a group delay of 71 real data points, while the analog one is in the bottom. They look very similar except in the initial (GRPDLY*DW) seconds, where in the digital case appears a dead time containing no significant information.

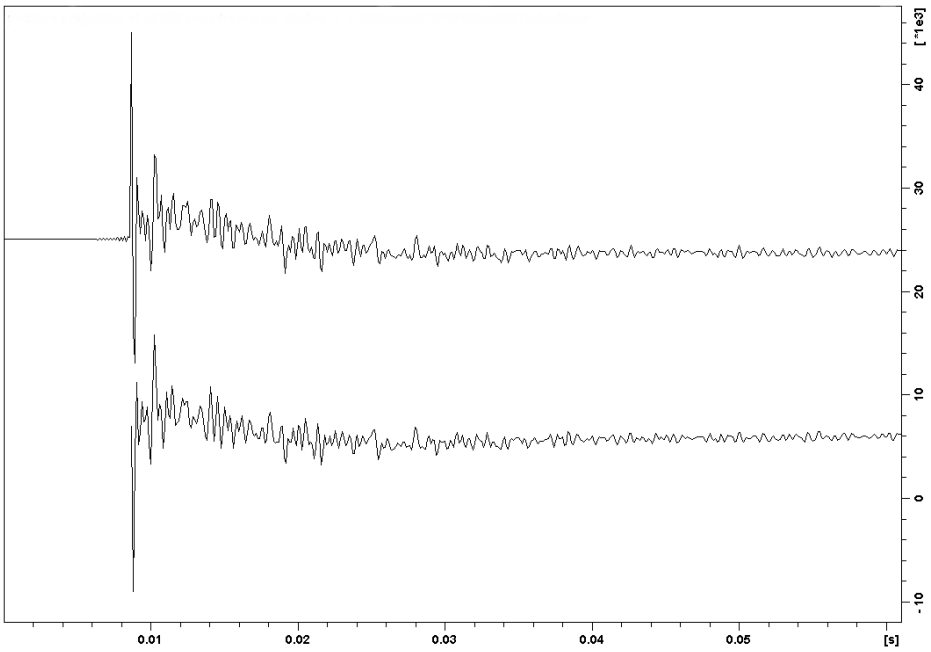


Figure 3.2 Oversampling and digital filtering effects: digital (upper) and analog (lower) FIDs (first rows) of the two-dimensional experimental NOESY spectra (par. 2.1.2.2.1) of HPr from *Staphylococcus aureus* (H15A).

A shift of N_g points in the time domain is translated to a very large first-order phase change after Fourier transforming which is clearly recognizable in the baseline wiggles of the spectrum. The Bruker manufacturer automatically calculates such correction using the decimation parameter that avoids showing those wiggles.

Generally, a circular shift of the points belonging to the group delay is applied in the time domain and a consequent phase correction (180° for each point of the GRPDLY) is performed after Fourier transforming the data. If the number of points belonging to the group delay is not exactly defined a baseline distortion would appear after Fourier transforming due to an incorrect phase correction. The Fig. 3.3 (part *a*) represents what typically occurs in the frequency domain if the GRPDLY is not properly managed (i.e. the first order phase correction is not correctly applied). The *b* part of the same figure represents instead the Fourier transform of the digitally filtered FID with a proper GRPDLY management.

The Bruker manufacturer does not always reveal the number of points belonging to the group delay (in the acquisition files). The user simply can recognize that in case of a digitally filtered FID, the parameter PKNL is typically set to TRUE that implies an

underlying automated first order phase correction in accordance to the number of points of the group delay.

For the purposes of this project, it was necessary to exactly know the number of data points belonging to the group delay, thus it is automatically computed from the DECIM and DSPFVS parameters, contained in the acquisition files.

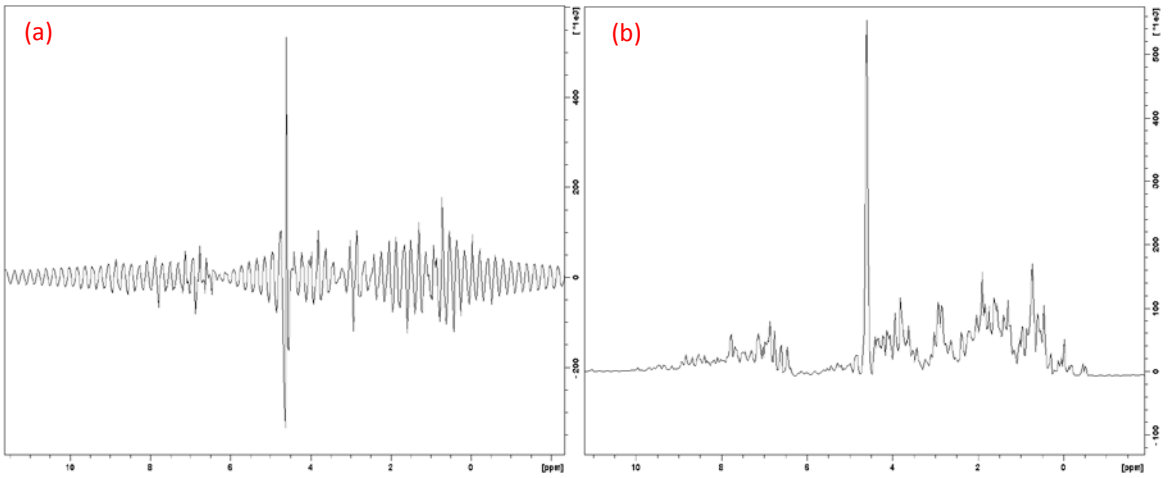


Figure 3.3 Group delay management: incorrect (left) and correct (right) first order phase correction in accordance to the number of time-domain data points belonging to the group delay. The first FID of the digitally filtered experimental two-dimensional NOESY NMR spectrum of HPr from *Staphylococcus aureus* (H15A) is reported (par. 2.1.2.2.1).

In particular, it has been experimentally proven that completely excluding those points from the SSA calculation leads to better solvent suppression. In order to obtain the optimal solvent removal without any distortion due to the presence of the group delay some attempts have been done. Initially the SSA has been performed on the complete FID, including the GRPDLY points. In this case the solvent removal is rather good but after the SSA, the initial part of the FID is distorted and it is no more zero, as typically happen in a digitally filtered case. The circular shift of such group delay points to the end of the FID before Fourier transforming leads to a step function yielding wiggles in the baseline. The proposed solution to overcome this problem is to completely exclude the GRPDLY part from the removal procedure. However, it has been empirically demonstrated that an equivalent result could be obtained including the GRPDLY in the SSA calculation and applying a baseline correction in the time domain (i.e. Bruker manufacturer BC_mode parameter set to quad) after SSA and before the circular shift in the following manner:

$$x_i = x_i - \frac{1}{\frac{TD}{4}} \left(\sum_{j=TD-\frac{TD}{4}}^{TD} x_j\right) \quad (3.1)$$

with i = 0,……,TD.

The average of the last quarter of the FID is calculated and subtracted from the whole FID (including the GRPDLY points) after the solvent removal, avoiding any step function connecting the final part of the FID and the circular shifted GRPDLY.

Considering that typically the user can decide whether to apply the time domain baseline correction (quad averaging in the Bruker BC_mode for zero filling compensation) or not and that the GRPDLY does not contain any additional information, it has been directly excluded from the SSA calculation [Malloni et al, 2010; De Sanctis et al, 2011].

Since the GRPDLY is not an integer number, it is rounded to the closest integer value representing the imaginary part of the complex number concluding the time delay, in order to avoid baseline distortions after SSA.

A complete substitution of the group delay with zero values has been attempted as well but it has not improved the solvent removal performance, yielding distortions and needing time domain baseline correction as described above. Moreover, it has been demonstrated that substituting the GRPDLY with zero values and applying the SSA on the whole FID produces the same results as keeping the original values of those points during the SSA since they do not contain relevant information.

A prior circular shift of the group delay to the end of the FID and the use of only the first half part of the signal as input to the SSA algorithm, has not improved the performance. It has generated wiggles in the frequency domain due to the step function connecting the part of the signal that undergoes to the algorithm and the second half of the FID. Signal averaging needs to be applied also in such case.

Therefore, for several time domain applications (as the SSA), the FID has to be left shifted by N_g data points, since the inclusion of these data leads to spectral artifacts. Before application of SSA, the data points belonging to the group delay are removed and stored for a subsequent reconstruction of the complete dataset.

As previously described (see par. 2.2.2) the dimensionality of each trajectory matrix is related to the embedding dimensions. Generally, fixing the shift of $\Delta n_t = 1$ between M lagged vectors, a trajectory matrix of (Mx(N-M)) dimensions is obtained. Excluding the group delay points, the zero-mean FID has a length of (1x(N-N_g)) and from it a trajectory matrix **X** of dimension MxQ with Q = (N-N_g) –M can be generated.

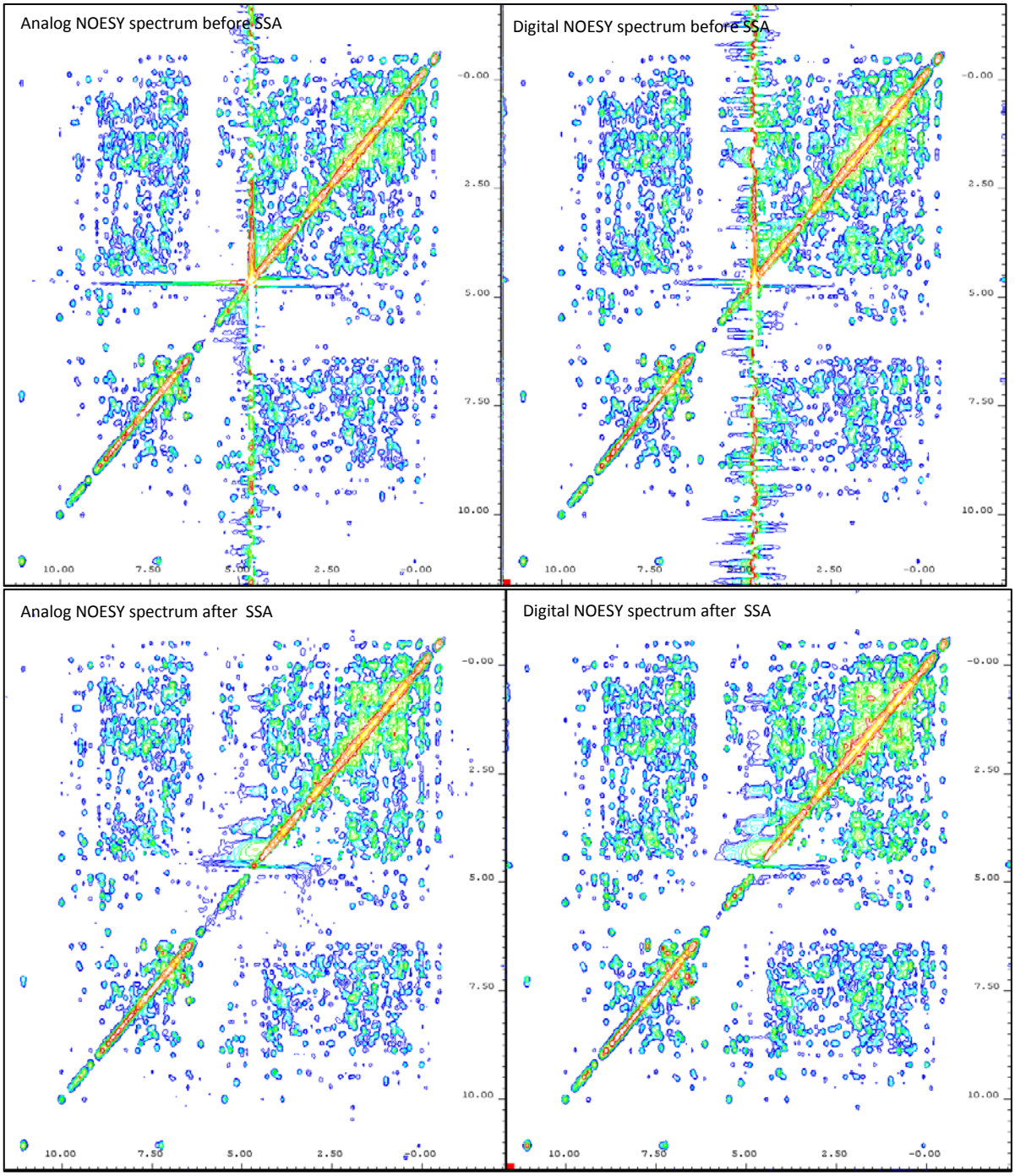


Figure 3.4 SSA application for solvent removal on digital and analog spectra: two-dimensional experimental NOESY spectrum of HPr from *Staphylococcus aureus* (H15A) with an analog (top left) and a digital (top right) acquisition (par. 2.1.2.2.1). SSA applied on the analog (bottom down) and on the digital (bottom right) spectra.

A digitally filtered two-dimensional NOESY spectrum of HPr from *Staphylococcus aureus* (H15A) has been measured (par. 2.1.2.2.1) and the SSA has been used to remove the solvent signal. The analog two-dimensional NOESY spectrum of the same protein (par. 2.1.2.2.1) has been obtained with the same acquisition parameters of the former and the SSA application on both spectra has been compared in Fig. 3.4. It is

evident that, if the group delay is excluded from the removal procedure, the performance of SSA to suppress the solvent becomes independent on the digitization mode used during the acquisition. The Fig. 3.4 shows the analog (top and down left side) and the digital (top and down right side) two-dimensional NOESY spectra before (upper part) and after (lower part) the solvent removal by means of the SSA.

3.1.1.2 PRE-PROCESSING: NORMALIZATION

As shown in Fig. 3.1 the group delay is excluded from the signal and normalization to unit norm (z-transform) is afterwards applied on each FID in order to avoid scale variations:

$$\hat{x}^i = \frac{x(t) - \mu}{\sigma} \tag{3.2}$$

with t = 1,….,N-N_g and

$$\sigma = \frac{1}{N - N_g} \sum_{t=0}^{(N-N_g)-1} (x(t) - \mu)^2 \tag{3.3}$$

where

$$\mu = \frac{1}{N - N_g} \sum_{t=0}^{(N-N_g)-1} x(t) \tag{3.4}$$

The trajectory matrix **X** is thus formed by M time-delayed copies of a zero-mean normalized FID of length (N-N_g):

$$\widehat{\mathbf{X}} = \begin{bmatrix} \hat{x}[M-1] & \hat{x}[M] & \cdots & \hat{x}\left[\left(N-N_g\right)-1\right] \\ \vdots & \ddots & \ddots & \vdots \\ \hat{x}[0] & \hat{x}[1] & \cdots & \hat{x}\left[\left(N-N_g\right)-M\right] \end{bmatrix} \quad (3.5)$$

The inverse z-transform is applied on the data after the SSA solvent removal before Fourier transforming.

3.1.1.3 SSA ON SPECTRA WHOSE SOLVENT SIGNAL IS NOT THE DOMINANT ONE

The SSA extracts the underlying components of each FID in accordance to the variance of the signal of each component. Typically, presaturation of the solvent signal [Hoult, 1976] is applied during the acquisition yielding a spectrum whose dominant signal is still the water. The SSA in such cases exploits the fact that the first component encompasses the greatest variance of the FID, thus it represents the solvent and it can be automatically discarded. In case that watergate solvent suppression [Piotto et al, 1992; Sklenar et al, 1993; Saudek et al, 1994; Liu et al, 1998] is applied during the measurement, the solvent signal is usually reduced to be no more the dominant one in the spectrum. If the SSA is applied in similar situations, the algorithm obviously detects the strongest solute signals and automatically removes them from the spectrum, while the water artifact is unchanged. In Fig. 3.5 is reported the application of the SSA on the one-dimensional spectrum of HPr protein from *Staphylococcus aureus* (H15A) with watergate solvent suppression (par. 2.1.2.3.1). It is evident that some resonances of the protein have been removed (red trace).

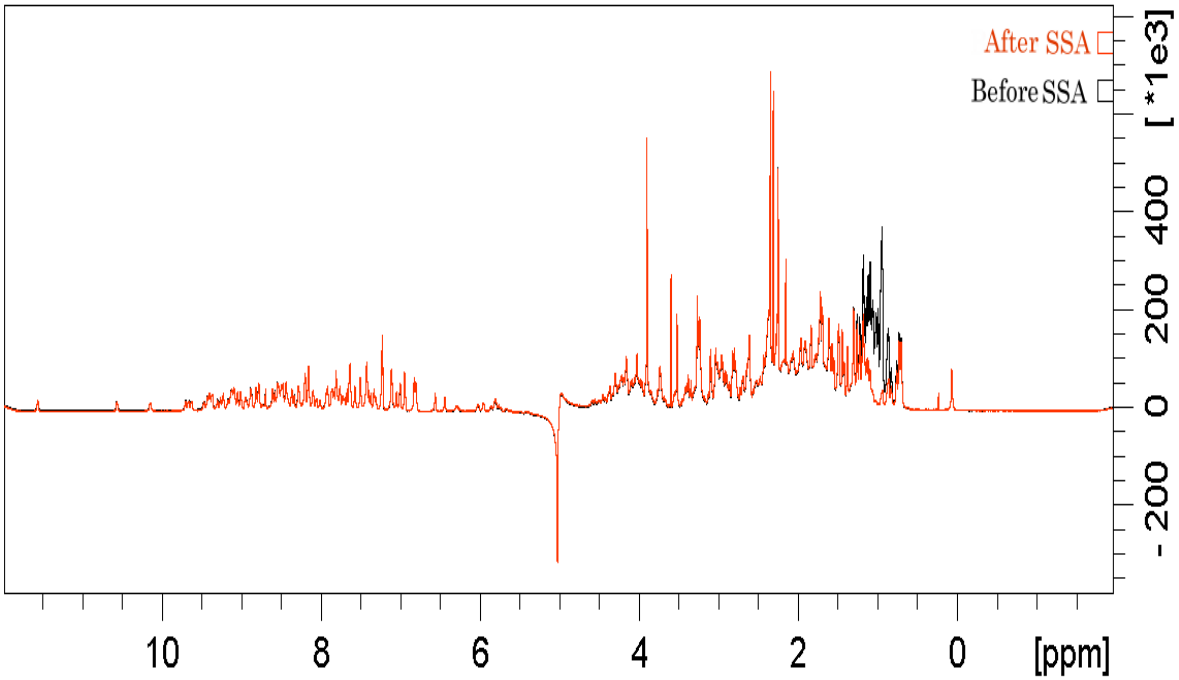


Figure 3.5 SSA of the one-dimensional spectrum acquired with watergate solvent suppression: a one-dimensional spectrum of HPr from *Staphylococcus aureus* (H15A) before (black trace) and after (red trace) SSA (par. 2.1.2.3.1). The solvent signal is not the dominant one in the spectrum, thus it is not properly recognized by the algorithm that automatically removes some resonances of the solute.

The SSA applied on the back-calculated (par. 2.1.1.1) two-dimensional spectrum of HPr from *Staphylococcus aureus* (H15A) obtained without adding the experimental water leads to general spectral distortions. In this case in fact the dominant signals are represented by some of the resonances lying along the diagonal that are automatically rejected by the algorithm. Whole stripes of artifacts are detected all over the spectrum along the w_1 direction spanned by the strongest diagonal peaks, as reported in Fig. 3.6.

If watergate solvent suppression [Piotto et al, 1992; Sklenar et al, 1993; Saudek et al, 1994; Liu et al, 1998] is applied on a multi-dimensional spectrum the SSA produces entire stripes of artifacts as well, thus it could be selectively applied only along those rows whose dominant signal is still the solvent. The algorithm for an automated identification of the rows of interest is proposed in the last chapter (in the discussion section).

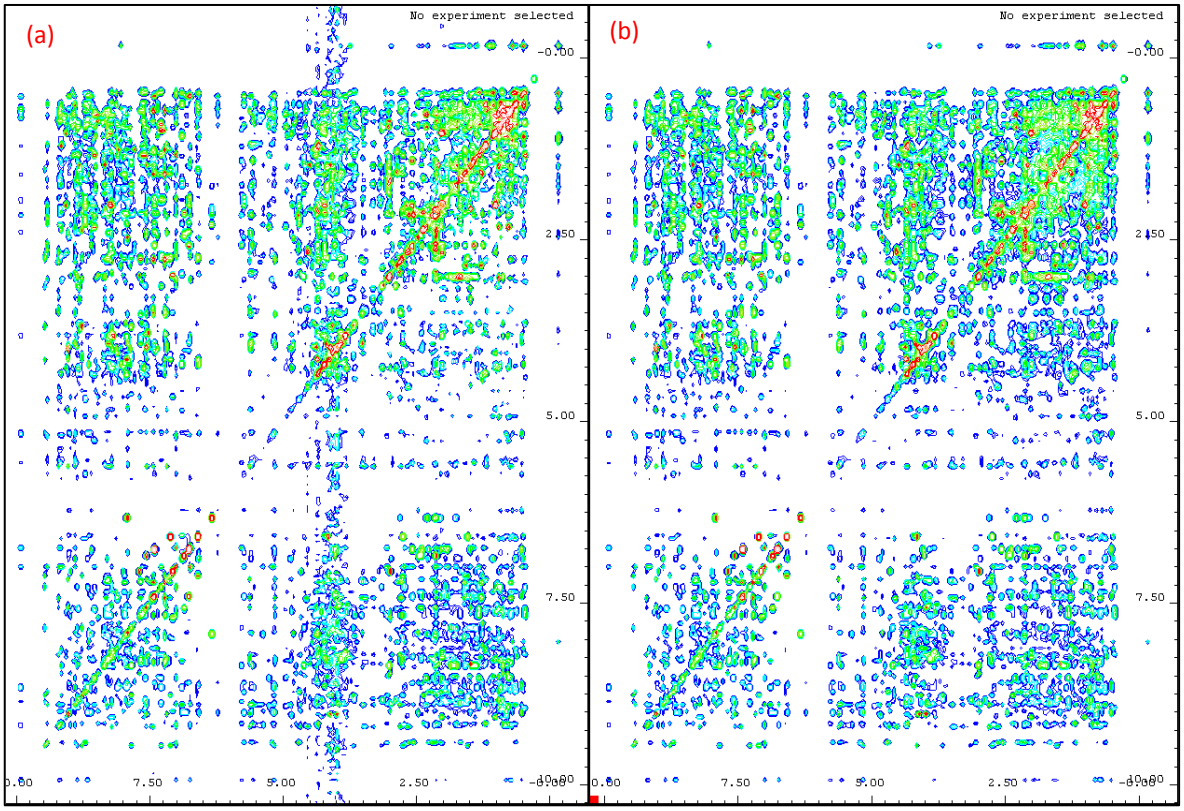


Figure 3.6 SSA of a back-calculated spectrum without solvent: a simulated two-dimensional spectrum (par. 2.1.1.1) of HPr from *Staphylococcus aureus* (H15A) after (a) and before (b) SSA. The experimental solvent signal has not been added to the protein spectrum, thus the SSA automatically removes some resonances of the solute, generating spectral distortions.

The qualitative assessment of the performance of SSA for solvent removal by inspection of the back-calculated spectra has been supported by a quantitative analysis. As starting point, the l2-norm (Euclidian norm) between the original one-dimensional simulated HPr spectrum without water (par. 2.1.1.2) and the same spectrum with the addition of a solvent signal 5000 times stronger than a typical amide protein resonance or 500 times stronger than the strongest protein signals (the superposed signals in methyl region) has been calculated after applying the SSA. The l2-norm of that spectrum has been arbitrarily set to 1 and the values obtained with other solvent/signal amplitude ratios were scaled correspondingly. As demonstrated in Fig. 3.7 the use of SSA even on a very distorted spectrum (where the solvent signal is 150 times stronger than the most intense protein resonance) can definitively improve the spectral analysis. As to be expected, when the relative intensity of the solvent resonance is reduced, the performance of the algorithm as measured by the l2-norm is improved reaching an optimum when the water artifact amplitude is about twice that one of the strongest protein resonance (see the region zoomed out from Fig. 3.7). Further reduction of the solvent signal mixed to the protein spectra leads to a slow increase of the l2-norm. This behavior directly follows from the method itself: SSA removes the component of the FID having the largest

variance. When the intensity of the solvent signal becomes of the order of the intensity of the protein resonances, it is not properly recognized by SSA. When its intensity is smaller than roughly the half of the most intense protein resonance, the algorithm removes just this component of the protein signal. Thus, a meaningful application of SSA is clearly related to the relative amplitude of the water artifact. A warning message about the solvent strength actually appears before starting the SSA calculation. Further improvements for dealing with such cases are described in chapter four.

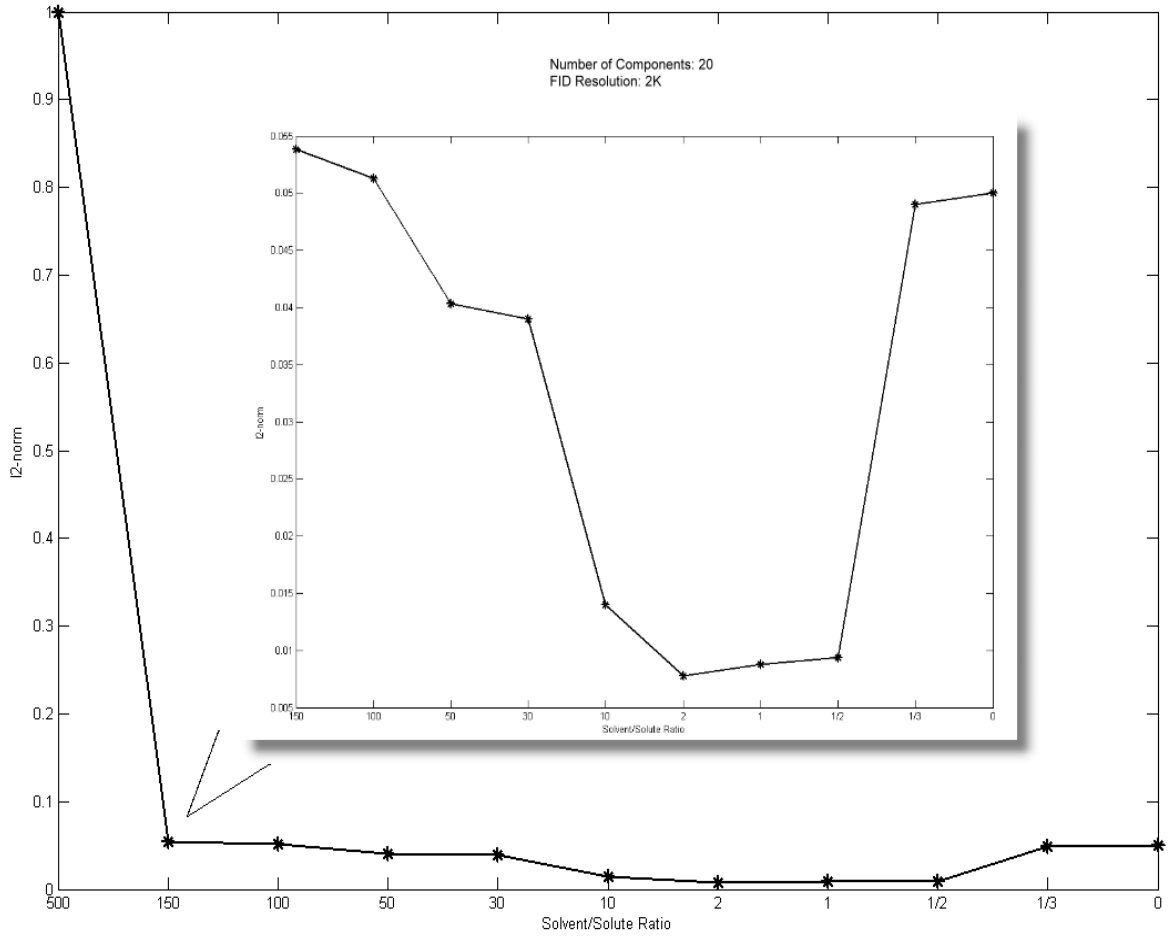
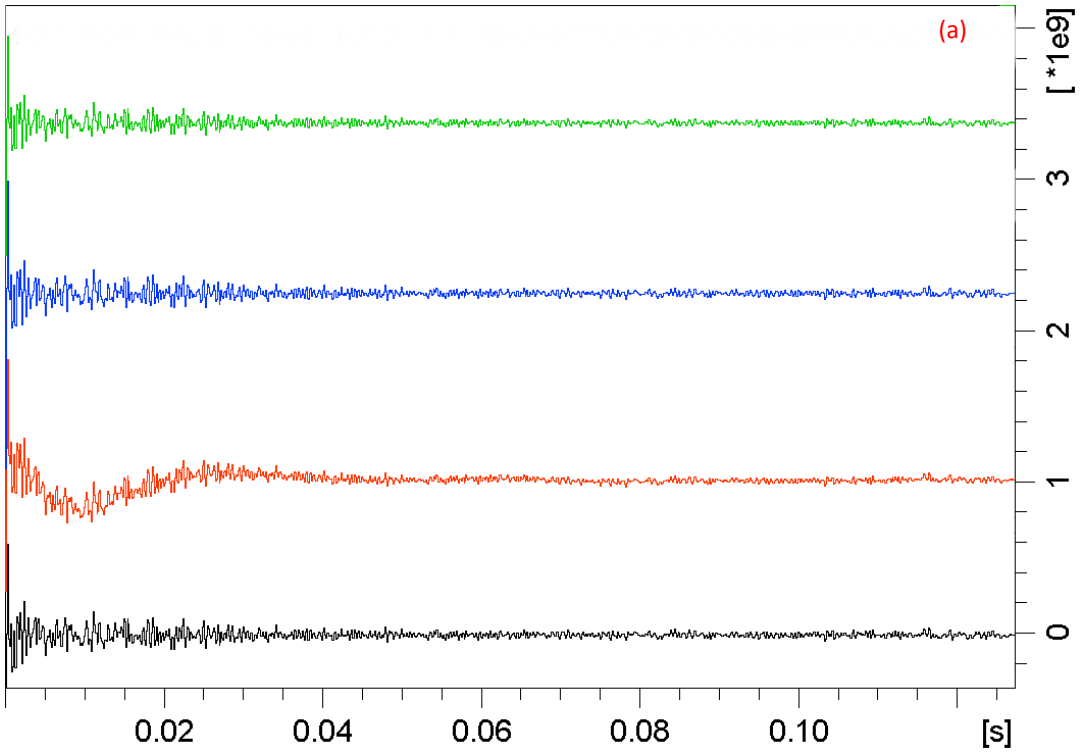


Figure 3.7 Quantitative analysis of the performance of SSA: a one-dimensional spectrum of HPr from *Staphylococcus aureus* was simulated (par. 2.1.1.2), the obtained FID was combined with an FID of an experimental water artifact signal with various relative intensities. As a measure of the performance of SSA for water suppression, the l2-norm was calculated between the original simulated HPr spectrum and an HPr spectrum where a water signal was added. The l2-norm of a spectrum where the solvent signal was 500-times stronger than the most intense protein signal was arbitrarily set to 1. Dependence of the l2-norm on the relative intensity of the solvent signal for a spectrum of HPr with 2 K complex data points is reported [De Sanctis et al, 2011].

3.1.1.4 SSA ON MIXED (TIME-FREQUENCY) DOMAIN

The two-dimensional spectrum is obtained applying a hypercomplex Fourier transformation on the acquired time-domain signals in both directions ($t_1 \rightarrow \omega_1$ and $t_2 \rightarrow \omega_2$). If the data are Fourier transformed only along the columns (ω_2), a mixed time-frequency domain is obtained (t_1, ω_2). The SSA for solvent suppression has been tested in such domain, where the rows are still in the time domain. In particular, the SSA has been applied on the first FID of the back-calculated two-dimensional NOESY spectrum (par. 2.1.1.2) of HPr from *Staphylococcus aureus* (H15A) and on the first FID of the same spectrum Fourier transformed only along ω_2 (mixed domain). The comparison of the resulting time (part *a*) and frequency (part *b*) domain signals is reported in Fig. 3.8. The same analysis has been conducted on the entire two-dimensional spectrum and the comparison of the SSA application in the time (part *b*) and in the mixed (part *a*) domain is shown in Fig. 3.9.



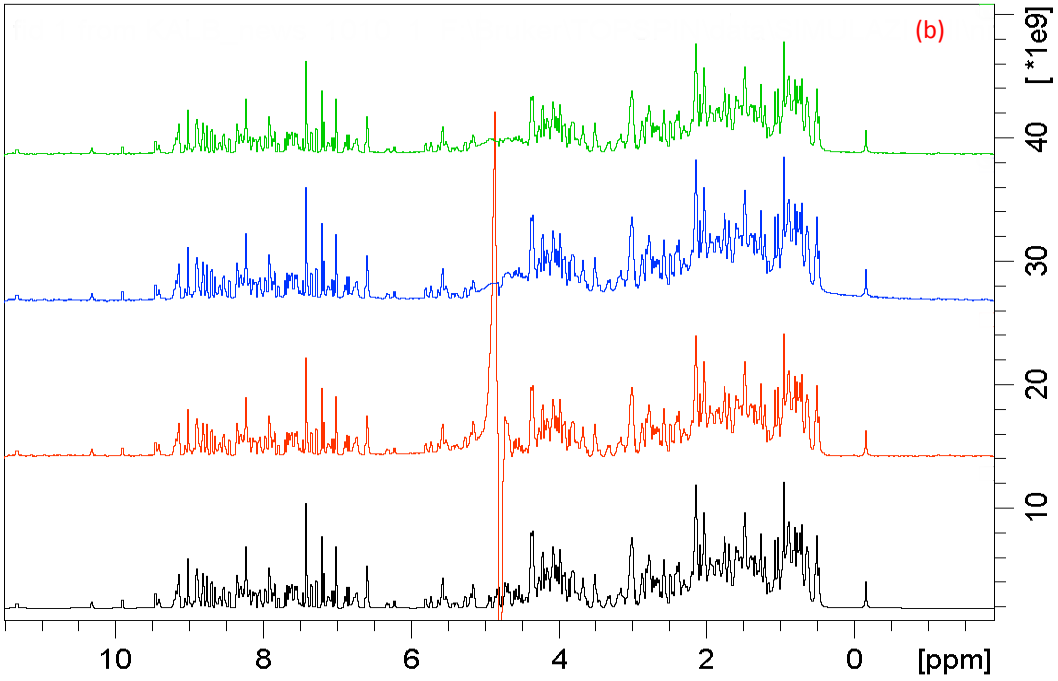


Figure 3.8 Solvent removal by means of the SSA applied in the time and in the mixed domain: (a) first FID of a synthetic two-dimensional NOESY spectrum (par. 2.1.1.2) of HPr from *Staphylococcus aureus* (H15A) (black trace); the same FID with additional experimental water (red trace); the same FID after SSA in the time domain (blue trace); the same FID after SSA in the mixed domain (green trace). (b) Fourier transform of the signals reported in part *a* of this figure.

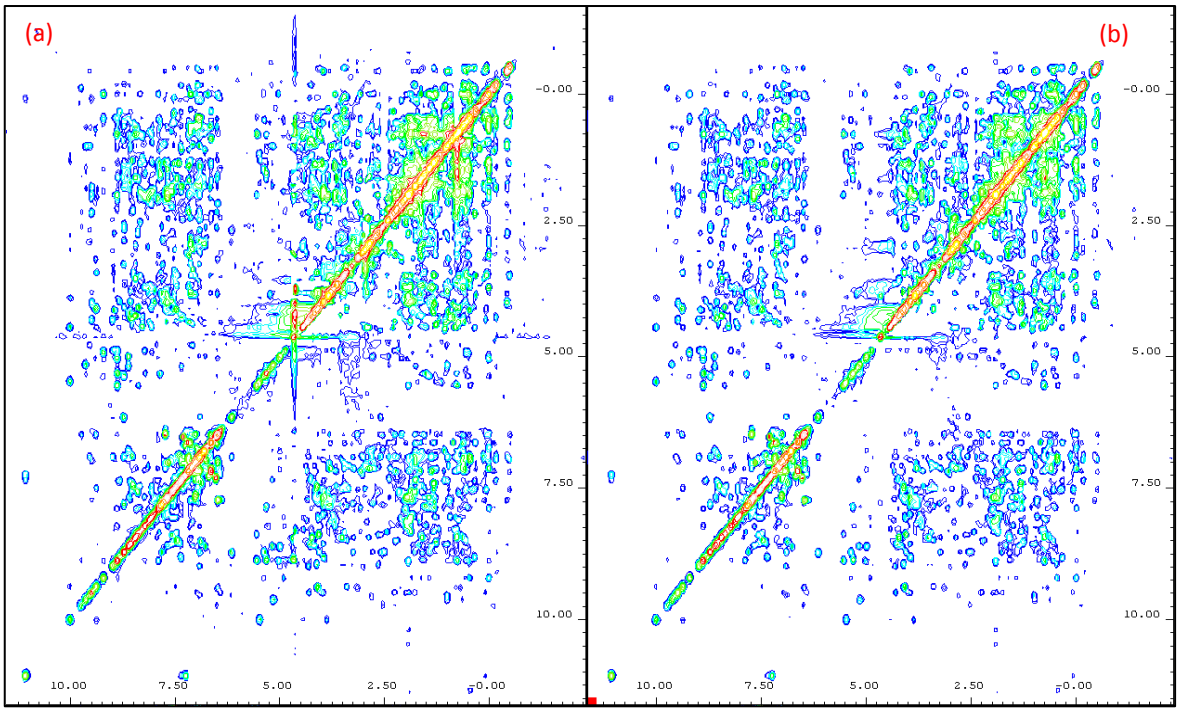


Figure 3.9 Solvent removal by means of the SSA applied in the time and in the mixed domain: synthetic two-dimensional NOESY spectrum (par. 2.1.1.1) of HPr from *Staphylococcus aureus* (H15A); SSA applied in the mixed (a) and in the time (b) domain.

The performance of the solvent suppression by means of the SSA is almost equivalent in both domains. In order to do not increase the computational time, the SSA is directly applied in the time domain.

3.1.1.5 SSA ON SPECTRA WHOSE SOLVENT SIGNAL IS NOT IN THE MIDDLE

In order to test the performance of the solvent suppression by means of the SSA, the back-calculated two-dimensional spectrum (par. 2.1.1.1) has been intentionally modified. In particular, the experimental solvent signal added to the synthetic protein spectrum has been positioned in some random locations in the spectrum. The SSA for solvent removal has demonstrated its capability to deal with such data, independently on the position of the dominant signal in the spectra. This analysis is reported in Fig. 3.10, where the original back-calculated spectrum (*c*) is compared with the same spectrum after the addition of the experimental water (*a*) and with itself after the solvent suppression obtained by means of the SSA (*b*).

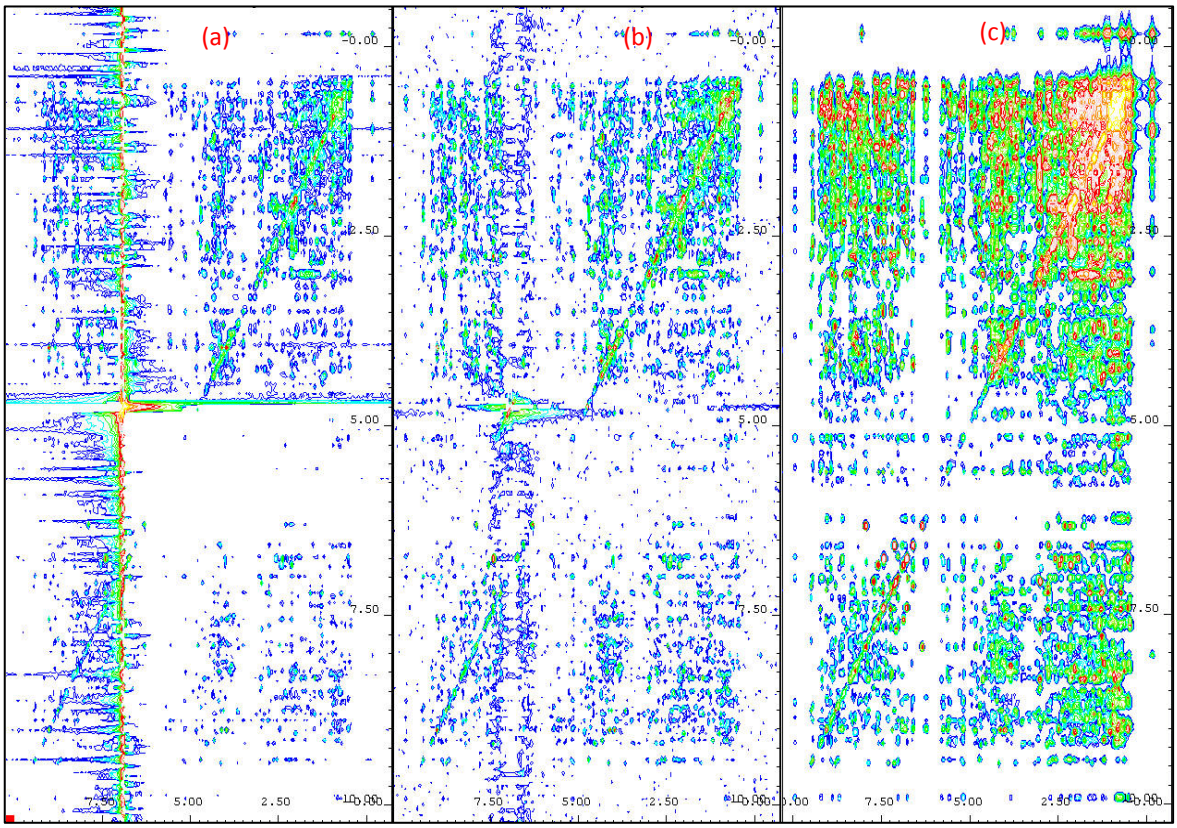


Figure 3.10 Solvent removal by means of the SSA in case of a not centered solvent signal: two-dimensional simulated NOESY spectrum (par. 2.1.1.1) of HPr from *Staphylococcus aureus* (H15A) before (c) and after (a) adding the experimental dislocated solvent signal, compared with the same spectrum after solvent removal by means of SSA (b).

This investigation shows that SSA can be applied also for other purposes, since it is not related to the location of the signal of interest but to the variance of such signal, as discussed in the last chapter.

3.1.2 SSA COMPONENTS EVALUATION

The eigenvalue decomposition of each trajectory matrix has to be performed. In particular, the projection of the columns of any of the Q trajectory matrices $\mathbf{X}^i$ (with i= 1,…,q) along the directions spanned by the eigenvectors of the covariance matrix **R** can be used to investigate the underlying components (see eq. 2.7). The variance of the components is described by the eigenvalues of the covariance matrix **R**, thus the projection of the eigenvector related to the first eigenvalue (the largest one) corresponds to the component with the highest variance (containing the dominant signal of the FID).

The decreasing order of the components (related to a decreasing order of the eigenvalues) allows an automated identification of the signal of interest. In particular, for denoising purposes the last components (with the smallest variance) can be discarded with a consequent nullification of the corresponding projections of the eigenvectors related to the smallest eigenvalues. For water suppression removal the first component is rejected and the new trajectory matrix $\mathbf{X}'$ is obtained after nullifying the projection of the eigenvector corresponding to the first eigenvalue. This procedure is valid only if the solvent artifact in the spectrum of interest is the dominant signal.

The number of extracted components is strictly related to the chosen embedding dimension. The number of time-delayed copies of the FID (namely the amount of the M rows of the trajectory matrix **X**) corresponds to the number of identified components. The embedding dimension has been empirically determined in accordance to the performance of the method (measured using the l2-norm) whose investigation is reported in Fig. 3.11. Such analysis has been conducted to determine the optimal embedding in dependence on the dimensionality of the dataset. In general, as Fig. 3.11 shows for spectra with higher digital resolution more than 20 components have to be used for the analysis. Starting with FIDs with an optimal solvent to protein intensity ratio of 2 (see Fig. 3.7 in par. 3.1.1.3), the dependence of the l2-norm on the number of components used in the SSA analysis was followed for a 2 K and a 16 K FID extracted from the two-dimensional back-calculated spectra of HPr protein (par. 2.1.1.2). In accordance with *Malloni et al, 2010*, optimal values are obtained for the low resolution FID with 20 components. Such a low resolution FID is usually recorded in multi-dimensional spectra. For the high resolution data (as in one-dimensional spectra) the optimal removal of the water signal is reached using 40

components (see the region zoomed out in Fig. 3.11). Therefore, the number of components must be automatically adapted to the size of the handled data.

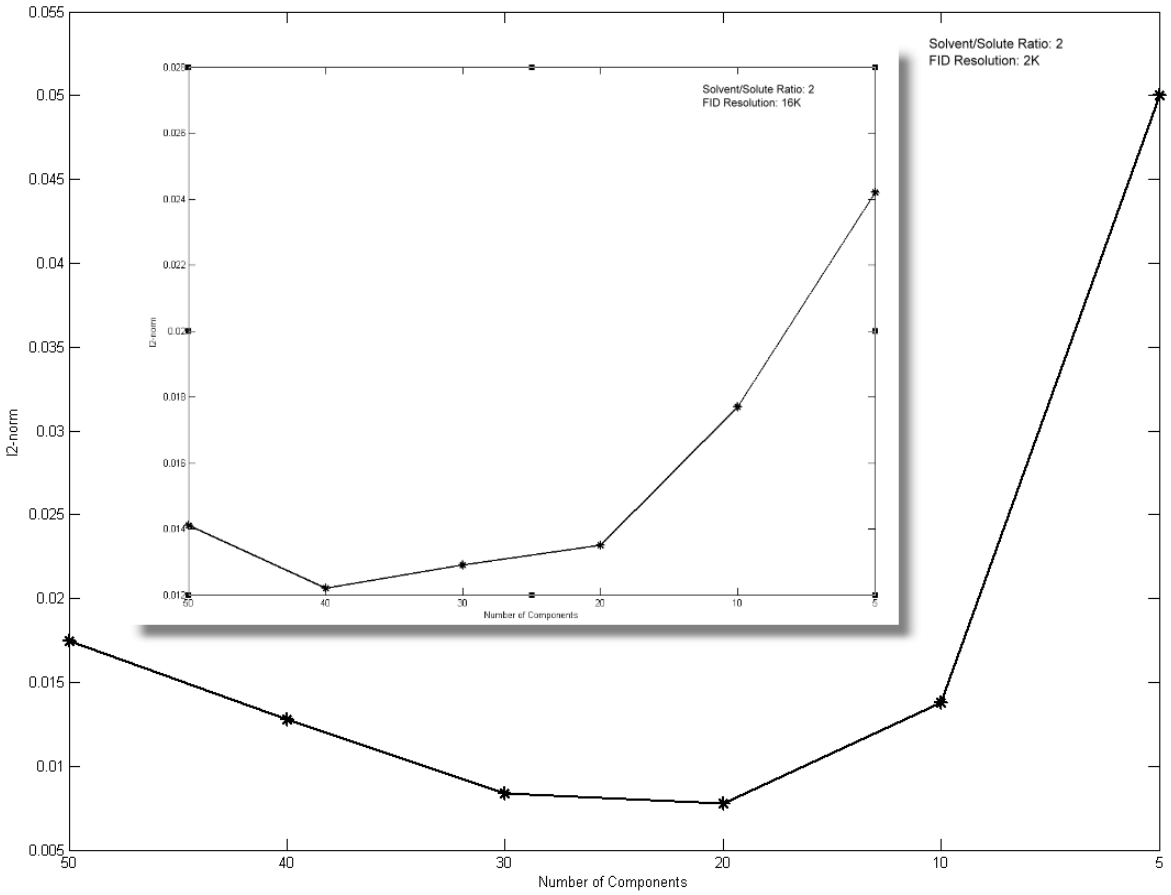


Figure 3.11 Quantitative analysis of the embedding dimension M: dependence of the performance of SSA on the number of extracted components. The l2-norm was calculated for a 2 K and 16 K FID (par. 2.1.1.2) with a ratio of solvent to protein signal intensity of 2 [De Sanctis et al, 2011].

A detailed description of the extracted components has been conducted on the trajectory matrix **X** built from the first FID (first row) of a digitally filtered two-dimensional spectrum (par. 2.1.2.2.1) of HPr from *Staphylococcus aureus* (H15A). Due to the low digital resolution (TD = 1K), 20 components have been extracted. They are reported in Fig. 3.12 demonstrating a clear separation between the water artifact and the protein signals in the time domain. Typically, as described in Fig. 3.13 (showing the corresponding components in the frequency domain), the first of the estimated components represents the solvent almost perfectly, the successive ten components identify the protein signal, while all the remaining ones contain just noise.

