AUREMOL-QTA, a program package for NMR based automated recognition and characterization of local and global conformational changes in proteins induced by ligand binding as external perturbation

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

vorgelegt von

Wilhelm Massimiliano Malloni

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. Chrisoph Oberprieler
Erstgutachter: Prof. Dr. Dr. Hans Robert Kalbitzer
Zweitgutachter: Prof. Dr. Wolfram Gronwald
Drittprüfer: Prof. Dr. Rainer Merkl

Abstract

Among all existing techniques for protein structure determination, NMR spectroscopy has the advantage to provide the most complete characterization of molecular structures in solution. On the other hand, the low sensitivity of NMR limits the available signal-to-noise ratio. This often leads to the disappearance of cross peaks and to the misinterpretations of noise peaks as peaks originating from the protein under consideration. These properties hamper the determination of the consistency between the two investigated spectra.

In this work, a completely new quality control (AUREMOL-QTA) package has been developed in order to automatically infer structural conformational changes from spectral modifications (in a set of investigated spectra). These differences may be induced by alterations of the external conditions (e.g. temperature, pressure, pH and ligand binding) as well as (partial) protein denaturation.

The complex task of manually collecting and interpreting vicinity relationships changes (through space and bonds) in order to extract structural modifications of the molecule is time demanding. The developed AUREMOL-QTA package facilitates this investigation by means of multi-dimensional NMR spectra and it is applicable to three main possible sceneries:

1. Quality control of a protein sample and assessment of the intact three-dimensional structure of the protein.
2. Automated identification of ligand binding sites by means of NMR spectroscopy.
3. Identification of structural changes (in the atomic resolution) induced by external perturbations such as pressure, pH and temperature.

The AURMOL-QTA begins with a pre-processing step where one (or more) reference spectra of the target protein and one or more test spectra are normalized (spectral width, offset, receiver gain and the number of scans) to be properly compared. A possible spectrum shift is evaluated and if it has occurred it is corrected. The simulated spectrum of

a protein can be used as an adjunctive reference spectrum and to recognize peak multiplets in the experimental spectra. If there are more reference spectra only the common peaks are retained and used for the requested comparison.

The routine continues with a low-level analysis involving a peak feature collection (volume, line width, chemical shift, cross-correlation in the time domain and shape) with a consequent association of cross peaks among the spectra. A mid-level analysis relies on the calculation of feature-related Bayesian probabilities that the associated peaks represent the same signal. It facilitates the detection of missing signals and the identification of feature differences between the compared spectra. The ratio of matching peaks is computed in order to quantitatively determine the similarity of the investigated spectra. A high-level analysis allows the identification of structurally altered parts of the molecule, mapping the feature variations and computing the fraction of residues involved in the modifications. In particular, if NOESY and TOCSY spectra are investigated the symmetrical properties are exploited in order to perform an adjunctive pattern analysis of the residues.

The method has been successfully tested on HSCQ spectra (pressurized with xenon) and on HSQC-TROSY spectra (with high pressure and temperature variations) of the human prion protein (*hu*PrP^{C}). It has been also used to compare the NOESY spectrum of the wild HPr protein from *Staphylococcus aureus* with the mutant (H15A) of the same protein and with the artificially denatured (partially and totally) ones.

Acknowledgements

- My Ph.D. supervisor Prof. Dr. Dr. Hans Robert Kalbitzer for his patience and his praiseworthy support demonstrated during the whole work. He has taught me how to become a scientist.
- Prof. Dr. Elmar W. Lang for his theoretical support given during the development of many methods of the project. Inestimable was also his care of me in some difficult moments of my life.
- Prof. Dr. Werner Kremer for his professional support during the analysis of the biological parts of the project.
- Prof. Dr. Wolfram Gronwald for his help and his suggestions particularly crucial at the beginning of this work.
- Prof. Dr. Mauro Barni for his useful suggestions regarding theoretical aspects of the FFT.
- Donaubauer Harald for his excellent support during the programming stage.
- Harsch Tobias for his help during the project.
- Dr. Trenner Jochen and Dr. Brunner Konrad for their support at the beginning of the project.
- Dr. Claudia Munte for providing the data of the HPr from *Staphylococcus aureus.*
- Marisa Barbosa de Aguiar for her contributions about the human prion protein.
- My wife Silvia for her love and for giving me a chance to have a better life.

Malloni Wilhelm Massimiliano

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3SA	**3**(three) **s**tep **a**lgorithm
acqus	**Acqu**isition state file
AUREMOL-FFT	**Auremol** Hyper complex **F**ast **F**ourier **T**ransformation
AUREMOL-QTA	**AUREMOL Q**uality **T**est **A**nalysis
AUREMOL-SSA/ALS	**AUREMOL S**ingular **S**pectrum **A**nalysis / **A**utomatic **L**inear **S**pline
BC_mod	**B**aseline **c**orrection **mod**e
DECIM	**Decim**ation Factor
DSPFVS	**D**igital **Sp**ectrometer **V**ers**i**on
FCOR	**F**irst data point **cor**rection
FFT	**F**ast **F**ourier **T**ransformation
FID	**F**ree **I**nduction **D**ecay
FWHM	**F**ull **W**idth at **H**alf **M**aximum
GB	**G**aussian **B**roadening
GRPDLY	**Gr**ou**p de**l**ay**
GSHIFT	**G**lobal **Shift** Analysis
KS-Test	**K**olmogorov-**S**mirnov **T**est
LB	**L**ine **B**roadening
LSHIFT	**L**ocal **Shift** Analysis
MSA	**M**ultiplets **S**earch **A**nalysis
NC_proc	Intensity scaling factor

NLST	**L**i**s**t of **n**eighbors
NS	**N**umber of **Scan** performed during the acquisition
OFF	Spectrum **Offset**
PKNL	Nonlinear phase correction
procs	**Proc**essing state file
RG	**R**eceiver **G**ain
SI	**Si**ze of real spectrum
SSB	**S**ine **b**ell **s**hift
SW	**S**pectral **W**idth of the acquired spectrum
TD	Size of the **t**ime **d**omain dataset

1

Introduction

1.1 Proteins

Proteins are biological macromolecules performing essential functions in the organisms such as binding, catalysis, molecular switching and serving as structural components of living systems. They can bind to other macromolecules as the DNA. This function is strictly connected to the shape complementarity and polarity interactions between the molecules. Some others are protein enzymes that catalyze chemical reactions. Conformational changes due to pH, temperature and pressure variations can be used as molecular switching to control cellular processes (e.g. GTPase Ras) [Bourne et al., 1990, Spörner et al., 2010]. Structural proteins are instead responsible of biomaterials constitution such as silk, collagen and keratin.

Proteins are made up of amino acids that are linked together by peptide bonds [Fischer, 1903] forming a polypeptide chain. The chemical diversity of the side chains of such amino acids, the flexibility of the polypeptide chain and the different way of folding in a tertiary structure, influence the functional diversity of the proteins. In particular, twenty canonical amino acids are defined as the building blocks (residues) of proteins whose size can vary from a few to thousands of amino acids. The main chain of a protein (backbone) is constituted as following:

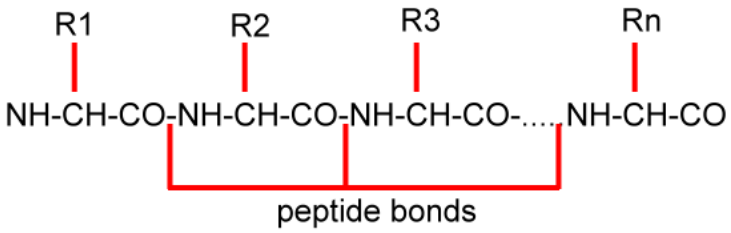


Fig. 1.1 Polypeptide chain: main (NH-CH-CO) and side chain (R) of proteins. Every amino acid possesses a different side chain (R). The number of residues can vary from a few dozens to several thousands.

The different nature of the side chains involves varying interactions between the residues and with the water. Hydrophobic amino acids (i.e. alanine, valine, phenylalanine, proline, leucine and isoleucine) have the tendency to avoid the water and to pack in the inner side of the tertiary structure, whereas the hydrophilic ones (i.e. arginine, aspartic acid, glutamic acid, serine, cysteine, asparagine, glutamine and histidine) are prone to make hydrogen bonds to one another and to water. The amphipathic residues (threonine, lysine, tyrosine, methionine and tryptophan) have both properties.

The sequence of those amino acids represents the primary structure of the protein. The regular patterns, known as alpha helices and beta sheets due to hydrogen-bonding interactions between the amino and the carboxyl group of the backbone, represent the secondary structures. The packing or folding of such elements defines the tertiary structure, while the association of several folded proteins is called quaternary structure. Therefore, the primary structure determines the tertiary structure that in turn determines the function of the protein. The polypeptide chain needs to have a stable tertiary structure in order to execute the protein functions. As a study case, the prion protein sometimes reveals a misfolded tertiary structure typically affecting the protein function.

Roughly speaking there are three main protein classes: fibrous, membrane and water soluble proteins. Studies on protein physics are manly conducted over the latter group because soluble proteins are easily isolatable and separable. During this project, multi-dimensional NMR spectra of soluble proteins have been used but there are no preclusive reasons to apply the same procedures on fibrous or membrane proteins spectra.

1.2 Nuclear Magnetic Resonance

The nuclear magnetic resonance [Rabi et al., 1938] experiment exploits the magnetic properties of the nuclei in order to provide structural information. The spinning nucleus is charged and it generates a local magnetic field possessing the magnetic moment $\vec{\mu}$. This magnetic moment depends on two main entities: the spin angular momentum $\vec{J}$ and the magnetogyric ratio γ (characteristic for each nucleus). In particular,

$$\vec{J} = \frac{h\,\vec{I}}{2\pi} \tag{1.1}$$

where $\vec{I}$ represents nuclear spin angular momentum quantum number and *h* is the Planck's constant with the relation:

$$\vec{\mu} = \gamma\,\vec{J} \tag{1.2}$$

implying an increasing magnetic moment proportional to larger γ and $\vec{I}$. In particular, the magnitude of $\vec{I}$ is given by

$$|I| = \frac{h}{2\pi}\sqrt{I^2 + I}. \tag{1.3}$$

In presence of a strong external magnetic field B_0 particles with spin I = $\frac{1}{2}$ align themselves, with respect to the B_0 field, only in well-defined arrangements described by the magnetic quantum numbers $m_I = \pm\frac{1}{2}$. The effect of energy splitting in presence of an external static magnetic field is called Zeeman effect. The torque excited by the external field produces a precessional motion of the nuclei. The precession happens with an angular frequency ω_0 called Larmor frequency. Each spin has a characteristic resonance frequency (Larmor frequency) in a given external magnetic field, B_o. If all the protons in a molecule had exactly the same precession frequency, the NMR technique would provide a unique

peak in the spectrum representing all of the protons. Some slight differences in resonance frequencies depending on the chemical environment (electron cloud) of the nucleus overcome this problem. As a consequence, the effective magnetic field B_{eff} experienced by each nucleus is directly related to an electron shielding constant. In particular, the external field induces a local field that is opposed to the former (diamagnetic shielding) that is

$$B_{eff} = B_0 + B_{loc} = (1 - \sigma)B_0 \quad (1.4)$$

with σ the shielding constant.

The chemical shift, which corresponds to the difference in resonance frequency of each nucleus, is thus one of the most accurate NMR parameters. In order to standardize the value of such shift independently on the field strength, it is reported in ppm (parts per million) as the difference between the resonance frequency of the considered nucleus and that one of a nucleus in a reference compound. The chemical shift is perturbed by the magnetic field induced by the orbiting electrons. As reported in eq. 1.4, the effective magnetic field B_{eff} at a given nucleus is the sum of the external field B_0 and the local magnetic field B_{loc}. In particular, a signal falling in the right side of the spectrum represents a relatively shielded nucleus (upfield) whereas that signal lying in the left side is related to a relatively de-shielded atom (downfield), as typically experienced by a proton bonded to nitrogen. A less shielded proton is more exposed to the external magnetic field, thus its resonant frequency is increased shifting the ppm position downfield.

The nuclear spins interact through chemical bonds giving rise to the so called scalar or J-coupling mediated by the electrons surrounding the spins. It is a through-bond interaction where one nucleus perturbs the spins of the electrons that in turn perturb the neighboring magnetic nuclei. Unlike the chemical shift, it is independent on the field strength and it provides information about the chemical connectivity between nuclei and on the conformation of rotatable bonds. This interaction is revealed by a splitting of the signal at both coupled spins. The constant J (measured in Hertz) is exactly the frequency separation between those splitting lines, also known as multiplet structure. The magnitude of J increases with the number of bonds separating the nuclei. In equilibrium the magnetization

$\vec{M}$ (that is the sum of the individual magnetic moments $\vec{\mu}$) is oriented parallel to the external field.

After irradiation with a weaker radiofrequency field B_1, that is perpendicular to the B_0 field, the magnetization (detected in the z-direction) is tuned into the x-y plane. After the perturbation, the equilibrium state is reestablished by relaxation. The return of the z-component of the magnetization M_z to the equilibrium value M_0 is described by the time constant T_1 (longitudinal relaxation time). The return to its equilibrium value of the transverse component $M_{x,y}$ is described by the transversal relaxation time T_2. The FID (free induction decay) time domain signal acquisition is performed during this relaxation.

1.3 NMR Experiments

1.3.1 1D-NMR Experiment

The 1D NMR experiment [Purcell et al., 1946] consists of two parts: the preparation and the detection. During the first step a 90° pulse is applied implying an excitation of all the nuclei contained in the sample. All the nuclei of a certain type (e.g. ^{1}H, ^{15}N and ^{13}C) are excited separately inducing a decaying signal into the coil (due to the transverse relaxation) that is recorded during the detection period. Depending on the chosen isotope the frequency range varies drastically. A schematic representation of this type of experiment is shown in Fig.1.2.

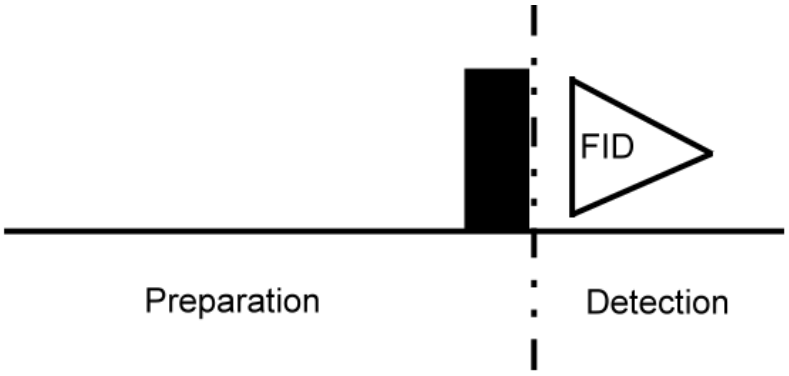


Fig. 1.2 1D NMR Experiment: it consists in two separate parts, the preparation and the detection.

The Fourier transform of such time signal creates the NMR spectrum where each intensity peak corresponds to a nucleus in the protein that precesses at that specific frequency ω around the external magnetic field. In order to increase the signal to noise ratio of the recorded signals the experiment is repeated several times adding them up.

1.3.2 2D-NMR Experiment

Generally, only some structural information can be extracted from one-dimensional (1D) NMR spectra such as the chemical shift, the peak intensity and through bond neighbors (via J coupling). For protein structure determination such experiments are not sufficient because of signals overlaps, thus multi-dimensional spectroscopy need to be performed. The two dimensional experiment [Aue et al., 1976] inherits the preparation and the detection step from the one dimensional case with two new items: the evolution and the mixing periods (see Fig.1.3). The evolution part is performed incrementing the time delay t_1 by a fixed number of steps. The recorded signal is modulated as a function of the indirect evolution period.

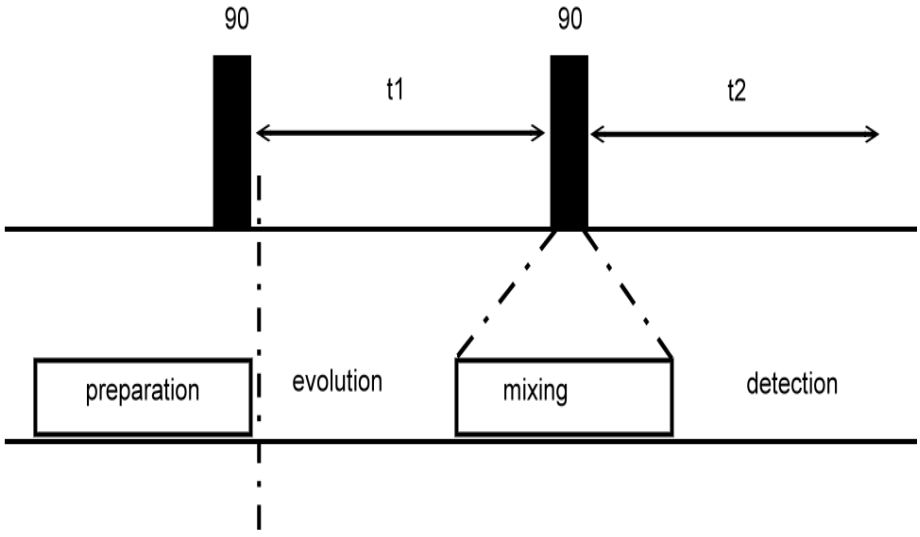


Fig. 1.3 General schema of the two-dimensional NMR experiment: the evolution and the mixing periods are additional with respect to the one-dimensional experiment.

During the mixing time step a combination of time delays and radio frequency pulses is applied. This combination induces a magnetization transfer from the first nucleus to the second one. This magnetization transfer is obtained through-bond (scalar coupling) or space interactions (dipolar coupling) depending on the experiment type. For example, the cross-peaks of a COSY experiment identify through-bond scalar correlations, while in the NOESY experiment cross-peaks represent through-space correlations. In case of molecules with more than 100 residues the spectral overlap and line width increase are considerable problems. To overcome these limitations the 3D NMR experiment [Oschkinat et al., 1988] has been proposed. In this experiment two evolution time periods (t_1 and t_2) are independently incremented with two different mixing periods and an additional detection period t_3. The same principle is recursively applied in order to increase the dimensionality of the measured spectrum obtaining high dimensional NMR data.

Moreover, the two-dimensional (2D) NMR spectra can be classified into homonuclear (magnetization transfer between two nuclei of the same type, as two protons) and heteronuclear (magnetization transfer between two different types of nuclei, as proton and nitrogen). The formers are typically COSY and TOCSY spectra where the magnetization is spread through bonds and thus it is constrained within each residue or NOESY and ROESY spectra characterized by through space distance interactions overcoming residual limitations. They all have cross peaks with symmetrical properties around the diagonal. The NOESY spectrum reveals some additional signals with respect to the COSY and TOCSY spectra and it is the most interesting one since it represents protons that are close in the three-dimensional molecular structure but not in the bonding network. The intensity (volume) of a NOE signal determines such distances. These three types of experiment can be used together to assign quite correctly NMR spectra of small proteins. Some of the heteronuclear spectra are instead HSQC, HMBC and HMQC and HSQC-TROSY. Dealing with larger molecule with a wide signal overlap involves the acquirement of other types of 2D experiments and the increase of the dimensionality [Cavanagh et al., 1996]. The three-dimensional structure of a protein can be determined extracting distance and structural restraints from cross peaks observed in the above mentioned experiments.

1.4 Protein Structure Determination

There are two main experimental methods used to investigate the tertiary (3D) structure of the proteins, namely X-ray [Bragg, 1907] crystallography and NMR spectroscopy. The former allows the recording of diffraction data of a protein crystal used to determine the electron density distribution. It performs the fitting of diffraction data to the electron density maps yielding a unique optimal structure where the R factor defines the quality of the fitting. The latter method consents essentially measures of the distances between atoms by means of a set of multi-dimensional magnetic resonance spectra of nuclear spins in the protein and it calculates the positions of the atoms from these data. In accordance to the protein data depository actually 80% of the known structures have been derived by the first technique, 15% by NMR and 5% by other methods. Both techniques reveal specific limitations. X-ray crystallography requires protein crystallization that cannot be always guaranteed. On the other hand, NMR can be carried out only for small molecules.

The NMR limit of 30 kDa due to the fast relaxation of spins and the complex overlap of signals is obsolete [Wüthrich, 1990]. In the past few years the molecular mass limit has been increased up to 70 kDa using isotope labeling techniques, triple-resonance and heteronuclear experiments. Today all the mass limits are overcome by the Transverse Relaxation-Optimized Spectroscopy (HSQC-TROSY) experiment [Pervushin et al., 1997]. The molecular mass is strictly related to the molecular rotational correlation time τ_C that in turn is affecting T_2 relaxation time, yielding a line broadening effect on the spectra of larger proteins.

In liquid-state NMR molecules are studied in solution that better represents native-like conditions allowing the evaluation of denatured-protein folding intermediates and transition states. Moreover, it provides information about internal motions as the presence of more possible conformations for the same flexible loops. Fluorescence spectroscopy can be applied to detect such motions as well, but it is limited by a small number of sites within the protein.

Some theoretical approaches have been developed as complementary methods for determining the protein structure. In particular, the native folding of the protein can be directly found using optimization algorithms that minimize the potential energy of the

structure. Comparative methods are also generally applied with the assumption that similarity between sequences of amino acids implies similar structures. They are known as homology modeling approaches [Chothia et al., 1986]. Additional techniques do not simply align the primary sequence but also similar structural domains.

1.4.1 Protein Structure Determination from NMR data

As a starting point for protein structure determination from NMR data, each resonance in the spectrum needs to be associated with a specific nucleus in the protein, in the so called resonance assignment. A strong relation between cross peaks and atomic structure of molecules is derivable from NMR spectra. Each resonance in the spectrum has a chemical shift (position) in ppm, a splitting pattern (single or multiple peaks) and an intensity (volume). In order to obtain a unique assignment, protons must be correlated to other protons or spins in the molecule either by bond relationships (scalar or J couplings) or by space relationships (dipolar coupling).

The most interesting parameter for protein structure determination from NMR spectra is the NOE effect, detectable between protons in a range not larger than 5 $\dot{A}$. It is directly extracted from NOESY spectra where the intensity of the cross peaks is proportional to the inverse sixth power of the distance between the involved spins. Given a known reference distance d_{ref} of a methylene group or an aromatic ring, all the other proton distances d_i are approximated by exploiting the intensity $V_i$of the peaks as follows:

$$d_i = d_{ref} \sqrt[6]{\frac{V_{ref}}{V_i}} \tag{1.5}$$

The distance information alone is not sufficient for a reliable structure determination, thus it must be completed by conformations derived from the known covalent structure. A resonance in the spectrum can be a single peaks or it can split in a multiplet pattern. The J coupling between two or three-bonds distant protons is typically observable in COSY spectra where the J value between peaks of a multiplet structure (i.e. H^N- H^α) can be

measured. The number and the equivalence of other interacting proton groups can be in fact extracted from multiplet structures. Moreover, the magnitude of the coupling constants reveals the geometric relationship of the angle bonds connecting the protons. Three-bond relationships are valuable since the $3J_{H^N H^\alpha}$ constant is directly related to the dihedral angle between the bonds (H^N-N-C^α-H^α), providing direct information about the secondary structure. The relationship between the J constant and the dihedral angles is described by the Karplus equation:

$$^3J = Acos^2\theta + Bcos\theta + C \tag{1.6}$$

where A, B and C are constant values depending on the nuclei considered and Θ represents the dihedral angle [Karplus, 1959]. Actually, many other experiments allow measurements of J coupling in isotope labeled samples.

The above mentioned parameters (e.g. chemical shifts, NOE, J constant) and some others (e.g. RDC, hydrogen bonding) build a set of restraints used for structure calculation. In particular, the chemical shifts are specifically used for predicting secondary structures since they exhibit regular patterns for specific secondary structures [Wishart et al., 1995, Wang et al., 2001]. The other parameters penalize directly the energy function of a specific conformation if they fall outside the boundaries defined by the NMR data, defining a violation case. In addition, specific J and NOE values correspond to certain secondary structures (small J and large NOE for α-helices; large J and weak NOE for β-sheets) [Wüthrich, 1986]. Hydrogen bonds stabilize the structure [Wagner et al., 1983] and are observable in deuterated HSQC spectra, whereas the accuracy is improved by residual dipolar coupling (RDC) orientational restraints [Clore et al., 1998]. This latter provides information about the angle relative to the external magnetic field and it can be obtained from decoupled HSQC spectra. The first step for structure determination by distance geometry algorithm [Havel, 1991] combined with rMSD (restraint Molecular Dynamics) [Güntert, 1998] is the generation of a group of conformations which satisfy such restraints. An energy minimization of the force field potential is then applied where a simulated annealing algorithm is necessary in order to avoid local minima. The quality of the fitting of such ensemble of low energy conformations is defined by the root mean square

deviation (RMSD) [Renugopalakrishnan et al., 1991, Hyberts et al., 1992]. The violations must not be simply discarded but used to correct possible estimation errors.

1.4.1.1 AUREMOL

The 12-year-old Auremol software [Gronwald et al., 2004] borne trough the collaboration between the department of Biophysics of the University of Regensburg and the NMR manufacturer BRUKER is one of the existing software for automated protein structure determination from NMR data. In particular, the AUREMOL software relies on AURELIA [Neidig et al., 1995] software package. Both were successfully used for the determination of three dimensional protein structures. Starting from a measured spectrum, it is possible through the AUREMOL Bayesian peak picking [Antz et al., 1995] to find all the resonances of interest. All these signals are automatically segmented [Geyer et al., 1995] to obtain volume information and used by the KNOWNOE module [Gronwald et al., 2002] in order to assign peaks of NOESY spectra when a chemical shift list is available. The REFINE module [Trenner, 2006] is used to automatically reproduce distance restraints that are required for a protein molecular dynamics (MD) [De Laplace et al., 1951, Alder et al., 1957, Gibson et al., 1960] simulation. Once the MD has been performed, the RFAC module [Gronwald et al., 2000] is applied to validate the obtained structures through the RELAX module [Görler et al., 1997, Ried et al., 2004, Görler et al, 1999a].

The whole AUREMOL toolbox was originally written in ANSI-C code. The need of a more powerful environment was evident during the project implementation: the lines of source code, the complexity of the algorithms and the necessity of graphics tended to the best possible solution, C++. Using this popular language it is manageable the use of objects instead of calling functions with lots of parameters. Each object has its own methods, constructors and destructors giving to the programmer the impression that they are transient. This has the beautiful advantage of an easy code handling and managing. In addition to the C++ code, the Qt [Molkentin, 2007] framework has been used in order to write powerful GUI (“Graphical user interface”) application. Another noteworthy aspect is the possibility to write multithreading routines using some utilities of the Qt framework. Considering the actual market trend of multi-core processors it is possible to handle long

time consuming calculation with relatively cheap machines. In this type of calculation each core of the CPU has its own task speeding up the whole calculus. This imposes other necessary operations like semaphores preventing data mishandling or deadlocks. In particular, this has been very useful in case of multiple spectra analysis.

1.5 Fast Fourier Transform

The aim of this paragraph is not to explain the Fourier transformation fundamentals (from the shift theorem to the Parseval's one) but to demonstrate the use of this beautiful mathematical concept during this project. In fact, the AUREMOL-QTA routine permits a hybrid (time-frequency) analysis in order to look for the features of the given spectrum. After the application of a radio frequency pulse RF on an NMR sample immersed in a strong static magnetic field B_0, the individual spins are instantly focused and they start to precess in a synchronized manner in the x-y plane. The x-component after a RF pulse with $\left(-\frac{\pi}{2}\right)_x$ can be written as follow:

$$M_x = M_0 \sin(2\pi\nu_0 t) e^{-t/T_2} \tag{1.7}$$

$$M_y = -M_0 \cos(2\pi\nu_0 t) e^{-t/T_2} \tag{1.8}$$

where M_x are the x and y component of the net magnetization vector **M.** M_0 is the magnitude of **M** while ν_0 is the Larmor frequency and T_2 is the transverse relaxation. Considering (1.7) and (1.8) together with the Euler's formula it is possible to write:

$$M_{xy}(t) = -M_0 \left[\cos(2\pi\nu_0 t) e^{-t/T_2} - i\sin(2\pi\nu_0 t) e^{-t/T_2}\right]$$

$$= -M_0 \, e^{-\frac{t}{T_2}} \, e^{-i2\pi\nu_0 t} \tag{1.9}$$

where the acquired time domain signal s(t) is proportional to $M_{xy}(t)$. As soon as they lose their coherence, the magnetization of the nuclei in the sample begin to decay at different rates that in the frequency domain is translated as different line widths. The signal of all spins gives the signal s(t) called free induction decay (FID). It means that the signals of all the nuclei in the sample can be measured simultaneously. Their individual resonance frequencies can be recovered by the FFT [Briggs et al., 1995; Kauppinen et al., 2002; Zonst, 2003]. Using the angular frequency $\omega_0 = 2\pi\nu_0$ the forward Fourier transformation is written as following:

$$S(\omega) = \int_{-\infty}^{\infty} s(t) e^{-i\omega t} dt \tag{1.10}$$

whereas the backward Fourier transformation is:

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) e^{i\omega t} d\omega \tag{1.11}$$

The Fourier transformation of an exponential decay of a cosine modulated function (see eq. 1.9) is a Lorentzian function representing the peak shape in the frequency domain. The transformation is expressed as following

$$S(\omega) = -\int_{0}^{\infty} M_0 (e^{-\frac{t}{T_2}} e^{-i\omega_0 t}) e^{-i\omega t} dt \tag{1.12}$$

Considering the initial sampling delay t_0 the equation becomes

$$S(\omega) = -\int_{0}^{\infty} M_0 e^{(-i\omega_0 - \frac{1}{T_2})(t + t_0)} e^{-i\omega t} dt \tag{1.13}$$

$$= -M_0 e^{(-i\omega_0 - \frac{1}{T_2}) t_0} \int_{0}^{\infty} e^{(-i\omega_0 - \omega - \frac{1}{T_2})t} dt$$

$$= -M_0 e^{(-i\omega_0 - \frac{1}{T_2})\, t_0} \left. \frac{e^{(-i\omega_0 - \omega - \frac{1}{T_2})t}}{i(\omega_0 - \omega) - \frac{1}{T_2}} \right|_0^\infty$$

$$= -M_0 e^{(-i\omega_0 - \frac{1}{T_2})\, t_0} \frac{1}{i(-\omega_0 - \omega) - \frac{1}{T_2}} \tag{1.14}$$

and multiplying the numerator and the denominator by

$$\frac{i(-\omega_0 - \omega) + \frac{1}{T_2}}{i(-\omega_0 - \omega) + \frac{1}{T_2}} \tag{1.15}$$

the equation is reduced to

$$= -M_0 e^{(-i\omega_0 - \frac{1}{T_2})\, t_0} \frac{i(-\omega_0 - \omega) + \frac{1}{T_2}}{\left(-(\omega_0 + \omega)^2 - \frac{1}{{T_2}^2}\right)} \tag{1.16}$$

where the real and the imaginary part of the equation are clearly visible. The derived equation describes the emission line after excitation by an RF-Pulse. Traditionally, the absorption line is depicted. In addition, $e^{(-i\omega_0 - \frac{1}{T_2})\, t_0}$ of eq. 1.16 represents the phase term. For $t_0 = 0$ the complex Lorentzian function can be described by

$$L(w) = \mathrm{A}(\omega) + \mathrm{iD}(\omega) = M_0 \left(\frac{{T_2}^{-1}}{(\omega_0 - \omega)^2 + \left(\frac{1}{T_2}\right)^2} + i\, \frac{(\omega_0 - \omega)}{(\omega_0 - \omega)^2 + \left(\frac{1}{T_2}\right)^2} \right) \tag{1.17}$$

where $\mathrm{A}(\omega)$ and $\mathrm{D}(\omega)$ represent respectively the absorptive (real) and the dispersive (imaginary) part of the Lorentzian peak in the frequency domain.

At this point the Fourier transform with its simple but extraordinary property of linearity allows the identification and the separation of several oscillating signals of the time domain. The information extracted by the Fourier transform from each resonant frequency

are intensities, phase and decay rate. The latter information is directly related to the line width of the signals, another useful feature analyzed through this project.

It is necessary to explain that in the multidimensional NMR spectroscopy, a multidimensional FFT is needed in order to collect the required frequency domain information. The eq. 1.10 in case of a two dimensional experiment becomes:

$$S(\omega_1, \omega_2) = \iint_{-\infty}^{\infty} s(t_1, t_2) e^{-i\omega_1 t_1} e^{-i\omega_2 t_2} dt_2 dt_1 \tag{1.18}$$

where the signal $s(t_1, t_2)$ is collected recording the FID signal in t_2 varying t_1. The same result is obtainable transforming the signal $s(t_1, t_2)$ first along the direction of t_2

$$S(t_1, \omega_2) = \iint_{-\infty}^{\infty} s(t_1, t_2) e^{-i\omega_2 t_2} dt_2 \tag{1.19}$$

and finally along the direction of t_1 as follow

$$S(\omega_1, \omega_2) = \iint_{-\infty}^{\infty} s(t_1, \omega_2) e^{-i\omega_1 t_1} dt_1 \tag{1.20}$$

During this project a hypercomplex (according to the Clifford algebra) multidimensional FFT has been introduced into the AUREMOL package. Hypercomplex numbers have been used for spectral representation of multidimensional NMR data for obvious reasons. Some interesting features of the real part of a spectrum signal are obtainable through a "phasing" step multiplying the complex signal by a factor $e^{i\vartheta}$. Moreover, the product operator used to represent NMR pulses can be directly expressed through the hypercomplex algebra (quaternions) [Delsuc, 1988]. In case of a 2D NMR experiment it is possible to separate the two independent time domains obtaining sharper lines. This is done using the hypercomplex algebra where each point of the spectrum is no more a complex number but a set of two complex numbers. It can be written as a bicomplex number as following

$$z = (x + iy) * (x_2 + jy_2) = re^{i\phi}e^{j\psi} = r(cos\phi + isin\phi)(cos\psi + jsin\psi) \quad (1.21)$$

The hypercomplex FFT (eq. 1.18) can be written as

$$Z(\omega_1, \omega_2) = \iint_{-\infty}^{\infty} z(t_1, t_2)e^{-i\omega_1 t_1}e^{-j\omega_2 t_2}dt_2 dt_1 \quad (1.22)$$

In a 2D NMR experiment the number of quadrants are four collected with this nomenclature:

- 2rr ($t_1 = real,\ t_2 = real$)
- 2ri ($t_1 = real,\ t_2 = imaginary$)
- 2ir($t_1 = imaginary,\ t_2 = real$)
- 2ii ($t_1 = imaginary,\ t_2 = imaginary$)

where the t_1 direction corresponds to *i* and the t_2 corresponds to *j*. One of the main advantages of using this algebra is that it is possible to phase correct the resulted spectrum in the frequency domain separately for $\omega_1$and ω_2 direction.

Particular care has been taken on the back transform, allowing the user (direct way) to manage the data transformation depending on his purposes through a user friendly interface. Although there is only one Fourier transformation, the data coming from a NMR spectrometer can vary depending on the device that has been used. The common detection types are:

1. single (older detection mode)
2. quadrature (new detection mode)

The most important difference between them is the ability of the latter to detect simultaneously the real and the imaginary part of the FID using the “trick” of the phase shift [Gengying et al., 1999] by ninety degrees with respect to the reference RF signal. This allows overcoming the sign problem for the measured frequencies distinguishing between positive and negative frequencies with the RF reference signal exactly in the middle of the

SW. In case of quadrature detection the bandwidth of the frequencies is the half with respect to the single detection increasing the SNR by a factor of $\sqrt{2}$. The typical time domain data set, the FID (one-dimensional experiments), is collected during the acquisition time. In this case the FID is a set of data points recorded in the time domain whose amount is identified by the TD parameter (NP for Varian manufacturer) counting both real and imaginary part separately as shown in Fig.1.4.

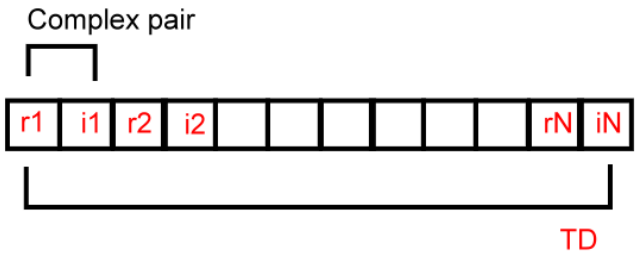


Fig. 1.4 Example of a 1D experiment FID: The TD (NP for Varian) parameter indicates the number of time domain points collected during the acquisition time counting all the real and imaginary parts separately.

In case of a multidimensional NMR experiment, according to the Bruker manufacturer, the FID is represented by a binary raw data file made up of a series of concatenated FIDs having a varying the evolution time t_1. In case of a 2D NMR experiment it is easy to imagine the dataset as a time domain matrix as shown in Fig.1.5.

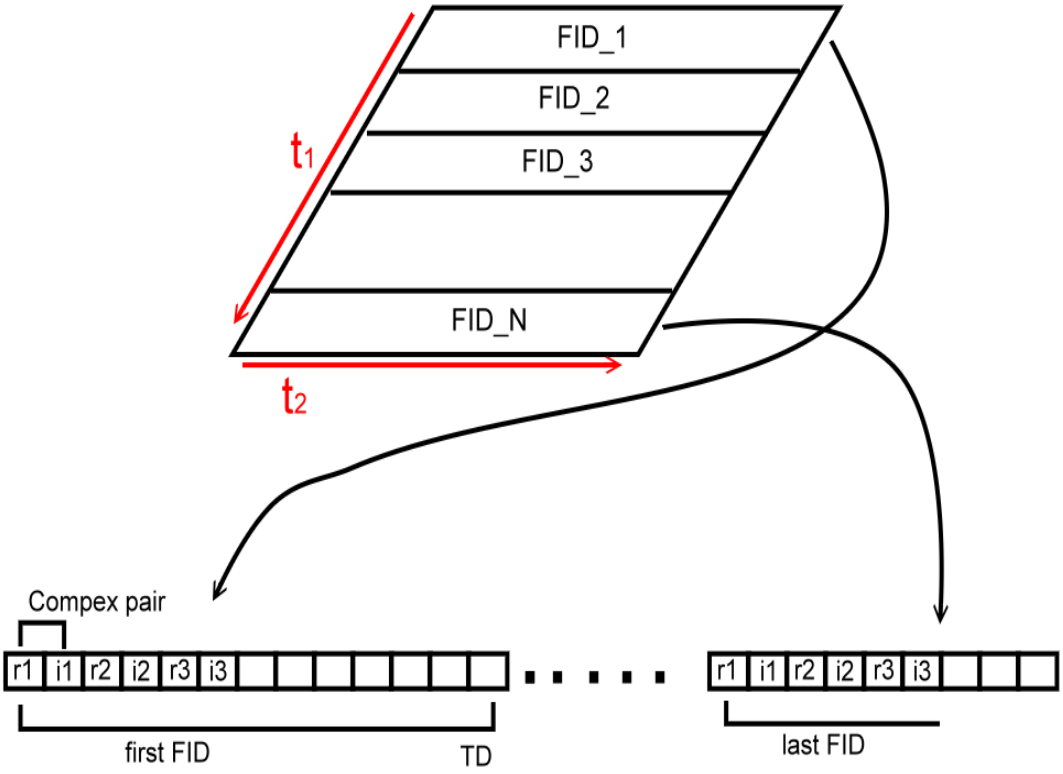


Fig. 1.5 Data collection of a 2D NMR experiment: The acquired FIDs are serially stored into the binary raw data file called *ser* file.

Each collected FID, one for each t_1 delay at $\frac{1}{2SW_1}$ steps, builds up the final user defined total number of increments (i.e. the resolution in ω_1) where SW_1 is the spectral width along the ω_1 direction. Once this time domain data has been collected it is necessary to consider that the Fourier transformation can be applied over two different types of data sampling:

1. simultaneous (i.e. qsim)
2. sequential (i.e. qseq)

where the real and the imaginary part are collected together instantly in the simultaneous case or separately in the sequential one. Depending on the acquired data architecture, different Fourier transformations have been written as shown on table 1.1.

Acquisition mode (direct direction)	Acquisition mode (indirect direction)	Fourier mode	Description
qf		Single real	Single channel detection
	QF	Quad real	No phase change between successive acquired FIDs
qseq		Quad real	Sequential quadrature detection
	QSEQ	Quad real	Successive FIDs sequentially acquired with phases 0° - 90°
qsim		Quad complex	Simultaneous quadrature detection
	States	Quad complex	Simultaneous acquisition of couple of FIDs with phases 0° (x) - 90° (y); 0° (x) - 90° (y)
DQD		Quad complex	Simultaneous digital quadrature detection
	States-TPPI	Single complex	Simultaneous acquisition of couple of FIDs with phases 0° (x) - 90° (y); 180° (-x) - 270° (-y)

	TPPI	Single real	Successive FIDs sequentially acquired with phases 0° (x) - 90° (y) -180° (-x) - 270° (-y)

Table 1.1 Types of Fourier transformation: the reported types are depending on the acquisition modes of the data.

As shown in table 1.1, there are mainly four classes of Fourier transformations:

1. Single (single detection)
2. Quad (quadrature detection)
3. Real (real FFT)
4. Complex (complex FFT)

In particular, the types of Fourier transformation on each direction are automatically inferred from some specific parameters listed in the acquisition files. However, the user can choose any other FFT mode because sometimes the indirect transformation type (i.e. multidimensional cases) is not specified in those files. If the chosen FFT mode is not in accordance with the data architecture, mirror signals would be obtained. For reasons of clarity, Fig.1.6 reports the most important differences among all the modes imposing the dwell time $\Delta = \frac{1}{2S\omega_1} = d\omega_1$ for each t_1 and $\Delta = \frac{1}{2S\omega_2} = d\omega_2$ for each t_2 delay respectively.

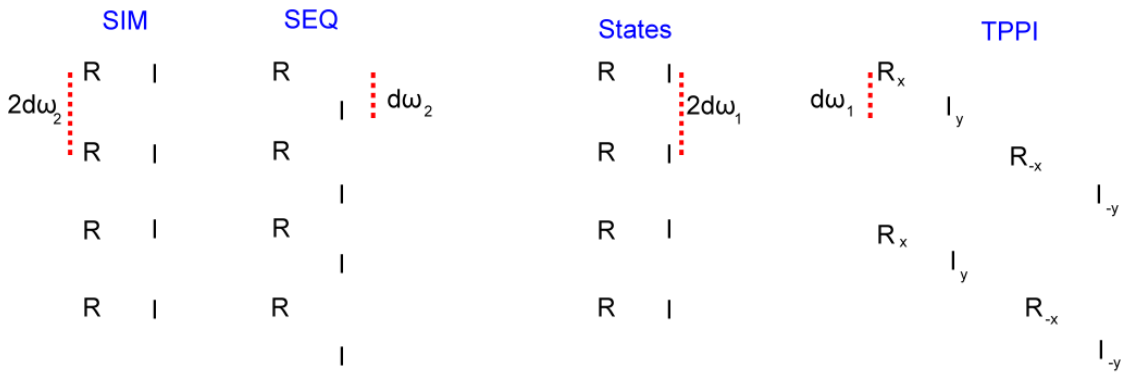


Fig. 1.6 Schematic representation of the simultaneous and sequential mode in the direct and the indirect direction: the TPPI mode is acquired incrementing the pulse phase by 90° during the time t_1 (x (0°), y (90°), -x (180°), -y (270°)) where the odd and even points are cosine and sine modulated signals respectively.

Moreover, the Fourier transformation order (i.e. transform along $\omega_1$direction then along ω_2) is not bound to the acquisition order. Unfortunately, it is not possible to say the same for the window functions commonly used in NMR spectroscopy.

Another processing parameter that must be taken into account when dealing with the developed FFT is the group delay [Moskau, 2002]. In this delay resides the main difference between analog and digital measured experiments. This group delay is the effect obtained when a sinusoidal input signal passes through a digital filter remaining still a sinusoidal signal but with a first high order phase shift. It creates a typical step response at the beginning of the signal as shown in Fig. 1.7. Digital filters combined with data oversampling produces better spectra, in term of signal folding, (fulfilling the Nyquist theorem) in term of baseline performance and SNR with respect to the analog measured ones. This time shift is no more perceptible in the frequency domain in fact it is annihilated through a strong first order phase correction automatically applied after the FFT by the developed routine. Not correcting this time shift would result in a strong wiggling signal whose oscillating rate is strictly connected to the number of points of the group delay.

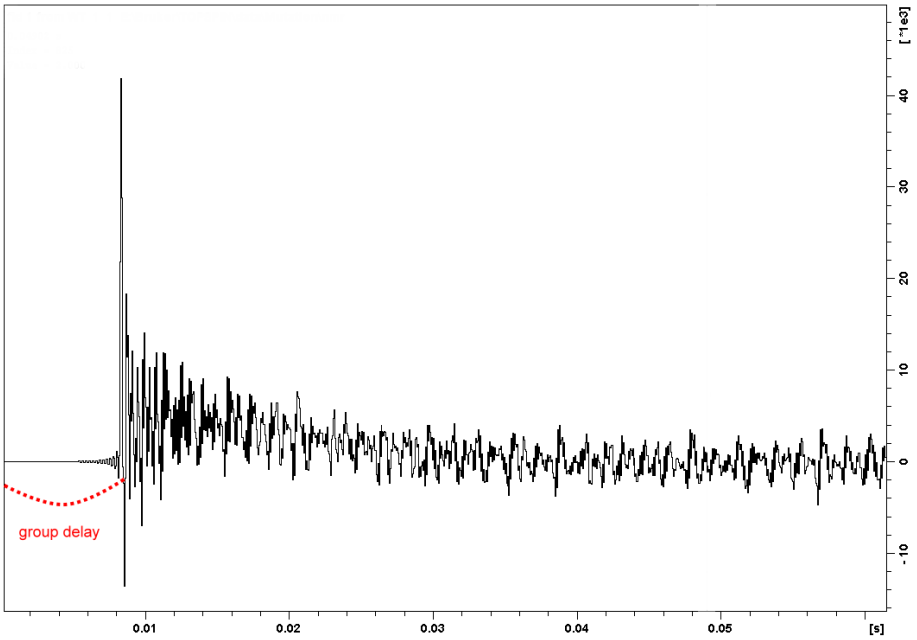


Fig. 1.7 The digital acquired FID: the typical step response of the signal due to the digital filter. The GRPDLY is clearly visible at the beginning of the FID. It corresponds to the time needed by the filter to slide over the data.

In particular, each time domain point belonging to the group delay yields a phase rotation of 180° in the frequency domain. The FFT routine automatically computes the number of points belonging to the group delay and thus the first order phase correction value using several parameters (i.e. DSPVS and DECIM) from the acquisition files. The PKNL parameter is in default active when dealing with digitally acquired FIDs, but it can be freely modified by the user.

Typically, the spectrum resolution is half of the time domain data point (i.e. SI = TD/2). Obviously, this value can be modified by the user in order to increase (only a few hertz for each data point) the digital resolution adding zeros to the end of the FID. In order to avoid step functions connecting the original FID with the added zeros, it is necessary to apply a base line correction in the time domain. It can be obtained computing the complex mean value of the last quarter of the FID and subtracting it from all the complex points of the signal. This procedure is automatically applied by the routine if the user selects a digital

resolution that implies the use of the zero filling techniques. Neglecting such correction would lead to a sync modulated frequency domain signal.

The FCOR (not related to the FFT) parameter is used in order to compensate digitizer idiosyncrasies. It is a multiplication of the first data point by a number between 0.5 and 2. Dealing with digital acquired data the FCOR parameter does not affect the signal in the direct direction since the first points of the group delay are always zero. In case of a Fourier back transformation improper effects should be taken into account, as shown in Fig.1.8 when the FCOR parameter differs from that one chosen during the forward FFT.

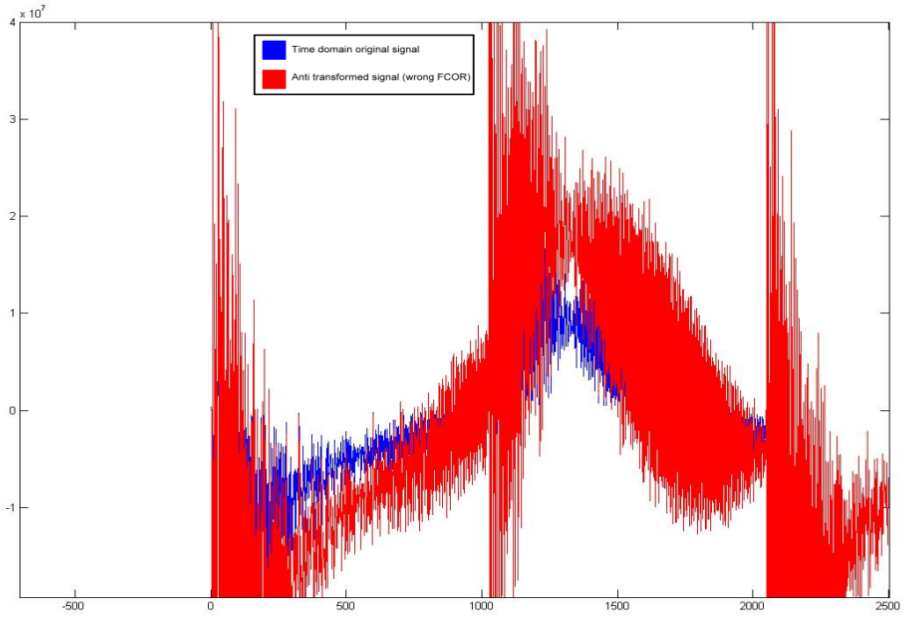


Fig. 1.8 The improper effect of selecting different FCOR values between back and forward FFT: the original time signal (blue) is compared with the back transformed one (red) using a different FCOR value.

In addition, the AUREMOL-FFT routine allows the user to select which parts of the FID to emphasize through some window functions (e.g. sine-bell, cosine-bell, Gaussian and exponential). Depending on the chosen function some other typical parameters can be freely set by the user (e.g. LB, GB and SSB).

1.6 Window functions

Window functions (weighting functions) are frequently applied on time domain datasets in order to accentuate some specific features (and logically deemphasizing some others). It is possible, to reduce the noise level of a spectrum (frequency domain) using an exponential decay function (in the time domain) because the noise is typically present in the last part of the FIDs. This window function is thus useful in order to emphasize the initial part of the FID. Since it increases the SNR, fewer scans are required during the experimental acquisition. The slope of this function is managed through the line broadening factor (LB) in Hz. This factor broadens the lines of each peak reducing the spectral resolution. If the line broadening is negative the opposite happens: the spectrum gain resolution with a loss of SNR. The exponential window function is computed as following:

$$g(t) = e^{-(\frac{(t-1)LB\pi}{2SW})} \tag{1.23}$$

where $t = nd\omega$ with n = 1,…N (number of complex points) and SW represents the spectral width parameter in Herz. In order to obtain a resolution enhancement the final part of the FID needs to be emphasized with a consequent reduction of the SNR. Such window functions sharpen the peak that is a particular useful effect for coupling constant measurements. For instance, an exponential growth (with a negative LB) multiplied by a Gaussian function can be applied in accordance to equation 1.24:

$$g(t) = e^{-\pi LBt(1+\frac{t}{2GBAQ})} \tag{1.24}$$

where GB is the Gaussian broadening factor that is used to position the filtering function everywhere all over the spectrum, $t = nd\omega$ with n = 1,…N (number of complex points) and AQ represents the acquisition time. Another useful window function is the sine bell that is represented by the simple $\sin x$ for x ∈ [0°- 180°]. As a typical sine function it emphasizes the first half of the FID decreasing on the second half. It yields a resonance enhancement observable by sharpened peaks that are also slightly distorted. In addition,

trapezoid functions are used with the aim to remove artifacts coming from the FID truncation, that is necessary when the time to reach the equilibrium is major than the acquisition time.