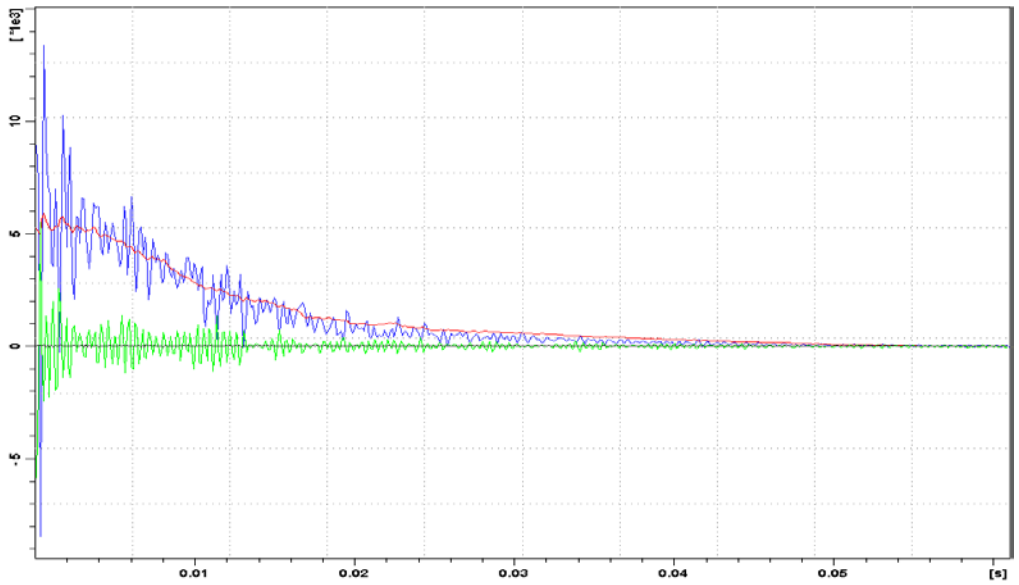


Figure 3.12 Time-domain extracted components by means of SSA: representation of some of the extracted components from the trajectory matrix of the first FID of the digitally filtered two-dimensional NOESY experimental spectrum (par. 2.1.2.2.1) of HPr from *Staphylococcus aureus* (H15A); time domain data, 1x512 complex data points; embedding dimensions of the trajectory matrix 20x492; number of extracted components 20. Superimposition in the time domain of the first component (red trace) related to the solvent signal, the second component (green trace) representing a portion of the protein signal, the last component (black trace) containing only noise and the first original FID before the decomposition (blue trace) [Malloni et al, 2010].

Investigating the components is mandatory to obtain a suitable solvent suppression. In particular, using more components than necessary, leads to undesired effects such as the splitting of the solvent signal in more than one component. Using twenty components for a FID with a resolution of 2K allows the direct rejection of the first component in order to suppress the solvent signal. Increasing this number to forty does not simply involves the nullification of the first two components for water removal purposes, since this signal is effectively shared between such components but the second one typically represents a mixture of solvent and solute signals. Therefore, the rejection of the second component leads to the removal of the strongest resonances of the protein as well, but discarding only the first one does not furnish a sufficient water artifact suppression (that still appears in the spectrum).

If the number of components is smaller than the necessary amount (e.g. twenty instead of forty components in case of a FID with 16 K of digital resolution), the first component inevitably contains a mixture of solute and solvent. Discarding such component leads to the loss of protein resonances.

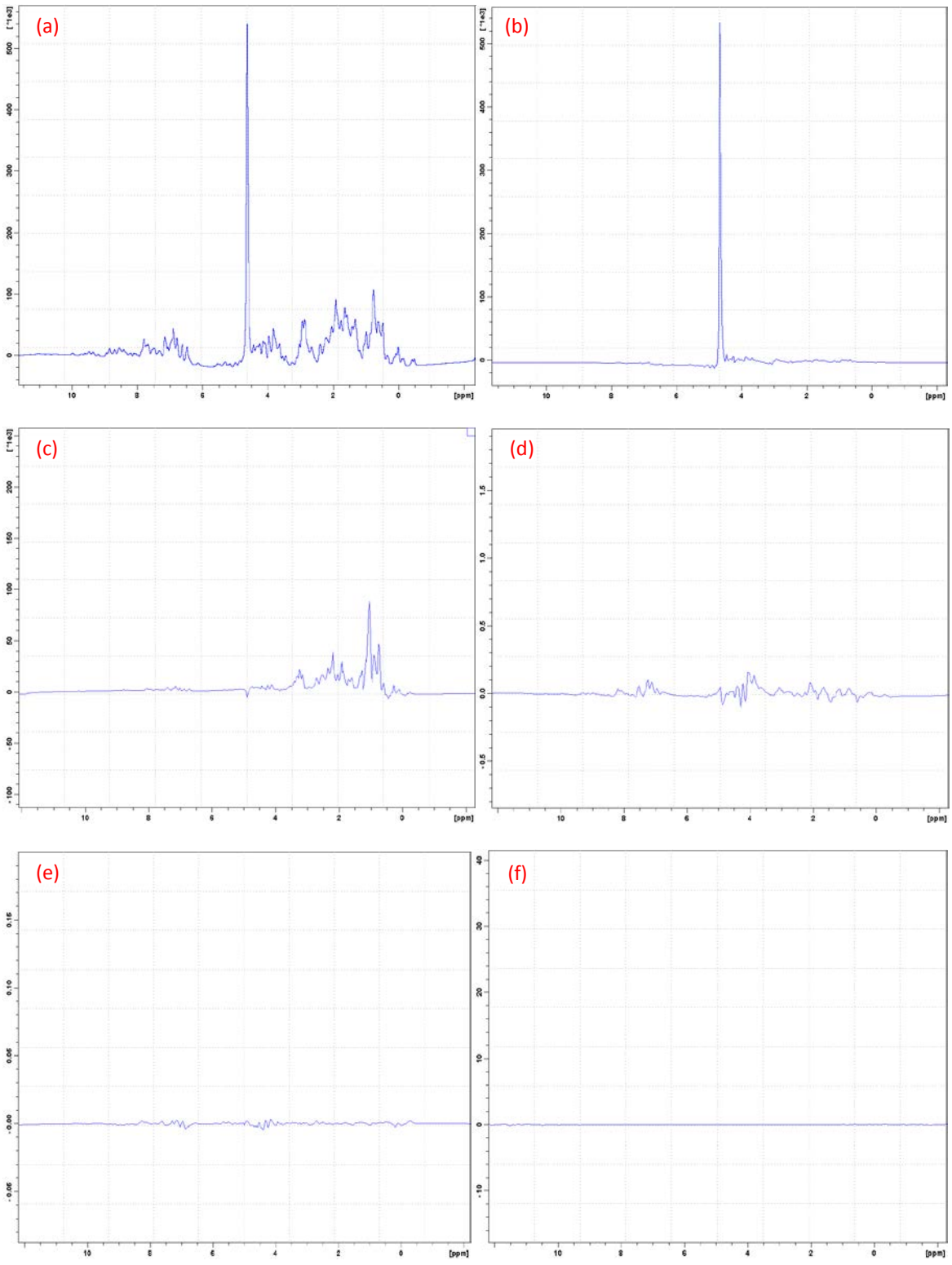


Figure 3.13 Frequency-domain extracted components by means of SSA: representation in the frequency domain of some of the extracted components from the trajectory matrix of the first FID of the digitally filtered two-dimensional NOESY experimental spectrum (par. 2.1.2.2.1) of HPr protein from *Staphylococcus aureus* (H15A); time domain data, 1x512 complex data points; embedding dimensions of the trajectory matrix 20x492; number of extracted components 20; size of the real data after Fourier transformation 20x492. 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. The

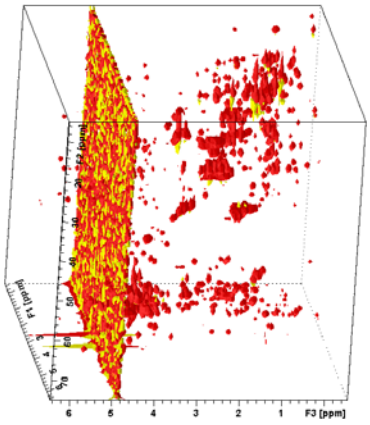
first component (b) contains the solvent signal, from the second to the tenth component (c-e) is instead identified the protein, while after the tenth component (f) there is only noise [Malloni et al, 2010].

3.1.3 SSA SOLVENT SUPPRESSION: TEST CASES

3.1.3.1 SSA OF THREE-DIMENSIONAL DATA

The result of the artifact removal procedure to an oversampled three-dimensional HCCH-TOCSY spectrum (par. 2.1.2.1) is reported in Fig. 3.14. SSA is applied in the direct (t_3)-dimension. After performing the solvent removal over all rows, the data were Fourier transformed. The water resonance and its tails were almost completely removed. The recovery of the peaks lying under the water is demonstrated in Fig. 3.15, showing the projection on the F_1-F_3 plane of the three-dimensional HCCH-TOCSY spectrum before (red colored) and after (green colored) the solvent removal.

(a)



(b)

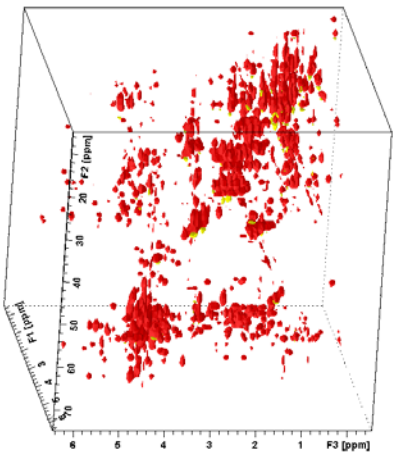


Figure 3.14 Solvent removal by means of SSA applied on a three-dimensional NMR spectrum: sub-cube of a three-dimensional $^{1}H^{13}C$ HCCH-TOCSY spectrum (par. 2.1.2.1) of the thioredoxine protein (Trx) from *Plasmodium falciparum* prior (a) and after (b) artifact removal. Size of the subcube 2048x96x128 real data points [Malloni et al, 2010].

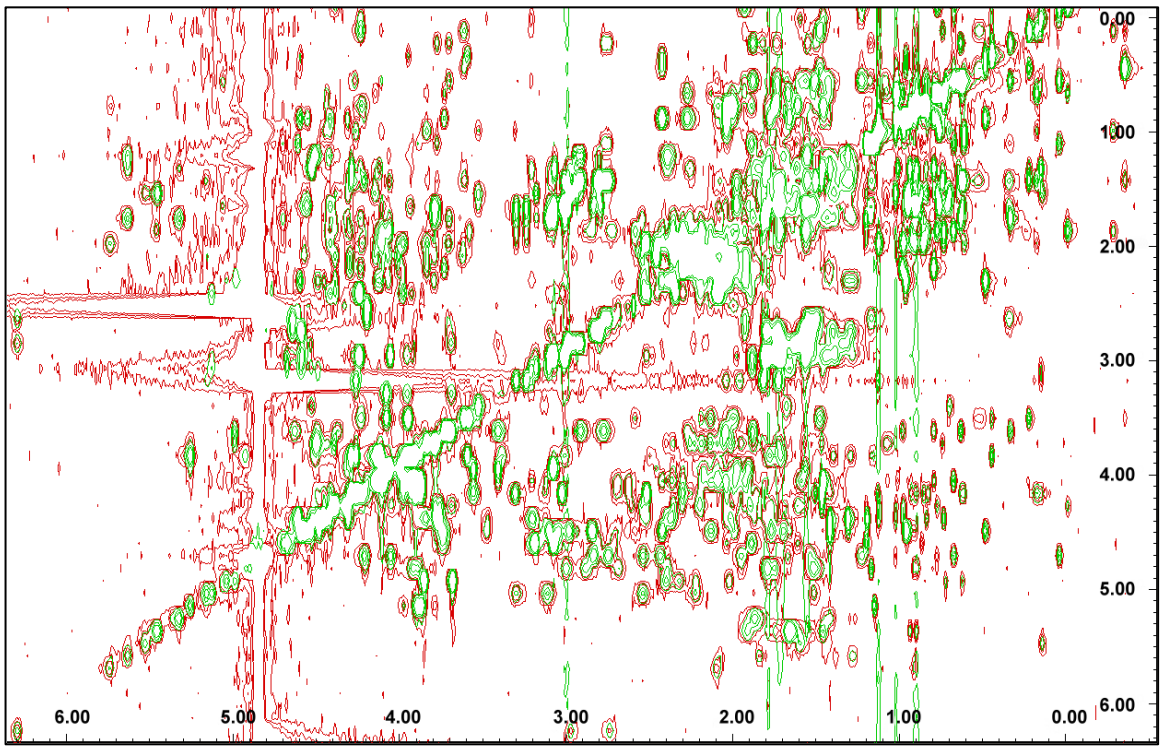


Figure 3.15 Solvent removal by means of SSA applied on a three-dimensional NMR spectrum: projection showing a portion of the F_1-F_3 plane of the three-dimensional $^{1}H^{13}C$ HCCH-TOCSY spectrum (par. 2.1.2.1) before (red) and after (green) the water artifact removal by means of SSA [Malloni et al, 2010].

3.1.3.2 SSA OF TWO-DIMENSIONAL DATA

The synthetic data have the advantage that the pure, artifact free spectrum is available and can be used as the standard for comparing the obtained results. In the case of experimental spectra, a reference spectrum does 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.

In Fig. 3.16 is demonstrated the performance of the method applied on the two-dimensional back-calculated NOESY spectrum (par. 2.1.1.1). The resulting spectrum (after water suppression) looked almost as the unperturbed original one (without additional experimental solvent signal). It is evident that water is strongly suppressed, whereas hidden protein resonances are recovered. Several processing

methods for solvent suppression tend to change the peak intensities. This cannot be accepted for a quantitative analysis of the data and for structural restrains determination. The residual is calculated as the difference between the processed spectrum and the original artifact-free spectrum and it is almost zero (within the limits of pure noise) for the protein cross peaks as shown in Fig. 3.16.

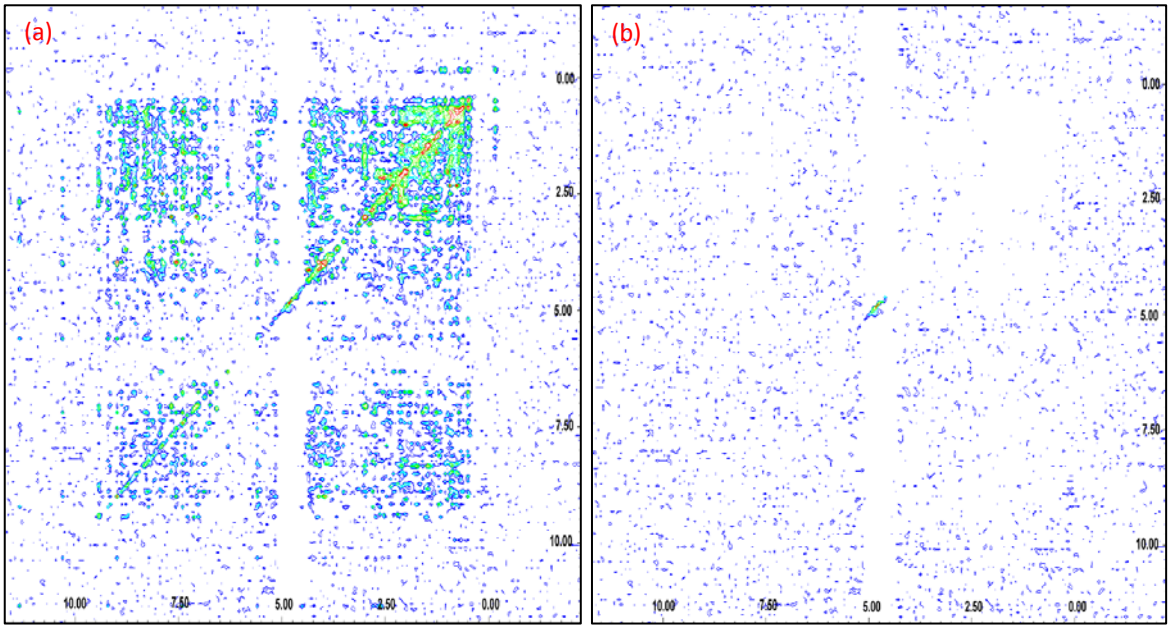


Figure 3.16 Artifact removal by means of SSA on a synthetic two-dimensional spectrum: synthetic NOESY spectrum (par. 2.1.1.1) of HPr from *Staphylococcus aureus* after solvent removal (a) and the result of the difference between the original simulated spectrum and the artifact free spectrum obtained by means of SSA (b). It is evident that the noise has not been removed as well as the small central portion of the water signal [Malloni et al, 2010].

The SSA application on an experimental two-dimensional TOCSY spectrum (par. 2.1.2.2.2) of the HPr protein from *Staphylococcus aureus* is reported in Fig. 3.17. The artifact is largely suppressed and hidden protein resonances lying underneath the water are recovered. The boxed part of the spectrum close to the water resonance is zoomed out in Fig. 3.18. It shows that a threonine $H_\alpha - H_\beta$ cross peak (at 5.430 ppm, 5.150 ppm) superposed by the water resonance has been recovered after the application of SSA.

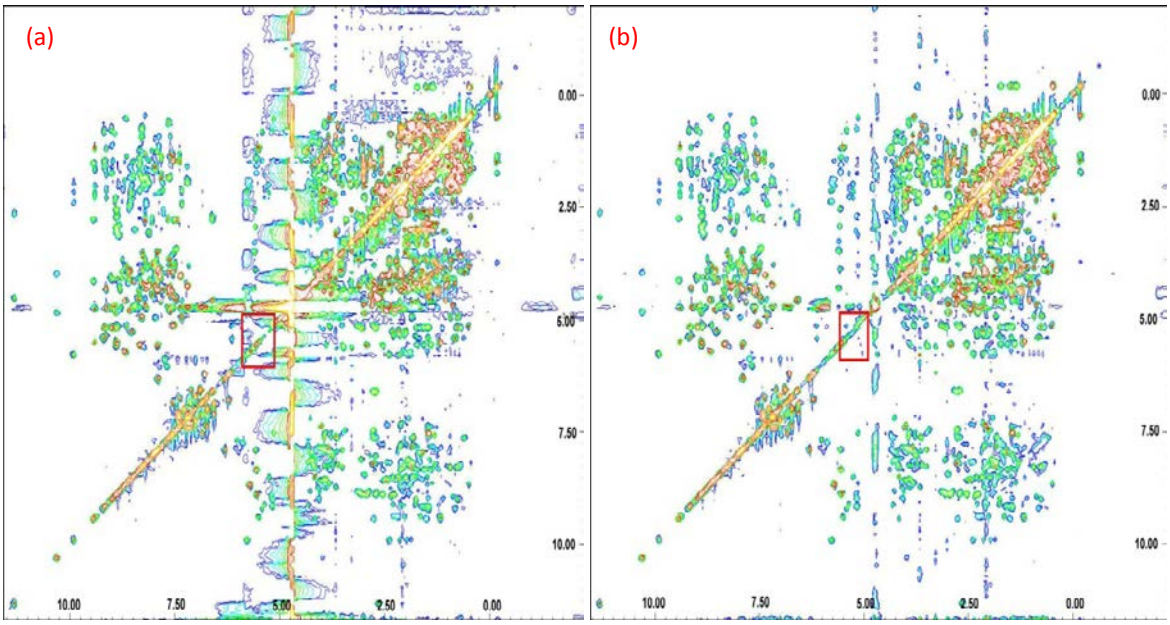


Figure 3.17 Solvent removal by means of SSA on an experimental two-dimensional spectrum: TOCSY spectrum (par. 2.1.2.2.2) of the histidine-containing phosphocarrier protein (HPr) from *Staphylococcus aureus*. (a) Original spectrum and (b) the spectrum after SSA. The regions in the red boxes are zoomed out in Fig. 3.18 [Malloni et al, 2010].

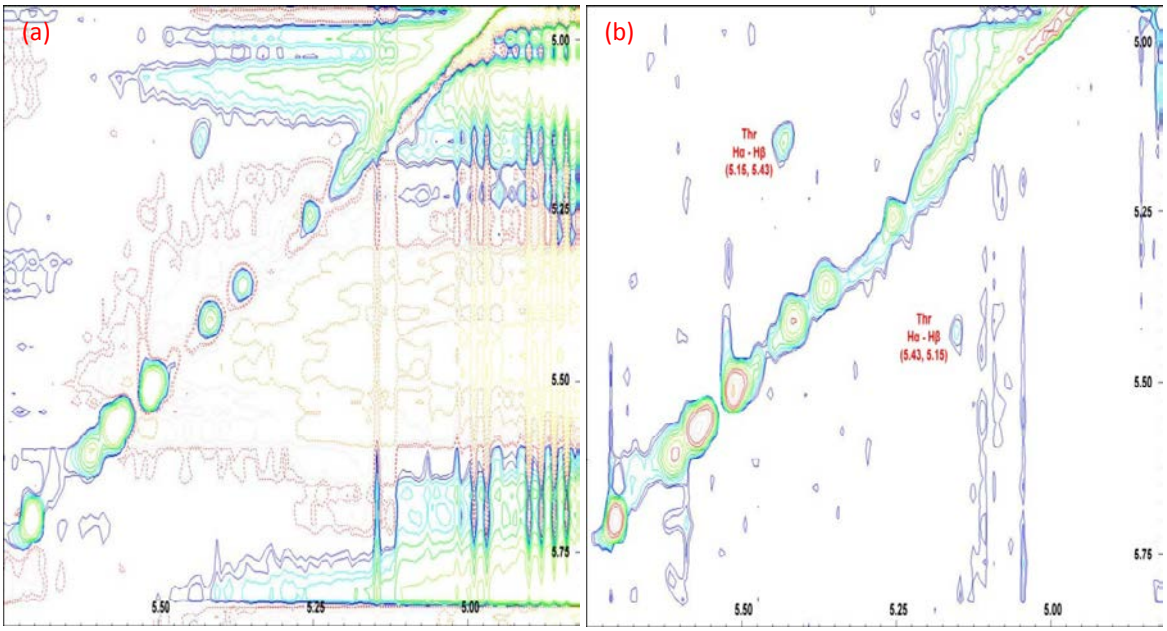


Figure 3.18 Enlargement of the red box regions depicted in Figure 3.17: TOCSY spectrum (par. 2.1.2.2.2) of the histidine-containing phosphocarrier protein (HPr) from *Staphylococcus aureus*. (a) Original spectrum and (b) the spectrum after SSA. Cross-peak resonance recovering at ω_2 = 5.430 ppm and ω_1 = 5.150 ppm revealed by symmetry criteria [Malloni et al, 2010].

3.1.3.3 SSA OF ONE-DIMENSIONAL DATA

The general method was initially developed for multi-dimensional NMR data of proteins [Malloni et al, 2010] but cannot be applied as such to one-dimensional spectra. In particular, the typical higher digital resolution of one-dimensional NMR spectra leads to the necessity of extracting much more components from the time domain signal in order to achieve an accurate separation of the solvent from the rest of the spectrum. Moreover, it is necessary to take into account that one-dimensional spectra of biomolecules are usually very crowded and may contain resonances with widely differing line widths that is determinant for a reliable baseline correction.

The synthetic one-dimensional spectrum (par. 2.1.1.2) allows a direct investigation of the performance of SSA since the original protein spectrum before adding the solvent signal can be compared with the result after application of SSA. Especially, the back-calculation of the one-dimensional time-domain signal of HPr protein allows the identification of the peaks of interest that are located in the water artifact region before adding the solvent signal. Since the back-calculated data are generated without any digital filtering, the algorithm automatically recognizes that no group delay data points have to be taken into account. After the addition of the partly saturated water signal to the time domain signal of the protein, the SSA removal procedure is applied. As default for 1D-spectra, 40 components would be extracted by SSA (par. 3.1.2). Since the simulated 1D spectrum of HPr has not a such high digital resolution (it is 2 K), 20 components showed to be sufficient. The central trace of Fig. 3.19 shows the simulated HPr spectrum after exponential filtering and Fourier transformation of the time domain data. The signal in the bottom was obtained after time domain filtering and Fourier transformation of the same data set with the addition of a very strong solvent artifact time domain signal (500 times stronger than the typical protein signals). The water signal obscures the protein resonances lying in the center of the spectrum and introduces strong baseline distortions with anti-phase contributions and also additional truncation artifacts especially visible in the high field and the low field regions of the spectrum. After application of the SSA (Fig. 3.19 top trace), the water signal is almost reduced to zero, the baseline is almost perfect, and most of the truncation artifacts have been removed. The intensities of the protein signals are not influenced by the procedure, a property very important for quantitative evaluations of NMR spectra.

A zoom of the central area of the spectrum is shown in Fig. 3.20. The HPr resonances in the range between 4.4 ppm and 5.5 ppm are severely compromised by the residual water signal and can hardly be evaluated. The situation changes when SSA is applied (Fig. 3.20 top trace): apart from a small region between 4.7 ppm and 4.8 ppm all HPr resonances are recovered with accurate intensities.

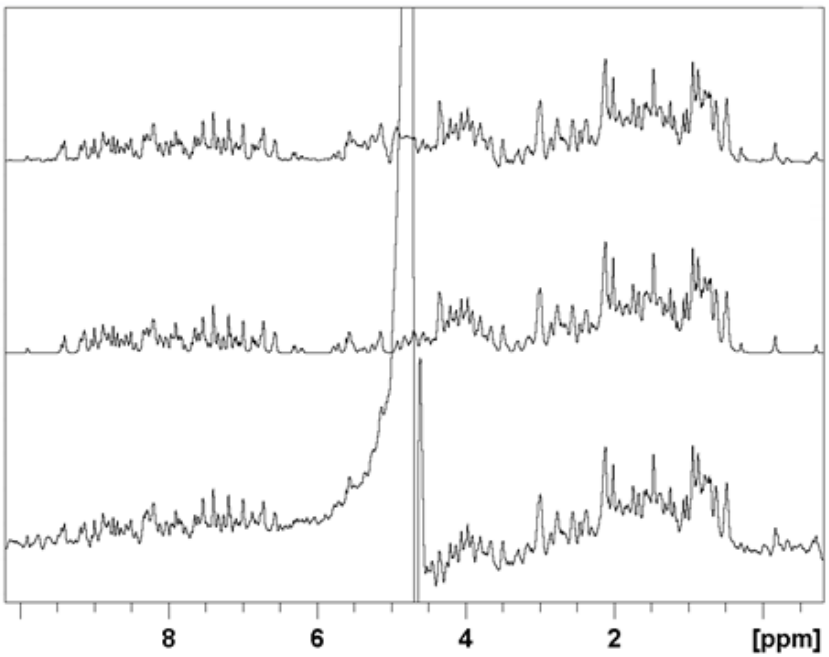


Figure 3.19 Solvent removal by means of SSA applied to a back-calculated one-dimensional spectrum: synthetic one-dimensional spectrum (par. 2.1.1.2) of HPr (H15A) from *Staphylococcus aureus*. An experimental one-dimensional solvent with solvent pre-saturation has been added to the simulated protein time domain data. The water artifact is approximately 500 times stronger than a typical protein resonance. The simulated protein spectrum (middle trace); the spectrum containing the water artifact (bottom signal) and the spectrum after application of SSA (top trace) [De Sanctis et al, 2011].

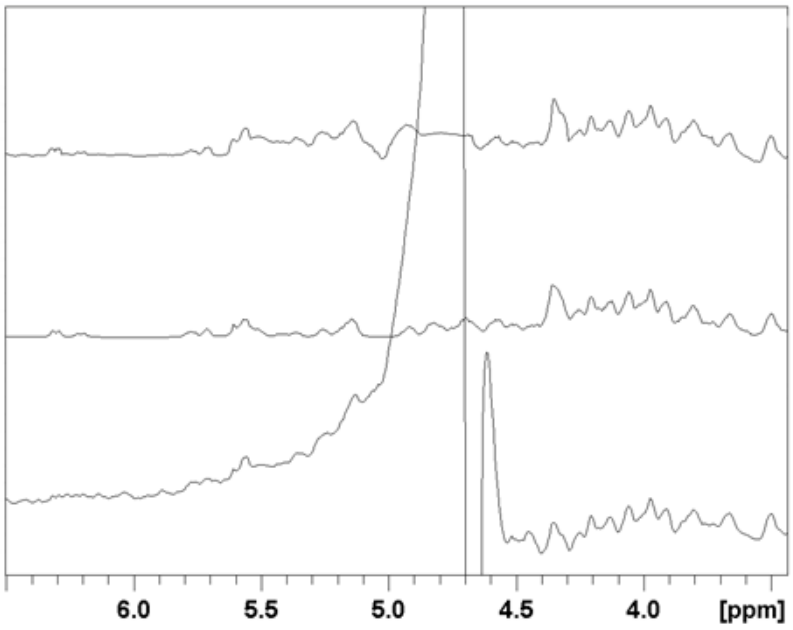


Figure 3.20 Zoom of the spectra shown in Figure 3.19: recovering of resonances close to the water line by means of SSA. Back-calculated Hpr spectrum with additional experimental water (bottom trace); the original spectrum without water (middle trace) and the spectrum after the application of SSA (top trace) [De Sanctis et al, 2011].

Solvent suppression by means of SSA has also been tested on one-dimensional NMR spectra obtained from biofluids (blood and urine) and cells yielding satisfactory results (par. 2.1.2.3.2, 2.1.2.3.3 and 2.1.2.4.1).

The result of the SSA applied on the one-dimensional NMR spectrum of blood plasma of a fasting patient is reported in Fig. 3.21, the SSA on one-dimensional human urine is described in Fig. 3.22, while the one-dimensional cell spectrum is shown in Fig. 3.23.

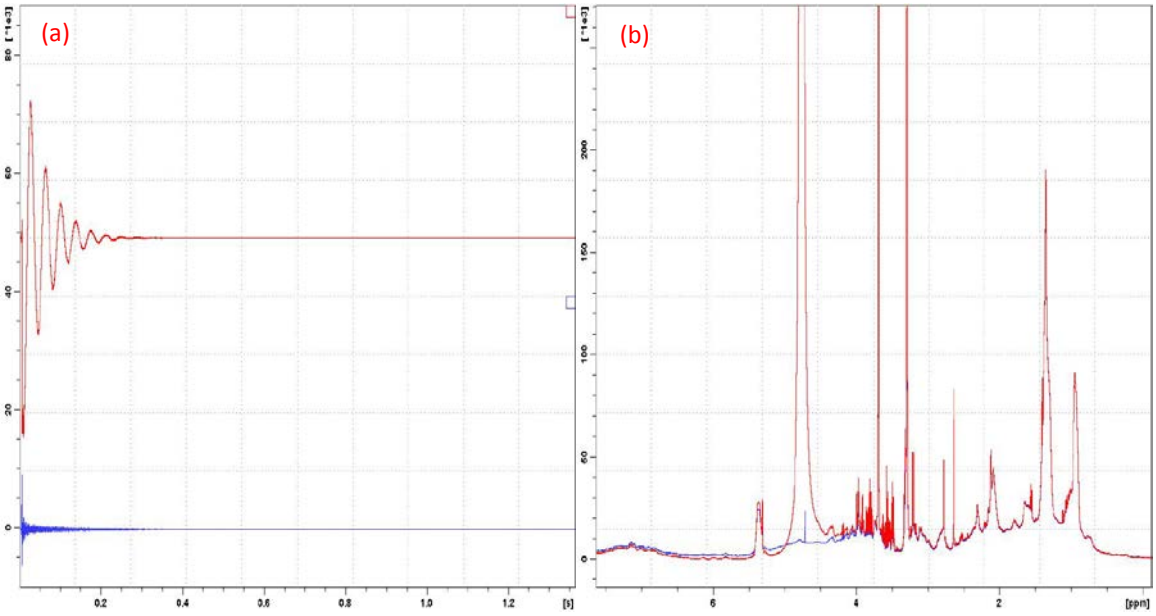


Figure 3.21 Solvent removal by means of SSA applied on the one-dimensional spectrum of blood plasma: the one-dimensional signal in the time (a) and in the frequency domain (b) before (red traces) and after (blue traces) SSA (par. 2.1.2.3.2).

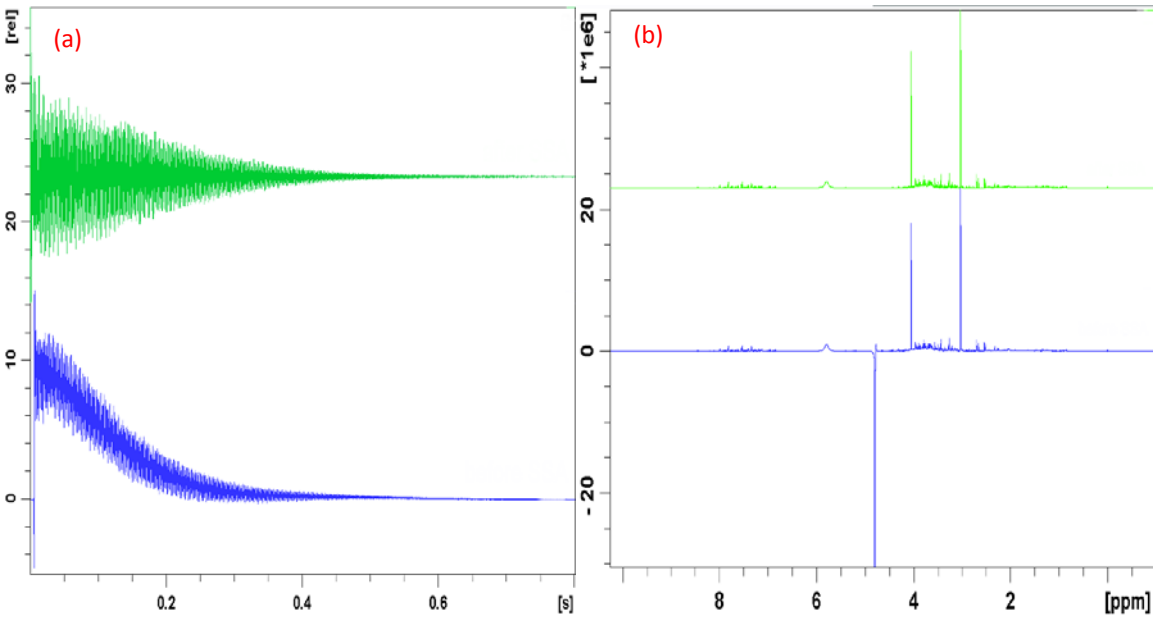


Figure 3.22 Solvent removal by means of SSA applied on the one-dimensional spectrum of human urine: the one-dimensional urine signal with a mixing time of 10 ms in the time (a) and in the frequency domain (b) before (blue traces) and after (green traces) SSA (par. 2.1.2.4.1).

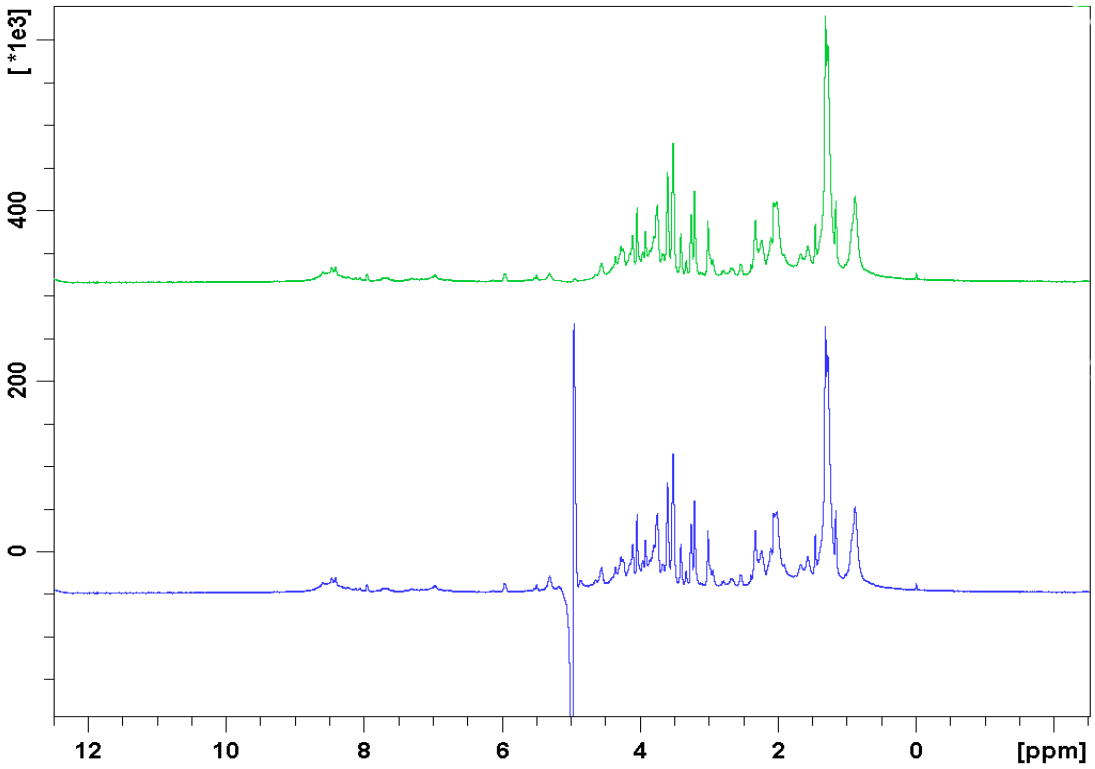


Figure 3.23 Solvent removal by means of SSA applied on the one-dimensional cell spectrum: the one-dimensional spectrum before (blue trace) and after (green trace) SSA (par. 2.1.2.3.2).

3.1.4 POST-PROCESSING: AUTOMATED BASELINE CORRECTION

After suppressing the solvent, an inverse normalization steps is applied at the output of the SSA, in order to avoid scaling problems on the data. In accordance to Fig. 3.1, 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 entire multi-dimensional time domain dataset 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.

After removing the strong solvent signal, the base plane usually needs to be corrected in the frequency domain. One of the most robust methods for baseline correction is the cubic spline [Zolnai et al, 1989] that, however, induces new artifacts in areas where only few baseline points can be defined. In this project the linear spline

interpolation [Saffrich et al, 1992] of the baseline points has been chosen since it is more efficient and simpler.

Typically, the base points where there are no relevant peaks are defined interactively by the user. This is not acceptable for a complete automated procedure, hence the points have to be identified by the program.

A method similar to that one developed by *Guentert and Wuethrich* in *1992* [FLATT algorithm] is used to automatically recognize the baseline regions in the spectrum (par. 1.2.1). The algorithm looks for contiguous pieces of row or column that can be well fitted by a straight line. This is possible only if they correspond to pure baseline regions. Obviously, the baseline correction depends on the ability of the automated recognition of the baseline points that must not contain valid signals.

Starting from any data point *k*, the size W of a window surrounding it, must be determined and it must be obviously larger than the line width of the protein resonance peaks. Therefore, it must automatically adapt this window in dependence on the spectrum of interest. The default value of 75Hz [Guenter and Wuethrich, 1992] is suitable only for homonuclear two-dimensional proton NMR spectra. The spectrum needs to be investigated in order to define the optimal window size. It is 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]. Dealing with two-dimensional data, the maximal values of the line width of the peaks are computed separately for each dimension in the following manner:

$$LW_1 = \max\ (\max\,(lw_r)) \tag{3.6}$$

$$LW_2 = \max\ (\max\,(lw_c)) \tag{3.7}$$

with r =1,……, number of rows and c = 1,….., number of columns.

In a two dimensional case, two line width histograms are generated. They represent the line width distributions within the frequency range (0, LW_1) and (0, LW_2) respectively and they contain the occurrence of the maximal line width values of the peaks for each dimension (i.e. the peaks with the maximum line width row by row or column by column). Only the peaks having intensities significantly larger ($> 3\sigma_N$) than the noise level σ_N are evaluated. The most frequently occurring line width is detected and it is used to establish the window size W along the considered dimension. Actually, it corresponds to the double of the most occurring line width values (i.e. the double of the maxima in the histograms). The Fig. 3.24 (part *a*) shows

the occurrence of the maximal line width values along the columns of the two-dimensional experimental spectrum of HPr protein (par. 2.1.2.2.1) from *Staphylococcus aureus* (H15A). In this case the line width value of 34 Hz has been the maximal value encountered in 37 rows, thus a window size of 72 Hz is generated.

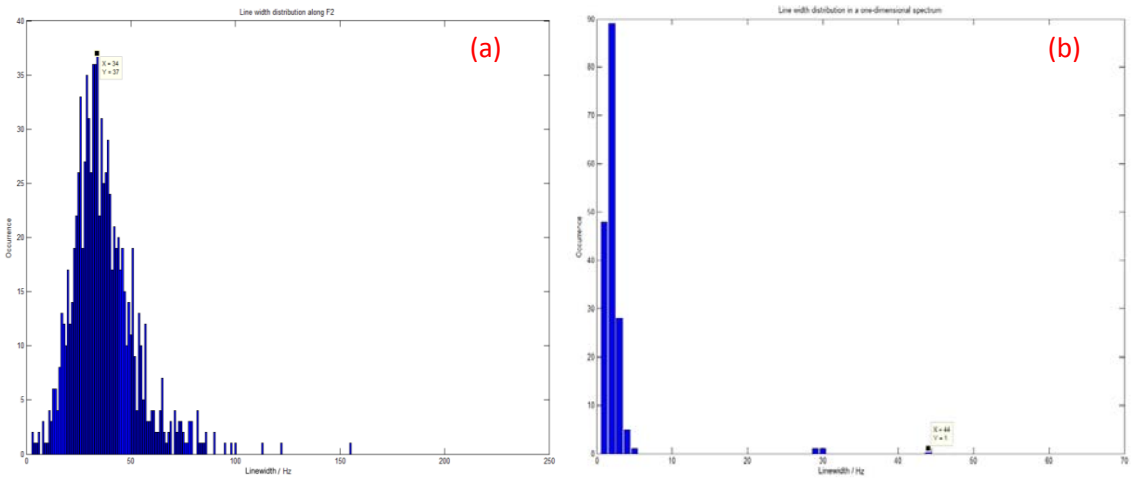


Figure 3.24 Line width distributions: (a) histogram of the maximal line width values along the columns of the two-dimensional experimental spectrum of HPr protein (par. 2.1.2.2.1) from *Staphylococcus aureus* (H15A); (b) histogram of the line width values of the one-dimensional spectrum of human urine (par. 2.1.2.4.1). The window size W is set to the double of the most occurring value (W = 72 Hz) in case of a two-dimensional dataset and to the double of the maximal value (W = 88 Hz) in the one-dimensional case.

This window size is especially important in the complicated spectra of biological samples with large variations in line widths. In one-dimensional spectra of such cases the previous definition of W does not work satisfactory since resonance lines with large line widths are not recognized correctly. It turned out that a window size W set to the double of the maximal occurring value (not the most occurring one) yielded optimal results in one-dimensional spectra, as described in the part *b* of Fig. 3.24. A unique histogram is generated in case of one-dimensional data and it contains the occurrence of all the line width values of the spectrum. In the one-dimensional human urine spectrum (par. 2.1.2.4.1) reported in Fig 3.24 (part *b*) the line width value of 44 Hz has been encountered only once and the window size is set to 88 Hz in accordance to this maximal computed line width.

The window slides over the data row-wise and column-wise. Within each sliding window (centered at the k^{th} point), the data points are fitted by a straight line. The mean square deviation χ_k^2 from the best fitting straight line is determined. The regions in the spectrum where it is lower than a certain threshold are determined as follows:

$$\chi_k^2 \leq \tau \chi_{min}^2 \quad (3.8)$$

The data points verifying eq. 3.8 are identified as pure baseline regions (where generally τ =10, as described by *Güntert* and *Wüthrich* in 1991). An example of the algorithm implementation is reported in Fig. 3.25, where the sliding widow appears with different colors and for each stretch of points included in every window the condition reported in eq. 3.8 is evaluated. If it is verified, the central point of a specific window joins the group of the baseline points.

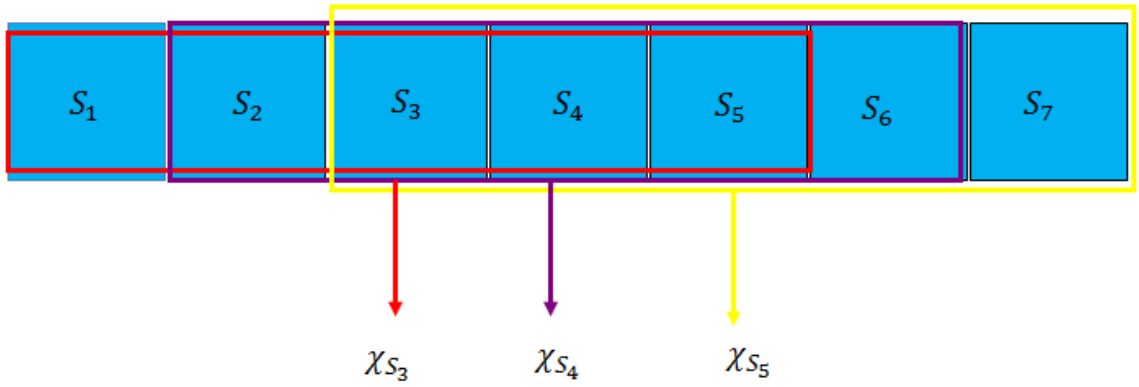


Figure 3.25 Automated baseline point identification: schematic representation of the sliding window (whose size depends on the most occurring line width value along the considered direction) and automated identification of baseline points by means of threshold setting on the mean square deviation of each central peak in the window from the best fitting to a straight line.

If the gap between two consecutive baseline points is larger than the 5% of the complete row or column size, then the threshold is iteratively modified as follows:

$$\tau = 1.5^k \tau \quad (3.9)$$

with k = 1,..,4.

The total set of pure baseline regions is linearly interpolated and subtracted row-wise and column-wise from the original dataset. However, if long stretches of baseline regions are interpolated straight lines of zeros appear in the spectrum. This happens when baseline points are direct neighbors and regions with a noise-less baseline are created by the linear spline. In order to avoid this, not all the consecutive baseline

points are interpolated. In stretches of five consecutive baseline points, only the middle one is used in the interpolation. In addition, the intensity value of such a baseline point is substituted by the mean value between its own intensity and the intensities of the two adjacent points. The described procedure is firstly applied on the rows and then repeated column-wise excluding the already corrected points (along the rows) from the search of baseline points.

The automated baseline point identification is highlighted in Fig. 3.26, where the left side of the one-dimensional HPr spectrum measured with watergate solvent suppression (par. 2.1.2.3.1) is reported before and after determining the baseline regions.

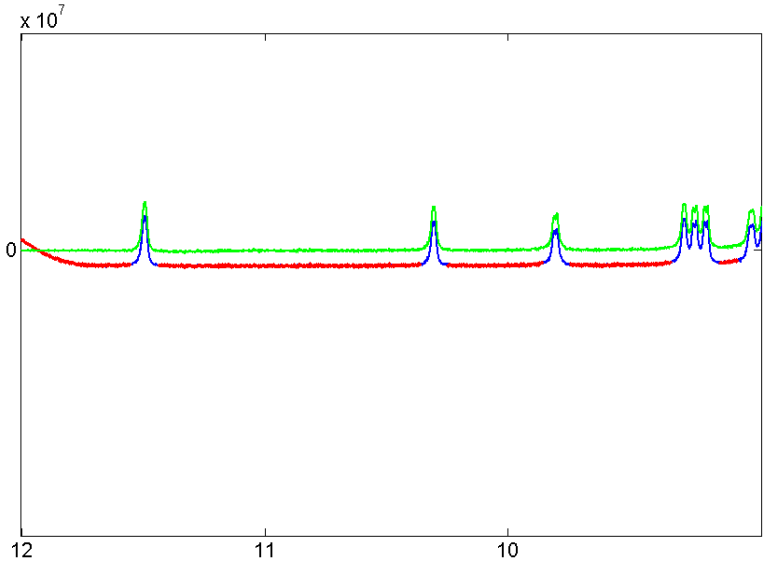


Figure 3.26 Example of automated baseline points identification: left side of the experimental one-dimensional spectrum of HPr from *Staphylococcus aureus* measured with watergate solvent suppression (par. 2.1.2.3.1) before (blue trace) and after (green trace) baseline correction. The automatically identified baseline points are highlighted with red stars.

In Fig. 3.27 is described an example of baseline correction by means of the developed algorithm (ALS, automated linear spline). The back-calculated two-dimensional NOESY HPr spectrum (par. 2.1.1.1) is shown without water and base plane distortions (*a*). The same spectrum with additional experimental solvent signal and severe base plane deviations (*b*) is reported as well and it is compared with the baseline corrected spectrum (*c*). This example demonstrates that the method is able to correct the baseline distortions outside the region of the water artifact but it alone cannot remove the solvent signal. This is in fact the domain of the SSA-module. In accordance to the schema reported in Fig. 3.1 the baseline correction (ALS-module) applied in cascade after the application of the SSA leads to significant improvements. A similar

example is reported in Fig. 3.28 where the automated baseline correction has been applied on an experimental three-dimensional spectrum (par. 2.1.2.1).