1.7 Line widths

One of the most interesting features analyzed through this work is the line width parameter. It contains vital information about the structure of the analyzed protein. As discussed before, the Fourier transformation of an exponential decay in time is a Lorentzian line shape function in the frequency domain. The full line width at half maximum (FWHM) of such function is used to compute the loss of coherence R_2 (transverse relaxation rate constant) that is obviously a crucial attribute for spectrum resolution and signal to noise ratio. The maximum evolution time $t_{1_{max}}$ should be selected according to the transverse relaxation rate. For example, in the indirect dimension it should not be bigger than $3T_2$, where T_2 is the spin-spin relaxation time. The line width is made up of two different parts: one depending on the nature of the molecular structure ("natural line width") and the other one due to instrumental limitations.

The line width feature is coupled to the τ_{rot} (rotational correlation time) parameter, representing the average tumbling rate of the considered molecule, in a good approximation by the Stokes-Einstein equation. A direct consequence of this equation, is that the τ_{rot} is directly proportional to molecular mass and viscosity whereas it is inversely proportional to temperature. It is possible to define a strict correlation among all the previously described features maintaining a constant temperature in the absence of internal mobility. A pseudo equation explaining the connections between the line width and some other entities is:

$$kDa \gg \rightarrow \tau_{rot} \gg \rightarrow T_2 \ll \rightarrow R_2 \gg \rightarrow \ lw \gg \rightarrow \ resol \ll \rightarrow \ overlap \ \gg \rightarrow \ inform \ll$$

The complete loss of coherence ("effective line width") is a sum of the above mentioned entities. The effective line width (FWHM) is written as following:

$$\Delta\nu_{\frac{1}{2}} = \frac{1}{\pi T_2^*} = \frac{R_2^*}{\pi} \tag{1.25}$$

with

$$R_2^* = R_2^{\Delta B_0} + R_2 \tag{1.26}$$

where $R_2^{\Delta B_0}$ is the effect of magnetic field inhomogeneity and R_2 is the sum over all individual rates caused by the dipolar coupling (R_2^{DD}) or the chemical shift anisotropy (R_2^{CSA}).

2

Materials and methods

2.1 Introduction

Many proteins fold into a well-defined characteristic tertiary structure, namely the native state that is essential for a proper function. A correct although partial folding is essential for functionality purposes. Misfolded proteins are associated to several diseases as Creutzfeldt-Jakob, bovine spongiform encephalopathy, Alzheimer disease and even cancer. Several types of factors typically lead to desired or undesired structural changes.

The interactions responsible of the formation of secondary and tertiary structure (e.g. hydrogen bonds, disulfide bridges between sulfur atoms of the cysteines, hydrophobic effects and van der Waals forces) may be disrupted by environmental changes implying the denaturation of the protein. Such a molecule does not possess a stable structure but it exists in a partially or totally unfolded state. In particular, an increasing temperature involves an increased molecular motion that means more instability. In addition, the protein structure may be broken by the use of denaturants or by extreme pH values. In fact, pH extremes below 5.0 and above 10.0 destabilize most proteins. Once the source of heat or the denaturants are removed, often the molecule spontaneously regains its native conformation with a decrease in free energy. More than one partially folded intermediate state may exist.

The primary structure of a protein determines its tertiary structure that in turn determines its function. This involves that a mutation of even a single amino acid may be reflected in the folding of the structure with a consequent malfunction of the protein. However, similar sequences do not always involve similar structures (and functions), thus the proteins need to be classified into families depending on other properties such as the common domains they contain. Every domain has its own biochemical function and the function of the entire protein results as the sum of the single domains.

Conformational changes may be also related to binding interactions between the protein and any other ligand, involving also modifications of denaturation temperature. The contact surfaces reveal complementary properties, thus for instance, polar ligand groups tend to bind with polar protein groups via hydrogen bond. The binding is performed by means of intermolecular forces (i.e. ionic and hydrogen bonds) or hydrophobic effects and it is usually reversible.

In particular, the developed project attempts to ease investigations of three main possible sceneries:

1. The protein itself can be a drug. In this case it is important that the 3D-structure remains intact.
2. Proteins are targets for drugs that modulate the function of the protein (activate or deactivate the protein). In order to obtain a correct modulation it is important to identify the interaction site.
3. Characterize conformational changes of the protein after a perturbation (e.g. pressure variation).

Actually, there is an increasing interest on predicting protein-drugs interactions. Considering that the protein acts as a receptor, it can bind to inhibitors that inactive its function. Therefore, the drugs designers have the intent to rationally inactivate those malfunctioning proteins keeping intact the spatial structure of the molecule. Even if the structure of a protein is available, determining the structure of its complex with ligands may be experimentally demanding [Dembowski et al., 1994]. This problem has stimulated the computational molecular modeling field, i.e. the molecular docking [Kitchen et al., 2004; Moitessier et al., 2008]. Identifying the functional surface of a protein where it interacts with the ligand is therefore an important phase in the pharmaceutical industry. The unaltered structural parts of the protein need to be identified and then evaluated if they belong to regions that must be absolutely kept intact. Consequently, the most involved residues in the structural changes (i.e. those binding to the inhibitor) must be defined verifying that they are not essential for a drug acceptance. In this manner the designed drugs can be discarded or retained in accordance to a certain threshold of the fraction of the unvaried molecule.

The main intent of a standardized quality control relies on the fact that a user can automatically measure and quantify structural differences on a set of samples of a certain substance. The user disposes just on a reference spectrum and/or the structure to be evaluated. Some solutions have been proposed in the last two decades with the aim to analyze and validate three-dimensional structures of macromolecules. Some of them relay on NMR techniques [Kalchhauser & Robien, 1984; Schröder & Neidig, 1999; Ross et al., 2000; Rossè et al., 2002] other on a combination of LC-MS and the former [Golotvin et al., 2006; Thiele et al., 2011].

Even though the LC-MS [Paul & Steinwedel 1953] techniques are widely recognized as the most powerful methods for analytically characterizing organic libraries [Sepetov & Issakova, 1999], they do not provide enough structural information. On the other hand NMR data contains a lot of useful structural information but the interpretation of the spectra is tedious requiring human experts. In particular, due to the strict relationship existing between NMR spectra and the 3D-Structure variations it is possible to identify spatial modifications by changes in the experimental spectra. LC-MS techniques can be complementary used in order to define the precise covalent structure.

2.2 Materials

2.2.1 Back-calculated dataset

2.2.1.1 HPr protein from *Staphylococcus aureus* (wild type)

The HPr protein plays an essential role transferring a phosphoryl group as part of the PTS (phosphotransferase system) for carbohydrate transport through the cell membrane. It is a medium size protein with 88 residues and formed by three α-helices and four stranded β-sheets.

The synthetic two dimensional NOESY spectrum of HPr protein from *Staphylococcus aureus* [Görler et al., 1999b, Maurer et al., 2004] has been back-calculated with the module RELAX-*JT2* [Ried et al., 2004] of AUREMOL starting from the three-dimensional

structure of HPr using the corresponding chemical shifts. In particular, the HPr protein 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 that have been used are 0.85, 0.80 and 0.65 respectively. The H_2O and the D_2O contents used are 90% and 10% (volume parts). A mixing time of 0.15 s, cut off of 0.5 nm, relaxation delay of 1.3 s and a Larmor frequency of 600.13 MHz have been chosen. A spectral width of 13.9791 ppm on both directions has been selected. The spectral offset of 11.7249 (ω_1) and 11.7205 (ω_2) has been used. The number of data points of the simulated spectrum in the frequency domain is 512x2048 (the same as the experimental one). An additional line broadening of 0.5 Hz has been used. The Lorentzian line shape has been selected in order to compute the simulated dataset with a simultaneous phase increment on both directions. The protein structure file (.pdb) *1ka5* has been downloaded from the webpage www.pdb.org and it has been used to simulate the NOESY spectrum. The three-dimensional structure of the bundle made up of the 16 conformations with the minimal energy has been used.

The synthetic spectrum of HPr has been used to generate an additional set of four spectra containing experimental solvent and increasing Gaussian distributed noise. The HPr spectrum has been converted into time domain data using the AUREMOL-CFID routine (FID simulation). The water artifact has been measured acquiring a 2D-NOESY spectrum of 90 % H2O/10 % D2O with solvent pre-saturation and with the same acquisition parameters of the protein spectrum. The time domain signals of the water artifact and HPr have been added. The solvent signal has been scaled in such a way that the maximum was about 500 times stronger than a typical protein resonance. Time domain Gaussian noise has been incrementally added to the HPr time domain signal containing experimental water. The dataset has been generated in order to obtain a signal to noise ratio of approximately 2, 4, 6 and 8σ (computed in the frequency domain) for a proton-proton pair in a distance of 0.5nm as described by Baskaran et al. (2009). The resulting dataset has been shifted (circular shift) by -0.08 and -0.03 ppm along ω_1 and ω_2 direction respectively.

2.2.1.2 HPr protein from *Staphylococcus aureus* (mutant H15A)

A single amino acid change on residue 15 in the protein HPr has been performed. In particular, the residue His (active-center of the protein), has been substituted with the residue Ala. As for the previous described wild type of HPr protein from *Staphylococcus aureus* the synthetic two-dimensional NOESY spectrum has been back-calculated with the module RELAX-*JT2*. The same parameters used to simulate the wild type have been adopted. The three-dimensional bundle of structures (*h15a*) made up of the 20 conformations with the lowest energies has been used.

2.2.1.3 HPr protein from *Staphylococcus aureus* (partially denatured)

A partial denaturation of the HPr protein from *Staphylococcus aureus* has been performed. In particular, the range of residues Gly13-Ser27 has been denatured, totally destroying the first α helix. This has been done computing structural restraints (e.g. distance, dihedral angle and H-bond) of the original HPr protein (wild type) via the AUREMOL routine PERMOL [Möglich et al., 2005]. Afterwards the set of restraints Gly13-Ser27 has been erased and new structures have been calculated by the CNS [Brünger et al., 1998] software package. In particular, the dynamical annealing protocol has been used with 4772 NOE restraints, 44 H-Bond restraints and 311 dihedral angle restraints starting from the extended strands. The torsion molecular dynamics has been chosen and no water refinement has been done. The computed bundle of the ten best structures has been used to compute the simulated spectrum. After the MD simulation the chemical shifts of the atoms belonging to the denatured part have been predicted as random-coil shifts [Schwarzinger et al., 2000, Schwarzinger et al., 2001, Arnold et al. 2002]. The adapted shifts list has been used to back-calculate the related NOESY spectrum.

2.2.1.4 HPr protein from *Staphylococcus aureus* (fully denatured)

A fully denaturation of the HPr protein from *Staphylococcus aureus* has been performed. This has been obtained via the MD simulation through the GROMACS [Berendsen et al., 1995; Lindahl et al., 2001] software package using the amber94 force field [Cornell et al., 1995, Sorin et al., 2005] and the TIP3P water model [Jorgensen et al., 1983; Miyamoto & Kollman, 1992]. A pressure of 0.1 MPa at 303 K has been used for the simulation. The protein has been simulated for 30 ps and one structure every 1 ps has been stored [Day & Garcia, 2008]. Each conformation was solvated in 11229 water molecules and 14 $Na^+$and 8 Cl^- ions. For controlling simulation temperature the Berendsen algorithm [Berendsen et al., 1984] has been used. A temperature coupling time of 0.1 ns and a pressure coupling time of 0.5 ns has been adopted. The water compressibility of 5.5 e-5 has been considered. The Long-range electrostatic interactions were calculated using particle-mesh Ewald24 with a grid spacing of 1.2 Å. A cut off of 9 Å has been used for Van der Waals energies. The fully denatured three-dimensional molecular structure of HPr obtained from the molecular dynamics simulation has been successively used to compute the simulated spectrum of HPr using the same parameters as described before. In order to back-calculate a spectrum representing the denatured HPr protein all the chemical shifts of each atom have been considered as random coil chemical shift taking into account neighbor residues [Schwarzinger et al., 2001].

2.2.2 Experimental data sets

2.2.2.1 HPr protein from *Staphylococcus aureus* (wild and mutant)

The two-dimensional experimental NOESY spectra (wild and mutant cases) were recorded from a sample containing 2.7 mM uniformly ^{15}N-enriched HPr protein from *Staphylococcus aureus* in 500 µL 95% H_2O/5% D_2O, pH 7.0. The mutant has been obtained substituting the residue histidine 15 with an alanine. The NMR spectra were recorded on a Bruker DMX-600 spectrometer operating at 600 MHz with a mixing time of

150 ms. The water signal was reduced by selective pre-saturation in the mutant version and with water-gate in the original wild HPr spectrum. Both spectra were recorded with a relaxation delay of 1.3 s and with 1024x4096 complex time domain points and successively Fourier transformed with a resolution of 512x2048 data points. The spectral widths were 13.9790 ppm and the number of scans was 32 in both cases. They have been measured at 303 K and have been acquired with the program TOPSPIN (Bruker, Karlsruhe). Each time domain data was filtered by exponential multiplication with a line broadening in the two dimensions of 0.5 Hz.

2.2.2.2 Human prion protein (Prp): Xenon binding

Prions are infectious pathogens that are responsible of fatal neurodegenerative diseases [Prusiner, 1998]. It is a soluble protein made up of 230 residues. The globular domain (121-230) contains two anti-parallel β-sheets and three α-helices [Zahn et al., 2000].

Five two-dimensional HSQC spectra of the recombinant human prion protein (*hu*PrP^{C}) have been measured in $^{1}H_2O$ on a sample containing 0.25 mM of ^{15}N-enriched *hu*PrP^{C} (residues range: 23-230) in 5 mM acetate buffer, pH 4.5 with 10% $^{2}H_2O$ and 0.1 mM of 2.2- dimethyl-2-silapentane-5-sulfonate (DSS) as internal reference. All the experiments have been recorded at 293 K on a Bruker Avance-500 spectrometer operating at 500 MHz. The Bruker pulse program used for the ^{1}H-^{15}N correlation spectra was hsqcetf3gpsi. A selective water pre-saturation has been applied to the spectra, that have been acquired with a relaxation delay of 1.2 s and with 2048 complex time domain points in the direct dimension ω_2 and 512 complex time points in the indirect dimension ω_1. They were recorded with a spectral width of 6.0161 ppm in ω_2 and 36.00 ppm in ω_1 except the first one (absence of xenon) having a spectral width of 8.0214 ppm in ω_2 and 36.00 ppm in ω_1. The number of scans was 80 in the first experiment, whereas it has been reduced to 48 on the other four cases. They have been recorded with the program TOPSPIN and they have been filtered in the time domain by a Gaussian multiplication with a line broadening of -6 Hz for the proton direction and -8 Hz along the nitrogen direction and a Gaussian broadening of 0.1 in ω_2 and 0.12 in ω_1. The assignment of the first spectrum was based on the data published by Zahn et al. [Zahn et al., 2000] and adapted as written by Kachel et

al. [Kachel et al., 2006]. The sample was pressurized with xenon gas (129 Xe) at 0.2, 0.4, 0.8 and 0.14 MPa. The first experiment at 0.1 MPa (absence of xenon) has been used as the assigned reference spectrum, whereas the other four unassigned spectra are all considered test data.

2.2.2.3 Human prion protein (Prp): High pressure NMR

The recombinant human prion protein (*hu*PrP^{C}) structure dependence on the pressure and temperature conditions has been evaluated using two different dataset of two-dimensional HSQC-TROSY spectra [Pervzsgub et al., 1997]. They have been measured using a Bruker Avance-600 spectrometer operating at a ^{1}H frequency of 600.13 MHz and at a ^{15}N frequency of 60.81 MHz, with a relaxation delay of 1 s. The solvent has been reduced by selective pre-saturation. All the spectra were acquired with 2048 complex time domain points in F2 (proton dimension) and 256 in F1 (nitrogen dimension), that were successively Fourier transformed with a size of 4096 real data points in F2 and 1024 in F1. The spectral width in F2 was 11.9705 ppm and 30.00 ppm in F1. The number of scans was 48 in every experiment. The time domain data have been filtered by exponential multiplication with a line broadening of 7 Hz in both directions.

For the NMR experiments a 1.1 mM solution of ^{15}N-enriched *hu*PrP^{C} (23-230, long sequence) and 1.2 mM solution of ^{15}N-enriched *hu*PrP^{C} (121-230, short sequence or folded core) each in 10 mM sodium acetate buffer, pH 4.8 has been used. They have been measured in $^{1}H_2O$ with 8% $^{2}H_2O$ added. The samples contained 0.1 mM of 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) as internal reference. In order to apply high pressure, an on-line variable pressure cell system has been used [Arnold et al., 2003, Kremer et al., 2004] with a sapphire capillary of 1.72 mm inner and 3.14 mm outer diameter. All the spectra were measured and processed with XWINNMR package (Bruker, Karlsruhe). High pressure is a valid tool to analyze the dynamics and the structure of the folding intermediates, since the transmissible spongiform encephalopathies (TSes) adopt alternative folds propagating the disease [Kuwata et al., 2002]. The short sequence spectra have been recorded at hydrostatic pressure steps of 0.1, 50, 100, 125, 150, 175 and 200 MPa, at a temperature of 293 K. The long sequence spectra have been acquired at

temperature steps of 293, 303, 313, 323 and 333 K, at a pressure of 200 MPa. The experiment at 0.1 MPa and 293 K have been used as assigned reference spectra and compared with all the other experiments with increasing pressure and temperature respectively.

2.3 Methods

2.3.1 Software

AUREMOL was initially developed using ANSI-C language. Currently, many routines are converted from ANSI-C to C++ code. Some existing AUREMOL routines have been used:

- AUREMOL-ALS for baseline correction [Malloni et al., 2010]
- AUREMOL-SSA for water removal [De Sanctis et al., 2011]
- SHIFTOPT for chemical shift optimization [Baskaran et al, 2009]
- RELAX-JT2 to back-calculate spectra [Ried et al., 2004]
- PERMOL to extract structural restraints [Möglich et al., 2005]
- PEAK-PICKING to identify true peaks [Antz et al., 1995]
- INTEGRATION to calculate peak volumes [Geyer et al., 1995].

As external software for data acquisition and processing has been used

- TOPSPIN 3.0 (Bruker BioSpin)

Some external software packages used for molecular dynamics simulations are:

- CNS [Brünger et al., 1998] has been used in order to obtain three-dimensional structures of the partially denatured HPr. The computation has been performed on a Linux cluster made up of eight nodes.
- GROMACS 8.0 has been used to compute the three-dimensional molecular structure of the fully denatured HPr [Berendsen et al., 1995; Lindahl et al., 2001].

3

Results

3.1 General considerations

The Quality Test Analysis (QTA) project has been developed in order to compare a set of spectra to discover if they are identical within the limits of error. In case differences are detected, they should be qualified and analyzed. An important application in protein science is the analysis of where and how much the structure of an isolated macromolecule or a macromolecular complex differs with respect to a target one.

In order to compare test and reference spectra, a previous standardization is required. Some of the acquisition and processing parameters of the test cases need to be adapted to those ones of the reference spectra. Some others have not to be necessarily adapted but any difference between them might compromise the results. In particular, the recorded experiment type must be the same and the resolution has to be identical, whereas the FIDs filtering type, the water suppression technique and the acquisition mode (digital or analog) may not coincide. The spectral widths, the offsets, the number of scans (NS), the receiver gain (RG) and the intensity scaling factor (NC_proc) need to be automatically adapted if they do not match each other. The spectral shift is evaluated and if it has occurred a proper correction is applied. It is assumed that the spectra in comparison have been previously phase corrected. The performance of the method is improved distinguishing a priori some possible artifacts (noise, baseline points and the solvent) through a Bayesian analysis [Antz et al., 1995] before carrying out the comparison. The AUREMOL-ALS routine for baseline correction [Malloni et al., 2010] has been introduced in the AUREMOL software. It allows

correcting the baseline of spectra involved before starting the analysis. In addition, if the evaluated spectra have a very strong solvent signal which hides many resonances of interest, the AUREMOL-SSA routine for solvent removal can be previously applied. The algorithm is able to handle more than one reference spectrum and several test spectra at once. If there are more reference spectra only the overlapping peaks occurring in all spectra are retained and used for the requested comparison. The synthetic spectrum back-calculated with the RELAX-JT2 [Ried et al., 2004] algorithm of any protein can be used as an adjunctive reference spectrum. Four different situations are automatically handled by the developed algorithm: neither the test nor the reference spectra are assigned; only the reference spectrum is assigned; both have been previously assigned and finally the three-dimensional molecular structure is present or not.

Depending on the type of spectra in hand, some main properties (position, volume, amplitude, multiplet structure and line width) are evaluated and used to estimate structural changes. Dealing with HSQC-type spectra, the chemical shifts, the volumes and the line width changes of single atoms are studied in detail. In the case of a NOESY spectral comparison, the chemical shifts and the volume variations are evaluated peak by peak, analyzing the level of symmetry of the residue patterns exploiting the symmetry and similarity properties of those patterns in the spectra.

During this project many definitions of "peak" are used. The terms $P_{m_c}$and P_{s_c}identify the center of a multiplet peak and of a singlet peak respectively. The term $P_c = P_{m_c} = P_{s_c}$ represents the central position of both P_{m_c} and P_{s_c}. The term P_I identifies the maximum intensity of a peak (multiplet and singlet), P_V corresponds to the peak volume (sum of all intensities P_D of a peak taking into account the overlap problem) whereas P_D represents all the peak voxels. The term P_X corresponds to the voxels belonging to the spectrum raw data. The term P_N represents the volume error as a function of the local noise level of the peak P. The terms $P_{ppm}$and P_{int} identify the peak position in the ppm unit and in the voxel unit (integers) respectively.

3.2 Project overview

The quality test analysis is not just a strict quality control of all types of spectra. The project has been developed in order to automatically adapt the quality analysis to different cases as the ligand screening (i.e. drug binding) and the determination of structural changes (variation of external conditions and mutations). As reported in Fig. 3.1 the quality test project is a two-way mode analysis. The standard mode has the ability to control the quality of any test sample (analyzing the peaks only in the range of the allowed shift of $\pm$ 1 voxel) and the other one performs the same control with the advanced possibility to map structural changes due to mutations, variations of external conditions and ligand binding.

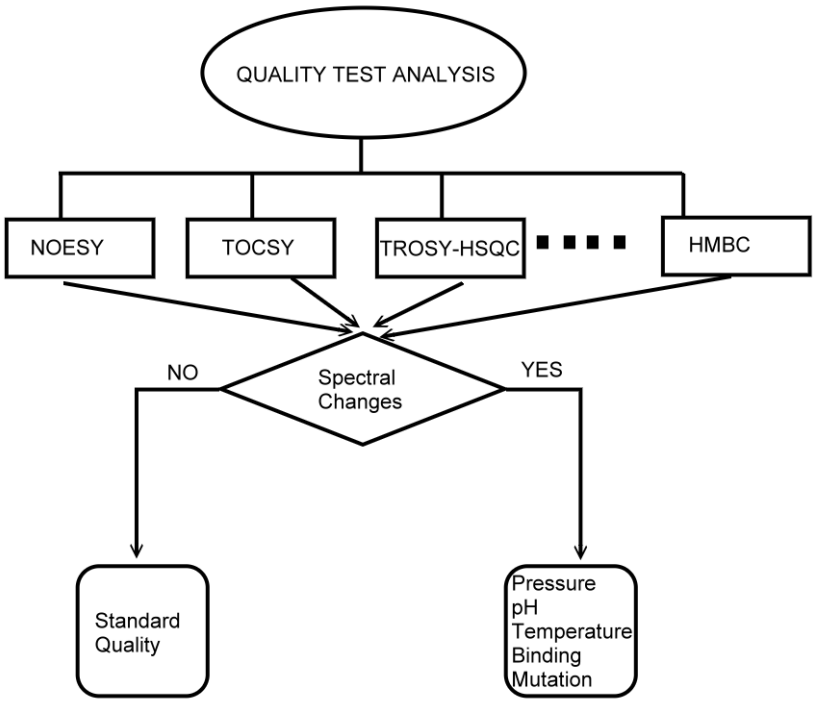


Fig. 3.1 General overview of the Quality Test Analysis project: The two-way mode analysis made up of the standard quality control (left) and the advanced one (right) including the mapping of the structural changes due to several reasons as: mutation, binding and variation of external conditions.

Efforts have been done in order to maintain the project and consequently the routines as general as possible. Special cases like NOESY and TOCSY spectra are particularly

analyzed in details in order to increase the available information. In Fig. 3.2 is reported the specific project description.

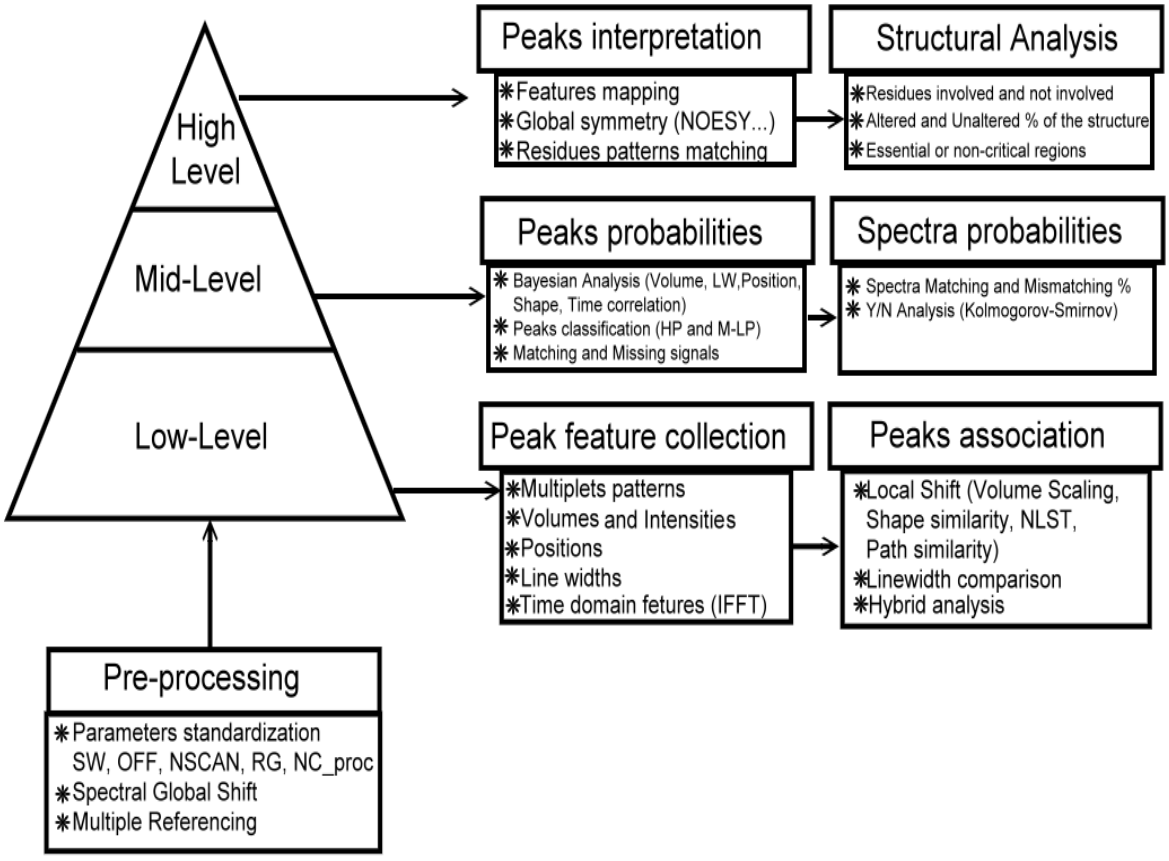


Fig. 3.2 Detailed overview of the Quality Test Analysis project: Standardization of the data during the pre-processing; peak feature collection and peaks association in the low-level; peaks and spectra probability computation of belonging to a certain class (matching and mismatching signals) the mid-level; peaks and patterns feature mapping with an additional structural analysis in the high-level.

The AUREMOL-QTA comprises four main procedures that are depicted in Fig. 3.2. It starts with a pre-processing part encompassing the identification of peak classes (noise, baseline, solvent and resonances of interest), spectra standardization (e.g. spectral width, offset, number of scans, receiver gain, intensity scaling factor and general spectral shift management) and multiple referencing (more than one spectrum used as a reference candidate). It continues with a low-level analysis involving a peak by peak feature collection (volume, intensity, line width, chemical shift, time domain cross-correlation and multiplet pattern) and a further signal comparison and association among the considered

spectra. Without taking into account the availability of the peak assignments it evaluates if a specific peak in the reference spectrum can be associated to another peak in the test cases. In particular, these peak connections are automatically investigated in a user definable ppm range. This range is defined for each measured direction. The associated peaks are compared feature-wise (position, volume, line width, cross-correlation and cosine similarity). The mid-level step focuses on those previously associated peaks and relies on the calculation of the probability of representing the same signal for each of them in dependence on the user selected feature. Consequently, each association allows defining if a matching between any two peaks in two different spectra has a high, mid or low probability according to some previously defined thresholds. It eases the discovering of those peaks that are located in the reference spectrum but which are, at the same time, missing into the test ones and vice versa that might represent significant variations in the structure. An overall probability of dealing with a test spectrum similar to the reference one is calculated and a ratio of the number of matching peaks over the total is obtained in order to define quantitatively the similarity between the spectra. The high-level phase is related to the structurally altered and unaltered parts of the molecule. For instance, it allows mapping the volume and the chemical shift behavior of every connected peak through a set of test spectra acquired at different external conditions. In particular, when HSQC or HSQC-TROSY spectra are analyzed, where almost each signal corresponds to an amino acid, it allows the identification and the quantification of the residues involved in the modification. When dealing with NOESY and TOCSY spectra the peaks are still evaluated separately, but nonetheless the pattern analysis of the residues and the symmetrical properties can be utilized as adjunctive information in order to discover complete pattern dislocations. During the high-level step a collective structural analysis is performed in order to express more than the simple similarity between pairs of spectra. The estimation of the portion of unaltered molecule structure is estimated in this level. This task is particularly interesting when comparing some spectra of molecules before and after drug binding. The drug designer can directly verify if a certain ligand can be rejected or retained depending on the residues involved in the application and on the altered molecular region.

As described above, the method relies on two different controls. The computational differences in HSQC and NOESY spectra are reported in Table 3.1.

Method	Description
init()	Read all spectrum parameters
loadspectra()	Create all the structures used through the computation
run_Simu()	Compute the simulated spectrum
load_external_pars()	Read all user defined external parameters
doStdCalculation()	Noise level analysis, NLST calculation, Global shift computation
findMultiplets()	MSA (3SA)
doLocalShift_v4()	Local shift algorithm
analyze_structural_changes()	Pattern analysis
pattern_identification()	Pattern probabilities
doLocalLinewidth()	Local width calculation
doLocalVolScore()	Score of volumes
ZO_Normalize_All()	Normalization method
doCalcBayes()	Bayesian probabilities computed for all the features
getKolmo()	Kolmogorov-Smirnov Test
analysis()	Collect all the information for showing graphical results
writeAll_XML()	Store all the parameters computed through the quality test in the XML data format

Table 3.1 Main methods implemented in the AUREMOL-QTA source code. Additional routines are used when dealing with NOESY (TOCSY) spectra in order to collect as much information as possible (see row 3, 6, 8, 9).

3.3 The developed interfaces

3.3.1 The AUREMOL-FFT interface

The AUREMOL-FFT interface has been written in order to transform measured data from the time domain to the frequency domain and vice versa. This hypercomplex FFT routine can be separately selected in the AUREMOL "Calculation" menu as shown in Fig. 3.3. Once it has been clicked, the AUREMOL-FFT module opens the main dialog requiring the input path where the time (FID or ser files) or the frequency domain files (1r, 2rr, 3rrr and so on) are located and the Fourier transform direction (i.e. forward or backward) required can be defined as reported in Fig. 3.4.

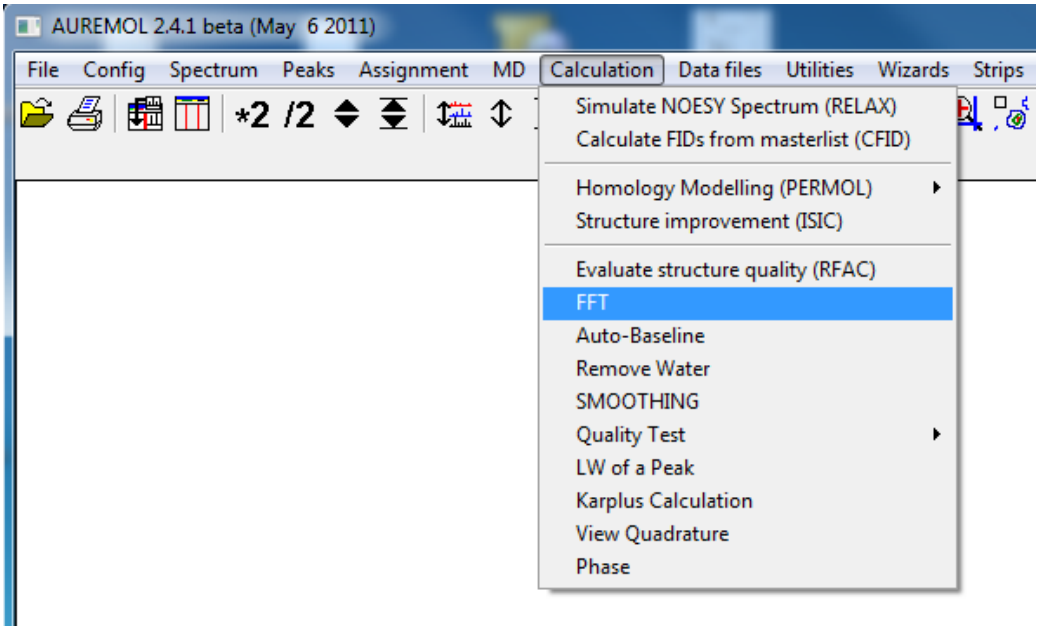


Fig. 3.3 Starting the AUREMOL-FFT module: The FFT module called via the "Calculation" menu in the AUREMOL software package.

In case of discordance between the provided file and the transformation's direction, the routine aborts reporting the error type that has triggered the termination of the procedure.

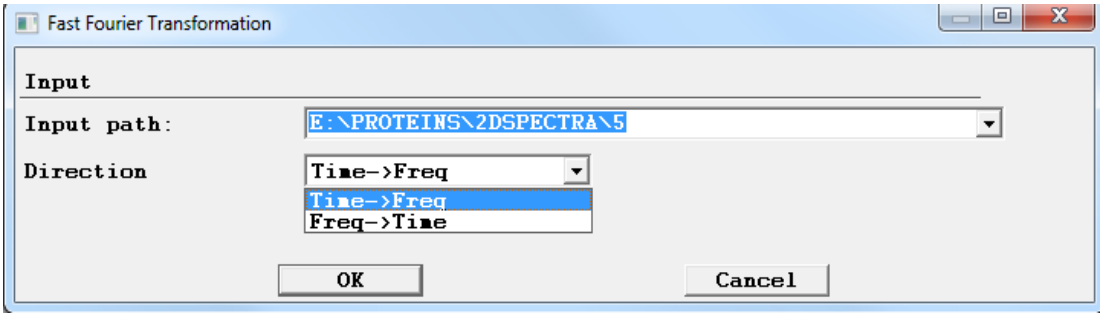


Fig. 3.4 AUREMOL-FFT main input interface: It is mandatory to select the input path of the data that need to be transformed and the FFT direction (time-frequency or frequency-time).

After having selected the input path and the FFT direction, in case of time to frequency transform, the user has to insert in the dialog (direction-dependent) reported in Fig. 3.5 all the needed parameters. Some of them are automatically (see parts A, B, C, E, L, M of Fig. 3.5) fulfilled by the routine since they are stored in the *acqus* (acquisition state) files whereas some others have to be manually set by the user (see parts D, F, G, H, I of Fig. 3.5). For example, the part A has been automatically set to the little endian order (meaning that in a measured time data point represented by a 4-bytes integer the least significant byte is stored first) after have checked the acquisition parameter BYTEORDA. This parameter (BYTEORDA) depends on the processor and on the OS of the machine that controls the spectrometer. If the time domain dataset is measured (and stored) with a machine using a big-endian processor than the BYTEORDA parameter is set to one. In order to properly manage this dataset in a little-endian machine (the *Windows* OS supports the little-endian order), the data is automatically converted by the routine. The BYTEORDP (parameter stored into the *procs* files) is set accordingly (zero and one in the little-endian and big endian case respectively). In addition, the user can change it depending on his necessity. The B part has also been automatically set to complex (the direct direction is even in quadrature) but it is possible to change it to a real FFT. The C part is set considering that $SI_i = \frac{TD_i}{2}$ where i is the direction of the transformation. The SI parameter represents the number of real data points expected in the Fourier transformed frequency domain, that can include the zero filling if $SI_i > \frac{TD_i}{2}$. The SI parameter varies from the lower limit of 16 data points and the upper one of 131072. By default the FCOR parameter (described in part

D) is set to 1 as explained in par. 3.4.2.8. In the reported example the FCOR parameter has been set to 0.5 in order to weight the first complex data point.

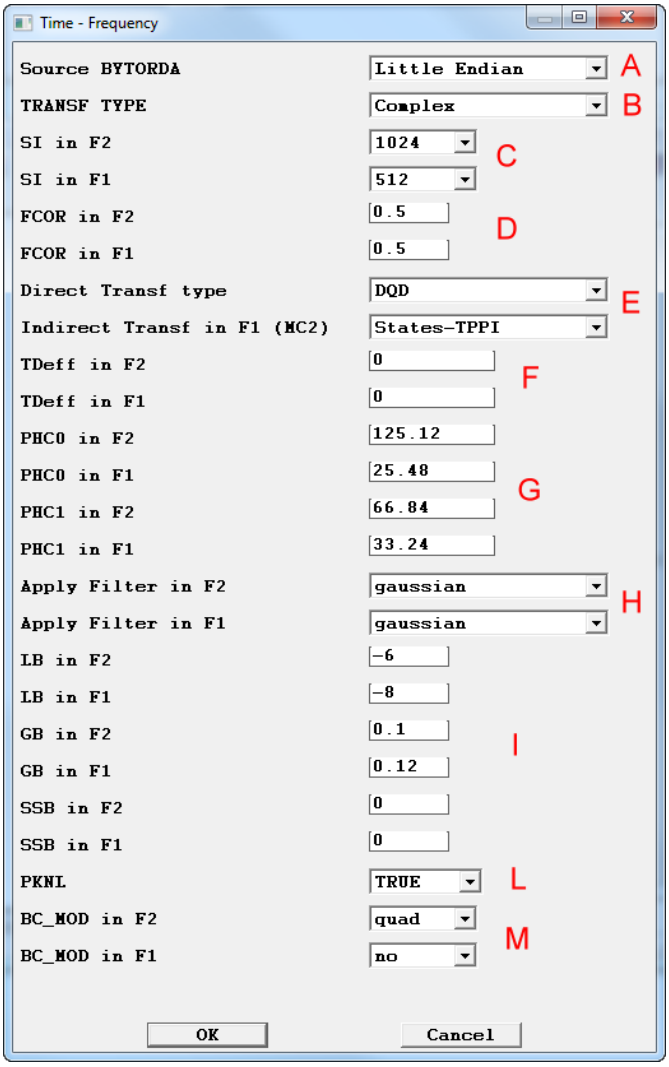


Fig. 3.5 The AUREMOL-FFT (Time-Frequency) interface: The dialog shows all the parameters that can be modified in order to perform the Fourier transform. Some parameters are automatically set by the routine (A – C, E, L and M) others are optional (D, F - I). All of them can be modified by the user.

The part E of Fig. 3.5 shows that the automatically selected FFT transform types are DQD-States-TPPI. Both combo boxes (in case of a 2D spectrum) are modifiable in order to perform all the different FFT transformations discussed in par. 1.5. The part F represents the number of points that are used during the processing of the data. The default value of zero indicates that the whole dataset is used. The G part represents the value of zero and

first order phase correction (in degrees) in all the spectrum directions. These parameters are settable by the user since this information is contained into the *procs* (processing state) file. If these files (*procs*) are already present into the previously defined folder, they are parsed and used easing the task of filling out the interface. The H and I parts show the filtering windows described in the first chapter (see par. 1.6).

The developed digital filters have been written considering the possibility of different acquisition modes (*sequential* or *simultaneous*). This is necessary in order to correct each time domain data point accordingly to the acquired modes. If the acquisition mode is *sequential* then the real and the imaginary part of the time domain complex data point are considered separately and multiplied by different filter coefficients (one after the other). In the *simultaneous* acquisition mode the real and the imaginary part of a time domain data point are multiplied by the same filter coefficient.

The L part in case of digital acquired spectra is automatically set to true but it is logically possible to disable the PKNL option avoiding the group delay correction (see par. 3.4.2.5). The M part is referred to the baseline correction applied in the time domain through the BC_mod option. If the user enables this option the routine subtracts the signal mean computed over the last quarter of considered FID (if the quad or single option is enabled). This parameter has to be set in accordance to the detection modes. In case of a single channel detection the BC_mod parameter has to be set to single and it has to be set to quad in case of a quadrature detection mode. The BC_mod parameter is automatically enabled by the routine according to the detection modes if the user performs a zero filling of the measured data ($SI_i \geq \frac{TD_i}{2}$). The term SI_i represents the number of real data point in the frequency domain along the direction *i* while TD_i is the number of data points in the time domain along the direction *i*.

3.3.2 The AUREMOL-LW (Line width) interface

The AUREMOL-LW module can be used independently from the AUREMOL-QTA routine. For example, if the user needs to compute the line width of all the spectrum peaks (in a batch processing mode) along all the directions, clicking “Calculate linewidth” under the menu “Peaks” this is done automatically as shown in Fig. 3.6.

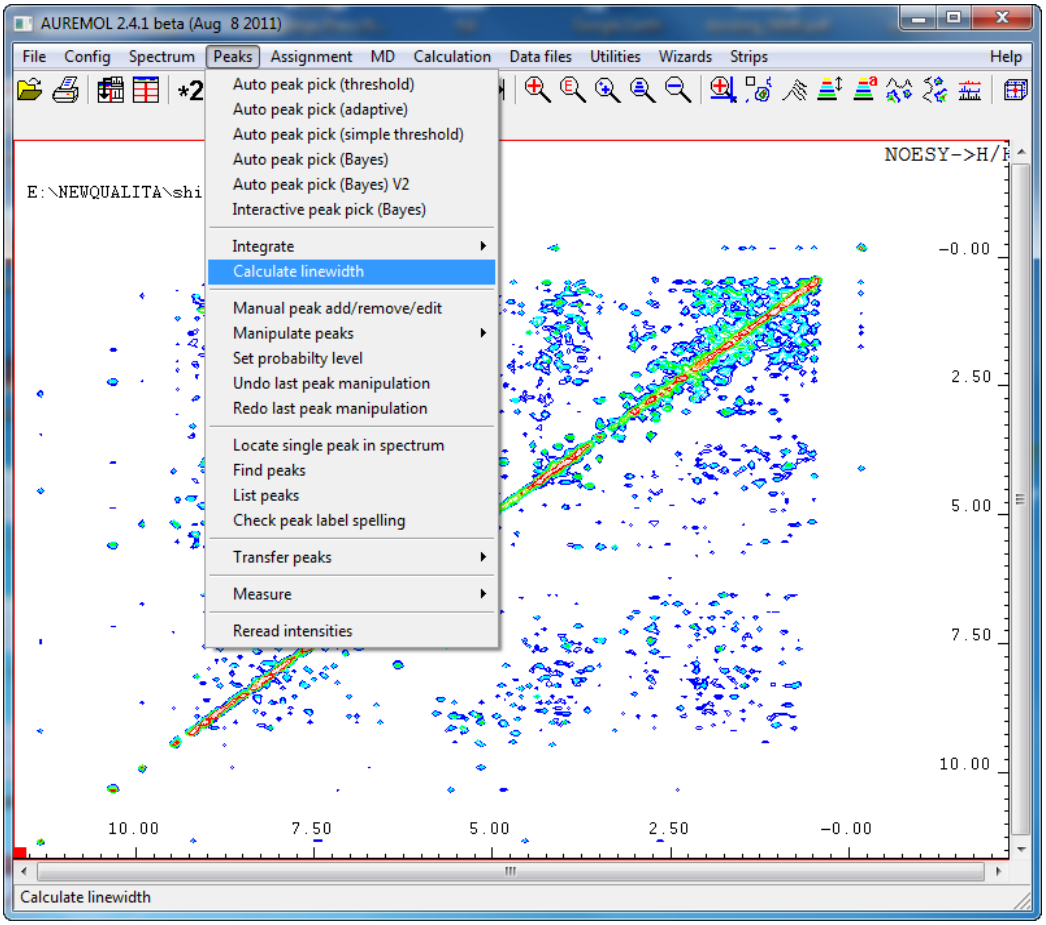


Fig. 3.6 Computation of the line width of all the spectrum peaks (batch mode): All the spectrum peaks are used to compute each line width along all the measured directions.

In case that the user needs to compute only the line width of a specific spectrum the command “LW of a Peak” under the AUREMOL menu “Calculation” can be used, as shown in Fig. 3.7.

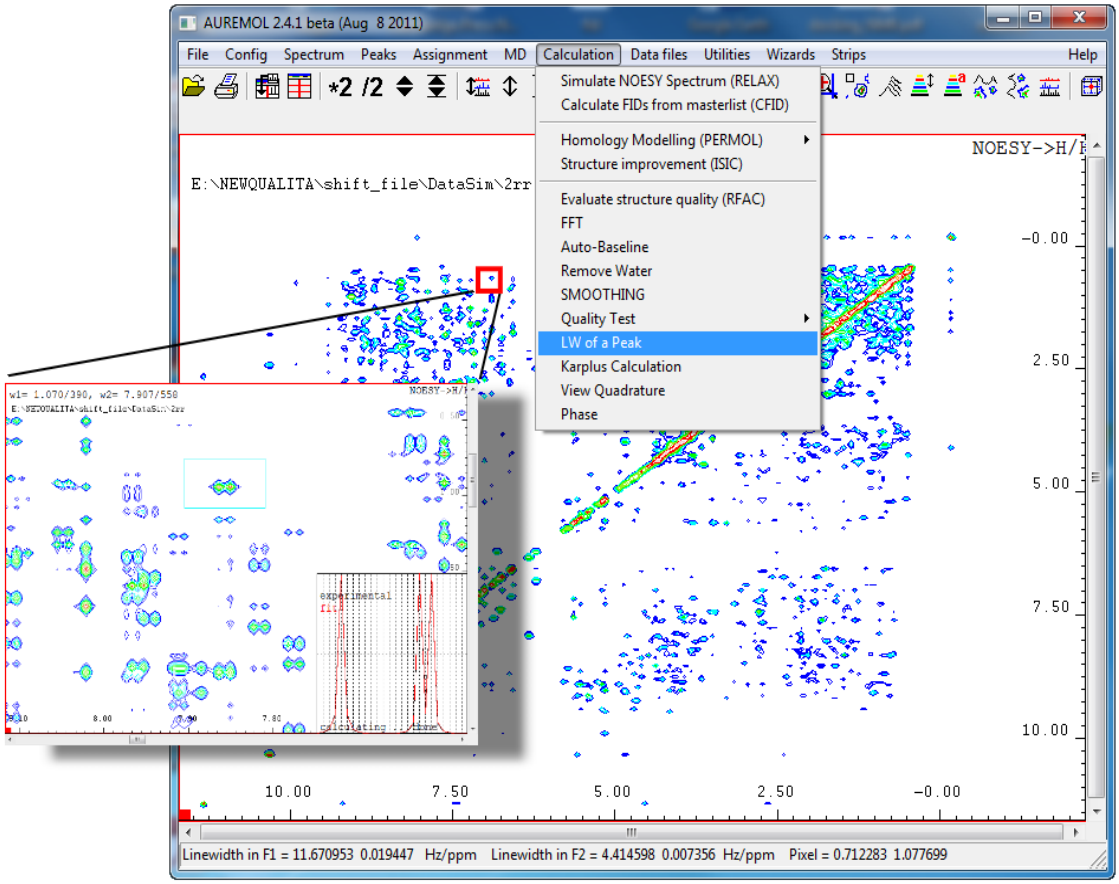


Fig. 3.7 Line width calculation of a single peak: The routine measures the line width only for the selected peak (HG2 80/HN 80). The result of the computation is visible on the down side of the AUREMOL main window. This result is given in the Hz, ppm and voxel units. In this example the inhomogeneous line width of a doublet peak has been computed.

3.4 Implementation of the AUREMOL-QTA module

3.4.1 The pre-processing level

As explained before, the AUREMOL-QTA is developed on four different levels. At the very beginning, the baseline of the spectra is corrected [Güntert & Wüthrich, 1992] and the solvent can be suppressed through the novel developed routine AUREMOL-SSA/ALS [Malloni et al., 2010; De Sanctis et al., 2011]. This routine eases the tasks of baseline

correction and solvent suppression since they are automatically performed without any user intervention.

The first stage is the pre-processing that involves the parameter standardization among the spectra (SW, OFF, NS, RG, NC_proc), the management of general or global spectral shift differences and of multiple referencing cases as resumed in Fig. 3.8.

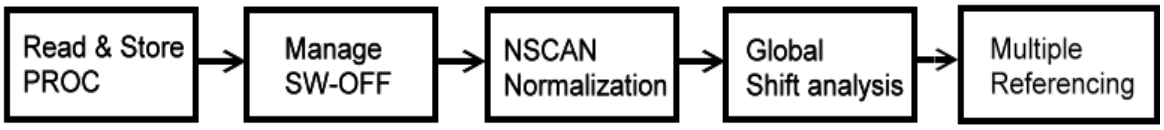


Fig. 3.8 AUREMOL-QTA pre-processing diagram: it starts reading and storing the spectrum parameters, it manages eventual spectral width and offset differences, it normalizes peak volumes according to number of scans, the receiver gain and the intensity scaling factor. It calculates, if present, the spectrum shift and it considers the presence of a set of reference spectra.

3.4.1.1 Standardization of the external parameters

The first phase of the pre-processing step starts reading all the processing parameters of each spectrum, it continues storing them into memory and in the case of spectral width and offset differences the routine is able to perform a conversion of the coordinate system. In particular, the ppm positions of each peak are converted to the correct voxel coordinate positions (integers) in order to obtain the exact raw data points to be used during the whole comparison procedure. The conversion is done using the following formula:

$$P_{m_c} = P_{s_c} = \frac{-\mathrm{P_{ppm}} + \mathrm{O}}{\mathrm{SW/SI}} \pm 0.5 \qquad (3.1)$$

where P_{m_c} is the central position of the multiplet in the voxel unit (integers), P_{s_c} is the central position of a peak singlet in the voxel unit, O is the offset of the spectrum (OFFSET parameter in the Bruker format), SW is the spectral width and SI is the size of the considered test dataset. The above mentioned calculation in case of multidimensional experiment is repeated over all the directions.

Successively the NS, the RG (receiver gain) and the NC_proc (intensity scaling factor) parameters are considered since the volume of the peaks is proportional to them. The peak volume P'_V of the considered test spectrum is adapted in the following manner:

$$P'_V = 2^{(C_T - C_R)} \left(\frac{G_R\, S_R}{G_T\, S_T}\right) P_{Vo} \tag{3.2}$$

where T and R represent the test and the reference spectrum respectively. The term C_T and C_R represent the NC_proc (intensity scaling factor) parameters of the test spectra T and R, $G_{R,T}$ identifies the RG (receiver gain) parameters whereas the $S_{R,T}$ represents the NS (number of scans) parameters. The term P'_V is the peak volume of the test spectrum scaled after the NS, the RG and the NC_proc normalization with respect to its initial peak volume P_{Vo}.

Afterwards, the recognition of real resonances of interest and possible remaining artifacts like noise, baseline points and solvent artifacts is carried out. For instance, the identification of classes of peaks is done using pre-existing AUREMOL modules such as AUREMOL peak picking [Antz et al., 1995], whereas the segmentation of those peaks, if not previously calculated, is computed by the AUREMOL integration routine [Geyer et al., 1995].

3.4.1.2 The global shift of the spectrum

A common underlying problem is the fact that different measured test spectra could display a spectral shift (GSHIFT): the so called “global shift”. It is not a shift of isolated peaks or a “local shift” but a tedious shift stretched out over the whole test spectrum due to errors in referencing or slightly different measurement conditions (e.g. temperature, pH and pressure). The user defined global shift parameter (in the ppm unit) of the spectrum is used in order to set the maximum allowed shift among investigated spectra in the voxel unit. It is mandatory to identify this shift because it might negatively influence the QTA procedure if not properly managed. It is possible to imagine the worst scenario where all the peaks are shifted by a few voxel points. In this condition not considering this hidden

effect might lead to a totally wrong result. The following diagram explains how to deal with this problem:

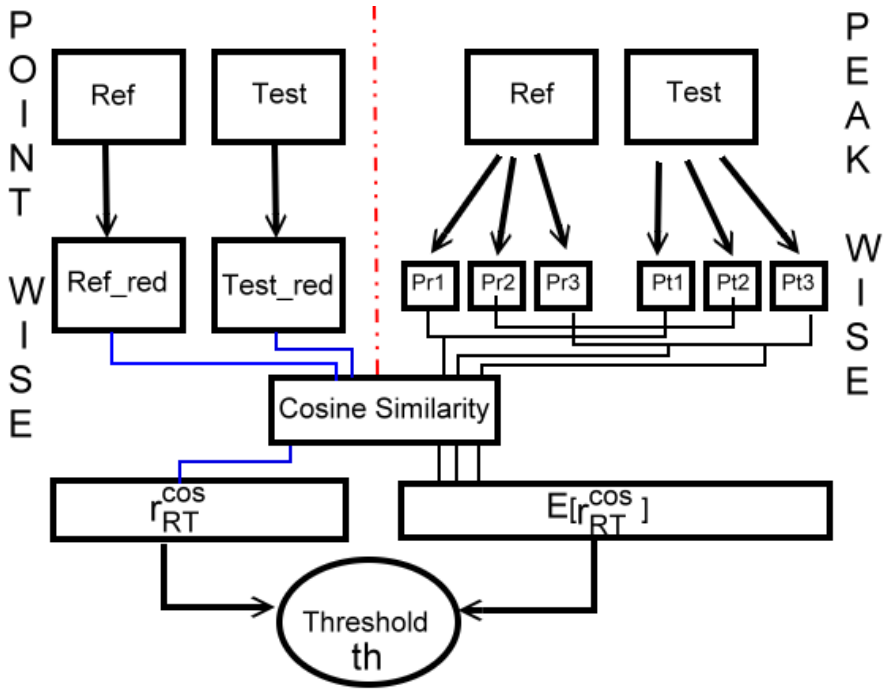


Fig. 3.9 Global shift schema: Two-way algorithm used to check the presence of global shift; point-wise cosine similarity (left side) and peak-wise cosine similarity (right side).

As shown in Fig. 3.9 the algorithm uses a two-way system. It takes as input the reference spectrum and the test spectrum. On the left side of the figure, the algorithm applies a reduction of both spectra. This reduction creates a temporary empty spectrum where each peak is segmented down until the 95% of the maximum intensity is reached and then stuck upon the empty dataset. This generates a temporary "pseudo noise free" spectrum. The same procedure is repeated for the test spectrum case. Once both reference and test spectra are reduced the cosine similarity criterion is calculated as following

$$r_{RT}^{cos} = \frac{\sum_{X=1}^{N} P_{X_R} \times P_{X_T}}{\sqrt{\sum_{X=1}^{N} {P_{X_R}}^2} \times \sqrt{\sum_{X=1}^{N} {P_{X_T}}^2}} \tag{3.3}$$

and $\mathrm{N} = \prod_{\mathrm{i}=0}^{\mathrm{dim}-1} \mathrm{SI_i}$ (3.4)

where r_{RT}^{cos} is the cosine similarity score, P_{X_R} and P_{X_T} are the intensities of each single voxel (P_X) contained into the reference *R* and test spectrum *T*, whereas N represents the total number of spectrum voxels. The parameter $\mathrm{SI_i}$ identifies the spectrum size along the measured direction *i*.

The dot product is applied point-wise over the whole reduced dataset. This new spectrum does not contain noise peaks and the majority of strong solvent artifacts are removed. The method is reliable but for safety reasons, either in case of no solvent suppression or of a very noisy spectrum, the algorithm calculates for each peak in both spectra the dot product between pairs of peaks (i.e. one peak from the reference spectrum and the other one from the test spectrum at the same ppm position considering possible offset and spectral width differences). Every peak has been previously segmented down until the 95% of the maximum intensity is reached. From this reduction the routine computes the size of each reference peak. The routine computes M cosine similarity distances between each reference-test peak pair (at the same ppm position). M is the number of peaks of the reference spectrum *R*. Every dot product $r_{j\,RT}^{cos}$ of each pair *j*, with j = 1,…M is saved into an array of temporary results that is used to calculate the mean cosine similarity value $\overline{r_{RT}^{cos}}$. This value has been computed as following:

$$\overline{r_{RT}^{cos}} = \frac{1}{M}\sum_{j=1}^{M} r_{j\,RT}^{cos} \tag{3.5}$$

The last step of the algorithm is a threshold comparison of both result, one coming from the point-wise similarity and the other result from the peak-wise one. In particular, if $\left|r_{RT}^{cos} - \overline{r_{RT}^{cos}}\right| > th$, ($th$ has been empirically set to 0.66) the routine informs the user with a warning message that the difference is too pronounced and that is possible either continue or abort the whole computation. The routine computes $(2S+1)^{DIM}$comparisons (point and peak-wise) where S is the maximum user allowed spectrum shift (in the integer voxel unit) and DIM is the dimension of the investigated reference spectrum. In addition, each r_{RT}^{cos} value is stored and used to optimize the global shift of the spectrum. In

particular, the routine selects the global shift value with the highest r_{RT}^{cos} value (in the ppm unit). Empirical tests have shown the ability of this kind of procedure to avoid either artifact or noise alignments between spectra under extreme conditions.

The method has been tested over a proper simulated dataset (see the second dataset described in par. 2.2.1.1) demonstrating a perfect reliability recognizing the correct spectrum shift amount in all the cases. An example of the comparison between the reference case and the test one is shown in Fig. 3.10.

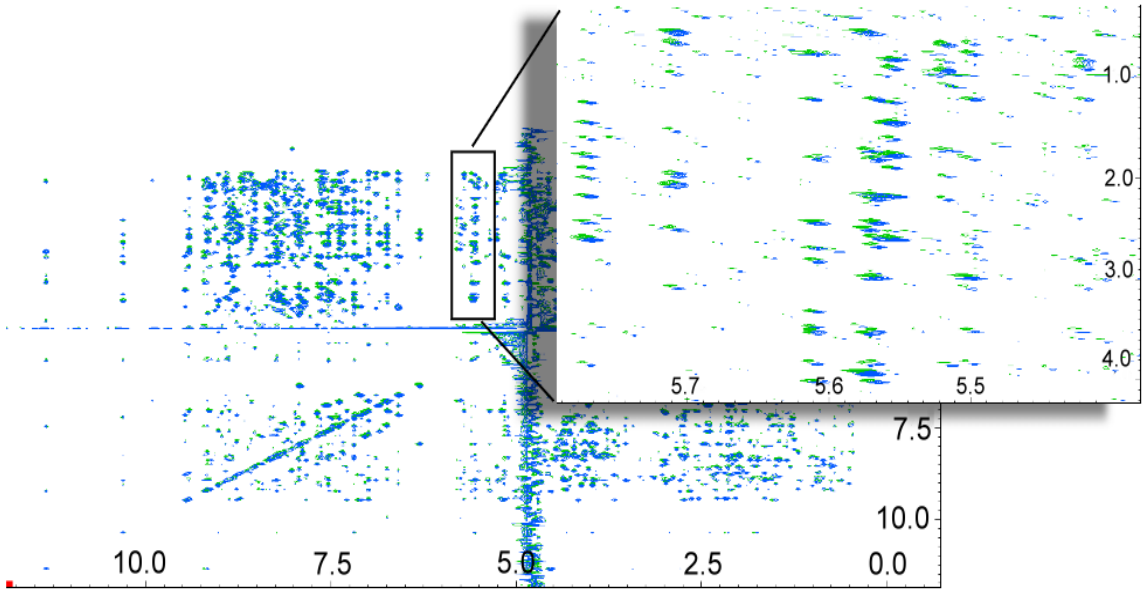


Fig. 3.10 Global shift reliability test: The second dataset described in par. 2.2.1.1 has been used in order to compare the not shifted reference spectrum (blue) with the test spectrum (green) that has been shifted by two voxels along δ_2 and one voxel along δ_1 direction.

The previous described algorithm has been used to control the global spectrum shift on multidimensional spectra.

3.4.1.3 Multiple spectrum referencing

The pre-processing module ends evaluating the number of spectra involved in the comparison. The QTA allows using more than one reference spectrum. In this case the program creates a temporary pseudo spectrum without peaks that is filled only with matching segmented peaks among at least two considered reference spectra as shown in Fig. 3.11. In particular, each peak of each reference spectrum is searched in all other reference spectra and it is added to the final pseudo spectrum if the same peak has been found (considering its voxel position $P_c \pm 1$) in at least two reference spectra. The purpose is to avoid ambiguities possibly coming from additional noise or artifact peaks. The proposed solution provides a more accurate and stable analysis.

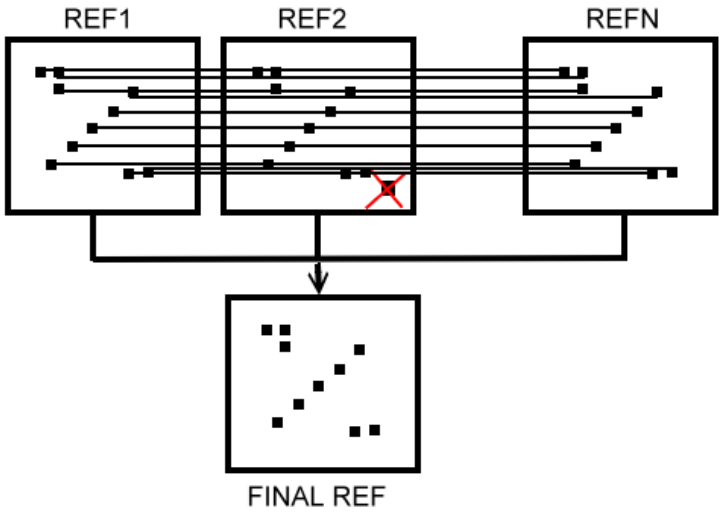


Fig. 3.11 Example of a quality control with multiple reference spectra: the presence of more than one reference spectra is managed considering only peaks found at least in two spectra.

In case that one reference peak has been found in at least another reference spectrum, a volume average value has been computed and stored. The same procedure is performed for the maximum peak intensity P_I.

3.4.2 Low-level analysis: a bottom-up approach

Once the pre-processing part is done, the QTA runs into the most computational time demanding task. The low-level analysis starts from the “peak level” point of view, so each peak is considered as an ideal viewer of the surrounding. The idea behind is to look for intrinsic properties on a “microscopic” scale trying to obtain a wider information literally known as a “bottom-up” approach. The low-level module starts to collect all information related to every peak in each spectrum. As shown in Fig. 3.12, this type of analysis considers each single modification of every evaluated peak as an important feature. For example, it takes into account volume changes, line width variations, shape behavior, splitting of the peaks tops, changes of relative positions in all measured directions, the symmetrical properties in case of a NOESY or TOCSY experiment, the relative peak intensities and the cross-correlations. They are analyzed by the AUREMOL-QTA method trying to obtain a wider point of view for all the considered peaks. Three possible cases are expected: the ideal case where all reference and test spectra are assigned, the case where the reference spectrum is assigned and all other ones are unassigned (most frequent) and the worst case where all the spectra are not assigned, neither references nor tests. The first option would provide more accurate information about the major differences between the compared data. If the user provides the three-dimensional structure of the studied protein (PDB file) the routine automatically back-calculates the desired spectrum according to all other experimental parameters, as explained in the introduction of this chapter, and uses this spectrum as an additional strong source of information. From this scenario, it is possible to extract relative volume fractions, peak assignments and J-coupling splitting at least. Each module of the low-level presented in Fig. 3.12 is a collection of other sub-modules that are described separately later.

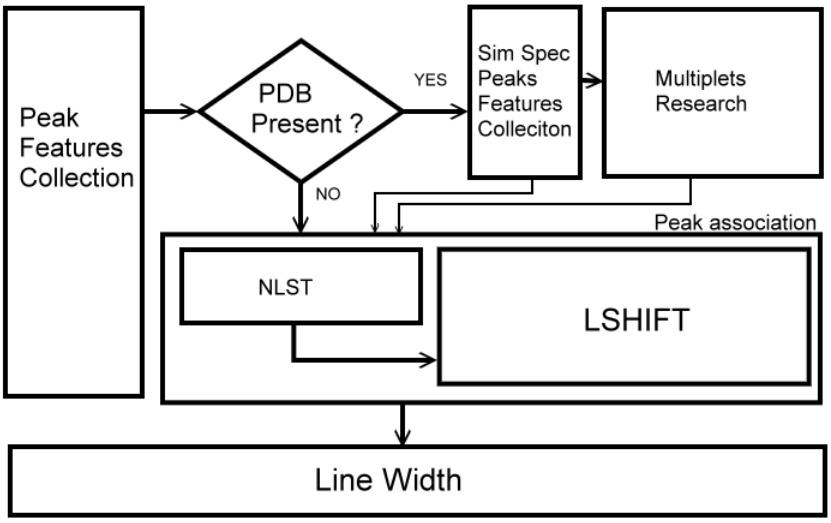


Fig. 3.12 The low-level analysis schema: the method starts collecting peak features (volume changes, line width variations, shape behaviors, position shifts and cross-correlations). If the three-dimensional molecular structure is present, the simulated spectrum is back-calculated and the method looks for multiplets inside all other spectra. In addition, the NLST (explained in par. 3.4.2.3.1) computed and the local shift of each peak is analyzed. Successively, the line width of each peak is calculated and used for comparison purposes. The time correlation is obtained by means of an inverse hypercomplex Fourier transformation. The size of the sub-modules is related to the required computational time.

3.4.2.1 Collection of peak features

All single peak features are collected by this module. Its intention is the attempt to accumulate so much information as possible from each considered element. During this step the AUREMOL-QTA routine creates all the data structures for saving the information needed successively to perform the analysis. For instance, during this phase QTA tries to know if a peak has been already assigned, if it has a volume integral P_V and in that case the integration error due to the presence of noise on each spectrum. The calculated point-wise local noise P_N [Trenner PhD, 2006] is stored to be used successively when the volume analysis is required. The parameters related to the ppm and voxel positions as well as local intensities are stored during this procedure. At this point, if the user has previously provided a 3D structure, AUREMOL-QTA starts to back-calculate the spectrum using the existing routine RELAX-JT2.

3.4.2.2 Search of multiplet peaks

One of the most challenging tasks of the whole QTA project is the possibility to find multiplets on measured spectra. This step is performed only if the three-dimensional structure and the chemical shift list are available (it is possible to back-calculate the spectrum of the investigated protein). The peak splitting behavior is the result of the J-coupling effect. It is a strong through-bond relationship measure. It represents the way of interaction between nuclei independently of the externally applied magnetic field. In case of peak splitting, a volume problem arises. Once the spectrum peaks are found and segmented, the critical situation where a peak multiplet is not recognized (by the peak-picking routine) should be taken into account. An example of this situation is reported in Fig. 3.13

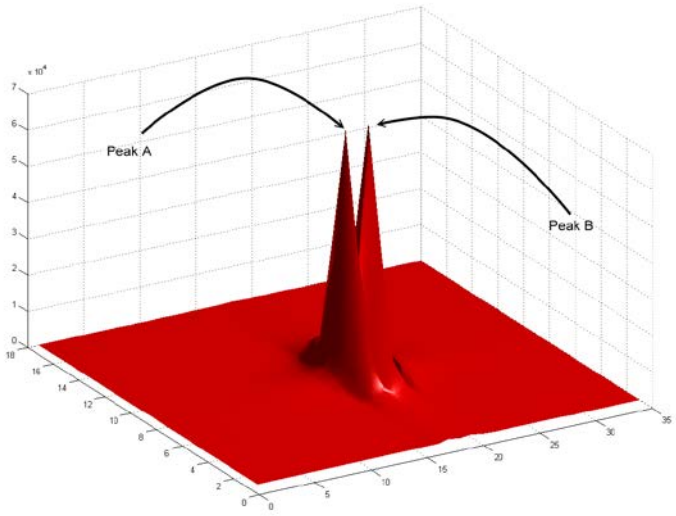


Fig. 3.13 Result after the automatic peak-picking routine of a peak multiplet: the doublet (HB3 36/HN 36 of the measured HPr protein from *Staphylococcus aureus*) is not recognized and instead of a single peak it is considered as two singlets (peak A and B respectively). The computed volume is not correct and the J-coupling information is missing.

As shown in Fig. 3.13, the automatic peak-picking routine is not able to distinguish between a peak doublet and two singlet peaks. The volume of the peak HB3 36/HN 36 is

shared between the peaks A and B. This leads the AUREMOL-QTA routine to numerous unwanted errors, both in the spectrum assignments and in the final structural calculations.

To solve these tedious problems the Multiplet Search Analysis (MSA) algorithm has been proposed. It relies on the comparison between the back-calculated spectrum obtained through the RELAX-JT2 module and the experimental ones. After having generated the simulated spectrum an additional file, the *"peakforms.dat"* file, is available. In this file each simulated peak is separately stored as a "box patch" (in a 2D case it is the square that defines the peak extension area). As reported in Fig. 3.14, this patch represents each peak before being merged into the spectrum with all others. In the output file (peakforms.dat) it is possible to collect important peak information like the peak assignment, the relative position (in voxel and in the ppm unit) the extension area of the peak (the so-called box patch), all the sub-peaks intensities that have been simulated and their total number for each peak. In addition, it is possible to extract the J-coupling information (splitting), the information of back-calculated volumes P_V and relative intensities P_I.

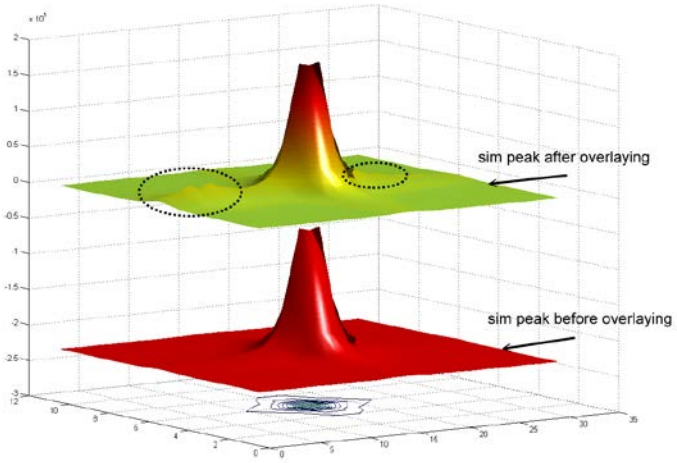


Fig. 3.14 Difference between the simulated peak before and after merging: The ideal simulated peak (red) and the same peak (colored shading) after having merged all the spectrum peaks (for generating the final simulated spectrum). After overlaying all the peaks in the same peak region some ridges are visible due to the presence of neighbor peaks.

The *Multiplets Search Analysis* module is made up of three different stages, the *three step analysis* (3SA) as shown in Fig. 3.15.

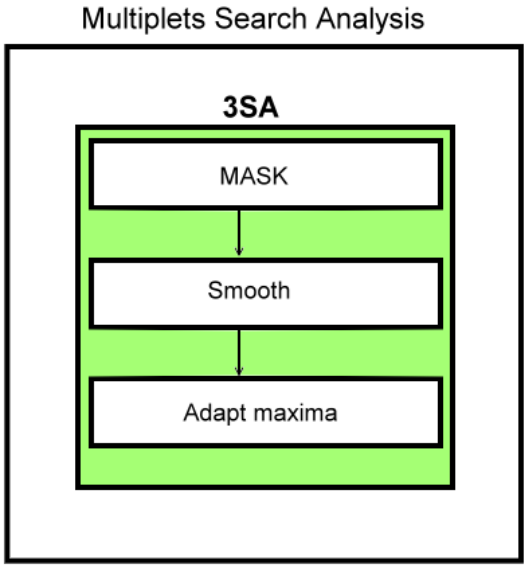


Fig. 3.15 The Multiplets Search Analysis module: the module is made up of the three steps algorithm (green highlighted). The peak MASK is computed in order to look only for peaks that are not superimposed. Successively, it is necessary to center the simulated peak relative to the experimental one after a dynamic filtering procedure performed through the Savitzky-Golay smoothing filter. Once the center of the multiplet has been found it is possible to locally adapt the peak maxima between the simulated and the experimental dataset.

The first step computes the peak *mask*, then the peaks (the simulated and the experimental one) are dynamically smoothed (zero order). This has been done in order to match their centers P_{m_c} in a correct manner. The last step looks for shifted peak maxima/minima all around. Each of these steps is analyzed in detail in the following paragraphs.

3.4.2.2.1 Multiplet search analysis: the *mask*

The reason of this step is that sometimes a multiplet peak with a relative small intensity with respect to its neighborhood is superimposed under other peaks. The algorithm starts to look for (into the back-calculated spectrum) peak maxima and minima with a local multidimensional search. This search is conducted for each simulated isolated sub-peak before merging it with all other sub-peaks in the simulated spectrum. This search is started

from the center of the simulated peak generating a peak *mask*. This mask (a multi-dimensional array) contains only the positions of peak maxima and minima that are not superimposed. This mask is then used to find the same peak structure in all other investigated spectra. As shown in Fig. 3.16, a relative small doublet has been recognized by the peak *mask*, also being partially superimposed by a tail of a neighbor peak.

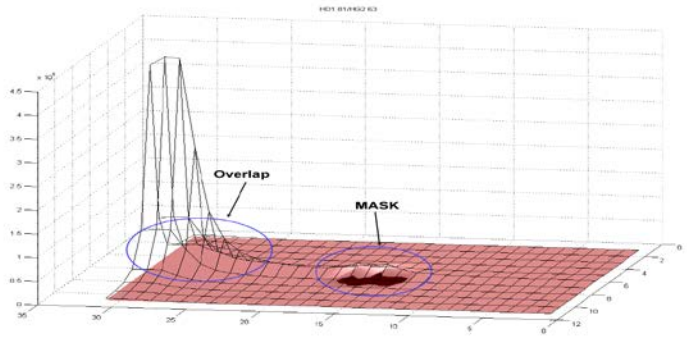


Fig. 3.16 Application of the peak *mask*: the simulated peak doublet HD1 81/HG2 63, of HPr protein from *Staphylococcus aureus* has been recognized through the *mask* procedure under pronounced neighbor peak overlapping.

According to the figure described above, if the same search procedure had been applied to the final simulated peak (where all the peaks and sub-peaks are superimposed together in the final spectrum) the algorithm would have found three maxima leading to a misinterpretation of the peak multiplet structure.

3.4.2.2.2 Multiplet search analysis: the dynamic data smoothing

The second step is needed in order to match the center of the simulated peak with the experimental one. The alignment algorithm, following the highest peak similarity (based on the peak shape) has empirically proven that sometimes wrong alignment tendencies occur. In particular, this happens in case of experimental multiplets with more than two maxima. In addition, the effect of the spectrum global shift (see par. 3.4.1.2) should be

considered. To overcome this limitation a digital smoothing filter has been proposed. The Savitzky-Golay [Savitzky & Golay, 1964] method is a polynomial FIR filter (low pass) that can be used to smooth the data maintaining higher moments unchanged. In particular, the peak position after filtering is preserved (the first moment). It is performed via a local averaging movement of a "time window" (leftward for the past and rightward for the future) fixing the highest conserved moment equal to the order of the smoothing polynomial. During this work, the filter width (crucial for the smoothing grade) has been estimated and dynamically adapted to the spectral resolution. In particular, the filter width has been imposed to be equal to the line width of each simulated peak as following:

$$n_L + n_R = (\Delta v_{\frac{1}{2}} - 1) \tag{3.6}$$

where n_L is the number of the points used leftward, n_R is the number of data points used rightward and $\Delta v_{\frac{1}{2}}$ is the full width at half maximum (FWHM) in the unit of number of voxels. The sum of the terms n_L and n_R has to be an even number. The polynomial order has been set to zero in order to increment the smoothing factor. Once the simulated and the experimental peaks have been smoothed, the alignment algorithm is applied looking for the position (of the experimental peak) showing the highest cosine similarity. If the similarity is major than an empirical value of 0.7 the centering procedure is accepted. A proper center alignment of a triplet by means of a Savitzky-Golay smoothing filter is shown in Fig. 3.17. The smoothing procedure has allowed a perfect matching of the two considered datasets. One limitation of this smoothing procedure is evident in case of antiphase absorptive cross-peak (COSY experiment).

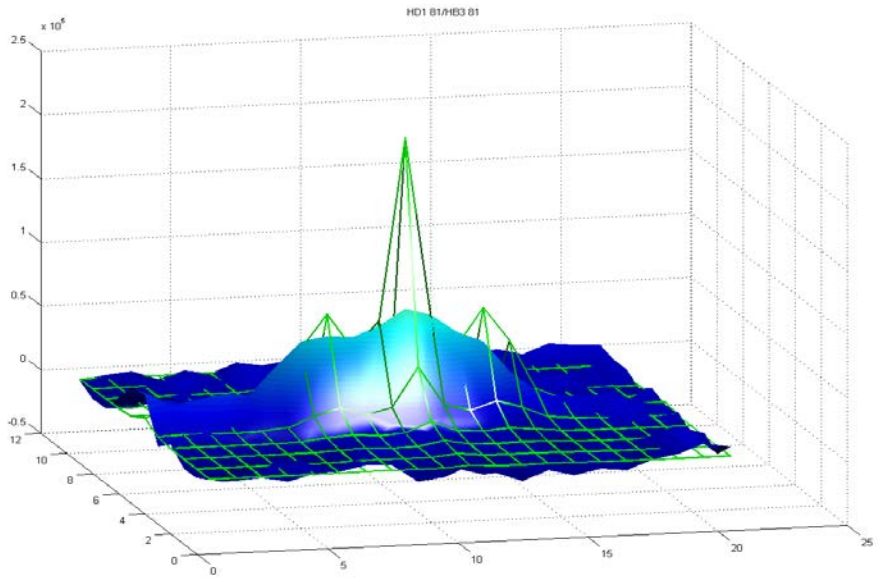


Fig. 3.17 The correctly recognized peak triplet of the simulated and the experimental peak HD1 81/HB3 81: this peak of HPr protein from *Staphylococcus aureus* has been correctly recognized as peak triplet in the test spectrum. Proper center alignment between the simulated (green) and the experimental peak (shading blue) by means of a Savitzky-Golay smoothing filter applied on both dataset.

3.4.2.2.3 Multiplet search analysis: local adaption of peak maxima

After the centering procedure, it is then mandatory to adapt the new centered initial *mask* over the test peak case in a sort of "local adapting problem" trying to match peak maxima between simulation and experimental spectra, as much as possible. In this step the routine calculates the simulated distance (J-coupling) between adjacent multiplet peaks and it verifies, in the range of a local shift ($\pm$ 1 voxel) if it is maintained in the experimental spectra. The heuristic that has been maximized in order to adapt the mask to the experimental peak is described in equation 3.7. In Fig. 3.17 it corresponds to a matching movement just of the lateral maxima keeping the center unmovable. The algorithm computes the following heuristic:

$$h = \left(\frac{N_f}{N_M}\right)\left(1 - \frac{E_f}{E_M}\right) \tag{3.7}$$

where N_f is the number of maxima/minima that have been found in the experimental spectrum, N_M is the number of maxima and minima belonging to the *mask* and E_f identifies the sum of all the Euclidean distances that have been found between maxima of the experimental and the simulated peak. These distances have been computed considering the nearest maximum/minimum between the experimental and the simulated datasets. E_M represents the maximum allowed Euclidean distance as described in eq. 3.8.

$$E_M = m_M\sqrt{D} \tag{3.8}$$

where D is the spectrum dimension and m_M is the total number of maxima and minima of the *mask.*

A graphical example of the Multiplets Search Analysis is shown in Fig. 3.18 where the simulated peak HG2 62/HD 64 has been recognized as a multiplet in an experimental spectrum (test spectrum). In particular, the (a) part of Fig. 3.18 shows the initial matching between the two datasets. It is evident, from the (b) part, that after filtering the multiplets structures are temporarily destroyed but the centers are optimally aligned between the two spectra. The optimized positions of the centers are stored and used to perform the last step involving the alignment of all the maxima surrounding the centers. As shown in (c), the simulation correctly locates the chemical shift of the peak HG2 62/HD 64 in the middle of the multiplet structure involving proper volume and line width parameters. The proposed method identifies multiplets in experimental spectra where no peaks have been picked near the same ppm positions and where they have been wrongly picked (i.e. possible maxima of multiplets are not identified as a unique peak). This avoids eventual problems related to wrong volumes and line widths related to experimental multiplet structures. For example, the doublet structure has been found in the experimental reference spectrum even without any picked peak in that region, whereas in the experimental test spectrum the multiplet shape is not identifiable and it is recognized as a singlet. Since the back-calculated data is available, the simulated doublet located in the range of ± 1 voxel local shift correctly allows the doublet identification in the experimental cases.

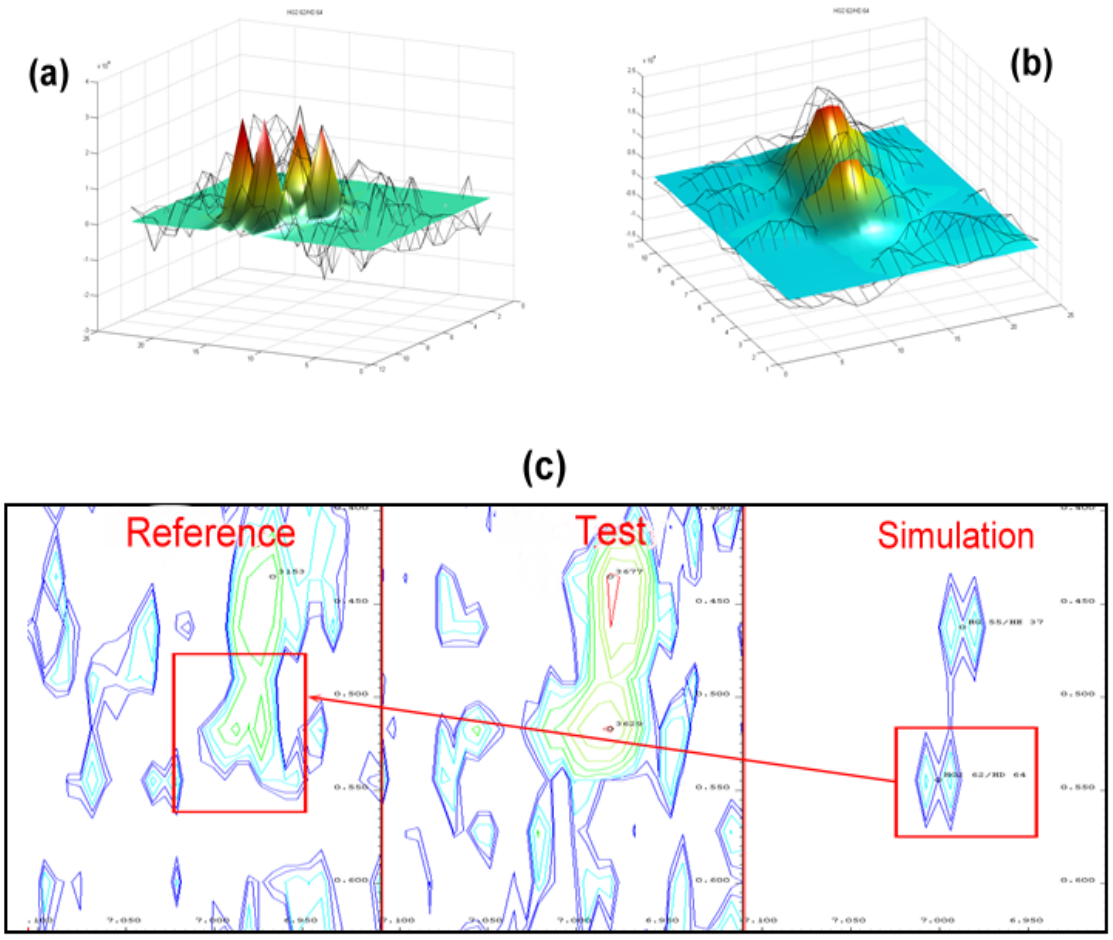


Fig. 3.18 **Multiplet recognition:** the experimental peak (black lines) and the simulated one (colored shaded) before (a) and after (b) the smoothing procedure. After centering, the *mask* has been adapted in order to match as much as possible the experimental peak. (c) The simulated peak is recognized in the experimental reference spectrum (HPr protein from *Staphylococcus aureus* wild type) whereas is not recognized in the experimental test spectrum (HPr protein from *Staphylococcus aureus* mutant) using the simulated spectrum of HPr protein from *Staphylococcus aureus* wild type.

In conclusion, the *mask* is used at the first step to find multiplet peaks that are not superimposed in the simulated spectrum and at the last step, after have aligned the centers through the Savitzky-Golay filter, to locate multiplet maxima in the experimental spectrum.

3.4.2.3 Analysis of the features for associating peaks

Once the multiplet peaks are properly recognized it is possible to collect more information from every signal appearing in the spectrum of interest independently on the associations of peaks among the considered spectra. In particular, through the AUREMOL integration routine the volumes of all peaks are extracted. Using the AUREMOL peak picking module, the chemical shifts and peaks intensity are stored, while the line width information and the time domain features are obtained respectively with the AUREMOL-LW and AUREMOL-FFT calculation developed during this work. All the previously mentioned features are initially collected separately in every considered spectrum. Once this has been done, they must be merged in order to have a wider view of the comparison. The hardest problem is related with the decisional algorithm designed to define which test peak is associated to certain reference peak and vice versa. The developed method builds connections based on shape, volume, position and comparisons of peak patterns, via a local and an external matching through the definition of an optimal peak surrounding region (i.e. local shift of the peak driven by a list of peak neighbors).

3.4.2.3.1 The list of neighbors (NLST) and the neighborhood distance

In order to obtain a reliable peak association between different spectra the neighborhood information of each peak has been analyzed. A minimum averaged neighborhood distance has been computed allowing a peak connection between spectra not only based on some peak features but also considering the peak neighbors. This analysis has been performed considering each peak as the center of a temporary cluster and the neighborhood distance represents its maximum length. The routine computes the minimum neighborhood distance and the list of neighbors (NLST) of each spectrum peak. This is done for two reasons:

1. instantly have a list of neighbors for each considered peak
2. have the maximal multidimensional distance between the considered peak and the neighbors

The first reason arises from a computational bottle neck due to the fact that peaks having similar index in the total spectrum master list does not mean spatial proximity in the multidimensional spectrum. In order to obtain a list of neighbors of each considered peak a loop through the complete master list (list of all the spectrum peaks) is required. The calculation of the list of neighbors (NLST) avoids that problem. The computed list (one for each peak of each spectrum) contains all information related to peak neighbors saving a lot of CPU cycles. The second reason upraises from the need to define a limit on the number of neighbors. Initially, the module was thought to calculate neighbors until the maximum distance of 0.3 ppm is reached not allowing the analysis of hetero-nuclear spectra. A proposed solution is to use a statistical approach to calculate for each peak P_j the minimum distance d_i^j (in the voxel unit P_{int}) to the next neighbor peak, with

$$d_i^j = \left| P_{I_j}(\delta_i) - P_{In_j}(\delta_i) \right| \tag{3.9}$$

where i represents the spectrum direction and i = 0,..,dim-1. P_{I_j} is the considered peak (having the maximum intensity P_I) with j = 0,.., N-1 where N is the total number of spectrum peaks. The resulting matrix

$$\boldsymbol{D} = \begin{bmatrix} d_i^j & \cdots & d_i^{N-1} \\ \vdots & \ddots & \vdots \\ d_{dim-1}^j & \cdots & d_{dim-1}^{N-1} \end{bmatrix} \tag{3.10}$$

is used to calculate the vectors representing respectively the mean and the standard deviation for each dimension row-wise as following:

$$\vec{\mu} = \begin{pmatrix} \sum_{j=0}^{N-1} d_i^j \Big/ N \\ \vdots \\ \sum_{j=0}^{N-1} d_{dim-1}^j \Big/ N \end{pmatrix} \tag{3.11}$$

$$\vec{\sigma} = \begin{pmatrix} \sqrt{(d_i^j - \langle d_i^j \rangle)^2 / (N-1)} \\ \vdots \\ \sqrt{(d_{dim-1}^j - \langle d_{dim-1}^j \rangle)^2 / (N-1)} \end{pmatrix} \quad (3.12)$$

Using the above formulas the optimal box size surrounding each peak is defined in this manner:

$$\vec{d}_{box} = 2(\vec{\mu} + \vec{\sigma}) + 1 \quad (3.13)$$

Applying such algorithm, the procedure is deadlock free, indeed every peak on each spectrum is considered only once and the neighbor growing search is trustworthy. A graphical representation of this algorithm is shown in Fig. 3.19.

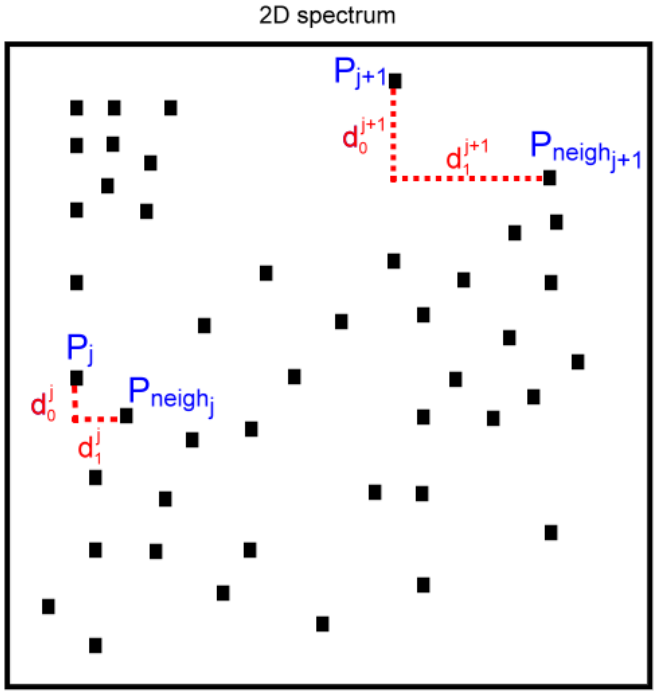


Fig. 3.19 The computation of the neighborhood distance: for all the spectrum peaks the algorithm stores the minimum distance for all the directions (dashed red) and uses this information in order to statistically compute the minimum neighborhood distance.

The calculated multidimensional box (in a 3D experiment is a cube) represents the optimal neighborhood conformation of the reference peak that can be used as additional information for successive peak associations.

3.4.2.3.2 The analysis of peak local shift

The most important step of the whole QTA, in absence of any peak assignments, is the *Local shift analysis* module, called here *LSHIFT*. This module is responsible for selecting the correct association between peaks belonging to different spectra in case of partially assigned spectra. The *LSHIFT* module automatically finds peak connections under two main different conditions:

- standard quality control where most of the peaks are not changed (between different spectra)
- advanced quality control under varied external conditions (e.g. pressure, temperature, pH, binding and mutation) where some or most of the spectrum peaks have changed their position and/or their characteristic features (e.g. volume, line width and maximum intensity).

Logically, using assigned reference and test spectra the routine has just to look for the same assigned peak in all other spectra, storing the needed parameters (e.g. volumes, shifts and intensities). It automatically warns the user in case that the same peak assignment has been found more than a user defined maximum ppm distance away among the spectra (in any direction).

The importance of this module in the second case is clear considering the hypothesis of perfectly aligned spectra, where each single peak can vary, for different reasons its ppm positions in one or all directions contemporarily. Not considering this effect, may cause a totally wrong final result due to the false belief that peaks are missing in one or more spectra. In particular, the amount in ppm of local shift allowed is set by the user at the very beginning of the QTA. Starting from this value, the *LSHIFT* module needs to evaluate all the peaks in the test spectra that are falling in this range, in order to associate the reference

peak with its correct corresponding one in the test. That is particularly difficult in very crowded regions.

As shown in Fig. 3.20, the standard quality control search is performed through the module STDS (standard search) whereas the advanced one is computed through the ADVS (advanced search). The STDS module looks for peak associations having the same ppm positions $\pm$ 1 voxel all over the spectrum in order to compute the volume scaling factor S_{ik} between spectra i and k respectively. After having computed the volume scaling factor, the ADVS module (if unassociated peaks have been detected) performs a wider peak association search involving the cosine criterion of associated peaks, the neighborhood distance (NLST) calculation, a comparison of peak patterns and a backtracking step.

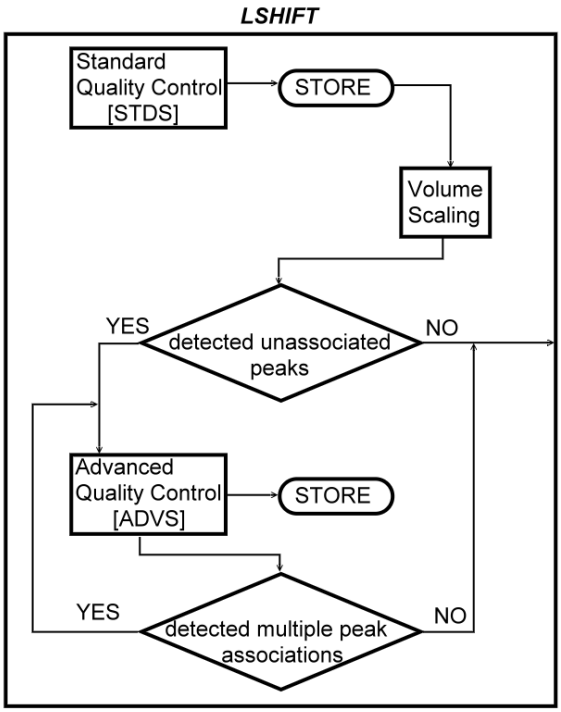


Fig. 3.20 The local shift algorithm (LSHIFT) schema: the standard quality control module (STDS) looks for peak associations in the deviation of the digital resolution. These previous associations are used in order to compute a volume scaling factor. Successively, the advanced quality control module (ADVS) performs a wider peak associations search involving shape, neighborhood and peak pattern matching. After the advanced quality control a backtracking loop is applied if multiple peak associations are still detected.

3.4.2.3.2.1 The *local shift* algorithm: the first peak associations and the volume scaling factor

In the standard quality control a reference spectrum (or a set of reference spectra) of the measured molecule is compared with one or many other test spectra. If the investigation is conducted on a set of test spectra every test spectrum is compared with the reference spectrum separately.

Every reference peak is searched in the test spectra at the same chemical shift position. If a reference peak has been found in the test spectrum at the same relative position (at the same chemical shift position considering eventual different spectral widths and offsets) the routine stores this new peak association. This association is unique between the reference and the test peak. This analysis is performed in a range of $\pm$ 1 voxel along each measured direction considering the possible spectral shift effect (GSHIFT, see par. 3.4.1.2) due to a wrong referencing between spectra. Once the routine has stored all the possible connections and the relative features (the volume, the maximum intensity, the ppm and the voxel positions in the multidimensional space) of each considered peak couple, it is possible to compute the volume scaling factor. It is calculated between the investigated spectra using volume ratios of the previously associated peak couples.

The volume ratio s_j is computed on the whole peak-by-peak correspondence over all the previously associated peaks in this manner:

$$s_j = \left(\frac{P_{V_T^j}}{P_{V_R^j}} \right) \tag{3.14}$$

where $P_{V_T^j}$ and $P_{V_R^j}$ are the volumes (P_V) of the test and the reference spectrum R respectively. The term j represents the association between peaks in both spectra. Moreover, j = 1,.., M where M is the number of peaks of the reference spectrum R. In order to obtain the volume scaling factor (different for each reference-test spectrum comparison) two different solutions have been proposed:

- If the quality control is performed on spectra containing at least five assigned peaks, the routine computes the volume ratio s_j only for them. The volume scaling factor S_J is then obtained as the average value of all the computed ratios (s_j).
- If the investigated spectra do not contain assigned peaks the routine computes the volume scaling factor S_J from a histogram-analysis. As reported in Fig.3.21 a set of ratios are available and a histogram is generated. The number of bins *b* is computed according to the following formula

$$b = \left\lceil \frac{\max(s_j) - \min(s_j)}{3.5\ \sigma\ n^{-1/3}} \right\rceil \qquad (3.15)$$

where σ is the sample standard deviation and *n* is the size of the sample (number of ratios s_j that have been computed). The denominator of eq. 3.15 is known in literature as the Scott's rule [Scott, 2010].

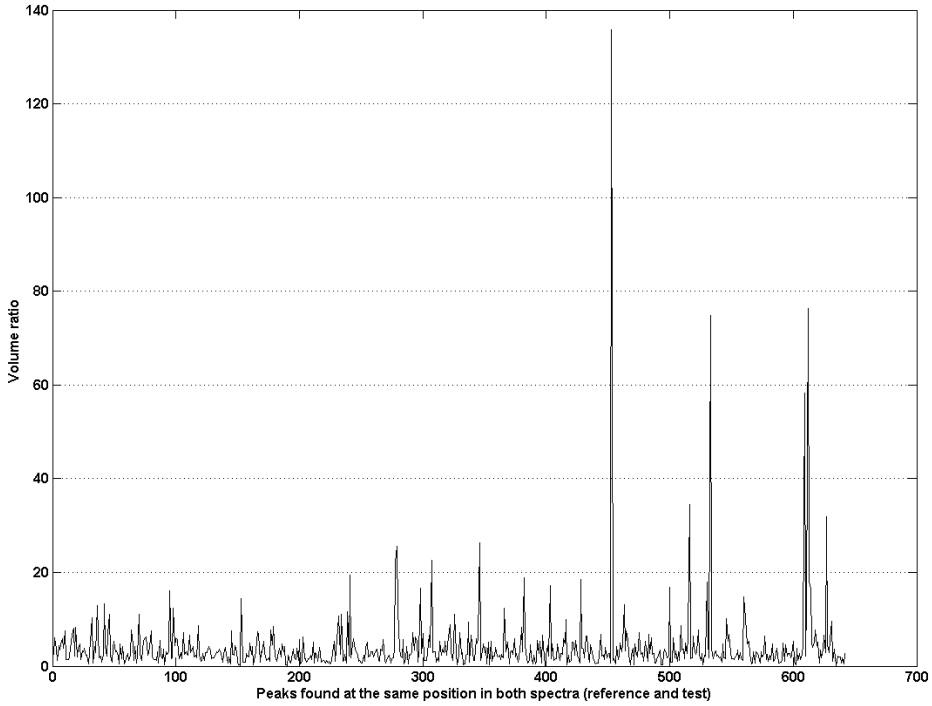


Fig. 3.21 The calculation of the peak volume ratios: peak ratios (black) are computed only for peaks at the same ppm positions (642 peaks). The reported dataset has been collected from ratios between HPr protein from *Staphylococcus aureus* (wild type) and HPr protein from *Staphylococcus aureus* mutant *(H15A)* type.

As shown in Fig. 3.22, the volume scaling factor S_J is obtained as the most frequent ratio in the histogram. The exact value of the volume scaling factor is computed as the mean value of the bin width with the maximum occurrences. This value has to be found in the first third part of the histogram otherwise the routine warns the user that it is not possible to compute this factor correctly. This limitation is used in order to avoid critical situations where the volume scaling factor is computed using solvent peaks.

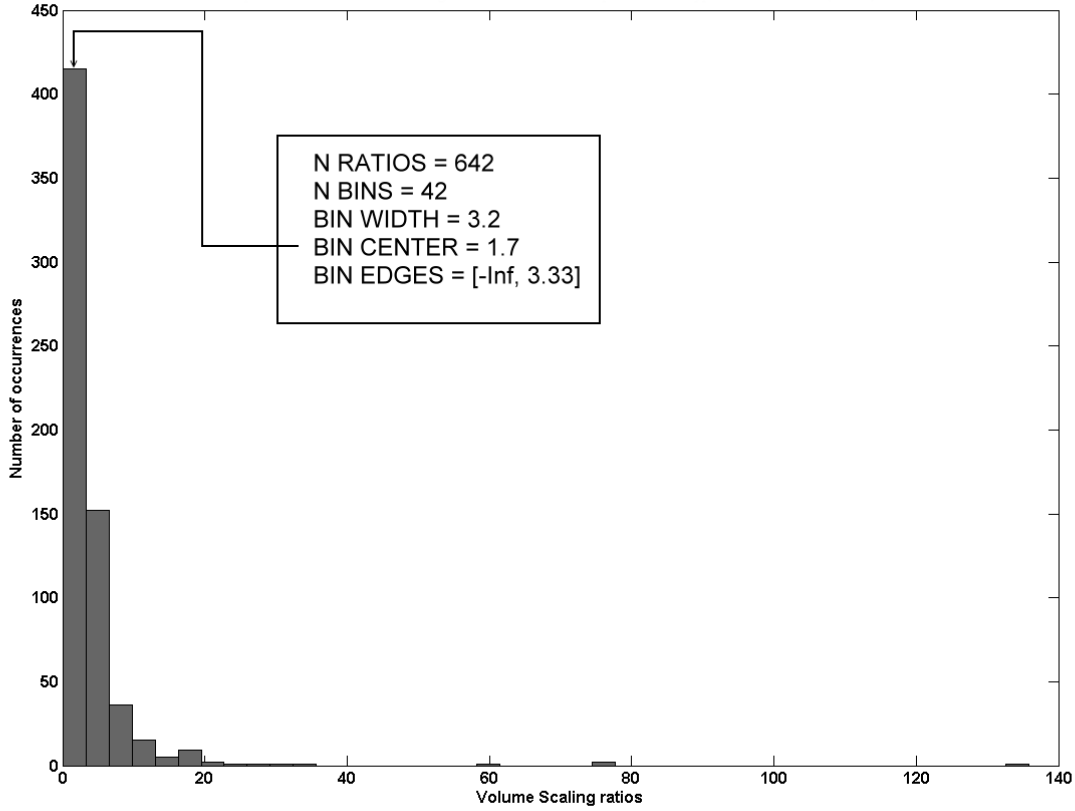


Fig. 3.22 Histogram analysis of volume ratios: bars (black) are computed from volume ratios (642) of peaks at the same ppm positions. The number of bins has been obtained according to the Scott's law. The mean value (i.e. the center of the bin) of the most occurring ratio (415 times) has been used as the volume scaling factor S_J of 1.7. The dataset has been obtained from 642 ratios computed between HPr protein from *Staphylococcus aureus* (wild type) and HPr protein from *Staphylococcus aureus* mutant *(H15A)* type. The value has been automatically accepted by the routine because is located in the first third part of the histogram.