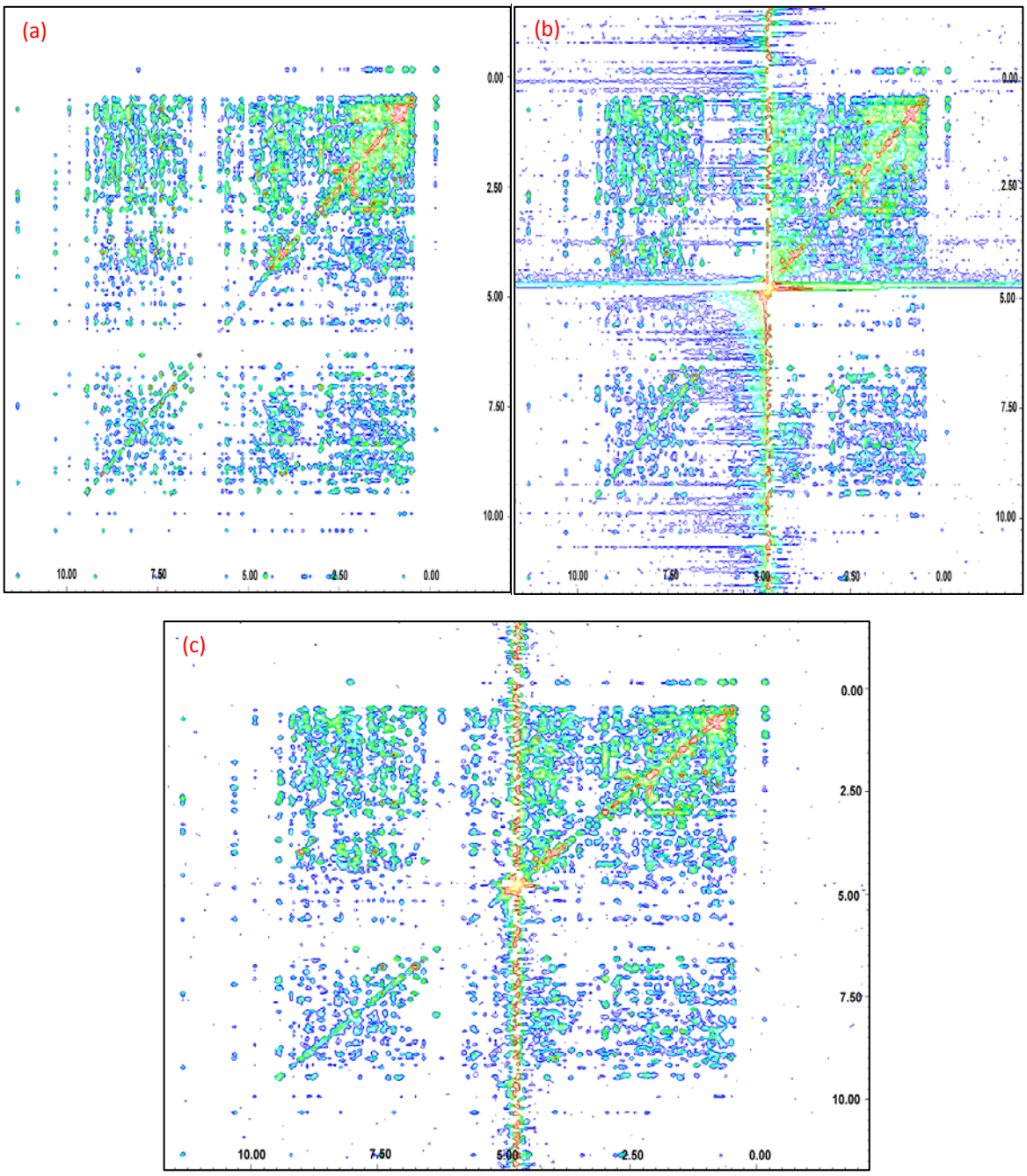


Figure 3.27 Two-dimensional automated baseline correction (ALS): synthetic two-dimensional NOESY spectrum (par. 2.1.1.1) of HPr from *Staphylococcus aureus* (a), the simulated NOESY spectrum with additional experimental water (b) and baseline correction on the same spectrum (c) [Malloni et al, 2010].

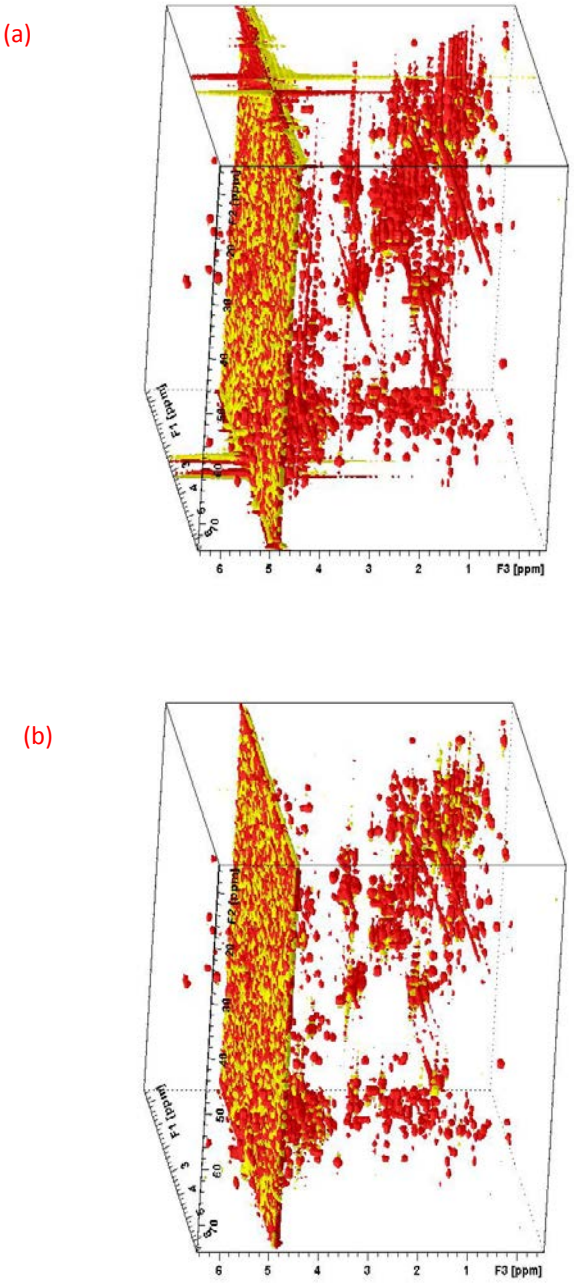


Figure 3.28 Multi-dimensional automated baseline correction (ALS): experimental $^{1}H^{13}C$ HCCH-TOCSY spectrum (par. 2.1.2.1) of the thioredoxine protein (Trx) from *Plasmodium falciparum* prior (a) and after (b) baseline correction. Size of the subcube, 2048 x 96 x 128 real data points.

A typical case where ALS has to be applied after solvent suppression by SSA would be on the urine (par. 2.1.2.4.1) spectrum (see Fig. 3.29, part *a*). Here, SSA as described in this application removes the strongest signal (that is the water signal) from the spectrum but has only a small effect on the baseline (see Fig. 3.29, part *b*). A sinusoidal baseline distortion as it occurs by data clipping was additionally

introduced before in the spectrum. As Fig. 3.29 (part *c*) shows the ALS module perfectly removes the baseline distortion. With the calculation method for the determination of the window size W used in multidimensional NMR a too small size of 6 Hz would result, however the window size calculated from the largest effective line width is 120 Hz and is necessary for a proper baseline correction.

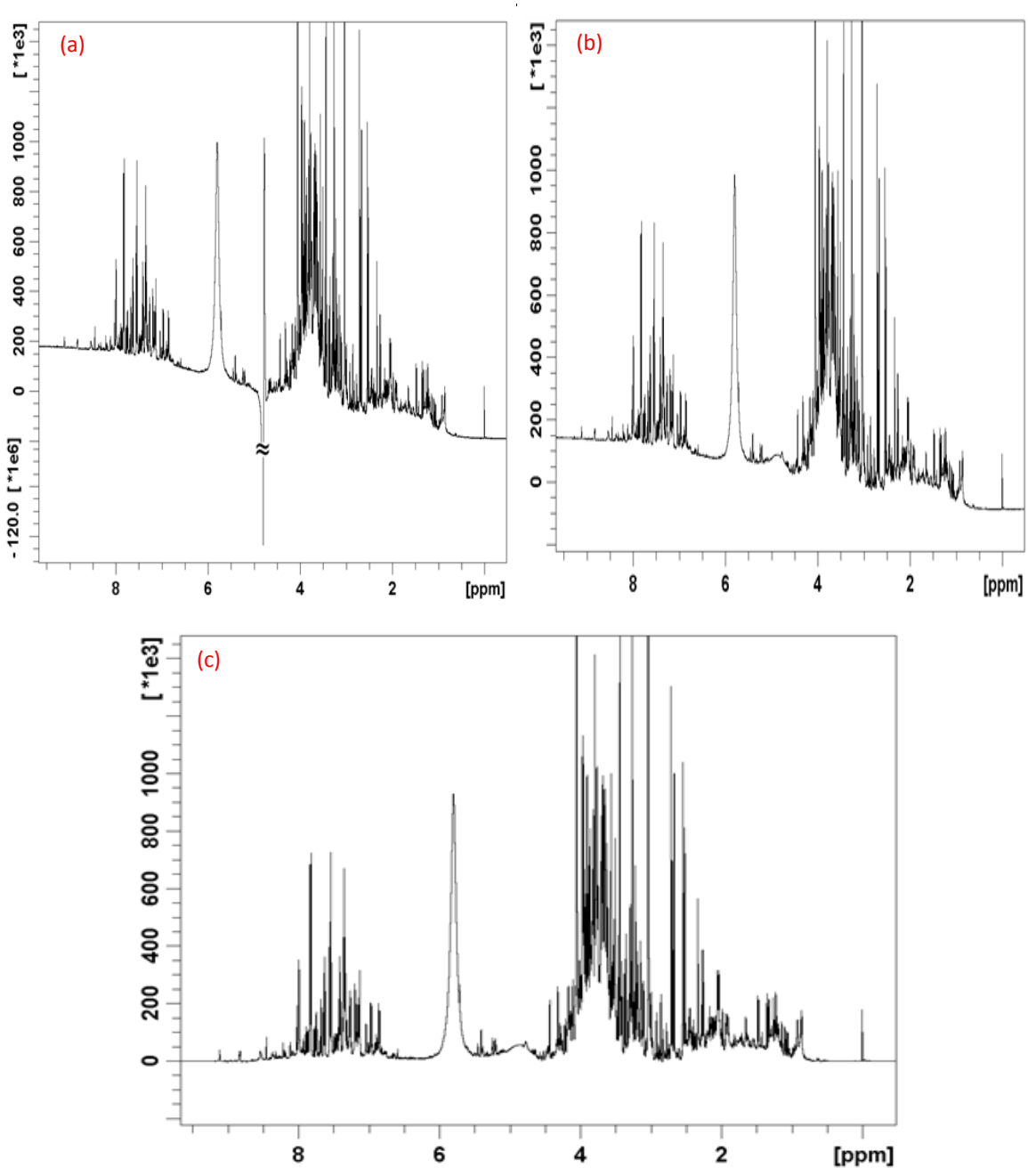


Figure 3.29 One-dimensional automated baseline correction (SSA/ALS): one-dimensional urine spectrum is recorded with a mixing time of 10 ms (par. 2.1.2.4.1). Solvent removal and baseline correction by means of SSA and ALS. Complete experimental spectrum of the urine (a), spectrum after application of SSA (b) and the spectrum after application of ALS (c) [De Sanctis et al, 2011].

The two-dimensional NOESY spectrum of the PSCD4-domain has been assigned for obtaining distance restraints (par. 2.1.2.2.3). The SSA and the ALS modules have been applied in cascade on this spectrum, as shown in Fig. 3.30.

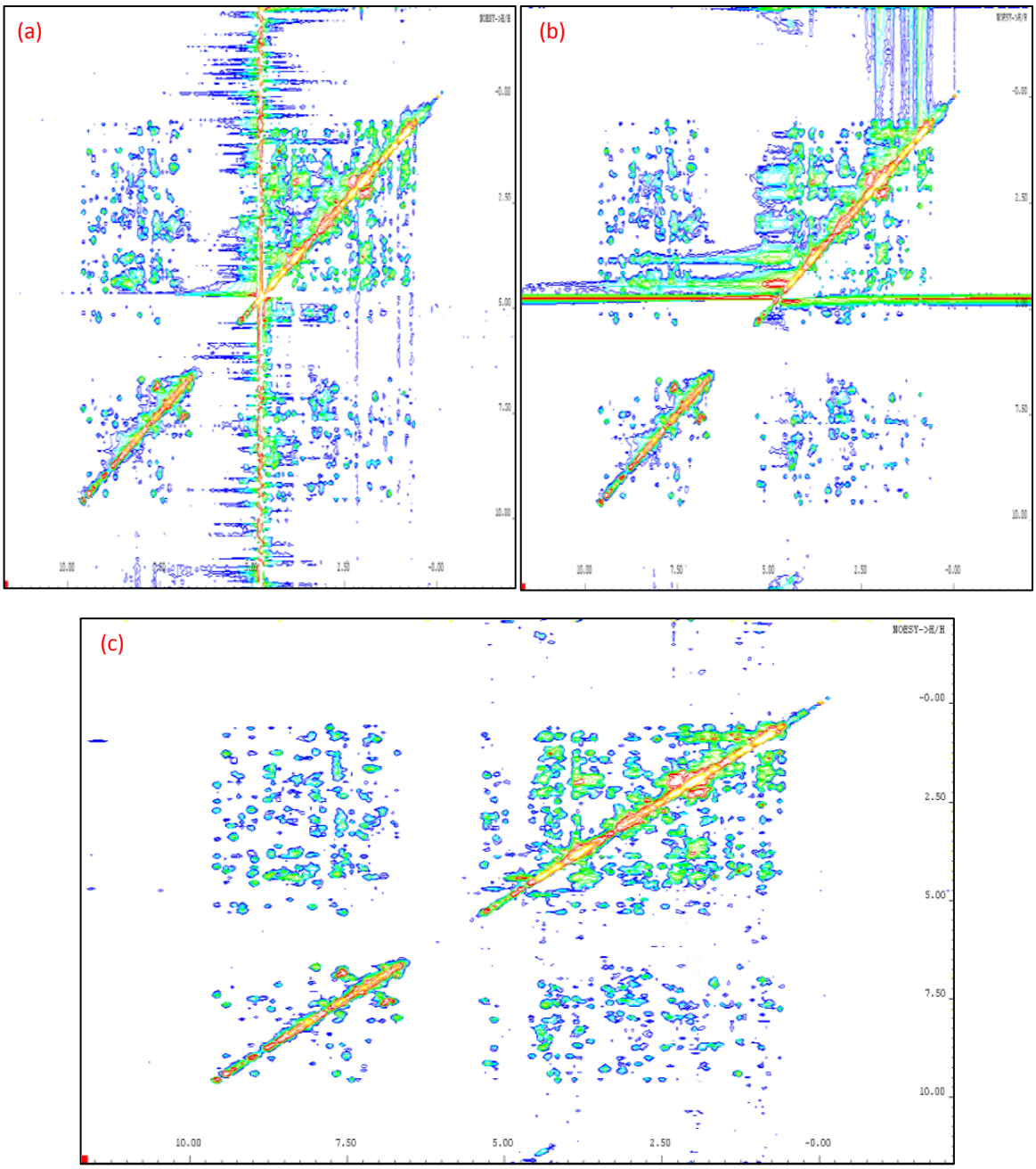


Figure 3.30 Two-dimensional solvent suppression and baseline correction (SSA/ALS): two-dimensional NOESY spectrum of PSCD4-domain (par. 2.1.2.2.3). Solvent removal and baseline correction by means of SSA and ALS. Complete experimental spectrum (a), spectrum after application of SSA (b) and the spectrum after application of ALS (c).

3.1.5 AUREMOL-SSA and AUREMOL-ALS DIALOGS

It is possible to start the developed AUREMOL-SSA routine from the "Calculation" menu as shown in Fig. 3.31. The main interface is launched once that the "Remove Water" submenu has been selected and it is represented in Fig. 3.32. The user must provide the input file path where the ser (multi-dimensional case) or the fid (one-dimensional spectra) files are located. The user may provide the frequency domain data in input to the routine (as the 2rr or 1r files). In this case an automated inverse Fourier transform is performed in order to obtain the time-domain data necessary to apply the SSA. It has been empirically demonstrated that starting directly from the time domain is more efficient since several procedural steps cannot be inverted from the frequency to the time domain (as the filtering) yielding spectral distortions after the SSA.

In addition, the user may need to store the time domain signal after applying the solvent suppression by means of the SSA. The last option in the dialog must be checked for such purposes, since otherwise the SSA provides directly frequency-domain results.

After launching the SSA routine, a warning message appears in order to continue the computation only with the user agreement. This dialog is reported in Fig. 3.33 and if the user indicates that the solvent signal is not the dominant one in the spectrum, the calculation is promptly interrupted.

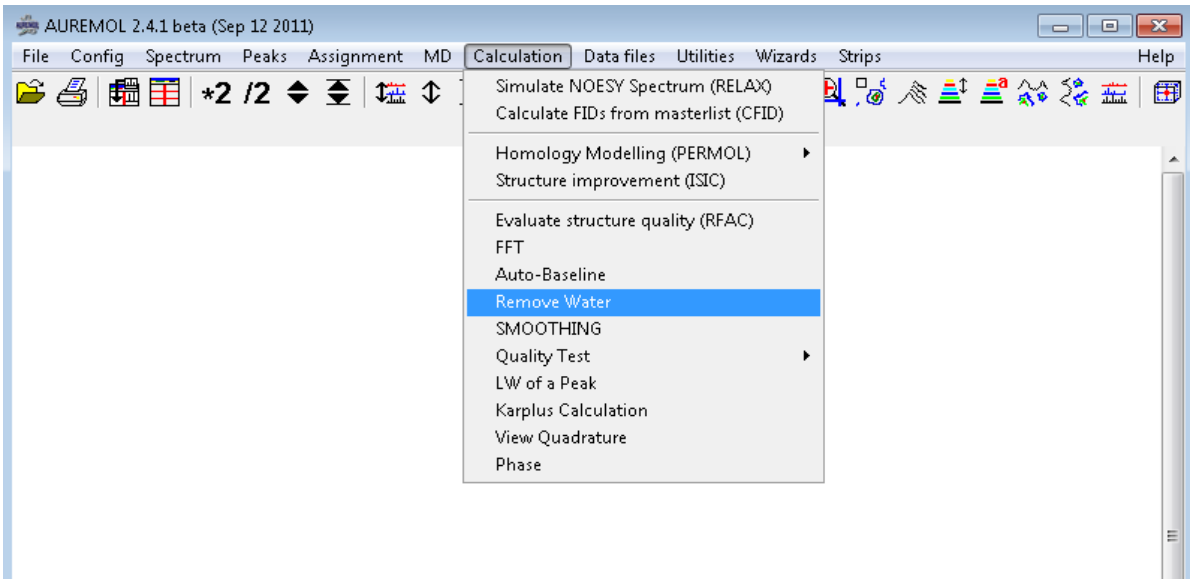


Figure 3.31 Starting the AUREMOL-SSA module to remove the water: the SSA module is in the "Remove Water" submenu of the "Calculation" menu in the AUREMOL software package.

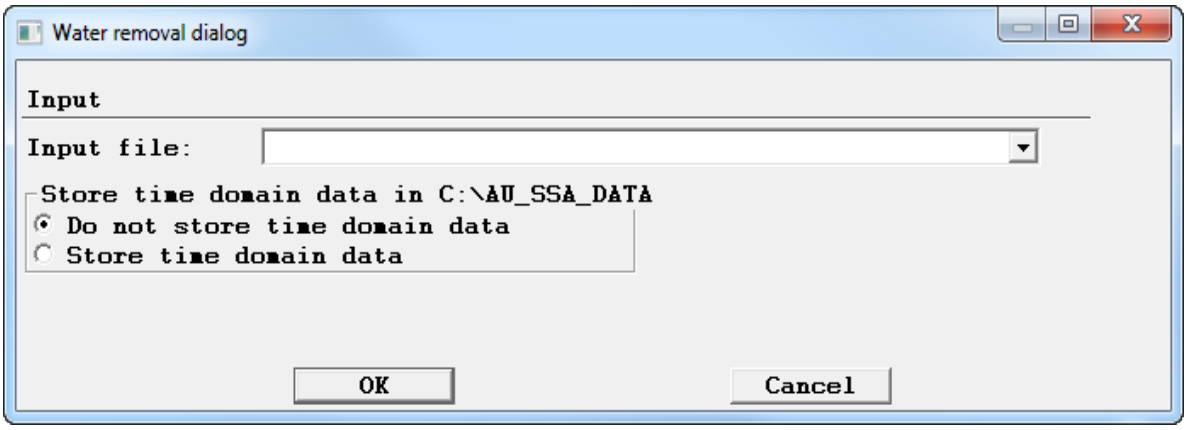


Figure 3.32 Main dialog of the AUREMOL-SSA: the time domain signal must be inserted as input file. If the user wishes to store the time-domain data after applying the solvent suppression by means of the SSA, the second option must be checked.

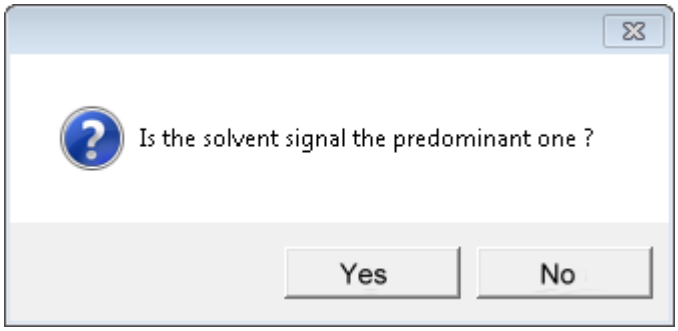


Figure 3.33 Warning message about the strength of the solvent signal in the investigated spectrum: the user must confirm that the solvent signal is the dominant one in the spectrum, otherwise the SSA calculation is promptly interrupted.

At the end of the computation for the solvent removal, the routine directly perfomes an hypercomplex Fourier transform. The processing parameters that had been eventually determined by the spectroscopist can be retained. If the processing files are not located in the same folder of the ser or fid file (see Fig. 3.34, part *a*), the user must indicate the folder path of such files (see Fig. 3.34, part *b*).

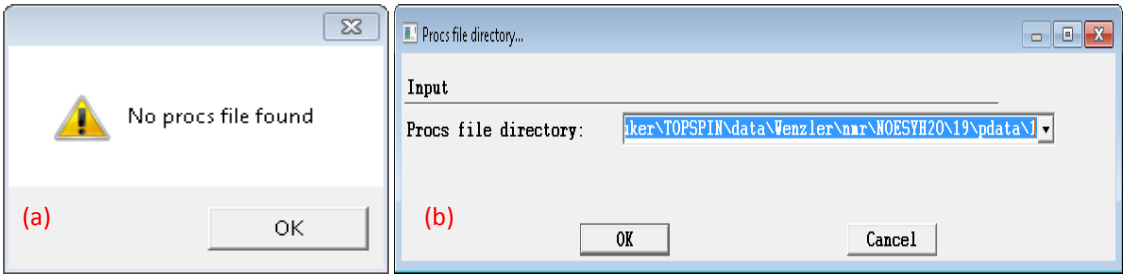


Figure 3.34 Processing files for Fourier transforming the data after the water removal: the user must provide the processing file path only if they are not in the same folder of the ser or the fid file.

If the processing files folder is not provided by the user and if the FnMODE parameter in the acquisition files describing the Fourier transformation type along the indirect direction (ω_1) is set to undefined, a message appears informing the user to verify the FnMODE directly in the processing files, using the MC2 parameter (see Fig. 3.35).

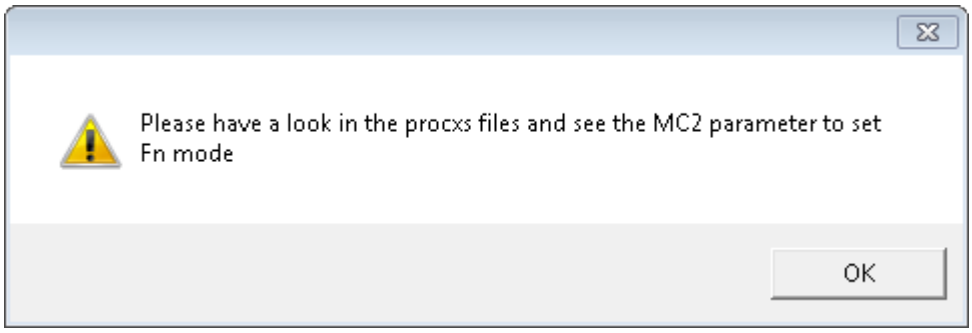


Figure 3.35 Message to identify the Fourier transformation type along the indirect direction: the MC2 parameter in the processing files contains such information in case that the FnMODE value in the acquisition files is set to undefined.

If the folder path containing the processing parameters has been provided the dialog reported in Fig. 3.36 is automatically filled for all the parameters. They can be anyway modified for obtaining different filtering types, data resolutions and phase corrections.

After the Fourier transformation, the spectrum without solvent appears to the user and the frequency domain files (2rr, 2ii, 2ri and 2ir in a two-dimensional case) and the processing files are generated and stored in a folder (called ssa) in the same directory of the ser or the fid files.

The post-processing step including the baseline correction is then applied if the user decides to perform it from the dialog reported in Fig. 3.37. In case that the baseline correction is requested the size of the sliding window (row-wise and column-wise) is automatically determined in accordance to the histograms of the line widths. The user is informed about these values (see Fig. 3.38) and he can eventually modify them. The final spectrum without solvent and baseline distortions appears at the end of the computation.

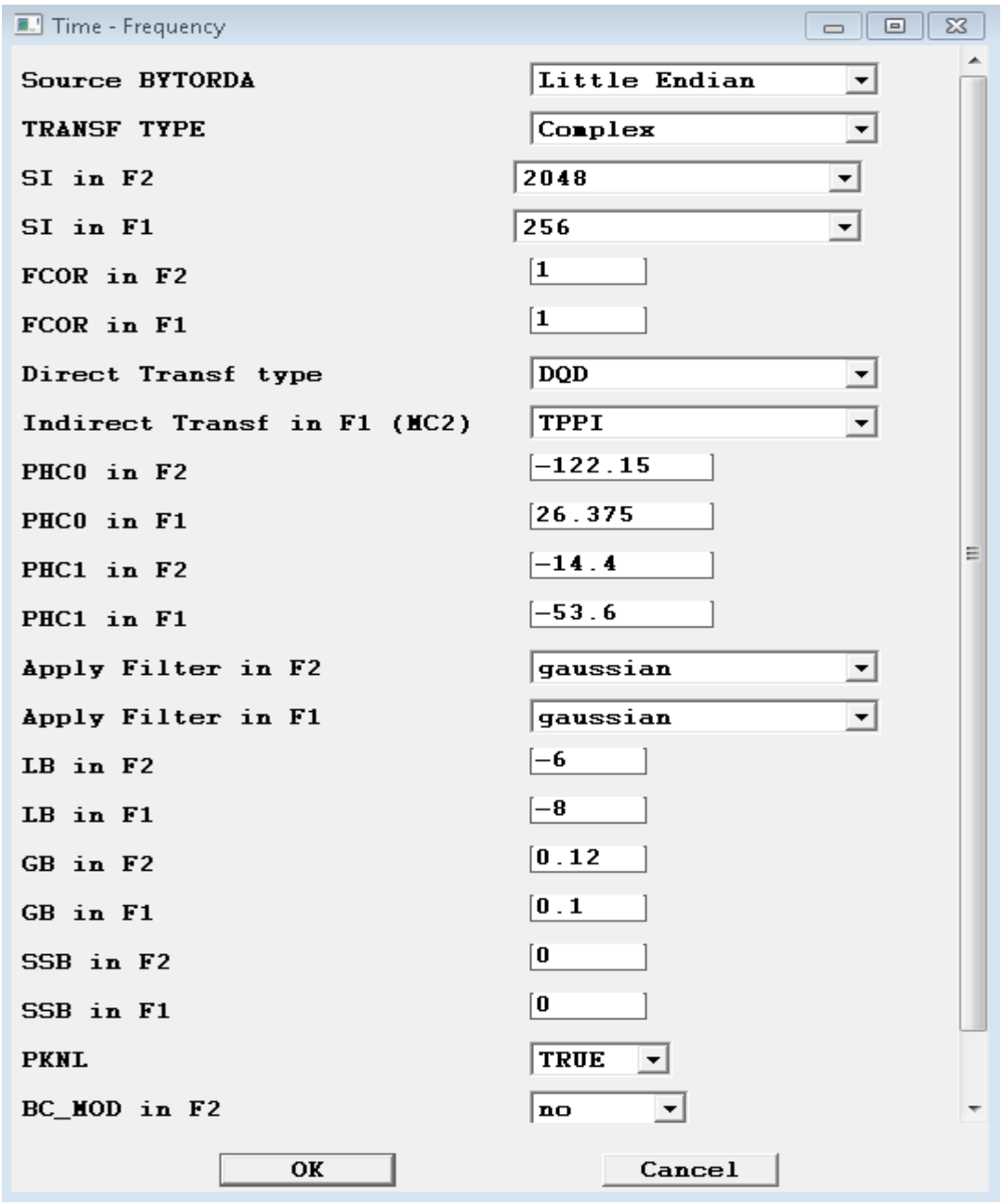


Figure 3.36 The AUREMOL dialog of the Fourier transformation: the dialog automatically contains all the previously defined processing parameters that can be modified in order to perform the Fourier transformation of the data in cascade with the solvent suppression by means of the SSA.

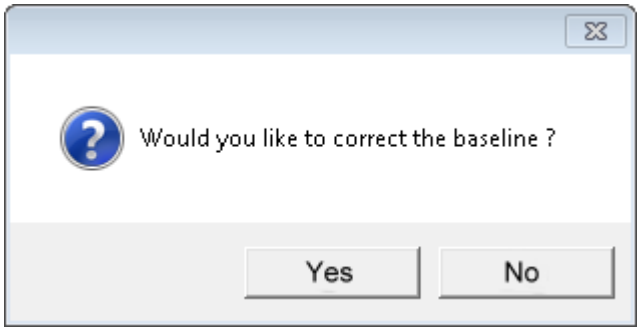


Figure 3.37 The ALS routine for baseline correction can be applied in cascade with the SSA: after removing the water from the time domain and after Fourier transforming, the user can decide to perform the baseline correction.

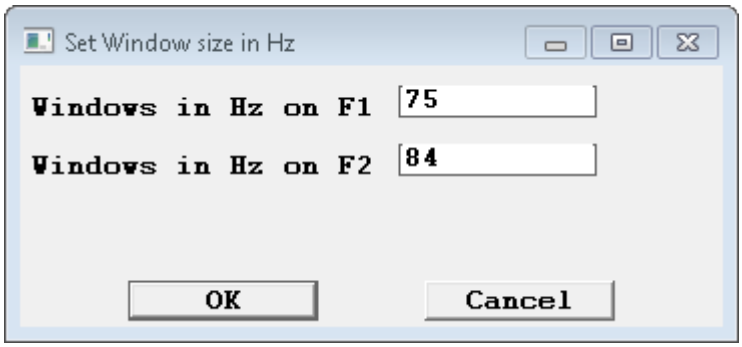


Figure 3.38 The automatically determined values of the window size for baseline point identification: the size of the window is obtained from the histograms of the line widths measured separately in the each direction.

The baseline correction can be applied out of the solvent suppression module. The "Auto-Baseline" submenu in the "Calculation" menu of AUREMOL can be used for such purposes (see Fig. 3.39). It performs the histogram analysis of the line widths and furnishes the window size values as previously shown in Fig. 3.38.

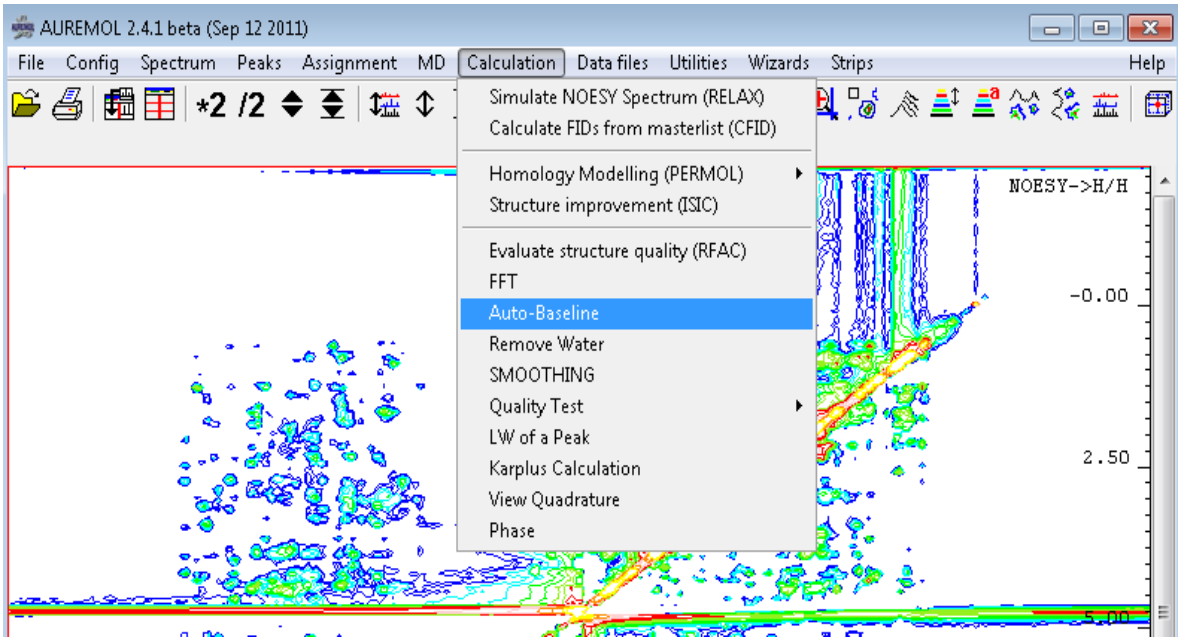


Figure 3.39 AUREMOL-ALS module: the baseline correction can be performed out of the solvent suppression routine selecting the Auto-Baseline option from the Calculation menu.

4 Alternative methods for solvent suppression

4.1 Comparison of methods

4.1.1 PCA OF TWO-DIMENSIONAL DATA

If the PCA is directly applied in the time domain on the whole set of FIDs without generating any trajectory matrix, satisfactory results are not obtainable. The number of projections along the directions spanned by the eigenvectors related to the largest eigenvalues does not correspond to the embedding dimensions (as in the SSA case) but it is equal to the number of measured FIDs. Dealing with such a larger number of estimated components (e.g. 512) and with many different time signals simultaneously, increases the computational time and generates the problem of identifying the components, since more than one projection could be related to the water artifact. Alternatively to the trajectory matrix, the autocorrelation matrix of the ensemble of FIDs may be generated in order to perform the PCA step [Mitschang et al, 1991]. If the data are not previously embedded in a specific manner in fact, the PCA cannot be satisfactorily applied to one-dimensional data.

The direct PCA application for solvent suppression (without any embedding) has demonstrated meaningful effects if performed in the frequency domain. This procedure relies on the direct eigenvalue decomposition of the covariance matrix (eq. 2.12) obtained from the multi-dimensional data matrix in the frequency domain after the two-dimensional Fourier transformation. As previously explained (in par. 2.2.2), the SSA is an extension of the PCA with the main difference concerning the trajectory matrix construction. In the PCA case this matrix is directly replaced by the multi-dimensional spectrum. The projections along the eigenvectors related to the largest eigenvalues are nullified (eq. 2.9) and a new data matrix is obtained (eq. 2.10).

The result of the PCA applied in the frequency domain (complex data) of the back-calculated two-dimensional (par. 2.1.1.1) NOESY spectrum of HPr from *Staphylococcus aureus* (H15A) is reported in Fig. 4.1. In this case the solvent suppression is not as strong as using the SSA. A visual or an automated inspection of the extracted components is mandatory in order to identify all of them representing the solvent artifact (in this case five of the 512 components have been discarded).

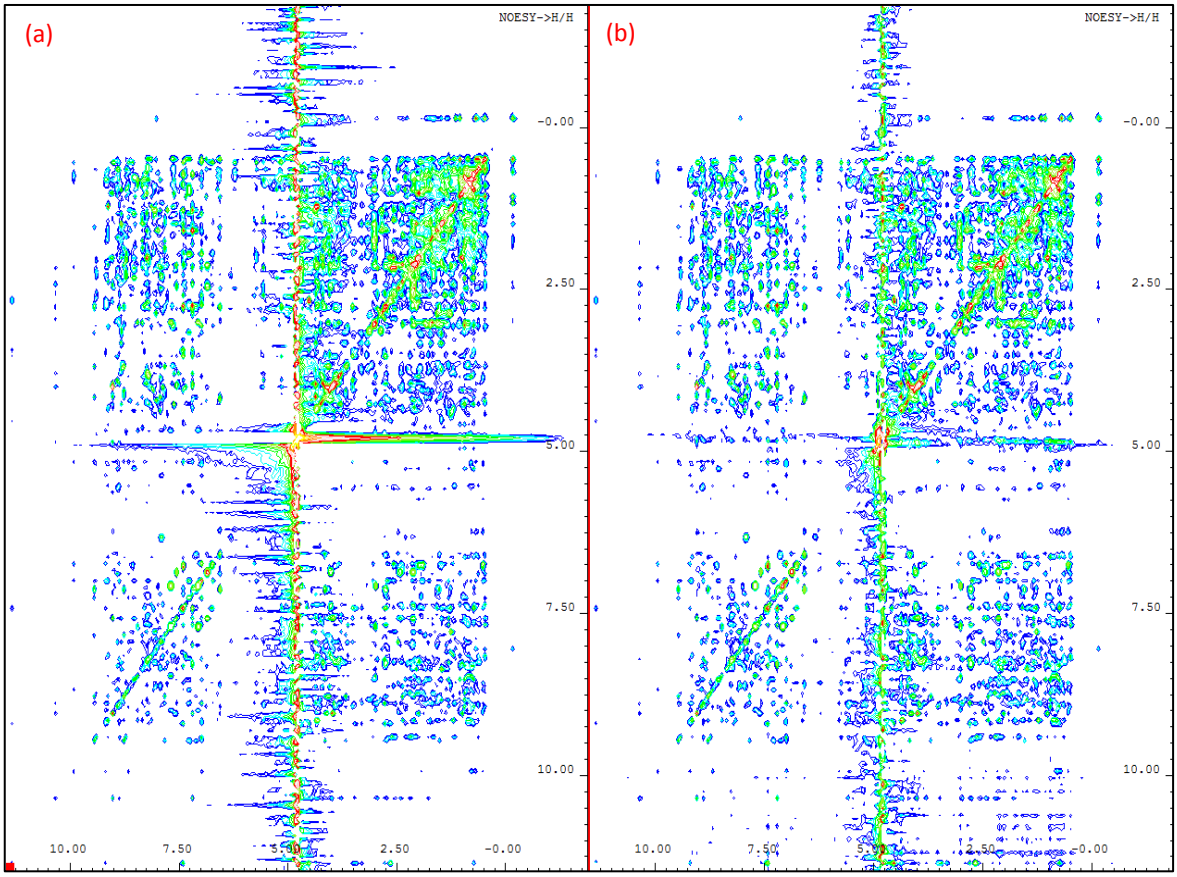


Figure 4.1 Solvent suppression by means of PCA applied in the frequency domain: back-calculated two-dimensional (par. 2.1.1.1) NOESY spectrum of HPr from *Staphylococcus aureus* (H15A). (a) original spectrum and (b) the spectrum after PCA.

4.1.2 ICA OF ONE-DIMENSIONAL DATA

Since SSA cannot be properly applied to spectra where the solvent artifact is not the dominant signal, the independent component analysis (ICA) represents a possible alternative. SSA is applied in 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 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 [Stadlthanner et al, 2006] 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.

The Fourier transform of a one-dimensional FID, of length N, is considered to be one of the n possible mixtures to be analyzed by the ICA. This number depends on the available measurements of the same dataset. The source separation problem consists in recovering a set of m independent source signals from (n >m) observed mixtures. It implies that more than one experimental FID of the same dataset needs to be Fourier transformed and then used as input to the separating algorithm. The simple collection of an arbitrary set of NMR experiments measured on the same sample is not enough to guarantee an optimal recovery of the resonances of interest. In order to reveal such resonances a proper protocol with a specific pulse sequence must be applied during the acquisition. In particular, the NMR mixtures must be properly generated (as a linear combination of the solute and the solvent signals) in order to obtain suitable ICA-tailored inputs (par. 2.1.2.5) during the experimental acquisition.

The number of sources m to estimate is for simplicity reduced at two, since a unique separation between the two signals of interest (i.e. the solvent and solute signals) is required. A general problem is the selection of the components to be removed. In SSA, usually the component with the largest eigenvalue is removed for solvent suppression, in ICA the component containing the water signal in the center of the spectrum must be removed after a visual 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 (see discussion section in the last chapter). SSA overcomes this problem since the natural ordering of the extracted components is strictly related to the variance of the involved signals.

In Fig. 4.2 the schematic representation of one-dimensional NMR data separation by means of ICA is compared with the SSA removal method. The ICA simply allows avoiding 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 (complex data). For instance, the group delay points at the beginning of the FIDs have not to 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 TOPSPIN software 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 take 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-mean and the unit-norm normalizations of the Fourier transformed signals are anyway 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.

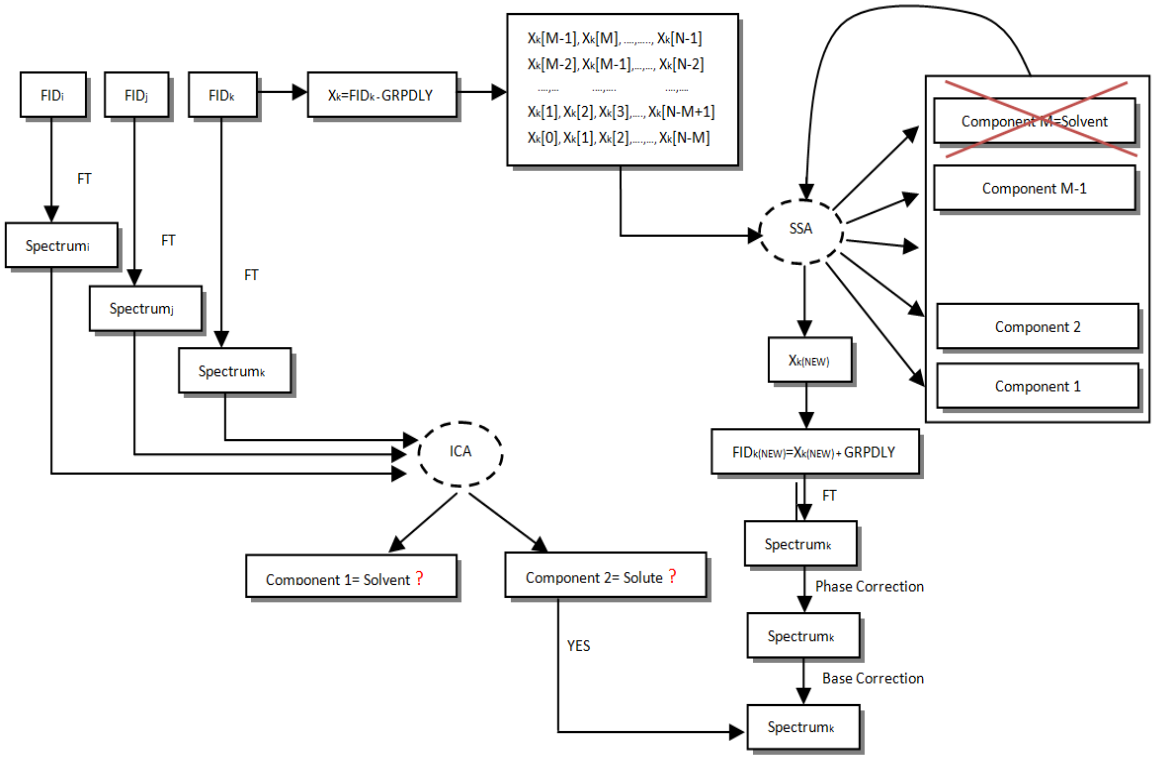


Figure 4.2 Schematic description of ICA and SSA applied to NMR data: ICA requires an ensemble of n one-dimensional spectra with different contributions of the solvent signal. ICA performs the separation of two main components on the frequency domain data: the solvent and the solute that nees to be automatically 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. ICA avoids all these steps, but it requires a visual inspection of the two extracted components in order to define which one must be retained (the solute signal).

4.1.2.1 ICA OF SIMULATED ONE-DIMENSIONAL SPECTRA

Both methods have been firstly applied on one-dimensional back-calculated HPr (H15A) protein spectra (par. 2.1.1.2). It is clear that for an optimal solvent removal by means of ICA a dataset of experiments must be generated with different weights of the solvent and of the solute signals. In order to investigate the most suitable input dataset for the ICA, the simulated HPr protein spectra have been added to experimental solvent spectra differing on the phase and/or on the intensity of the solvent signal. In particular, phase and intensity variations were performed in the frequency domain of the experimental solvent spectrum by different zero-order phases and by different scaling factors followed by an inverse Fourier transformation.

These data have been used to carry out a quantitative analysis of the performance of the ICA over a range of inputs with different solvent weights. Before the forward Fourier transformation the modified time domain water artifact signals have been added to the synthetic time domain signal of the protein scaled in such a way that the maximum of the water signal was 2.5 times smaller than the strongest protein resonance. As described by *De Sanctis et al, 2011*, the SSA algorithm has some limitations when it deals with spectra with a solvent signal having an intensity minor than the half of the strongest resonance of the protein in hand. A relatively small intensity of the solvent signal has intentionally produced with the aim to show the comparison between the ICA and SSA in such unfavorable cases.

The Fig. 4.3 shows such an ideal synthetic data set where a water signal with different phases and/or intensity was added to a simulated spectrum of HPr protein (par. 2.1.1.2) containing in addition artificial white noise. The zoom of the solvent peak is reported in each spectrum. In particular, in Fig. 4.3 the experimental water has been directly added to the simulated HPr spectrum (*a*); the phase of the experimental solvent has been corrected by forty-six degrees before adding it to the protein (*b*); the intensity of the water has been reduced by an adjunctive factor before mixing it with the solute (*c*); a phase change of nineteen degrees and an additional scaling factor have been applied on the solvent signal before generating the mixture (*d*). The reported cases represent only a part of the dataset created for applying a successive quantitative analysis of the number and of the type of the inputs used by the ICA algorithm. The SSA has been tested separately on anyone of the simulated spectra with an embedding of M= 20, whereas ICA has been applied on any combination of those spectra, varying the number and the type of inputs.

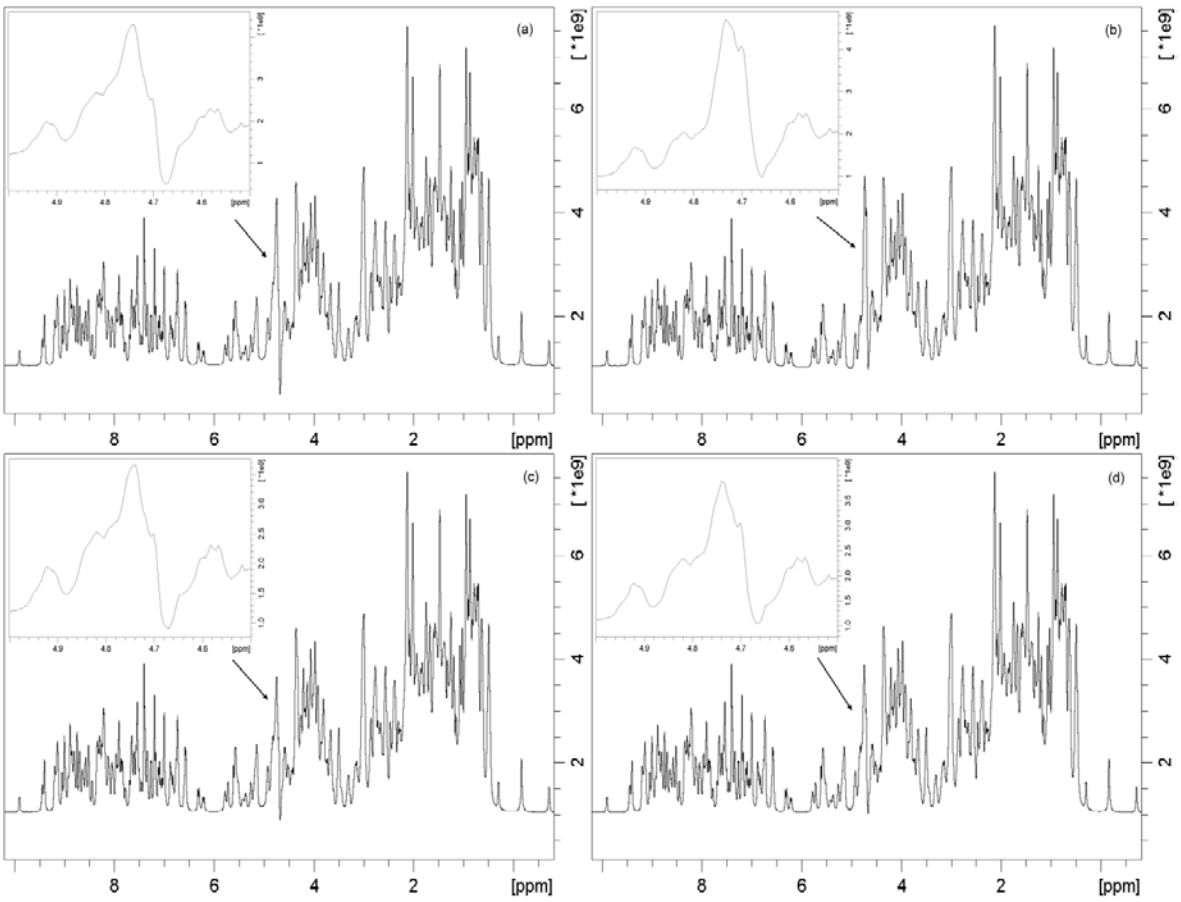


Figure 4.3 Fourier transforms of the back-calculated one-dimensional FIDs of the HPr protein (par. 2.1.1.2) from *Staphylococcus aureus* (H15A) added to four different experimental solvent signals: water artifact 2.5 times smaller than the strongest resonance of interest having the same acquisition parameters of the simulated protein, with no phase and intensity change (a), with a phase change of 46° (b), with an intensity variation (c) and with an intensity change and a phase correction of 19° (d). A zoom of the solvent is reported in each spectrum.