This solution takes the advantage of using only "true" peaks with a maximum deviation of $\pm$ 1 voxel (i.e. the digital resolution deviation).

3.4.2.3.2.2 The *local shift* algorithm: the second peak associations

Once the volume scaling factor has been computed and some peak associations have been stored the routine looks for peak associations not recognized in the range of the maximum deviation of $\pm$ 1 voxel. If the sample and the measurement conditions are identical, the spectra are identical and the local shift (LSHIFT) algorithm (considering the effect of wrong referenced spectra) has already associated all the peaks between the reference and the test spectra. If the sample has been measured under different conditions the second local shift algorithm finds these association in the maximum range defined be the user.

During this project some questions have been raised. For example, what would happen if the considered region is too crowded and many peaks have quite the same shape, like in a NOESY experiment? Is the cosine similarity, that is size independent, the best usable discriminant to select a relationship one to one among many similar peaks in different spectra? Supposing that the volume is another good feature of comparison, in which experiment type has to be used mainly? Is it useful, when possible, to consider residues patterns? These and many other questions were faced by this module.

During this work many attempts have been analyzed in order to obtain the optimal solution to this search problem.

The first idea was trying to find the correct peaks associations using the cosine similarity criterion (eq. 3.3). Each reference peak has been compared with the test peaks that are inside of the user defined local shift range. The algorithm has shown a good performance (finding exact peak associations) on isolated regions of the spectrum and a low reliability on spectral crowded zones. In addition, this method has not shown a good performance in cases where the volume changes (e.g. due to a pressure change) are too large.

An improvement of the previous proposed algorithm is to include the information obtained from the neighborhood distance. For each considered peak, the list of neighbors (NLST) between peaks belonging to the same cluster has been used to analyze the surrounding neighborhood within the limits imposed by the user through the maximum allowed local shift. Including the NLST involves not only a one-to-many (one reference peak compared with many test peaks) signal comparison, but a many-to-many (reference

peak cluster compared with the test peak cluster) approach. It means that both the reference and the test surrounding of the peak of interest are compared. Starting from the center of the considered reference peak a "pseudo" energy term E has been computed as following:

$$E = \sum_{n=1}^{N} \left(\left(\sum_{i=1}^{dim} \left| d_{cn}^{i} \right| \right)^{P_{In}} / P_{Ic} \right) \quad (3.16)$$

where *dim* is the considered direction and *n* = 1,.., N, with N representing the number of peaks of the considered spectrum. The term P_{In} represents the intensity of the peak's neighbor, P_{Ic} is the intensity of the analyzed peak (center of the cluster) and d_{cn}^{i} defines the distance (radius of the peak cluster) between them. In order to select the best connection between a reference and a test peak a measure of goodness (based on the neighborhood of each peak) has been proposed. The following criterion has been used:

$$E_{RT}^{nei} = 1 - \left| (E_R - E_T)/(E_R + E_T) \right| \quad (3.17)$$

where E_R and E_T are the energy terms of the reference and the test spectra respectively. The association between peaks is accepted if the cosine similarity of the considered couple of peaks is > 0.6 and the E_{RT}^{nei} > 0.5. These values have been obtained from an empirical evaluation of some test cases. This method has shown a good reliability in case of not assigned peaks and in crowded regions of the spectra. Unwanted associations have been shown because the algorithm tends to create couples of peaks (between the reference and the test spectrum) assuming as a good candidate the first nearest peak (in the ppm unit) with higher cosine similarity, as shown in Fig. 3.23. A backtracking algorithm is required in order to compare all the possible connections.

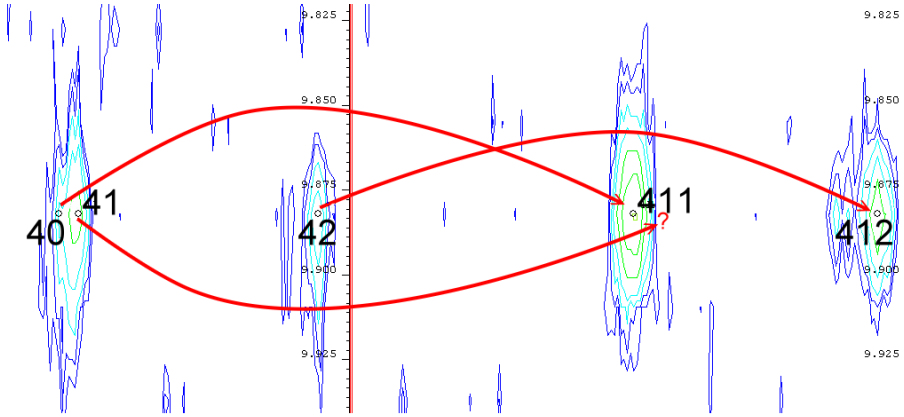


Fig. 3.23 Incorrect peak association: the peak association is not efficient when not performing a possible backtracking step: it associates peak 40 of HPr protein from *Staphylococcus aureus* (wild type) with peak 411 of HPr protein from *Staphylococcus aureus* mutant *(H15A).* The peak 41 is wrongly marked as missing.

As reported in Fig. 3.23 the routine associates the peak number 40 (in the reference spectrum) with the peak number 411 (in the test spectrum). The peak number 41 cannot be evaluated for other possible further associations, thus it wrongly appears to be not found (missing). The main important reason of this failure was the not considered backtracking factor, sometimes necessary, thus successively introduced in the algorithm. Using a back tracking algorithm it is possible to replace the association between peaks 40→411 with the correct one 41→411.

A further improvement of the *LSHIFT* routine is depicted in Fig. 3.20, where the backtracking step and the two described local shift modules are visible (one for the digital resolution shift and the other user defined one that uses the volume scale information previously computed). In order to describe the backtracking step it is necessary to define the optimal measure of goodness between peak associations. For each possible peak association the algorithm computes a score H_{ass} (see eq. 3.18) made up of three distinct parts:

1. The comparison based on the cosine criterion between peaks (see eq. 3.3)
2. The measure of goodness E_{RT}^{nei} between peak neighborhood (see eq. 3.17)
3. The peak pattern matching (see eq. 3.21).

The optimal goodness score is computed as following:

$$H_{ass} = \max\left(r_{RT}^{cos} \, E_{RT}^{nei} \, E_{RT}^{pat} \right) \tag{3.18}$$

where

$$r_{RT}^{cos} = \frac{P_{D_R} \cdot P_{D_T}}{\|P_{D_R}\| \|P_{D_T}\|}, \tag{3.19}$$

$$E_{RT}^{nei} = 1 - \left| (E_R - E_T)/(E_R + E_T) \right| \tag{3.20}$$

and

$$E_{RT}^{pat} = f\left(R_{pat}, T_{pat} S_{RT}\right) = \frac{1}{N_{pat_R}} \left(1 - \frac{|SH|}{SH_{max}}\right) (1 - Z) \tag{3.21}$$

with

$$\mathrm{Z} = \left| \frac{P_{V_R}^h - \left(P_{V_T}^k S_{RT}\right)}{P_{V_R}^h + \left(P_{V_T}^k S_{RT}\right)} \right| \tag{3.22}$$

The terms P_{D_R} and P_{D_T} (see eq. 3.19) represent the peak voxels (raw data) reordered as multidimensional arrays in case of reference *R* and test spectrum *T* respectively. The terms R_{pat} and T_{pat} (see eq. 3.21) are multidimensional arrays containing all the peak volumes (P_V) that are encountered passing through all the directions spanned from those peaks while S_{RT} is the previously computed volume scale between the reference and the test spectrum. The spanning width of the pattern allows a $\pm$ 2 voxel shift with an additional unitary shift due to the deviation of the digital resolution. The term SH represents the measured difference in spectrum voxels between the reference and the test peak found in both peak patterns whereas the term SH_{max} (default value of 3) represents the maximum allowed difference between peaks belonging to the compared patterns. The terms P_{V_R} and P_{V_T} (see eq. 3.22) represent the volume of the reference and the test peak respectively, with

h = 1,..., N_{pat_R} and k = 1,..., N_{pat_T}, where N_{pat_R} and N_{pat_T} are the number of encountered peaks (peaks with volume P_V) along the spanned pattern of the reference and the test respectively. In order to generate each peak pattern one coordinate of the peak positions is varied while the others are maintained constant. This procedure is repeated changing the direction of the coordinate one by one. For example, in a 2D experiment the routine varies the δ_1 direction (collecting all the peaks that are encountered along the variation $\pm$ 2 voxel) leaving the δ_2 direction steady. Successively, the routine varies the δ_2 direction leaving the δ_1 fixed.

In case that the user defines a large local shift (larger than the maximum distance computed through the neighborhood distance list NLST see par. 3.4.2.3.1) only the best half of matching cases contained in the NLST are evaluated according to eq. 3.18.

A schematic representation of the pattern alignment evaluated in the proposed algorithm is described in Fig. 3.24.

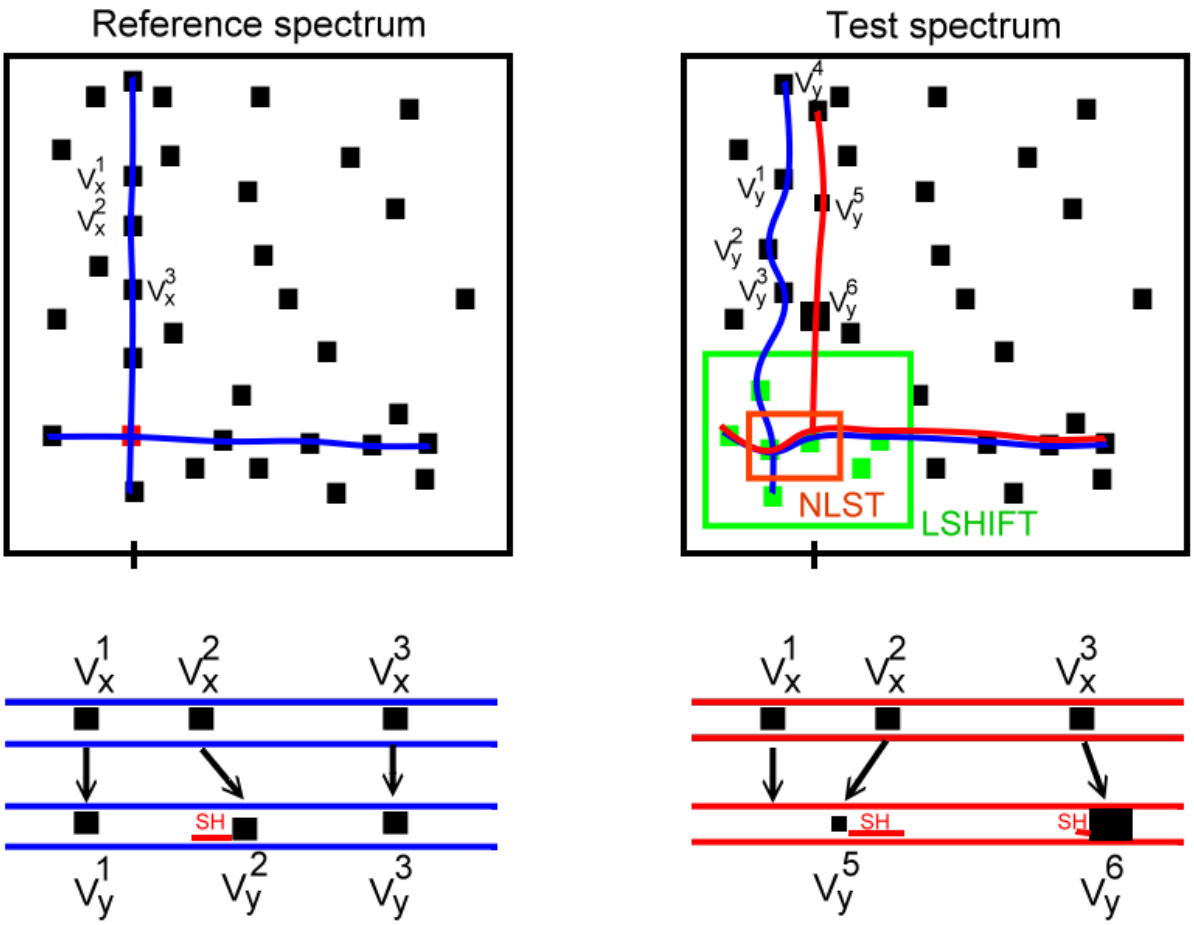


Fig. 3.24 The pattern alignment evaluation in the local shift algorithm: the chemical shift positions, the normalized volumes and amount of peaks along the blue pattern (spanned starting by the red signal in the reference spectrum) are compared with all the possible patterns originating from every best neighbor peaks (orange) in neighborhood distance list (NLST) of the test spectrum. The pattern alignment shift (SH) between the red and the blue pattern has a lower score due to the different number of encountered peaks along the pattern, to the different volume and to the larger pattern shift.

As shown in Fig. 3.24, the simple connection of one signal in the reference spectrum with a peak at the same ppm coordinates in the test spectrum, even with a strong cosine similarity, would lead to a wrong association. The algorithm evaluates the whole pattern considering the peak surrounding, the cosine criterion, the number of peaks in the pattern, their shifts and their normalized volumes.

Incrementing the information used to select the correct peaks association and avoiding the discriminant application of previously defined thresholds it is possible to perform an optimal peak association. In both cases a backtracking step is needed in order to avoid temporarily wrong connections that have been corrected maximizing the heuristic H for all the investigated test peaks.

The described procedure is repeated for every reference peak (around the computed cluster radius) creating a connection list because multiple connections can happen. For example, as reported in Fig. 3.25 more than one reference peak (A and C) can be associated with the same test signal (B). For this reason a backtracking algorithm based on the eq. 3.18 has been applied selecting only that connection H* (e.g. H_{AB}, where $H_{AB} > H_{CB}$) with the maximum value. In this case some connections are lost (C with D) requiring, as shown in Fig. 3.20, a recalling of the LSEARCH2 module allowing a connection of the signal (C) with the second best option (D) in accordance to eq. 3.18. If this peak had only one possible connection, obviously it could not be associated to any other one becoming a missing peak. This analysis represents the most important step for a correct peak-by-peak classification performed during the mid-level evaluation.

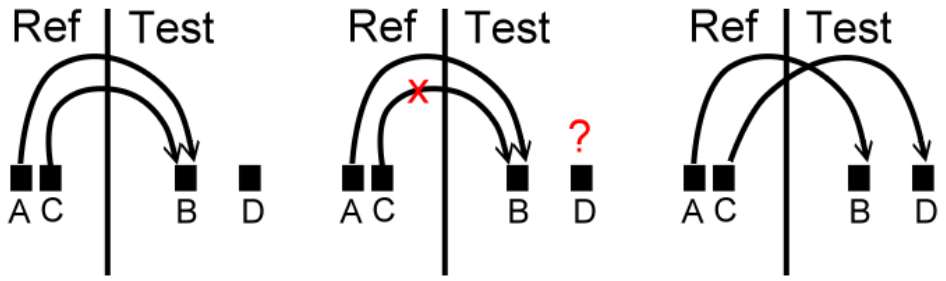


Fig. 3.25 The backtracking algorithm of the LSHIFT module: in case of multiple associations (peak A and C with peak B) the routine retains the connection with the highest score (peak A with peak B) leaving the peak C orphan. The backtracking assumes that the connection A→C is true and it tests again only the reference peak C. The second best possible connection of the reference peak C is then with the peak D.

Once the peaks have been optimally associated, it is possible to compute a matching score based on the peak shift variations. This score is created as a function of the user allowed local shift as following:

$$r_{RT}^{sh} = \prod_{i=0}^{DIM-1}\left(1 - \frac{|D_i|}{LSHIFT_i}\right) \quad (3.23)$$

where r_{RT}^{sh} is the matching score of the chemical shift variation between the reference *R* and the test spectrum *T*. The term $|D_i|$ represents the chemical shift difference along the *i*-th direction whereas $LSHIFT_i$ represents the normalized local shift variation along the direction *i* (in the ppm chemical shift unit).

Once peaks have been associated (necessary in case of unassigned spectra), their multiplet structure has been identified and their volumes are normalized, is then possible to evaluate the line width of each associated peak in a correct manner.

3.4.2.4 Line width comparison

The line width of a peak is a strong information that can be used in order to extract further peak features. During the QTA procedure, the shapes, positions, volumes and line width variations are considered in order to classify mutual peak associations and to compute peak probabilities.

Magnetic field inhomogeneties induce peak shape alterations. This should be taken into account together with noise and artifact distortions. The calculation of line widths on experimental spectra requires an optimization method (nonlinear least-squares) due to Lorentzian peak shapes.

The developed AUREMOL-LW routine allows a completely automated line width calculation of every type of peak by means of two fitting functions: the Lorentzian (default function) and the Gaussian one in any given experimental dimension (see Appendix A). In order to calculate the proper line width of peaks in a multidimensional space, each peak (singlet or multiplet) must be sliced through to the center along all experimental directions. Each slice, starting from a guessed set of parameters (amplitude, mean and standard deviation) is fitted in order to obtain the correct intensity, the correct center position and the full width at half maximum (FWHM). The routine optimizes (simultaneously) as many fitting functions as the number of maxima/minima found in each single slice. An example is shown in Fig. 3.26.

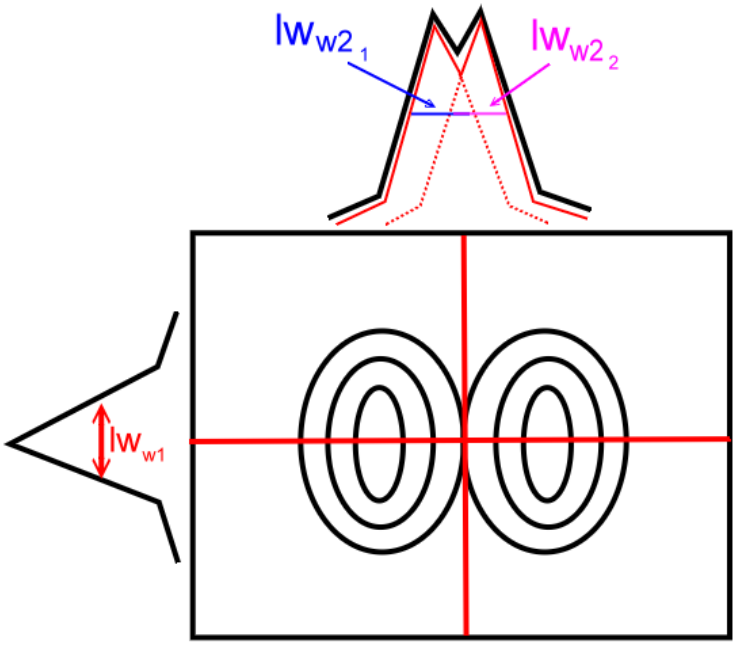


Fig. 3.26 Calculation of the line width of a peak doublet: the considered peak is sliced along all the measured directions. The resulting slices are simultaneously fitted with as many Lorentzian functions as the number of maxima/minima found in the analyzed slice. The method computes the fitting function/functions by means of the non-linear Levenberg-Marquardt optimization algorithm. The outputs of the algorithm are the fitted intensities, the correct positions of the peak centers and the line widths. In the proposed example the final line width along the direction δ_2 is the sum of $lw_{\delta 2_1}$ and $lw_{\delta 2_2}$.

As reported in Fig. 3.26, the AUREMOL-LW routine computes the peak line width along each measured direction for all the peak maxima/minima. In case of a peak showing a multiplet structure the final line width has been calculated as the sum of the computed line widths (one for each single multiplet component) along each measured direction.

Obviously, when an experimental multiplet is not properly recognized, not a unique central peak is picked, leading to a wrong peak volume. The MASK described in par. 3.4.2.2.1 avoids that problem. In addition, the routine calculates the optimal multidimensional box around each peak along all the directions using the previously computed line width. The box has been obtained in a mathematical manner computing the full width until it reaches one percent of the maximum height of the considered peak.

This procedure has been used to calculate volumes of multiplet peaks previously recognized via the multiplets search tool. In this case the slicing is performed through a dynamic sliding window along all the rows of the whole box previously computed. Each slice has been fitted and the output has been superimposed and summed obtaining the final volume. Noteworthy, the outside part of the "superimposed peak" is theoretically noise free without overlaps as shown in Fig. 3.27.

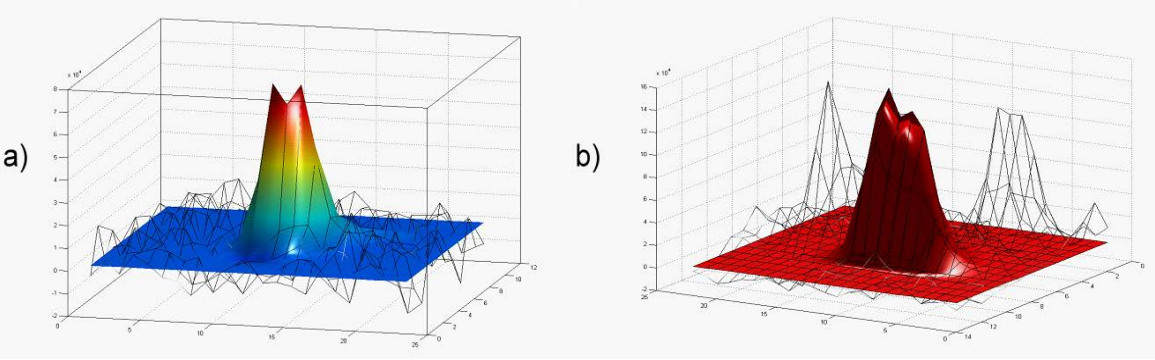


Fig. 3.27 Segmentation of peak multiplets: The algorithm used to compute the line widths of each peak is recursively applied (row-wise in a 2D case) over the whole peak producing a noise free segmentation of an isolated doublet (a) and of a doublet in a crowded region coming through overlap problems (b).

Once the line width parameters of the peaks belonging to the reference and to the test spectra have been computed, the routine automatically performs a comparison between them. The comparison furnishes a matching score between the investigated peaks based on the line width feature (in the Hz unit) as following

$$r_{RT}^{lw} = 1 - \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{dim}\left(lw_{\delta_i}(R) - lw_{\delta_i}(T)\right)^2}}{lw_{max} - lw_{min}} \tag{3.24}$$

where r_{RT}^{lw} is the normalized root mean square deviation, $lw_{\delta_i}(R)$ and $lw_{\delta_i}(T)$ are the computed line widths (in the Hertz unit) along the direction δ_i (i represents the *i-th* direction) using the reference *R* and the test spectrum *T* respectively. The terms lw_{max} and lw_{min} represent the maximum and the minimum line width measures of the investigated peaks.

3.4.2.5 The hybrid time-frequency domain analysis

All the previously described methods and computations are developed in the frequency domain after the Fourier transformation [Briggs & Henson, 1995; Kauppinen & Partanen, 2002; Zonst, 2003]. This transformation has the ability to sort all the frequencies in a mathematical manner "building" the final NMR spectrum. For a human being, the Fourier transformation is the best way to split information that is mixed up in the time domain earning benefits from its linear properties. During this project, efforts have been done to extract useful information from the human hard interpretable time domain data becoming a hybrid analysis tool.

In particular, this calculation is not perceptible (indirect way) by the user but it is automatically applied in order to extract additional features that are helpful during the comparison.

Once all the acquisition and processing parameters of the recorded spectrum are known (e.g. dimensions, acquired data architecture, group delay, filter types, FCOR and phase corrections), it is possible to apply an inverse Fourier transformation of the multidimensional box (surrounding each peak in the spectra) in the frequency domain as shown in Fig. 3.28.

Some problems occurred while trying to match signals in the time domain. From the Parseval's theorem it is well known that the integral of the square of a function is equal to the integral of the square of a Fourier transformed function. As a consequence, the application of the similarity criterion (see eq. 3.3 with $P_X = P_D$) in the time domain does not bring out additional information with respect to the similarity criterion applied in the frequency domain.

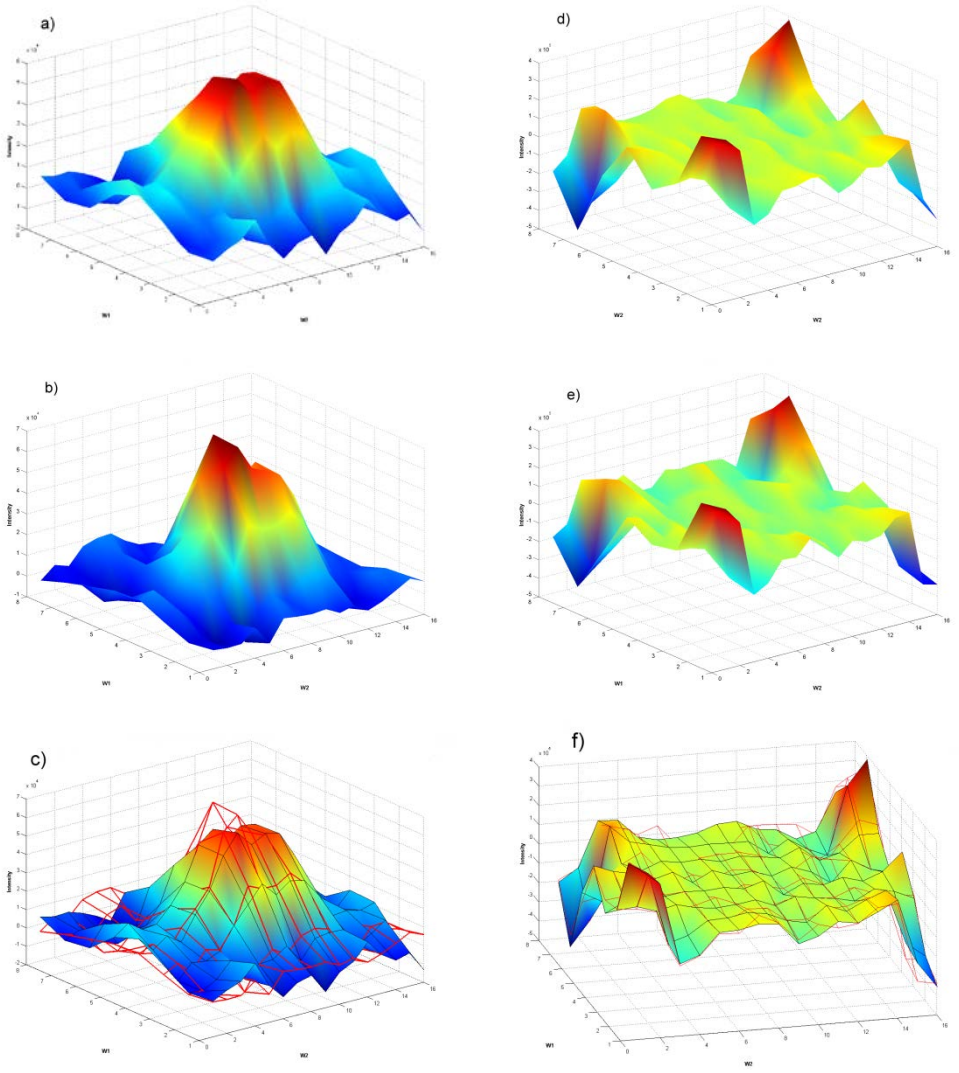


Fig. 3.28 Comparison of frequency and locally back-transformed time domain peak: Reference (a), test (b) and overlapped (c) peaks of HPr protein from *Staphylococcus aureus* in the frequency domain. Inverse Fourier transform of the reference (d), the test (e) and the overlapped (f) peaks in the time domain.

The Pearson product-moment coefficient (cross-correlation coefficient) r_{RT}^{corr} is the proposed alternative method. It calculates the strength of the linear associations between two datasets (in the range $\pm$ 1), as following:

$$r_{RT}^{corr} = \frac{\sum_{D=1}^{N}(P_{D_R} - \overline{P_{D_R}})(P_{D_T} - \overline{P_{D_T}})}{\sqrt{\sum_{D=1}^{N}(P_{D_R} - \overline{P_{D_R}})^2}\sqrt{\sum_{D=1}^{N}(P_{D_T} - \overline{P_{D_T}})^2}} \quad (3.25)$$

where P_{D_R} and P_{D_T} are the voxels (P_D) belonging to the reference and the test peaks respectively, with D = 1,..,N and N is the number of bytes forming the considered peaks.
Considering the eq. 3.3 and eq. 3.25 it is straightforward that these two indices differ by the mean subtraction from the latter. The mean values of time and frequency domain signals are obviously different. The FFT routine has been used in order to obtain a local inverse Fourier transformation of each peak in the spectra. In particular, the size of each peak has been adapted to the next integer number in the power of two. This has been done in order to manage the data properly according to the Fourier transform (Radix-2 Cooley-Tukey). For example, if the size of a peak box is 5x12 voxels, the routine automatically resizes the box in 8x16 voxels.
In order to compare the sensitivity of both proposed methods (cosine similarity and cross-correlation in time and frequency domain) three different conditions (varying the peak box size) have been analyzed:

1. Presence of weak noise distortions
2. Presence of strong noise distortions
3. Presence of strong baseline distortions

The first comparison has been conducted analyzing the peak HD22 38/HD22 38 of the back-calculated NOESY spectrum of the HPr protein from *Staphylococcus aureus* (wild type). Gaussian noise has been added to the previously described simulated spectrum yielding four different datasets (see par. 2.2.1.1). These spectra differ each other just with respect to the SNR:

1. SNR_{dB} of 38.258 decibel
2. SNR_{dB} of 32.2056 decibel
3. SNR_{dB} of -8.7206 decibel
4. SNR_{dB} of -14.7679 decibel

The reported signal-to-noise ratios have been computed considering a proton-proton pair in a distance of 0.5 nm. The diagonal peak HD22 38/HD22 38 (without noise) has been compared with itself in the first and in the second dataset. The result of the comparison is shown in Fig. 3.29.

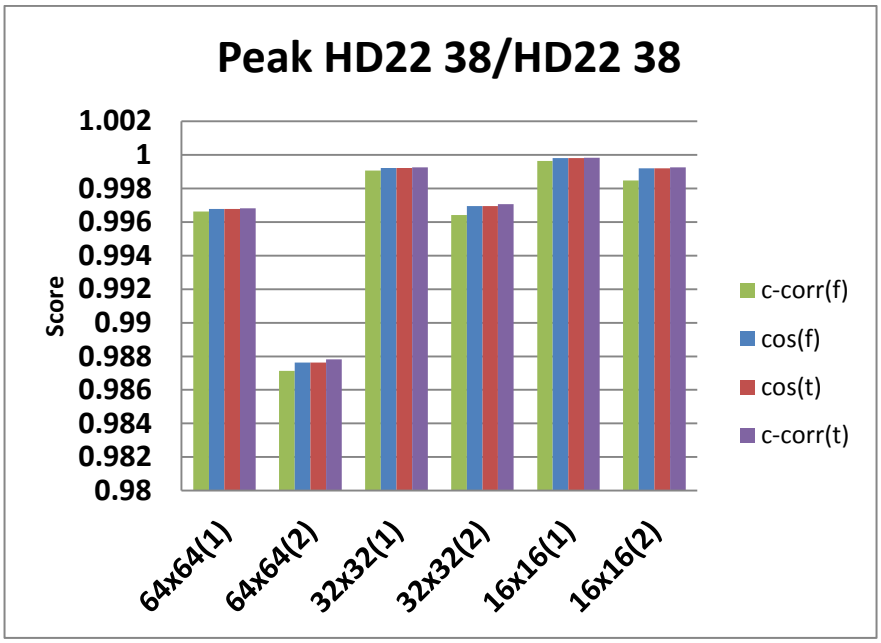


Fig. 3.29 Comparison between the cosine similarity and the cross-correlation in the time and in the frequency domain using the first two datasets (1, 2). The comparison has been performed varying the size of the analyzed peak box: the first dataset (1) has a SNR_{dB} of 38.258 decibel whereas the second (2) of 32.2056 decibel. The cosine similarity behavior (red and blue traces superimposed by the violet one) is unchanged between the time and the frequency domain. The cross-correlation computed in the time domain (violet trace) is more sensitive with respect to the one calculated in the frequency domain (green trace).

As reported in Fig. 3.29, the comparison has been performed varying the size of the analyzed peak box starting from 64x64 to 16x16 voxels. The cosine similarity behavior (red and blue traces) is unchanged in the time and in the frequency domain. The cross-correlation computed in the time domain (violet trace lying over the red and blue traces) is marginally more sensitive with respect to the one calculated in the frequency domain (green trace).

A second test been conducted using the peak HD1 74/HN 64 of the HPr protein from *Staphylococcus aureus* (wild type). This simulated peak (noiseless) has been compared with itself in the third and in the fourth datasets. The result is shown in Fig. 3.30.

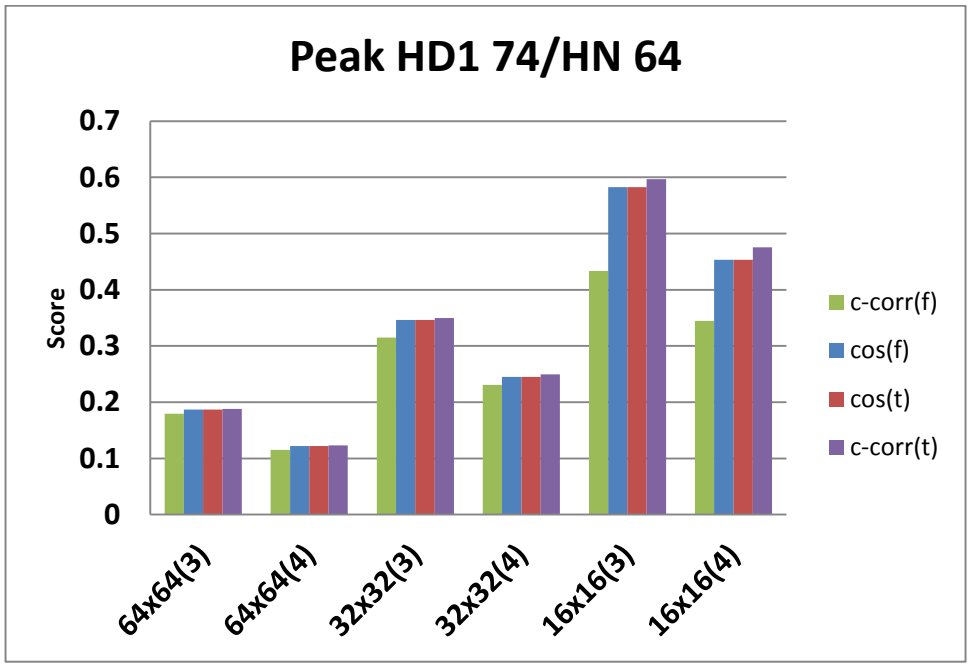


Fig. 3.30 Comparison between the cosine similarity and the cross-correlation in the time and in the frequency domain using the last two datasets (3, 4). The comparison has been performed varying the size of the analyzed peak box: the first dataset (3) has a SNR_{dB} of -8.7206 decibel whereas the second (4) of -14.7679 decibel. Increasing the noise level the cosine similarity behavior (red and blue traces) is unchanged between the time and the frequency domain. The cross-correlation computed in the time domain (violet trace) is more sensitive (more than 12%) with respect to the one calculated in the frequency domain (green trace).

The example reported in Fig. 3.30 shows that the cross-correlation computed in the time domain is less sensitive to the noise with respect to the same criterion computed in the frequency domain. In addition, this criterion is more suitable (more than 12%) showing a higher score. The cross-correlation computed in the frequency domain demonstrated to be the less reliable criterion than all others. This is particularly evident decreasing the peak box size and the SNR.

A third test has been conducted analyzing the peak HD22 4/HA3 of the HPr protein from *Staphylococcus aureus* (wild type) as reported in Fig. 3.31. This peak (noiseless) has been compared with itself in the fourth dataset. This peak has been intentionally selected in the water region showing strong baseline distortions. The peak size is 16x16 (in the voxel unit) along the δ_2 and the δ_1 direction (2D-case). The result of the comparison is shown in Fig. 3.32.

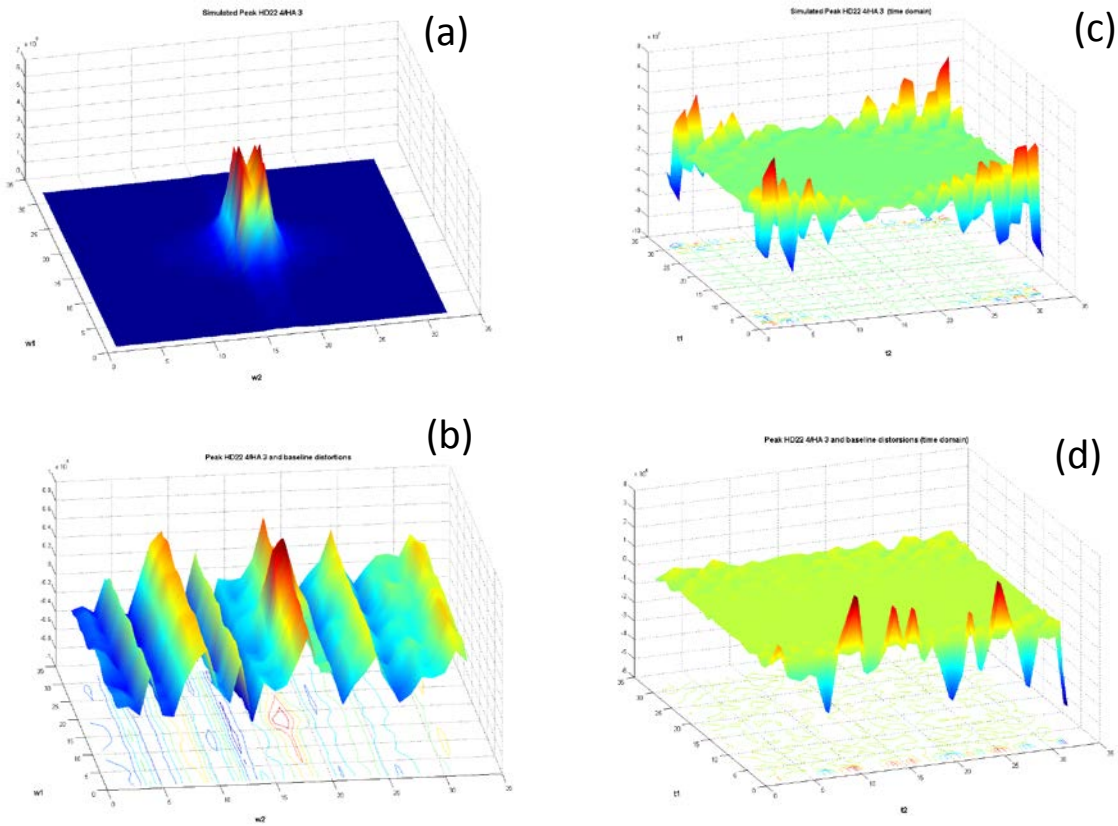


Fig. 3.31 The simulated and the baseline distorted peak HD22 4/HA3 of the NOESY spectrum of the HPr protein from *Staphylococcus aureus* (wild type) in the time and in the frequency domain: the simulated peak (a) and the very same peak distorted by strong baselines (b). The simulated (c) and the baseline distorted (d) peak in the time domain. The peak size is 16x16 (in the voxel unit) along the δ_2 and δ_1 direction.

The performed comparison of the peak HD22 4/HA 3 of the HPr protein from *Staphylococcus aureus* (wild type) shows that in presence of strong baseline distortions the cross-correlation computed in the time domain recognizes the same peak more accurately than all the other criteria. The same criterion computed in the frequency domain is instead more affected by the presence of noise and baseline distortions.

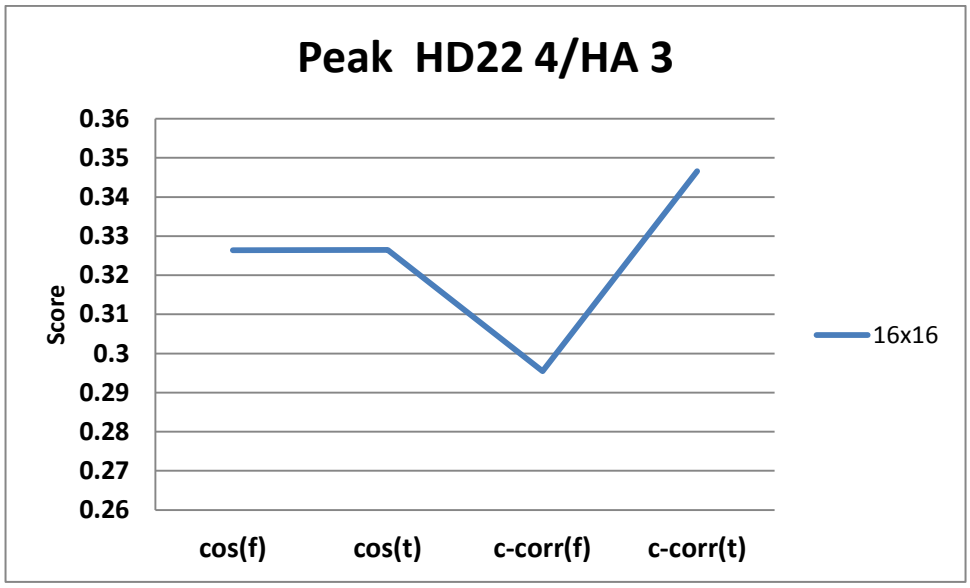


Fig. 3.32 Comparison between the cosine similarity and the cross-correlation in the time and in the frequency domain. The comparison has been performed analyzing the peak HD22 4/HA 3 (size of 16x16 voxels) from the NOESY spectrum of the HPr protein from *Staphylococcus aureus* (wild type): in presence of strong baseline distortions the cross-correlation criterion computed in the time domain recognizes the peak with a higher score than all other criteria. The cross-correlation criterion computed in the frequency domain is instead more affected by the presence of noise and baseline distortions.

The analysis performed using the three different test cases shows that the cross-correlation criterion computed in the time domain is more reliable than all others, thus it has been chosen as an additional feature for spectral comparison. Back-transforming the data involves considering two main aspects:

1. The data has to be back-transformed according to the dimensionality of the measured spectrum (in a 3D experiment the data has to be back-transformed along the δ_3, the δ_2 and the δ_1 direction)
2. The size of the peak box has to be adapted dynamically with respect to the original size to the closest number in the power of two.

3.4.3 Mid-level analysis: from a score-like to a probability-like system

In the previously described low-level procedure a lot of information has been collected. For example, the peak associations and the peak features have been detected and stored. Missing peaks have been identified. Volume information differences have been collected and ppm position shifts of the center as well. The line width variation and the degree of cross-correlation of time domain data are used as additional information together with all the other analyzed parameters. In order to reveal data structures and to draw a general conclusion, an interpretation of the computed results (stored as scores) is needed. The basic idea is to analyze all values from a score-like system (low-level) to a more interesting probability-like one (mid-level).

The main difference relies on the fact that a score is just a number telling the user how good the matching is between the considered peaks commonly in a normalized range of ± 1 (i.e. plus one in case of a perfect match and minus one in the contrary case). The probability is instead the expression of a knowledge that an event could happen. This understands is obtained through the statistical inference.

3.4.3.1 Peak probabilities

Using the Bayesian [Cornfield, 1967; Cornfield, 1969; Schulte et al., 1997, Antz et al., 1995] model the *prior* (typically a value of 0.5) distribution is subjective offering to the analyzer the possibility to set options on relative weights. Since the uncertainty of our collected data is always present it is trustworthy to consider them as random variables, controlling unidentified external factors and giving a measure of the uncertainty. At the

end, with the Bayes' theorem the posterior distribution is computed through the prior and the observed one.

During the mid-level stage of this project, random distributions have been computed in order to classify observed peaks distributions through the prior $P(C_k)$ probability. According to Bayes' theorem, the probability that the cross peak *i*, known its local property F_j^i (with *j* = volume, shape, position, line width and cross-correlation), belongs to the class C_k (where *k* = same or different protein) is:

$$P\left(C_k \middle| F_j^i\right) = \frac{P(C_k)P\left(F_j^i \mid C_k\right)}{\sum_{k=1}^{K} P(C_k)P\left(F_j^i \mid C_k\right)} \tag{3.26}$$

The routine computes all the Bayesian probabilities using the previously collected scores. For example, it is possible to obtain for each peak a probability as function of the preferred features as the position, the volume, the shape, the line width and the linear correlation (cross-correlation). To obtain these probabilities, it is necessary to combine (through the Bayes' theorem) the observed values, representing the probability that the selected peak belongs to the same protein in the other spectrum, with the random ones. They represent the probability that the peak belongs to another protein, in the other spectrum. In order to obtain these latter values, couples of random positions (according to the experimental parameters) have been generated. Each couple identifies a new peak in the test spectrum (no restriction on the location of the new peak in the spectrum is applied) that is compared with the investigated reference spectrum peak. If a set of reference spectra is available the distribution of "true" peaks is obtained from these spectra.
Once these probabilities have been computed, two main applications have been analyzed:

- Determine if it is the same peak with respect to the other spectra
- Determine if the considered peak has the same specific feature with respect to the same peak in the other spectra

The QTA routine allows the detection of direct probabilities with the possibility to separate them through a threshold system. One possibility is to classify peaks in three classes:

1) good matching peaks (high probability) with a probability p ranging in the interval $0.66 < p \leq 1.0$.
2) rather good matching peaks (mid probability) with a probability p ranging in the interval $0.33 < p \leq 0.66$
3) not good matching peaks (low probability) with a probability p ranging in the interval $0 < p \leq 0.33$

Obviously, the more peaks belong to the first class the higher the probability is that the compared spectra are referring to the same protein.

3.4.3.2 Assessment of individual feature of peaks

Frequency distributions have been collected for the cosine similarity feature, the cross-correlation feature, the line width the volume and the position features. They are computed for every spectra's couple (e.g. reference-test, simulation-reference, simulation-test). The score of each peak comparison as in function of the investigated feature.

For example, the cosine similarity feature has been computed comparing the NOESY spectrum of the HPr protein from *Staphylococcus aureus* wild type with the mutant (H15A) one. According to the eq. 3.3 two sets (the measured and the random one) of cosine similarity scores have been measured from which two distributions have been generated as shown in Fig. 3.33.

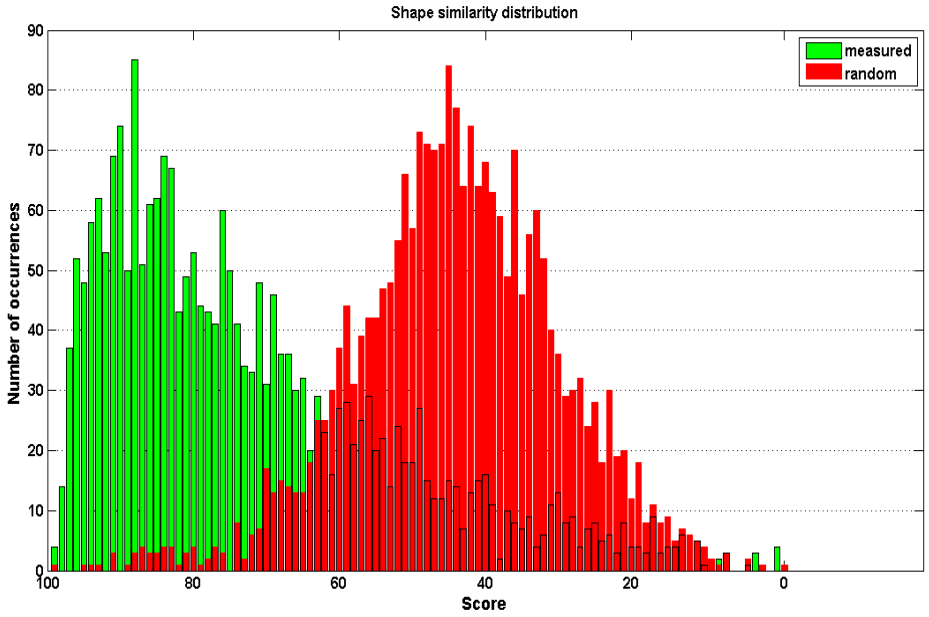


Fig. 3.33 The measured (green) and the random (red) distributions based on the cosine similarity feature: the distributions are computed comparing the NOESY spectrum of the HPr protein from *Staphylococcus aureus* wild type with the mutant (H15A) one: the observed values (green bars) and a random generated ones (red bars) are shown based on a score scale ranging from 0 to 100% (perfect match).

The same procedure has been used to compare the cross-correlation feature. In particular, each reference (NOESY spectrum of the HPr protein from *Staphylococcus aureus* wild type) and test (NOESY spectrum of the HPr protein from *Staphylococcus aureus* mutant H15A) peak has been back-transformed (i.e. real IFFT) to the time domain generating two datasets of scores according to the eq. 3.25. The resulting distributions are shown in Fig. 3.34

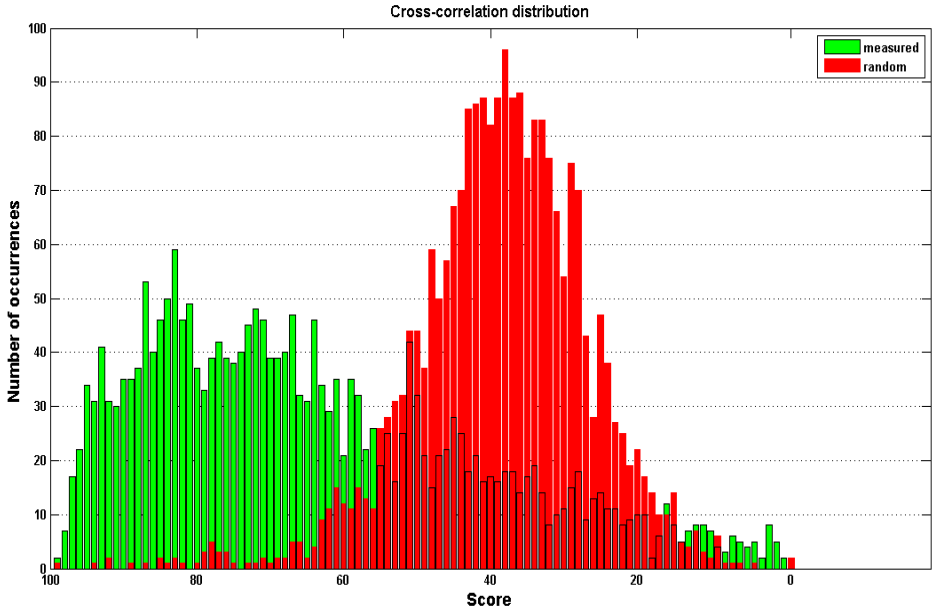


Fig. 3.34 The measured (green) and the random (red) distributions based on the time domain cross-correlation feature: the distributions are computed comparing the NOESY spectrum of the HPr protein from *Staphylococcus aureus* wild type with the mutant (H15A) one: the observed values (green bars) and a random generated ones (red bars) are shown based on a score scale ranging from 0 to 100% (perfect match).

The line width feature has been analyzed according to the eq. 3.24 and the resulting distributions are shown in Fig. 3.35.

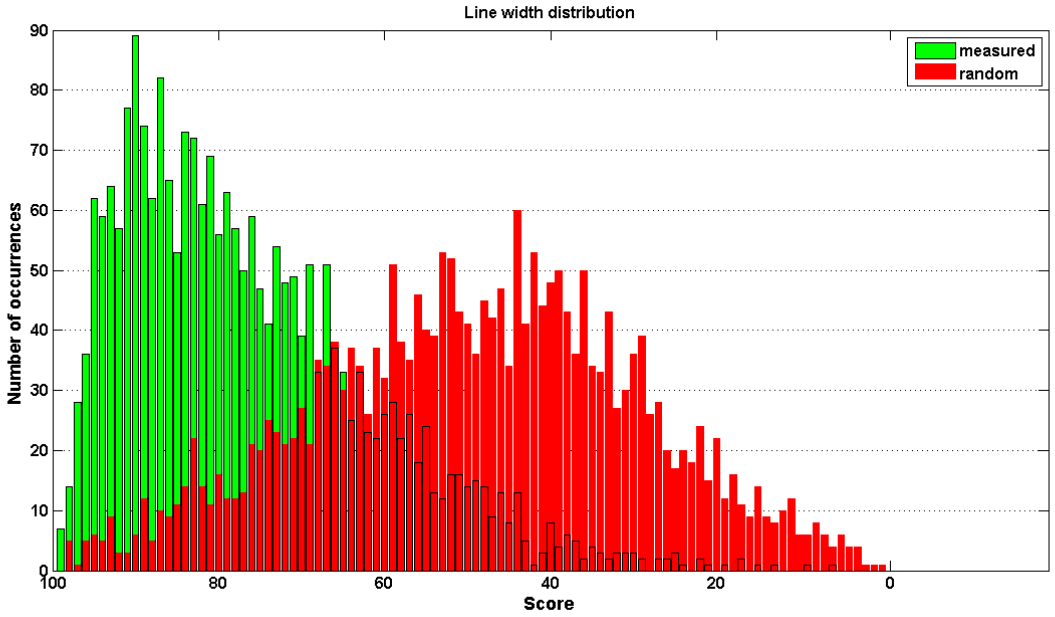


Fig. 3.35 The measured (green) and the random (red) distributions based on the line width feature: the distributions are computed comparing the NOESY spectrum of the HPr protein from *Staphylococcus aureus* wild type with the mutant (H15A) one: the observed values (green bars) and a random generated ones (red bars) are shown based on a score scale ranging from 0 to 100% (perfect match).

The peak position variation feature has been considered according to the eq. 3.23. The distributions obtained comparing the NOESY spectrum of the HPr protein from *Staphylococcus aureus* wild type and the NOESY spectrum of the HPr protein from *Staphylococcus aureus* mutant (H15A) are shown in Fig. 3.36.

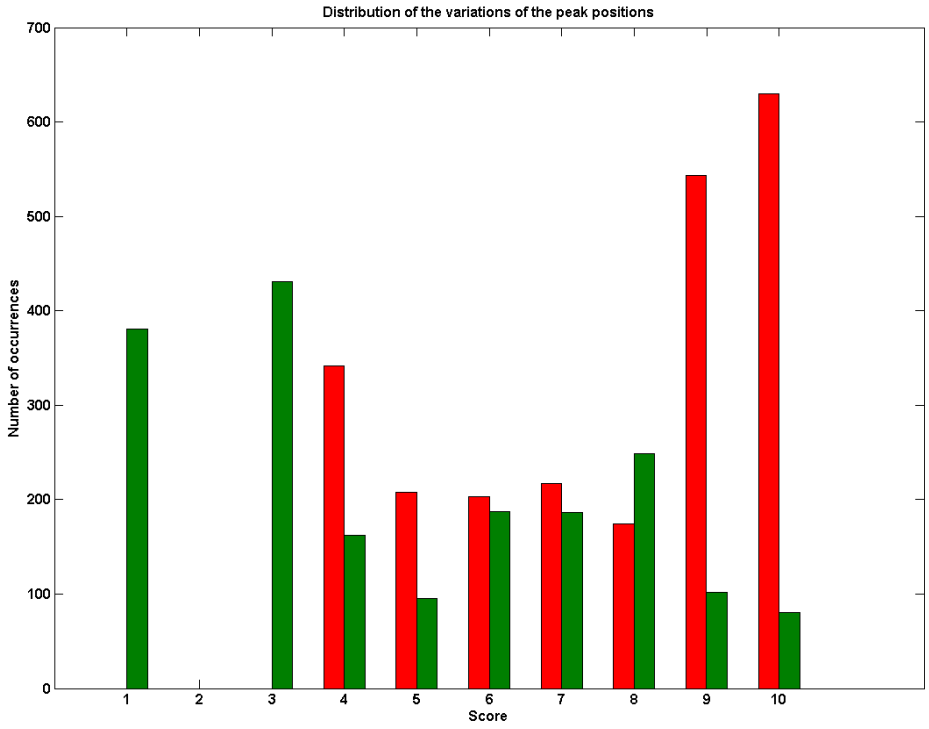


Fig. 3.36 The measured (green) and the random (red) distributions based on the peak feature of shift variations: the distributions are computed comparing the NOESY spectrum of the HPr protein from *Staphylococcus aureus* wild type with the mutant (H15A) one: the observed values (green bars) and a random generated ones (red bars) are shown based on a score scale ranging from 0 to 100% (perfect match).

In order to compute the peak probability based on the volume feature additional assumptions must be considered. For example, the level of noise present in the investigated spectra varies in every experimental measurement. In order to compare proper peak volumes it is necessary to correct the volume of each peak depending on the local noise level [Trenner, 2006]. The two datasets (two peaks associated between the reference and the test spectra respectively) are assumed to be both sampled from Gaussian populations relying on the central limit theorem. The size of the considered populations is sufficiently large to be considered tending to the normal distribution (sample size of at least 30 peak voxels P_X). The two populations are not assumed to have the same standard deviation (volume error P_N due to the presence of noise). The main purpose of such peak volume (P_V) comparison is to quantify how far apart the two volumes are. According to the eq. 3.2

and 3.14, the minimum volume error P_N contributed by the local noise level of each considered peak has been computed and scaled. Under such assumption each peak volume (P_V) varies in a range defined by the volume error ($P_V \pm P_N$) depending on the local noise level. Confidence intervals (C_R, C_T) around the peak volumes have been generated. The unequal-variance (Welch) t-test [Welch, 1938] (see appendix B) has been used in order to verify the null hypothesis (two population means are the same) and to compute the probability that the volumes are drawn from the same distribution (is the same cross-peak of the same protein in accordance to the volume feature). An example of the test is shown in Fig. 3.37.

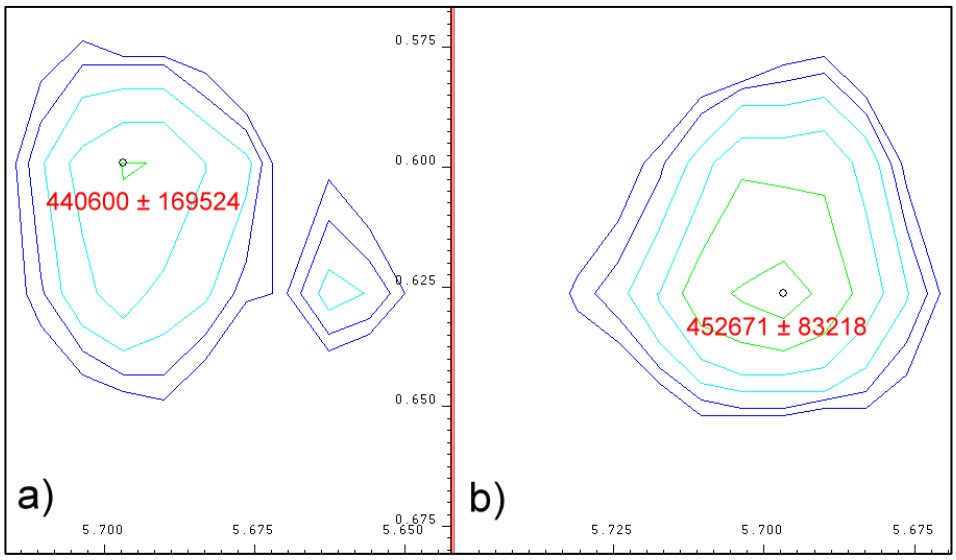


Fig. 3.37 Comparison of the peak volume feature based on the Welch test: the reference peak (a) and the associated test peak (b) of the NOESY spectrum of the HPr protein from *Staphylococcus aureus* wild type (a) and mutant (b) (H15A) one. The computed t value of -0.55 associated to a critical value of 5% in a two-tailed t-test corresponds to a matching probability of 0.58%.

The Fig. 3.37 shows an example of a peak volume feature analysis. The peak of the reference spectrum (a) is compared with the associated one in the test spectrum (b) showing different volumes (P_V) and a different volume errors (P_N). The test peak (b) has

been automatically scaled by a factor of 2.11 (see eq. 3.2 and 3.14). The result of the comparison using the Welch test is a matching probability of 58% (t value of 0.55). The unequal-variance (Welch) t-test has been computed for all the reference peaks generating matching probabilities based on the volume feature.

3.4.3.3 Assessment of complete spectra

Once the peak classification has been performed through the Bayesian analysis it is possible to define if and with which probability the test spectra are generally similar to the reference one. In particular, two main results can be extracted: the spectral matching ratio based on the Bayesian probability of all the features and the general yes/no answer that relies on the feature scores analyzed separately by the Kolmogorov-Smirnov analysis.

3.4.3.4 Spectral matching ratios

In a first analysis, the user is allowed to choose a feature of interest whose Bayesian probability is automatically computed by the routine. Once the feature has been selected, the spectroscopist has at his disposal color highlighted peaks (peaks whose Bayesian probability is in a range defined by the user) that eases the task of spectral analysis. In particular, dealing with HSQC-type spectra allows the identification of two main categories of peaks: the high probability signals (with a probability > 0.66) that are not highlighted at all and the mid-low probability signals (including the completely missing peaks) identified with red boxes. This color based signal classification helps the spectroscopist to perform a fast identification of those more different peaks when comparing spectra of interest. In case of NOESY and TOCSY spectra the global symmetrical properties may be exploited in order to obtain a further classification of the low probability peaks (see par. 3.4.4.2). They can be distinguished in missing, new and ambiguous signals identified by different colored boxes.

One example of this classification based on the cosine similarity feature, is reported in the three-dimensional plot of the Bayesian peak distribution shown in Fig. 3.38. It represents the cosine similarity probability distribution between HPr protein from

Staphylococcus aureus wild and mutant (H15A) types. It is evident that mutating only one residue, the majority of the peaks are unaltered (red circles) from a shape point of view, whereas some of them show mid-low probabilities (green and blue circles). They are investigated in the next (high-level) analysis level.

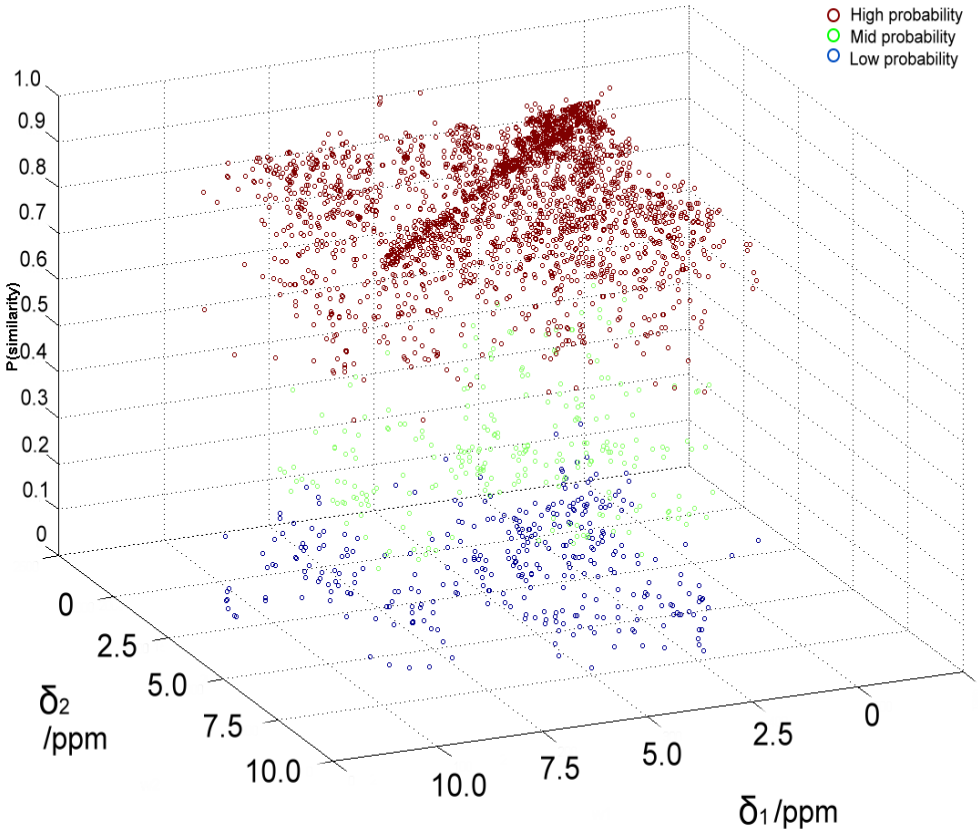


Fig. 3.38 Three-dimensional scattering plot of the Bayesian probability distribution based on the cosine similarity criterion: The plot represents the three different probability levels between the experimental HPr protein from *Staphylococcus aureus* wild type and the experimental HPr protein from *Staphylococcus aureus* mutant *(H15A)* type. Three different probabilities are reported: high probabilities (red circles), mid probabilities (green circles) and low probabilities (blue circles).

As shown in Fig. 3.39, applying a projection of the z-axis it is possible to recognize some interesting features like missing peaks (blue stars), due to the fact that either the local shift selected by the user was not enough in order to find the corresponding peak in the test spectrum or the peaks are really not present. The most important information is to find peaks of the test spectrum with a notable shape variation with respect to the reference one

(green and blue stars). This measure is given by the probability of the cosine similarity calculated for each peak. The same procedure has been calculated for all five analyzed features (volume, shift, shape, line width and cross-correlation in the time domain). It is possible, under realistic conditions, to recognize base line distortions, artifact peaks, peaks raised from a wrong phase correction and so on.

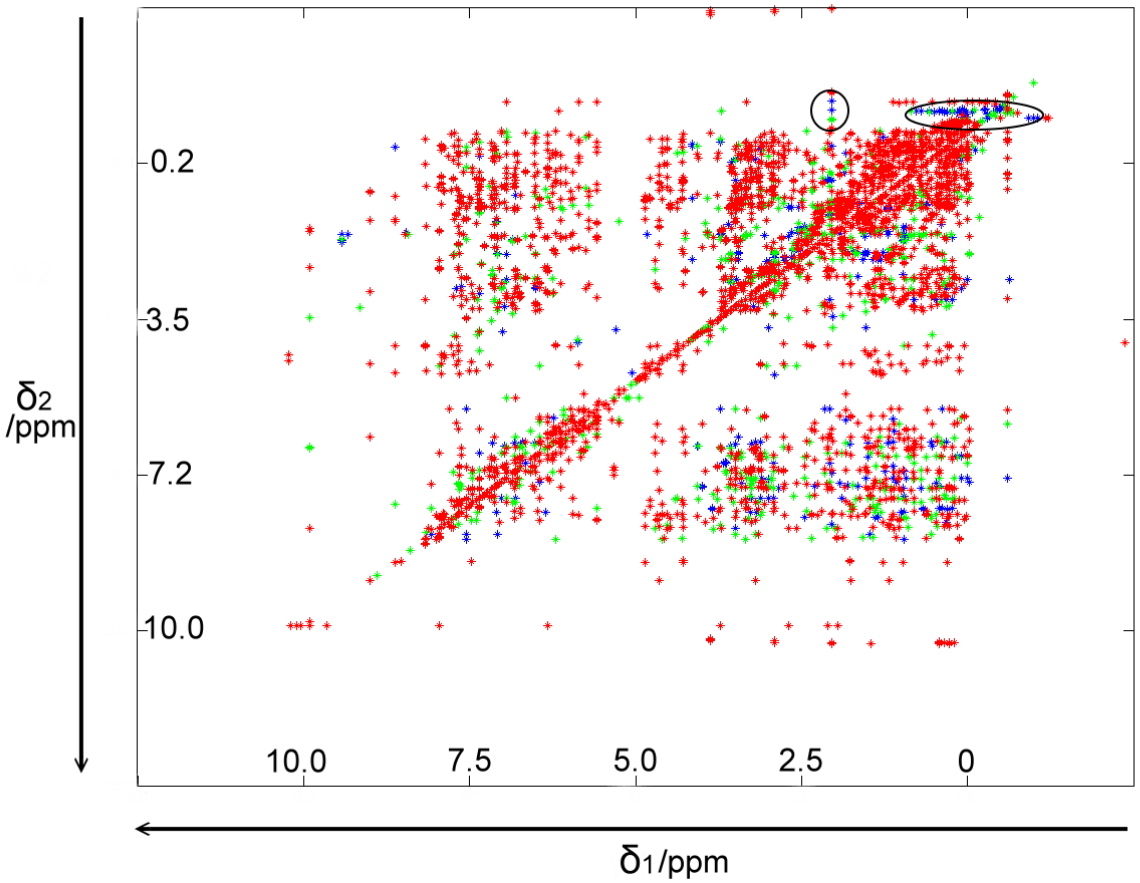


Fig. 3.39 Z-axis projection of the two dimensional Bayesian probability distribution: the figure shows the projection of the probability computed between the HPr protein from *Staphylococcus aureus* wild type and the HPr protein from *Staphylococcus aureus* mutant type *(H15A)* using the cosine similarity criterion. Baseline distortions are highlighted (black circles). In particular, three different probabilities are reported: high probabilities (red crosses), mid probabilities (green crosses) and low probabilities (blue crosses).

A step further is to combine such results among them, mixing the probability of the cosine similarity with the probability of the correct peak position. In this manner is possible to obtain a probability that is dependent on more than one feature as shown in Fig.

3.40

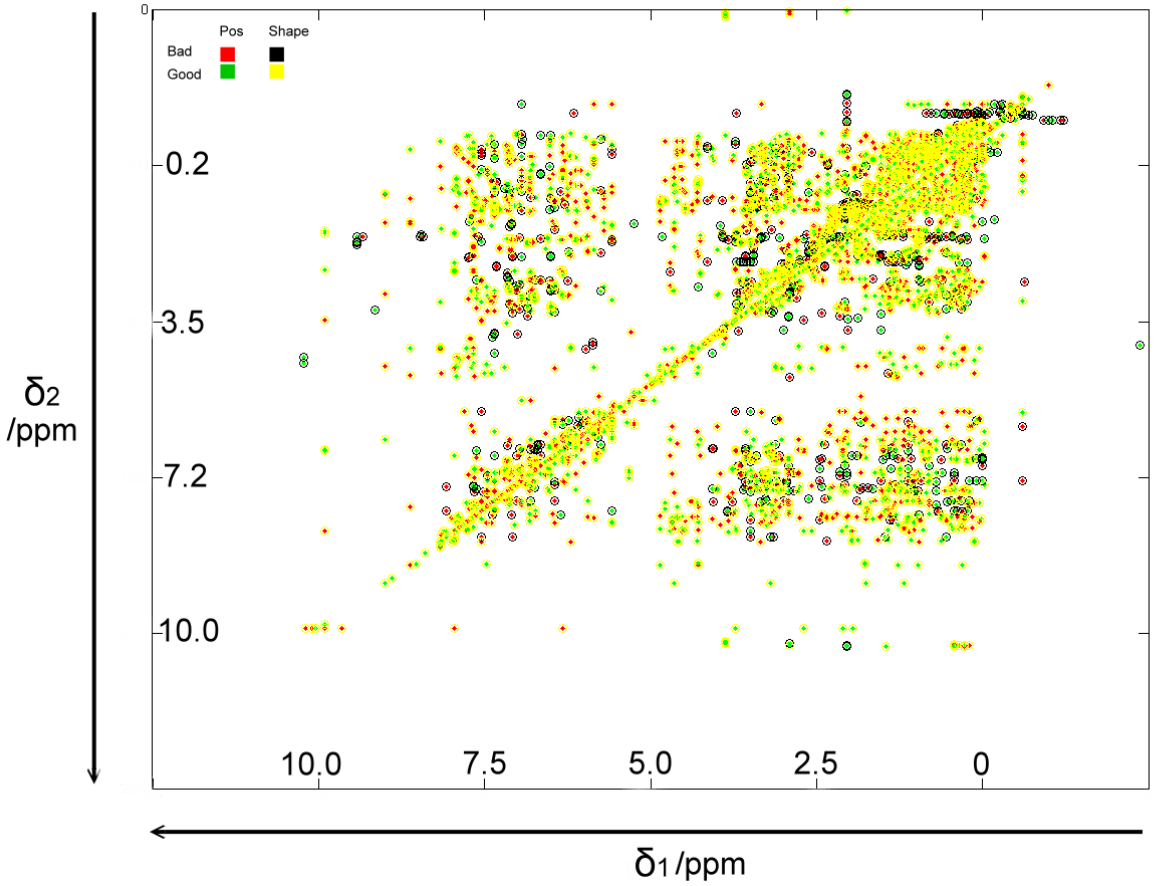


Fig. 3.40 Cross probability of the Bayesian distribution comparing the position and the cosine similarity features: The two-dimensional projection of the Bayesian probability between HPr protein from *Staphylococcus aureus* wild type and measured HPr protein from *Staphylococcus aureus* mutant (H15A) type using the cosine similarity criterion and the peak position shift. Some peaks reveal high probabilities on both features (yellow-green circles) while other show a low probability on one or two features (black-red circles).

As shown in Fig. 3.40, the majority of the peaks has a high position and cosine similarity probability (yellow-green circles). Some of them reveal a good shape probability after a large local shift (yellow-red circles), while some others have an almost perfectly matching position with a strong shape variation (black-green circles). A few of them show a very low probability in both cases (black-red circles). They can be used to detect baseline distortions and artifacts signals.

In addition, the user can interactively select any peak in any spectrum. By pushing the keyboard character “q” a window appears. This latter, for every selected peak, contains information related to the names (over all spectra, if available), the ppm positions, the ppm shift differences, the volumes before and after scaling, the line widths (in Hz), the presence or absence of multiplet structure and the Bayesian probabilities of every feature. Beyond this first visual check the user can obtain global spectral matching and mismatching ratios. They define the number of signals that has a probability major than a threshold of 0.66 for all the features simultaneously with respect to the total number of peaks in every considered spectrum. The matching factor F is defined by

$$F = \frac{N_{HP}}{N} \tag{3.27}$$

whit N_{HP} the number of all peaks with p > 0.66 (for all the features simultaneously) and N the number of peaks of the investigated spectrum.

3.4.3.5 Kolmogorov-Smirnov analysis

Another important step of the mid-level analysis routine is to compute a probability (additional to the matching ratio) that has a wider point of view: the probability that two spectra are the same (i.e. different measurements of the same protein). The idea behind is to obtain an answer telling the user if the measured spectra are the same with respect to the reference or not. This question can be defined more precisely: “Are the peaks in the reference spectrum or the reference spectra identical to those in the test spectrum within a given limit of error (e.g. 5%) ?”

The proposed solution is to compute the Kolmogorov-Smirnov test [Kolmogorov, 1933; Massey, 1951] that is applicable to distributions that are not binned with respect to previously selected categories and obviously non-parametric. It tests the null hypothesis that data are drawn from the same distribution. The power of this test relies on the fact that it is a “distribution free” test not requiring the normally distributed population. This test is sensitive to shape changes (i.e. skewness) between two distributions. In particular, the considered dataset (one for each feature) must be represented as a function $S_P(z)$ with z =

1,.., P, of the cumulative fraction of data points minor than a value z. The KS-test calculates the maximum distance D in absolute value, between cumulative distributions of the two samples (i.e. same feature distributions between reference and test spectra):

$$D = \max_{0 \le z \le 1} \left| F_{i_T}(z) - F_{i_R}(z) \right| \tag{3.28}$$

where F_{i_T} and F_{i_R} represent the score distribution of the feature *i* considering the reference spectrum *R* and the test spectrum *T* (see Fig. 3.41). The significance of the distance D (within the limit of 5% of error) is calculated as following:

$$K(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2} \tag{3.29}$$

where

$$\lambda = D\left(0.12 + \sqrt{\frac{n_T n_R}{n_T + n_R}} + \frac{0.11}{\sqrt{\frac{n_T n_R}{n_T + n_R}}}\right) \tag{3.30}$$

where n_T and n_R represent the number of computed scores (the amount of scores is the same as the number of peaks of each considered spectrum) for the reference and the test spectrum respectively. If $\lambda \to 0$ then $K(\lambda) \to 1$, while if $\lambda \to \infty$ then $K(\lambda) \to 0$. Thus, the null hypothesis is true when $K(\lambda)$, representing the probability that the two datasets are drawn from the same distribution, gets close to one (i.e. the distance D approaches zero). The null hypothesis is then validated in the error limit of five percent.

As shown in Fig. 3.41 the KS-test computes the cumulative fraction function of the two shape distribution datasets having a maximum vertical distance D = 0.1138. In the error limit of five percent, the null hypothesis has been rejected. The same decision has been reached using the Student's t-test in the error limit of five percent.

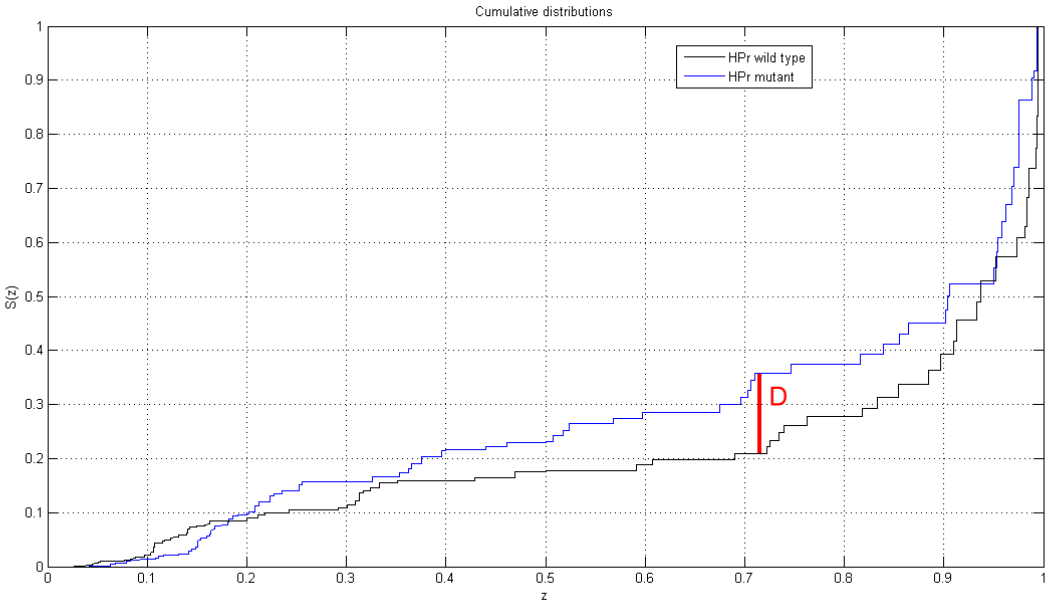


Fig. 3.41 The cumulative fraction plot of the KS-test: the plot shows the cumulative fraction of the KS-test based on the cosine similarity criterion. The test has been performed using the feature scores computed between the experimental HPr protein from *Staphylococcus aureus* wild type (black) and the mutant type (blue).

This test has been performed analyzing the scores of the cosine similarity feature. The same KS-Test has been conducted over all the previously collected features separately in order to differentiate the results having a broader statistical view of the comparison. For example, it is possible to compare spectra by means of the volume distributions of two datasets using the KS-test. This test gives relevance to the volume feature instead of cosine similarity although all features are interconnected each other (i.e. volume increase probably means a peak shape change that leads to a line width difference and so on).

Once, all information coming from the low-level have been converted into statistical entities it is still necessary to expand the comparison including biochemical properties of the compound. The high-level stage has been developed for this reason.

3.4.4 High-level analysis: investigating structural changes

The core of the whole routine lies in the low-level analysis where all the main algorithms have been developed to automatically collect and associate relevant information. In the mid-level analysis the values related to each feature (volume, position, line width, time correlation and shape) are expressed as probabilities in order to perform an automated peak classification (i.e. distinguish between high and mid-low probability peaks). This latter is important to ease the spectroscopic analysis and it is used to obtain a general overview on the considered dataset. After computing spectral matching ratios and statistical probabilities, it is possible to define if the structures represented by the spectra in hand are identical or similar and in which percentage. The high-level stage has been developed in order to express spectral properties into molecular features.

Independently on the available spectral assignment, the routine allows to recognize and to store all the signal variations and in the worst case where even the reference spectrum has not been previously assigned, the ppm positions of the most altered resonances (for any feature of interest) can be automatically identified. This final level has the goal to ease the task of recognizing the residues which correspond to the strongest signal variations in the spectrum. In this manner, both a qualitative and a quantitative study of the amino acids involved in the external changes has been conducted in order to define the percentage of altered three-dimensional molecular structure and to identify if those residues belong either to critical or to non-essential regions.

3.4.4.1 Interpretation of peaks

The first part of the high-level stage has been differently developed depending on the type of considered experiment. In particular, as shown in Fig. 3.42 the routine follows two different approaches after identifying the type of considered spectra:

1. after the initial signal association, in case of HSQC-type spectra, it evaluates the average variation of the user selected feature (e.g. volume and chemical shift) of all the resonances (residues). In addition, it identifies those ones whose changes

exceed the double of the standard deviation computed over the entire dataset generating a list of most changed residues.

2. dealing with NOESY and TOCSY spectra, before creating the list of most altered resonances, it evaluates the global symmetry properties of the spectrum. The additional residue pattern recognition of the peaks classified as new, missing and ambiguous is performed in order to verify if the typical pattern of a certain residue has moved somewhere else in the spectrum, even with a remarkable volume variation.

In both considered cases the analyzed peaks can be clustered in two classes (with a threshold of 0.66): high probability signals whose features have been conserved (all features simultaneously) among the spectra and low probability resonances showing a very strong variation of any feature. This separation in two classes has been applied in order to simplify the evaluation of the data. Further different classification may be eventually analyzed.

When HSQC-type spectra are evaluated, the high probability signals are considered as matching signals, while the low probability resonances are identified as not matching peaks. These latters include also peaks that have been picked in the reference spectrum but not in the test spectra or vice versa, thus defined as missing signals. Since these types of spectra do not possess global symmetry properties, it is not possible to distinguish the missing resonances as new, ambiguous or completely missing peaks. The global symmetry of NOESY and TOCSY spectra allows this further classification as explained in the next paragraph.

Independently on the type of spectra in hand, this initial classification in two clusters can be easily visualized by the user in a colored fashion (colored boxes around low probability peaks).

The behavior of the feature of interest (e.g. chemical shift) is investigated for every signal in both classes. Actually it is performed only on HSQC-type spectra where almost each peak corresponds to a residue. In principle it would be applicable also to NOESY and TOCSY spectra where the feature behavior can be mapped both peak-wise and pattern-wise.

The signals included in the high probability cluster typically show a feature variation that is not particularly relevant, whereas the other ones possess a strong feature change that eventually can be greater than the double of the standard deviation of all the considered peaks. In case of HSQC-type spectra they can be identified as the most altered residues. They are accordingly displayed in a list containing the residue name (if the reference spectrum was previously assigned) or the atom positions of the reference spectrum (if all the spectra in hand were not assigned at all). In case of NOESY and TOCSY spectra this list would contain the name of the peak (independently on the assignment) and the ppm position in both directions.

The spectroscopist can personally decide to analyze either the volume or the chemical shift behavior obtaining a different list of residues involved in the process.

Dealing with heteronuclear data the chemical shift changes can be evaluated in both directions simultaneously (Hamming distance) with a variety of weighting factors different for each nucleus and for each residue [Schumann et al., 2007]. This step is particularly interesting when a series of test spectra has been acquired with a gradual variation of any possible external factor (e.g. pressure, temperature and pH) thus, the behavior of every signal can be mapped automatically without any user intervention. The structural modifications due to protein-ligand interactions can be investigated as well. In homonuclear spectra (NOESY, TOCSY and COSY types) the available information can be deeper analyzed. In such cases it is not possible to correlate almost every peak signal with a residue (as in HSQC-type) but they have at least two strong points that can be exploited for our purposes: they possess global symmetry properties and they contain residue patterns that allow the identification of a wider dislocation of any residue through the spectra. Using the global symmetry properties the low probability peaks are distinguished in three classes: new, ambiguous and missing signals. Entire new, ambiguous and missing patterns are detected and compared among the spectra. This further classification of the low probability peaks is particularly interesting for a visual analysis of the compared spectra since the peaks will be surrounded by boxes with different colors depending on their belonging class (red for missing peaks, blue for ambiguous signals and green in case of new peaks).

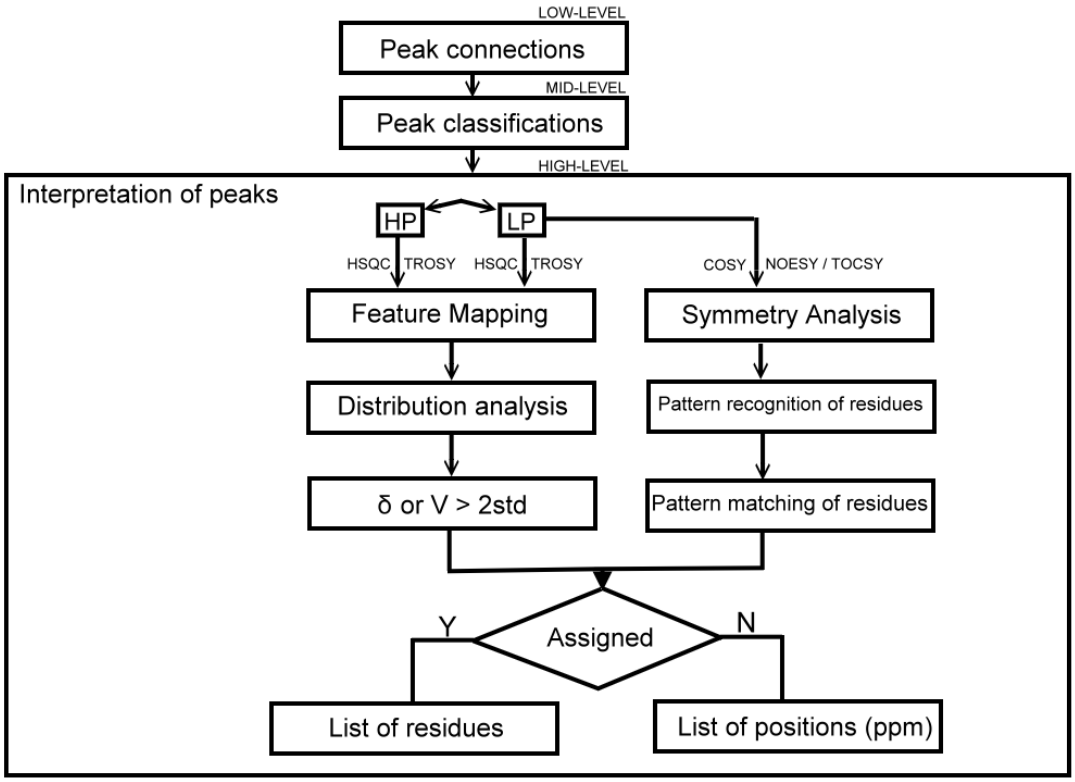


Fig. 3.42 High level analysis: after the low (peak association) and mid-level (peak classification) analysis the high-level (peak interpretation) is performed. The feature mapping of the high and low probability peaks is applied in HSQC-type spectra. A successive analysis of the average behavior is evaluated and the signals whose feature variation is greater than the double value of the standard deviation are retained. Low probability (LP) peaks belonging to NOESY, TOCSY and COSY spectra are further classified by means of symmetry properties. Entire new, ambiguous and missing patterns are detected and compared among the spectra. A list (dependently on the availability of a previously assigned reference spectrum) is generated that contains the residues or the ppm atom positions mostly involved in the spectral and in the structural changes.

3.4.4.2 Global symmetry analysis and refinement of peaks

The data collected in the low-level are used to perform the Bayesian analysis in accordance with the feature of interest. This procedure allows classifying each peak in order to verify if it is a matching signal among the spectra and with which probability. In particular, if this probability related to the feature of interest is higher than a threshold value of 0.66 it automatically considers the association as a correct matching case. When dealing with multidimensional NOESY, TOCSY and COSY spectra, some mismatching errors can be

taken into account for those cases falling below that threshold in order to guarantee an effective distinction between real cross peaks and noise or artifacts signals.

In these latter cases or if the considered peaks have not been connected at all the routine exploits the symmetry properties of cross peaks with respect to the diagonal to check the correctness of the association. The method looks for symmetrical cross peaks using the position information, thus it neglects shape, line width and volume differences between each couple of cross signals due to inhomogeneous magnetic fields, relaxation effects, insufficient digital resolutions, improper base line corrections and inappropriate solvent artifact suppressions. As shown in Fig. 3.43, it is rather easy to combine symmetry-related cross peaks. The case described below is not related to any type of spectral comparison since it just intends to represent the symmetrical pairs found in the reference two dimensional NOESY spectrum of the wild type HPr protein from *Staphylococcus aureus*.

It is evident that some of the non-symmetrical signals belong to base line distortions, some others are gathered in the solvent regions, a few of them represent pure noise peaks and some peaks do not possess any symmetrical association. Many of these latter do not match the symmetrical signal because its line width has been distorted and thus it has been no more recognized even as a peak.

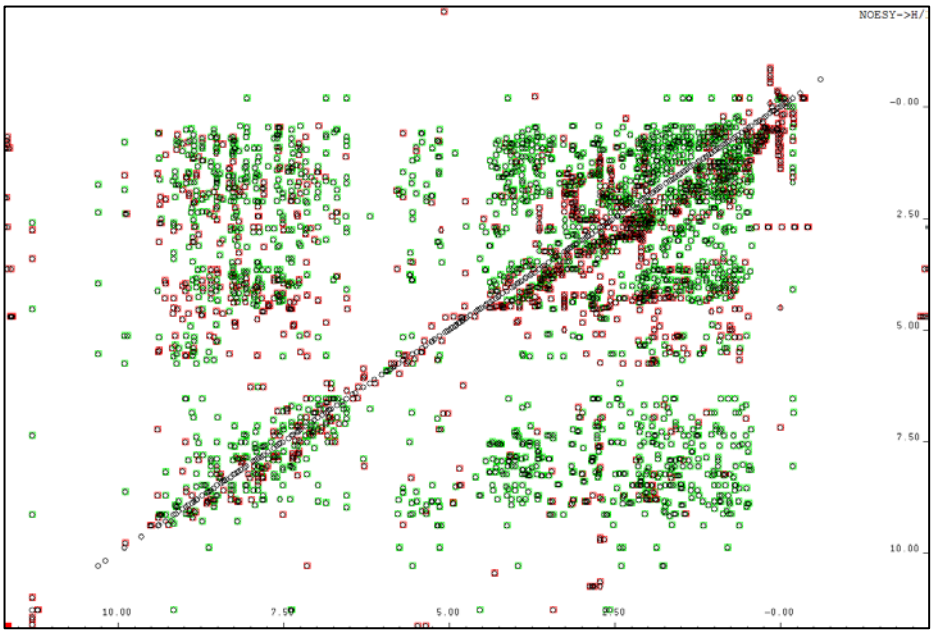


Fig. 3.43 Symmetry property in a NOESY spectrum: Symmetrical pairs (green) and not symmetrical signals (red) from the two dimensional NOESY spectrum of the wild type HPr protein from *Staphylococcus aureus.*

The search and the analysis of the symmetrical cross peaks is of particular interest when the association of the considered signals has a low probability or when it has not been found. In particular, starting from a not associated peak P_{ij}^{R} of the reference spectrum *R*, three different situations are evaluated:

1. its symmetrical peak P_{ji}^{R} is not associated to any other signal P_{kh}^{T} in the test spectra
2. its symmetrical peak P_{ji}^{R} is connected to the peak P_{kh}^{T} in the test spectra
3. it has no symmetrical peak P_{ji}^{R}

where *i* and *j* represent the ppm positions of a reference peak while *h* and *k* are the chemical shifts of the peak belonging to the test spectrum.
In the first case both peaks P_{ij}^{R} and P_{ji}^{R} join the list of the completely new signals that has not been found in all the other considered spectra. In the second option the peak P_{ij}^{R} is

identified as a “pseudo-missing” or ambiguous peak meaning that some external influences, described above, have led to the loss of its connected test signal P_{kh}^{T}. In the third case the peak P_{ij}^{R} is classified as a missing signal (typically noise or artifacts). An exception is done when the not found symmetrical peak P_{ji}^{R} should lie in the 0.2 ppm range around the typical solvent region (4.6-5.0 ppm): the correspondent P_{ij}^{R} peak does not represent a missing signal but it is considered as a completely new one. The same principles are applied over all the low probability peak associations.

The above described refinement of the investigated low probability signals allows a more complicated color-related class display of the NOESY, TOCSY and COSY spectra. As defined in the previous section, the HP peaks are not highlighted, while the MP-LP signals are sub-classified in the following manner:

1. mid-low probability signals and the missing ones whose symmetrical cross peak is a mid-low probability or a missing signal too are highlighted with green boxes (they are candidate to be considered new arising signals).
2. A blue box surrounds the peak when dealing with ambiguous peaks (also called pseudo-missing peaks).
3. A red box is used when the symmetrical cross peak has not been found.

The equation 3.27 needs to be modified as following

$$F_{HP} = \frac{N_{HP} + N_{PM}}{N} \tag{3.31}$$

where N_{HP} denotes the number of peaks with $p > 0.66$ (for all the features simultaneously), N_{PM} represents the total pseudo-missing signals and N is the number of peaks of the spectrum. When dealing with almost identical spectra, the matching ratio must get closer to one. If the compared spectra have been partially or completely assigned, the term N_{HP} of eq. 3.31 includes the peaks with a matching assignment inde.

3.4.4.2.1 Signal patterns recognition of residues in NOESY and TOCSY spectra

Signal pattern analysis is performed on patterns containing resonances with a low probability (according to the Bayesian analysis) that have been classified as new, ambiguous or missing peaks by means of global symmetry properties (thus, it is applicable only on NOESY, TOCSY and COSY spectra). Considering the case that both the reference and the test spectra are not assigned, the patterns are evaluated according to the following steps:

1) the algorithm looks for each diagonal peak starting from the down-left to the upper-right side of the reference spectrum (2D experiment).
2) In a two-dimensional reference spectrum the pattern spanned in both directions (vertical and horizontal patterns) from the considered diagonal peak is analyzed in order to verify the amount and the class of peaks lying on these patterns. If no peaks are detected the pattern (vertical and horizontal patterns) is discarded. A peak is considered to be part of the vertical pattern when it keeps the same ppm position of the diagonal peak in the indirect direction (δ_1) allowing a ± 1 voxel variation in the other direction (δ_2). The peak joins instead the horizontal pattern if the opposite conditions are fulfilled.
3) From the mid-level analysis each peak belonging to a certain pattern (in the reference spectrum) is already classified as high or low probability signal. The symmetrical properties of the low probability peaks of a considered pattern are evaluated in order to classify them as new, ambiguous or missing peaks (as described in the previous paragraph). This further classification enables the spectroscopist to obtain visual results which simplify spectral investigation (see par. 3.4.4.1).
4) The amount of low probability peaks in a reference pattern is computed. If this amount corresponds at least to the 80% of the peaks found in the pattern, the reference pattern is identified as a new one.

5) Each reference pattern must be associated with its corresponding test pattern spanned from the ppm position of the reference diagonal peak ($\pm$ 1 voxel).
6) The same analysis (from step one) is repeated beginning from the test spectrum. This is particularly useful to discover new test patterns that would not be reached otherwise.
7) A pattern matching is performed:
    a) Patterns not recognized as new are compared one by one between the reference and the test spectrum.
    b) Each new pattern in the reference spectrum is compared with all the new ones in the test spectrum.

In both cases a pseudo energy *E* term representing the matching ratio (similar to the eq. 3.21) between compared patterns has been computed as following

$$E = \frac{1}{\mathrm{MAX}\left(N_{path_R}, N_{path_T}\right)} \sum_{i=1}^{N_{path_R}} \left(1 - \frac{|SH|_i}{SH_{max}}\right)(1 - Z_i) \qquad (3.32)$$

where $|SH|_i$ is the term representing the existing shift between the position of the *i*th peak along each pattern (reference and test) and the respective ppm position of the diagonal signal. SH_{max} is the maximum allowed shift ($SH_{max} = 2$ in case of a $\pm$ 1 voxel variation). N_{path_R} and N_{path_T} are the number of encountered peaks (peaks with volume P_V) along the spanned patterns of the reference and the test spectrum respectively and Z_i represents the symmetry factor of scaled volumes of the *i*th couple of compared peaks along the analyzed patterns (see eq. 3.21).

The pseudo-energy *E* is in function of the maximum number of low probability peaks encountered in the two compared patterns, since this maximum number is more discriminating than taking into account only the common peaks between the compared patterns. It is also related to the amount of volume symmetry between each couple of compared peaks along the considered patterns. Moreover, the shift SH_i determines the optimal choice of peaks along each pattern.

The pseudo-energy *E* must be almost one in case of completely unchanged patterns between the reference and the test spectra, whereas it decreases when the compared

patterns are more different. This score is particularly useful for the spectroscopist since it eases the identification of those resonances that have experienced volume variations (corresponding to structural distance variations) along almost identical reference and test patterns.

In the former case (7a) the matching ratio defines the similarity score between two considered patterns.

In the latter case (7b) the matching ratio is used to identify the best matching between a new reference pattern and any possible new test pattern. A list of scores is stored and the highest one is considered the best matching case. This wide comparison allows finding a specific new reference pattern even at very different ppm positions in the test spectrum.

If the reference spectrum has not been previously assigned the detected patterns from diagonal peaks can be evaluated in order to discover the amino acid correspondence. Some residues in fact, have a unique pattern (e.g. Gly, Ala, Thr, Ile, Val and Leu), while others possess very similar patterns. Overlapping, baseline and solvent artifact may cause problems on the complete identification of those patterns and must be taken into account during the visual inspection.

3.4.4.2.2 Application of the signal pattern recognition (NOESY case)

An example of the above described algorithm for pattern identification is reported in Fig. 3.44. The algorithm represented in Fig. 3.44 attempts to show some of the typical situations encountered when comparing similar spectra of the same protein. In accordance to the step number four, it starts looking for patterns containing mostly (major than 80%) low probability peaks in the previously assigned reference spectrum. In particular, it has found the typical patterns of alanine (Ala19) and glycine (Gly24) in the reference spectrum made up exclusively of peaks classified as new. After performing the step number six, the routine matches those reference patterns with any new pattern in the test spectrum. In this example, the algorithm has been able to recognize them in the test spectrum even with a

strong dislocation (downfield) and with additional peaks along the compared patterns (as the peak in the HN ppm range along the test Ala19 pattern).

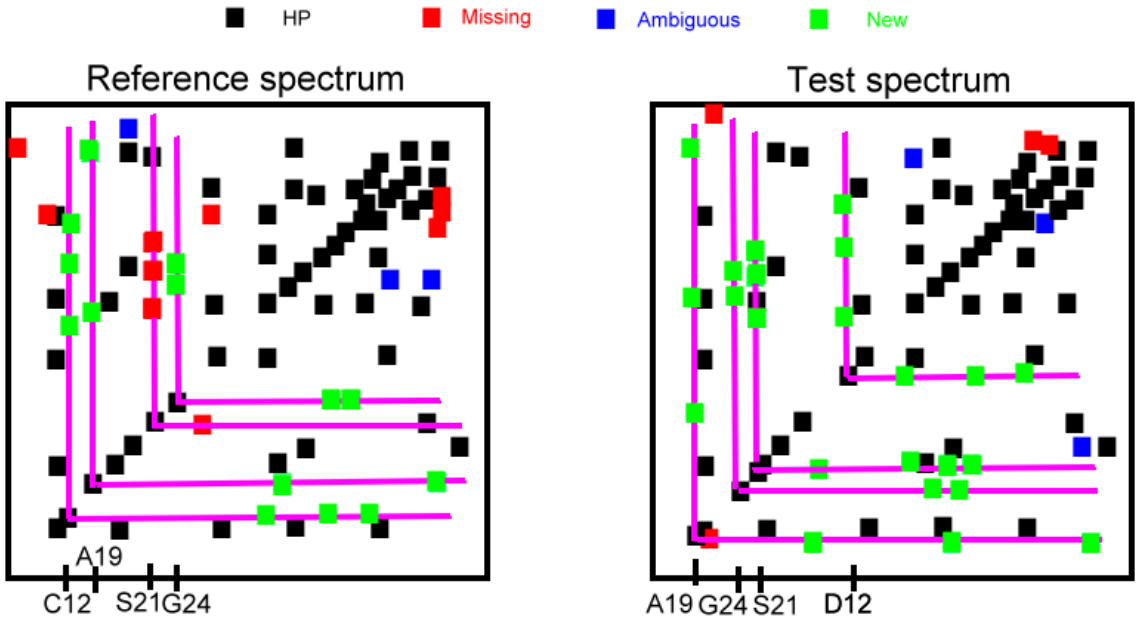


Fig. 3.44 A toy example of pattern matching in NOESY spectra: residue pattern matching between the reference spectrum and the test spectrum. The reference patterns of the Ala19, Ser21 and Gly24 have been found at shifted position in test spectrum; the pattern of the residue Cys12 is completely changed due to a mutation thus it has not been detected in the test spectrum.

The algorithm does not simply look for completely identical patterns all over the spectrum, but as in the case of Ala19 it allows the presence of any type of peak (high probability, new, missing or ambiguous signals) along the matching patterns with the discrimination described in the eq. 3.32. Obviously, additional peaks may be found along the considered patterns due either to pattern overlapping or to structural changes. As in the case of Ala19, the considered pattern in the test spectrum shows an adjunctive transfer of magnetization between its HN proton and another one belonging to a certain other amino acid representing a structural variation. In the case of the serine (Ser21) the algorithm is able to recognize the pattern composed of missing peaks (they have not been associated to any other signal in the test spectrum and they do not possess any symmetrical peak that

eventually is disappeared after a very strong baseline correction or solvent reduction) and it finds the same pattern shifted downfield in the test spectrum. The pattern of the residue Ser21 has been intentionally positioned in overlap with other ones both in the reference and in the test spectra in order to show the capability of the method of dealing with low probability patterns hidden by other already well associated ones. The pattern of the cysteine (Cys12) has very low matching scores with any other new test pattern since this residue was previously involved in a mutation. It is no more present in the test case and it must not be wrongly associated to the pattern of the aspartic acid (Asp12).