The application of ICA on two back-calculated spectra (parts (*a*) and (*b*) of Fig. 4.3) with the water signals having different phases leads to a not optimal recovery of the protein spectrum. The use of a third input (part (*c*) or part (*d*) of Fig. 4.3) allows an almost perfect removal as described in Fig. 4.4 (upper trace) when compared to the spectrum before adding the experimental solvent artifact (Fig. 4.4, central trace). This result is independent on the selection of the three spectra. Therefore, the same is true when the water signal has different intensities relative to the protein signal or differs in phase as well as in the intensity.

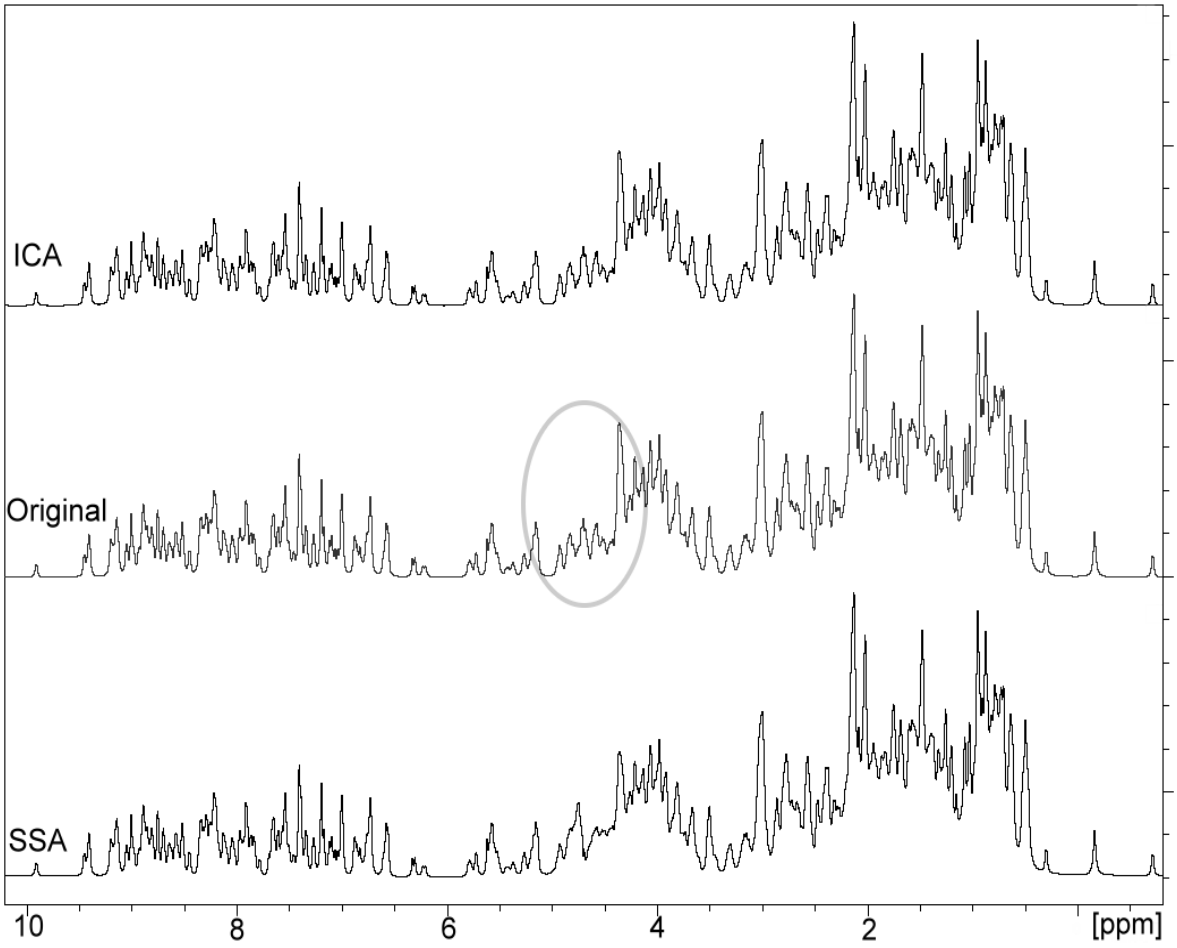


Figure 4.4 Application of SSA and ICA to the one-dimensional HPr synthetic data set (par. 2.1.1.2): (bottom trace) SSA has been applied to the FID of a back-calculated one-dimensional HPr protein spectrum with an experimental solvent signal added (part *a* of Fig. 4.3); (top trace) ICA has been applied to three simulated spectra with variations on the phase and on the intensity of the solvent as shown in Fig. 4.3 (part *a*, *b* and *c* or *d*); (central trace) original synthetic spectrum without additional water signal.

In summary, the application of ICA works always equally good independently on the relative phases and intensities of the water signals, as long as they are significantly different. From a quantitative analysis of the ICA applied on the simulated dataset it has been demonstrated that the performance can differ in dependence on the number of inputs or on the strength of solvent variation. For a detailed analysis see Fig. 4.5 and the corresponding Table 4.1. It is evident that using only two inputs to the ICA algorithm does not lead to extremely good results, especially when the phase or the intensity variation of the solvents in the experiments is very weak. Increasing such variation improves the recovery, overcoming the SSA performance with a phase variation major than ninety degrees or when three inputs are used by the ICA algorithm (see fig. 4.5).

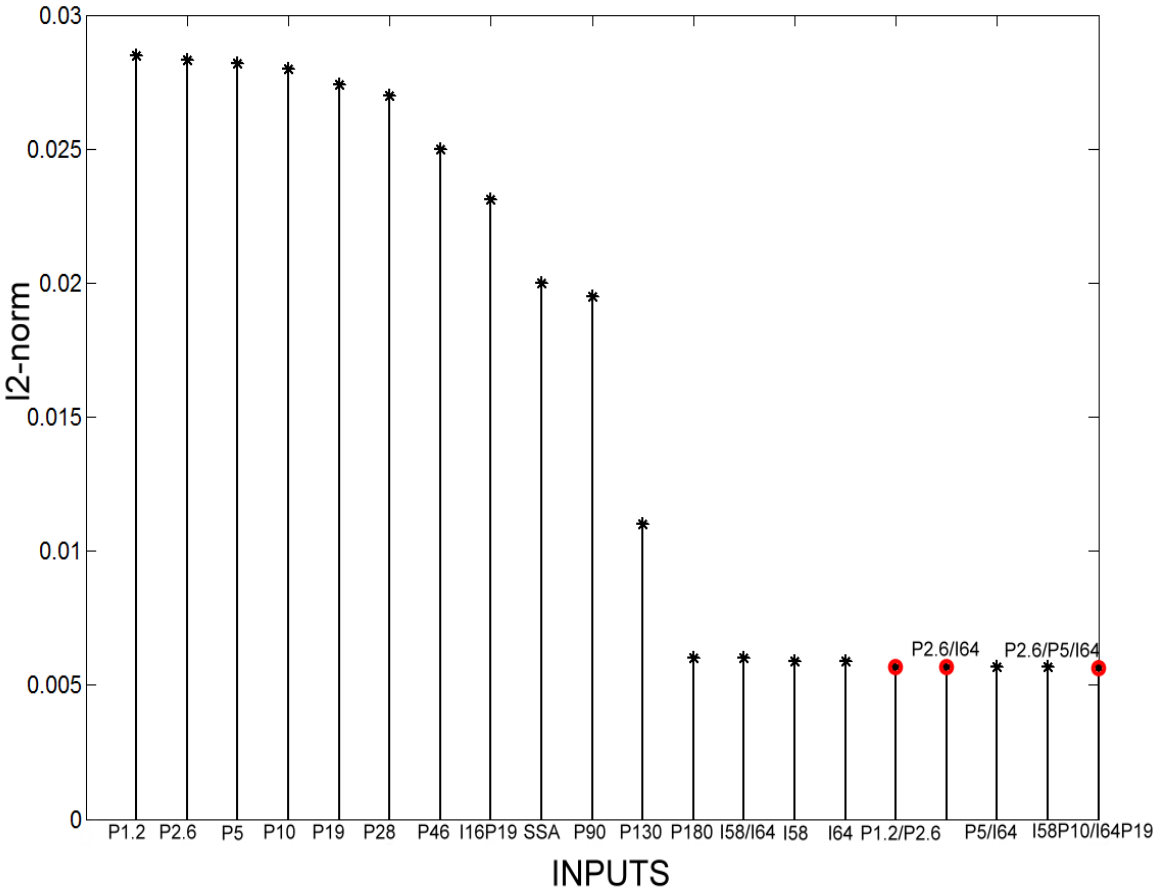


Figure 4.5 Dependence of the performance of ICA on the number and on the type of inputs: the data shown in Fig. 4.3 were used for the analysis with some additional cases (see Table 4.1). As a measure for the performance of the ICA, the l2-norm, calculated on the difference between the original simulated HPr spectrum and an HPr spectrum where the water signal was 500 times stronger than the most intense protein signal, was arbitrarily set to 1. The ICA has been applied on the simulated HPr spectrum with additive experimental water (part *a* of Fig. 4.3) and on the simulation mixed with a different solvent signal having a modified phase (indicated by P) and/or a different intensity (represented by the letter I). The performance improves increasing either the number of inputs or the strength of the phase and/or of the intensity variation (red highlighted cases).

Symbol	Number of Inputs	First input: PHC0	Second input: PHC0	Third input: PHC0	Fourth input: PHC0	First input: Intensity	Second input: Intensity	Third input: Intensity	Fourth input: Intensity
P1.2	2	0°	1.2°	-	-	S(P)	S(P)	-	-
P2.6	2	0°	2.6°	-	-	S(P)	S(P)	-	-
P5	2	0°	5°	-	-	S(P)	S(P)	-	-
P10	2	0°	10°	-	-	S(P)	S(P)	-	-
P19	2	0°	19°	-	-	S(P)	S(P)	-	-
P28	2	0°	28°	-	-	S(P)	S(P)	-	-
P46	2	0°	46°	-	-	S(P)	S(P)	-	-
I16P19	2	0°	19°	-	-	S(P)	S(P)/16	-	-
SSA	1	0°	-	-	-	S(P)	-	-	-
P90	2	0°	90°	-	-	S(P)	S(P)	-	-
P130	2	0°	130°	-	-	S(P)	S(P)	-	-
P180	2	0°	180°	-	-	S(P)	S(P)	-	-
I58/I64	3	0°	0°	0°	-	S(P)	S(P)/58	S(P)/64	-
I58	2	0°	0°	-	-	S(P)	S(P)/58	-	-
I64	2	0°	0°	-	-	S(P)	S(P)/64	-	-
P1.2/ P2.6	3	0°	1.2°	2.6°	-	S(P)	S(P)	S(P)	-
P2.6/I64	3	0°	2.6°	0°	-	S(P)	S(P)	S(P)/64	-
P5/I64	3	0°	5°	0°	-	S(P)	S(P)	S(P)/64	-
I58P10/ I64P19	3	0°	10°	19°	-	S(P)	S(P)/58	S(P)/64	-
P2.6/P5 /I64	4	0°	2.6°	5°	0°	S(P)	S(P)	S(P)	S(P)/64

Table 4.1 Symbol interpretation of Figure 4.5: number of inputs to the ICA algorithm; phase P and intensity I variation of the solvent in the first, second, third and fourth input. The zero-order phase PHC0 of the solvent is set to 0° in case of direct sum of the simulated solute and the experimental solvent signals. The intensity scale of the solvent is set to S(P), corresponding to a water maximum 2.5 times smaller than the strongest protein resonance, in the default case. The best solvent removal performances are highlighted in red.

This assertion involves two possible sceneries: when the spectroscopist acquires the data, he does not need to measure more than two spectra of the same sample, but the solvent signal on those spectra must be as much different as possible between the two experiments; the spectroscopist must acquire three experiments from the same sample with a weak variation in the phase and/or in the intensity of the solvent signal in accordance to some ICA-tailored inputs. As described in Fig. 4.5 the best results have been achieved either using three inputs with very small phase variations (i.e. 0, 1.2 and 2.6 degrees in each experiment) or two inputs with strong differences (i.e. 130 degrees). When SSA is applied to one of the spectra, the solvent resonance is strongly suppressed but very close to the former position of the water signal, the resonance recovery is not perfect (Fig. 4.4, top trace). On the other hand, using ICA the calculation becomes computationally more involving and one has to inspect more than one component interactively or solve the non-trivial task to select the valid components.

Inspecting the results obtained from the simulated dataset, one can conclude that it is advisable to apply ICA in the unfavorable case of a spectrum with a very weak solvent signal and that using ICA-tailored inputs leads to perfect recovery of the resonances of interest.

The first assertion has been confirmed applying the ICA and the SSA on the experimentally acquired one-dimensional human urine spectra. In order to confirm the second assertion, the ICA and the SSA have been then applied on the experimentally acquired one-dimensional HPr protein spectra from *Staphylococcus carnosus* designed as ICA-tailored inputs.

4.1.2.2 ICA OF EXPERIMENTAL ONE-DIMENSIONAL SPECTRA

4.1.2.2.1 HUMAN URINE SPECTRA

The algorithms (SSA and ICA) have been applied on two datasets of experimental one-dimensional spectra of human urine recorded with a 1D-NOESY pulse sequence [McKay, 2011] with different mixing times τ, as shown in Fig. 4.6. The first urine dataset (mixing time of 1500 and 2000 ms) has been used to show the SSA limitations and to propose ICA as a valid alternative. The second urine dataset (mixing time of 10 and 20 ms) has been chosen in order to demonstrate the ICA limitations if applied on experimental spectra without a proper data acquisition for a suitable ICA application.

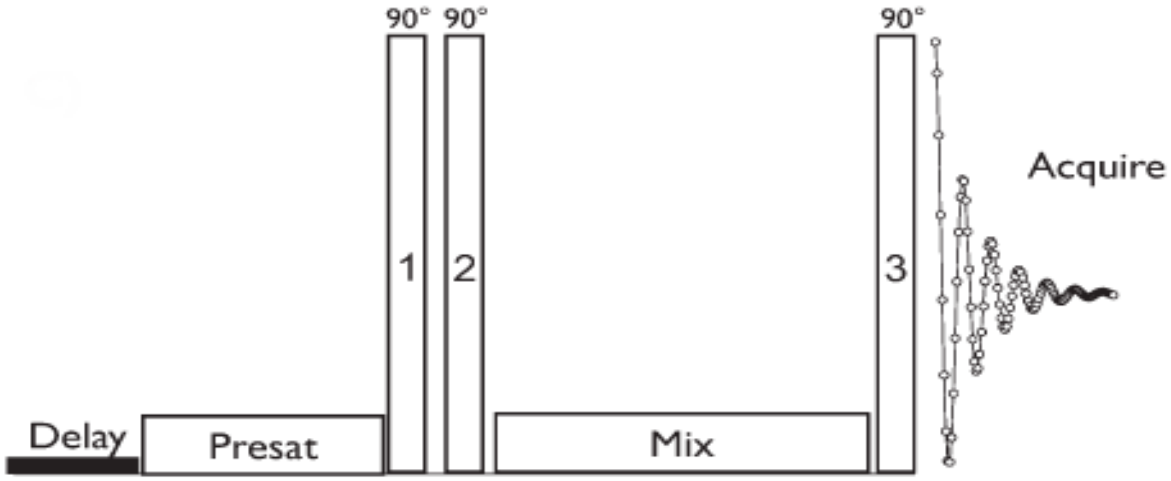


Figure 4.6 Pulse sequence of the human urine dataset: 1D-^{1}H-NOESY-type [McKay, 2011]. After a long low power saturation period, two initial 90 degree pulses, another long low power saturation period and a final 90 degree pulse are applied before starting the FID acquisition.

The Fig. 4.7 shows four experimental NOESY-type 1D spectra of human urine recorded with different mixing times. Fig. 4.7 part *a*, shows the Fourier transform of the FIDs acquired with a mixing time of 10 ms (left side) and 20 ms (right side), whereas the *b* part shows the urine spectra acquired with a mixing time of 1500 ms (left side) and 2000 ms (rights side), including a zoom of the solvent signals of each spectrum. The SSA has been applied separately on the experiments with a mixing time of 10 ms and on the 2000 ms, with M= 40. ICA has been instead applied separately on the two datasets.

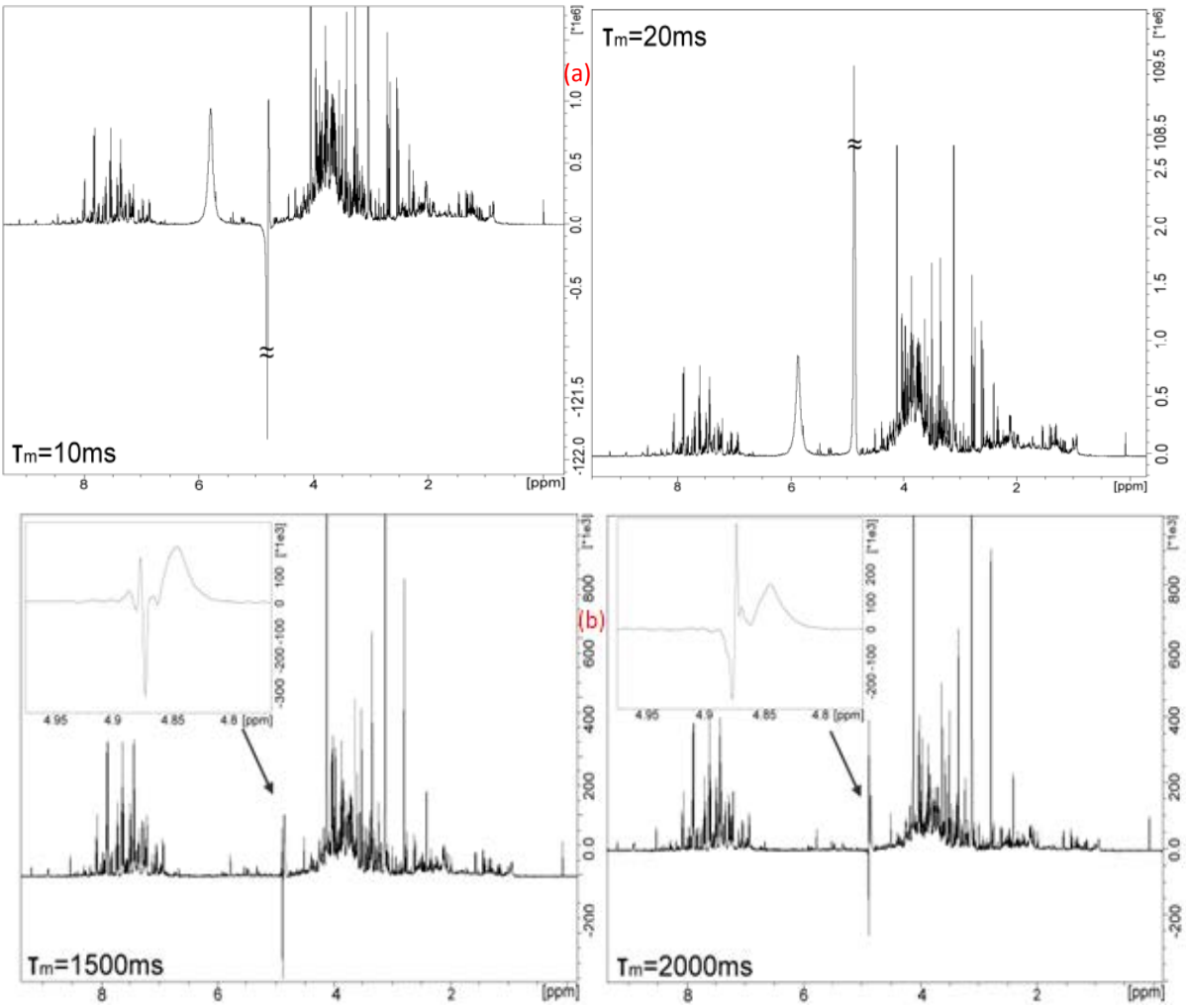


Figure 4.7 One-dimensional human urine spectra (par. 2.1.2.4.1) with different mixing times: (left side) 10 ms and (right side) 20 ms (a); (left side) 1500 ms and (right side) 2000 ms including a zoom of the solvent signal (b).

In Fig. 4.8, part *a*, the urine spectrum has been displayed after the ICA solvent removal procedure (in the first urine dataset) and it has been compared with SSA. In this case the very strong solvent signal has been better removed by the SSA that does not compromise the intensities of the solute resonances of interest. ICA in such case

does not remove the solvent signal as properly as the SSA does (see the box depicted in Fig. 4.8 part *a*). However, as previously described SSA has some limitations. This work has the aim to propose a valid alternative method that allows overcoming such problems. In Fig. 4.8 part *b* the comparison between SSA and ICA has been thus reported in a typical unfavorable case for the SSA algorithm (with a weak solvent signal as in the second urine dataset). It is evident that also when dealing with experimental data, the ICA allows a better removal than the SSA if the solvent signal is not the dominant one. As depicted in Fig. 4.8 part *b*, the SSA removes the component with the highest variance in the spectrum that in this case was not the solvent.

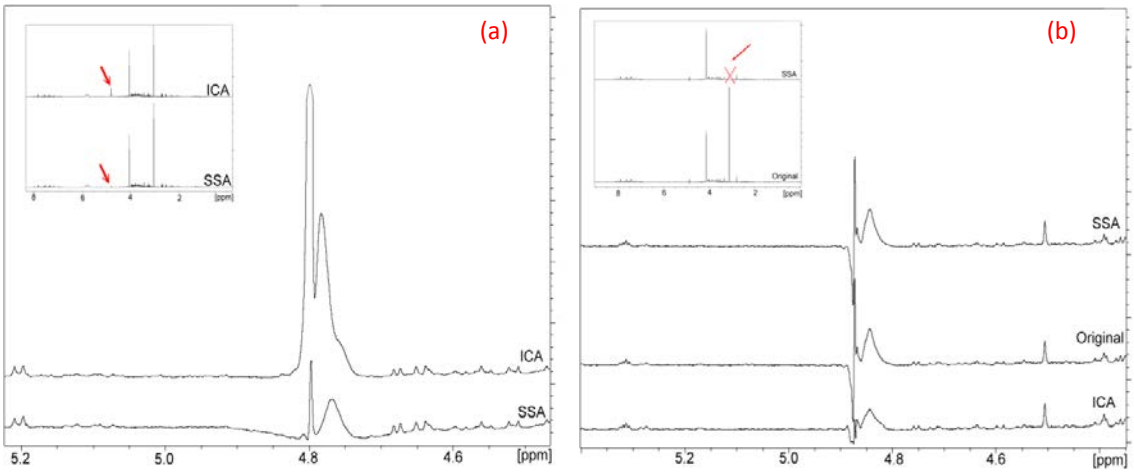


Figure 4.8 ICA and SSA application on the two human urine dataset (par. 2.1.2.4.1): (a) zoom of solvent artifact removed by SSA applied on the FID of the one-dimensional human urine spectrum measured with a mixing time of 10 ms (lower trace) and zoom of solvent artifact removed by ICA applied on the two experimental one-dimensional spectra with a mixing time of 10 ms and 20 ms (upper trace). The complete spectra are depicted in the box on the left top of the figure. The results of both methods can be compared with the experimental spectra having a mixing time of 10 ms and 20 ms described in Fig. 4.7 part *a*; (b) zoom of solvent artifact removed by SSA used on the FID of the one-dimensional human urine spectrum measured with a mixing time of 2000 ms (upper trace); zoom of solvent artifact removed by ICA used on the two experimental one-dimensional spectra with a mixing time of 1500 ms and 2000 ms (lower trace). Comparison of the results of both methods with the experimental spectrum having a mixing time of 2000 ms without any solvent removal (middle line). On the left top of the *b* part of the figure the complete original spectrum (lower trace) and the spectrum after SSA removal (upper trace) are shown in order to highlight the loss of the strongest resonance of interest due to the SSA failure.

4.1.2.2.2 HPr ICA-TAILORED SPECTRA

In order to obtain suitable linear combinations of solute and solvent signals, some additional ICA-tailored inputs have been acquired (par. 2.1.2.5). In particular, two

dataset of one-dimensional HPr protein spectra from *Staphylococcus carnosus* have been measured.

In Fig. 4.9 the first dataset of ICA-tailored inputs is highlighted: it is made up of two one-dimensional experiments acquired with a different phase cycling, whose pulse sequence in reported in Fig. 4.10. ICA has been applied on both spectra (blue trace). The performance of the ICA is not satisfactory, thus a second dataset made up of two experiments has been generated with different diffusion times. In Fig. 4.11, the Fourier transform of the second dataset is represented, while the pulse sequence is reported in Fig. 4.12. SSA has been applied only on the first spectrum with M= 40, whereas ICA has used all the two measured spectra.

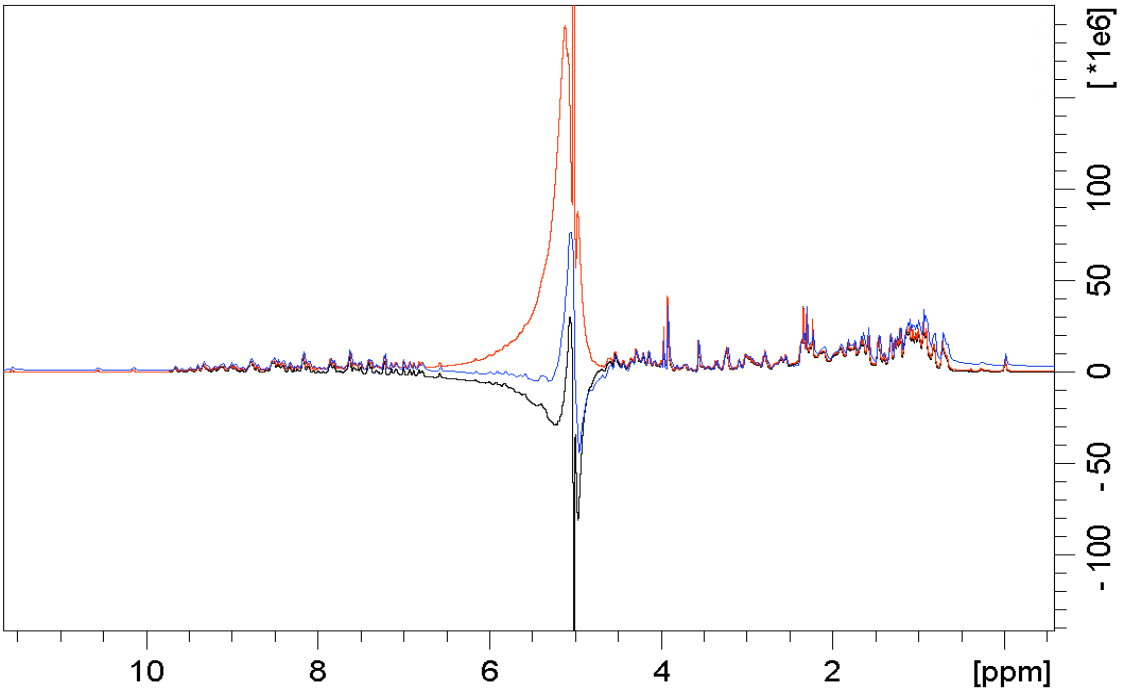


Figure 4.9 Two one-dimensional experimental HPr protein spectra (first dataset) with a different phase cycling used as ICA-tailored inputs (par. 2.1.2.5): complete spectrum of the first (black trace) and of the second (red trace) experiment. They are compared with the signal after solvent removal by means of ICA (blue trace).

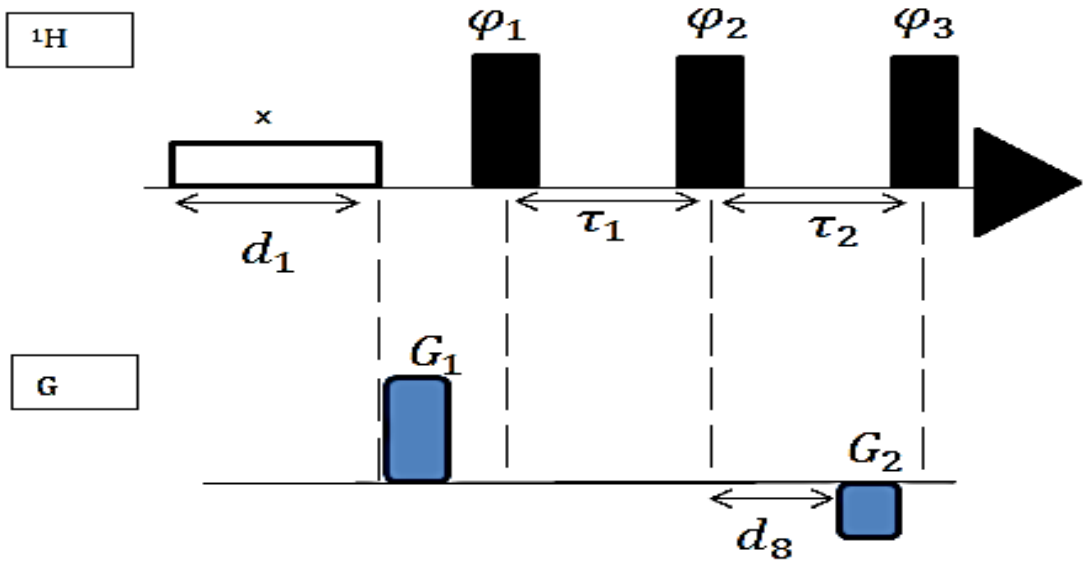


Figure 4.10 Pulse sequence of the first ICA-tailored dataset: two one-dimensional spectra of HPr protein from *Staphylococcus carnosus* have been measured with a different phase cycling. The solid rectangular bars represent 90° selective pulses. Gradient pulses applied along the z-axis are represented by blue boxes. The phase cycles are: $\varphi_1 = (x)$, $\varphi_2 = (x)$, $\varphi_3 = 2(x)2(-x)2(y)2(-y)$ and receiver $= 2(x)2(-x)2(y)2(-y)$ (first experiment); $\varphi_1 = (-x)$, $\varphi_2 = (x)$, $\varphi_3 = 2(x)2(-x)2(y)2(-y)$ and receiver $= 2(x)2(-x)2(y)2(-y)$ (second experiment). Relaxation delay d_1, 1 s; G_1 gradient length, 1 ms; G_1 gradient strength, 50 G/cm; G_2 gradient length, 1 ms; G_2 gradient strength, -10 G/cm; mixing time d_8, 10 ms; delay for gradient recovery d_{16}, 0.5 ms; gradient shape, sinus.

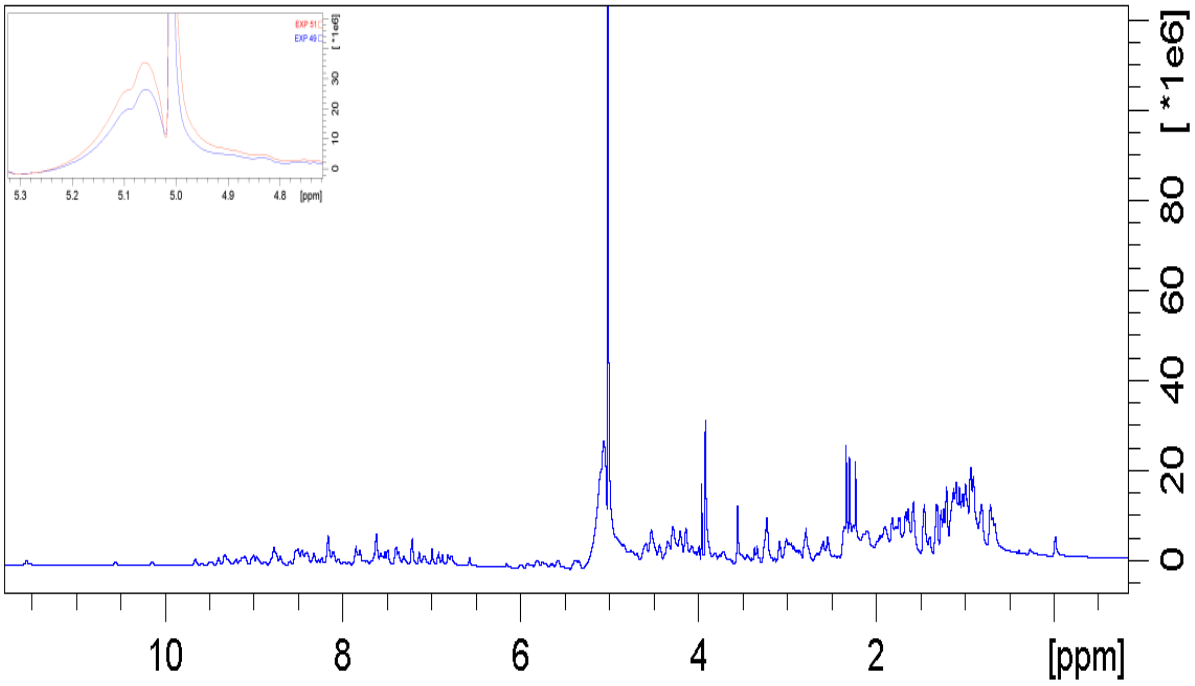


Figure 4.11 Two one-dimensional experimental HPr protein spectra (second dataset) with different diffusion times used as ICA-tailored inputs (par. 2.1.2.5): complete spectrum of the first experiment and zoom of the solvent artifacts of the two spectra. Gradient weights G_2: 80 G/cm (blue trace) and 50 G/cm (red trace).

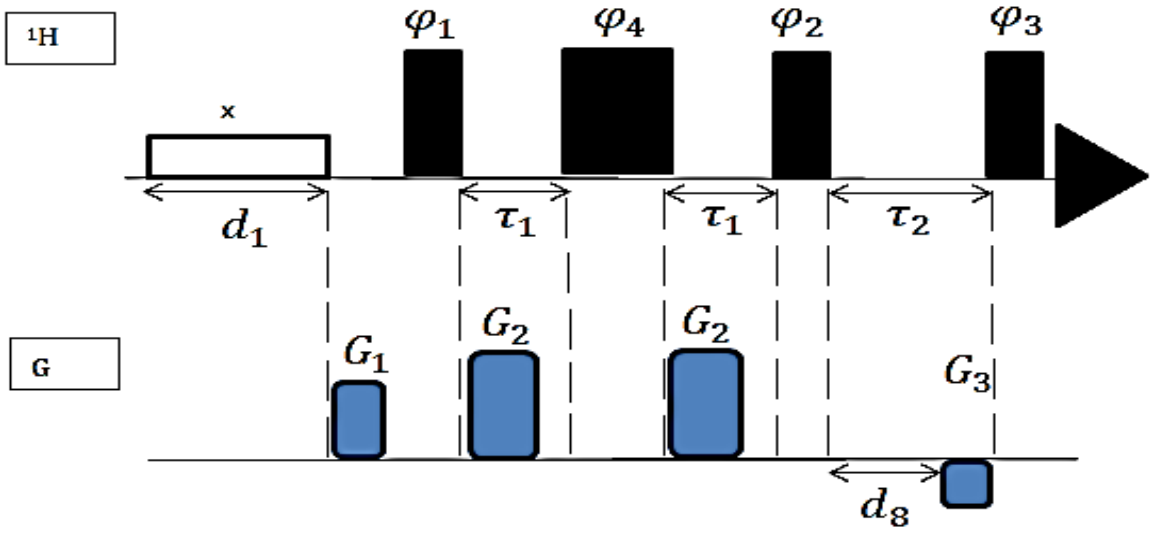


Figure 4.12 Pulse sequence of the second ICA-tailored dataset: two one-dimensional spectra of HPr protein from *Staphylococcus carnosus* have been measured with different diffusion times. The thin and the thick solid rectangular bars represent 90° and 180° selective pulses respectively. Gradient pulses applied along the z-axis are represented by blue boxes. The phase cycles are: $\varphi_1 = (x)(-x)$, $\varphi_4 = (y)(-y)$, $\varphi_2 = 8(x)8(-x)$, $\varphi_3 = 2(x)2(-x)2(y)2(-y)$, and receiver $=(x)2(-x)(x)(y)2(-y)(y)(-x)2(x)(-x)(-y)2(y)(-y)$. Relaxation delay d_1, 1 s; G_1 gradient length, 1 ms; G_1 gradient strength, 50 G/cm; G_2 gradient length, 4 ms; G_2 gradient strength, 80 G/cm; G_3 gradient length, 1 ms; G_3 gradient strength, -10 G/cm. G_2 gradient weights: 80 G/cm (first experiment) and 50 G/cm (second experiment); mixing time d_8, 10 ms; delay for gradient recovery d_{16}, 0.5 ms; gradient shape, sinus.

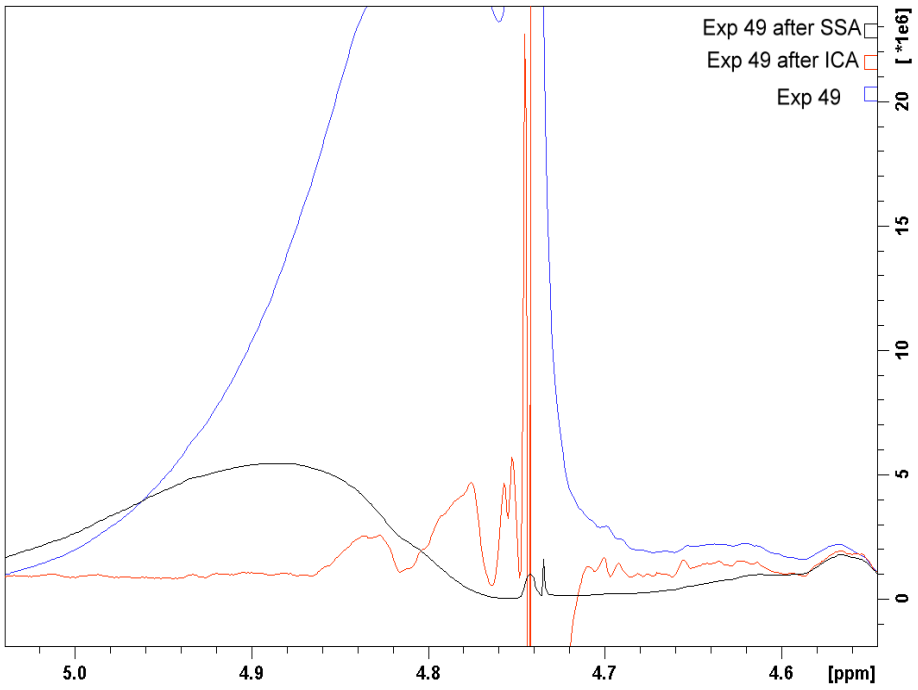


Figure 4.13 ICA and SSA application on the ICA-tailored one-dimensional experimental HPr protein from *Staphylococcus carnosus* with different diffusion times (second dataset): zoom of the solvent artifact removed by SSA (black trace) applied on the first experiment; zoom of the ICA applied on the second dataset (par. 2.1.2.5) made up of two ICA-tailored experiments (red trace); first of the two ICA-tailored spectra (blue trace).

As evident from the comparison reported in Fig. 4.13, the SSA (black trace) removes almost completely the dominant solvent signal, but does not reveal any resonance of interest previously hidden by the solvent artifact. The ICA recovers almost all the resonances in the solvent area (red trace). Moreover, a detailed inspection of the H^{α} chemical shifts of the HPr protein from *Staphylococcus carnosus* may allow the recognition and the assignment of such resonances.

4.1.3 EMD OF ONE-DIMESIONAL SPECTRA

The EMD of time-domain signals (complex data) has been performed on the one-dimensional spectrum of a sample containing a mixture of five amino acids (par. 2.1.2.6.1), shown in Fig. 4.14. The group delay has been excluded before applying the decomposition. The routine has generated sixteen different IMFs. It has been initially applied with solvent removal purposes but since its lower performance than the SSA, this algorithm revealed to be more suitable for automated phase correction procedures.

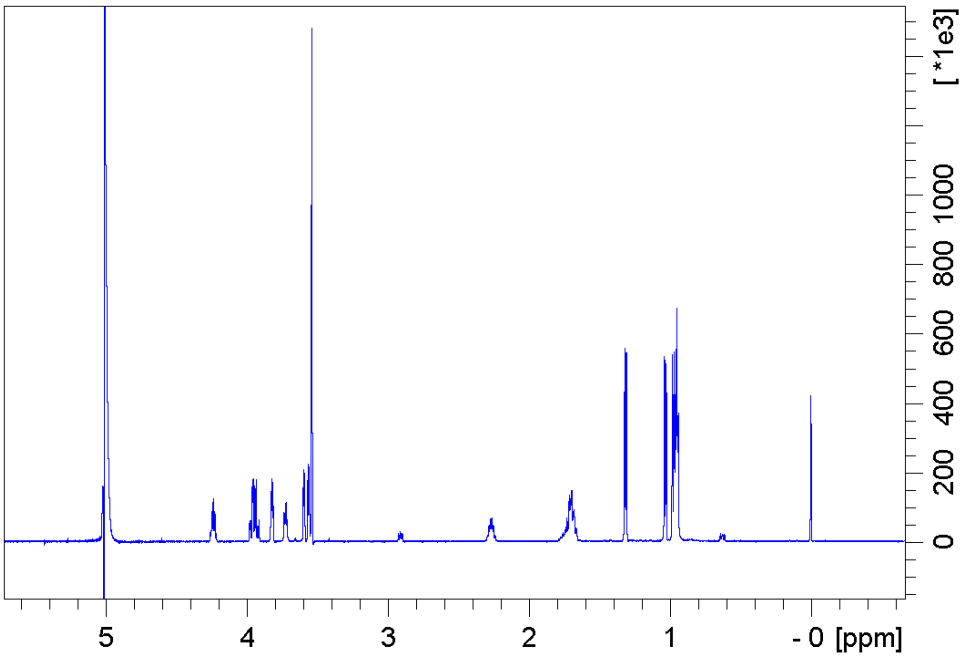
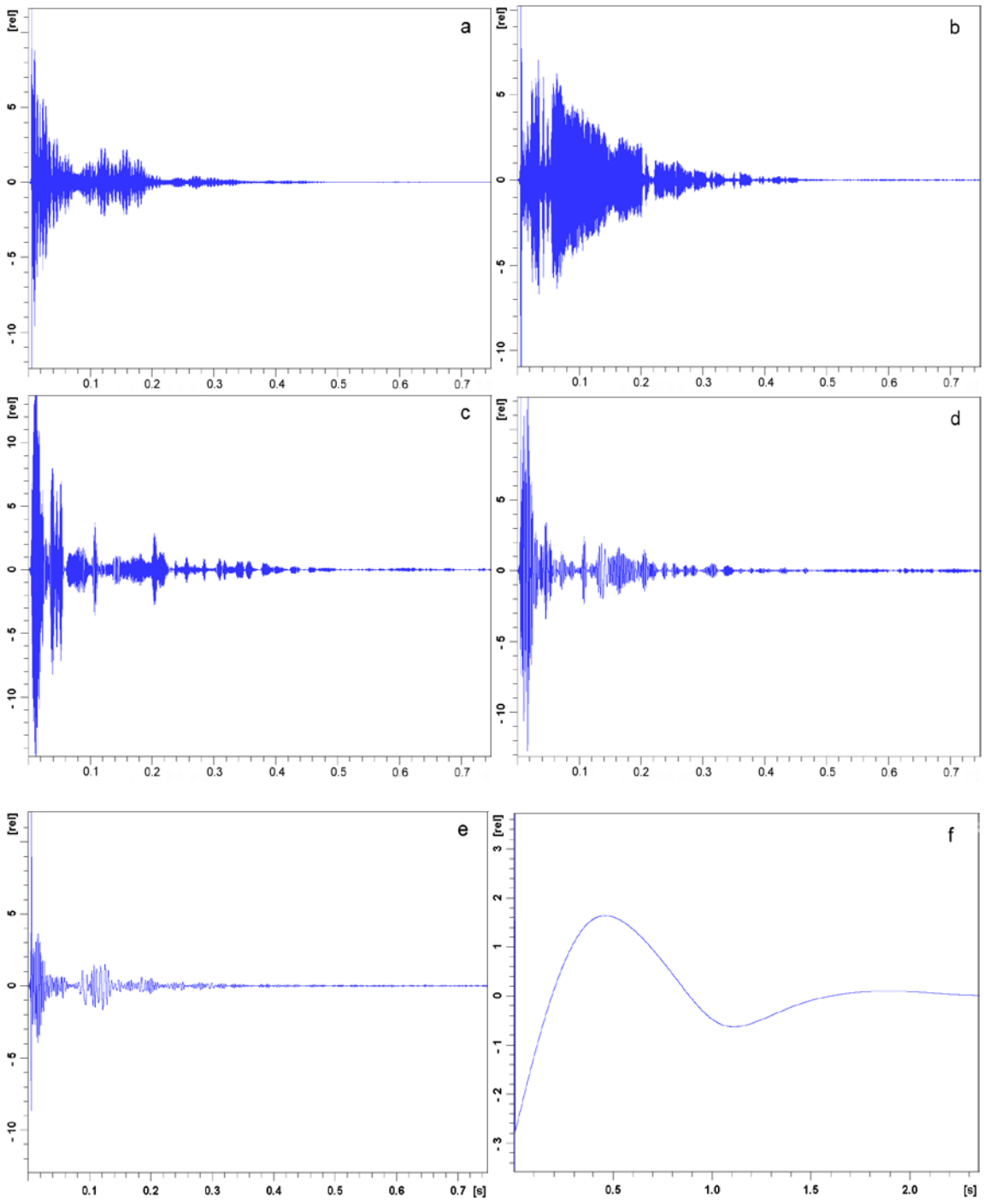


Figure 4.14: One-dimensional spectrum of a sample with a mixture of five amino acids: spectrum description in par. 2.1.2.6.1.

In Fig. 4.15 some of the IMFs are reported. The detailed analysis of the oscillating components extracted from the time domain signal is a necessary step for the rejection of a specific signal as the solvent artifact. It is evident that the frequency of the oscillations decreases from the first to the last component. In particular, the first five IMFs contain all the high frequency oscillations of the recorded FID, whereas the last IMFs reveal the low ones.



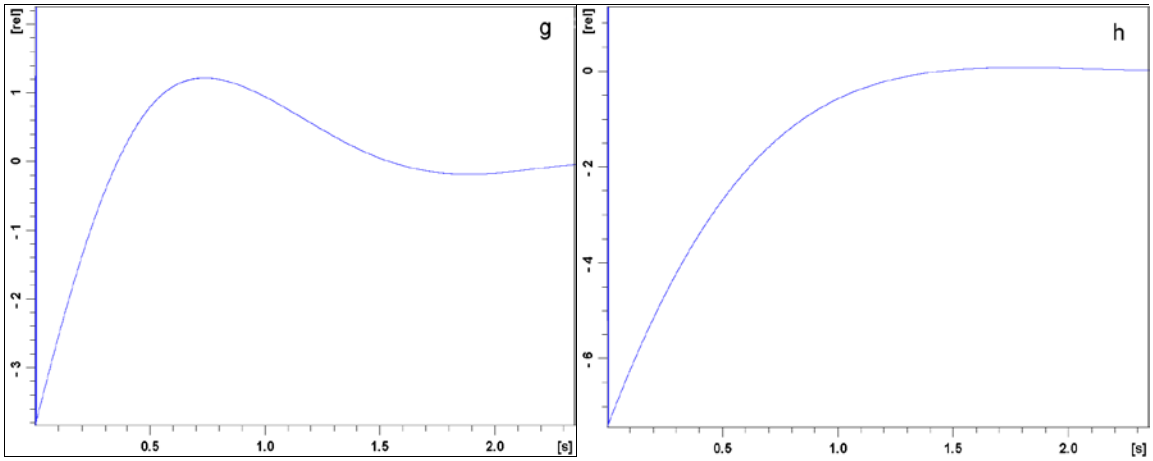


Figure 4.15 IMFs extracted from the one-dimensional spectrum measured from the mixture of five amino acids: it represents the first (a), the second (b), the third (c), the fourth (d), the fifth (e), the fifteenth (f) and the sixteenth (g) IMF and the residual (h).