When the reference new pattern matches more than one pattern in the test spectrum, the association with the highest score is retained according to the eq. 3.32. If the reference spectrum was previously assigned, the routine stores a list of residue names and their ppm proton positions in both spectra in order to identify the shift of the whole pattern between the considered spectra. When dealing with not assigned datasets the routine produces only the list of the ppm proton positions of the associated patterns in both spectra without any residue identification. The list generated by the developed algorithm applied in the case reported in Fig. 3.44 is shown in Table 3.2.

Residue	**Reference ppm (w2)**	**Test ppm (w2)**
Cys12	8.76	-----
Ala19	8.23	8.96
Gly24	7.34	8.14
Ser21	7.81	8.03

Table 3.2 Pattern associations of the toy example reported in Fig. 3.44: The pattern of the residue Cys12 present in the reference spectrum is no more recognizable. The pattern of the residue Ala19 has been found shifted by 0.73 ppm. The same effect has been found for the residue Gly24 and the residue Ser21.

3.4.4.3 Structural analysis

The final part of the project attempts to give a general overview on the structural changes extractable comparing the reference and the test spectra. In particular, the peak interpretation previously described, allows the direct detection of both altered and unaltered signals. The method has been developed to ease the spectroscopic task of finding the strongest variations between two or more dataset of interest. It yields colored boxes around the peaks in each spectrum in order to give a first reliable visual impact on the data, but it has also been thought to allow the storing of any variations of all the considered features.

In accordance to Fig. 3.42, two different spectral analyses are performed on the data producing various results. Actually, the routine generates histograms and corresponding output files only when HSQC-type spectra are compared. The histograms of the selected feature (e.g. chemical shift, volume) for all the peaks (in these cases residues) in the reference spectrum are used to investigate the average and the standard deviation of that feature. In particular, three different histograms of chemical shifts may be generated:

1. Shift along δ_2(proton chemical shift).
2. Shift along δ_1(nitrogen or carbon chemical shift)
3. Combined shift $\Delta\delta_{comb}$ in both directions [Schumann et al., 2007].

The user can decide to obtain volume histograms representing the normalized volume (NS, RG and NC_proc) change between the spectra with or without considering the volume scaling S_{TR} (in accordance to eq. 3.14). In specific cases (e.g. pressure, temperature mapping or ligand screening) the volume scaling can be intentionally avoided in order to analyze the volume variation. In other cases (i.e. standard quality control) it can be intentionally used.

Since the HSQC-type spectra contain almost as many peaks as the number of residues, it is possible to obtain such histograms showing the feature behavior in a crescent order based on the protein sequence.

When the peaks are associated between the reference and test spectra (either with a high or with a low probability) the histogram shows the feature change. In case that a specific peak is not connected to any compared one (it is a missing peak) the feature variation is not expressed in the histogram.

When more than one test spectrum is compared with the reference one, the user can obtain more histograms: (1) the histogram containing the feature mapping between the reference and the selected test spectrum (one bar for each residue); (2) the histogram of the feature investigation among the reference and all the available test cases (as many bars for each residue as the number of test spectra).

In addition, the histograms can express different information:

1. They represent the differences (e.g. in ppm unit for the chemical shift) between the values (of the feature of interest) of each residue in the reference and in the test spectra.
2. They contain the slope of every linear interpolation (one for each residue) of the differences of the selected feature.

When the chemical shift is the feature of interest, the histogram can furnish additional information about the direction of the peak dislocation that is directly related to the sign either of the chemical shift difference or of the slope. For instance, a negative sign corresponds to a downfield shift of a proton in the direct direction (in case of a two-dimensional experiment).

Dealing with HSQC-type spectra yields a further dialog containing the name of the peaks (or residues) exceeding the standard deviation in accordance to the histograms. Additional ascii files are produced containing the variations either of the normalized volume (NS, RG and NC_proc) with or without volume scaling or of the chemical shift of every signal (residue) in the spectra. In particular, these files provide a list of information in accordance to the histograms requested by the user. These files are structured as a data matrix as follows:

a) The rows represent the investigated spectra
b) The columns represent the peaks
c) The data matrix contains the values of the user selected feature.

If NOESY and TOCSY spectra are compared, actually the routine does not furnish the histograms and the above described output files. It yields only the list of shifted patterns as shown in the previous section (see Table 3.2). The computation of fractions of structural variations represents a further procedural method that still needs to be implemented. These additional routines may be introduced to produce more information when dealing with those types of spectra (see the discussion section).

The residue patterns dislocation can actually be well identified and in case of previously assigned reference spectrum it allows the extraction of structural information. When the comparison is performed over a set of spectra representing the same protein before and after binding with a newly designed drug, the analysis is particular useful since it reveals not only the percentage of altered three-dimensional structure but also the residues involved in the process verifying the criticality of the most implicated regions.

4

Test case: HPr protein from *Staphylococcus aureus*

This chapter shows the practical application of the previously described methods (see chapter 3) to the HPr protein [Görler et al., 1999b, Maurer et al., 2004] from *Staphylococcus aureus* (see par. 2.2.2.1). The HPr wild type and the mutant one (H15A) have been investigated by the AUREMOL-QTA routine under four different conditions. In particular, the method has been applied to test its ability for characterizing site-directed mutagenesis and conformational changes by means of measured and back-calculated NOESY spectra.

4.1 Introduction

The AUREMOL-QTA routine automatically performs the pre-processing steps as described in par. 3.4.1. The user needs to determine which ones of the provided spectra must be considered as reference and those ones representing the test cases. The datasets that have been used are described in par. 2.2.1 and par. 2.2.2. In particular, four different cases (2D NOESY spectra in all cases) have been investigated:

1. the spectrum of the HPr protein (wild type) has been compared with that one of a mutant case of HPr (H15A). The chemical shift assignments of these proteins are available but this case has been analyzed without using this information. This has been done in order to show the abilities of the AUREMOL-QTA method to compare not assigned spectra and to analyze the effects of a small perturbation (e.g.

a point of mutation at position 15) on the considered spectra. The available three-dimensional structure of the HPr protein has not been used. The method automatically detects complete peak pattern dislocations between spectra.

2. The second case is similar to the first one with the additional reference spectrum (2D NOESY) obtained from the known three-dimensional structure and the chemical shifts of the HPr protein (wild type). Here, the simulated multiplet structure can be used for obtaining more precise line widths.

3. In the third case a natively folded protein is compared with a partially denatured one. In particular, the simulated 2D NOESY spectrum of the folded protein has been compared with the simulated 2D NOESY spectrum of the partially denatured one. The partially denatured spectrum of the folded protein has been obtained simulating the three-dimensional structure with a MD package and excluding the atom restraints between the residue Gly13 and the residue Ser27 described in par. 2.2.1.3.The structural parts mostly involved in the denaturation are automatically identified comparing the spectra.

4. The last case is a comparison between the simulated spectrum of the natively folded HPr protein and the simulated spectrum obtained from the three-dimensional structure of the HPr protein artificially denatured as described in par. 2.2.1.4.

At the low-level the signals are automatically associated (see par. 3.4.2.3.2), independently on the available assignment of the spectra. Dealing with NOESY spectra involves much more comparisons than considering HSQC data thus increases the computational time.
Peak classes are defined through the Bayesian analysis (see par. 3.4.3) performed in the mid-level. In particular, not only single peaks, but entire patterns of residues can be identified as completely new, exploiting the symmetrical properties typically recognized in those types of spectra (see par. 3.4.4.2). A matching ratio between the signals in the reference and in each of the test case is computed, in order to determine the fraction of

unaltered spectrum (see eq. 3.31). The KS-test (see par. 3.4.3.5) provides a direct yes/no answer to evaluate if the compared spectra identify the same protein.

During the high-level step the reference residue patterns recognized as completely new, missing or ambiguous are automatically associated to any residue pattern of the analyzed test spectra. Their chemical shift positions are detected and showed to the user in the final dialog containing the global results of the comparison.

As described in the discussion section, the fraction of altered molecular structure is not actually computed (in NOESY and TOCSY cases) but it represents a further development of the project (see eq. 6.1 and eq. 6.2). The mapping of the peak features could be performed but since NOESY spectra typically contain many resonance peaks, the computation is time demanding and the visual inspection by means of histogram represents an hard task (the peak does not corresponds to a specific residue as in the HSQC case).

4.2 AUREMOL-QTA main interface

It is possible to start the developed AUREMOL-QTA routine from the “Calculation” menu as shown in Fig. 4.1. Once this has been done, the user selects the “Quality Test” submenu from where the computation can be started and the main interface appears as described in Fig. 4.2.

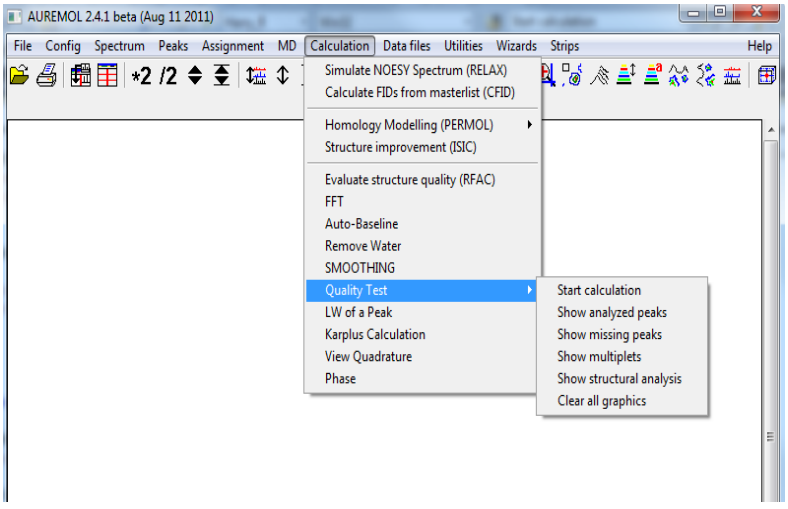


Fig. 4.1 Starting the AUREMOL-QTA module: the QTA module is called via the “Calculation” menu in the AUREMOL software package.

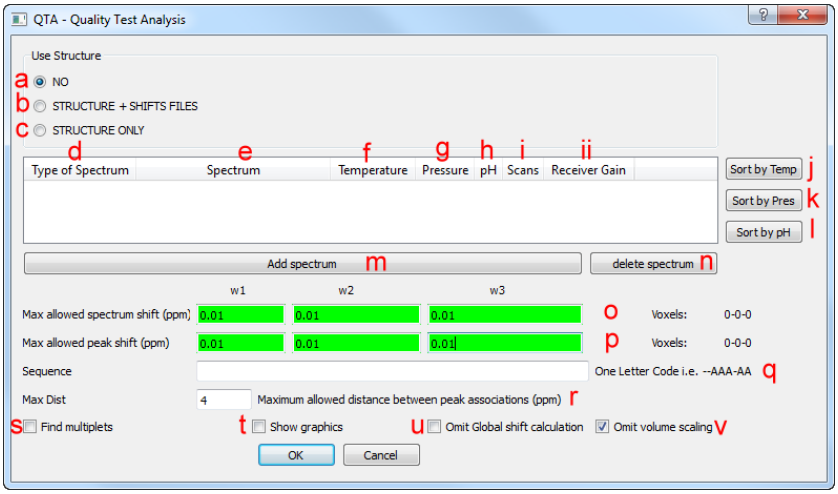


Fig. 4.2 Main interface of the AUREMOL-QTA: the dialog contains several parameters whose description is reported in the following text.

The AUREMOL-QTA main interface contains various customizable options. The option (a) is applied when the three-dimensional structure is not available. The routine uses only the information contained in the spectra thus the multiplet search analysis cannot be performed. The option (b) is checked when the user has the three-dimensional molecular structure of the protein and the list of chemical shifts. The additional simulated spectrum obtained through the RELAX-JT2 algorithm [Görler et al., 1997; Görler et al, 1999; Ried et al., 2004] is automatically included into the quality control calculation and used to recognize multiplet structures. This latter case would be the optimal one for a more accurate comparison. The option (c) is selected in case the user has only the structure (.pdb file) of the inspected molecule. In order to obtain the chemical shifts information the routine estimates them via a structure-based chemical shift prediction algorithm [Xu & Case, 2001; Neal et al., 2003]. The options (d - ii) are used to define the spectrum type (d), the spectrum file path (e), the temperature (f) of the measured sample, the pressure (g), the pH (h), the number of scans (i) and the receiver gain (ii) applied during the acquisition of the spectra. All these options are settable after having pushed the button “Add spectrum” (m option). In case that the user needs to remove one or more spectra from the Quality Test Analysis, the button “delete spectrum” has to be used (n option). The options (j - l) can be

used in order to sort the data according to the feature of interest. The option (o) defines the maximum allowed spectrum shift (described in par. 3.4.1.2) while the option (p) identifies the user defined local limit search (LSHIFT) of each peak (see par. 3.4.2.6). The text editor (q option) can be filled up in order to define the primary sequence of the investigated compound. It is possible to set the character "-" in case of an unknown residue. The total number of residues is shown and the routine controls the primary sequence nomenclature. The option (r) is used in order to set a warning messages threshold. A warning message pops up in case of assigned spectra where the distance (in ppm) between the connected peaks has been found larger than the threshold limit defined by the option (r). This option is particularly useful in case of strong conformational changes. The option (s) is activated by the user that needs to recognize peaks having a multiplet structure (see par. 3.4.2.2). The option (t) should be enabled if the user desires to follow the computation with graphical screenshots. In some cases the user does not want to compute the global spectrum shift either to save computational time or because no spectrum shifts are present. This can be obtained enabling the (u) option. Finally, if the user does not want to compute the volume scaling factor automatically (see par. 3.4.2.6.1) the option (v) should be checked. It is important to consider this latter case especially in presence of modified external conditions and ligand binding. In such cases, an automated volume scaling is not advisable. The volumes of all the peaks analyzed by the quality control are anyway normalized respect to the *NSCAN*, the *RG* and the *NC_proc* parameter (see par. 3.4.1.1).

4.3 HPr protein from *Staphylococcus aureus*

4.3.1 The wild and the mutant H15A HPr protein

The histidine-containing phosphocarrier protein (HPr) is essential for regulating the metabolism of carbohydrates in bacteria (Gram-positive bacteria) [Viana et al., 2000; Maurer et al., 2004]. This protein plays an essential role transferring a phosphate group to the carbohydrate transported through the membrane of the cell. In addition, it is an activator of gene repression. The tertiary structure of this protein has been solved by NMR

[Maurer et al., 2004]. The three-dimensional structure of the HPr protein from *Staphylococcus aureus* wild type and mutant (H15A) one are shown in Fig. 4.3.

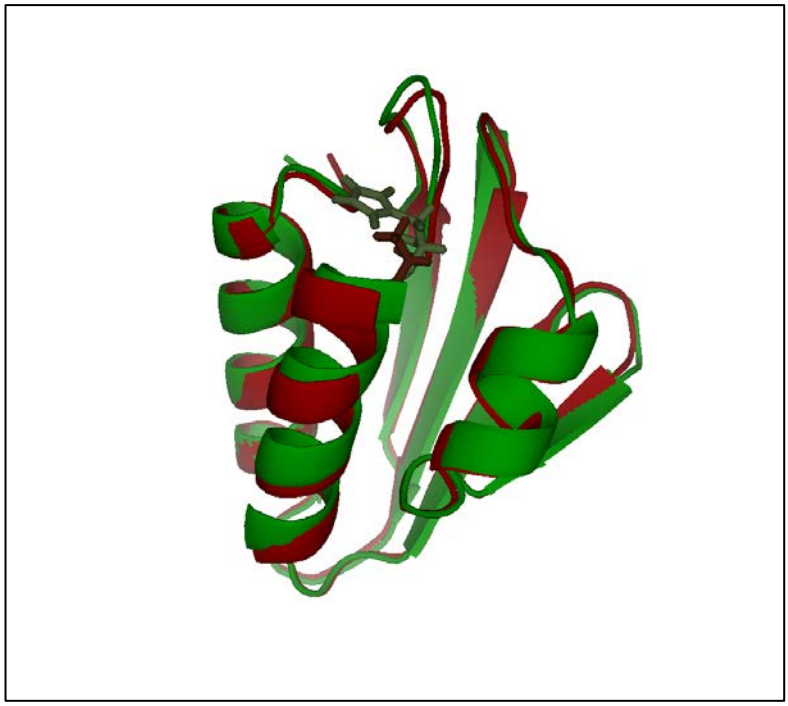


Fig. 4.3 The three-dimensional structure of the HPr protein from *Staphylococcus aureus*: the folded structure of the HPr wild type (green) and the H15A mutant one (red). The residue number fifteen has been stick-highlighted (green His15 and brown Ala15).

The experimental spectra of the wild and mutant HPr are shown in Fig. 4.4. They have been superimposed in order to visually identify the main spectral differences.

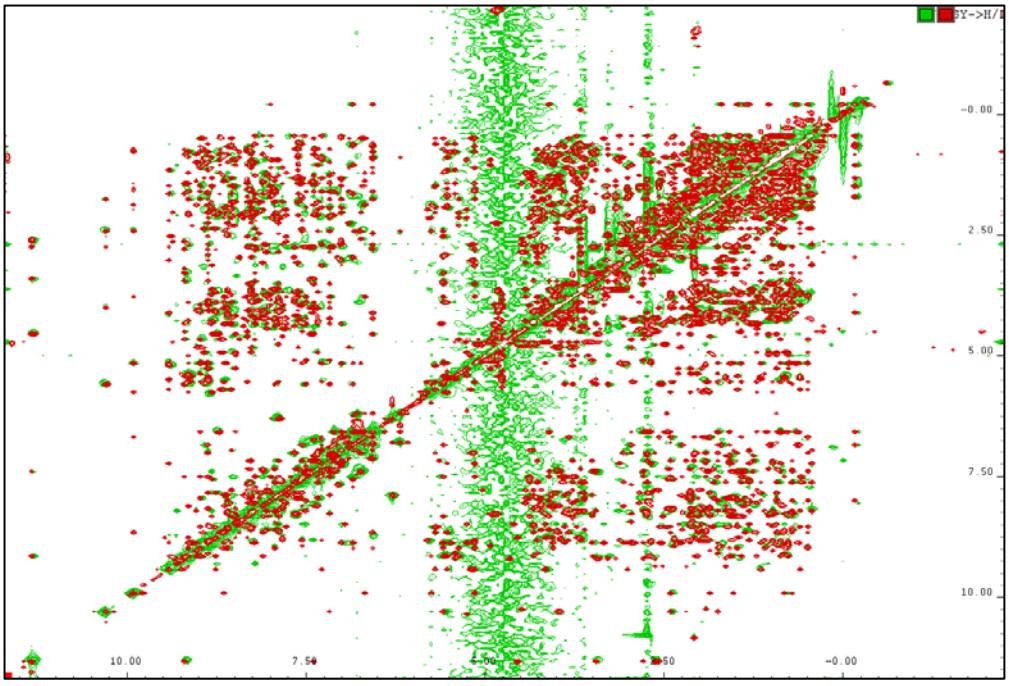


Fig. 4.4 The measured spectra of the HPr protein from *Staphylococcus aureus*: The spectrum of the HPr wild type (green signals) and the H15A mutant one (red signals).

4.3.1.1 The Quality control of the wild and mutant H15A measured spectra of the HPr protein from *Staphylococcus aureus*

A set of two spectra (obtained from the simulation of the three-dimensional structure of the native folded HPr protein as the reference spectrum and the slightly changed three-dimensional structure of the mutant HPr as the test case) has been loaded into the main interface, as shown in Fig. 4.5. The chemical shift assignments of the spectra were available but not used. The spectra have been measured with the same number of scans (NS) and receiver gains (RG) and with different NC_proc parameters, involving the data normalization (see eq. 3.2). The spectral width and the offset are not changed between the spectra. In case of spectral width differences a warning message appears informing the user that the computation can continue only if this variation is automatically adjusted by the routine (see Fig. 5.4 in par. 5.2.2).

The maximum allowed local shift (of every peak) option has been set to 0.04 ppm in both directions (corresponding to a shift of two and six voxels along the δ_1 and δ_2 direction respectively). A larger local shift may be introduced when HSQC data experiencing ligand binding or external condition variations are treated. A global shift (see par. 3.4.1.2) of 0.01 ppm (corresponding to a shift of one voxel along both directions) has been set by the user.

The volume scaling must not be omitted in the evaluated case since the goal of the comparison is not the feature mapping (as in the HSQC and the HSQC-TROSY datasets), but the conformational change detection.

In the quality control of a NOESY spectrum the primary sequence must not be provided. That is particularly useful when dealing with HSQC data.

The check box “Find multiplets” is particularly relevant when a back-calculated spectrum is available.

The “Max Dist” check box represents the maximal difference in ppm between the same assigned peaks in the compared spectra. This option can be used in order to warn about discordances between the assignments of the spectra. In case that the same peak name is found at a distance larger that the allowed one, a warning message appears containing the peak assignment in order to visually verify such difference in the spectra. The “Show graphic” check box is used if the user desires to observe the computational progress of the routine.

Several results (detailed and general) are produced by the routine as spectra with colored boxes around the peaks in dependence of the classification, matching ratios between the signals in the compared spectra, yes/no answer (if the investigated spectra are representing the same protein), pattern dislocation, matching fraction between the structures and the list of the most altered residues.

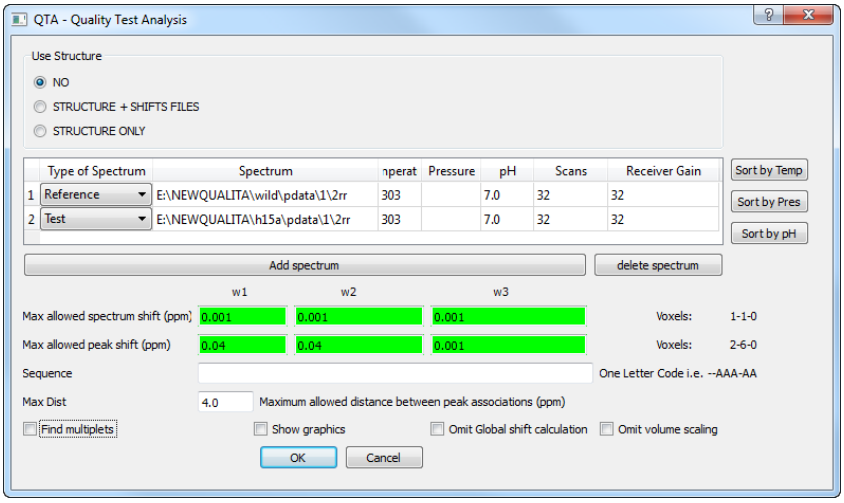


Fig. 4.5 Main interface of the quality test control on the HPr protein from *Staphylococcus aureus* (wild and mutant H15A): the main dialog shows two NOESY spectra and all the others user defined parameters as the type (reference or test) and the folder of the investigated data. The multi-dimensional global and the local shifts can be determined after a visual inspection of the spectra and they are automatically converted in terms of voxels.

4.3.1.2 Quality control detailed results

After completing the computation, the user can visually inspect the spectra peak by peak. A detailed comparison of any peak of interest can be obtained pushing the keyboard character “q”. In Fig. 4.6 some peaks have been investigated. It shows the associations of those peaks between the spectra automatically determined by the routine. If the user positions the mouse over the reference peak number 106 and pushes the keyboard character “q” the dialog reported in Fig. 4.7 appears. It contains the names of the associated peaks in the spectra, the shifts positions (in ppm) of every peak along both directions, the shift differences between the reference and the test peak, the volumes of the considered peaks before and after the normalization (eq. 3.2 and eq. 3.14) and the line width values in Hz of each peak in both direction. The probability of every peak association with respect to any of the five computed features (shape similarity, line width, cross-correlation, volume and chemical shift variation) is reported as well. If the

probability values are smaller than 0.5 they are highlighted in red, otherwise they are green.

As an example for the performance of the routine, in Fig. 4.7 the reference peak number 106 is automatically associated with the test peak HA2 54/HN 54. From the ppm difference is possible to detect a downfield shift of the peak in both directions in the test spectrum (it is recognizable from the negative signs of the chemical shift differences). The volume has been appropriately normalized and it results to be larger in the test than in the reference spectrum (according to the VOL_NEW parameter). The VOL OLD and the VOL NEW reveal the existing difference between the NC_proc parameters of the two experiments.

The Bayesian analysis has allowed providing high probabilities between the considered reference and the test peak for the shape, the correlation and the line width variations, while very low probabilities (red highlighted) have been obtained considering the volume and the shift.

In this case the MULTIPLET parameter shown in Fig. 4.7 is set to NO, demonstrating that it is not possible to analyze the peak multiplet structure due to the lack of the simulated spectrum.

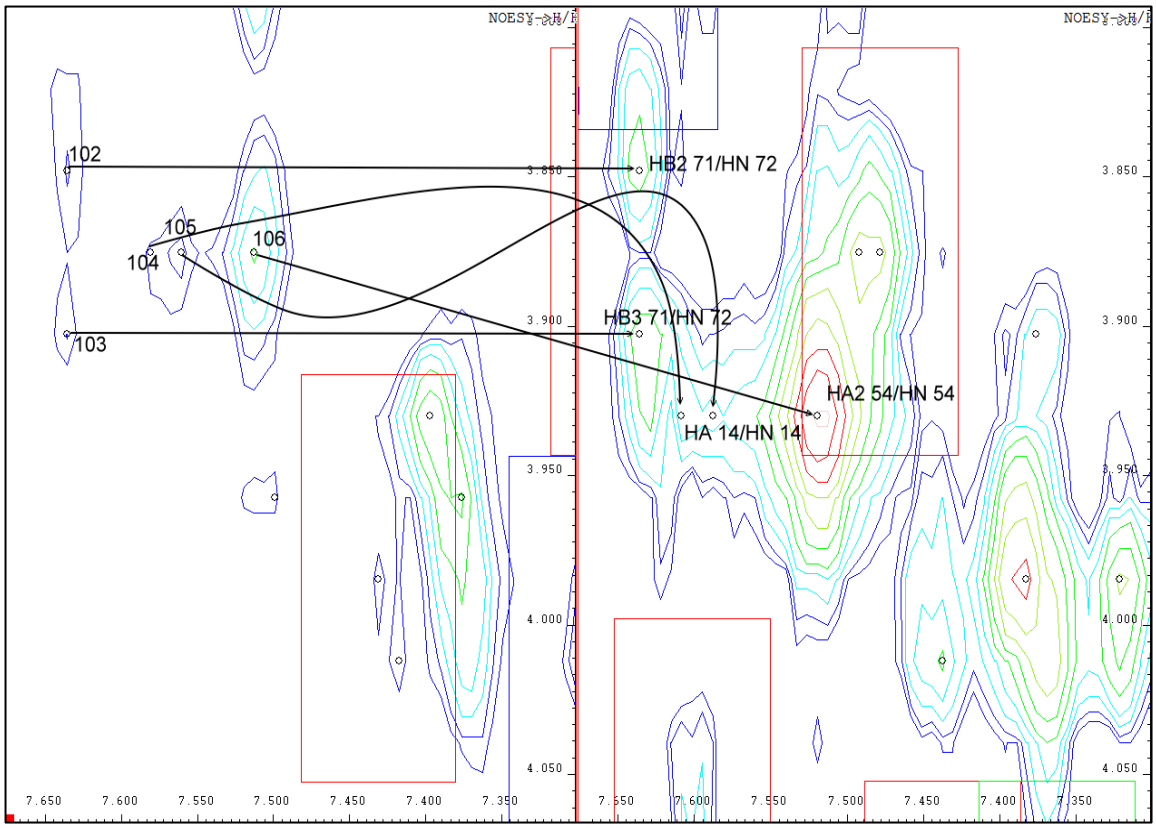


Fig. 4.6 Zoom of the HPr protein spectra: the peaks of the reference spectrum (HPr from *Staphylococcus aureus* wild type) in the right part, have been correctly assigned to the test signals (HPr from *Staphylococcus aureus* mutant H15A) in the left side.

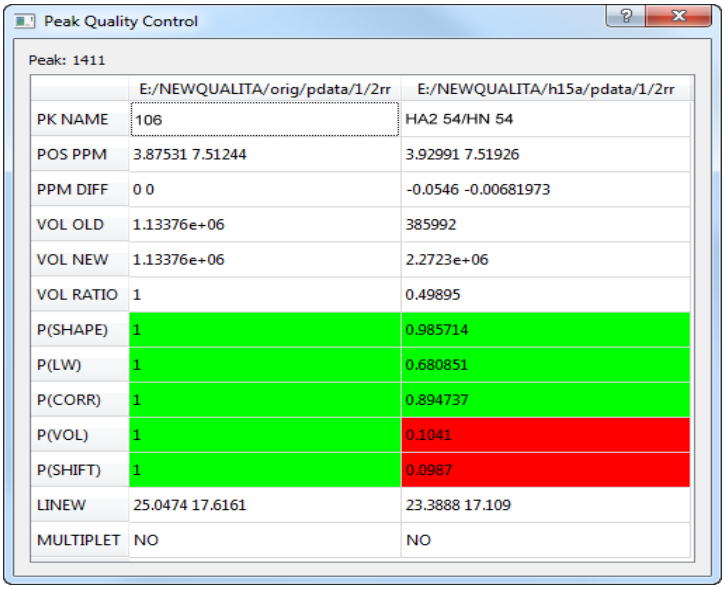


Fig. 4.7 The peak quality control dialog of the reference peak number 106 of the HPr protein from *Staphylococcus aureus* (wild type): the dialog shows in the first row the peak association between the spectra that has been automatically determined by the routine. In the second row of the dialog the ppm positions of each peak is reported. The third row represents the ppm differences among all the connected peaks. The row named VOL OLD represents the volume of each peak before applying the scaling factor. The row VOL NEW represents the volume after the normalization (due to the NC_proc). The row named VOL RATIO represents the volume ratio between VOL OLD and VOL NEW of each peak after the normalization. The rows from seven to eleven represent the values of the Bayesian analysis of each investigated feature (shape, line width, time cross-correlation, volume and peak position variation). The penultimate row shows the measured inhomogeneous line widths in Hz of each peak in both directions. The last row named MULTIPLET is yes defined if the peak has a multiplet structure.

4.3.1.3 Analysis of the peak features through the Bayesian probabilities

The “Show analyzed peaks” option in the “Quality test” menu (Fig. 4.1) opens a dialog where the user can select the desired feature of interest to be shown in the spectra with colored boxes in accordance to the selected probability and chosen color. The dialog is represented in Fig. 4.8, where the user can determine one or more features to be highlighted simultaneously with boxes of different colors having for instance a probability smaller than 0.5.

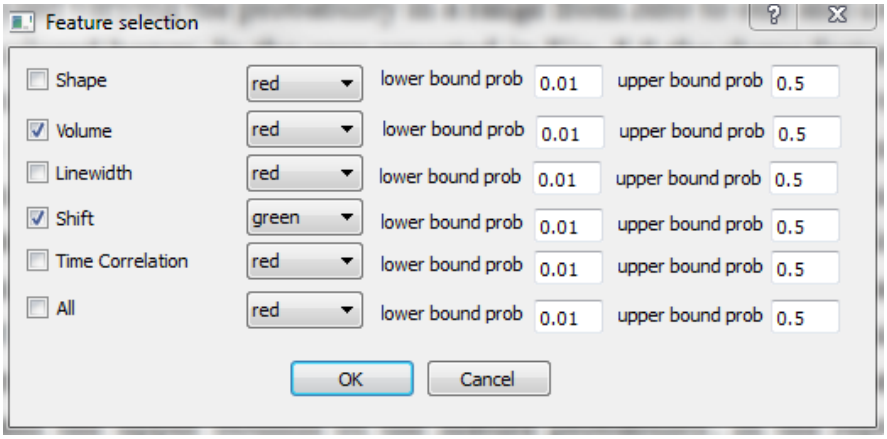


Fig. 4.8 The feature selection dialog: the user can select any of the desired features (shape similarity, volume, line width, chemical position variation, cross-correlation or all the features together) in the range defined by the upper and the lower probability bounds. In addition, the user can choose one of the seven available colors to highlight the feature of interest. In this case the volume and the shift of those peaks whose probability is minor than 0.5 have been selected to be highlighted with red and green boxes respectively.

If the features listed in Fig. 4.8 are all selected then different boxes appear in dependence of the feature verifying the defined probability, while if the option “All” is checked then the boxes are visible only around those peaks that verify the user determined probability simultaneously for all the features.

In the example shown in Fig. 4.9 the reference peaks surrounded by red boxes and by green boxes have been obtained in accordance to the parameters defined in the dialog reported in Fig. 4.8. Obviously some peaks have both volume and shift variations that can appear as a unique one.

Dealing with NOESY spectra implies a more complicated interpretation of the spectral signals than the HSQC and the HSQC-TROSY cases, where almost every peak corresponds to a residue. If the compared spectra have been previously assigned, the routine can generate a list containing the names of the peaks that are most altered, in accordance to the selected feature and probability, otherwise it contains only a list of ppm positions. Moreover, only if the NOESY spectra have been previously assigned it is possible to obtain a list containing the names of the residues that possess a most altered pattern in the spectra, otherwise only a list of ppm describing the possible patterns can be produced.

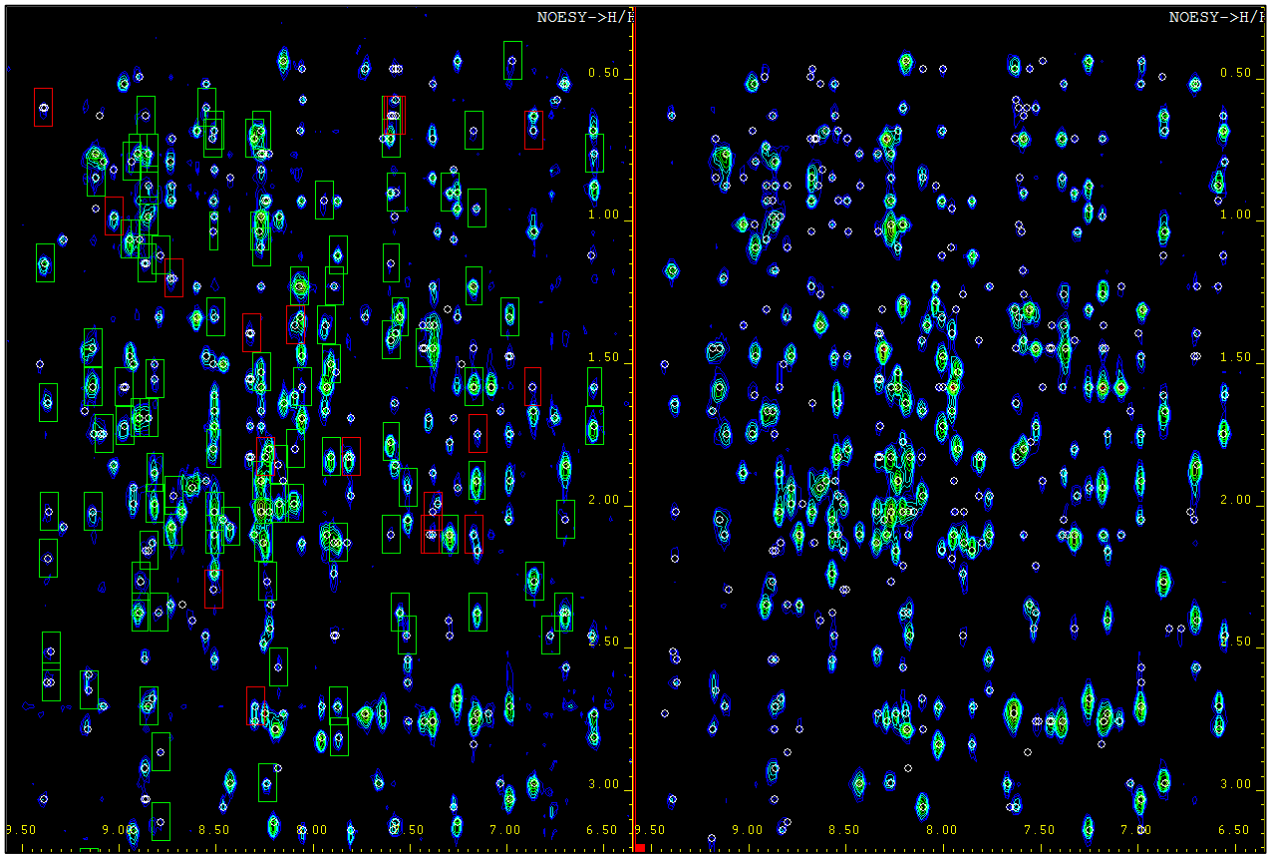


Fig. 4.9 Identification of the chemical shift and volume variations by means of the feature selection routine applied on the HPr protein from *Staphylococcus aureus:* the red- and green-colored boxes in the reference spectrum (wild type on the left side) identify peaks whose volume and shift probabilities are respectively smaller than 0.5 when compared to the test spectrum (H15A mutant on the right side).

4.3.1.4 Peaks that have not been associated between the spectra (new, missing and ambiguous signals)

After the AUREMOL-QTA computation, all the low-probability peaks are automatically classified by means of symmetrical analysis of the signals (see par. 3.4.4.2) and they are color-highlighted according to the following rules:

- A missing peak is highlighted with a red box
- A new peak is highlighted with a green box
- A false missing peak (ambiguous) is highlighted with a blue box

The color-highlighted classification, automatically obtained at the end of the quality computation performed on the wild and the mutant (H15A) HPr spectra is reported in Fig. 4.10. The classification reported in this figure has been performed in both directions, thus comparing the reference with the test spectrum and vice versa. Entire patterns of signals can be classified as new, missing or ambiguous, implying not only a local conformational change but a global one encompassing a specific residue. In Fig. 4.11 is shown the zoom of Fig. 4.10 demonstrating the capability of the routine to classify the signals in both directions. In Fig. 4.12 is instead zoomed out the lower part of the diagonals of both spectra, where several new, missing and ambiguous peaks have been detected.

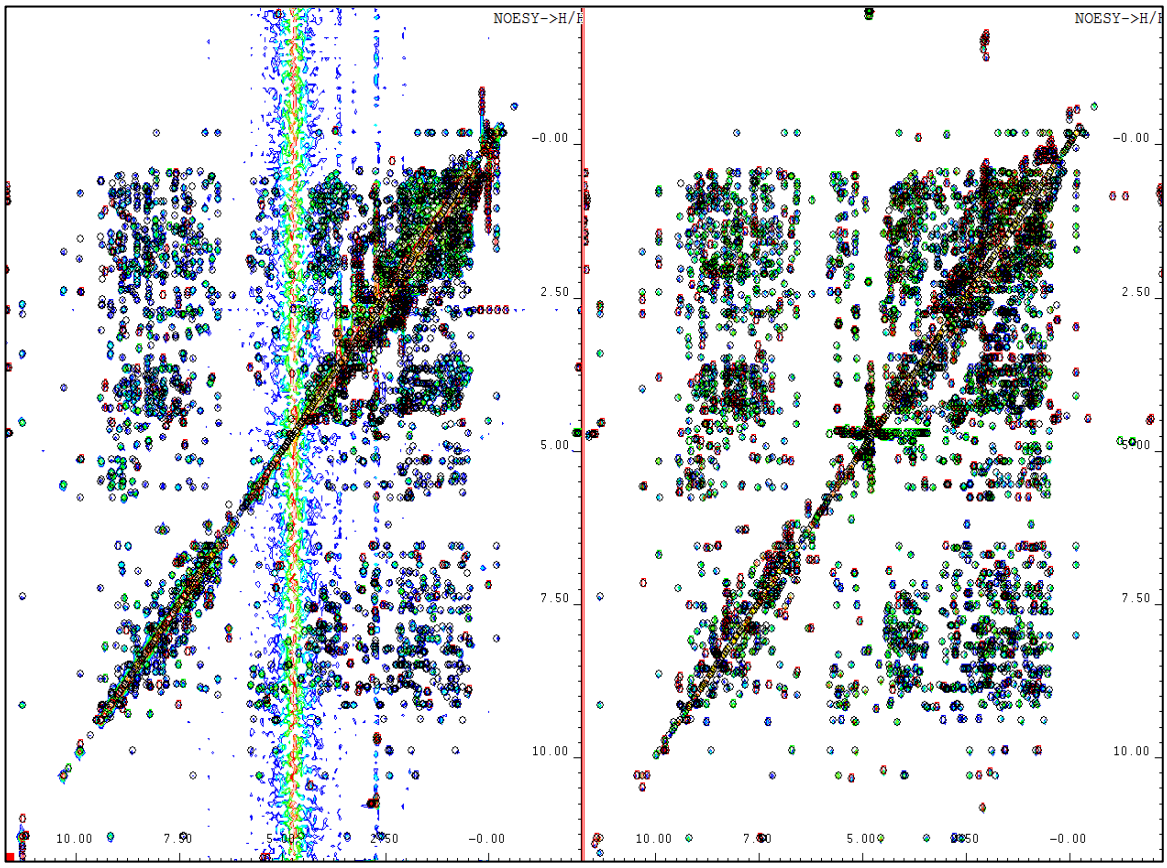


Fig. 4.10 The graphical result of the quality control when comparing experimental spectra of the wild and the mutant (H15A) HPr protein from *Staphylococcus aureus*: red, green and blue boxes have been automatically drawn by the quality control method in order to show new, missing and ambiguous peaks in the reference (wild type in the left side) and in the test (H15A mutant in the right side) spectrum.

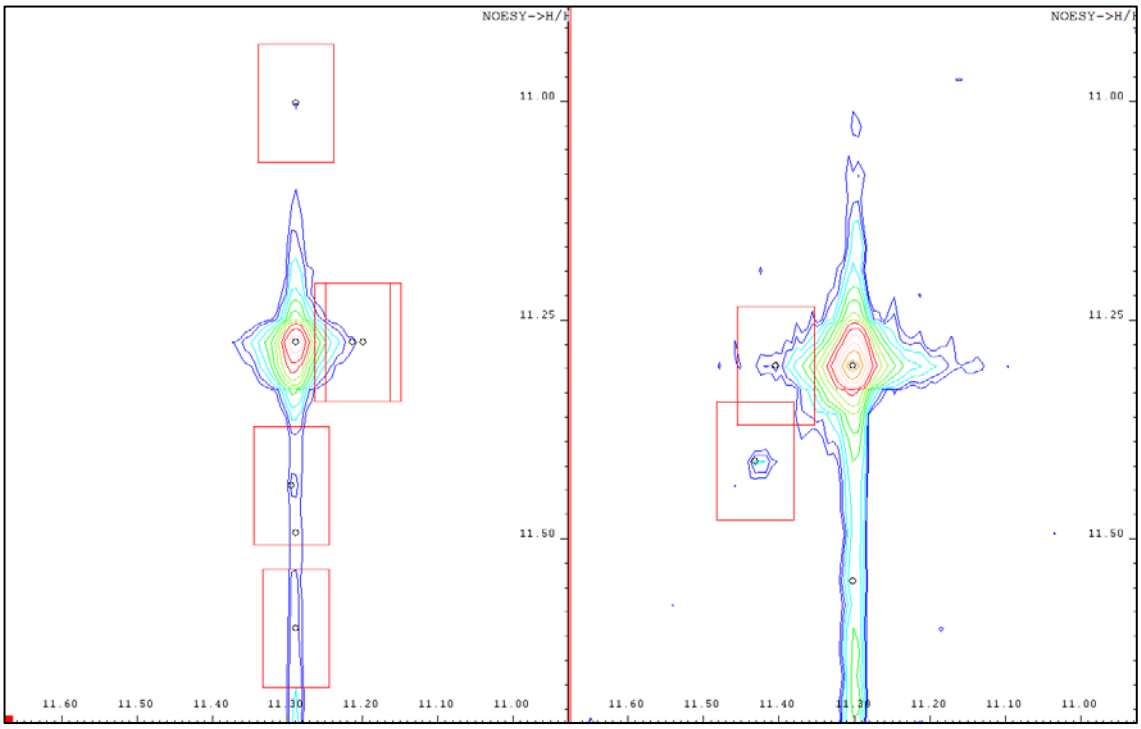


Fig. 4.11 Zoom of figure 4.10: red boxes have been automatically drawn by the quality control method in order to highlight missing peaks. The peak classification, thus the color of the box, depends on the global symmetrical properties of the signal (par. 3.4.4.2). The missing signals have been detected performing the comparison from the reference (wild type in the left side) to the test spectrum (H15A mutant in the right side) and vice versa.

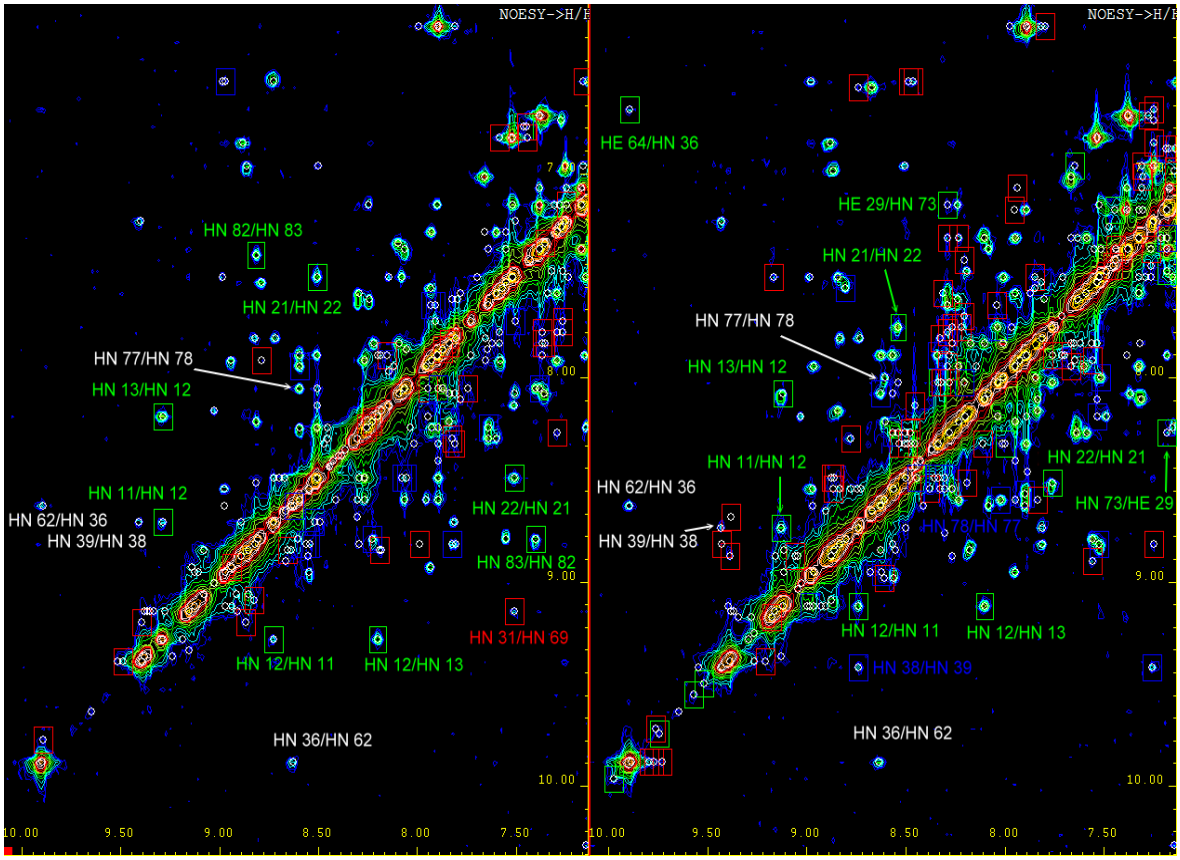


Fig. 4.12 Zoom of the lower part of the diagonal reported in figure 4.10: red, green and blue boxes have been automatically drawn by the quality control method in order to show new, missing and ambiguous peaks in the reference (wild type in the left side) and in the test spectrum (H15A mutant in the right side). The peak classification, thus the color of the box depends on the symmetrical properties of the signal (par. 3.4.4.2). See the following text for a more detailed explanation.

The green highlighted peaks in Fig. 4.12 represent those signals that have not been associated between the reference and the test spectra of the HPr protein from *Staphylococcus aureus*. For example, the reference peak HN 12/HN 11 and its symmetrical corresponding peak HN 11/HN 12 do not possess any association in the test spectrum (in the range of the user allowed local shift search), thus they can be identified as new signals. The red highlighted peaks are those signals that have not been found in the compared test spectrum. In particular, the reference missing peak HN 31/HN 69 does not possess any association in the test spectrum and in addition it has no symmetrical reference peak that

renders it suitable to be classified as a missing peak. The blue colored peaks identify those signals that have only a partial association in the compared test spectrum. For example, the test peak HN 39/HN 38 has been correctly associated in the upper part of the diagonal, but its symmetrical corresponding signal (HN 38/HN 39) has not been found in the reference spectrum. It could be erroneously classified as missing signal, whereas it represents a false missing or ambiguous case.

Entire patterns can be classified as new, missing or ambiguous ones involving a specific analysis of the pattern dislocation between the spectra (par. 3.4.4.2.1). In particular, the example reported in Fig. 4.13 shows the automated identification of the reference pattern of the residue Thr12 in a different ppm position in δ_2 direction. The reference pattern is constituted of 5 missing signals and 2 new peaks, thus it has been candidate to be entirely compared with all the new, missing and ambiguous test patterns. Six of the reference peaks lying along the Thr12 pattern have been recognized in the test spectrum. The association between these two specific patterns in the reference and in the test spectra has been identified according to eq. 3.32 described in par. 3.4.4.2.1. This pattern analysis is particularly favorable when a large shift variation has been occurred between the spectra.

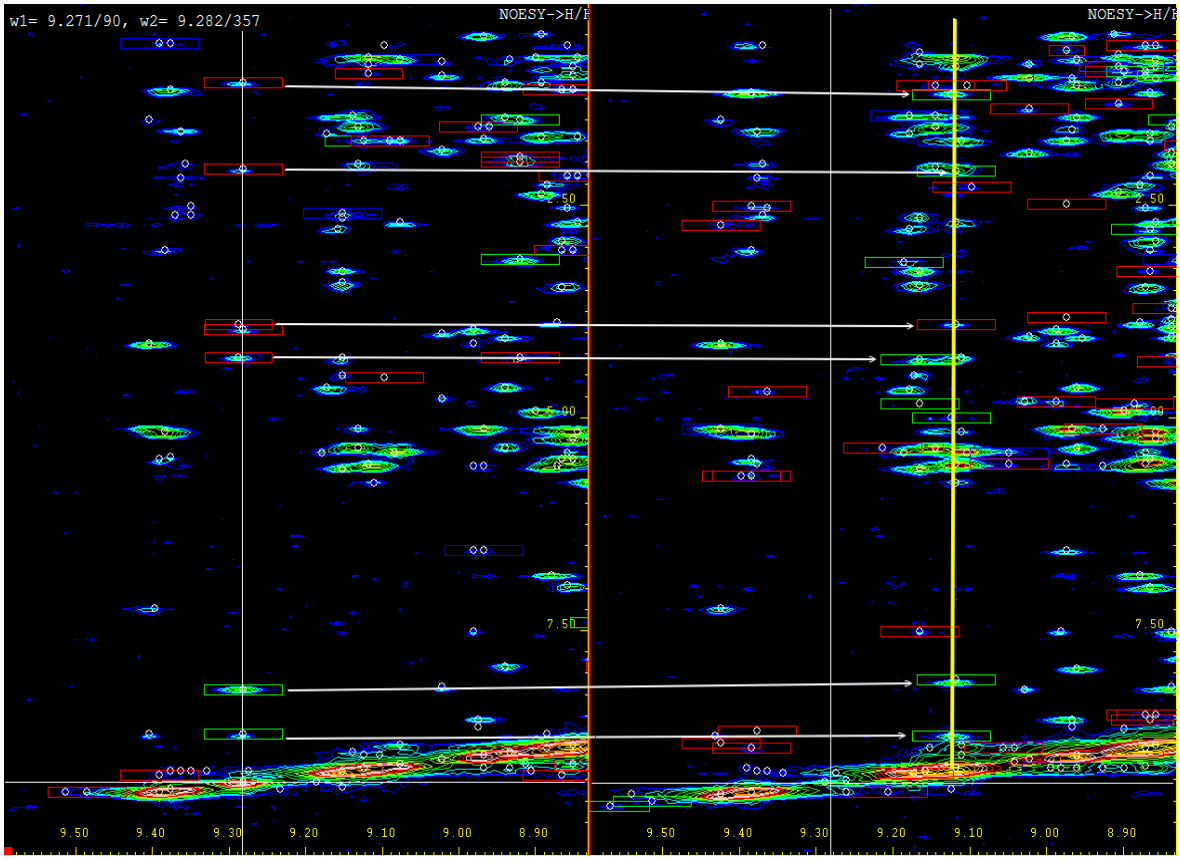


Fig. 4.13 Pattern recognition of the residue Thr12 of HPr protein from *Staphylococcus aureus*: the whole reference pattern (wild type in the left side) is constituted of new and missing peaks, thus it is used in the pattern recognition routine (par. 3.4.4.2.1). It has been automatically recognized at a different ppm position in the δ_2 direction in the test spectrum (H15A mutant in the right side) associating six of its seven peaks with the corresponding test signals.