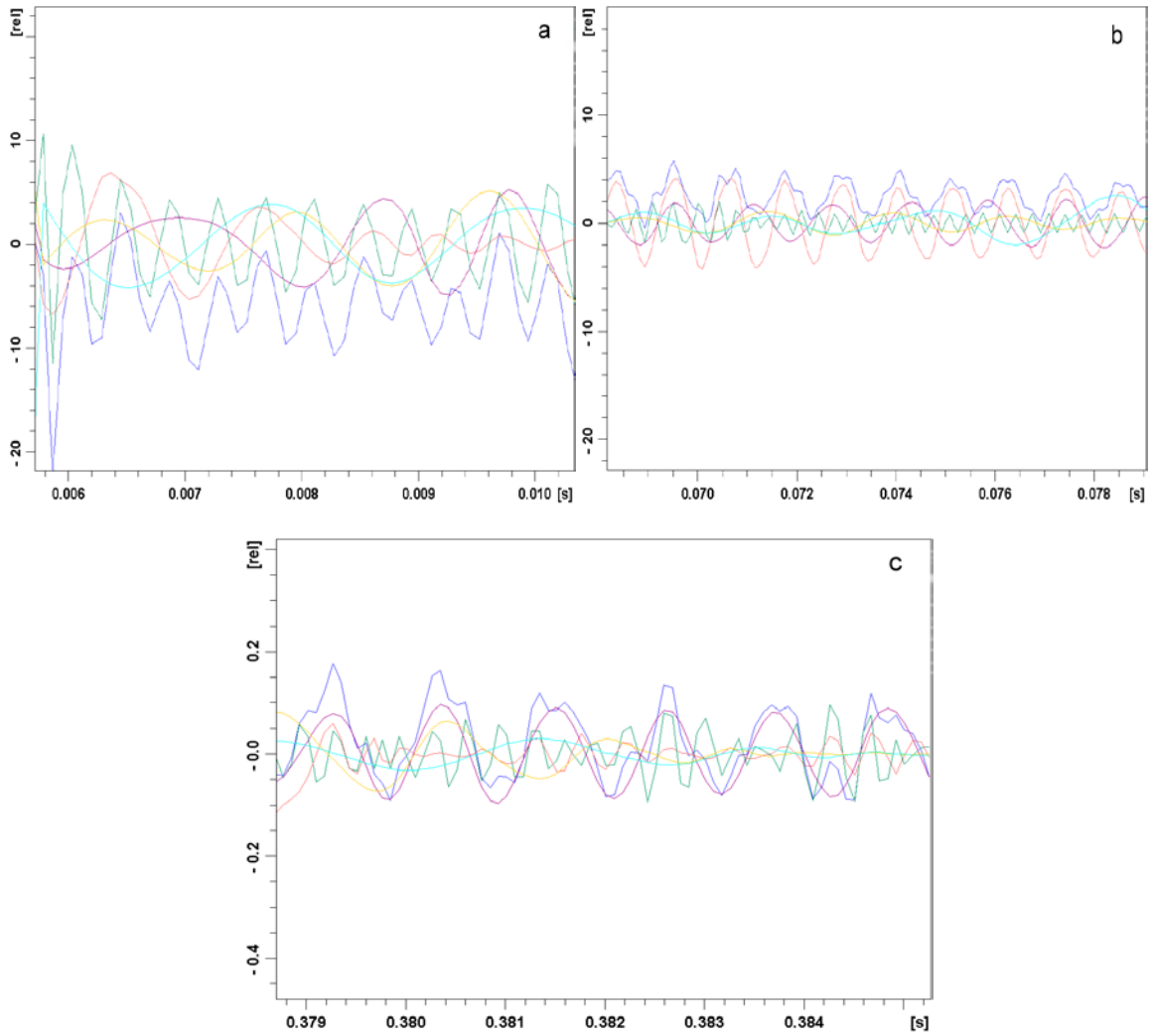
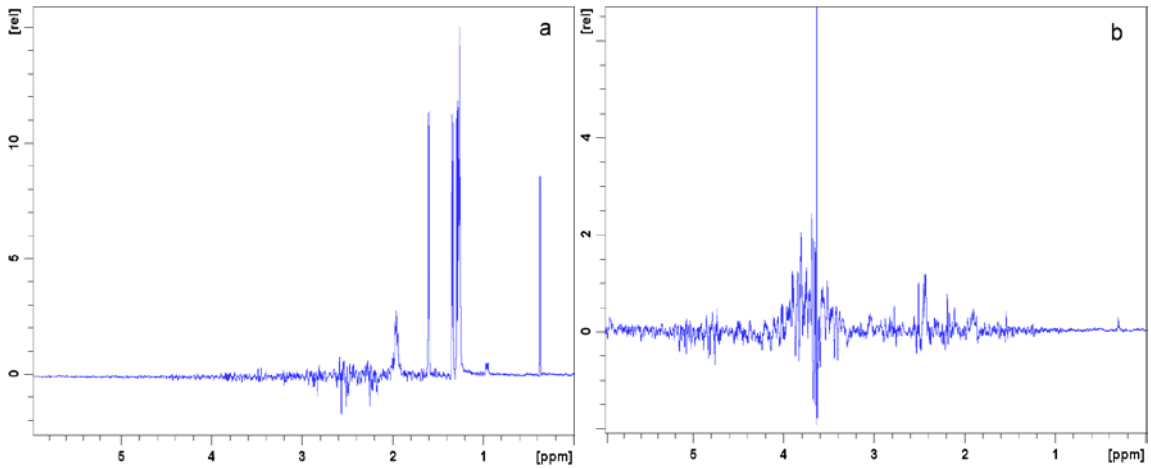


Figure 4.16 Superposition of the first five IMFs with different sections of the original time domain signal: the first (green trace), the second (red), the third (purple), the fourth (yellow) and the fifth (cyan) IMF compared with different parts of the original time domain signal (blue trace). The best

fitting of the initial part of the FID is obtainable using only the first IMF (a); the successive FID time window is well reconstructed by the sum of the first and the second IMFs (b); the successive section of the FID is well fitted summing the first, the second and the third IMFs (c).

A detailed comparison among the first five components and the original time domain signal is depicted in Fig. 4.16. Clearly, the data may be well fitted with only those IMFs. However, a deeper inspection reveals an interesting trend. The early time segment of the FID is well represented by only the first IMF. In the subsequent segment, the first two IMFs need to be summed to fit properly the signal (see part *b* of Fig. 4.16). The third segment needs the sum of the first three IMFs for a perfect fit (see part *c* of Fig. 4.16) and so on. The aim is to show that the contribution of the first IMF is not sufficient to well describe the solute signal, but a fusion with some of the successive IMFs may improve the performance of the suppression.

The Fourier transformed spectra of the IMFs (Fig. 4.15) are shown in Fig. 4.17. In particular, the Fourier transform of the first IMF well represents the low ppm range of the original spectrum (from 0 to 1 ppm almost perfectly), while the FFTs of the second up to the fifth IMF do not provide any recognizable spectral resonances. They contain several signal distortions and if they are summed up they allow an almost perfect reconstruction of the solute spectrum. The successive IMFs (from the sixth to the sixteenth) identify several signals belonging to the solvent artifact. However, they also need to be summed up in order to reconstruct the original water signal.



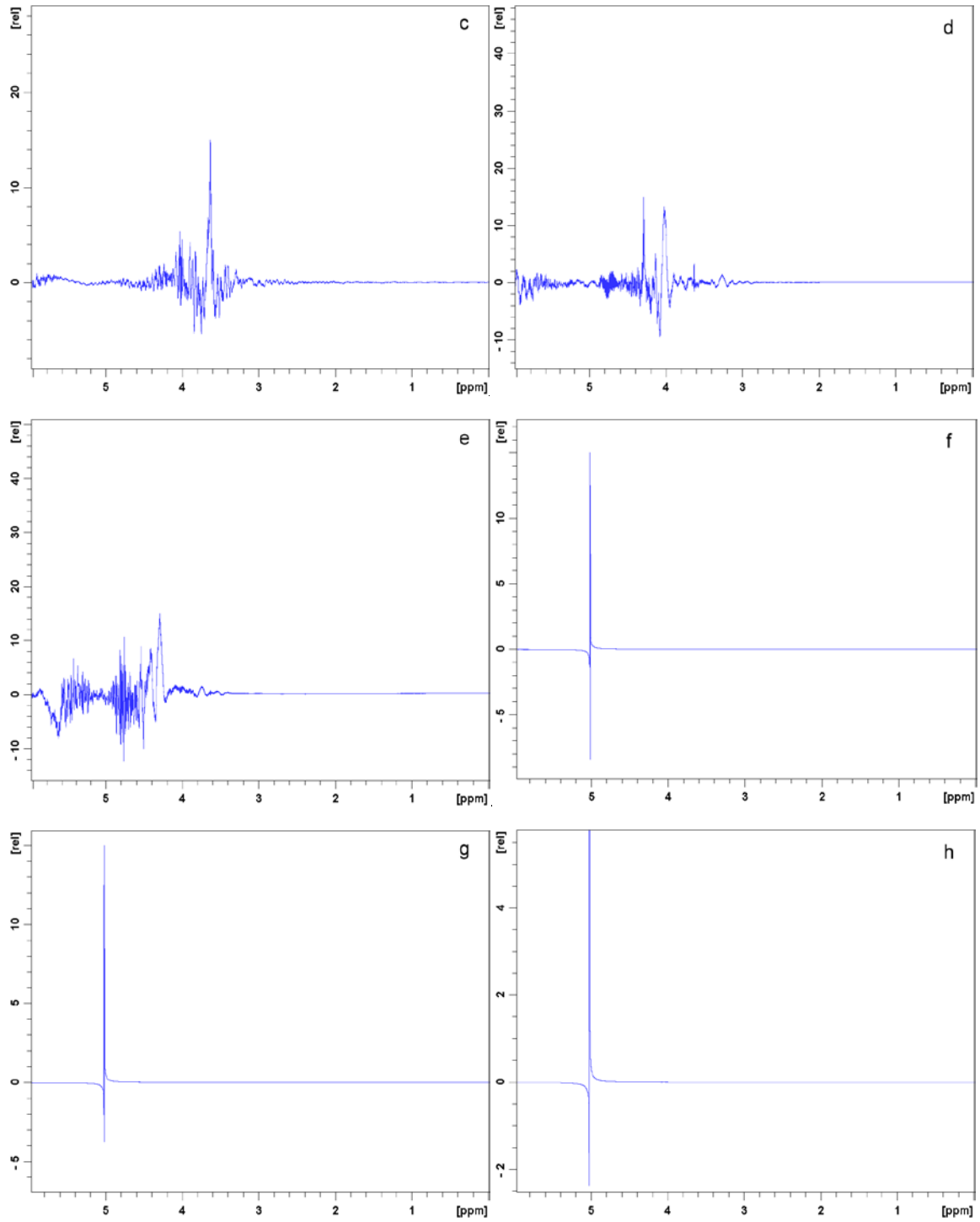


Figure 4.17 Fourier transform of some of the extracted IMFs from the one-dimensional spectrum measured from a sample containing a mixture of five amino acids: it represents the Fourier transform of the first (a), the second (b), the third (c), the fourth (d), the fifth (e), the fifteenth (f) and the sixteenth (g) IMF and of the residual (h).

In Fig. 4.18 the sum of the first five time domain IMFs (highlighted in green) and the sum of the remaining ones (in red) are overlapped with the original time domain signal (blue trace). It is interesting to observe that the fusion of the final IMFs (with a low frequency of oscillation) optimally reflects the general trend of the signal

behavior due to the solvent artifact, whereas the sum of the first five IMFs encompasses the solute time domain signal.

The Fourier transforms of the extracted IMFs can be easily summed up or fused as well in the frequency domain in order to prove the effectiveness of the solute signal reconstruction. Equivalently, the Fourier transforms of the two dataset of fused time domain IMFs (the first dataset including the first to the fifth IMFs and the second dataset from the sixth to the sixteenth IMF) yield the same results. They are described in Fig. 4.19 where in particular, the part *a* shows the overlap (the summation) of the first five components in the frequency domain, while the part *b* identifies the superposition of the remaining IMFs (the second dataset). It is evident that many distortions are still visible in the solvent area after summing up the first five components. Therefore, the fusion of some of the extracted components allows an almost perfect reconstruction of the solute signal but does not furnish a reliable identification of resonances of interest close to the solvent artifact.

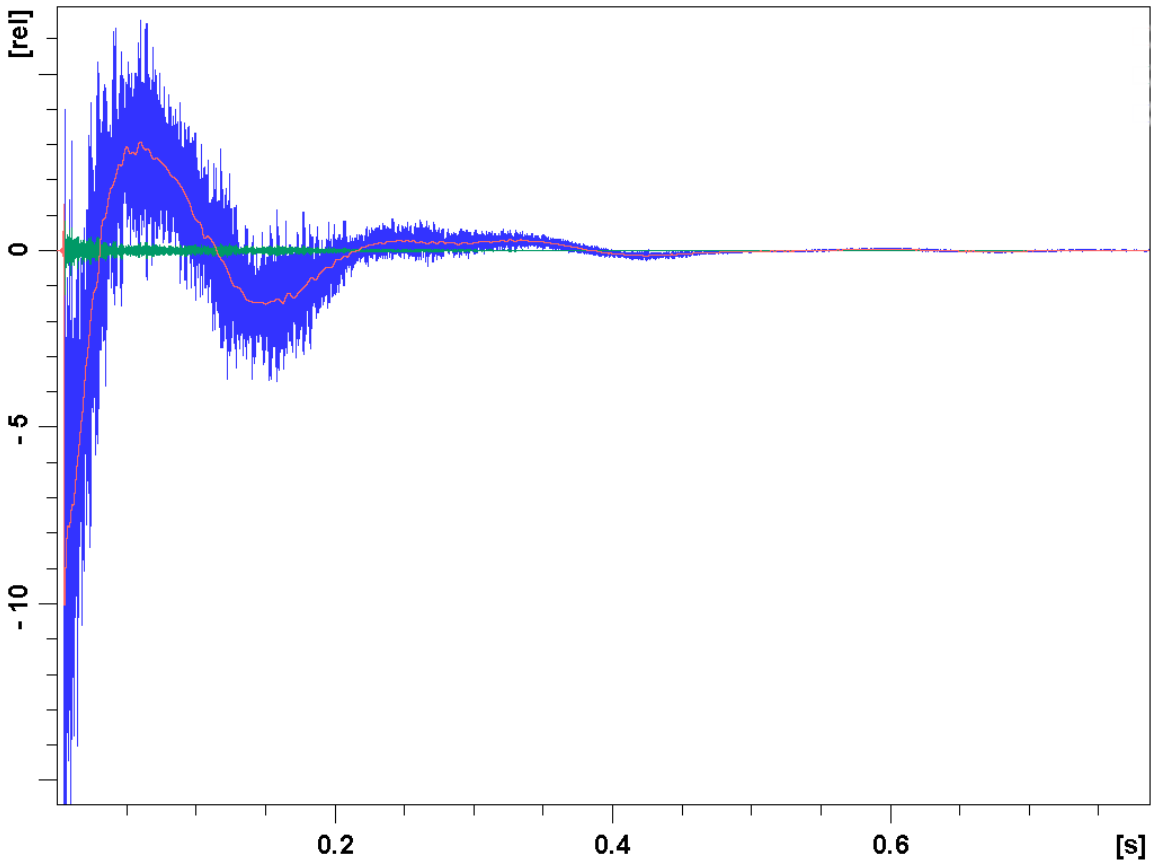


Figure 4.18 IMFs summation in the time domain: the original signal (blue trace) is compared with the fusion of the first five components (green trace) and with the sum of the remaining IMFs, from the sixth to the sixteenth component (red trace).

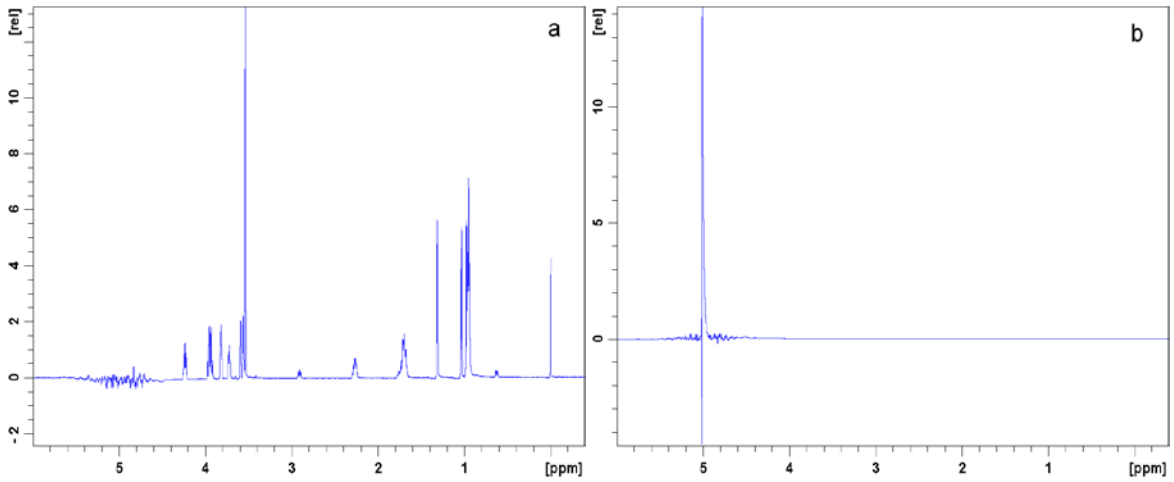


Figure 4.19 Fourier transforms of two datasets of extracted IMFs: the FFT of the summation of the first five components (a); the FFT of the fusion of the latest IMFs, from the sixth to the sixteenth component (b).

The validation of the reconstructed spectrum is performed comparing it with the original one as shown in Fig. 4.20. The spectral features detected around 1 ppm (red box) are zoomed out in Fig. 4.21 (part *a*) as well as the solvent artifact (part *b*). It demonstrates that the reconstructed intensities of the compound do not differ from the original ones and this assertion holds true for all the investigated spectral range. This high reconstruction quality is a fundamental prerequisite for a correct protein structure determination.

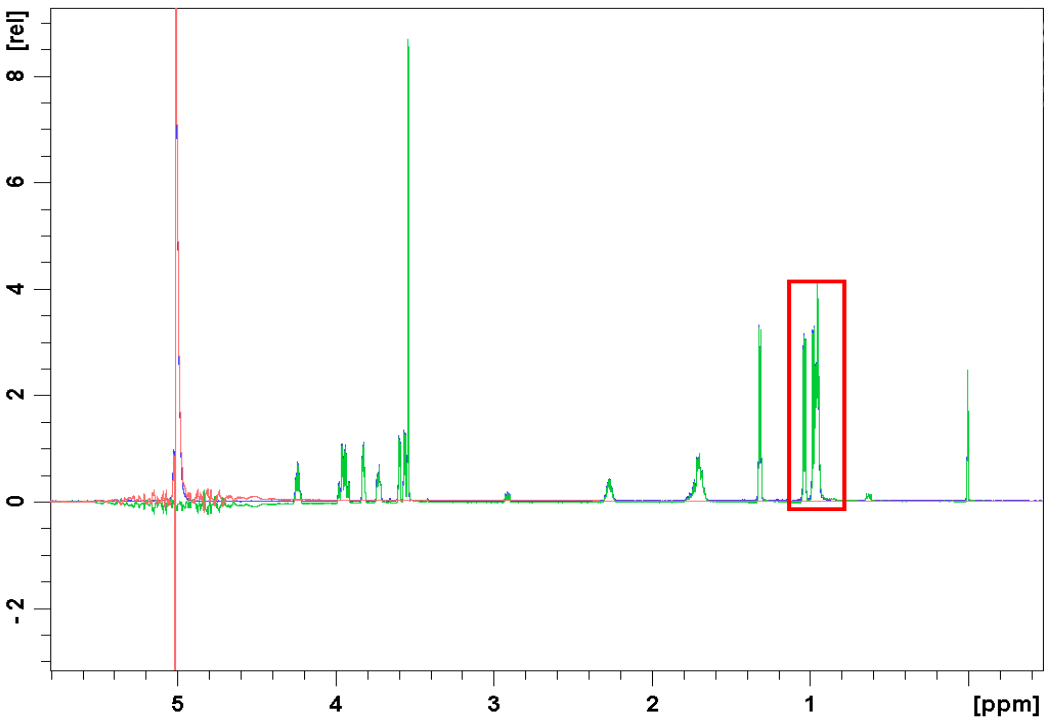


Figure 4.20 Comparison of the Fourier transforms of two datasets of extracted IMFs with the original spectrum: the FFT of the summation of the first five components (green trace); the FFT of the fusion of the last IMFs, from the sixth to the sixteenth component (red trace); the original spectrum (blue trace).

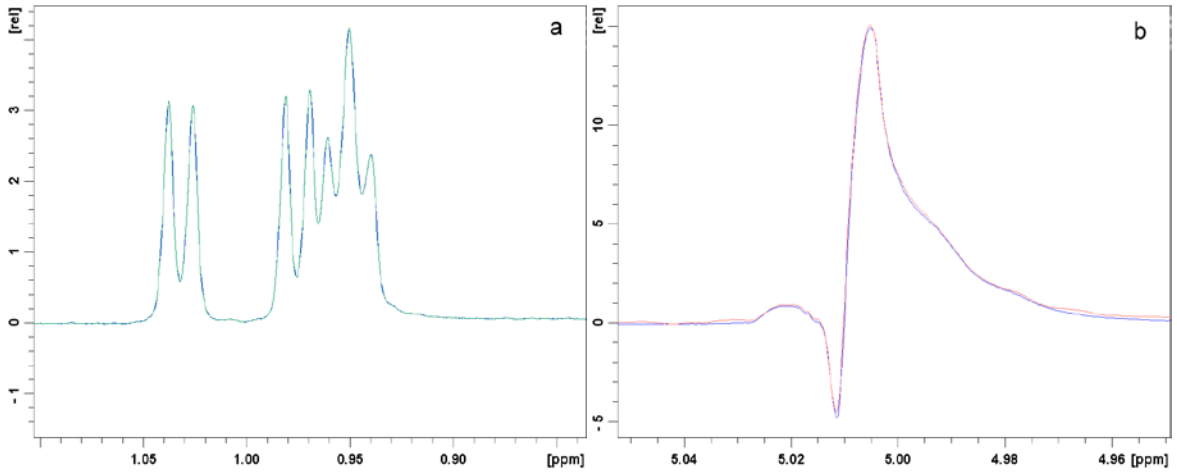


Figure 4.21 Zoom of the red box depicted in Figure 4.20 and of the solvent artifact: (a) the original spectrum (blue trace) compared with reconstructed residue resonances (green trace) obtained fusing the first five IMFs; (b) the original solvent spectrum (blue trace) and the reconstructed solvent signal (red trace) obtained summing up the IMFs from the sixth to the sixteenth component.

In order to evaluate the correctness of the empirically chosen number of fused IMFs used to reconstruct the solute signal, the first five Fourier transformed components have been inspected. In Fig. 4.22 a stepwise superimposition of the first five components on the original spectrum is described and evaluated peak by peak. In particular, it shows that the resonances between 0 to 1 ppm can be reconstructed using only the first IMF, as depicted in part *a* of Fig. 4.23. To correctly rebuild the peaks in the range between 1 and 3 ppm, the first and the second IMFs must be summed up (see part *b* of Fig. 4.23). The sum of the second, the third and the fourth components is necessary to reconstruct the spectrum in the interval between 3.5 and 4 ppm, while the first IMF is no more useful in such range. As illustrated in part *c* of Fig. 4.23 the fusion of the second and the third IMFs in this area also reduces the baseline distortions. The final resonance at 4.25 ppm close to the solvent signal can instead be reconstructed by fusing only the third, the fourth and the fifth IMFs as shown in part *d* of Fig. 4.23.

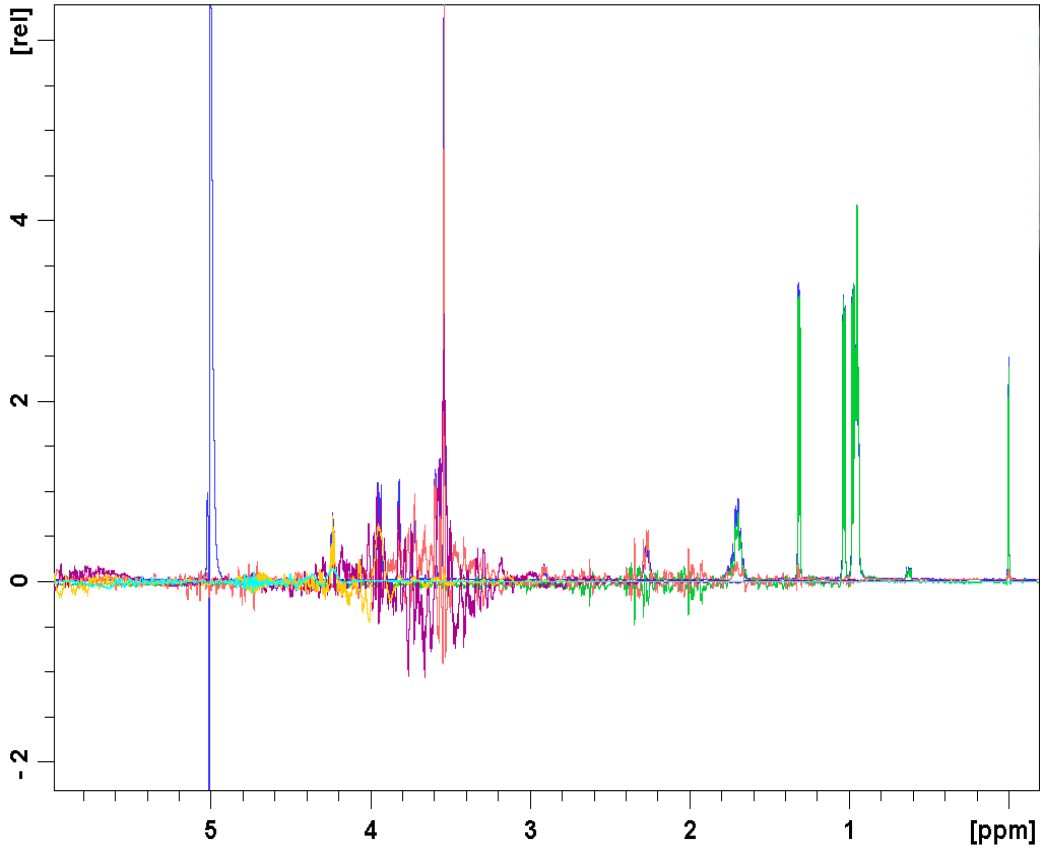
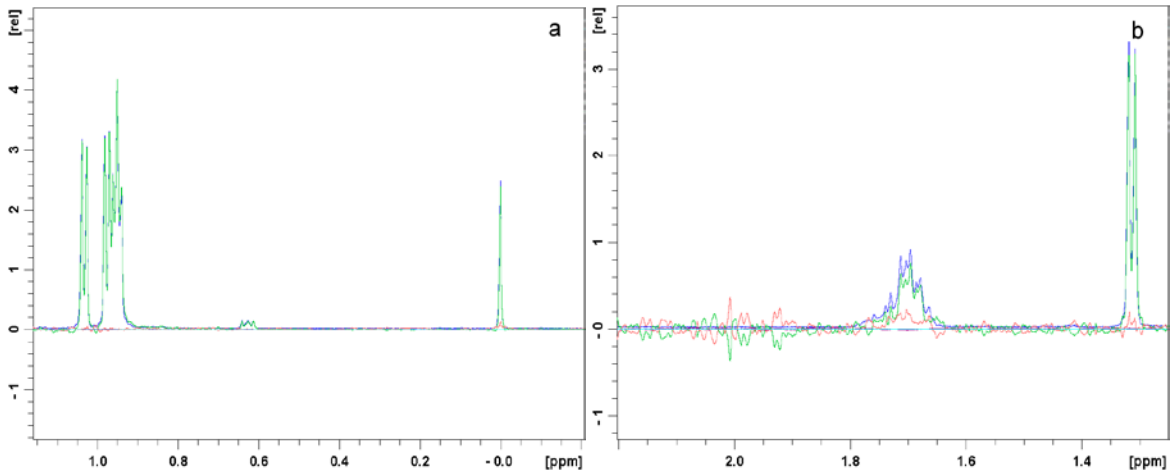


Figure 4.22 Superimposition of the Fourier transform of the first five IMFs with the original spectrum: first (green), second (red), third (purple), fourth (yellow) and fifth (cyan) component compared with the original spectrum (blue trace).



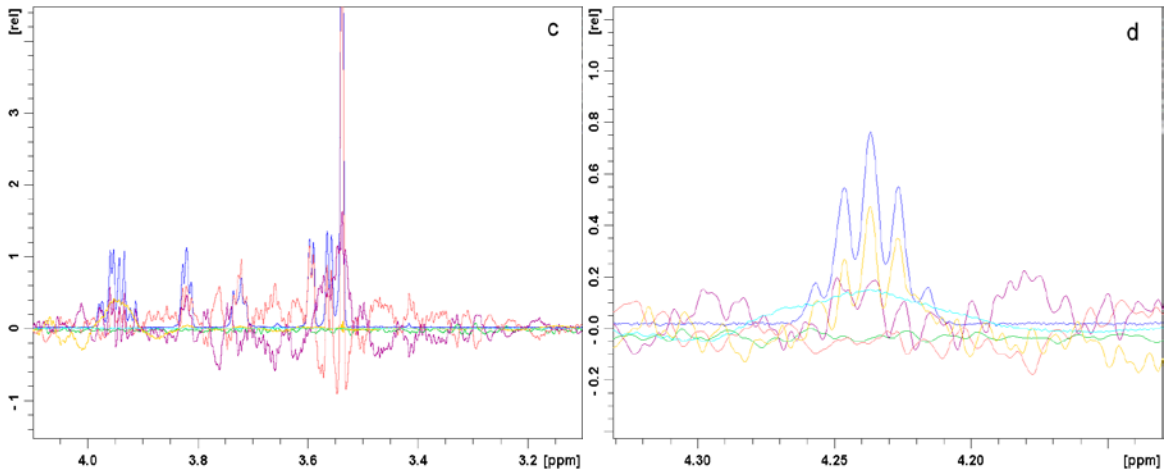


Figure 4.23 Detailed comparison of the first five IMFs with the original signal: first (green), second (red), third (purple), fourth (yellow) and fifth (cyan) component compared with the original spectrum (blue trace). The residue resonances in the interval from 0 to 1 ppm where only the first IMF well fits the original data (a); the zoom of the spectral range between 1 and 3 ppm where the summation of the first and the second IMFs is needed (b); the residue resonances in the interval between 3.5 and 4 ppm where the fusion of the second, the third and the fourth IMFs well fits the original resonances (c); the threonine peak at 4.25 ppm is well reconstructed summing up the third, the fourth and the fifth IMFs.

The EMD has been demonstrated to be able to decompose the FID into several oscillatory modes with physical meanings. An automated inspection and fusion of the IMFs of interest may represent one of the further developments of such method as described in the discussion section in the last chapter.

4.1.3.1 AUTOMATED PHASE CORRECTION BY MEANS OF EMD

Typically the measured time-domain signals need to be phase corrected after Fourier transforming. The following procedure has been developed in order to perform an automated phase correction:

1. EMD of a complex time domain signal (FID) is applied.
2. The IMFs are visually inspected and those ones containing solute resonances are summed up.
3. The sum of the IMFs of interest is Fourier transformed.
4. The baseline regions are automatically identified (see par. 3.1.4).
5. Each non-baseline region is considered as a true peak and a box is built around it. The mean value of the surrounding baseline regions of every peak constitutes a local threshold σ.

6. Ideally all the points in the box should have an intensity value over such threshold in order to determine the optimal individual phase correction (in absorption mode):

$$\max \sum_{i=1}^{K} P_i \qquad (4.1)$$

where P defines the point of a peak in a box (containing K points), whose intensity value is above the local threshold σ.

7. The algorithm maximizes the number of points above each local threshold simultaneously for all the peaks in the spectrum.
8. Genetic algorithms [e.g. Holland, 1975] are used to determine the optimal combination of zero- and first-order phase correction parameters that verify the fitness function described in eq. 4.1 for every peak. A population of two-thousand individuals is randomly generated. Each individual is made up of two random genes, where the first one can have any value between -180 and 180 (representing the zero-order PHC0 value) and the second one has a value range between -500 and 500 (identifying the first-order PHC1 parameter). The individuals are sorted in a decreasing order depending on their fitness value. Each consecutive couple of individuals produces two children applying the one-point crossover (mixing their two genes). The 5% of the generated children in the entire population are randomly mutated in both genes. The fitness function is then newly computed and the doubled population is reduced by an half in accordance to this value. The reproduction procedure is repeated for one-hundred generations. The first individual in the final population (that one with the highest fitness value) contains the optimal phase correction values in its genes [De Sanctis, 2006].

The phase correction algorithm cannot be applied on the entire original spectrum containing the solvent signal, since this latter would compromises the performance of the correction. It is also not possible to apply the algorithm to each IMF separately, because as shown in Fig. 4.17 only the first IMF is interpretable, while the others do not allow even a visual recognition of the resonances of interest. The limitations of this algorithm are discussed in the last chapter.

The EMD has been initially performed on the one-dimensional time domain signal acquired from a sample containing a mixture of five amino acids (par. 2.1.2.6.1). The Fourier transform of this signal (without any phase correction) is reported in Fig. 4.24 (part *a*). The group delay has been excluded before applying the decomposition. The routine has generated sixteen different IMFs. The necessary zero and first order phase correction parameters were obtainable from the processing parameter files (PHCO = -93.40 and PHC1 = 0.00). They have been used in order to validate the PHC0

and PHC1 values, automatically calculated by the algorithm. The first five IMFs have been summed up. The baseline regions have been identified in the frequency domain. The genetic algorithms have produced a final optimal individual with the following genes: -95.80 (PHC0) and 3.20 (PHC1). The application of such values for correcting the phase in the original spectrum yields remarkable results as described in part *b* of Fig. 4.24.

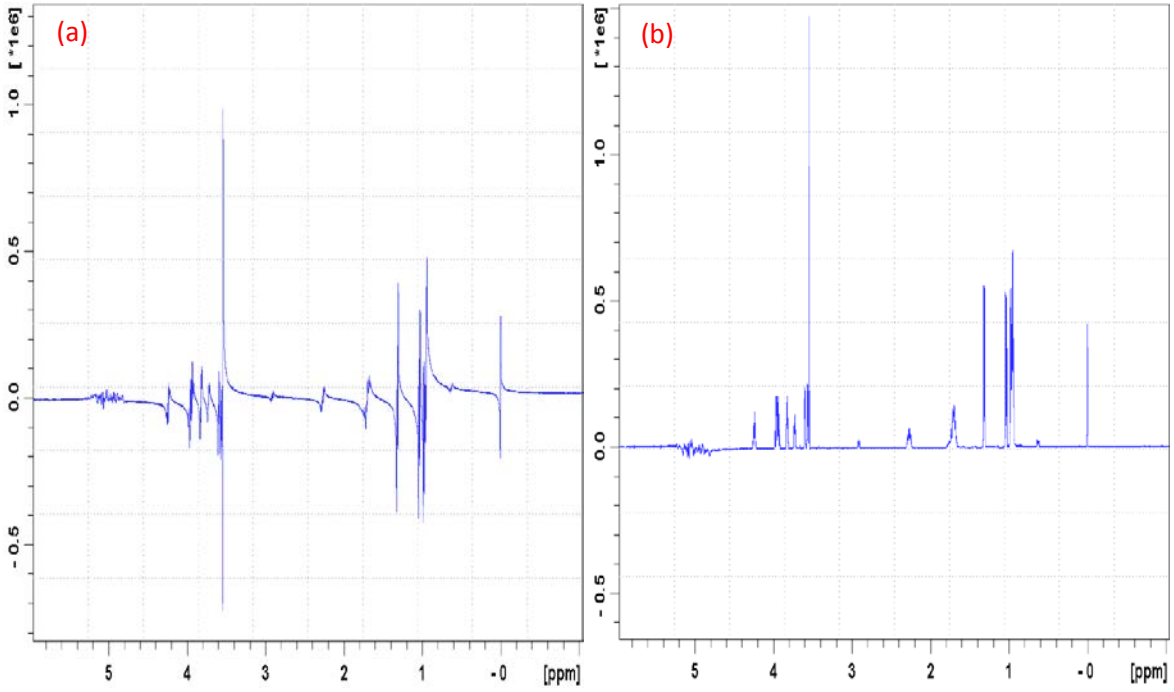


Figure 4.24 Automated phase correction of the one-dimensional spectrum obtained from a sample containing a mixture of five amino acids (par. 2.1.2.6.1): spectrum before (a) and after (b) automated phase correction.

The same procedure has been applied starting from the one-dimensional time domain signal acquired from a sample containing a mixture of twenty amino acids (par. 2.1.2.6.2). The Fourier transform of this signal (with an intentional very large phase distortion of PHC0 = 44.00 and PHC1 = -246.00 obtained manually with Topspin software) is reported in Fig. 4.25 (low trace). The routine has generated sixteen different IMFs. The first six IMFs have been summed up and their Fourier transform is shown in the upper trace of Fig. 4.25. The baseline regions of this sum have been identified in the frequency domain. The genetic algorithms have produced a final optimal individual having the following genes: -57.50 (PHC0) and -246.80 (PHC1). The spectrum resulting from the automated phase correction is described in Fig. 4.26 (blue trace) and it is compared with the spectrum obtainable with a manual phase correction (using Topspin software). The better performance of the automated method is particularly evident in Fig. 4.27 where several resonances of interest are zoomed out.

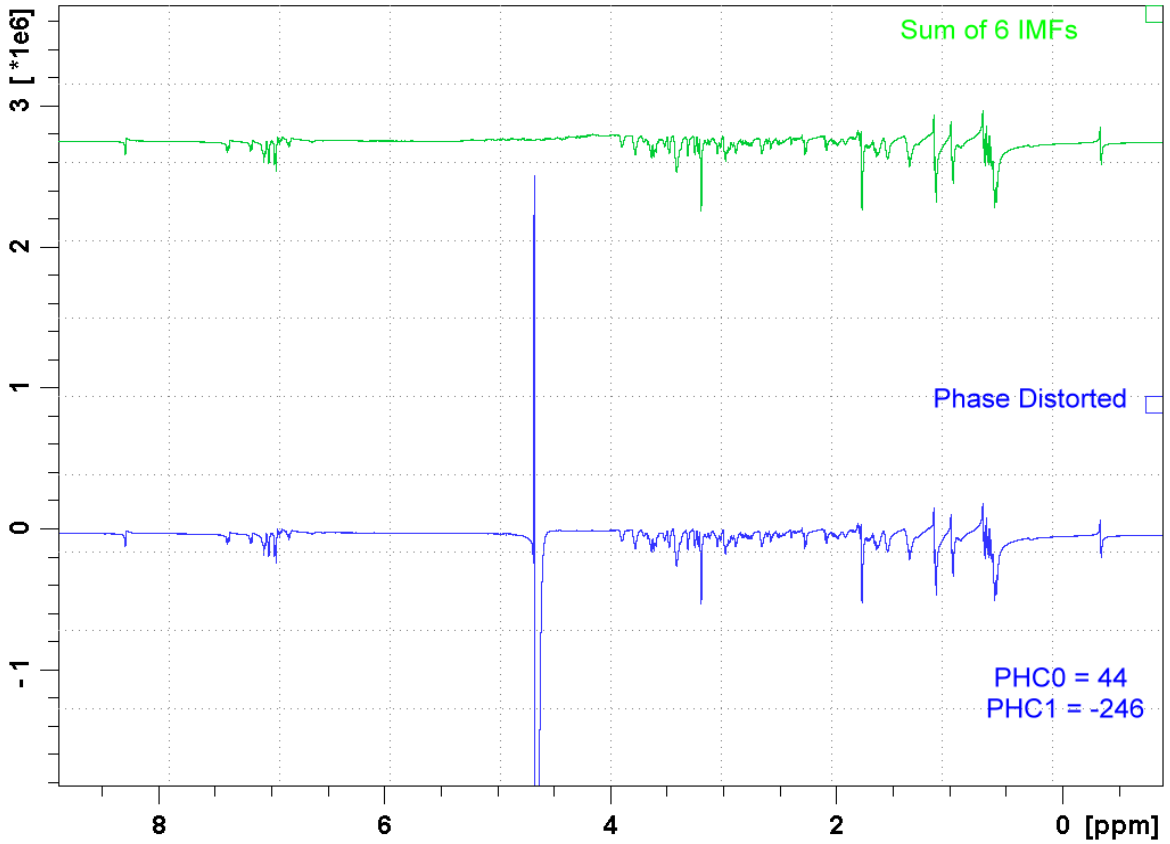


Figure 4.25 Intentional phase distortion of a one-dimensional spectrum obtained from a sample containing a mixture of twenty amino acids (par. 2.1.2.6.2): the entire intentionally phase distorted spectrum (lower trace) manually obtained modifying the PHC0 and PHC1 parameters in the Topspin software (with PHC0 = 44 and PHC1 = -246); the spectrum containing the sum of the first six IMFs obtained from the inverse Fourier transform of the intentionally phase distorted spectrum (upper trace).

Several aspects of such application (automated phase correction) may be further investigated. In particular, the performance of this method is strictly related to the identification of the IMFs of interest, to the automated recognition of baseline points and to the initial parameters defined by the genetic algorithms (as the population size, the range of the gene values, the selection of the couples, the mutation type, the mutation percentage and the number of generations). The weak and strong points of these aspects are evaluated in the discussion section in the last chapter.

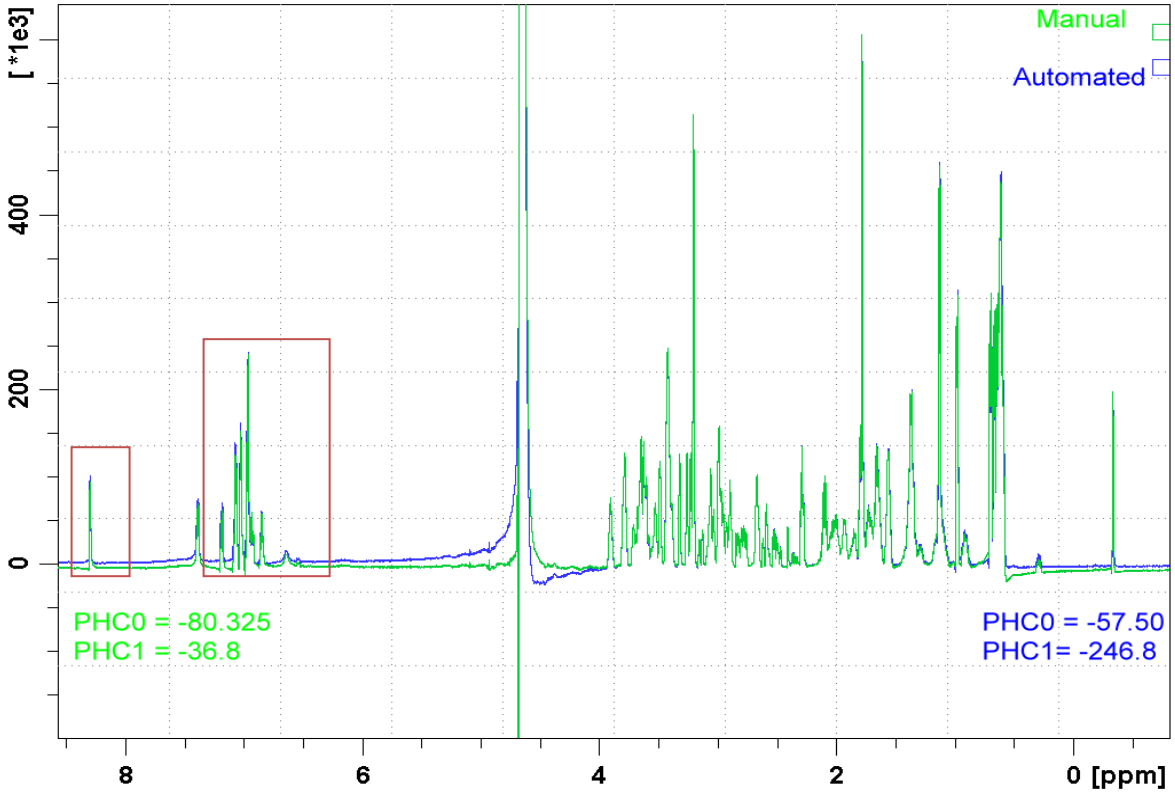


Figure 4.26 Automated phase correction of the one-dimensional spectrum obtained from a sample containing a mixture of twenty amino acids: comparison of the automated (blue trace) phase correction (PHC0 = -57.50 and PHC1 = -246.80) with the manual phase correction (green trace) obtainable with Topspin software (PHC0 = -80.32 and PHC1 = -36.80).

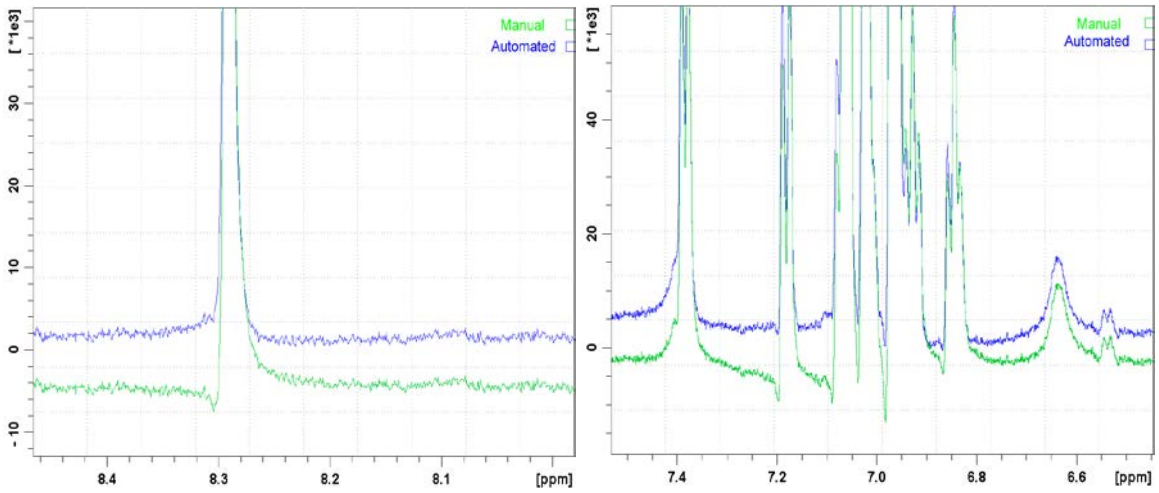


Figure 4.27 Zoom of the boxes reported in Figure 4.26: detailed comparison of the automated (blue trace) phase correction (PHC0 = -57.50 and PHC1 = -246.80) with the best manual phase correction (green trace) obtainable with Topspin software (PHC0 = -80.32 and PHC1 = -36.80).

5 Protein structure determination

5.1 PSCD4-DOMAIN OF PLEURALIN PROTEIN

5.1.1 SPECTRAL ASSIGNMENT OF CHEMICAL SHIFTS

The spectra used for the sequential assignment of the atoms in the backbone and in the side chains have been described in par. 2.1.3. The assigned $^{1}H^{15}N$-NOESY-HSQC spectrum of the recombinant His_6PSCD4-domain of the pleuralin protein is reported in Fig. 5.1. It is evident that the resonances are notably superimposed in the central part of the spectrum, namely between 115 to 120 ppm and from 8.2 to 8.7 ppm in the ^{15}N and ^{1}H dimension respectively. Some residues in the C-terminal reveal split signals (as the Ala 106) due to the flexibility of this terminal that can possess different conformations. The list of the newly detected chemical shifts and the primary sequence are reported in Appendix A.

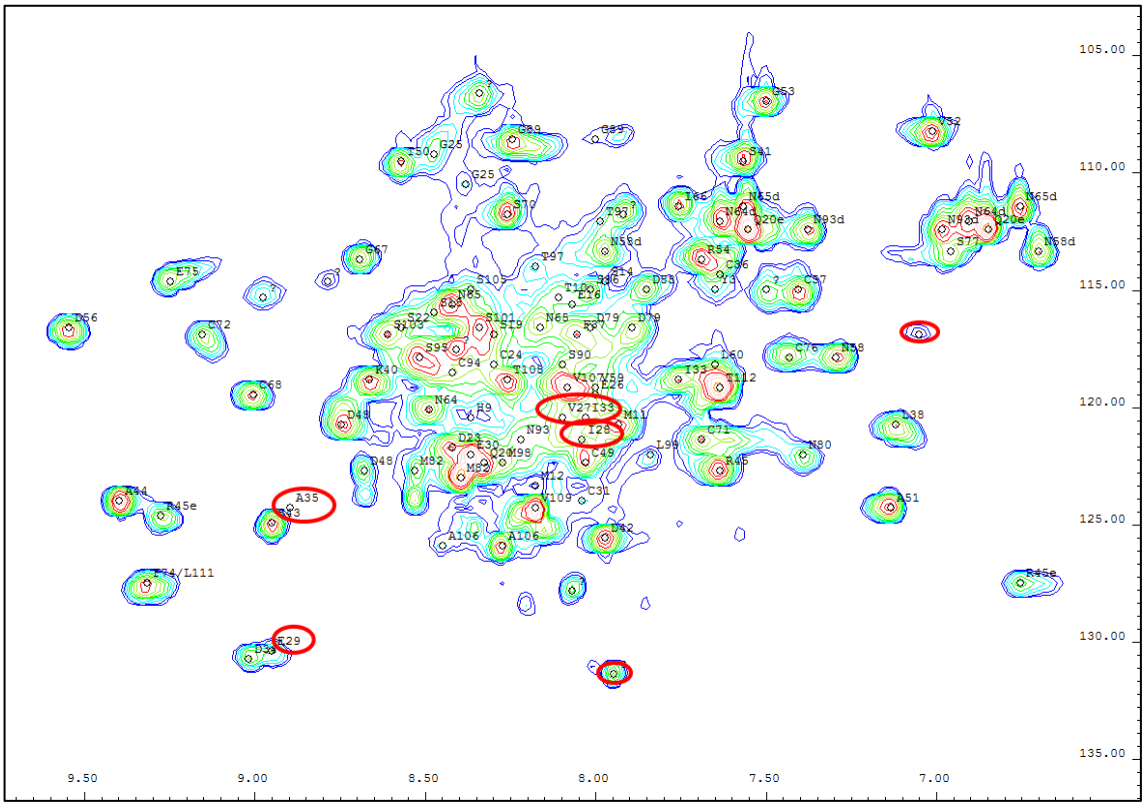


Figure 5.1 Assigned $^{1}H^{15}N$-NOESY-HSQC spectrum of the PSCD4-domain of the pleuralin protein: the F2-F3 projection of the three-dimensional spectrum. The assignment of the signals is described by

the one-letter code of the amino acids and by their position in the primary sequence. The small letters d and e define the amide signals of the side chains of glutamine and arginine respectively. The main differences with respect to a previous work [Wenzler et al, 2001] are highlighted in red: Val27, Ile28, Glu29 and Ala35.

The following residues could not be assigned in any of the measured spectra: serine 1, histidine 4, histidine 6, histidine 7, proline 61, threonine 84 and proline 91. Only the last three amino acids, among them, are part of the PSCD4-domain. The glutamine acid 29, the alanine 35 and the valine 27 have been instead newly assigned. The chemical shifts have been analyzed by the TALOS+ software in order to predict the presence of canonical secondary structures. In particular an α-helix has been predicted to be formed from the residue Cys 49 to the Val 52. Three beta sheets should be located instead, from the residue Cys 71 to the Phe 74. The TALOS+ software generates a sequence window (Fig. 5.2) where the residues are highlighted with different colors depending on the classification: green represents a good shift classification, yellow an ambiguous one, blue a dynamic part, while blank defines a not possible interpretation. The sequence window obtained from the observed chemical shifts (reported in Appendix A) is compared with that one generated with the previously observed shifts [Wenzler, 2003], as shown in Fig. 5.2. The residues Ser 19, Gln 20, Pro 21, Pro 96 and Thr 97 are in both cases identified as dynamical parts. Using the newly observed shifts the residues Ser 22, Asp 23, Glu 26, Glu 29, Asp 34, Arg 54, Asn 58, Asn 65, Ser 86, Cys 94 are no more ambiguously classified.

Figure 5.2 Sequence window of TALOS+ software for predicting structural motives using the existing and the newly observed chemical shifts: the residues are differently colored dependently on the classification. Green, good; yellow, ambiguous; blue, dynamical; blank, not classified. Sequence windows obtained with the observed (Appendix A) and with the existing [Wenzler, 2003] chemical shifts in the right and in the left side respectively.

5.1.2 EXPERIMENTAL RESTRAINTS

The positions of five disulfide bonds have been determined by *Wenzler* in 2003: from Cys 24 to Cys 94, from Cys 31 to Cys 76, from Cys 36 to Cys 72, from Cys 49 to Cys 71 and between Cys 57 and Cys 68, as described in Fig. 5.3. All of them have been determined by biochemical methods.

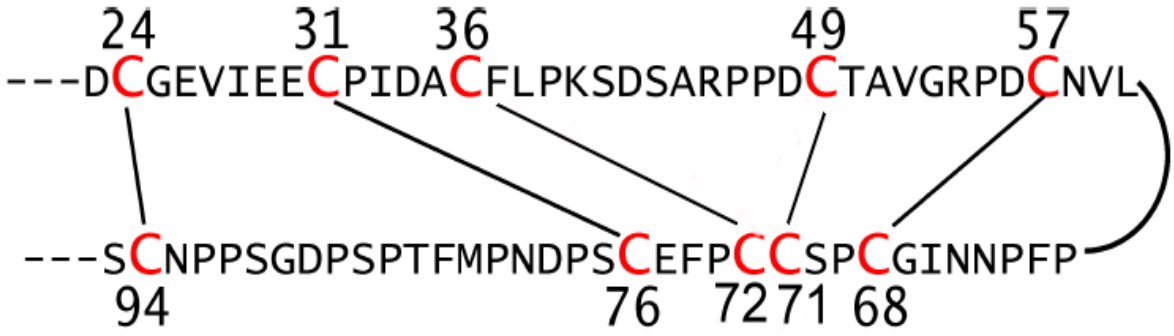


Figure 5.3 Disulfide bonds: connections of the ten cysteine residues of the PSCD4-domain.

5.1.2.1 THREE-BOND SCALAR COUPLING RESTRAINTS

The TALOS+ program yields a list of dihedral angle restraints. The Karplus relation has been exploited in order to derive 38 $^{3}J_{H}{}^{N}{}_{H}{}^{\alpha}$ coupling constants from the obtained φ torsion angles. The A, B and C constants of the Karplus equation (see eq.1.17) have been set to 7.13, -1.31 and 1.56 respectively [Habeck et al, 2005]. From the HNCA-E.COSY 27 $^{3}J_{H}{}^{N}{}_{H}{}^{\alpha}$ coupling restraints have been observed (reported in Appendix B) and 8 of them reveal different values with respect to the previous work of *Wenzler* (gray shaded in Appendix B). They have been added or substituted to those ones predicted by the TALOS+ software, as reported in Appendix C (obtaining 65 $^{3}J_{H}{}^{N}{}_{H}{}^{\alpha}$ coupling restraints totally).

The routine used to compute the Karplus equation has been introduced in the AUREMOL software. In particular, it can be started from the "Calculation" menu as shown in Fig. 5.4. The main interface (reported in Fig. 5.5) is launched as the "Karplus calculation" submenu is selected. The user must provide the input file containing either the dihedral angle or the ^{3}J coupling restraints, whose formats are displayed in Fig. 5.6 part *a* and part *b* respectively. He must, accordingly to the provided restraints, select the direction of conversion. In particular, if the user furnishes the dihedral angle restraint file and he wishes to obtain ^{3}J coupling restraints from it, accordingly

to the Karplus curve [Karplus, 1963] a unique value would be obtained for every restraint. If the dihedral angle restraints must be derived from the ^{3}J coupling restraints then up to four different values would be produced for each restraint. Depending on the bonds involved in the scalar coupling different values of the A, B and C constants of the Karplus equation have been empirically obtained [Habeck et al, 2005]. Therefore, the user must select the atoms involved in the restraints among the following ones: $H^N H^\alpha$, $H^N C^\alpha_{i-1}$, $H^N C^\beta$ and $C^\alpha_{i-1} C^\alpha_i$.

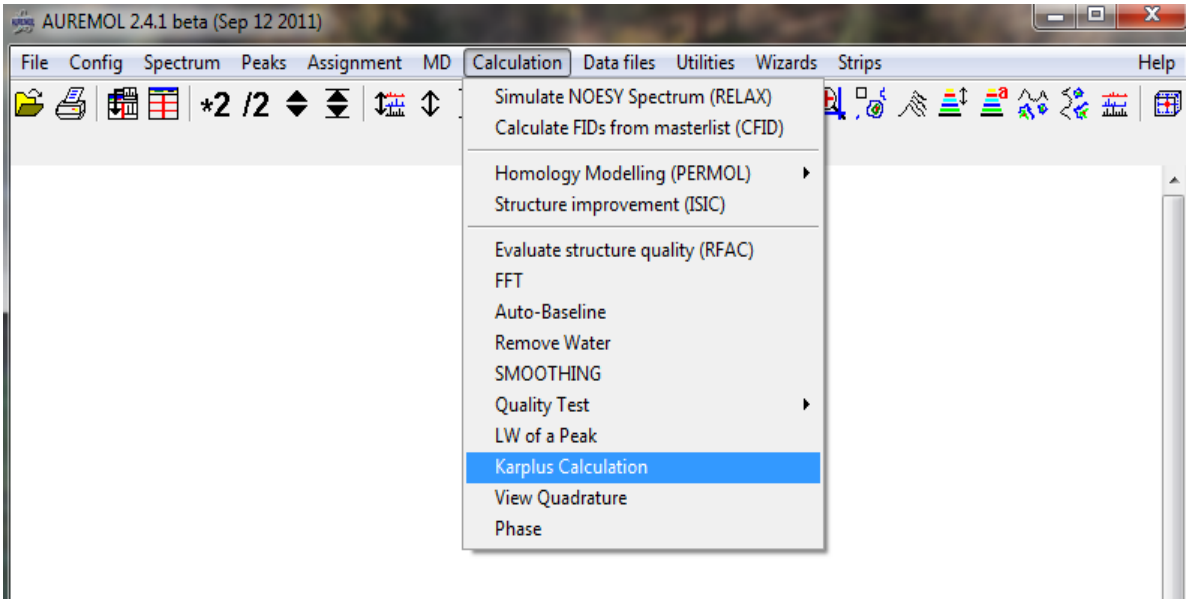


Figure 5.4 Starting the Karplus routine: this module is in the "Karplus calculation" submenu of the "Calculation" menu in the AUREMOL software package.

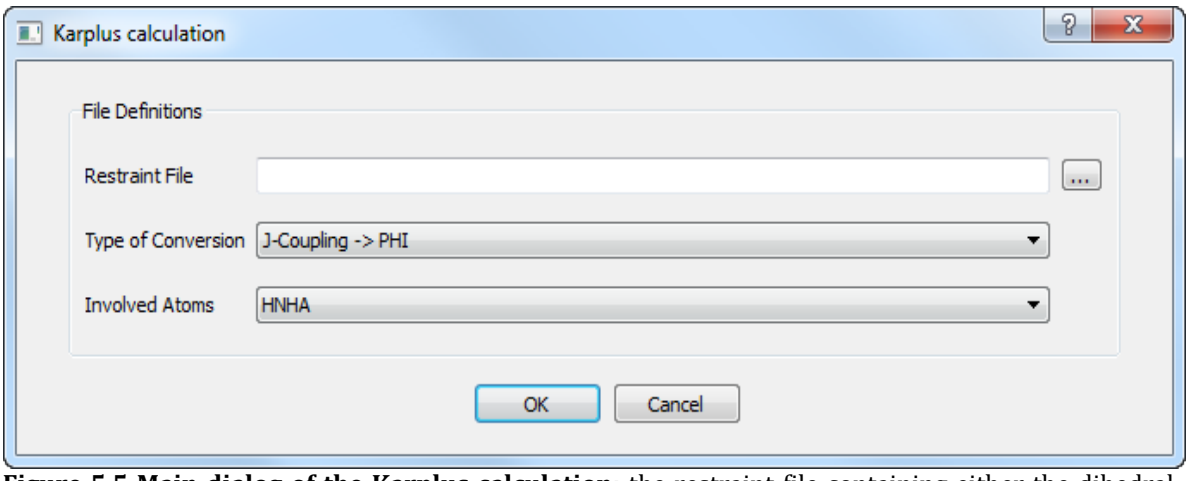


Figure 5.5 Main dialog of the Karplus calculation: the restraint file containing either the dihedral angles φ or the ^{3}J coupling restraints must be provided with the formats reported in Fig. 5.6 part *a* and part *b* respectively. The type of the conversion (from ^{3}J coupling to φ torsion angle or vice versa) must be selected accordingly to the provided restraint file. The atoms involved in the observed scalar coupling must be chosen.

(a)
```
assign (resid   12 and name C    ) (resid   13 and name N    )
 (resid   13 and name CA   ) (resid   13 and name C    )    -66.5

assign (resid   14 and name C    ) (resid   15 and name N    )
  (resid   15 and name CA   ) (resid   15 and name C    )    -70.3
```

(b)
```
assign (resid  10 and  name C    ) (resid  11 and  name  N   )
(resid  11 and  name CA    ) (resid  11 and  name  C)        8.5

assign (resid  18 and  name C    ) (resid  19 and  name  N   )
(resid  19 and  name CA    ) (resid  19 and  name  C    )    4.7
```

(c)
```
assign (resid   12 and name C    ) (resid   13 and name N    )
  (resid   13 and name CA   ) (resid   13 and name C    )      4.9

  assign (resid   14 and name C    ) (resid   15 and name N    )
   (resid   15 and name CA   ) (resid   15 and name C    )     5.4
```

(d)
```
assign (resid  10 and  name C    ) (resid  11 and  name  N   )
(resid  11 and  name CA    ) (resid  11 and  name  C    )

-146.0 -94.0

assign (resid  18 and  name C    ) (resid  19 and  name  N   )
(resid  19 and  name CA    ) (resid  19 and  name  C    )

100.4 -174.7 -65.3 19.6
```

Figure 5.6 Karplus file formats: the user provides either the torsion angle φ (a) or the ^{3}J coupling (b) restraint file obtaining the corresponding output files (c and d respectively).

In Fig. 5.6 is reported an example of the input file formats that must be provided by the user (part *a* and part *b*) and the corresponding output files (part *c* and part *d*).

5.1.2.2 HYDROGEN BONDS RESTRAINTS

The hydrogen bonds have been detected from the long-range HNCO experiment (15 restraints as described in Appendix D). The distance between the proton (H) and the acceptor (O) has been defined from 0.18 to 0.25 nm, while it varies from 0.23 to 0.35 nm between the donor (N) and the acceptor (O) accordingly to Fig. 1.9.

5.1.2.3 RDC RESTRAINTS

The isotropic and the anisotropic coupled $^{1}H^{15}N$ HSQC experiments have been used to identify 50 $^{1}H^{15}N$ RDCs (as reported in Appendix E). In order to use such restraints to calculate the three-dimensional structure of the protein, the three components A_{ij} of the molecular magnitude tensor must be determined. The smallest and the biggest measured RDC represent the orientation of the atom connecting vector parallel to the y- and to the z-axis of the tensor respectively. The most frequently occurring RDC value identifies instead the orientation parallel to the x-axis. In Fig. 5.7 is reported the histogram of the distribution of the observed RDCs (from residue Glu 29 to Asp 79) where $A_{yy} = -20$, $A_{zz} = 18$ and $A_{xx} = 2$. The rhombicity R (see eq. 1.20) and the axiality A_a (see eq. 1.21) parameters have been computed and introduced in the CNS program as $\frac{a_2}{a_1}$ = 0.814 and a_1=9 respectively. Generally, the A_{yy} and A_{zz} are underestimated thus the latter has been computed accordingly to eq. 1.22 [Wenzler, 2003].

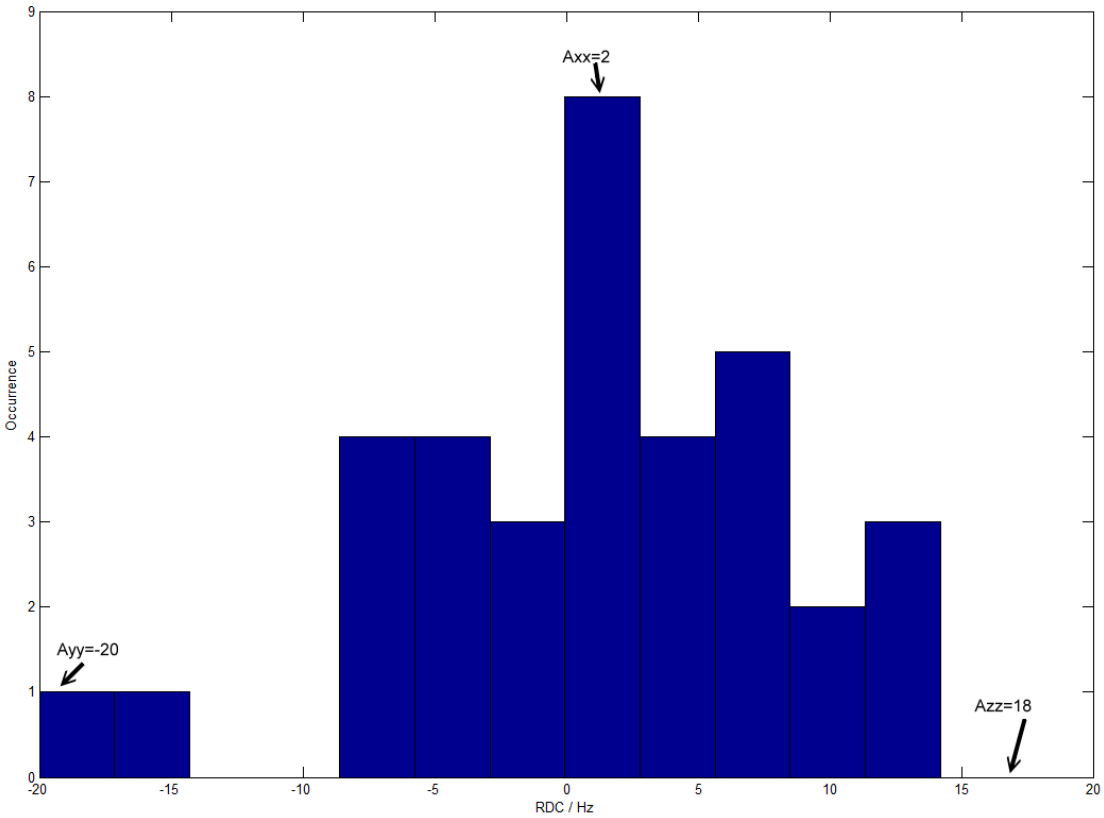
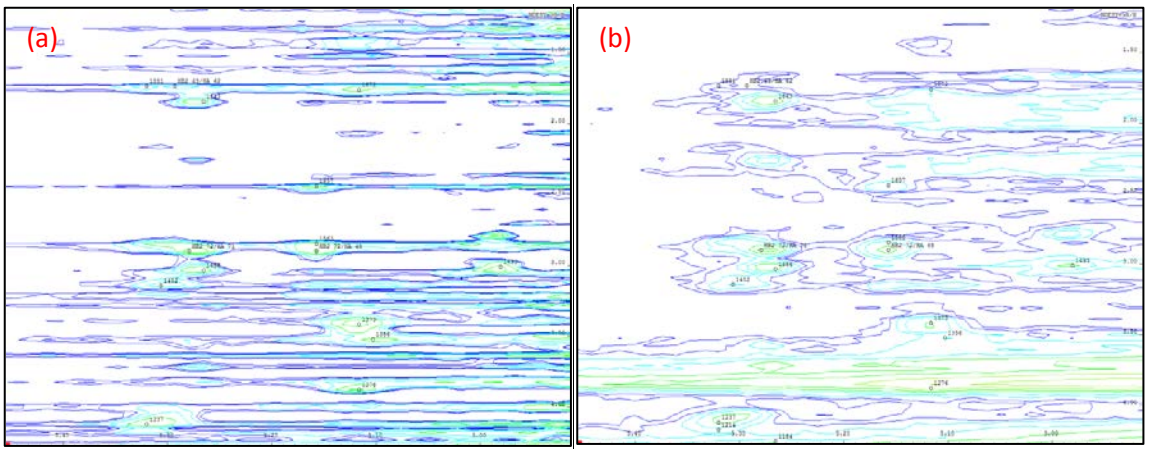


Figure 5.7 Histogram of the observed RDCs (35 restraints from residue Glu29 to Asp79): the three components A_{ij} of the molecular magnitude tensor correspond to the extrema (A_{yy} is the minimum and A_{zz} is the maximum) and to the most frequently occurring RDC value (A_{xx}). The rhombicity R and the axiality A_a are obtained from these components (eq. 1.20 and eq. 1.21).

5.1.2.4 NOE DISTANCE RESTRAINTS

The distance restraints have been obtained applying the REFINE routine [Trenner, 2006] on the previously assigned (by means of KNOWNOE [Gronwald et al, 2002]) two-dimensional NOESY spectrum (par. 2.1.2.2.3). The KNOWNOE routine has been applied on the previously existing pdb file of *Wenzler* imposing 5 iterations for performing the assignment. The REFINE routine applies the default values on the distance error. The SSA for solvent suppression and the ALS for baseline correction have been applied on the spectrum (see Fig. 3.30) before performing the automated assignment in order to reveal some resonances lying under the strong solvent artifact (Fig. 5.8). However, the large amount of certain amino acids as proline, serine, cysteine and aspartic acid in the primary sequence and the lack of aromatic ones as phenylalanine lead to many signal overlaps and to multiple assignments of the chemical shifts (that have been visually inspected). Therefore, the three-dimensional $^{1}H^{15}N$- and $^{1}H^{13}C$-NOESY-HSQC experiments have been additionally used for such purpose. Totally, 590 signals have been assigned in the spectrum and 926 NOE distance restraints have been obtained with REFINE.



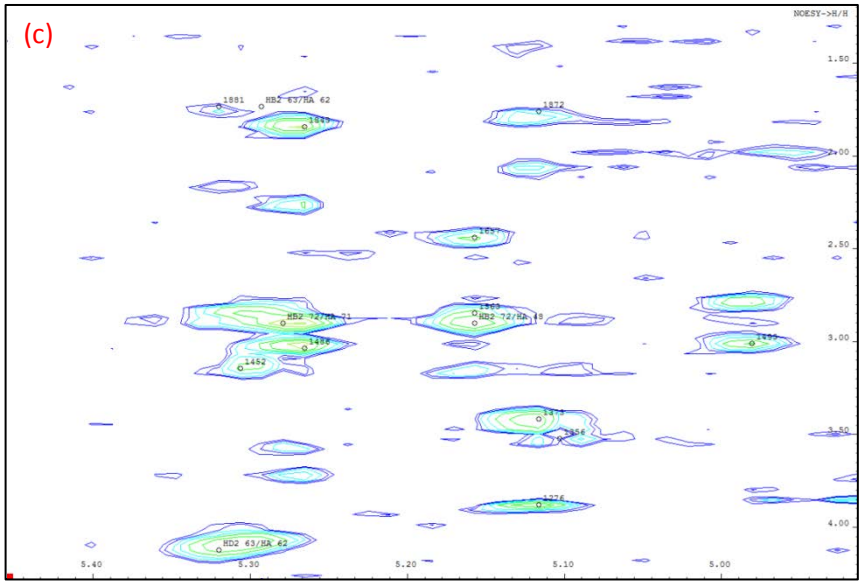


Figure 5.8 SSA and ALS on the two-dimensional NOESY spectrum of the PSCD4-domain: zoom in the range between 5.4 and 4.9 ppm in $\omega_2$and between 4.1 and 1.5 ppm in ω_1. The original spectrum (a), the spectrum after SSA (b) and after ALS in cascade with SSA (c).

In Fig. 5.8 is reported a detailed comparison among the original acquired NOESY spectrum (part *a*), the same spectrum after suppressing the solvent by means of SSA (part *b*) and the same spectrum after SSA and ALS for baseline correction (part *c*). The developed algorithm allows a better identification of the resonances of interest.

5.1.3 STRUCTURE DETERMINATION

The standard simulated annealing [Kirkpatrick et al, 1983] protocol of CNS 1.21 has been used in order to calculate the structure of the PSCD4-domain.

Five different calculations have been performed in the following manner:

1) Using only the previously existing restraint files of *Wenzler* [Wenzler, 2003].
2) Modifying the upper and lower bounds of the NOE distances of *Wenzler* [Wenzler, 2003] accordingly to *Kalbitzer and Hengstenberg, 1992*.
3) Substituting the $^{3}J_{H^N H^\alpha}$ coupling restraints of *Wenzler* with those ones obtained from TALOS+.
4) Using the above described restraints (without those ones of Wenzler) and merging (as described in par. 5.1.2.1) the detected $^{3}J_{H^N H^\alpha}$ coupling restraints with those ones obtained from TALOS+. In particular, five disulfide bonds, 65 $^{3}J_{H^N H^\alpha}$ coupling, 35 RDCs (from Glu 29 to Asp 79), 15 hydrogen bonds and 926 NOE distance restraints have been used.

5) Using all the above (without those ones of Wenzler) described restraints without considering those ones obtained from TALOS+ (see par. 5.1.2.1): five disulfide bonds, 27 $^{3}J_{H}{}^{N}{}_{H}{}^{\alpha}$ coupling, 35 RDCs (from Glu 29 to Asp 79), 15 hydrogen bonds and 926 NOE distance restraints.

In the first three cases, the molecular dynamics simulation has been performed from residue Pro15 to Ser103, while in the last two cases it has been performed up to the residue Thr112 since several additional NOE restraints have been detected in this range. From the third to the fifth case, the newly detected chemical shifts (see Appendix A) have been used.

In particular, the number of restraints and the chemical shifts (existing and newly detected) used for each described case is reported in table 5.1.

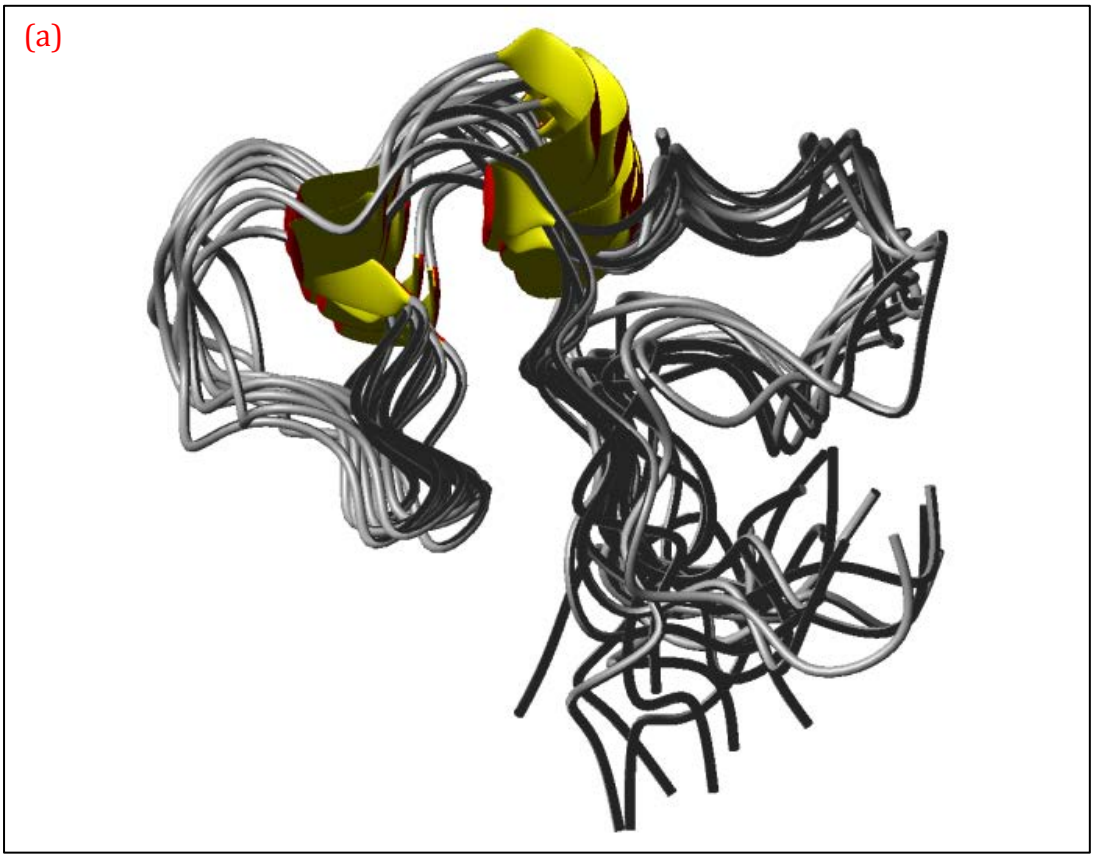
	1	**2**	**3**	**4**	**5**
Disulfide bonds	5	5	5	5	5
$^{3}JH^{N}H^{\alpha}$ couplings	29	29	0	27	27
TALOS angles	0	0	38	38	0
RDCs	32	32	32	35	35
H_bonds	2	2	2	15	15
NOE	460	460	460	926	926
TOT	530	530	539	1046	1008
CHEMICAL SHIFTS	OLD	OLD	NEW	NEW	NEW

Table 5.1 Number of restraints used in the five different computations of the PSCD4-domain structures: the numbers on the first row identify the different structure calculation. (1) using only the previously existing restraint files of Wenzler; (2) modifying the upper and lower limits of the existing NOE distance restraint file of *Wenzler* [Kalbitzer and Hengstenberg, 1992]; (3) substituting the existing J coupling restraint file of Wenzler with that one obtained with TALOS+; (4) using only the completely new detected restraints (different from those ones of *Wenzler*), including the observed J coupling and the TALOS+ dihedral angle restraints; (5) using only the new detected restraints (different from those ones of *Wenzler*) without including the TALOS+ dihedral angle restraints. Five disulfide bonds have been used in all the cases, while different amounts of $^{3}J_{H}{}^{N}{}_{H}{}^{\alpha}$ coupling, RDC, hydrogen bond and NOE restraints have been detected. The spectral assignment led to new (different from those ones of *Wenzler*) chemical shifts that have been used from the third to the fifth case.

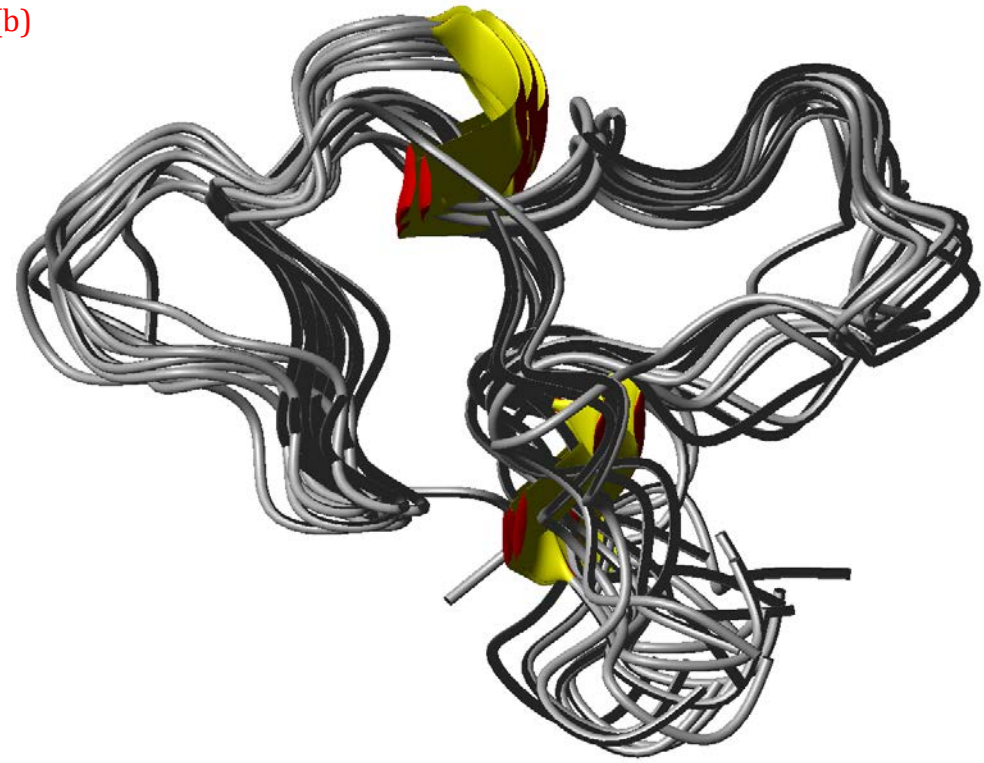
Moreover, 1024 structures have been obtained for each case and the first 10 have been retained (those ones with the minimal energy). Every bundle of 10 structures has been investigated separately with MolMol and it has been compared with the previously determined structure of *Wenzler*, as reported in Fig. 5.9. In particular, the *a* part of this figure shows the first described case, the *b* part contains the result of the second case, the *c* part represents the third one, the *d* part shows the fourth obtained structure, while the *e* part describes the structure generated in accordance to the fifth case.

The RMSD of each bundle of structures has been computed fitting the first model with the successive nine structures of the bundle (using MolMol). In particular, the RMSD has the following values starting from the first to the fifth case: 1.887, 1.576, 1.447, 1.291 and 1.276 (see table 5.2).

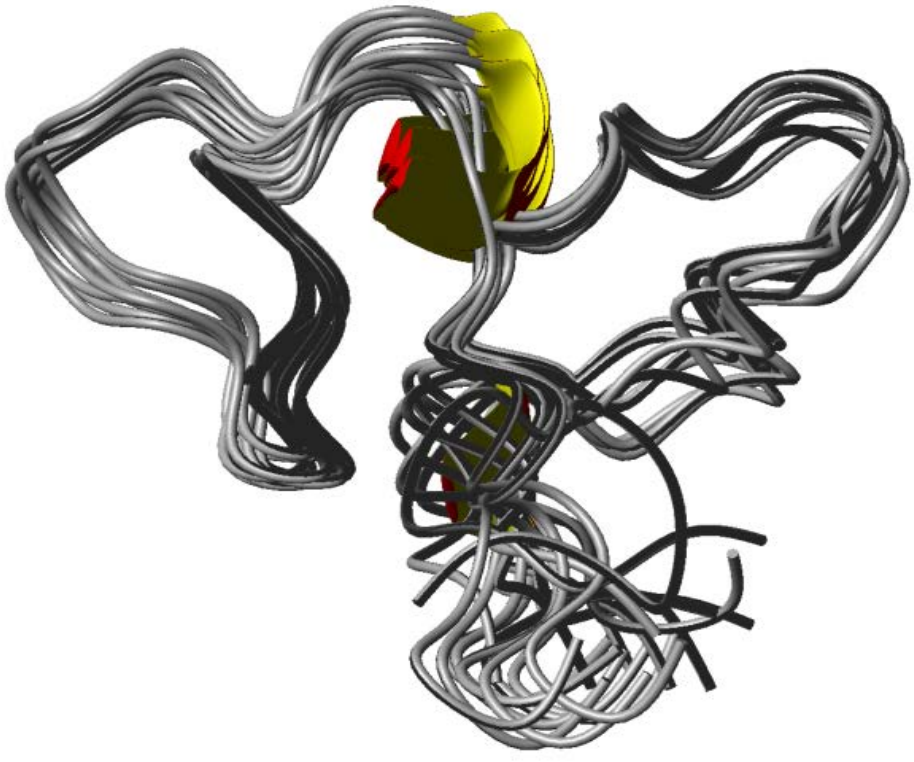
The violation files of the first ten structures of every bundle generated by CNS have been inspected. The largest violations of the detected NOE distances (> 0.5 Å) have been evaluated in order to correct overlapping assignment of chemical shifts.

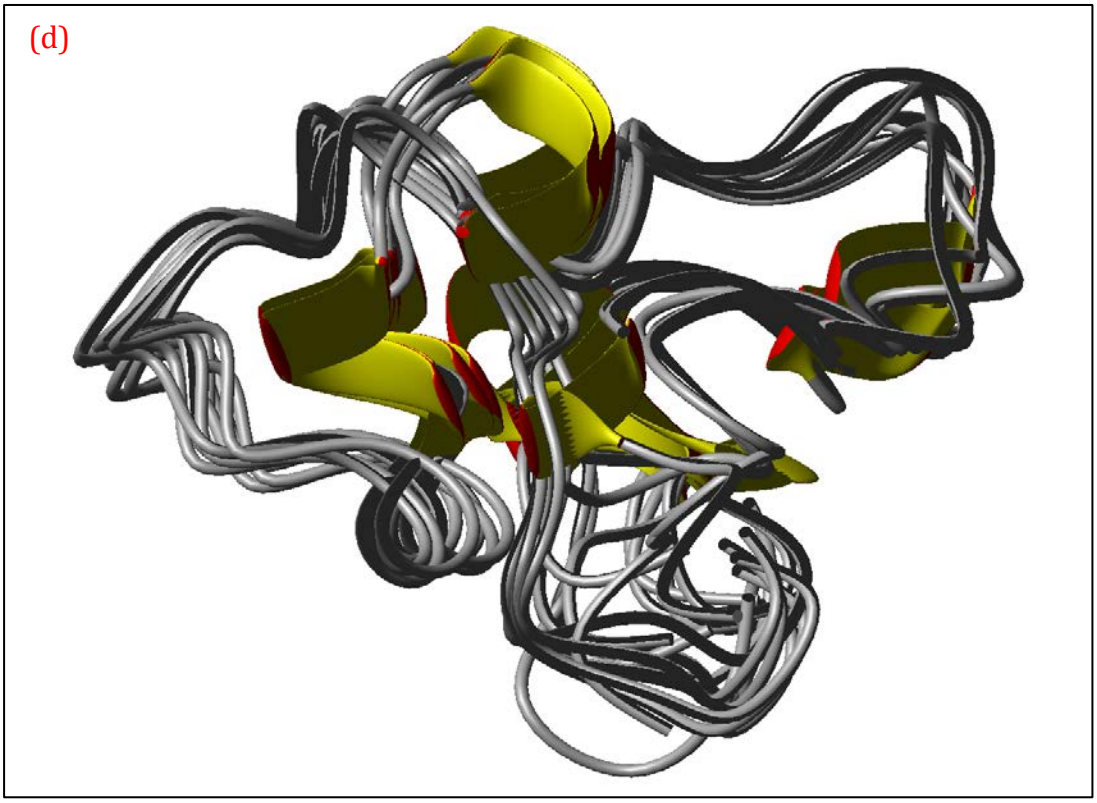


(b)



(c)





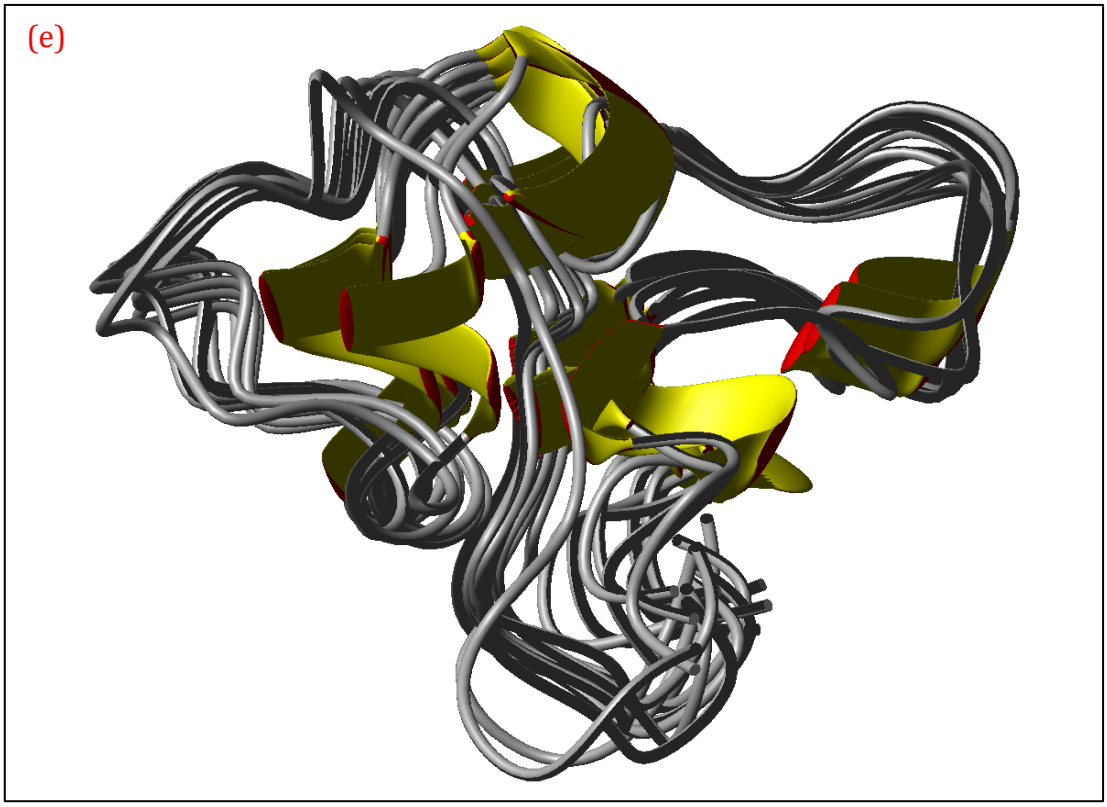


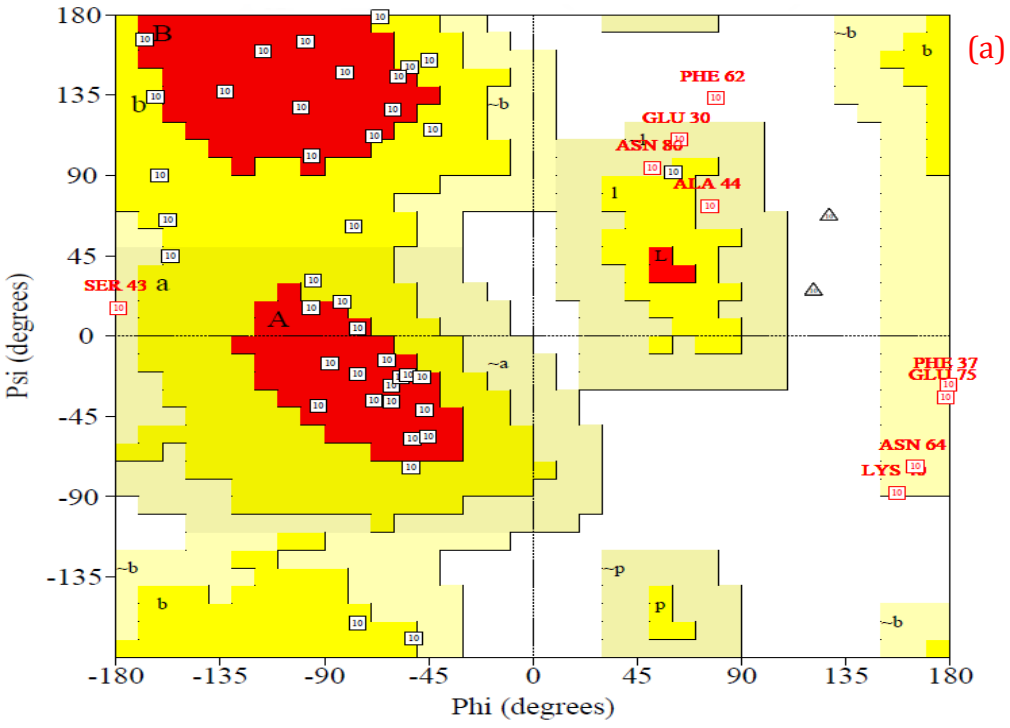
Figure 5.9 PSCD4-domain structure determination (from residue 30 to 80) without water refinement: structures obtained with different restraints: using only the previously existing restraint files of *Wenzler* (a); modifying the upper and lower limits of the existing NOE distance restraint file of *Wenzler* [Kalbitzer and Hengstenberg, 1992] (b); substituting the existing J coupling restraint file of *Wenzler* with that one obtained with TALOS+ (c); using only the completely new detected restraints, including the observed J coupling and the TALOS+ dihedral angle restraints (d); using only the new

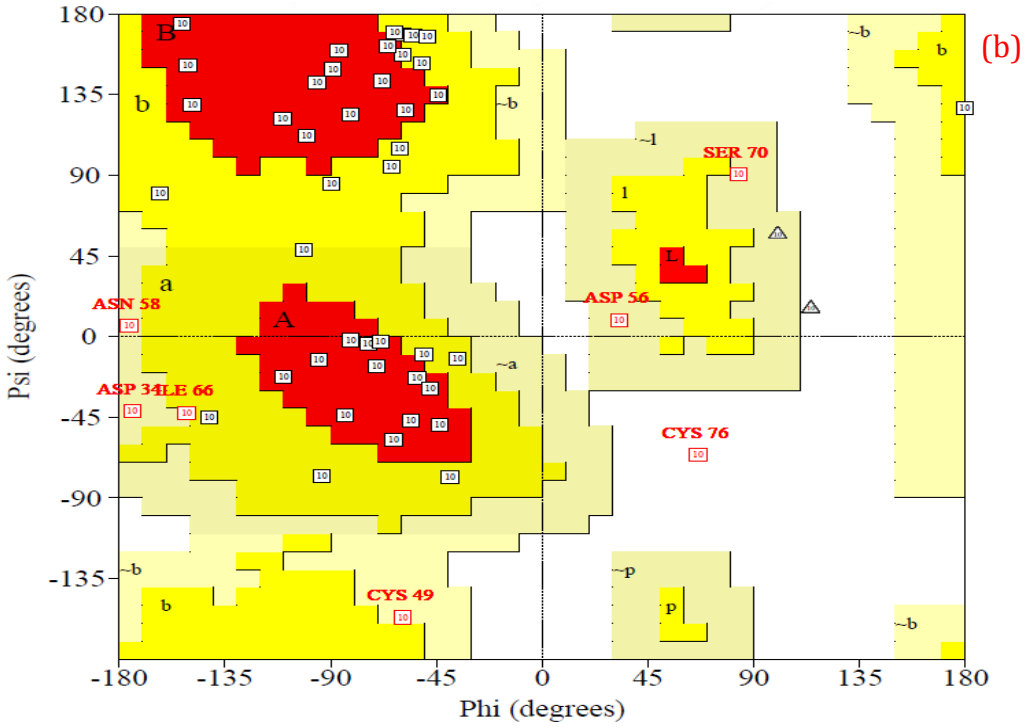
detected restraints without including the TALOS+ J coupling restraints (e). For a detailed description see Table 5.1.

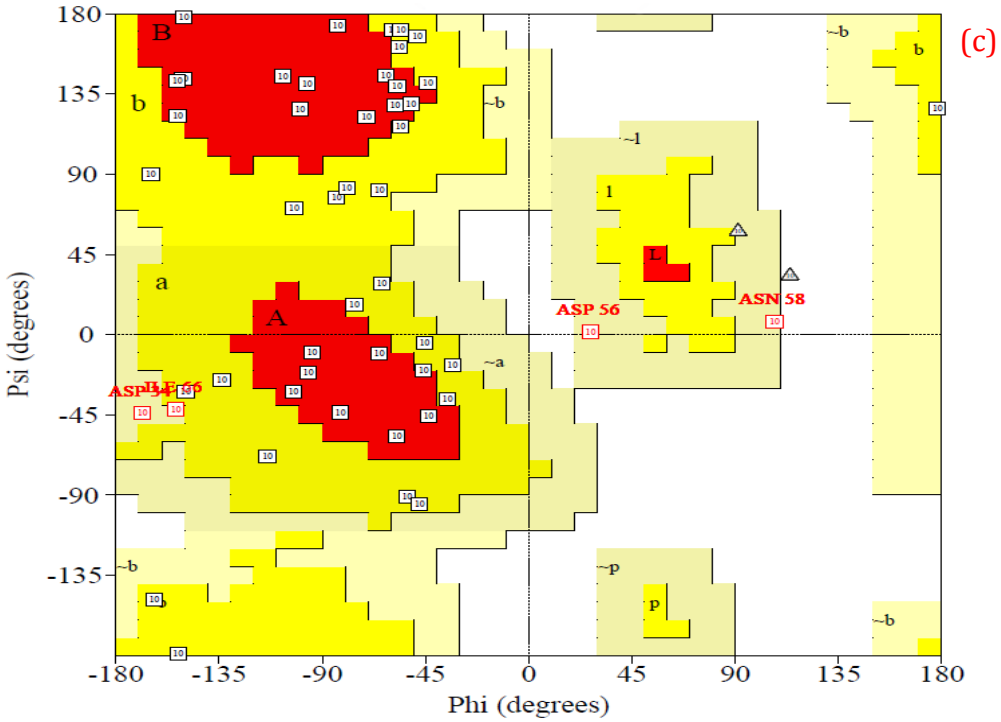
The first structure contains two α-helices in the residue ranges from Cys 49 to Gly 53 and from Pro 55 to Asn 58. The first α-helix is kept through all the calculated structures, as was originally predicted by the TALOS+ program (par. 5.1.1). In the second case another α-helix is defined in the residue range from Phe 74 to Ser 77, while in the third structure there are no other structural motives. In the fourth and in the fifth cases two additional α-helices (to the two α-helices found in the first structure) have been detected in the residue ranges from Ile 33 to Cys 36 and from Pro 39 to Asp 42.

5.1.4 STRUCTURE VALIDATION

The Ramachandran plots of the PSCD4-domain have been generated with PROCHECK for each evaluated case and they have been compared with the previously determined one of *Wenzler*. In particular, in Fig. 5.10 the Ramachandran plot related to the structure obtained accordingly to the first case (described in par. 5.1.3) is shown in the *a* part of the figure, the second case is reported in the *b* part, the third result is represented in the *c* part, the fourth case is described in the *d* part, while the *e* part shows the Ramachandran plot of the fifth case.