4.3.1.5 General results of the quality control

General results are obtained in a final main window that contains the number of picked peaks, of missing signals, of ambiguous (false missing) peaks and of new signals in every investigated spectrum. The matching ratio between the high probability peaks in the compared spectra is reported in the P(FOUND) parameter. The Kolmogorov-Smirnov test is applied separately with respect to three features (peak volume, chemical shift and shape variation) in order to determine whether the compared spectra are different measurements

of the same protein. The analysis of the residue patterns is also encompassed in this final window, where a list of ppm positions is displayed with the aim to define the specific pattern dislocation among the compared spectra. In particular, considering the comparison of the wild type of the HPr protein with the mutant H15A (see Fig. 4.14), the method has automatically detected 3187 and 3730 peaks respectively, by means of the AUREMOL peak picking [Antz et al., 1995] routine. Moreover, 1373 and 821 peaks in the test and in the reference spectra are classified as missing (thus, 821 reference peaks has not been encountered in the test H15A spectrum), 329 and 178 resonances have been recognized as ambiguous (par. 3.4.4.2), while 285 and 126 are completely new signals (par. 3.4.4.2). All the peaks in both spectra have been segmented through the AUREMOL integration routine [Geyer et al., 1995]. The computation has been performed on a 64-bit Windows system with 4 Gb of system memory using an Intel Core2 Quad CPU Q9550 (four cores) with a single core clock speed of 2.83 GHz. The computational time of the whole AUREMOL-QTA procedure when analyzing the previously described data was 387 s.

In the (a) part and in the (b) part of Fig. 4.14 are reported respectively the results of the comparison performed starting from the reference (wild HPr) to the test (H15A) and vice versa. The matching ratios between the high probability peaks of both spectra are 79% starting the comparison from the reference and 72% starting from the test. Accordingly to the KS-test applied separately on the volume, the shape and the position score distributions, the spectra have not been obtained from the same conformation of the analyzed compound. The reference patterns of the Thr12 at 9.28 ppm and of the Gly56 at 7.72 ppm have been automatically recognized in the test spectrum at 9.11 and 7.62 ppm respectively.

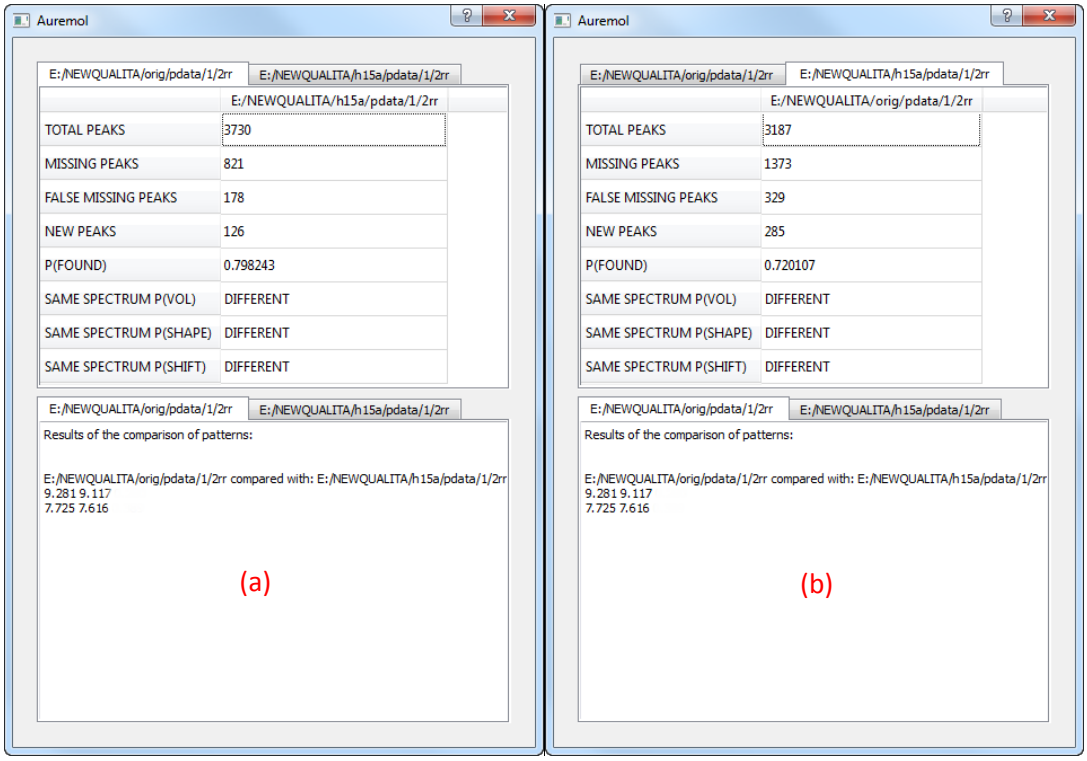


Fig. 4.14 General results of the comparison between the wild HPr protein and the mutant H15A: in the former and in the latter 3187 and 3730 peaks have picked, 821 and 1373 signals have been classified as missing (thus, 821 picked reference peaks has not been encountered in the test H15A spectrum), 178 and 329 resonances have been recognized as ambiguous (par. 3.4.4.2), while 126 and 285 are completely new signals (par. 3.4.4.2). In addition, the spectra are not recognized as the same for all the reported features (volume, shape and peak position change). The probability that the spectra are the same (based on the number of high probability peaks) is 72% and 79% performing the comparison from the test to the reference spectrum and vice versa. Two residue patterns have been detected with a large dislocation: Thr12 (found at 9.281 ppm and 9.117 ppm in the reference and in the test spectrum respectively) and Gly56 (found at 7.725 ppm and 7.616 ppm in the reference and in the test spectrum respectively) as listed in the lower part of the dialog (in the δ_2 direction).

4.3.2 The Quality control using the wild, the measured mutant (H15A) and the simulated spectrum (wild type) of the HPr protein from *Staphylococcus aureus*

The measured spectra of the wild, the mutant and the simulated (wild type) HPr protein from *Staphylococcus aureus* are shown in Fig. 4.15. The simulated spectrum of the HPr protein has been obtained from the known chemical shifts and the available three-dimensional structure. They have been superimposed in order to highlight the main spectral differences.

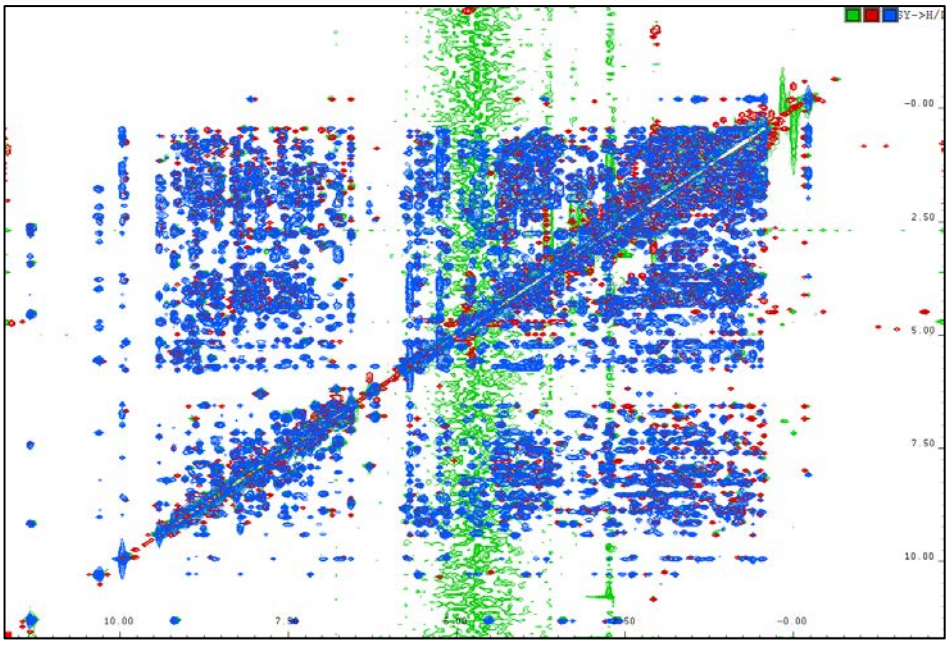


Fig. 4.15 The measured and the simulated spectra of the HPr protein from *Staphylococcus aureus* superimposed: the experimental spectrum of the HPr wild type (green signals), the H15A mutant (red signals) and the simulated spectrum of the wild HPr (blue spectrum).

4.3.2.1 The quality control of the wild-mutant-simulated spectra of HPr

A set of two spectra (the wild HPr as the reference spectrum and the mutant HPr as the test case) has been loaded into the main interface, as shown in Fig. 4.16 (see par. 4.3.1.1).

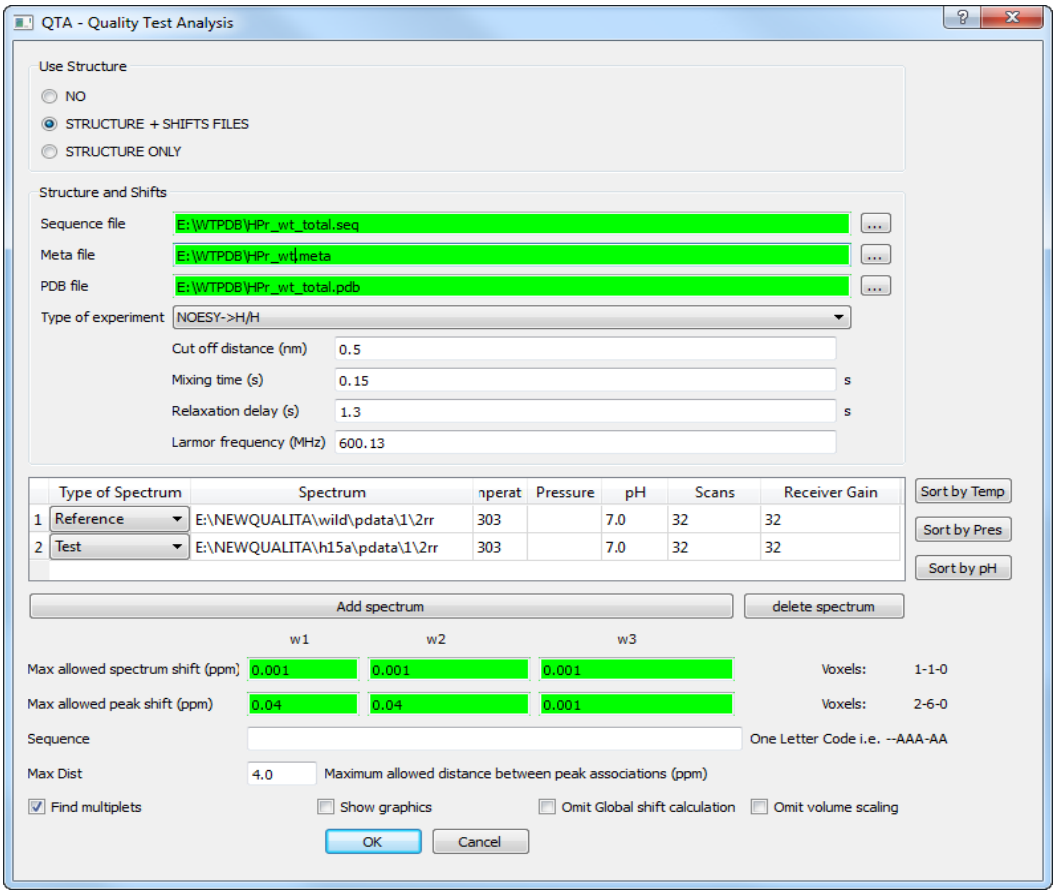


Fig. 4.16 Main interface of the quality control analysis applied on the HPr protein from *Staphylococcus aureus:* the user provides the parameters (temperature, pH, number of scans and receiver gain) and defines the reference and the test spectra (the experimental wild and mutant HPr). The simulated spectrum of the wild HPr is automatically back-calculated checking the "STRUCTURE + SHIFT FILE" radio button and inserting user provided files (the protein sequence file, the files containing the atom chemical shifts and the protein structure) and additional parameters (cut off distance, mixing time, relaxation delay and Larmor frequency).

The simulated spectrum of the HPr protein from *Staphylococcus aureus*, if not provided together with the measured spectra (reference and test case), is automatically computed by the routine. In this case, some parameters (cut off distance, mixing time, relaxation delay and Larmor frequency) must be specified by the user in the main dialog (see Fig. 4.16). The user must also provide a file containing the primary sequence, a file containing the chemical shift of some previously assigned atoms and the .pdb file (containing the three-dimensional coordinates of the protein structure). Once the user has provided all the requested parameters and has clicked the OK button the routine computes the simulated spectrum. This spectrum is then successively used as a simulated reference spectrum in order to detect peaks with a multiplet structure (see par. 3.4.2.2). Examples of recognized peak multiplets are shown in Fig. 4.17 and Fig. 4.18.

The peak multiplet HD 53/NH 42 (in Fig. 4.17) of the simulated spectrum has been automatically associated with the peak 3241 (belonging to the measured reference spectrum of the HPr protein wild type). The same peak (HD 53/NH 42) has not been recognized in the spectrum of the mutant HPr.

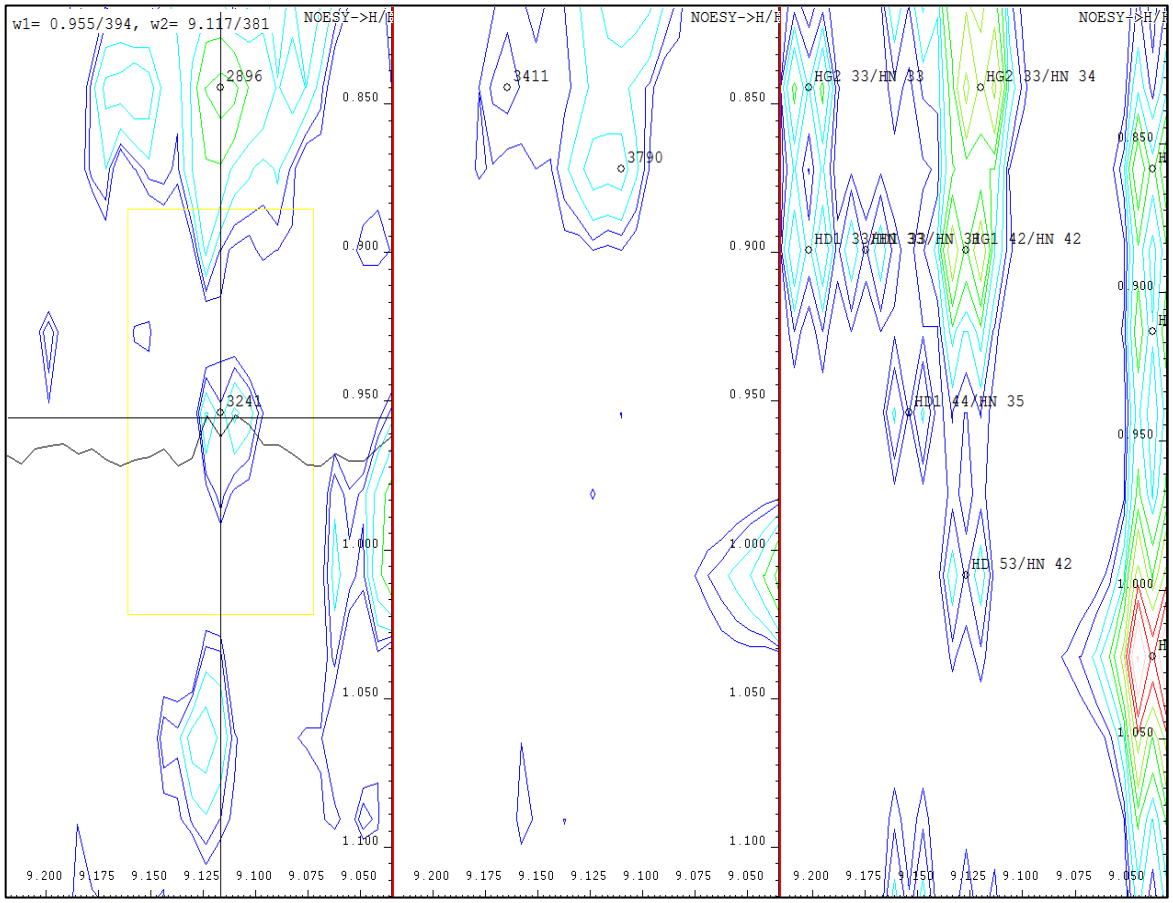


Fig. 4.17 Recognition of a peak multiplet in the experimental spectrum (wild type) of the HPr protein: the simulated spectrum (right side) of the wild HPr protein from *Staphylococcus aureus* has been used in order to identify and associate the multiplet peak HD 53/NH 42 with the peak 3241 (belonging to the measured spectrum of the HPr protein wild type in the left side). It is not possible to recognize the same peak (HD 53/NH 42) in the spectrum of the mutant HPr (center). The peak multiplet has been automatically highlighted by the routine with a yellow box. In addition, the center of the peak multiplet has been correctly positioned by the routine.

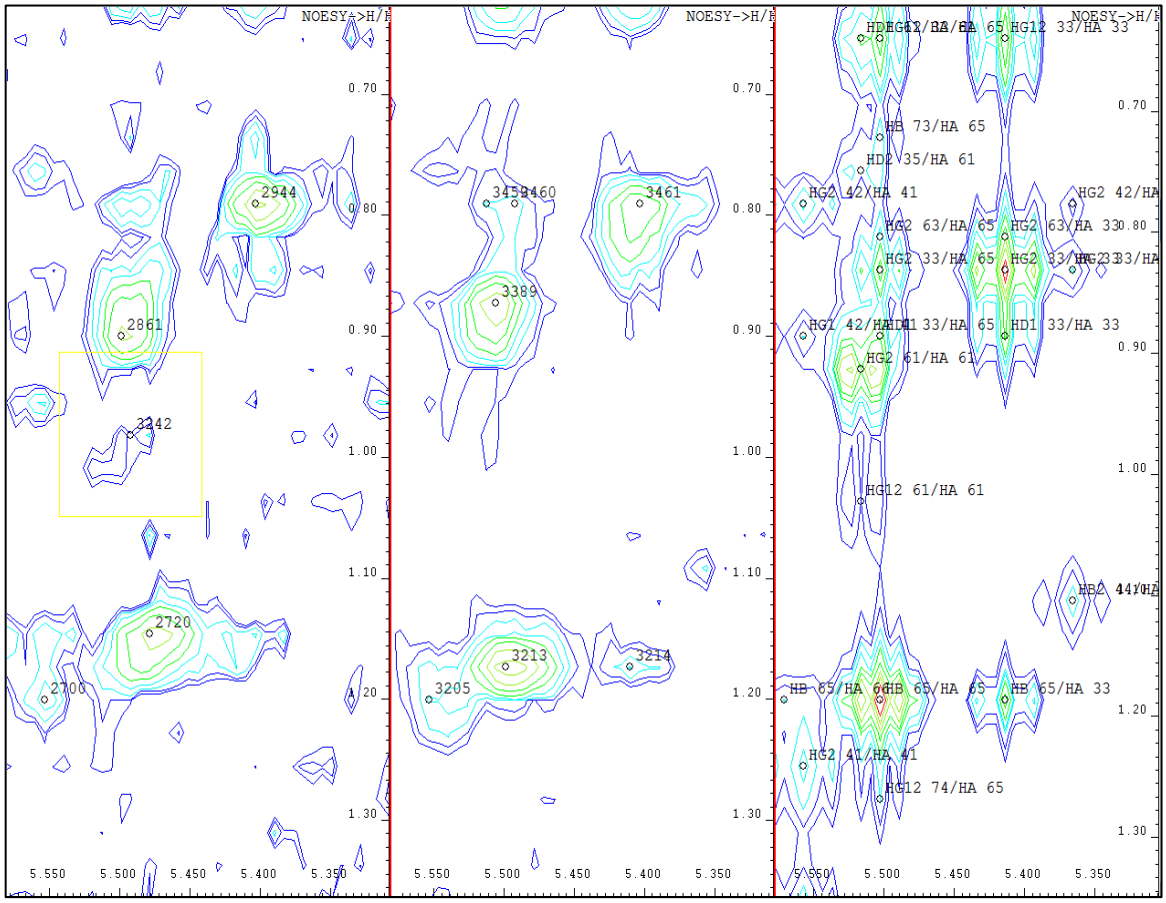


Fig. 4.18 Recognition of a peak multiplet in the experimental spectrum (wild type) of the HPr protein: the simulated spectrum (right side) of the wild HPr protein from *Staphylococcus aureus* has been used in order to identify and associate the multiplet peak HG12 61/HA 61 with the peak 3242 (belonging to the measured spectrum of the HPr protein wild type in the left side). It is not possible to recognize the same peak (HG12 61/HA 61) in the spectrum of the mutant HPr (center). The peak multiplet has been automatically highlighted by the routine with a yellow box. In addition, the center of the peak multiplet has been found correctly by the routine.

The peak multiplet HG12 61/HA 61 (in Fig. 4.18) of the simulated spectrum has been automatically associated with the peak 3242 (belonging to the measured reference spectrum of the HPr protein wild type). The same peak (HG12 61/HA 61) has not been recognized in the spectrum of the mutant HPr.

4.3.3 Recognition of partial denaturation of a protein

In order to investigate structural changes of the HPr protein from *Staphylococcus aureus* (wild type) under critical conditions, a partial denaturation (residue range of Gly13-Ser27) has been simulated by restrained molecular dynamics (see par. 2.2.1.3) as shown in Fig. 4.19.

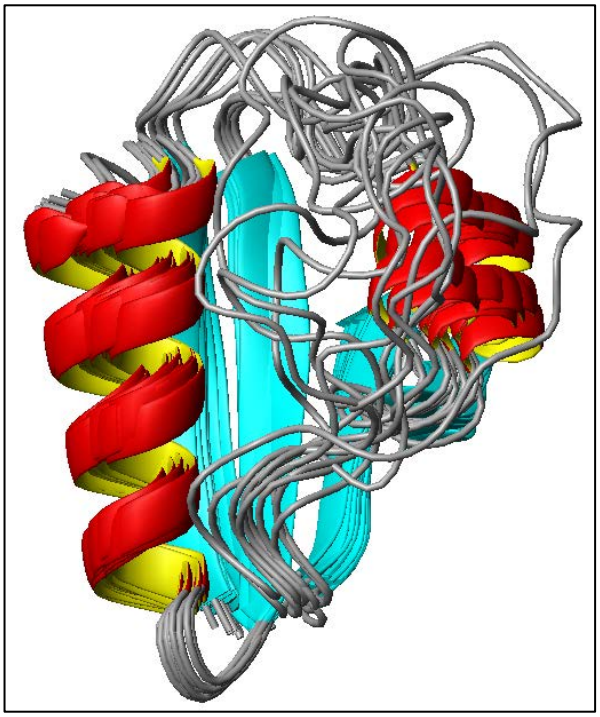


Fig. 4.19 The partially denatured HPr protein from *Staphylococcus aureus*: the residue range between Gly13 and Ser27 has been denatured. The chemical shifts of the involved residues are comparable with random coil chemical shifts.

The simulated spectra (obtained from the known three-dimensional structure and the available chemical shifts) of the folded and the partially denatured HPr protein from *Staphylococcus aureus* (wild type) are shown in Fig. 4.20. They have been superimposed in order to highlight the main spectral differences.

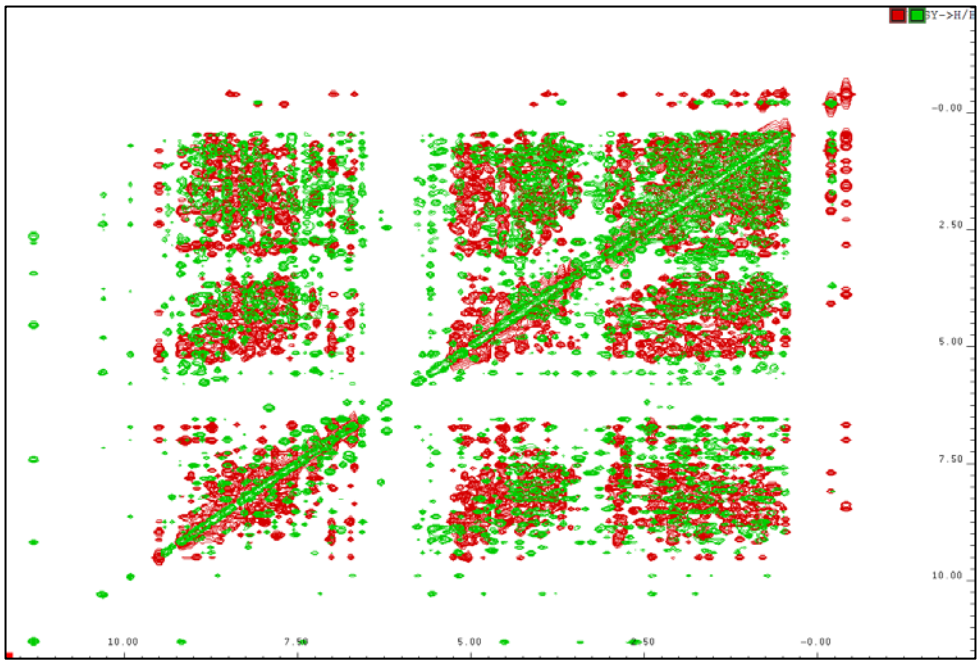


Fig. 4.20 The folded and the partially denatured spectra of the HPr protein from *Staphylococcus aureus* superimposed: the spectrum of the folded HPr protein wild type (green signals) and the spectrum of the partially denatured HPr protein (red spectrum).

The partial denaturation has been computed in order to analyze the reliability of the AUREMOL-QTA routine under extreme situations. Changes of the external conditions (e.g. temperature, pressure and pH) can lead to a denaturation of the investigated protein.

4.3.3.1 The quality control of the folded and the partially denatured spectra of the wild HPr protein from *Staphylococcus aureus*

A set of two spectra (the simulated folded HPr protein as the reference spectrum described in par. 2.2.1.1 and the partially denatured HPr described in par. 2.2.1.3 as the test case) has been loaded into the main interface (see Fig. 4.5). Both spectra were previously assigned. They have the same number of scans (NS), the same receiver gains (RG) and the same NC_proc parameters. No data normalization (see eq. 3.2) is needed. The SW (spectral width) and the OFF (spectrum offset) are identical in both cases. The maximum allowed

local shift (of every peak) option has been set to 0.08 ppm in both directions (corresponding to a shift of three and twelve voxels along the δ_1 and δ_2 direction respectively). This larger local shift (larger with respect to the comparison of the wild and the mutant HPr) has been introduced in order to test the routine under such difficult condition. A global shift (see par. 3.4.1.2) of 0.01 ppm (corresponding to a shift of one voxel along both directions) has been set by the user.

The volume scaling has been omitted. The test has been conducted enabling the multiplet recognition. The "Max Dist" has been set to 4.0 ppm (option "r" in Fig. 4.2).

Several results (detailed and general) are produced by the routine as spectra with colored boxes around the peaks in dependence of the classification, matching ratios between the signals in the compared spectra, yes/no answer (if the investigated spectra are representing the same protein), pattern dislocation, matching fraction between the structures and the list of the most altered residues. In particular, as shown in Fig. 4.21 the routine is able to detect many differences between the investigated spectra.

The folded HPr (reference spectrum) has been compared with the partially denatured spectrum of the same protein (test spectrum). In the row "TOTAL PEAKS" is reported the number of peaks (4596) belonging to the investigated test spectrum (the total number of reference peaks is 8447). The row "MISSING PEAKS" contains the number of reference peaks (2119) that have been not found in the test spectrum. 100 reference peaks have been considered ambiguous (see par. 3.4.4.2) while 3735 have been classified as new peaks. A matching ratio of 30.69% has been automatically computed.

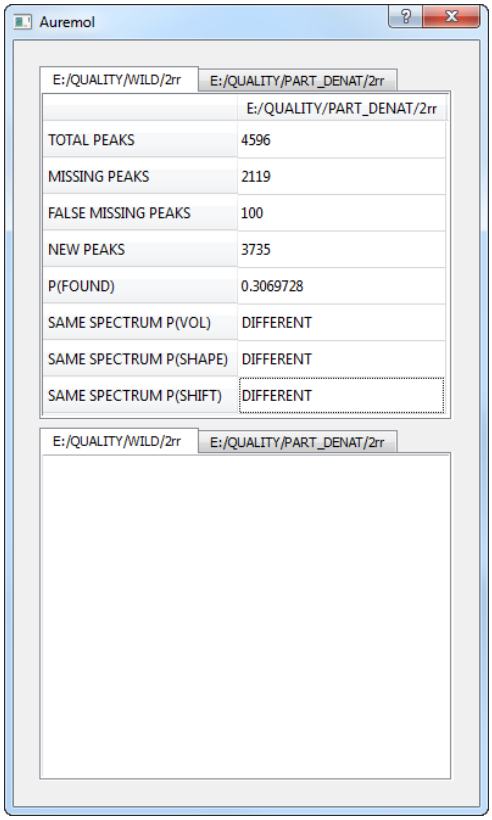


Fig. 4.21 General results of the comparison between the simulated folded HPr protein (wild type) and the partially denatured one (wild type): the folded HPr (reference spectrum) has been confronted with the partially denatured spectrum of the same protein (test spectrum). In the row "TOTAL PEAKS" is reported the number of peaks (4596) belonging to the investigated test spectrum (the total number of reference peaks is 8447). The row "MISSING PEAKS" contains the number of reference peaks (2119) that have been not found in the test spectrum. 100 reference peaks have been considered ambiguous (see par. 3.4.4.2) while 3735 have been classified as new peaks. The row P(FOUND) reveals that a probability of 30.69 % of matching signals has been found. The KS-test reported in the last three rows indicates that it is not the same protein by means of volume, shape and chemical shift variations.

4.3.4 The completely denatured HPr protein from *Staphylococcus aureus*

A full denaturation of the HPr protein has been analyzed (see par. 2.2.1.4). As shown in Fig. 4.22, the HPr protein from *Staphylococcus aureus* (with a characteristic extended strand structure) has been analyzed.

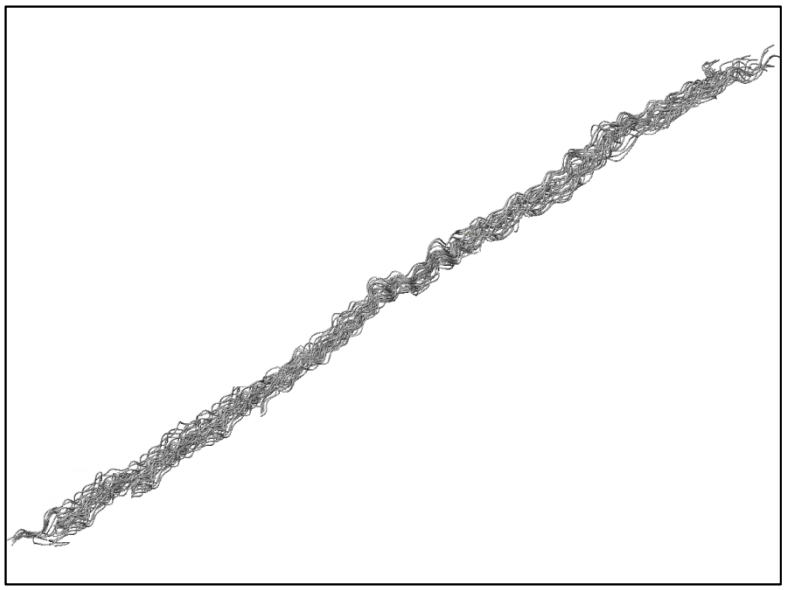


Fig. 4.22 The fully denatured HPr protein (wild type) from *Staphylococcus aureus*: The extended strand structure of the HPr wild type after the denaturation.

The simulated spectra of the folded and the fully denatured HPr protein (wild type) from *Staphylococcus aureus* are shown in Fig. 4.23. They have been superimposed in order to highlight the main spectral differences.

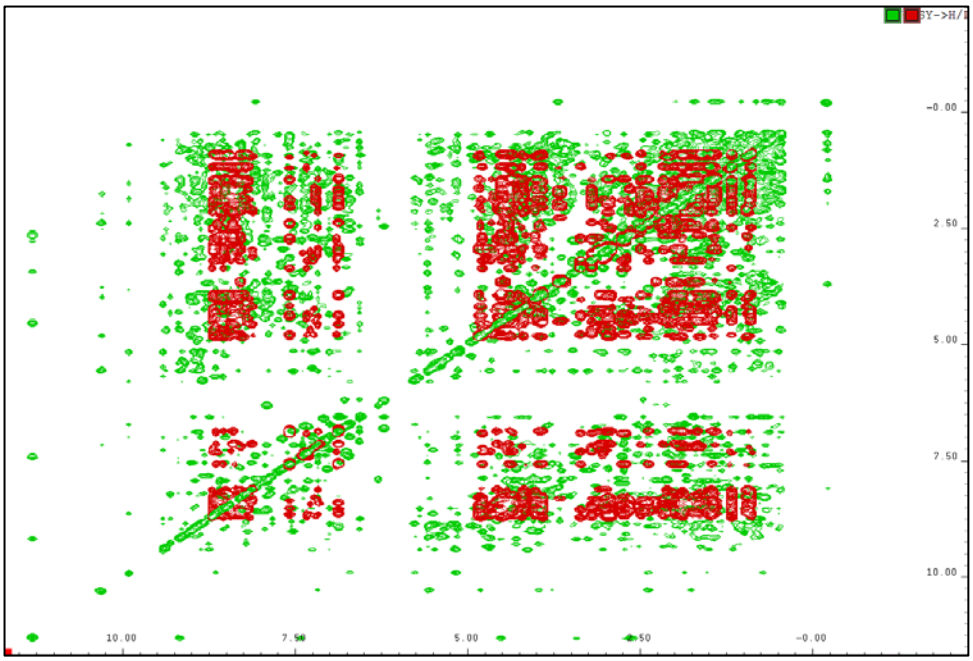


Fig. 4.23 The folded and the totally denatured spectra of the HPr protein from *Staphylococcus aureus* superimposed: the spectrum of the folded HPr protein wild type (green signals) and the spectrum of the fully denatured HPr protein (red spectrum).

4.3.4.1 The quality control of the native and the fully denatured spectra of the wild HPr protein from *Staphylococcus aureus*

A set of two spectra (obtained from the back-calculation of the native folded HPr protein as the reference spectrum described in par. 2.2.1.1 and the spectrum back-calculated from the totally denatured three-dimensional structure of the HPr protein described in par. 2.2.1.4 as the test case) has been loaded into the main interface (see Fig. 4.5). The chemical shifts of both spectra have been used. They have the same number of scans (NS), the same receiver gains (RG) and the same NC_proc parameters. No data normalization (see eq. 3.2) is needed. The SW (spectral width) and the OFF (spectrum offset) are identical in both cases. The maximum allowed local shift (of every peak) option has been set to 0.08 ppm in both directions (corresponding to a shift of three and twelve voxels along the δ_1 and δ_2 direction respectively). This larger local shift (larger with respect to the comparison of the wild and the mutant HPr) has been introduced in order to test the routine under such

difficult condition. A global shift (see par. 3.4.1.2) of 0.01 ppm (corresponding to a shift of one voxel along both directions) has been set by the user.

The volume scaling has been omitted. The test has been conducted enabling the multiplet recognition. The “Max Dist” has been set to 6.0 ppm (option “r” in Fig. 4.2).

As shown in Fig. 4.24, the routine is able to detect many differences between the investigated spectra.

The native folded HPr (reference spectrum) has been compared with the fully denatured spectrum of the same protein (test spectrum). In the row “TOTAL PEAKS” is reported the number of peaks (5330) belonging to the investigated test spectrum (the total number of reference peaks is 8447). The row “MISSING PEAKS” contains the number of reference peaks (3485) that have been not found in the test spectrum. 128 reference peaks have been considered ambiguous (see par. 3.4.4.2) while 3673 have been classified as new peaks. A matching ratio of 15.25 % has been detected.

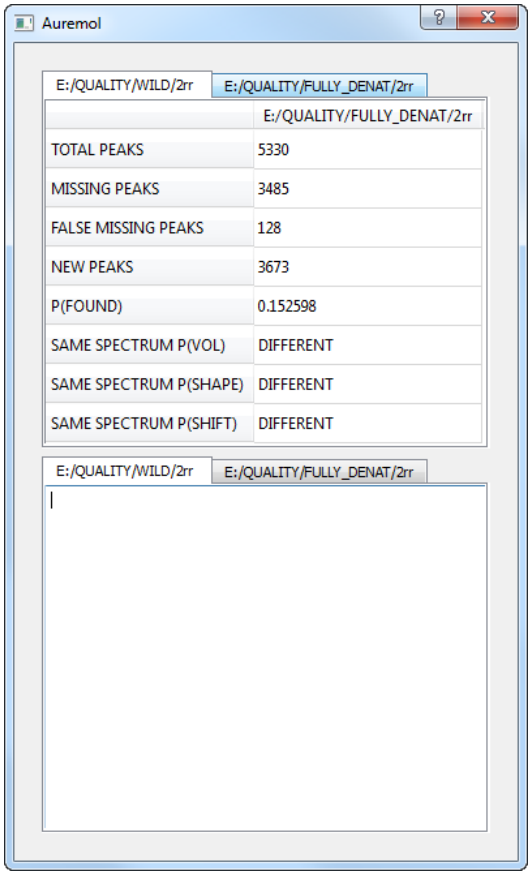


Fig. 4.24 General results of the comparison between the simulated folded HPr protein (wild type) and the totally denatured one (wild type): the folded HPr (reference spectrum) has been confronted with the fully denatured spectrum of the same protein (test spectrum). In the row "TOTAL PEAKS" is reported the number of peaks (5330) belonging to the investigated test spectrum (the total number of reference peaks is 8447). The row "MISSING PEAKS" contains the number of reference peaks (3485) that have been not found in the test spectrum. 128 reference peaks have been considered ambiguous (see par. 3.4.4.2) while 3673 have been classified as new peaks. The row P(FOUND) reveals that a probability of 15.25% of matching signals has been found. The KS-test reported in the last three rows indicates that it is not the same protein by means of volume, shape and chemical shift variations.

5

Test case: human prion protein

This chapter shows the practical application of the previously described methods (see chapter 3) to the human Prion protein *hu*PrP^{C} (see par. 2.2.2.2 and par. 2.2.2.3). The prion has been investigated in two particular cases: ligand binding with xenon and high pressure. In particular, the AUREMOL-QTA routine is used to test its ability in ligand screening and in conformational change identification due to variations of external conditions (as the pressure and the temperature).

5.1 Introduction

All the pre-processing steps are automatically managed by the AUREMOL-QTA module. The user provides all the spectra of interest personally deciding which ones must be considered as test and reference cases and in which order (in accordance with the altered external condition of interest as temperature, pressure or pH). Some parameters (SW, OFF and NC_proc) are automatically obtained by the routine from the processing files, whereas some others (that are stored in the acquisition file) are provided directly by the user filling out the main graphical interface (see Fig. 4.2). These are the number of scans, the maximum allowed global shift of the spectra and the receiver gain (RG).
The low-level management, that includes the collection of several peak features (volume, position, line width and shape) and the associations of peaks among the considered spectra, is totally automated. The local shift limit can be adjusted directly by the user after a first visual inspection of the dataset.

The Bayesian analysis performed during the mid-level stage over any user selected feature allows the visualization of peak classes in different colors and the consequent computation of matching and mismatching ratios between signals in the spectra. Moreover,

the Kolmogorov test furnishes the general yes/no answer when comparing different spectra of the same compound.

The user is able to know the fraction of altered three-dimensional structure through the high-level analysis that is particularly straightforward when HSQC and HSQC-TROSY spectra are evaluated. It is possible to map the behavior of specific peak features (chemical shift and volume variation) by means of histograms. The list of residues that are mostly involved in the conformational change is available only when at least the reference spectrum has been previously assigned, otherwise it contains just the most altered ppm atom positions with respect to the considered feature. The routine automatically detects, from the very beginning, if the data at hand were already assigned in order to perform different computations.

The datasets are described in par. 2.2.2.2 and in par. 2.2.2.3. In particular, three cases are evaluated:

1. The prion spectrum measured without ligand binding with xenon is compared with four prion spectra incrementally pressurized with xenon (par. 2.2.2.2).
2. The prion spectrum acquired at 0.1 MPa (293 K) is compared with six other spectra measured at incrementing pressure (par. 2.2.2.3).
3. The prion spectrum measured at 293K (200 MPa) is compared with four prion spectra obtained with incrementing temperature (par. 2.2.2.3).

5.2 Prion protein (*hu*PrP^{C})

5.2.1 General considerations

The recombinant human prion protein (*hu*PrP^{C}) [Prusiner, 1998] has been pressurized with xenon and used as a trial case for the AUREMOL-QTA module (see par. 2.2.2.2). High pressure and temperature structure dependence of the same protein has been evaluated as well (par. 2.2.2.3). High pressure is a valid tool to analyze the dynamics and the structure of the folding intermediates, since the transmissible spongiform encephalopathies (TSes) adopt alternative folds propagating the disease [Kuwata et al., 2002].

As reported in Fig. 5.1 the structure of the human prion protein has been solved only in the residue range 121-230 since the rest (23-120) is a flexible disordered tail [Zahn et al., 2000].

The developed method is able to identify structural changes due to external conditions as pressure and temperature and to ligand binding with xenon, comparing respectively seven TROSY spectra *hu*PrP(121-231) with increasing pressure (0.1, 50, 100, 125, 150, 175 and 200 MPa at 293 K), five TROSY spectra *hu*PrP(23-231) with increasing temperature (293, 303, 313, 323 and 333 K at 200 MPa) and five HSQC spectra at different pressures (0.1, 0.2, 0.4, 0.8 and 0.14 MPa).

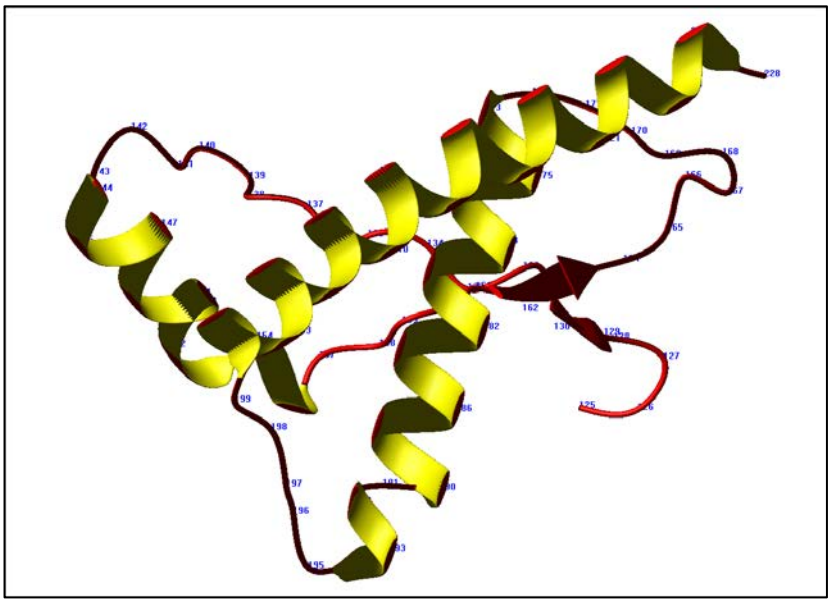


Fig. 5.1 The human prion protein *hu*PrP^{C}: the folded structure (1qm2 .pdb file) from residues 125 to 228.

The behavior of the signals is typically evaluated manually by the spectroscopist. The developed routine has been thought to ease such task, since it recognizes and stores automatically all the detected feature changes among the compared spectra.

As reported in Fig. 5.2, the peaks Gly131 and Thr199 experience chemical shift variations due to the high pressure (part *a*) and to xenon-binding (part *b*). The xenon-binding pressurized spectra reveal a smaller shift variation with respect to the high pressure case. Moreover, the signals corresponding to the residues located in the proximity of the cavities containing the xenon tend to show the strongest chemical shift and volume variations. High pressure application results in chemical shift changes of all resonances, in volume decrease of some signals, in a complete disappearing of some peaks and in a not significant volume variation of some others. The tedious task of a manual identification, quantification and storage of all the altered features is automatically performed by the quality control routine.

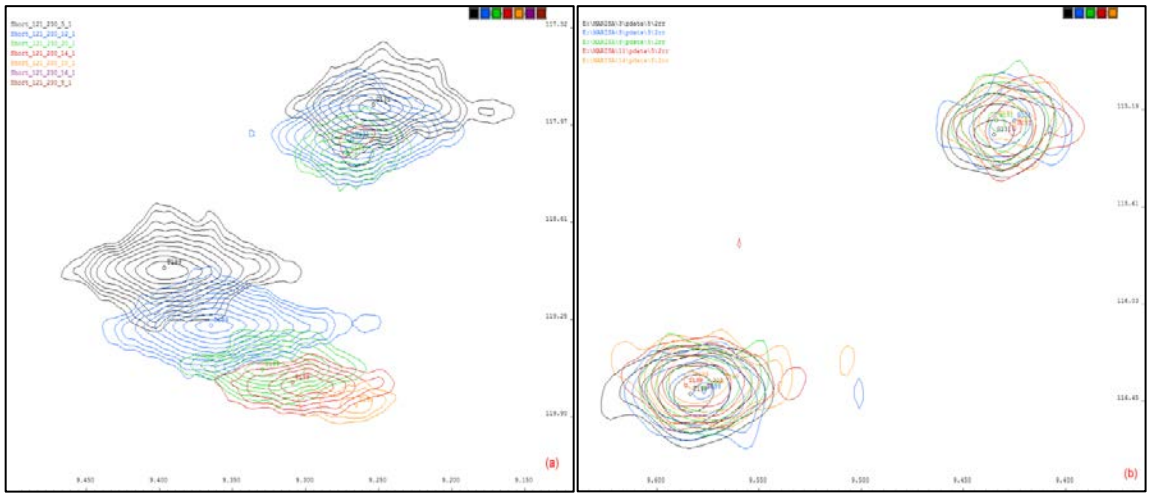


Fig. 5.2 High pressure and ligand binding effects (with xenon) on the Thr199 and Gly131 residues of (*hu*PrP^{C}): the ligand binding with xenon (b) reveals a smaller shift variation with respect to the high pressure case (a).

5.2.2 Quality control of the xenon-binding dataset

The set of five spectra (varying the pressure from 0.1 to 1.4 MPa) has been loaded into the main interface correctly setting all the previously described parameters (see Fig. 4.2) as shown in Fig. 5.3.

In this case, the reference spectrum (experiment PRION_HSQC_M\3\pdata\5\2rr) was previously assigned, while the four test spectra were not assigned at all.

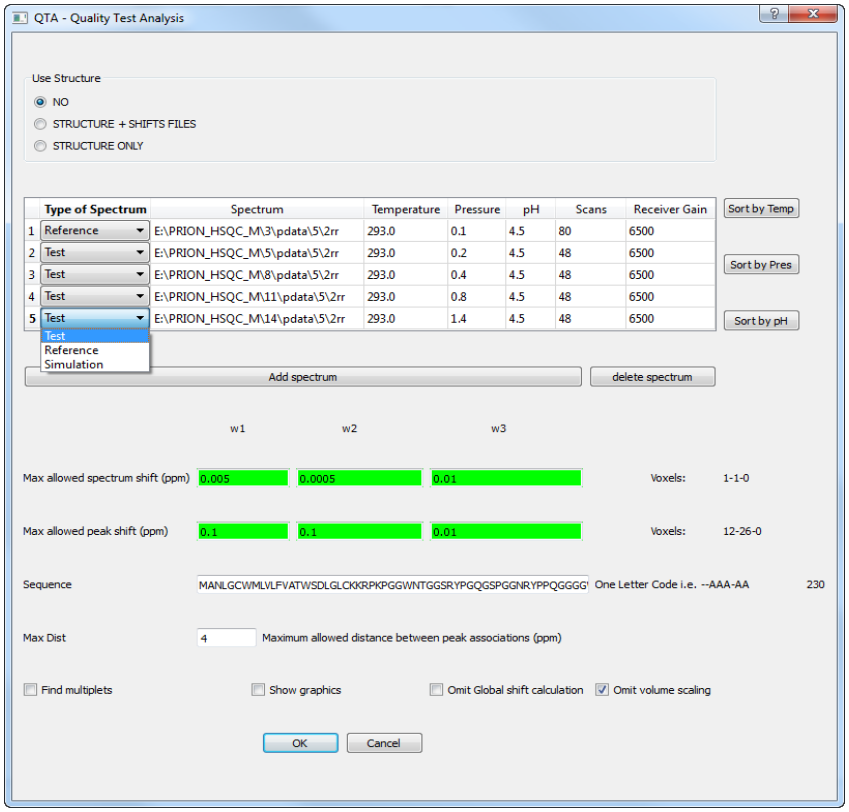


Fig. 5.3 Main interface of the quality test control on the xenon-binding human prion test case (*hu*PrP^{C}): the main dialog shows the five HSQC spectra and all the other user defined parameters as the type and the folder of the investigated spectra. The parameters of temperature, pressure, pH, NS (number of scans) and RG (receiver gain) are necessary to perform a reliable comparison. For more explanations see the following text.

The reference experiment has been collected with a different number of scans (*NS*) with respect to all others (the *RG* and the *NC_proc* parameters have the same value), involving the volume normalization of each test spectrum as described in par. 3.4.1.1. If the user does not provide any information about the number of scans applied on each experiment and no acquisition files are available, they are meant to be identical all over the considered spectra. The same rule is used for the *RG* parameter. In addition, the SW (thus the OFF) of the reference experiment was differing along the δ_2 direction by 2 ppm with respect to all other cases. This difference is automatically managed after a prior warning message asking for the user consent as shown in Fig 5.4. In case of agreement of the user, the routine continues the computation otherwise the quality test analysis is aborted. The SW and the OFF parameters can be directly extracted from the processing files, thus no acquisition files are needed.