(b)
Psi (degrees)
Phi (degrees)
SER 70
ASP 56
ASN 58
ILE 66
CYS 76
CYS 49


(c)
Psi (degrees)
Phi (degrees)
ASP 56
ASN 58

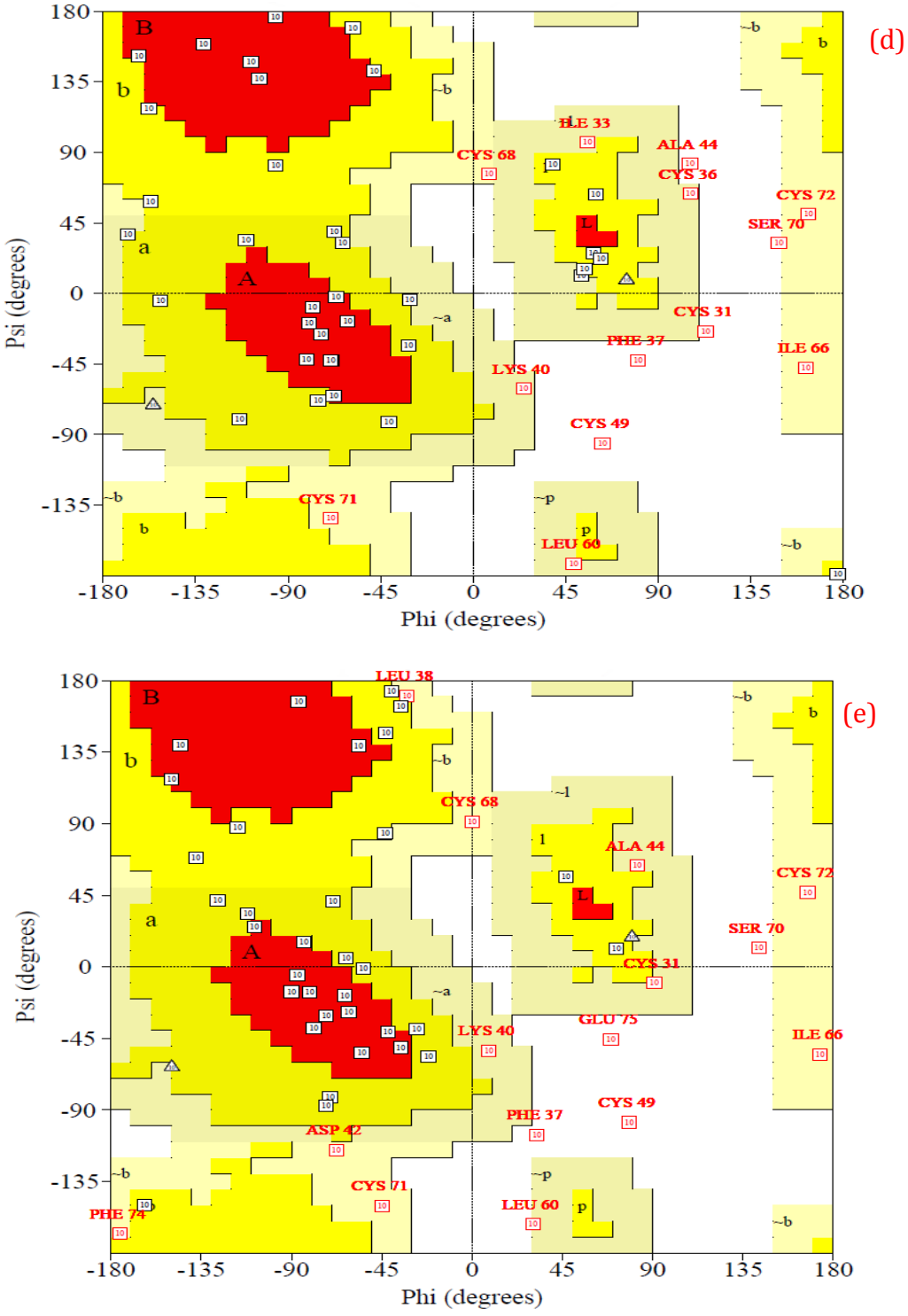


Figure 5.10 PSCD4-domain structure validation (from residue 30 to 80) without water refinement: the Ramachandran plots related to different structures obtained with different restraints. Using only the previously existing restraint files of *Wenzler* (a); modifying the upper and lower limits of the existing NOE distance restraint file of *Wenzler* [Kalbitzer and Hengstenberg, 1992] (b); substituting the existing J coupling restraint file of *Wenzler* with that one obtained with TALOS+ (c); using only the completely new detected restraints, adding the observed J coupling with the TALOS+ dihedral angle restraints (d); using only the new detected restraints without including the TALOS+ J coupling restraints (e). For a detailed description see Table 5.1.

The Ramachandran plots of the second and the third structures reveal that the majority of the residues are positioned in the most favorable regions. All the described Ramachandran plots have been analyzed with the PROCHECK software in order to determine the percentage of residues located in the most favored regions, in the strictly and generously allowed regions and in the disallowed regions (see table 5.2).

	1	2	3	4	5
E_{tot}	132.139	811.044	1537.69	8175.63	6464.66
E_{bond}	5.14995	19.6276	40.7299	262.59	263.313
E_{angle}	51.7283	104.467	235.661	1214.06	1214.66
E_{imp}	4.9956	22.6409	119.141	620.587	620.477
E_{vdw}	32.6574	86.6475	138.608	952.559	948.825
E_{elec}	4.673186E-05	3.223294E-04	6.702447E-05	2.898619E-04	9.820044E-04
E_{noe}	17.7141	82.9034	388.895	2138.93	2135.89
E_{coup}	10.5527	53.1972	40.3706	2390.25	685.02
E_{sani}	9.34075	20.4002	17.8155	10.8139	10.8292
NOE violations > 0.05nm	0	0	3	18	18
RDC violations	30	29	29	0	0
vdw violations	0	9	13	0	0
dihed violations	35	716	727	856	854
H_Bond violations > 0.05nm	0	0	0	0	0
RMSD	1.887	1.576	1.447	1.291	1.276
Residues in the most favored regions	42.9%	60.3%	47.6%	25.7%	22.4%
Residues in the strictly allowed regions	38.1%	27.0%	39.7%	41.4%	44.3%
Residues in the generously allowed regions	12.7%	11.1%	9.5%	22.9%	21.9%
Residues in the disallowed regions	6.3%	1.6%	3.2%	10.0%	11.4%
R-factor (interresidual signals)	0.290826	0.246233	0.326509	0.234169	0.245834
R-factor (all experimental and calculated signals)	0.88	0.83	0.73	0.67	0.64

Table 5.2 Energy contributions, RMSD values and Ramachandran plot results of the five different computations (see table 5.1) without water refinement: several energy terms are reported as well as the total energy and the number of NOE and hydrogen bonds violations with respect to the total amount of restraints. The average RMSD of each bundle of structures is described. The results of the Ramachandran plots are visualized as percentage of residues located in the most

favored, in the strictly and generously allowed and in the disallowed regions. The R-factor (for inter-residual signals and for all experimental and calculated signals) is reported as well.

In table 5.2 the specific energy contributions (the energy values of the every first structure obtained from the five different computations), the NOE violations, the RMSD values and the PROCHECK results of all the five datasets of generated structures are reported. The numbers on the top of every column identify the differently generated bundles of structures in accordance to the description reported in par. 5.1.3.

The first three datasets of structures reveal an overall lower energy contribution due to the smaller amount of restraints used to calculate the structure and to the shorter considered sequence (from Pro15 to Ser103). In particular, only 460 NOE-contacts, 2 hydrogen bonds and 29 $^{3}J_{H^N H^\alpha}$ coupling restraints were detected.

The R-factor [Gronwald et al, 2000] has been calculated from the two-dimensional NOESY spectrum of the PSCD4-domain of the pleuralin protein (par. 2.1.2.2.3). It decreases from the first to the last computation.

6 Discussions and Conclusions

6.1 General considerations

The time-domain measured NMR experiments typically need to be accurately processed both before and after the Fourier transformation (par. 1.2). They can be multiplied by a weighting window function (exponential, Gaussian, etc.) in the time domain, whereas phase correction are usually performed in the frequency domain (par. 1.2.1 and par.1.2.2). Baseline correction can be applied in both domains. An interactive baseline and phase correction is generally required, but it is not acceptable for a fast and completely automated procedure of protein structure determination.

Some algorithms have been developed during this work in order to consent a reliable automated phase (EMD in the time domain combined with GA, par. 4.1.3.1) and baseline correction (ALS in the frequency domain, par. 3.1.4) of multi-dimensional NMR data. The former has been tested on strong phase-distorted one-dimensional NMR spectra obtained from samples containing mixtures of various amino acids (par. 2.1.2.6). The latter has been instead applied to metabolomics (human urine) NMR one-dimensional spectra with a strong baseline distortion (par. 2.1.2.4.1), to a two-dimensional back-calculated NOESY spectrum of the HPr protein from *Staphylococcus aureus* (par. 2.1.1.1) and to the experimental three-dimensional $^{1}H^{13}C$ HCCH-TOCSY spectrum of Trx protein from *Plasmodium falciparum* (par. 2.1.2.1).

Several experimental (during the acquisition of the data) and post-processing (in both domains) methods for suppressing the solvent artifact (par. 1.2.3) are generally used for a more reliable detection of the resonances of interest in the investigated spectra. The existing procedures typically are not able to recover the solute signals hidden by the solvent artifact.

In this work, some algorithms have been developed and compared (SSA and ICA in the time and in frequency domain respectively, par. 2.2.2 and par. 2.2.3) in order to establish the most efficient automated method. The SSA has been chosen due to its ability in recovering the resonances of interest in spectra severely affected by the solvent artifact.

In particular, the SSA has been successfully applied to experimental NMR data (three-dimensional $^{1}H^{13}C$ HCCH-TOCSY spectrum of Trx protein from *Plasmodium falciparum* (par. 2.1.2.1), two-dimensional NOESY and TOCSY spectra of HPr protein from *Staphylococcus aureus* (par. 2.1.2.2.1 and par. 2.1.2.2.2), one-dimensional metabolomics data as blood, urine and cell spectra (par. 2.1.2.3.2, par. 2.1.2.3.3 and par. 2.1.2.4.1) and to back-calculated (RELAX-JT2 algorithm) NMR spectra (one- and

two-dimensional NOESY spectra of HPr protein from *Staphylococcus aureus*, par. 2.1.1).

The application of the SSA followed by ALS (AUREMOL_SSA/ALS) is mathematically rather simple and provides as good results as those ones obtained with more complicated methods. In particular, the SSA has been totally automated and introduced in the AUREMOL package followed by the ALS that is applied in cascade in the investigated data for an automated baseline correction. This latter can be also used out of the SSA algorithm (par. 3.1.4).

The AUREMOL_SSA/ALS has been tested successfully for higher-dimensional NMR spectra earlier [Malloni et al, 2009]. In this work, it has been shown that some modifications had to be introduced (as the number of extracted components in the SSA and the window size in the ALS) in order to apply it properly also to one-dimensional spectra. In particular, it has been demonstrated that the higher digital resolution of one-dimensional spectra involves the extraction of a larger amount of components (par. 3.1.2) in order to better distinguish the solvent signal (represented by the component with the highest variance) from the rest. In principle, it could be useful to nullify more than one component before reconstructing the signal, an idea that will be worked out in the future.

It is clear that using modern NMR spectrometers an excellent water suppression and almost perfect baselines can be achieved with selective excitation techniques such as WET [Smallcombe et al, 1995], watergate [Piotto et al, 1992; Sklenar et al, 1993; Saudek et al, 1994; Liu et al, 1998] and excitation sculpting [Hwang et al, 1995] or by applying more complicated selective pre-saturation sequences such as PURGE [Simpson and Brown, 2005] instead of the simple NOESY-type pre-saturation sequence used here. Selective excitation methods have always the disadvantage, that a rather large spectral range around the water signal is attenuated or not visible at all that may contain valuable information. In addition, the quality of the water suppression also depends on the concentration of solute under consideration. At mM concentrations these methods can attenuate the water signal to such a degree that it is not stronger than a typical resonance of the solute. In contrast at μM concentrations often relevant in biology these methods fail and the application of SSA/ALS would lead to a significant improvement of the spectral quality.

The complete AUREMOL_SSA/ALS algorithm has been tested on the two-dimensional NOESY spectrum of the PSCD4-domain of the pleuralin protein (par. 1.3.2, par. 2.1.2.2.3 and par. 3.1.4) in order to perform a more reliable signal assignment (KNOWNOE algorithm) and consequently to extrapolate more and better defined distance restraints (REFINE algorithm) from the spectrum. These and some other restraints (par. 2.1.3 and par. 5.1.2) have been used to calculate (CNS) the three-dimensional structure of the considered domain.

The limitations of the SSA have been investigated (par. 3.1.1.3) and the ICA has been proposed as a valid alternative to properly deal with such unfavorable cases (par. 4.1.2.2.1 and par. 4.1.2.2.2). Since the ICA needs in input at least as many different spectra as the number of desired components (solute and solvent components) the definition of a suitable protocol for generating the dataset of ICA-tailored data is straightforward (par. 4.1.2.2.2). Some specific pulse sequences have been designed and successfully applied to one-dimensional NMR spectra of HPr protein from *Staphylococcus carnosus* demonstrating even a better performance of the SSA.

Advantages and disadvantages of both methods have been described (par. 4.1.2). The ICA will become part of the AUREMOL package as well, with the advantage of recovering resonances of interest, especially in those SSA non-manageable cases (where the solvent signal is not the dominant one in the spectrum).

6.1.1 SOLVENT SUPPRESSION BY MEANS OF SINGULAR SPECTRUM ANALYSIS

The SSA has been developed in order to remove the solvent artifact from NMR spectra of any dimensionality, digitally and analogically acquired, decomposing every time-domain signal (FID) into one solvent and several solute related components. Time signal embedding (par. 2.2.2) in the space of time-delayed coordinates (building one trajectory matrix for each investigated FID) is applied and the eigenvalue decomposition is performed on these data (par. 2.2.1). The embedding dimension (i.e. the number of components) has been determined empirically in accordance to a qualitative and quantitative analysis of the extracted components (par. 3.1.2). The projection of the eigenvector related to the largest eigenvalue (describing the highest variance in the data, i.e. the solvent signal) is nullified and the embedding process is then reverted. The pre-processing and the post-processing of the time-domain and the frequency domain data are mandatory. The former includes the automated management of the group delay data points (par. 3.1.1.1) and the signal normalization (par. 3.1.1.2). The latter is instead related to the phase correction according to the group delay and to the baseline correction (par. 3.1.4).

The SSA cannot satisfactorily be applied on spectra whose dominant signal is not the solvent (as in case of watergate experimental solvent suppression and in case of artificial data not including the experimental solvent signal, par. 3.1.1.3). The optimal solvent-to-solute ratio for a reliable solvent suppression has been investigated by means of a quantitative analysis of the resulting spectra (after SSA) compared with the corresponding one-dimensional back-calculate one of the HPr protein from *Staphylococcus aureus* (par. 3.1.1.3).

The application of certain experimental solvent suppression methods (par. 1.2.3.2) during the measurement could lead to multi-dimensional spectra whose solvent signal is still the dominant one only in some rows of the spectrum. If the SSA were

conducted over such data, it would yield unsatisfactory results (vertical stripes of artifacts would appear in the spectrum). It should be selectively applied only along specific rows that should be automatically identified. Therefore, this additional application could be taken into account for a further development of the method that must be able to deal with this type of data. For instance, the spectrum could be automatically evaluated row by row in order to determine if the solvent artifact is the dominant signal in the considered row. This task could be easily performed in the frequency domain, where typically the solvent signal is recognized in the middle of the spectrum. The SSA algorithm is applied in the time domain, thus an initial Fourier transform of each FID would be automatically performed. At this point, the algorithm should verify if the strongest intensity of the spectrum is located in the middle of this latter and in such case the FID corresponding to the considered row could undergo the SSA for solvent suppression. The entire procedure would require a previously applied phase correction to do not compromise the recognition of the strongest intensity in every row of the spectrum. Moreover, a specific range of the solvent region could be interactively defined.

Actually, the ICA has revealed to be a valid alternative to suppress the solvent signal in spectra whose dominant signal is not the water artifact, but a further improvement of the SSA in not to exclude.

6.1.1.1 AUTOMATED BASELINE CORRECTION BY MEANS OF LINEAR SPLINE

The automated recognition of baseline points (par. 3.1.4) in the spectra has been obtained with a method similar to FLATT [Güntert and Wüthrich, 1991]. It searches for contiguous pieces of row or column that can be well fitted by a straight line (as it happens in baseline regions). A sliding window must be used to look for such regions and it must be larger than the line width of the protein peaks. As originally proposed by *Güntert and Wüthrich*, the fixed window size of 75 Hz has revealed to do not be suitable for all the types and all the dimensionalities of experiments. Therefore, the method has been modified in such a way that the window size is automatically adapted in dependence on the considered spectrum. In particular, if two- and three-dimensional data are analyzed the maximal values of the line width of the peaks are computed in each direction separately (e.g. row by row) and some histograms are built containing the occurrence of the line width values in each direction. The most frequently occurring line width value is used to establish the window size. In one-dimensional spectra (with a larger degree of signal overlap) this definition does not properly work (e.g. in case of biological sample with large variations of line width values) thus the method has been re-arranged (for such cases) in a way that the

window size is determined as the maximal occurring line width (not the most frequent one).

The Automated Linear Spline (ALS) has been developed and introduced in the AUREMOL package in order to perform an automated baseline correction with an automated identification of the baseline points (differently determined in dependence on the dimensionality of the data) and a consequent linear spline interpolation of these points that is subtracted from the original data row- and column-wise.

In particular, in the AUREMOL package it is used in a fully automated way in cascade after the application of the SSA without any user intervention. Considering the simplicity and the efficiency of this automated implementation, the AUREMOL_SSA/ALS should become a widely diffused tool for the treatment of multi-dimensional spectra.

6.1.2 SOLVENT SUPRRESSION BY MEANS OF INDEPENDENT COMPONENT ANALYSIS

The SSA cannot be properly applied (to suppress the water signal) to spectra whose solvent artifact is not the dominant signal (par. 3.1.1.3). Therefore, the ICA has been investigated as a valid alternative to properly manage these cases. The latter decomposes the overlapping signals in the frequency domain and it needs at least as many different inputs or mixtures (spectra with different weights for each underlying signal) as the number of source signals (solvent and solute components). Dealing with one-dimensional data implies the creation of a set of at least two spectra tailored for the application of the ICA. A proper protocol with a specific pulse sequence must be followed in order to generate suitable ICA-tailored data (par. 4.1.2.2.2).

The ICA (par. 2.2.3) determines the components maximizing the non-Gaussianity of the sources. In particular, several cost functions can measure the non-Gaussianity (e.g. kurtosis, negentropy) of the components, thus several algorithms (e.g. FastICA, JADE, InfoMax) are generally used. The FastICA algorithm has been chosen being fast and reliable. It has been applied to the ICA-tailored one-dimensional spectra of the HPr from *Staphylococcus carnosus* in order to recover resonances of interest showing a better performance than the SSA (par. 4.1.2.2.2).

Both methods could be combined with the aim to overcome the limitations of every one of them.

The SSA provides an easily interpretable set of components as output. In particular, they are directly disposed in a decreasing order in accordance to the variance of the

contained signal (par. 2.2.1). This is useful for an automated and fast recognition of the solvent artifact that has the highest variance in all the cases that it represents the dominant signal in the spectrum. Using ICA does not furnish a specific order, scale and sign of the extracted components (par. 2.2.3). This fact involves the spectroscopist to the direct visual inspection of the components and it is not acceptable for a complete automation of the structure determination procedure.

A further development of the ICA algorithm could include an automated evaluation of the two extracted components. The recognition of the solvent component could be performed either as described in the case of the SSA applied only to some specific rows (par. 6.1.1) or with a method similar to those ones previously proposed by Joyce [Joyce et al, 2004] or by Nicolaou [Nicolaou and Nasuto, 2004]. In the former case, the algorithm should consider that the highest intensity of the data is typically located in the middle of the spectrum or alternatively that the spectrum of interest (between the components) contains significant intensity values only outside the solvent region.

6.1.3 AUTOMATED PHASE CORRECTION BY MEANS OF EMPIRICAL MODE DECOMPOSITION

The Empirical Mode Decomposition (EMD) combined with genetic algorithms (GA) has been proposed as a reliable method to perform an automated phase correction (par. 4.1.3.1) and it has shown to be able to properly deal with strongly phase distorted data.

The AUREMOL_SSA/ALS for practical applications has the advantage of a complete automation. Phase correction must be typically performed interactively since still no stable methods exist. The EMD_GA combination has shown to be a promising tool for such purposes.

The EMD relies on a process called sifting which allows the decomposition of the signal into a finite set of oscillatory components (with a decreasing frequency of oscillation from the first to the last component). The sum of the first Intrinsic Mode Functions (IMFs) extracted from the time-domain signal is Fourier transformed and the non-baseline region are determined. The genetic algorithms are applied in order to optimize the fitness function that maximizes the number of points whose intensity value is over a locally defined threshold in every non-baseline region separately. The individuals of the random population are identified by two random genes representing the zero- and first-order values of phase correction. The individual optimizing the fitness function after a previously determined number of generations contains the correct phase correction values in its genes.

The EMD method does not allow the recovery of resonance of interest hidden underneath the solvent signal since several baseline distortions and artifacts appear

in the solvent region (par. 4.1.3). Further investigation of the method could lead to more satisfactory results in this field of application. The EMD has anyway shown to be useful to deal with a spectrum containing all the relevant resonances of interest that can be used to properly correct the phase of the investigated spectrum. The amount of the extracted IMFs may differ depending on the size and on the type of considered data, thus the amount of IMFs to be summed can differ.

An automated recognition of the necessary IMFs could be further developed in order to improve the performance of the phase correction. In particular, an automated comparison of the original spectrum with that one obtained summing some of the extracted IMFs should yield a maximal overlap (e.g. comparing the intensity of the signals) except in the solvent region, where it must be minimal. The choice to apply the genetic algorithms on the sum of the first IMFs (instead of applying them separately to each IMF) has come from the investigation of the components. The first one (with the highest frequency of oscillation) could be used to detect only some of the solute resonances that need to be phase corrected. The second IMF allows a similar analysis over some other resonances in the spectrum but due to the presence of baseline distortions and noise, the identification of the signals of interest start to be more complex with the risk to compromise the complete automation of the method. This problem becomes particularly relevant on the successive IMFs (see Fig. 4.17). The sum of certain IMFs provides instead a well-defined solute spectrum without remarkable distortions except in the solvent region. This condition is sufficient to properly recognize all the non-baseline regions that need to be phase corrected.

The performance of the phase correction is not only related to the ability of selecting the correct amount of IMFs to be summed, but it also depends on the capability of identifying true signals and baseline points (par. 3.1.4), that may represent an hard task in very distorted spectra. The genetic algorithm application (par. 4.1.3.1) involves the definition of some initial parameters: the population size, the allowed range of gene values (PHC0 and PHC1, zero- and first-order phase correction respectively) representing each individual, the method for selecting the reproducing couples, the type of mutation (either both values or only one and in which range), the percentage of mutation (the amount of mutated individuals in each generation) and the number of generations. The modification of these parameters could lead to different performances of the phase correction.

6.2 PSCD4-domain of the pleuralin protein

6.2.1 PROTEIN STRUCTURE DETERMINATION

The pleuralin cell wall protein is obtained from the diatom *Cylindrotheca fusiformis* organism [Kröger et al, 1997; Wenzler et al, 2001]. In particular, the HEP200 has a modular construction with an N-terminal, a proline-rich domain, five proline-rich conserved PSCD-domains (with 87 or 89 amino acids in each domain) ad a C-terminal. The PDSC-domains share the 73-91% of the sequence with ten cysteine residues at exactly the same positions in all the domains. They are separated by different short sequences of amino acids. The recombinant His_6PSCD4 (with 112 amino acids) contains the PSCD4-domain and it has been newly investigated (in accordance to the previous work of *Wenzler*, 2003).

Several spectra (triple resonance experiments) have been used for the sequential assignment of the backbone and side chain atoms (par. 2.1.3). The SSA and the ALS have been applied on the two-dimensional NOESY spectrum that has been automatically assigned (KNOWNOE algorithm) and it has been evaluated to automatically determine distance restraints (REFINE algorithm). They have incremented the number of those previously detected ones [Wenzler, 2003]. The TALOS+ software has produced a list of dihedral angle restraints from the new chemical shifts (Appendix A). The routine to convert the $^3J_{H^N H^\alpha}$ coupling to torsion angle restraints and vice versa has been introduced in the AUREMOL package (par. 5.1.2). The TALOS+ restraints have been investigated and integrated with the new detected $^3J_{H^N H^\alpha}$ coupling restraints (Appendices B and C). The newly determined hydrogen bonds and residual dipolar coupling restraints (Appendices D and E) have been also used to calculate (CNS) several three-dimensional structures of the PSCD4-domain.

In particular, five different calculations of the PSCD4-domain structure have been performed using different restraints (par. 5.1.3). Comparing the results reported in Table 5.2 (par. 5.1.4), the Ramachandran plots, the RMSD values and the R-factors, the best structures have been obtained using the newly detected chemical shifts (Appendix A) and either the existing restraints [Wenzler, 2003] with the addition of the TALOS+ torsion angle restraints (third case in par. 5.1.3) or using only the newly detected restraints (RDC, H_Bond, $^3J_{H^N H^\alpha}$ coupling and NOE) with or without those ones obtained with TALOS+ (fourth and fifth cases in par. 5.1.3).

All the calculated structures have been obtained without water refinement. The percentage of residues lying in the not allowed region has increased from the first to the last case (with completely new determined restraints), whereas the R-factor and the RMSD have decreased. The water refinement could lead to better performances improving the Ramachandran plot of the last evaluated structures.

The potential model of all the five PSCD-domains of the pleuralin protein has been described by *Wenzler*. Those domains are separated by different sequences of residues of various lengths. The disulfide bridges are conserved over the five domains, thus they can be used as restraints for the computation of the other domains. The NOE restraints of the PSCD4-domain are instead kept through the

domains if the involved amino acids are identical between the sequence of the PSCD4 and the other considered domains. Using such information it could possible to calculate the bundle of structures of the entire pleuralin protein (from the amino acid 88 to the 581) with CNS. In particular, the residues lying in the sequences connecting the domains do not possess any restraints, but a refined structure could be obtained adding some no-NOE-contacts between the single PSCD-domains.

References

[Adler and Wagner, 1991] M.Adler and G.Wagner (1991) Removal of dispersive baseline distortions caused by strong water signals. *J.Magn.Reson*. **91**(2):450-454.

[Antoine et al, 2000] J.P.Antoine, A.Coron and J.M.Dereppe (2000) Water peak suppression: time-frequency vs time-scale approach. *J.Magn.Reson.* **144**:189-194.

[Antz et al, 1995] C.Antz, K.P.Neidig and H.R.Kalbitzer (1995) A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. *J.Biomol.NMR* **5**: 287-296.

[Aranibar et al, 2011] N.Aranibar, M.Borys, N.A.Mackin, V.Ly, N.Abu-Absi, S.Abu-Absi, M.Niemitz, B.Schilling, Z.J.Li, B.Brock, R.J.Russell II, A.Tymiak and M.DReily (2001) NMR-based metabolomics of mammalian cell and tissue cultures. *J.Biomol.NMR* **49**(3-4):195-206.

[Aue et al., 1976] W.P.Aue, E.Bartholdi and R.R.Ernst (1976) Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *J.Chem.Phys.* **64**(5): 2229-2246.

[Balacco, 1994] G.Balacco (1994) A new criterion for automatic phase correction of high-resolution NMR spectra which does not require isolated or symmetrical lines *J.Magn.Reson.* **110**(1): 19-25.

[Balacco and Cobas, 2009] G.Balacco and J.C.Cobas (2009) Automatic phase correction of 2D NMR spectra by a whitening method. *Magn.Reson.Chem*. **47**(4): 322-327.

[Barache et al, 1997] D.Barache, J.P.Antoine and J.M.Dereppe (1997) The continuous Wavelet transform, an analysis tool for NMR spectroscopy. *J.Magn.Reson.* **128**(1):1-11.

[Bartels et al, 1995] C.Bartels, P.Güntert and K.Wüthrich (1995) IFLAT-A new automatic baseline-correction method for multidimensional NMR spectra with strong solvent signals. *J.Magn.Reson.* A **117**: 330-333.

[Barsukov and Arseniev, 1987] I.L.Barsukov and A.S.Arseniev (1987) Base-plane correction in 2D NMR. *J.Magn.Reson.* **73**(1):148-149.

[Baskaran et al, 2009] K.Baskaran, R.Kirchhöfer, F.Huber, J.Trenner, K.Brunner, W.Gronwald, K.P.Neidig and H.R.Kalbitzer (2009) Chemical shift optimization in multidimensional NMR spectra by AUREMOL-SHIFTOPT. *J.Biomol.NMR* **43**(4):197-210.

[Baskaran et al, 2010] K.Baskaran, K.Brunner, C.E.Munte and H.R.Kalbitzer (2010) Mapping of protein structural ensembles by chemical shifts. *J.Biomol.NMR* **48**(2): 71-83.

[Bax et al, 1991] A.Bax, M.Ikura, L.E.Kay and G.Zhu (1991) Removal of F1 baseline distortion and optimization of folding in multi-dimensional NMR spectra. *J.Magn.Reson*. **91**: 174-178.

[Bax and Pochapsky, 1992] A.Bax and S.Pochapsky (1992) Optimized recording of heteronuclear multi-dimensional NMR spectra using pulsed field gradients. *J.Magn.Reson.* **99**: 638–643.

[Beckonert et al, 2007] O.Beckonert, H.C.Keun, T.M.D.Ebbels, J.Bundy, E.Holmes, J.C.Lindon and J.K.Nicholson (2007) Metabolic profiling, metabolomics and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat.Prot.* **2**:2692-2703.

[Bell and Sejnowski, 1995] A.J.Bell and T.J.Sejnowski (1995) An information-maximization approach to blind separation and blind deconvolution. *Neur.Comp*. **7**:1129-1159.

[Bell and Sejnowski, 1997] A.J.Bell and T.J.Sejnowski (1997) The Independent Components of natural scenes are edge filters. *Vis.Res.* **37**(23): 3327.

[Beloucharani et al, 1997] A.Beloucharani, K.A.Meriam, J.F.Cardoso and E.Mouline (1997) A blind source separation technique using second order statistics. *IEEE Trans.Sign.Proc.* **45**(2): 434-444.

[Böhm et al, 2006] M.Böhm, K.Stadlthanner, P.Gruber, F.J.Theis, E.W.Lang, A.M.Tomé, A.R. Teixeira, W.Gronwald and H.R.Kalbitzer (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.

[Borgognone et al, 2001] M.G.Borgognone, J.Bussi and G.Hough (2001) Principal component analysis in sensory analysis: covariance or correlation matrix? *Food qual.pref.* **12**(5-7):323-326.

[Bragg, 1907] W.H.Bragg (1907) The nature of Röntgen rays. Trans.Royal Soc.Sci.Australia 31:94

[Braun and Go, 1985] W.Braun and N.Go (1985) Calculation of protein conformations by proton-proton distance constraints: a new efficient algorithm. *J.Mol Biol.* **186:**611–626.

[Brooks et al, 1983] B.R.Brooks, R.E.Bruccoleri, B.D.Olafson, D.J.States, S.Swaminathan and M.Karplus (1983) CHARMM: a program for macromolecular energy minimization and dynamics calculations *J. Comp. Chem.* **4**:187–217.

[Broomhead and King, 1986] D.S.Broomhead and G.P.King (1986) Extracting qualitative dynamics from experimental data. *Physica* **20**D:217-236.

[Brown et al, 1989] D.E.Brown, T.W.Campbell and R.N.Moore (1989) Automated phase correction of FT NMR spectra by baseline optimization. *J.Magn.Reson.* **85**(1): 15-23.

[Brown and Campbell, 1990] D.E.Brown and T.W.Campbell (1990) Enhancement of 2D NMR spectra using singular value decomposition. *J.Magn.Reson.* **89**:255-264.

[Brown, 1995] D.E.Brown (1995) Fully automated baseline correction of 1D and 2D NMR spectra using Bernstein polynomials. *J.Magn.Reson.* **114**:268–270.

[Brown et al, 1977] F.F.Brown, I.D.Campbell, P.W.Kuchel and D.C.Rabenstein (1977) Human erythrocyte metabolism studies by ^{1}H spin echo NMR. *FEBS Lett.* **82**(1): 12–16.

[Bruenger, 1993] A.T.Brunger (1993) XPLOR: a system for x-ray crystallography and NMR. 3.1 Ed.Yale Univ.Press, CT.

[Bruenger et al, 1998] A.T.Brünger,P.D.Adams, G.M.Clore, W.L.DeLano, P.Gros, R.W.Grosse-Kunstleve, J.S.Jiang, J.Kuszewski, M.Nilges, N.S.Pannu, R.J.Read, L.M.Rice, T.Simonson and G.L.Warren (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr.*D **54**:905–921.

[Bruenger and Nilges, 1993] A.T.Brünger and M.Nilges (1993) Computational challenges for macromolecular structure determination by x-ray crystallography and solution NMR spectroscopy. *Q. Rev.Biophys.* **26**, 49–125.

[Bryce and Bax, 2004] D.L.Bryce and A.Bax (2004) Application of correlated residual dipolar couplings to the determination of the molecular alignment tensor magnitude of oriented proteins and nucleic acids. *J.Biomol.NMR* **28**:273-287.

[Cardoso, 1999] J.F.Cardoso (1999) High-order contrasts for independent component analysis. *Neur.Comp.* **11**(1):157-192.

[Cavagnero et al, 1999] S.Cavagnero, H.J.Dyson, P.E.Wright (1999) Improved low pH bicelle system for orienting macromolecules over a wide temperature range. *J.Biomol.NMR* **13**(4):387-391.

[Cavanagh et al., 1996] J.Cavanagh, W.J.Fairbrother, A.G.Palmer III, M.Rance and N.JSkelton (1996) Protein NMR Spectroscopy Principles and Practice. Acad.Press Inc., San Diego.

[Chang et al, 2007] D.Chang, C.D.Banck and S.L.Shah (2007) Robust baseline correction algorithm for signal dense NMR spectra. *J.Magn.Reson*. **187**(2): 288-292.

[Chen et al, 2002] L.Chen, Z.Weng, L.Y.Goh and M.Garland (2002) An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. *J.Magn.Reson.* **158**:164-168.

[Chothia et al, 1986] C.Chothia and A.M.Lesk (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J*. **5**(4):823–826.

[Chylla and Markley, 1993] R.A.Chylla and J.L.Markley (1993) Simultaneous basepoint correction and signal recognition in multidimensional NMR spectra. *J.Magn.Reson.* B **102**:148-154.

[Cieslar et al, 1988] C.Cieslar, G.M.Clore and A.M.Gronenborn (1988) Automatic phasing of pure phase absorption two-dimensional NMR spectra. *J.Magn.Reson.* **79**(1):154-157.

[Clore et al, 1998a] G.M.Clore, M.R.Starich and A.M.Gronenborn (1998) Measurement of residual dipolar couplings of macromolecules aligned in the nematic phase of a colloidal suspension of rod-shaped viruses. *J.Am.Chem.Soc.* **120**:10571–10572.

[Clore et al, 1998b] G.M.Clore, A.M.Gronenborn and A.Bax (1998) A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *J. Magn.Reson.* **133**(1):216–221.

[Cobas et al, 2006] J.C.Cobas, M.A.Bernstein, M.M.Pastor and P.G.Tahoces (2006) A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *J.Magn.Reson.* **183**(1):145-151.

[Comon, 1994] P.Comon (1994) Independent Component Analysis, a new concept? *Sign. Proces*. **36**(3): 287-314.

[Cong et al, 2009] F.Cong, T.Sipola, T.Huttunen-Scott, X.Xu, T.Ristaniemi, and H.Lyytinen (2009) Hilbert-Huang versus Morlet wavelet transformation on mismatch negativity of children in uninterrupted sound paradigm. *Nonlin.Biom.Phys.* **3**(1):1-8.

[Cordier, and Grzesiek, 1999] F.Cordier and S.Grzesiek (1999) Direct observation of hydrogen bonds in proteins by interresidue $^{3h}J_{NC'}$ scalar couplings. *J.Am.Chem.Soc.* **121**:1601-1602.

[Cordier et al, 2008] F.Cordier, L.Nisius, A.J.Dingley and S.Grzesiek (2008) Direct detection of N-H[...]O=C hydrogen bonds in biomolecules by NMR spectroscopy. *Nat.Protoc.* **3**(2):235-41.

[Corneliescu et al, 1999] G.Cornilescu, F.Delaglio and A.Bax (1999) TALOS: Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J.Biom.NMR* **13**(3):289-302.

[Cornell et al, 1995] W.D.Cornell, P.Cieplak, C.I.Bayly, I.R.Gould, Jr.K.M.Merz, D.M.Ferguson, D.C.Spellmeyer, T.Fox, J.W.Caldwell, and P.A.Kollman (1995) A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J.Am.Chem. Soc*. **117**:5179–5197.

[Coron et al, 2001] A.Coron, L.Vanhamme, J.P.Antoine, P.V.Hecke and S.Van Huffel (2001) The filtering approach to solvent peak suppression in MRS: a critical review. *J.Magn.Reson.* **152**(1):26-40.

[Coughlin and Tung, 2004] K.T.Coughlin and K.K.Tung (2004) 11-Year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Adv.Space Res.* **34**:323–329.

[Couillard-Depres et al, 2004] S.Couillard-Depres, G.Uyanik, S.Ploetz, C.Karl, H.Koch, J.Winkler and L.Aigner (2004) Mitotic impairment by dublecortin is diminished by dublecortin mutations found in patients. *Neurogen.* **5**(2): 83-93.

[Craig and Marshall, 1988] E.C.Craig and A.G.Marshall (1988) Automated phase correction of FT NMR spectra by means of phase measurement based on dispersion versus absorption relation (DISPA). *J.Magn.Reson.* **76**: 458-475.

[Crippen and Havel, 1988] G.M.Crippen and T.F.Havel (1988). Distance Geometry and Molecular Conformation (Chemometrics Series) Research Studies Press.

[Danilov and Zhigljavsky, 1997] D.Danilov and A.Zhigljavsky (1997) Principal components of time series: the Caterpillar method. Univ.St.Petersburg Press.

[Daubenfeld et al, 1985] J.M.Daubenfeld, J.C.Boubel, J.J.Delpuech, B.Neff and J.C.Escalier (1985) Automatic intensity, phase and baseline corrections in quantitative carbon-13 spectroscopy. *J. Magn. Reson*. **62**(2): 195-208.

[Delorme et al, 2001] A.Delorme, S.Makeig and T.Sejnowski (2001) Automatic artifact rejection for EEG data using high-order statistics and independent component analysis. Proc. Third Intern. ICA Conference, 457-462.

[De Sanctis, 2006] S.De Sanctis (2006) Previsione dell' evoluzione di sistemi naturali tramite un modello ibrido AG-NN applicato ad immagini satellitari.

[De Sanctis et al, 2011] S.De Sanctis, W.M.Malloni, W.Kremer, A.M.Tomé, E.W.Lang, K.P.Neidig and H.R.Kalbitzer (2011) Singular spectrum analysis for an automated solvent artifact removal and baseline correction of 1D NMR spectra. *J.Magn.Res*. **210**(2), 177-183.

[Dietrich et al, 1991] W.Dietrich, C.H.Rüdel and M. Neumann (1991). Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra. *J.Magn.Reson.* **91**(1):1–11.

[Dzakula, 2000] Z.Dzakula (2000) Phase angle measurement from peak areas (PAMPAS). *J.Magn.Reson.* **146**(1):20-32.

[Eilers, 2003] P.H.Eilers (2003) A perfect smoother. *Anal. Chem.* **75**(14):3631–3636.

[Ernst, 1969] R.R.Ernst (1969) Numerical Hilbert transform and automatic phase correction in magnetic resonance spectroscopy. *J.Magn.Reson.* **1**:7-26.

[Friedrichs et al, 1991] M.S.Friedrichs, W.J.Metzler and L.Mueller (1991) Removal of diagonal peaks in two-dimensional NMR spectra by means of digital filtering. *J.Magn.Reson*. **95**:178-183.

[Gardner et al, 1998] K.H.Gardner and L.E.Kay (1998) The use of 2H, 13C, 15N multidimensional NMR to study the structure and dynamics of proteins. *Annu.Rev.Biophys.Biomol.Struct*. **27**:357–406.

[Gerbrands, 1981] J.J.Gerbrands (1981) On the relationships between SVD, KLT and PCA. *Patt. Recogn.* **14**:375-381.

[Geyer et al, 1995] M.Geyer, K.P.Neidig, and H.R.Kalbitzer (1995) Automated peak integration in multidimensional NMR spectra by an optimized iterative segmentation procedure. *J.Magn.Reson*. B **109**(1):31-38.

[Ghil et al. 2002] M.Ghil, M.R.Allen, M.D.Dettinger and K.Ide (2002) Advanced spectral methods for climatic time series. *Rev.Geoph.* **40**(1):1003.

[Golotvin and Williams, 2000] S.Golotvin and A.Williams (2000) Improved baseline recognition and modeling of FT NMR spectra. *J.Magn.Reson.* **146**(1):122–125.

[Golyandina et al, 2001] N.Golyandina, V.Nekrutkin and A.A.Zhigljavsky (2001) Analysis of time series structure: SSA and related techniques. Chapman and Hall/CRC.

[Görler and Kalbitzer, 1997] A.Görler and H.R.Kalbitzer (1997) Relax, a flexible program for the back calculation of NOESY spectra based on complete-relaxation-matrix formalism. *J. Magn.Reson.* **124**(1): 177-188.

[Görler et al, 1999] A.Görler, W.Gronwald, K.P.Neidig and H.R.Kalbitzer (1999) Computer assisted assignment of 13C or 15N edited 3D-NOESY-HSQC spectra using back calculated and experimental spectra. *J.Magn.Reson.* **137**:39-45.

[Grahn et al, 1988] H.Grahn, F.Delaglio, M.A.Delsuc and G.C.Levy (1988) Multivariate data analysis for pattern recognition in two-dimensional NMR. *J.Magn.Reson.* **77**(2):294-307.

[Griesinger et al, 1987] C.Griesinger, O.W.Sørensen and R.R.Ernst (1987) Practical aspects of the E.COSY technique. Measurement of scalar spin-spin coupling constants in peptides. *J.Magn.Reson.* **75**:474-492.

[Gronwald et al, 2000] W.Gronwald, R.Kirchhofer, A.Gorler, W.Kremer, B.Ganslmeier, K.P.Neidig and H.R.Kalbitzer (2000) RFAC, a program for automated NMR R-factor estimation. *J. Biomol. NMR* **17**(2):137-151.

[Gronwald and Kalbitzer, 2004] W.Gronwald and H.R.Kalbitzer (2004) Automated structure determination of proteins by NMR spectroscopy. *Prog. NMR Spectrosc.* **44**:33-96.

[Gronwald et al, 2002] W.Gronwald, S.Moussa, R.Elsner, A.Jung, B.Ganslmeier, J.Trenner, W.Kremer, K.P.Neidig and H.R.Kalbitzer (2002) Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE). *J.Biomol.NMR* **23**(4) 271-287.

[Grzesiek and Bax, 1993a] S.Grzesiek and A.Bax (1993) The importance of not suturing H2O in protein NMR. Application to sensitivity enhancement and NOE measurements. *J.Am.Chem.Soc.* **115**:12593-12594.

[Grzesiek and Bax, 1993b] S.Grzesiek and A.Bax (1993) Measurement of amide proton exchange rates and NOE with water in 13C/15N enriched calcineurin B. *J.Biomol.NMR* **3**:627-638.

[Güntert et al, 1991] P.Güntert, W.Braun and K.Wüthrich (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J.Mol.Biol*. **217**(3):517–530.

[Güntert and Wüthrich, 1992] P.Güntert and K.Wüthrich (1992) FLATT- A new procedure for high quality baseline correction of multidimensional NMR spectra. *J.Magn.Reson.* **96**:403-407.

[Güntert et al, 1992] P.Güntert, V.Dötsch, G.Wider and K.Wüthrich (1992) Processing of multi-dimensional NMR data with the new software PROSA. *J.Biomol.NMR* **2**:619-629.

[Günther et al, 2002] U.L.Günther, C.Ludwig and H.Rüterjans (2002) WAVEWAT—improved solvent suppression in NMR spectra employing wavelet transforms. *J.Magn.Reson.* **156**(1):19-25.

[Habeck et al, 2005] M.Habeck, W.Rieping and M.Nilges (2005) Bayesian estimation of Karplus parameters and torsion angles from three-bond scalar couplings constants. *J.Magn.Res.* **177**(1):160-165.

[Halamek et al, 1994] J.Halamek, V.Vondra and M.Kasal (1994) The elimination of baseline distortions induced by audio filters. *J.Magn.Reson*. A **110**:194-197.

[Hansen et al, 1998] M.R.Hansen, L.Mueller and A.Pardi (1998) Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat. Struct. Biol.* **5**(12):1065–1074.

[Hardy and Rinaldi, 1990] J.K.Hardy and P.L.Rinaldi (1990) Principal component analysis for artifact reduction in COSY Spectra. *J.Magn.Reson*. **88**:320-333.

[Havel, 1991] T.F.Havel (1991) An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog.Biophys.Mol.Biol.* **56**(1):43–78.

[Hecky et al, 1973] R.E. Hecky, K.Mopper, P.Kilham and T.E.Degens (1973) The amino acid and sugar composition of diatom cell-walls. *Mar.Biol.* **19**(4):323-331.

[Heuer and Haeberlen, 1989] A.Heuer and U.Haeberlen (1989) A new method for suppressing baseline distortions in FT NMR. *J.Magn.Reson.* **85**(1):79-94.

[Heuer, 1991] A.Heuer (1991) A new algorithm for automatic phase correction by symmetrizing lines. *J.Magn.Reson.* **91**(2): 241–253.

[Hoffman et al, 1992] R.E.Hoffman, F.Delaglio and G.C.Levy (1992) Phase correction of two-dimensional NMR spectra using DISPA. *J.Magn.Reson.* **98**(2):231-237.

[Holland, 1975] J.H.Holland (1975) Adaptation in natural and artificial systems, Universityof Michigan Press, Ann Arbor, Michigan.

[Holmes et al, 1992] E.Holmes, J.K.Nicholson, F.W.Bonner, B.C.Sweatman, C.R.Beddell, J.C.Lindon and E.Rahr (1992) Mapping the biochemical trajectory of nephrotoxicity by pattern recognition of NMR urinanalysis. *NMR Biomed* **5**:368-372.

[Hore, 1983] P.Hore (1983) Solvent suppression in Fourier transform NMR. *J.Magn.Reson.* **55**(2):283-300.

[Hoult, 1976] D.I.Hoult (1976) Solvent peak saturation with single phase and quadrature Fourier transformation. *J.Magn.Reson.* **21**(2):337-347.

[Hoult et al, 1983] D.I.Hoult, C.-N.Chen, H.Eden and M.Eden (1983). Elimination of baseline artifacts in spectra and their integrals. *J.Magn.Reson*. **51**(1):110-117.

[Huang et al, 1998] N.E.Huang, Z.Shen, S.R.Long, M.C.Wu, H.H.Shih, Q.Zheng, N.C.Yen, C.C.Tung, and H.H.Liu (1998) The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proc.R.Soc.Lod.* A **454**:903–995.

[Huber, 1985] P.J.Huber (1985) Projection pursuit. *Ann.Statis.* **13**(2):435–475.

[Hwang et al, 1995] T.L.Hwang and A.J.Shaka (1995) Water suppression that works. Excitation sculpting using arbitrary wave-forms and pulsed-field gradients. *J.Magn.Reson*. A **112**:275–279.

[Hyberts et al, 1992] S.G.Hyberts, M.S.Goldberg, T.F.Havel and G.Wagner (1992) The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. *Protein Sci.* **1**(6):736–751.

[Hyvärinen and Oja, 1997] A.Hyvärinen and E.Oja (1997) A fast fixed-point algorithm for independent component analysis. *Neur.Comp.* **9**(7):1483-1492.

[Hyvärinen et al, 2001] A.Hyvärinen, J.Karhunen and E.Oja (2001) Independent Component Analysis, Wiley.

[Hyvärinen, 1998] A.Hyvärinen (1998) New approximations of differential entropy for independent component analysis and projection pursuit. *NIPS* **10**:273–279.

[Joliffe, 1986] I.T.Jolliffe (1986) Principal component analysis. New York, Springer-Verlag.

[Jones and Sibson, 1987] M.C.Jones and R.Sibson (1987) "What is projection pursuit?" *J.Royal Stat.Soc.* A **150**(1):1–37.

[Joyce et al, 2004] C.A.Joyce, I.F.Gorodnitsky and M.Kutas (2004) Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychop*. **41**(2):313-325.

[Jung et al, 1998] T.P.Jung, C.Humphries, T.W.Lee, S.Makeig, M.J.Mckeown, V.Iragui and T.J. Sejnowski (1998) Extended ICA removes artifacts from electroencephalographic recordings. *Adv.Neur.Inform.Proc.Sys.* **10**:894-900.

[Kalbitzer and Hengstenberg, 1992] H.R.Kalbitzer and W.Hengstenberg (1993) The solution structure of the histidine-containig protein (HPr) from *Staphylococcus aureus* as determined by two-dimensional ^{1}H-NMR spectroscopy. *Eur.J.Biochem*. **216**:205-214.

[Kalk and Berendsen, 1976] A.Kalk and H.J.C.Berendsen (1976) Proton magnetic relaxation and spin diffusion in proteins*. J.Magn.Reson.* **24**:343–366.

[Karplus, 1963] M.Karplus (1963) Vicinal proton coupling in Nuclear Magnetic Resonance. *J.Am.Chem.Soc.* **85**(18):2870–2871.

[Kay et al, 1992] L.E.Kay, P.Keifer and T.Saarinen (1992) Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc*. **114**:10663–10665.

[Kendrew et al, 1958] J.C.Kendrew, G.Bodo, H.M.Dintzis, R.G.Parrish, H.Wyckoff and D.C.Phillips (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**(4610): 662-666.

[Kirkpatrick et al, 1983] S.Kirkpatrick, Jr.C.DGelatt and M.P.Vecchi (1983) Optimization by simulated annealing. *Science* **220**:671–680.

[Koehl et al, 1995] P.Koehl, C.Ling and J.F.Lefevre (1995) Automatic phase correction of NMR spectra: statistics and limits. *J.Chim.Phys*. **92**:1929-1938.

[Kraskov et al, 2004] A.Kraskov, H.Stögbauer and P.Grassberger (2004) Estimating mutual information. *Phys.Rev.* **69**(6):066138.

[Krishnaveni et al, 2006] V.Krishnaveni, S.Jayaraman and K.Ramadoss (2006) Application of mutual information based least dependent component analysis (MILCA) for removal of ocular artifacts from electroencephalogram. *Int.J.Biomed.Sci.* **1**(1):63-74.

[Kröger et al, 1997] N.Kröger, G.Lehmann, R. Rachel and M. Sumper (1997) Characterization of a 200-kDa diatom protein that is specifically associated with a silica-based substructure of the cell wall. *Eur.J.Biochem.* **250**:99-105.

[Kumar et al, 1980] A.Kumar, R.R.Ernst and K.Wüthrich (1980) A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem.Biophys.Res.Comm*. **95**:1–6.

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

[Lee and Sejnowski, 1996] T.W.Lee and T.J.Sejnowski (1996) Independent Component Analysis for sub Gaussian and super-Gaussian mixtures. *Proc.4th Joint Symp. on Neur.Comp.* **7**:132-139.

[Levenberg, 1944] K.Levenberg (1944) A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quart.Appl.Mat.* **2**:164–168.

[Liavas and Regalia, 2001] A.P.Liavas and P.A.Regalia (2001) On the behavior of information theoretic criteria for model order selection. IEEE *Trans.Sign.Proc*. **49**(8):1689-1695.

[Lin et al, 1997] Y.Y.Lin, P.Hodgkinson, M.Ernst and A.Pines (1997) A Novel detection–estimation scheme for noisy NMR signals: applications to delayed acquisition data. *J.Magn.Reson.* **128**(1):30-41.

[Lindon et al, 1999] J.C.Lindon, J.K.Nicholson, J.R.Everett (1999) NMR spectroscopy of biofluids. *Ann.Rep.NMR Spect.* **38**:1–88.

[Lippens et al, 1995] G.Lippens, C.Dhalluin and J.M.Wieruszeski (1995) Use of a water flip-back pulse in the homonuclear NOESY experiment. *J.Biomol.NMR* **5**(3):327-331.

[Liu et al, 1998] M.Liu, X.Mao, C.Ye, H.Huang, J.K.Nicholson and J.C.Lindon (1998) Improved WATERGATE pulse sequence for solvent suppression in NMR spectroscopy. *J.Magn.Reson.* **132**(1):125-129.

[Lo et al, 2008] M.-T.Lo, K.Hu, Y.Liu, C.-K.Peng, and V.Novak (2008) Multimodal pressure-flow analysis: Application of Hilbert-Huang transform in cerebral blood flow regulation. *EURASIP J. Adv. Sig. Proc.* **2008**:785243.

[Lo et al, 2009] M.-T.Lo, V.Novak, C.-K.Peng, Y.Liu, and K.Hu (2009) Nonlinear phase interaction between non-stationary signals: a comparison study of methods based on Hilbert-Huang and Fourier transforms. *Phys. Rev. E* **79**: 061924.

[Lobel et al, 1996] K.D.Lobel, J.K.West and L.L.Hench (1996) Computational model for protein-mediated biomineralization of the diatom frustule. *Marine Biol.* **126**(3):353-360.

[Makeig et al, 1996] S.Makeig, T.P.Jung, A.J.Bell and T. J.Sejnowski (1996) Independent Component Analysis of electroencephalographic data. *Adv.Neur.Inf.Proc.Sys.* **8**:145-151.

[Malloni et al, 2010] W.M.Malloni, S.De Sanctis, A.M.Tomé, E.W.Lang, C.E.Munte, K.P.Neidig and H.R. Kalbitzer (2010) Automated solvent artifact removal and base plane correction of multidimensional NMR protein spectra by AUREMOL-SSA. *J.Biomol.NMR* **47**(2), 101-111.

[Marion and Bax, 1988] D.Marion and A.Bax (1988) Baseline distortion in real-Fourier-transform NMR spectra. *J.Magn.Reson.* **79**:352-356.

[Marion et al, 1989] D.Marion, L.E.Kay, S.W.Sparks, D.A.Torcia and A.Bax (1989) Three-dimensional heteronuclear NMR of 15N labeled proteins. *J.Am.Cjem.Soc.* **111**:1515-1517.

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

[Marion and Bax, 1989] D.Marion and A.Bax (1989) Baseline correction of 2D FT NMR spectra using a simple linear prediction extrapolation of the time-domain data. *J.Magn.Reson*. **83**:205-211.

[Marquardt, 1963] D.Marquardt (1963) An algorithm for least-squares estimation of nonlinear parameters. SIAM J.Appl.Mat. 11(1):431-441.

[Maurer et al, 2004] T.Maurer, S.Meier, N.Kachel, C.E.Munte, S.Hasenbein, B.Koch, W.Hengstenberg and H.R.Kalbitzer (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.

[McKay, 2011] R.T.McKay (2011) How the 1D-NOESY suppresses solvent signal in metabolomics NMR spectroscopy: an examination of the pulse sequence components and evolution. *Conc.Magn.Reson.* A **38**(5):197-220.

[Mertz et al, 1991] J.E.Mertz, P.Güntert, K.Wüthrich and W.Braun (1991) Complete relaxation matrix refinement of NMR structures of proteins using analytically calculated dihedral angle derivatives of NOE intensities. *J.Biomol.NMR*. **1**:257–269.

[Messerle et al, 1989] B.A.Messerle, G.Wider.,G.Otting, C.Weber and K.Wüthrich (1989) Solvent suppression using a spin lock in 2D and 3D NMR spectroscopy with H_2O solutions. *J.Magn.Reson.* **85**:608-613.

[Mitschang et al, 1990] L.Mitschang, K.P.Neidig and K.H.Kalbitzer (1990) Suppression of oscillatory artifacts in two-dimensional NMR spectra. *J.Magn.Reson*. **90**(2):359-362.

[Mort and Lamport, 1997] A.J.Mort and D.T.A.Lamport (1977) Anhydrous hydrogen fluoride deglycosylates glycoproteins. *Anal.Biochem*. **82**(2):289-309.

[Moskau, 2002] D.Moskau (2002) Application of real time digital filters in NMR spectroscopy. *Conc.Magn.Reson.* **15**(2):164-176.

[Munte et al, 2009] C.E.Munte, K.Becker, R.H.Schirmer and H.R.Kalbitzer (2009) NMR assignments of oxidized thioredoxin from *Plasmodium falciparum*. *Biomol.NMR* Assign. **3**(2):159-161.

[Neff et al, 1977] B.L.Neff J.L.Ackerman, and J.S. Waugh (1977) Fully automatic software correction of Fourier transform NMR spectra *J.Magn.Reson.* **25**(2):335-340.

[Neidig et al, 1995] K.P.Neidig, M.Geyer, A.Görler, C.Antz, R.Saffrich, W.Beneicke and H.R. Kalbitzer (1995) AURELIA, a program for computer-aided analysis of multidimensional NMR spectra. *J.Biomol.NMR* **6**(3):255-270.

[Neuhaus and Williamson, 1989] D.Neuhaus and M.Williamson (1989) The Nuclear Overahauser Effect in structural conformational analysis. Wiley-VCH.

[Macura and Ernst, 1980] S.Macura and R.R.Ernst (1980) Elucidation of cross relaxation in liquids by two-dimensional NMR spectroscopy. *Mol. Phys*. **41**:95–117.

[Nicholson et al, 1999] J.K.Nicholson, J.C.Lindon JC and E.Holmes (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenob*. **29**(11):1181–1189.

[Nicolaou and Nasuto, 2004] N.Nicolaou and S.J.Nasuto (2004) Temporal Independent Component Analysis for automatic artefact removal from EEG signals". *Procs. MEDSIP* 5-8.

[Nilges et al, 1988] M.Nilges, G.M.Clore and A.M.Gronenborn (1988) Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. Circumventing problems associated with folding. *FEB Lett.* **239**(1):129-136.

[Nowick et al., 2003] J.S.Nowick, O.Khakshoor, M.Hashemzadeh and J.O.Brower (2003) DSA: A new internal standard for NMR studies in aqueous solution. *Org.Lett.* **5**(19):3511–3513.

[Oschkinat et al., 1988] H.Oschkinat, C.Griesinger, P.J.Kraulis, O.W.Sarensen, R.R.Ernst, A.M.Gronenborn and G.M.Clore (1988) Three-dimensional NMR-spectroscopy of a protein in solution. *Nature* **332**:374-376.

[Ottiger and Bax, 1998] M.Ottiger and A.Bax (1998) Characterization of magnetically oriented phospholipid micelles for measurement of dipolar couplings in macromolecules. *J.Biomol.NMR* **12**:361–372.

[Otting et al, 1986] G.Otting, H.Widmer, G.Wagner and K.Wüthrich (1986) t1- and t2-ridges in 2D-NMR Spectra: origin and procedures for suppression. *J.Magn.Reson* **66**:187-193

[Pardi et al, 1984] A.Pardi, M.Billeter, and K.Wüthrich (1984) Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, ^{3}JHN alpha, in a globular protein: use of ^{3}JHN alpha for identification of helical secondary structure. *J.Mol.Biol.* **180**(3):741–751.

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

[Pearson, 1901] K.Pearson (1901) On lines and planes of closest fit to systems of points in space. *Philos.Mag.* **2**(6):559-572.

[Pearson, 1977] G.A.Pearson (1977) A general baseline-recognition and baseline-flattening algorithm. *J.Magn.Reson*. **27**(2):265-272.

[Pervushin et al, 1997] K.Pervushin, R.Riek, G.Wider and K.Wuthrich (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci. USA* **94**(23):12366–12371.

[Pijnappel et al, 1992] W.W.F.Pijnapple, A.Van Den Boogaart, R.De Beer and D.Van Ormondt (1992) SVD-based quantification of magnetic resonance signals. *J.Magn.Reson.* **97**:122-134.

[Piotto et al, 1992] M.Piotto, V.Saudek and V.Sklenar (1992) Gradient-tailored excitation for single-quantum NMR Spectroscopy of aqueous solutions. *J.Biomol.NMR* **2**(6):661-665.

[Plateau and Gueron, 1982] P.Plateau and M.Gueron (1982) Exchangeable proton NMR without baseline distortion using new strong-pulse sequences. *J.Am.Chem.Soc.* **104**(25):7310-7311.

[Purcell et al., 1946] E.M.Purcell, H.C.Torrey and R.V.Pound (1946) Resonance absorption by nuclear magnetic moments in a solid. *Phys.Rev.* **69**:37-38.

[Quiroga et al, 2000] R.Q.Quiroga, J.Arnhold, and P.Grassberger (2000) Learning driver-response relationships from synchronization patterns. *Phys.Rev*. E **61**:5142.

[Quiroga et al, 2002] R.Q.Quiroga, A.Kraskov, T.Kreuz, and P.Grassberger (2002) Performance of different synchronization measures in real data: a case study on EEG signals. *Phys.Rev.* E **65**:041903.

[Ramm et al, 2009] P.Ramm, S.Couillard-Depres, S.Plotz, F.J.Rivera, M.Krampert, B.Lehner, W.Kremer, U.Bogdahn, H.R.Kalbitzer and L.Aigner (2009) A NMR biomarker for neural progenitor cells: is it all neurogenesis? *Stem Cells* **27**(2):420-423.

[Renugopalakrishnan et al, 1991] V.Renugopalakrishnan, P.R.Carey, I.C.P.Smith, S.G.Huans and A.C.Storer (1991) Proteins: structure, dynamics and design. Springer, 1 edition.

[Ried et al, 2004] A.Ried, W.Gronwald, J.M.Trenner, K.Brunner, K.P.Neidig and H.R.Kalbitzer (2004) Improved simulation of NOESY spectra by RELAX-JT2 including effects of J-coupling, transverse relaxation and chemical shift anisotrophy. *J.Biomol.NMR* **30**(2):121-131.

[Rouh et al, 1993] A.Rouh, M.A.Delsuc, G.Bertrand and J.Y.Lallemand (1993) The use of classification in baseline correction of FT NMR spectra. *J.Magn.Reson.* A **102**:357-359.

[Saffrich et al, 1992] R.Saffrich, W.Beneicke, K.P.Neidig and H.R.Kalbitzer (1993) Baseline correction in *n*-dimensional NMR spectra by sectionally linear interpolation. *J.Magn.Reson* B **101**:304-308.

[Sanders and Schwoneck, 1992] C.R.Sanders and J.P.Schwoneck (1992) Characterization of magnetically oreintable bilayers in mixtures of dihexanoylphosphatidylcholine and dimyristoylphosphatidylcholine by solid-state NMR. *Biochem*. **31**:8898-8905.

[Sanders et al, 1994] C.R.Sanders, B.J.Hare, K.P.Howard and J.H.Prestegard (1994) Magnetically-oriented phospholipids micelles as a tool for the study of membrane associated molecules. *Prog.NMR.Spectrosc*.**26**:421-444.

[Sass et al, 1999] J.Sass, F.Cordier, A.Hoffmann, M.Rogowski, A.Cousin, J.C.Omichinski, H.Lowen and S.Grzesiek (1999) Purple membrane induced alignment of biological macromolecules in the magnetic field. *J.Am.Chem.Soc.* **121**:2047-2055.

[Sass et al, 2000] J.Sass, G.Museo, S.J.Stahl, P.T.Wingfield and S.Grezsiek (2000) Solution NMR of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *J.Biomol.NMR* **18**:303-309.

[Sattler et al, 1999] M.Sattler, J.Schleucher and C.Griesinger (1999). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Progr.Nucl.Magn.Reson.Spect*. **34**(2):93-158.

[Saupe, 1964] A.Saupe (1964) Nuclear Magnetic Resonance in liquid crystals and in liquid-crystal solutions. *Naturforsch.* **19**:161-171.

[Saupe, 1968] A.Saupe (1968) Recent results in the field of liquid crystals. *Angew.Chem.Int.Ed* **7**(2):97-112.

[Saudek et al, 1994] V.Saudek, M.Piotto and V.Sklenar (1994) WATERGATE: applications in homonuclear and heteronuclear nD NMR Experiments. *Bruker Report* **140**(94):6-9.

[Schulze et al, 2005] G.Schulze, A.Jirasek, M.M.L.Yu, A.Lim, R.F.B.Turner and W.B.Michael (2005) Investigation of selected baseline removal techniques as candidates for automated implementation. *App.Spectrosc.* **59**(5):545–574.

[Simpson and Brown, 2005] A.J.Simpson and S.A.Brown (2005) Purge NMR: effective and easy solvent suppression. *J.Magn.Reson.* **175**(2):340-346.

[Sklenar et al, 1993] V.Sklenar, M.Piotto, R.Leppik and V.Saudek (1993) Gradient-tailored water suppression for 1H-15N HSQC experiments optimized to retain full sensitivity. *J.Magn.Reson*. A **102**:241-245.

[Smallcombe et al, 1995] S.H.Smallcombe, S.L.Patt and P.A.Keifer (1995) WET solvent suppression and its applications to LC-NMR and high-resolution NMR spectroscopy. *J.Magn.Reson.* A **117**:295–303.

[Stadlthanner et al, 2006] K.Stadlthanner, A.M.Tome, F.J.Theis, E.W.Lang, W.Gronwald and H.R.Kalbitzer (2006) Separation of water artifacts in 2D NOESY protein spectra using congruent matrix pencils. *Neurocomp.* **69**:497-522.

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

[Tang, 1994] C.Tang (1994) An analysis of baseline distortion and offset in NMR spectra *J.Magn.Reson*. A **109**:232–240.

[Tolman et al, 1995] J.R.Tolman, J.M.Flanagan, M.A.Kennedy and J.H.Prestegard (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc. Natl. Acad. Sci. USA* **92**:9279–9283.

[Trenner, 2006] J.M.Trenner (2006) Accurate proton-proton distance calculation and error estimation from NMR data for automated protein structure determination in AUREMOL.

[Tsang et al, 1990] P.Tsang, P.E.Wright and M.Rance (1990) Signal suppression in the frequency domain to remove undesirable resonances with dispersive lineshapes. *J.Magn.Reson.* **88**:210-215.

[Tjandra et al, 1997] N.Tjandra, J.G.Omichinski, A.M.Gronenborn, G.M.Clore and A.Bax (1997) Use of dipolar 1H-15N and 1H-13C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat.Struct.Biol.* **4**:732–738.

[Tjandra and Bax, 1997] N.Tjandra and A.Bax (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science.* **278**(5340):1111-1114.

[Van den Hoek et al, 1993] C.Van den Hoek, H.M.Jahns and D.G.Mann (1993) Algen, 3rd edn Thieme-Verlag, Stuttgart.

[Vigario et al, 2000] R.Vigario, J.Sarela, V.Jousmaki, M.Hamalainen and E.Oja (2000) Independent component approach to the analysis of EEG and MEG recordings. IEEE *Trans.Biom.Eng.* **47**(5): 589-593.

[Volcani, 1981] B.E.Volcani (1981) Cell wall formation in diatoms: morphogenesis and biochemistry, in Silicon and siliceous structures in biological systems. Springer-Verlag, 157-200.

[Wang et al, 2001] B.Wang, U.Fleischer, J.F.Hinton and P.Pulay (2001) Accurate prediction of proton chemical shifts. *J.Comp.Chem.* **22**(16):1887-1895.

[Wang et al, 2008] Z. Wang, A. Maier, N. K. Logothetis, , and H. Liang (2008) Single-trial classification of bistable perception by integrating empirical mode decomposition, clustering, and support vector machine. EURASIP *J.Adv. Sig. Proc.* 592742.

[Wei and Werner, 2006] Y.F.Wei and M.H.Werner (2006) iDC: a comprehensive toolkit for the analysis of residual dipolar couplings for macromolecular structure determination. *J.Biomol.NMR* **35**(1):17-25.

[Wenzler et al, 2001] M.Wenzler, E.Brunner, N.Kröger, G.Lehmann, M.Sumper and H.R. Kalbitzer (2001). Letter to the editor: ^{1}H ^{13}C and ^{15}N sequence-specific resonance assignment of the PSCD4-domain of diatom cell wall protein pleuralin-1. *J.Biomol.NMR* **20**(2):191-192.

[Wenzler, 2003] M.Wenzler (2003) Die erste struktur eines zellwandproteins einer diatomee: NMR-spektroskopie charakterisierung der PSCD4-domäne von pleuralin-1 aus *Cylindrotheca fusiformis* unter verwendung von methoden zur partiellen orientirung.

[Wider, 1990] G.Wider (1990) Elimination of baseline artifacts in NMR spectra by oversampling. *J.Magn.Reson.* **89**:406-409.

[Wilson et al, 1987] I.D.Wilson, J.Fromson, I.M.Ismail and J.K.Nicholson (1987) Proton magnetic resonance spectroscopy of human urine: excretion of 1-(3'-carboxypropyl)-3,7-dimethylxanthine by man after dosing with oxpentifylline. *J.Pharm.Biomed.Anal.* **5**(2):157-163.

[Wishart et al, 1995] D.S.Wishart, C.G.Bigam, A.Holm, R.S.Hodges and B.D.Sykes. "1H, 13C and 15N random coil NMR chemical shifts of the common amino acids: I. investigations of nearest neighbor effects. *J.Biom.NMR* **5**(1):67-81.

[Wittaker, 1923] E.T. Whittaker (1923) On a new method of graduation. *Proc.Edinburgh Math.Soc*. **41**:63–75.

[Wu et al, 2001] Z.Wu, E.Schneider, Z.Hu and L.Cao (2001) The impact of global warming on ENSO variability in climate records. COLA Technical report. 110:25.

[Xi and Rocke, 2008] Y.Xi and D.M.Rocke (2008) Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinf.* **9**(1):324-333.

[Zolnai et al, 1989] Z.Zolnai, S.Macura and J.L.Markley (1989) Spline method for correcting baseplane distortions in two-dimensional NMR spectra. *J.Magn.Reson.* **82**(3):496–504.

[Zhu et al, 1997] G.Zhu, D.Smith and Y.Hua (1997) Post-acquisition solvent suppression by Singular-Value Decomposition*. J.Magn.Reson.* **124**(1):286-289.

Appendix A

RESIDUE ID	RESIDUE NAME	ATOM NAME	SHIFT
2	Y	CA	57.75
2	Y	CB	40.12
2	Y	C	175.42
3	Y	N	115.11
3	Y	HN	7.64
3	Y	CA	59.30
3	Y	HA	4.77
3	Y	C	176.1
5	H	CA	56.70
5	H	HA	3.97
5	H	CB	30.07
5	H	HB3	3.20
5	H	HB2	3.11
5	H	C	176.39
8	H	CA	55.82
8	H	HA	4.57
8	H	CB	30.60
8	H	HB	2.98
8	H	C	173.92
9	H	N	120.44
9	H	HN	8.35
9	H	CA	56.17
9	H	CB	30.74
9	H	C	175.59
9	H	HA	4.65
9	H	HB	3.08
10	T	N	114.95
10	T	HN	8.12
10	T	CA	61.36
10	T	HA	4.12
10	T	CB	70.00
10	T	HB	4.02
10	T	CG2	21.20
10	T	HG2	1.20
10	T	C	172.91
11	M	N	120.41
11	M	HN	7.91
11	M	CA	58.74
11	M	C	175.61
12	M	N	123.73
12	M	HN	8.08
12	M	CA	52.83
13	P	CA	63.04
13	P	HA	4.38
13	P	CB	32.15

13	P	HB3	2.19
13	P	HB2	1.89
13	P	CG	26.98
13	P	HG	1.92
13	P	CD	50.57
13	P	HD3	3.77
13	P	HD2	3.65
14	S	N	114.78
14	S	HN	7.98
14	S	CA	55.28
14	S	HA	4.39
14	S	CB	65.53
15	P	CA	62.94
15	P	HA	4.40
15	P	CB	32.2
15	P	HB3	2.24
15	P	HB2	1.88
15	P	CG	26.92
15	P	HG	1.98
15	P	CD	50.69
15	P	HD3	3.70
15	P	HD2	3.64
15	P	C	177.97
16	E	N	115.59
16	E	HN	8.08
16	E	CA	55.18
16	E	HA	3.80
16	E	CB	31.00
16	E	HB3	2.18
16	E	HB2	2.11
16	E	CG	31.79
16	E	HG	2.63
16	E	CD	177.06
17	P	CA	63.35
17	P	HA	4.46
17	P	CB	32.27
17	P	HB3	2.32
17	P	HB2	1.95
17	P	CG	26.89
17	P	HG	1.91
17	P	CD	48.88
17	P	HD3	3.85
17	P	HD2	3.75
17	P	C	176.94
18	S	N	115.75
18	S	HN	8.43
18	S	CA	58.12
18	S	HA	4.34
18	S	CB	63.43
18	S	HB	3.80
18	S	C	172.88
19	S	N	117.04

19	S	HN	8.32
19	S	CA	58.28
19	S	HA	4.40
19	S	CB	63.82
19	S	HB	3.75
19	S	C	174.30
20	Q	N	122.5
20	Q	HN	8.30
20	Q	CA	53.35
20	Q	HA	4.71
20	Q	CB	29.13
20	Q	HB3	2.14
20	Q	HB2	1.98
20	Q	CG	33.29
20	Q	HG	2.40
20	Q	CD	180.45
20	Q	NE2	111.74
20	Q	HE21	7.56
20	Q	HE22	6.86
20	Q	C	174.09
21	P	CA	63.03
21	P	HA	4.80
21	P	CB	32.20
21	P	HB3	2.27
21	P	HB2	1.95
21	P	CG	27.12
21	P	HG	1.85
21	P	CD	49.15
21	P	HD1	3.80
21	P	HD2	3.70
21	P	C	177.07
22	S	N	116.72
22	S	HN	8.58
22	S	CA	58.66
22	S	HA	4.40
22	S	CB	63.87
22	S	HB3	3.93
22	S	HB2	3.84
22	S	C	174.28
23	D	N	121.85
23	D	HN	8.44
23	D	CA	54.32
23	D	HA	4.57
23	D	CB	41.03
23	D	HB3	2.64
23	D	HB2	2.51
23	D	CG	180.07
23	D	C	176.72
24	C	N	118.16
24	C	HN	8.32
24	C	CA	56.88
24	C	HA	4.52

24	C	CB	40.83
24	C	HB3	3.17
24	C	HB2	3.04
24	C	C	175.26
25	G	N	109.17
25	G	HN	8.50
25	G	CA	45.87
25	G	HA	3.86
25	G	C	174.86
26	E	N	120.24
26	E	HN	8.00
26	E	CA	55.83
26	E	HA	4.34
26	E	CB	33.13
26	E	HB	2.08
26	E	CG	35.9
26	E	HG	2.18
26	E	C	174.98
27	V	N	120.7
27	V	HN	8.09
27	V	CA	63.44
27	V	HA	4.15
27	V	CB	32.07
27	V	HB	2.14
27	V	CG1	20.85
27	V	HG1	0.97
27	V	CG2	18.72
27	V	HG2	0.84
27	V	C	176.19
28	I	N	121.5
28	I	HN	8.02
28	I	CA	60.15
28	I	HA	4.27
28	I	CB	38.59
28	I	HB	1.81
28	I	CG1	26.79
28	I	HG13	1.46
28	I	HG12	1.18
28	I	CG2	17.65
28	I	HG2	0.90
28	I	CD1	12.04
28	I	HD1	0.84
28	I	C	175.97
29	E	N	130.5
29	E	HN	8.97
29	E	CA	57.8
29	E	HA	4.35
29	E	CB	31.17
29	E	HB	2.18
29	E	HG	2.20
29	E	C	176.68
30	E	N	122.2

30	E	HN	8.37
30	E	CA	55.16
30	E	HA	4.35
30	E	CB	33.08
30	E	HB3	2.01
30	E	HB2	1.94
30	E	HG	2.28
30	E	C	173.97
31	C	N	123.94
31	C	HN	8.05
31	C	CA	52.70
31	C	HA	4.48
31	C	CB	40.60
31	C	HB3	2.85
31	C	HB2	2.80
31	C	C	174.95
32	P	CA	63.70
32	P	HA	4.47
32	P	CB	32.60
32	P	HB3	2.27
32	P	HB2	1.97
32	P	CG	26.91
32	P	HG	1.92
32	P	CD	50.70
32	P	HD3	3.82
32	P	HD2	3.70
32	P	C	175.71
33	I	N	120.41
33	I	HN	8.02
33	I	CA	59.45
33	I	HA	4.31
33	I	CB	39.37
33	I	HB	1.76
33	I	CG1	26.54
33	I	HG13	1.32
33	I	HG12	0.95
33	I	CG2	17.38
33	I	HG2	0.85
33	I	CD1	13.10
33	I	HD1	0.78
33	I	C	176.24
34	D	N	130.69
34	D	HN	9.00
34	D	CA	54.24
34	D	HA	4.74
34	D	CB	43.76
34	D	HB3	3.01
34	D	HB2	2.40
34	D	CG	184.11
34	D	C	178.05
35	A	N	124.42
35	A	HN	8.90

35	A	CA	53.60
35	A	HA	4.78
35	A	CB	19.32
35	A	HB	1.33
35	A	C	177.94
36	C	N	114.47
36	C	HN	7.65
36	C	CA	57.85
36	C	HA	4.25
36	C	CB	38.53
36	C	HB3	3.30
36	C	HB2	2.35
36	C	C	174.93
37	F	N	117.04
37	F	HN	8.06
37	F	CA	59.50
37	F	HA	4.08
37	F	CB	37.54
37	F	HB3	3.27
37	F	HB2	2.75
37	F	HD	7.16
37	F	C	174.6
38	L	N	120.89
38	L	HN	7.15
38	L	CA	53.64
38	L	HA	4.42
38	L	CB	40.87
38	L	HB3	1.72
38	L	HB2	1.46
38	L	CG	26.45
38	L	HG	1.17
38	L	CD	24.85
38	L	HD	0.84
39	P	CA	62.30
39	P	HA	4.33
39	P	CB	32.20
39	P	HB3	2.35
39	P	HB2	1.83
39	P	CG	27.65
39	P	HG	2.09
39	P	CD	49.5
39	P	HD3	3.84
39	P	HD2	3.63
39	P	C	177.95
40	K	N	118.96
40	K	HN	8.67
40	K	CA	59.32
40	K	HA	3.86
40	K	CB	31.67
40	K	HB3	1.89
40	K	HB2	1.75
40	K	CG	26.45

40	K	HG3	1.60
40	K	HG2	1.27
40	K	CD	28.35
40	K	HD	1.70
40	K	CE	41.41
40	K	HE	2.96
40	K	HZ	8.29
40	K	C	176.92
41	S	N	109.65
41	S	HN	7.57
41	S	CA	57.86
41	S	HA	4.23
41	S	CB	63.21
41	S	HB3	4.07
41	S	HB2	3.76
41	S	HG	4.74
41	S	C	174.09
42	D	N	125.71
42	D	HN	7.98
42	D	CA	54.52
42	D	HA	4.55
42	D	CB	43.90
42	D	HB3	3.01
42	D	HB2	2.40
42	D	CG	178.96
42	D	C	178.16
43	S	N	125.07
43	S	HN	8.95
43	S	CA	60.8
43	S	HA	4.17
43	S	CB	63.06
43	S	HB	3.95
43	S	C	174.5
44	A	N	124.1
44	A	HN	9.40
44	A	CA	51.75
44	A	HA	4.26
44	A	CB	19.34
44	A	HB	1.31
44	A	C	177.63
45	R	N	122.82
45	R	HN	7.65
45	R	CA	54.98
45	R	HA	4.16
45	R	CB	30.73
45	R	HB3	1.77
45	R	HB2	1.57
45	R	CG	27.52
45	R	HG3	2.28
45	R	HG2	1.87
45	R	CD	43.01
45	R	HD	2.98

45	R	NE	127.48
45	R	HE	6.75
45	R	CZ	159.21
45	R	C	174.80
46	P	CA	61.17
46	P	HA	4.52
46	P	CB	29.12
46	P	HB3	2.28
46	P	HB2	1.88
46	P	CD	50.49
46	P	HD3	3.92
46	P	HD2	2.94
47	P	CA	61.13
47	P	HA	4.39
47	P	CB	31.95
47	P	HB3	2.21
47	P	HB2	1.85
47	P	CG	26.99
47	P	HG	1.95
47	P	CD	48.88
47	P	HD3	3.75
47	P	HD2	3.59
47	P	C	177.36
48	D	N	120.73
48	D	HN	8.75
48	D	CA	54.92
48	D	HA	5.15
48	D	CB	39.64
48	D	HB3	2.89
48	D	HB2	2.44
48	D	CG	178.35
48	D	C	178.45
49	C	N	122.34
49	C	HN	8.03
49	C	CA	57.54
49	C	HA	4.25
49	C	CB	39.95
49	C	HB3	3.31
49	C	HB2	2.33
49	C	C	176.84
50	T	N	109.49
50	T	HN	8.57
50	T	CA	65.00
50	T	HA	3.63
50	T	CB	66.64
50	T	HB	4.27
50	T	CG2	22.18
50	T	HG2	1.27
50	T	C	176.75
51	A	N	124.26
51	A	HN	7.15
51	A	CA	53.98

51	A	HA	4.39
51	A	CB	19.63
51	A	HB	1.65
51	A	C	178.70
52	V	N	108.20
52	V	HN	7.01
52	V	CA	59.81
52	V	HA	4.69
52	V	CB	30.65
52	V	HB	2.59
52	V	CG1	21.14
52	V	HG1	0.94
52	V	CG2	18.44
52	V	HG2	0.92
52	V	C	176.06
53	G	N	106.92
53	G	HN	7.50
53	G	CA	46.22
53	G	HA3	4.01
53	G	HA2	3.88
53	G	C	174.07
54	R	N	113.66
54	R	HN	7.69
54	R	CA	49.45
54	R	HA	5.12
54	R	CB	28.06
54	R	HB3	1.80
54	R	HB2	1.51
54	R	CG	24.32
54	R	HG	1.43
54	R	CD	39.27
54	R	HD	3.41
54	R	NE	124.58
54	R	HE	9.27
54	R	CZ	159.054
54	R	C	174.22
55	P	CA	64.97
55	P	HA	4.09
55	P	CB	31.07
55	P	HB3	2.28
55	P	HB2	2.14
55	P	CG	26.95
55	P	HG3	2.03
55	P	HG2	1.98
55	P	CD	49.95
55	P	HD3	3.81
55	P	HD2	3.74
55	P	C	177.98
56	D	N	116.72
56	D	HN	9.55
56	D	CA	54.62
56	D	HA	4.14

56	D	CB	37.10
56	D	HB3	2.74
56	D	HB2	2.50
56	D	CG	181.70
56	D	C	175.48
57	C	N	115.11
57	C	HN	7.41
57	C	CA	56.90
57	C	HA	4.28
57	C	CB	43.6
57	C	HB3	3.30
57	C	HB2	3.02
57	C	C	172.65
58	N	N	118.00
58	N	HN	7.30
58	N	CA	52.25
58	N	HA	4.70
58	N	CB	37.68
58	N	HB3	3.28
58	N	HB2	2.19
58	N	CG	177.30
58	N	ND2	113.34
58	N	HD21	7.97
58	N	HD22	6.71
58	N	C	177.50
59	V	N	119.28
59	V	HN	8.06
59	V	CA	60.54
59	V	HA	4.34
59	V	CB	35.50
59	V	HB	1.91
59	V	CG1	20.05
59	V	HG1	0.83
59	V	CG2	20.05
59	V	HG2	0.78
60	L	N	118.32
60	L	HN	7.65
60	L	CA	51.42
60	L	HA	4.77
60	L	CB	43.54
60	L	HB3	1.64
60	L	HB2	1.28
60	L	CG	26.99
60	L	HG	1.70
60	L	CD1	23.78
60	L	HD1	0.95
60	L	CD2	25.39
60	L	HD2	0.89
62	F	CA	55.29
62	F	HA	5.29
62	F	CB	39.8
62	F	HB3	3.21

62	F	HB2	3.03
62	F	HD	7.52
62	F	HE	7.58
62	F	HZ	7.87
62	F	C	174.91
63	P	CA	63.03
63	P	HA	4.28
63	P	CB	33.04
63	P	HB3	2.17
63	P	HB2	1.74
63	P	CG	27.52
63	P	HG3	2.20
63	P	HG2	1.80
63	P	CD	50.49
63	P	HD3	4.10
63	P	HD2	2.95
63	P	C	177.44
64	N	N	120.25
64	N	HN	8.50
64	N	CA	53.02
64	N	HA	4.76
64	N	CB	39.5
64	N	HB3	3.10
64	N	HB2	2.90
64	N	CG	176.82
64	N	ND2	111.74
64	N	HD21	7.64
64	N	HD22	6.90
64	N	C	176.92
65	N	N	115.59
65	N	HN	8.41
65	N	CA	53.30
65	N	HA	4.97
65	N	CB	38.28
65	N	HB3	3.02
65	N	HB2	2.80
65	N	CG	178.04
65	N	ND2	111.09
65	N	HD21	7.57
65	N	HD22	6.77
65	N	C	175.89
66	I	N	111.58
66	I	HN	7.76
66	I	CA	60.60
66	I	HA	4.76
66	I	CB	38.50
66	I	HB	2.12
66	I	CG1	26.45
66	I	HG13	1.28
66	I	HG12	0.65
66	I	CG2	16.84
66	I	HG2	0.78

66	I	CD1	13.64
66	I	HD1	0.63
66	I	C	177.01
67	G	N	113.82
67	G	HN	8.70
67	G	CA	46.86
67	G	HA2	3.79
67	G	HA3	3.85
67	G	C	176.42
68	C	N	119.61
68	C	HN	9.01
68	C	CA	56.26
68	C	HA	4.39
68	C	CB	41.94
68	C	HB3	3.19
68	C	HB2	2.72
68	C	C	173.48
69	P	CA	61.75
69	P	HA	4.78
69	P	CB	32.20
69	P	HB3	2.26
69	P	HB2	2.00
69	P	CG	25.92
69	P	HG3	2.19
69	P	HG2	1.50
69	P	CD	49.95
69	P	HD	3.80
69	P	C	174.35
70	S	N	111.9
70	S	HN	8.26
70	S	CA	58.21
70	S	HA	4.4
70	S	CB	61.63
70	S	HB	3.8
70	S	C	173.14
71	C	N	121.53
71	C	HN	7.69
71	C	CA	53.35
71	C	HA	5.28
71	C	CB	43.24
71	C	HB3	3.03
71	C	HB2	2.94
71	C	C	170.8
72	C	N	117.04
72	C	HN	9.16
72	C	CA	52.06
72	C	HA	5.14
72	C	CB	41.94
72	C	HB3	3.18
72	C	HB2	2.89
72	C	C	172.78
73	P	CA	62.73

73	P	HA	4.77
73	P	CB	33.40
73	P	HB3	2.73
73	P	HB2	1.84
73	P	CG	28.05
73	P	HG3	2.91
73	P	HG2	2.32
73	P	CD	52.09
73	P	HD3	4.35
73	P	HD2	3.61
73	P	C	176.93
74	F	N	127.47
74	F	HN	9.32
74	F	CA	60.78
74	F	HA	4.21
74	F	CB	38.98
74	F	HB3	3.40
74	F	HB2	3.14
74	F	HD	7.23
74	F	HE	7.37
74	F	HZ	7.32
74	F	C	176.47
75	E	N	114.79
75	E	HN	9.27
75	E	CA	58.51
75	E	HA	3.88
75	E	CB	31.04
75	E	HB3	2.18
75	E	HB2	1.89
75	E	CG	37.13
75	E	HG3	2.39
75	E	HG2	2.33
75	E	CD	177.22
75	E	C	177.43
76	C	N	118.00
76	C	HN	7.43
76	C	CA	51.76
76	C	HA	4.42
76	C	CB	37.14
76	C	HB3	2.91
76	C	HB2	2.33
76	C	C	170.8
77	S	N	113.5
77	S	HN	6.98
77	S	CA	53.66
77	S	HA	4.85
77	S	CB	64.62
77	S	HB3	3.89
77	S	HB2	3.36
77	S	C	176.93
78	P	CA	63.7
78	P	HA	4.53

78	P	CB	31.4
78	P	HB3	2.28
78	P	HB2	1.99
78	P	CG	26.57
78	P	HG	1.70
78	P	CD	50.34
78	P	HD3	3.80
78	P	HD2	3.71
78	P	C	175.64
79	D	N	117.04
79	D	HN	7.90
79	D	CA	53.06
79	D	HA	4.63
79	D	CB	40.16
79	D	HB3	2.74
79	D	HB2	2.46
79	D	CG	179.85
79	D	C	175.9
80	N	N	122.18
80	N	HN	7.40
80	N	CA	51.57
80	N	HA	4.67
80	N	CB	38.11
80	N	HB	2.78
80	N	ND2	112.06
80	N	HD21	7.56
80	N	HD22	6.93
80	N	C	173.51
81	P	CA	62.9
81	P	HA	4.41
81	P	CB	32.30
81	P	HB3	2.28
81	P	HB2	1.93
81	P	CG	27.03
81	P	HG3	2.04
81	P	HG2	1.98
81	P	CD	50.89
81	P	HD3	3.85
81	P	HD2	3.68
81	P	C	176.67
82	M	N	123.14
82	M	HN	8.40
82	M	CA	55.3
82	M	HA	4.33
82	M	HB3	2.00
82	M	HB2	1.88
83	F	CA	56.28
83	F	HA	4.34
83	F	CB	39.23
83	F	HB3	2.74
83	F	HB2	2.71
83	F	C	176.38

85	P	CA	62.82
85	P	HA	4.45
85	P	CB	32.1
85	P	HB3	2.27
85	P	HB2	1.95
85	P	CG	24.85
85	P	HG	1.92
85	P	CD	50.27
85	P	HD3	3.80
85	P	HD2	3.70
85	P	C	175.46
86	S	N	115.45
86	S	HN	8.02
86	S	CA	55.28
86	S	HA	4.53
86	S	HB	3.68
87	P	CA	65.00
87	P	HA	4.40
87	P	CB	31.74
87	P	HB3	2.31
87	P	HB2	1.97
87	P	CG	27.26
87	P	HG	2.06
87	P	CD	50.96
87	P	HD	3.67
87	P	C	176.84
88	D	N	115.11
88	D	HN	7.86
88	D	CA	53.88
88	D	HA	4.58
88	D	CB	40.36
88	D	HB3	2.87
88	D	HB2	2.68
88	D	CG	180.56
88	D	C	176.86
89	G	N	108.69
89	G	HN	8.25
89	G	CA	45.06
89	G	HA2	3.66
89	G	HA3	4.26
89	G	C	174.7
90	S	N	118.51
90	S	HN	8.10
90	S	CA	57.67
90	S	HA	4.65
90	S	CB	62.22
90	S	HB	3.87
92	P	CA	63.40
92	P	HA	4.43
92	P	CB	32.18
92	P	HB3	2.23
92	P	HB2	1.91

92	P	CG	26.86
92	P	HG	1.98
92	P	CD	50.65
92	P	HD	3.69
92	P	C	176.41
93	N	N	121.53
93	N	HN	8.26
93	N	CA	54.01
93	N	HA	4.62
93	N	CB	40.93
93	N	HB	2.72
93	N	CG	176.42
93	N	ND2	112.06
93	N	HD21	7.42
93	N	HD22	6.98
93	N	C	176.73
94	C	N	118.82
94	C	HN	8.32
94	C	CA	54.14
94	C	HA	4.72
94	C	CB	41.02
94	C	HB3	3.26
94	C	HB2	3.00
94	C	C	173.92
95	S	N	118.32
95	S	HN	8.43
95	S	CA	56.26
95	S	HA	4.76
95	S	CB	63.24
95	S	HB3	3.88
95	S	HB2	3.78
96	P	CA	63.58
96	P	HA	4.43
96	P	CB	32.24
96	P	HB3	2.24
96	P	HB2	1.90
96	P	CG	27.06
96	P	CD	50.65
96	P	C	177.06
97	T	N	114.15
97	T	HN	8.18
97	T	CA	62.40
97	T	HA	4.50
97	T	CB	69.28
97	T	HB	4.17
97	T	C	174.27
98	M	N	122.5
98	M	HN	8.27
98	M	CA	55.27
98	M	HA	4.43
98	M	CB	32.98
98	M	HB	1.95

98	M	CG	31.62
98	M	HG	2.47
98	M	C	174.26
99	L	N	122.18
99	L	HN	7.88
99	L	CA	52.33
99	L	HA	4.76
99	L	CB	41.2
99	L	HB	1.68
99	L	CG	26.35
99	L	HG	1.68
99	L	CD1	23.32
99	L	HD1	1.06
99	L	CD2	25.09
99	L	HD2	0.93
100	P	CA	62.76
100	P	HA	4.41
100	P	CB	32.11
100	P	HB3	2.22
100	P	HB2	1.86
100	P	CG	27.04
100	P	HG	1.98
100	P	CD	50.49
100	P	HD3	3.78
100	P	HD2	3.66
100	P	C	176.00
101	S	N	116.72
101	S	HN	8.35
101	S	CA	55.95
101	S	HA	4.68
101	S	CB	63.93
101	S	HB	3.78
101	S	C	173.42
102	P	CA	62.77
102	P	HA	4.16
102	P	CB	34.14
102	P	HB3	2.36
102	P	HB2	2.08
102	P	CG	24.85
102	P	HG	1.91
102	P	CD	49.47
102	P	HD	3.53
102	P	C	177.18
103	S	N	117.04
103	S	HN	8.34
103	S	CA	58.52
103	S	HA	4.71
103	S	HB3	3.87
103	S	HB2	3.80
104	P	CA	63.55
104	P	HA	4.77
104	P	CB	32.20

104	P	HB3	2.25
104	P	HB2	1.99
104	P	CG	27.03
104	P	HG	1.95
104	P	CD	50.49
104	P	HD	3.69
104	P	C	176.93
105	S	N	115.43
105	S	HN	8.43
105	S	CA	58.21
105	S	HA	4.43
105	S	CB	63.24
105	S	HB	3.87
105	S	C	174.26
106	A	N	126.03
106	A	HN	8.27
106	A	CA	52.38
106	A	HA	4.16
106	A	CB	19.8
106	A	HB	1.38
106	A	C	173.63
107	V	N	119.28
107	V	HN	8.09
107	V	CA	62.06
107	V	HA	4.10
107	V	CB	32.90
107	V	HB	2.06
107	V	CG1	20.35
107	V	HG1	0.93
107	V	CG2	20.35
107	V	HG2	0.88
107	V	C	176.35
108	T	N	118.96
108	T	HN	8.27
108	T	CA	61.75
108	T	HA	4.33
108	T	CB	69.65
108	T	HB	4.11
108	T	HG2	1.18
108	T	C	174.07
109	V	N	124.42
109	V	HN	8.19
109	V	CA	59.80
109	V	HA	4.43
109	V	HB	2.09
109	V	HG2	0.91
110	P	CA	62.73
110	P	HA	4.42
110	P	CB	33.46
110	P	HB3	2.46
110	P	HB2	2.00
110	P	CG	25.43

110	P	HG	1.89
110	P	CD	50.34
110	P	HD	3.50
110	P	C	175.45
111	L	N	127.64
111	L	HN	9.35
111	L	CA	55.9
111	L	HA	4.42
111	L	CB	42.27
111	L	HB3	1.78
111	L	HB2	1.65
111	L	CG	26.92
111	L	CD1	24.58
111	L	HD1	1.02
111	L	CD2	23.61
111	L	HD2	0.94
111	L	C	176.95
112	T	N	119.28
112	T	HN	7.64
112	T	CA	62.75
112	T	HA	4.13

Appendix B

RESIDUE	$^{3}J_{H^{N}H^{\alpha}}$ COUPLING in Hz
Ser 19	4.7
Asp 23	5.5
Cys 36	4.3
Leu 38	4.8
Lys 40	2.8
Ser 41	6.0
Asp 42	5.4
Ser 43	2.6
Arg 45	3.3
Asp 48	3.5
Cys 49	2.4
Thr 50	3.9
Ala 51	5.1
Val 52	8.7
Arg 54	5.9
Asp 56	3.9
Cys 57	7.4
Val 59	5.6
Ile 66	8.2
Cys 68	1.8
Cys 71	6.0
Cys 72	7.2
Phe 74	2.0
Glu 75	3.8
Asp 79	8.1
Ser 90	4.2
Asn 93	5.5

Appendix C

RESIDUE	$^{3}J_{H^{N}H^{\alpha}}$ COUPLING in Hz
Tyr 2	7.7
Pro 13	4.9
Pro 15	5.4
Ser 18	8.7
Ser 19	4.7
Gln 20	8.6
Pro 21	4.8
Asp 23	5.5
Ile 28	8.6
Glu 29	8.1
Pro 32	5.6
Asp 34	6.9
Ala 35	8.2
Cys 36	4.3
Leu 38	4.8
Pro 39	4.9
Lys 40	2.8
Ser 41	6.0
Asp 42	5.4
Ser 43	2.6
Arg 45	3.3
Pro 46	5.6
Pro 47	5.1
Asp 48	3.5
Cys 49	2.4
Thr 50	3.9
Ala 51	5.1
Val 52	8.7
Gly 53	6.8
Arg 54	5.9
Asp 56	3.9
Cys 57	7.4
Asn 58	7.5
Val 59	5.6
Pro 61	5.7
Phe 62	8.5
Pro 63	4.6
Ile 66	8.2
Gly 67	7.1
Cys 68	1.8
Pro 69	5.6
Cys 71	6.0

Cys 72	7.2
Pro 73	5.6
Phe 74	2.0
Glu 75	3.8
Pro 78	5.3
Asp 79	8.1
Asn 80	5.7
Thr 84	7.9
Pro 85	4.8
Pro 87	4.4
Asp 88	7.2
Ser 90	4.2
Pro 91	5.4
Pro 92	4.6
Asn 93	7.4
Leu 99	7.7
Pro 100	4.8
Ser 101	8.5
Pro 102	4.9
Ser 103	8.4
Pro 104	4.4
Ser 105	8.0
Pro 110	5.1

Appendix D

RESIDUE	ATOM	RESIDUE	ATOM	H_Bond in nm
Cys 71	O	Phe 62	HZ	2.0
Asp 42	O	Ser 86	HN	2.0
Asp 42	O	Ser 86	N	3.0
Asp 34	O	Phe 37	HN	2.0
Asp 34	O	Phe 37	N	3.0
Pro 47	O	Cys 49	HN	2.0
Pro 47	O	Cys 49	N	3.0
Lys 40	O	Arg 45	HN	2.0
Lys 40	O	Arg 45	N	3.0
Pro 96	O	Asn 93	HN	2.0
Pro 96	O	Asn 93	N	3.0
Ile 33	O	Phe 37	HN	2.0
Ile 33	O	Phe 37	N	3.0
Asp 56	O	Val 59	HN	2.0
Asp 56	O	Val 59	N	3.0

Appendix E

RESIDUE	RDC in Hz
Ser 19	-3.5
Ser 22	-6.9
Asp 23	-0.4
Glu 29	-20.1
Ile 33	-0.4
Asp 34	-15.0
Cys 36	-7.15
Leu 38	2.0
Lys 40	2.8
Ser 41	8.2
Asp 42	13.79
Ser 43	3.3
Ala 44	2.8
Arg 45	1.0
Asp 48	-5.6
Cys 49	12.4
Thr 50	6.7
Ala 51	-1.5
Val 52	10.1
Gly 53	-6.2
Arg 54	2.4
Asp 56	14.2
Cys 57	1.2
Asn 58	-3.8
Val 59	-3.8
Leu 60	-6
Asn 65	-2.1
Ile 66	6.5
Gly 67	-7.9
Cys 68	6.7
Ser 70	10.0
Cys 71	2.0
Cys 72	3.9
Phe 74	4.3
Glu 75	1.6
Cys 76	5.6
Ser 77	-3.4
Asp 79	7.4
Met 82	-7.7
Ser 86	-1.9
Ser 90	-4.0
Ser 95	-14.0

Met 98	-0.9
Ser 105	-3.8
Ala 106	-7.5
Val 107	-9.3
Thr 108	-7.0
Val 109	-7.2
Leu 111	-8.2
Thr 112	-3.4

Erklärung

Hiermit erkläre ich, das ich die vorliegende Arbeit selbständig angefertigt, und keine Hilfsmittel, außer den angegebenen, benutzt habe.

Regensburg, 21-11-2011

Silvia De Sanctis