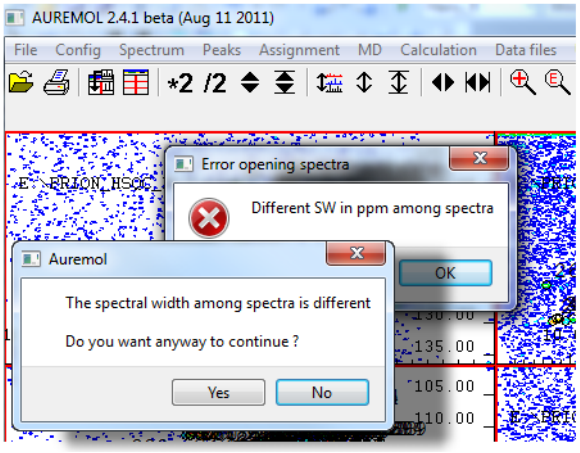


Fig. 5.4 Warning messages of the AUREMOL-QTA: in case of different acquisition parameters among the compared spectra (spectral width and offset), the user can decide whether to continue or to abort the computation.

The maximum allowed local shift (peak shift) option (described in par. 3.4.2.3.2) has been set to 1.0 ppm (corresponding to 113 and 227 voxels shift along δ_2 and δ_1 direction respectively). In particular, the user defined shift of 1.0 ppm is applied along both directions. Such large shift is advisable when dealing with pressure dependent dataset

where a higher local peak shift must be allowed in order to correctly associate signal resonances among spectra. The automated volume scaling has been avoided due to the nature of the investigated test case.

As reported in Table 5.1, the computed neighborhood distances (see par. 3.4.2.3.1) have shown a mean width of 0.8 ppm ($\pm$ 0.033) along δ_1 and 0.4 ($\pm$ 0.076) along δ_2 direction. To allow a fast computation of the neighborhood distances a multi-threaded code has been written [Molkentin, 2007].

Spectrum	Neighborhood list distance (NLST) width (δ_1)	Neighborhood list distance (NLST) width (δ_2)
Exp 3	0.782319 ppm	0.3403 ppm
Exp 5	0.817479 ppm	0.441998 ppm
Exp 8	0.799899 ppm	0.488936 ppm
Exp 11	0.85264 ppm	0.535874 ppm
Exp 14	0.764739 ppm	0.39506 ppm

Table 5.1 List of the computed neighborhood distances: for each experiment of the investigated human prion protein (*hu*PrP^C) the neighborhood distances have been calculated along all the acquired directions.

As shown in Fig. 5.3, the AUREMOL-QTA routine does not necessarily require the primary sequence, but when provided it is used to combine the chemical shift changes in both directions as described in par. 3.4.4.3 [Schumann et al, 2007]. As a consequence of the provided primary sequence, the total number of residues appears in the right side of the dialog (Fig. 5.3), beside the sequence, with the aim to consent a visual control of the furnished information. The check box “Find multiplets” (see option “s” of Fig. 4.2) has not been used during the computation of this quality control (it is typically used when the back-calculated spectrum is available).

After the whole computation has been completed (all the steps described in the low, mid and high-levels have been performed), a series of different results are provided to the spectroscopist (e.g. spectra containing peaks surrounded by colored boxes, matching and mismatching signal ratios, residues ratios, yes/no answers, histograms of feature variations and lists of most altered residues).

5.2.2.1 Quality control detailed results

The quality routine does not simply furnish a general result defining the probability of comparing reference and test spectra of the same compound. It also provides detailed information regarding every picked signal in any considered spectrum.

In Fig. 5.5 the residue Gln212 has been zoomed out and a detailed comparison of this reference peak with all the associated test peaks is available to the user pushing the keyboard character “q”, as depicted in Fig. 5.6.

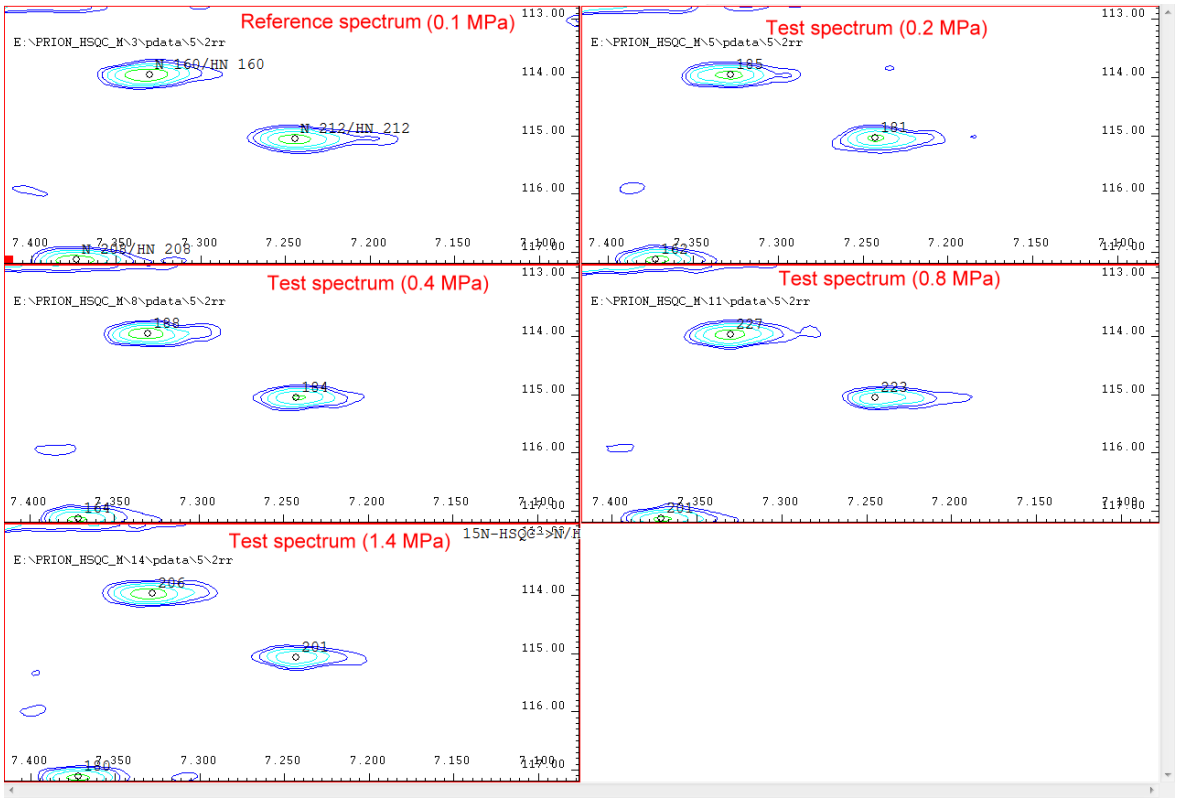
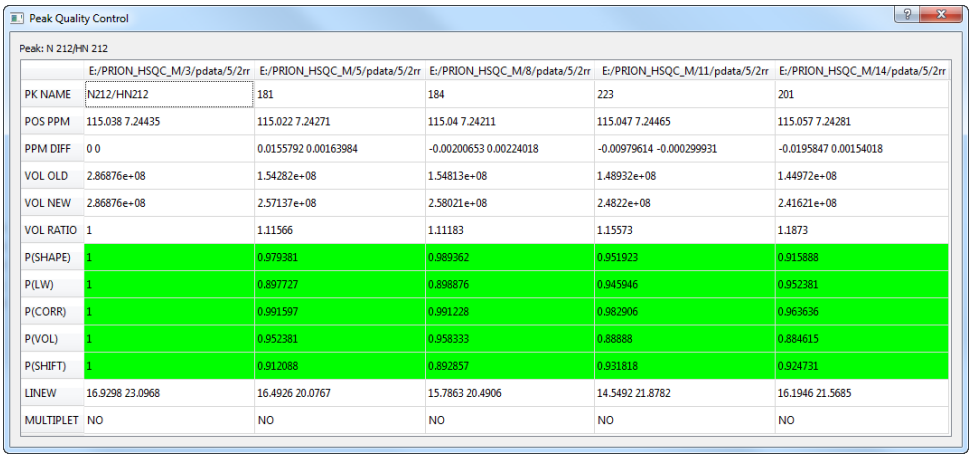


Fig. 5.5 Zoom of the residue Gln212 of the human prion protein (*hu*PrP^{C}) with xenon-binding: it shows all the compared spectra involved in the quality control analysis. The top left spectrum (reference) has been compared with all other test spectra (increasing the pressurization with xenon).


Peak Quality Control

Peak: N 212/HN 212

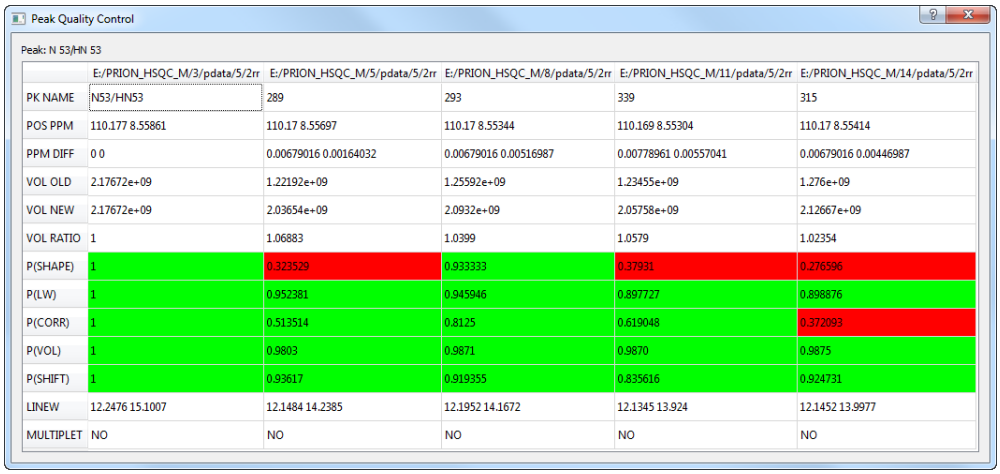
	E:/PRION_HSQC_M/3/pdata/5/2rr	E:/PRION_HSQC_M/5/pdata/5/2rr	E:/PRION_HSQC_M/8/pdata/5/2rr	E:/PRION_HSQC_M/11/pdata/5/2rr	E:/PRION_HSQC_M/14/pdata/5/2rr
PK NAME	N212/HN212	181	184	223	201
POS PPM	115.038 7.24435	115.022 7.24271	115.04 7.24211	115.047 7.24465	115.057 7.24281
PPM DIFF	0 0	0.0155792 0.00163984	-0.00200653 0.00224018	-0.00979614 -0.000299931	-0.0195847 0.00154018
VOL OLD	2.86876e+08	1.54282e+08	1.54813e+08	1.48932e+08	1.44972e+08
VOL NEW	2.86876e+08	2.57137e+08	2.58021e+08	2.4822e+08	2.41621e+08
VOL RATIO	1	1.11566	1.11183	1.15573	1.1873
P(SHAPE)	1	0.979381	0.989362	0.951923	0.915888
P(LW)	1	0.897727	0.898876	0.945946	0.952381
P(CORR)	1	0.991597	0.991228	0.982906	0.963636
P(VOL)	1	0.952381	0.958333	0.88888	0.884615
P(SHIFT)	1	0.912088	0.892857	0.931818	0.924731
LINEW	16.9298 23.0968	16.4926 20.0767	15.7863 20.4906	14.5492 21.8782	16.1946 21.5685
MULTIPLET	NO	NO	NO	NO	NO

Fig. 5.6 The peak quality control dialog of peak Gln212 of the human prion protein (*hu*PrP^{C}) with xenon binding: the dialog shows in the first row the peak association among the spectra that has been automatically determined by the routine. In the second row of the dialog the ppm positions of each peak is reported. The third row represents the ppm differences among all the connected peaks. The row named VOL OLD represents the volume of each peak before applying the scaling factor (including NS, RG and NC_proc). The row VOL NEW represents the volume after having normalized the volumes (see eq. 3.2). The row named VOL RATIO represents the volume ratio between VOL OLD and VOL NEW of each peak after the normalization. The rows from seven to eleven (green highlighted) represent the values of the Bayesian analysis of each investigated feature (shape, line width, time cross-correlation, volume and peak position variation). The penultimate row shows the measured line widths in Hz of each peak in both directions. The last row named MULTIPLET tells the user if the peak is a multiplet.

The dialog reported in Fig. 5.6 shows the peak connections among the spectra that have been determined by the routine. For each peak of each spectrum the shift (in ppm unit) along each direction (δ_1, δ_2) is reported. The shift differences are computed between the reference spectrum peak and all the associated test peaks. The information of the volume of each peak before and after the normalization (see eq. 3.2) is reported as well. The spectroscopist has the possibility to analyze the probability of each peak association with respect to the five features (shape similarity, line width, time cross-correlation, volume and position variation). If the probability values of a specific feature are smaller than 0.5 the dialog shows these critical values highlighting them in red as shown in Fig 5.7. If the probability of each feature in each comparison between different spectra is major than the value of 0.5, the boxes are green highlighted. The dialog provides also information about

the line width in Hz of each peak in both directions and about the detection of multiplet structures.

In Fig. 5.6 the peak Gln212 of the reference spectrum has been automatically associated with the peaks number 188, 184, 233 and 201 of the test spectra. The shift, volume, shape, correlation and line width differences among Gln212 and the other test peaks is very small implying high probability results (as reported in the green colored rows containing the probability of the shape, of the line width, of the time cross-correlation, of the volume and of the position). The VOL OLD row reveals that the reference volume was originally about the double of the test cases due to the different *NS* used during the acquisition. The *RG* and the *NC_proc* were instead unaltered among the reference and all the test cases (see Fig. 5.3), thus the VOL NEW row contains the scaled test volumes after *NS* correction (see paragraph 3.4.1.1). It is evident that there is a slight volume decrement from the reference to the last test case, while the chemical shift reveals small variations only along the indirect direction. It is straightforward that all the features have a high probability.



	E:/PRION_HSQC_M/3/pdata/5/2rr	E:/PRION_HSQC_M/5/pdata/5/2rr	E:/PRION_HSQC_M/8/pdata/5/2rr	E:/PRION_HSQC_M/11/pdata/5/2rr	E:/PRION_HSQC_M/14/pdata/5/2rr
PK NAME	N53/HN53	289	293	339	315
POS PPM	110.177 8.55861	110.17 8.55697	110.17 8.55344	110.169 8.55304	110.17 8.55414
PPM DIFF	0 0	0.00679016 0.00164032	0.00679016 0.00516987	0.00778961 0.00557041	0.00679016 0.00446987
VOL OLD	2.17672e+09	1.22192e+09	1.25592e+09	1.23455e+09	1.276e+09
VOL NEW	2.17672e+09	2.03654e+09	2.0932e+09	2.05758e+09	2.12667e+09
VOL RATIO	1	1.06883	1.0399	1.0579	1.02354
P(SHAPE)	1	0.323529	0.933333	0.37931	0.276596
P(LW)	1	0.952381	0.945946	0.897727	0.898876
P(CORR)	1	0.513514	0.8125	0.619048	0.372093
P(VOL)	1	0.9803	0.9871	0.9870	0.9875
P(SHIFT)	1	0.93617	0.919355	0.835616	0.924731
LINEW	12.2476 15.1007	12.1484 14.2385	12.1952 14.1672	12.1345 13.924	12.1452 13.9977
MULTIPLET	NO	NO	NO	NO	NO

Fig. 5.7 The peak quality control dialog of the residue Gly53 of the human prion protein (*hu*PrP^{C}) with xenon binding showing variations of features: the dialog shows red-colored boxes where the probability value of the investigated feature is smaller than 0.5 (as the shape probability of the first, the third and the fourth test cases and the correlation probability of the fourth test spectrum).

In Fig. 5.7 the detailed analysis of the reference peak Gly53 is reported. It has been associated with the peaks number 289, 293, 399 and 315 of the test spectra. It shows a VOL OLD affected by the different number of scans used during the acquisition that is corrected as reported in VOL NEW. The dialog shows red-colored boxes where the probability value of the investigated feature is smaller than 0.5, as the shape probability of the first, the third and the fourth test cases and the correlation probability of the fourth test spectrum.

5.2.2.2 Bayesian feature analysis

The user can visually analyze each feature of interest in any selected spectrum through the “Quality test” menu under the “Calculation” menu. As shown in Fig. 4.1 the option “Show analyzed peaks” opens a dialog where the user can select the desired feature of interest. Every feature can be analyzed varying the probability in a range from zero to one and can be highlighted with user defined colored boxes. In the case reported in Fig. 5.8 the shape feature is selected and as a result red boxes will surround the peaks having a shape probability lower than 0.5. The probability results may be obtained either on each feature separately or considering them all together. In particular, in the latter case the option “All” must be checked and as a result the spectrum will appear with colored boxes around the peaks having a probability lower than 0.5 for all the features simultaneously. The routine verifies the probability values of the selected feature in all the opened test spectra and it furnishes colored boxes around the references peaks whose probability is lower than 0.5 at least in one of the opened test spectra. The user can personally define the lower and the upper bounds of the feature probability. In the reported case a low probability was particularly significant for the identification of the most changed signals in the spectra, thus the most altered residues in the structure. A probability range between 0.5 and 1 is instead useful for determining the most stable parts of the protein.

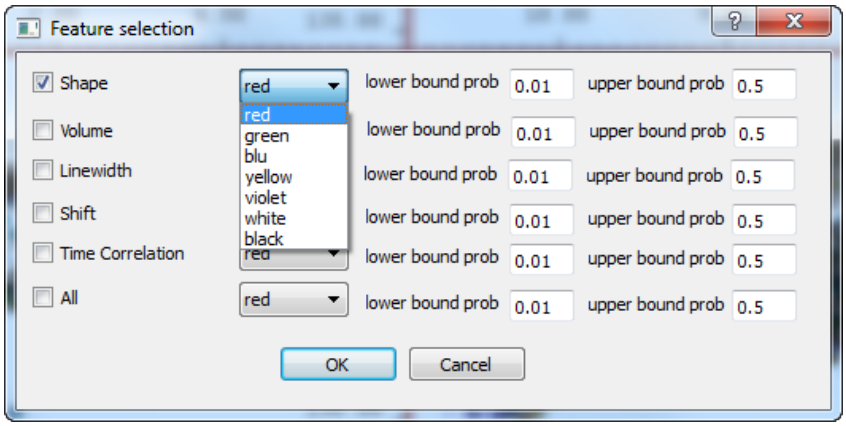


Fig. 5.8 The feature selection dialog: the user can select any of the desired features (shape similarity, volume, line width, position, time cross-correlation or all the features together) in the range defined by the upper and the lower probability bounds. In addition, the user can choose one of the seven available colors to highlight the feature of interest.

The routine controls the spectra that are open (in the AUREMOL main window) in order to build up the visual results only for the opened spectra. An example of the results regarding the chemical shift variations of the reference spectrum is reported in Fig. 5.9. This spectrum can be obtained if the shift feature box of Fig. 5.8 has been checked and the probability bounds are established between 0.1 (lower bound) and 0.5 (upper bound). Red boxes surround the peaks whose shift probability is lower than 0.5.

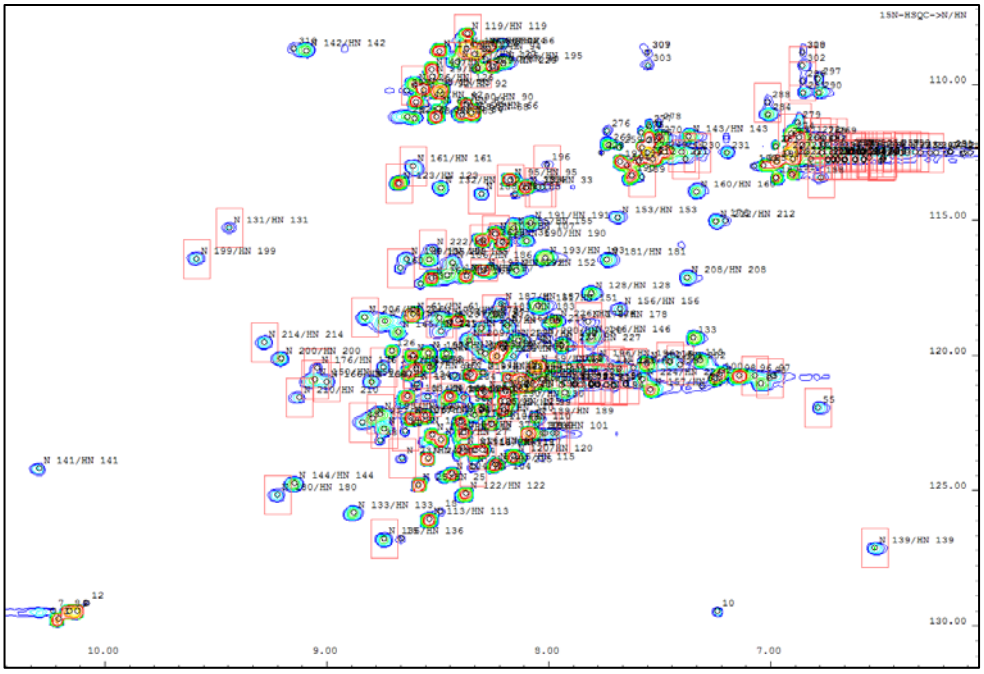


Fig. 5.9 Identification of the chemical shift variations by means of the feature selection routine of the human prion protein (*hu*PrP^{C}) with xenon binding: the red-colored boxes identify peaks whose chemical shift probability is smaller than 0.5. This comparison has been performed between each reference peak and all the other associated peaks of the four test spectra.

The same reference spectrum can be analyzed using the shape similarity feature. All the peaks having a probability, with respect to the shape similarity feature, smaller than 0.5 are red-highlighted as reported in Fig. 5.10.

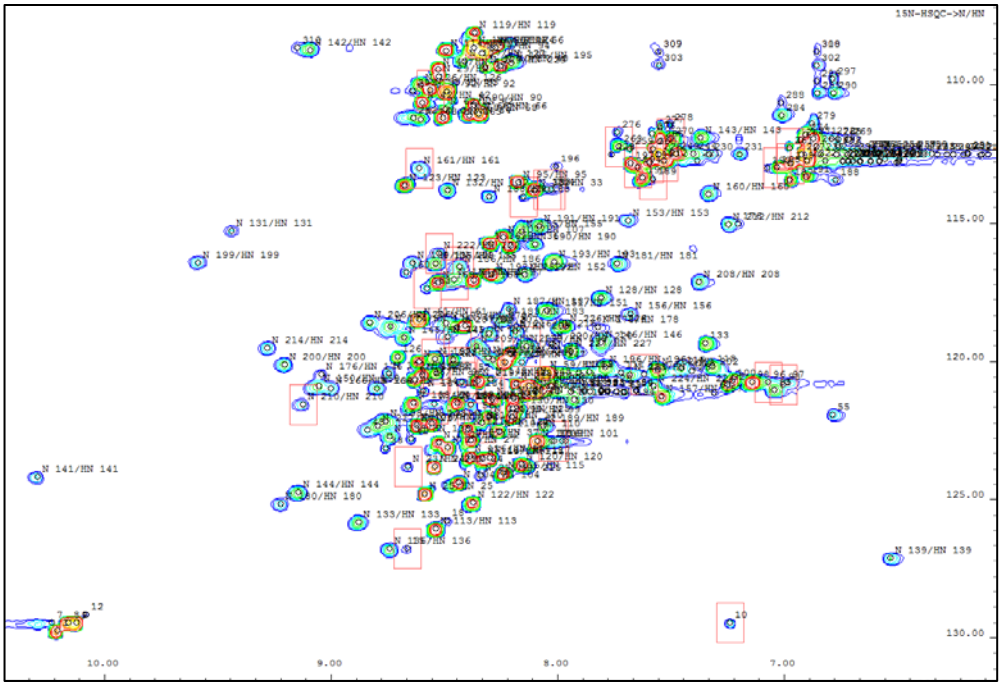


Fig. 5.10 Identification of the shape variation of the reference spectrum by means of the feature selection routine of the human prion protein (*hu*PrP^{C}) with xenon binding: the red boxes are showing the peaks whose shape similarity probability is smaller than 0.5 in at least one of all the four investigated test spectra.

The reference spectrum can be analyzed also using the line width feature. As showed before in this case the peaks having a probability (with respect to the line width feature) smaller than 0.5 are visualized with red-colored boxes. This is reported in Fig. 5.11.

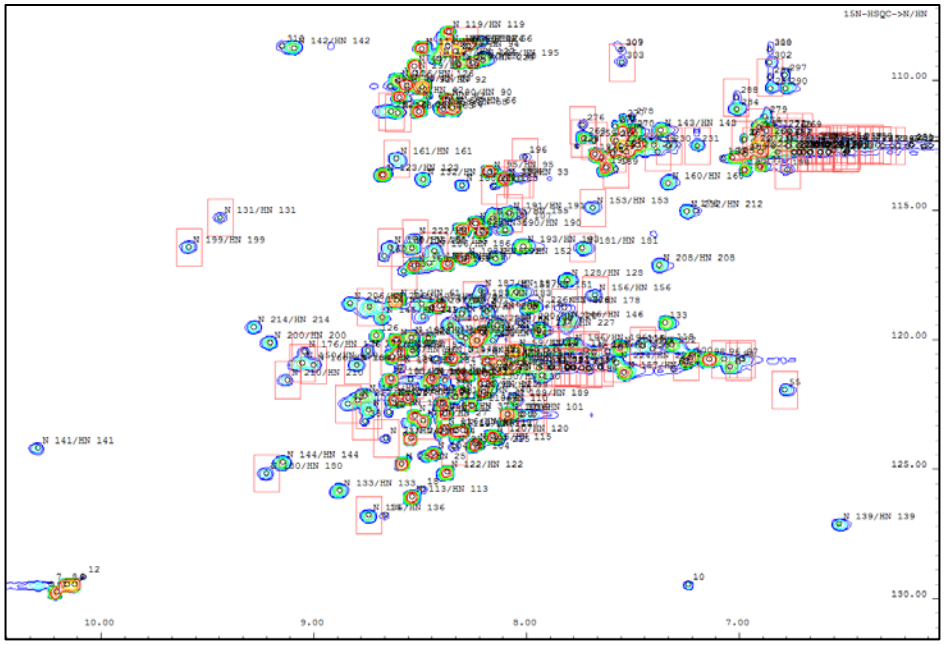


Fig. 5.11 Identification of the line width variation of the reference spectrum by means of the feature selection routine of the human prion protein (*hu*PrP^{C}) with xenon binding: the red boxes are showing the peaks whose line width probability is smaller than 0.5 in at least one of all the four investigated test spectra.

The volume variation of each reference peak with respect to all the other investigated test spectra is shown in Fig. 5.12.

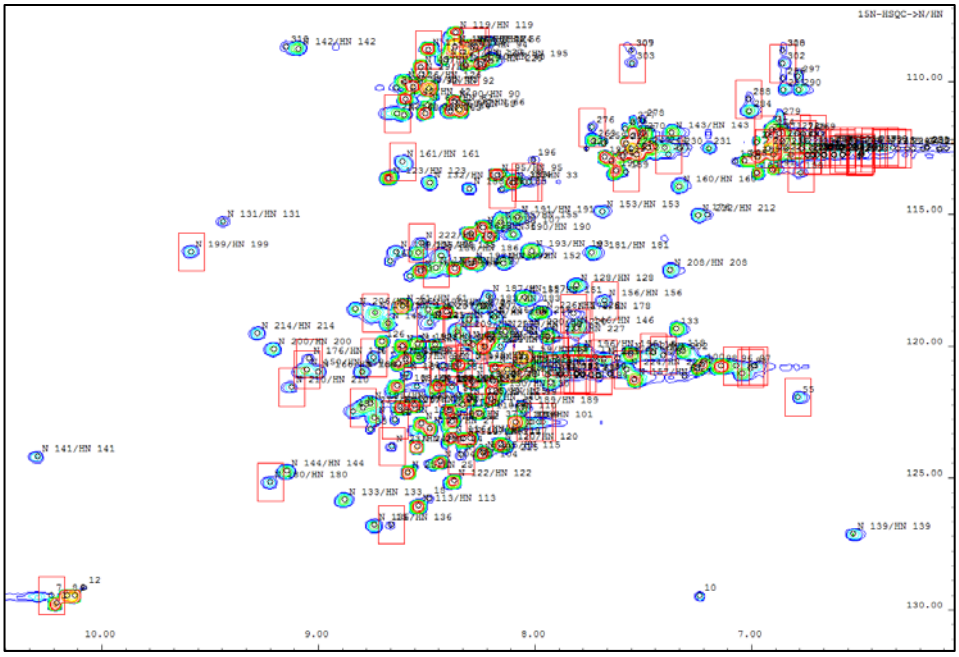


Fig. 5.12 Identification of the volume variation of the reference spectrum by means of the feature selection routine of the human prion protein (*hu*PrP^{C}) with xenon binding: the red boxes are showing the peaks whose volume probability is smaller than 0.5 in at least one of all the four investigated test spectra.

It is possible to show the probability results of the superimposition of all the features. Peaks having each of the selected features with a probability smaller than 0.5 are identified and shown as reported in Fig. 5.13. Blue, green, yellow, black and red boxes correspond respectively to volume, line width, chemical shift, time cross-correlation and shape probabilities lower than 0.5.

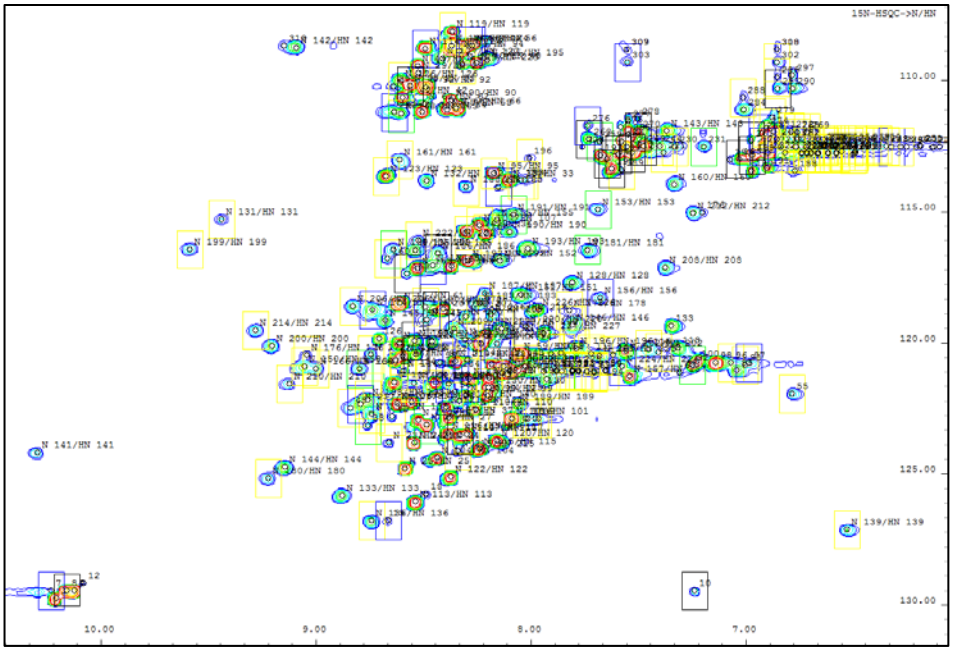


Fig. 5.13 Identification of all the peak features with a probability smaller than 0.5 of the human prion protein (*hu*PrP^{C}) with xenon binding: peaks of the reference spectrum are highlighted with different colors with respect to the investigated feature. Highlighted are peaks having a volume probability (blue boxes), a line width probability (green boxes), a chemical shift probability (yellow boxes), a cross-correlation probability (black boxes) and a shape-similarity probability (red boxes) smaller than 0.5 when compared with all the four investigated test spectra.

Obviously, some features intersect each other. For example a peak showing a pronounced shape variation has also a remarkable line width variation and consequently a volume change. In this case the shape red-colored boxes have been completely superimposed by correlation (black) and line width (green) boxes, thus they are not visible. In Appendix C the list of the residues whose feature probabilities are lower than 0.6 is reported, including the experiment number where it occurred.

The AURMOL-QTA method allows a comparison for each feature of each peak in every analyzed spectrum. The result of this analysis is used to infer the effect of these variations on the three-dimensional structure (see Appendix C). As shown in Fig. 5.14, the three-dimensional structure is changed in the color-highlighted regions. The residues whose features probabilities are lower than 0.6 have been detected and differently colored (orange

for volume, light pink for position, cyan for line width and magenta for a joint representation of these three features as reported in the (*a*) part of Fig. 5.14) in the structure dependently on the feature. In the (*b*) part of the same Figure the magenta colored regions identify those residues whose features (shape, correlation, position, volume and line width) have simultaneously a probability lower than 0.6 (red highlighted in Appendix C). In particular, six residues experience a simultaneous variation of all the features: Lys23, Met109, Val161, Val210, Met213 and Ser222. They are all involved in a slow exchange with xenon (on the NMR time scale) since their signals split through the experiments.

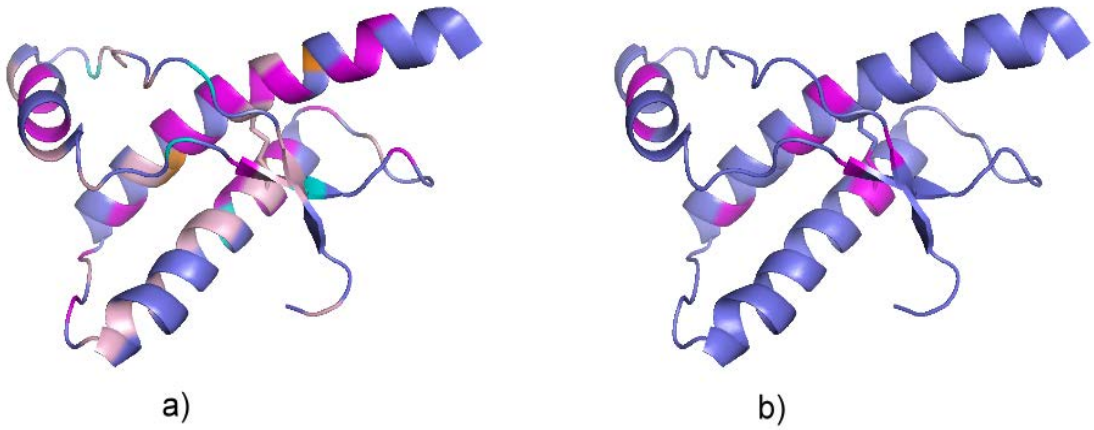


Fig. 5.14 Conformational changes of the human prion protein (*hu*PrP^{C}) with xenon binding in the three-dimensional structure identified by means of the analyzed features: residues whose probability of the volume (orange), chemical shift (light pink), line width (cyan) and all the three together (magenta) is lower than 0.6 (a); residues involved in a simultaneous variation of all the features (volume, chemical shift, line width, time correlation and shape similarity) whose probability is lower than 0.6 (magenta) (b): K23, M109, V161, V210, M213 and S222 .

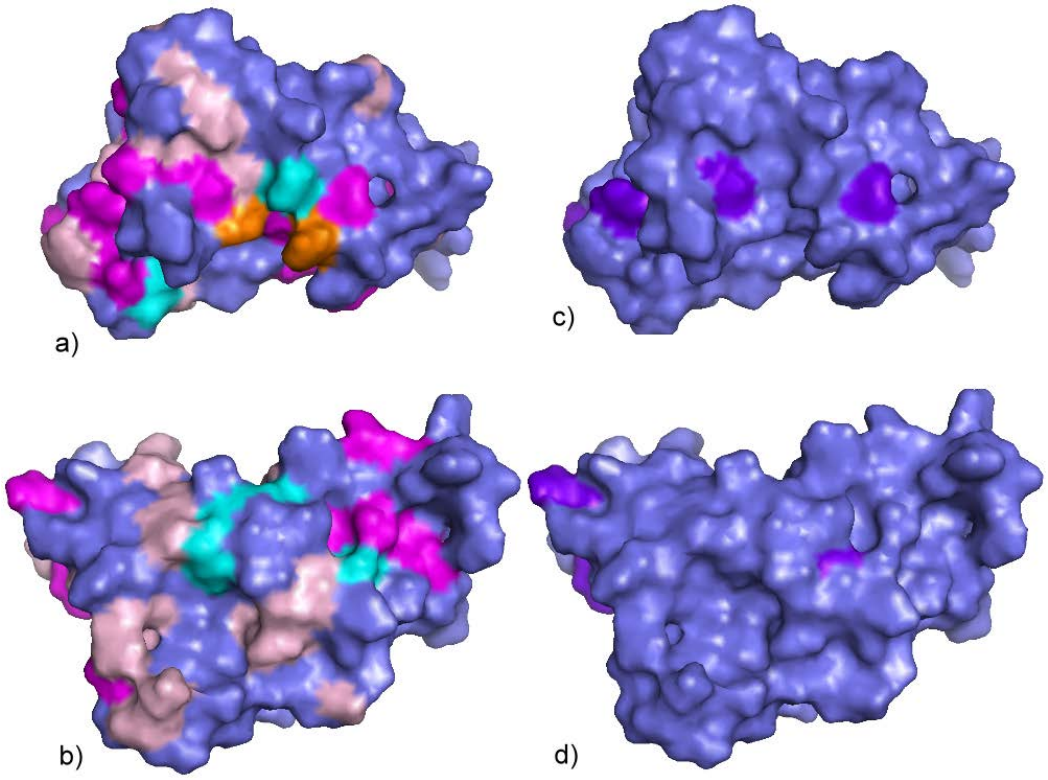


Fig. 5.15 Surface changes identified by means of the analyzed features of the human prion protein (*hu*PrP^{C}) with xenon binding: residues whose probabilities of volume (orange), chemical shift (light pink), line width (cyan) and all the three together (magenta) are lower than 0.6, in the front (a) and back view (b); residues involved in a simultaneous variation (volume, chemical shift, line width, time correlation and shape similarity) of all the features (with a probability lower than 0.6) are highlighted in magenta, in the front (c) and back view (d).

As shown in Fig. 5.15, the routine correctly identifies the most important conformational changes of the xenon binding on the three-dimensional structure of the human prion protein. The highlighted residues are in fact those ones located close to the cavities where the xenon can be buried. This automated identification can be used to speed up the determination of the ligand binding sites.

5.2.2.2.1 Peaks that have not been associated among the spectra (missing signals)

Some of the picked peaks of the reference spectrum cannot be associated to any other peak in all the test spectra and vice versa. As shown in Fig. 4.1 those not associated or missing peaks can be visualized by the user via the option “Show missing peaks” from the “Quality test” menu. The missing signals detected in the reference spectrum appear surrounded by red boxes. In this case all of them represent not assigned or baseline peaks, as shown in Fig. 5.16. They are peaks that at least have not been found in one of the opened test spectra.

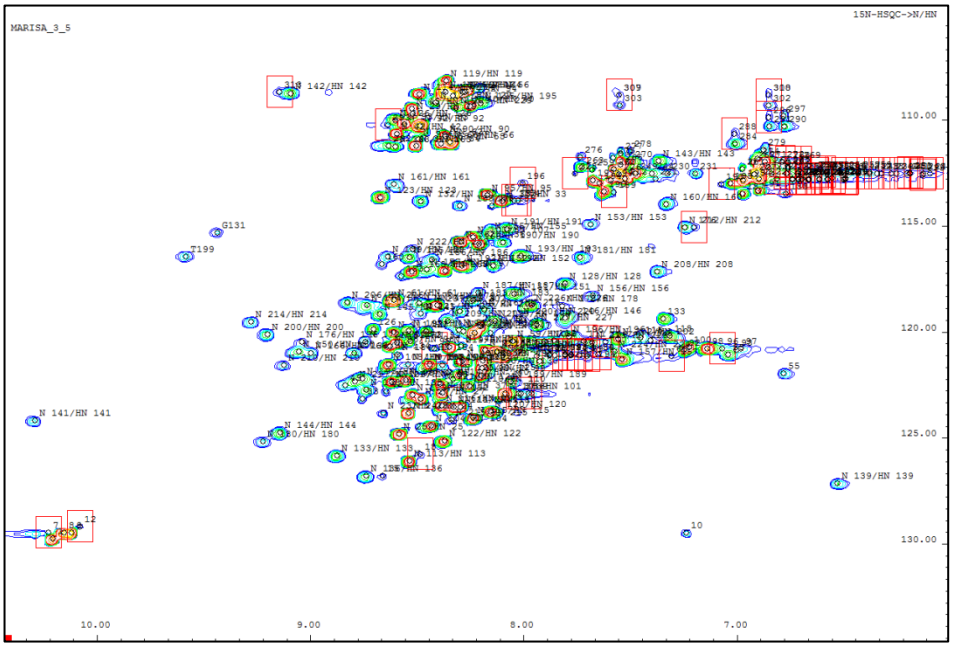


Fig. 5.16 Identification of not associated (missing) peaks between the reference spectrum and all the test cases of the human prion protein (*hu*PrP^{C}) with xenon binding: missing peaks of the reference spectrum are surrounded by red boxes.

5.2.2.3 Structural analysis by means of histograms

The spectroscopist has the possibility to further analyze the volume and the chemical shift variations among the reference spectrum and all the test spectra. As described in Fig. 3.42 and in paragraph 3.4.4 when dealing with HSQC data the feature mapping can be performed by means of histograms containing the average and the standard deviation of the selected feature. This is possible through the option “Show structural analysis” from the “Quality test” menu (see Fig. 4.1). As shown in Fig. 5.17 the option “Show structural analysis” opens a dialog used to conduct these specific analyses.

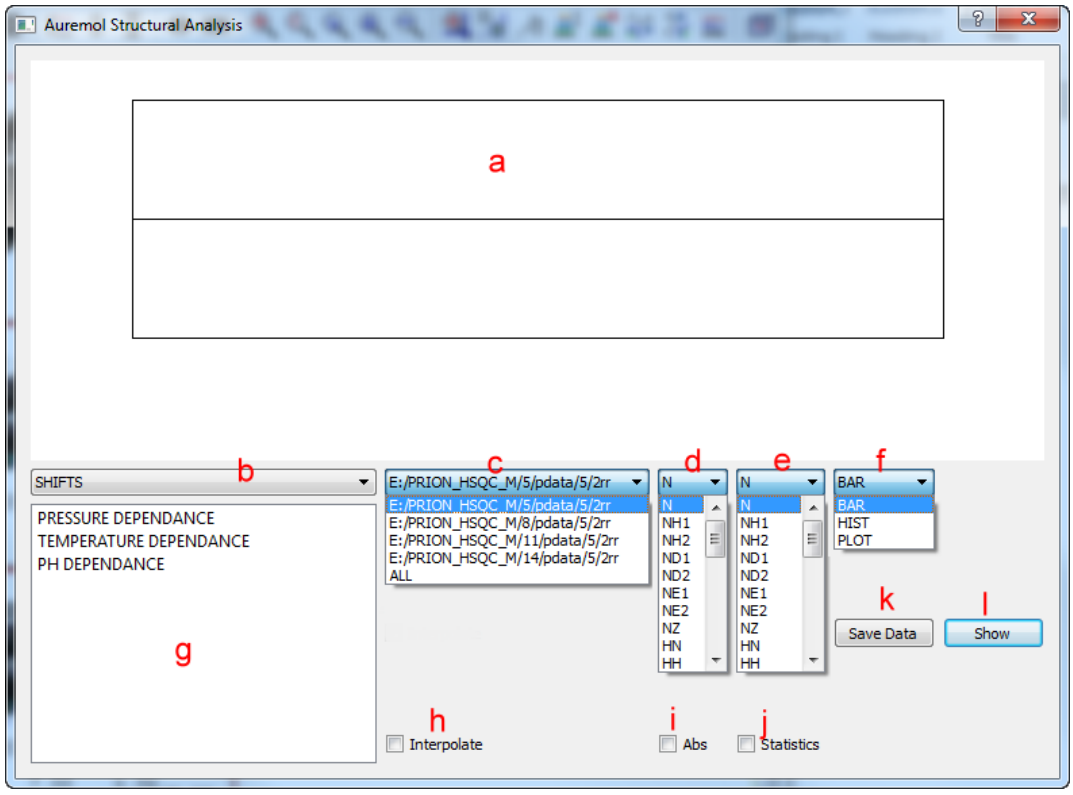


Fig. 5.17 The AUREMOL Structural Analysis dialog: the dialog shows the residual changes via a graphical interface. The upper part is reserved to the graphical representations (a) through bars, histograms and standard plot (f) of the investigated feature (chemical shift or volume). In the lower part, the feature to be analyzed has to be selected (b) like the volume and the chemical shift variation. The user can select one or all the considered test spectra to perform the analysis (c). In case of a chemical shift structural analysis the user

can select the atoms that have to be analyzed (d-e). The user can define the external condition that has experienced a relevant variation among the spectra (g) as the temperature, pressure or pH. It is possible, if the user has selected all the spectra (option c) to show a linear interpolation of the data (h). In this case the interpolated slope (one for every residue) will be plotted. The user is enabled both to show absolute values or not (i) and to generate the corresponding statistics including average and standard deviation (j). It is possible to store the result of the analysis through the button "Save Data" (k). To start the analysis the user has to push the button "Show" (l).

The dialog reported in Fig. 5.17 shows the variations of the residue features (chemical shift or volume) via a graphical interface. The upper part of the figure (a) is reserved to the graphical representations of bars, histograms and standard plots (f). In the lower part the feature that has to be analyzed is selected (b). In the considered case the volume and the chemical shift variations are considered, but as reported in the discussion section also some other features may be added in this analysis (shape and line width). This analysis can be performed using one or more spectra that have been investigated (c). If the user chooses to analyze the chemical shift it is possible to select the atoms that have to be considered (d-e). In particular, the shifts of the considered HSQC may be analyzed either in each direction separately (only the proton or only the nitrogen direction) or combining them according to the work of *Schumann et al., 2007*. In the former case the "d" option must contain HN or N and the "e" option must be blank, while in the latter case in "d" the HN term is selected and in "e" the N term is chosen or vice versa. The user can define the external condition that has experienced a relevant variation among the spectra (g) as the temperature, pressure and pH. In order to analyze the structural changes, an additional option is available: the data can be interpolated through a linear interpolation (h). If the user selects this option the routine will show slope (one for every residue) of the linear interpolation instead of mere differences. The user is enabled to show absolute values or not (i) and to generate the corresponding statistics including average and standard deviation (j). It is possible to store the result of the analysis through the button "Save Data" (k). To start the analysis the user has to push the button "Show" (l).

This analysis has been applied on the human prion protein dataset and one of the possible results is reported in Fig. 5.18.

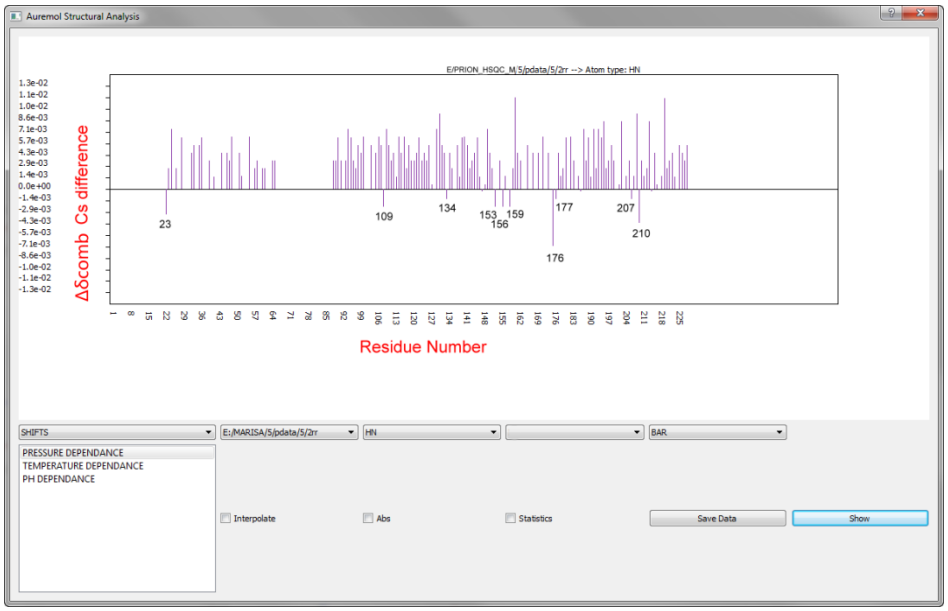


Fig. 5.18 The chemical shift dialog of the AUREMOL Structural Analysis of the human prion protein (*hu*PrP^{C}) with xenon binding: the dialog shows chemical shift variations (H^N) without linear interpolation and without absolute values between the reference and the first test spectrum (one bar for each residue).

The bars in Fig. 5.18 represent the chemical shift variations (H^N) without linear interpolation and without absolute value between the reference and the first test spectrum. The peak shift Δ has been obtained as following

$$\Delta = (\delta - \delta_0) \tag{4.1}$$

The shift Δ is positive if it is a downfield shift and negative in the opposite case (see negative bars in Fig. 5.18). The bars of some residues are not visible due to several reasons: the shift difference is almost zero; the reference residue peak has not been assigned; the reference peak has not been associated to any other signal in the first test spectrum (see par. 3.4.4.3).

The analysis of the normalized volume differences (*NS*, *RG* and *NC_proc*) without considering the volume scaling s_j (see eq. 3.14) is reported in Fig. 5.19.

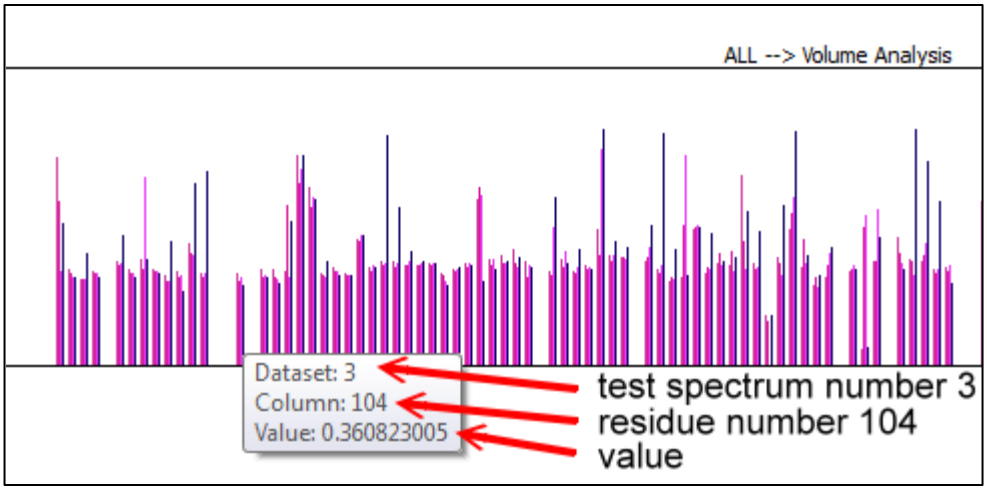


Fig. 5.19 The volume dialog of the AUREMOL Structural Analysis of the human prion protein (*hu*PrP^{C}) with xenon binding: the dialog shows the volume variations between the reference and all the other test spectra (four bars for each residue). The user can graphically select the bar corresponding to a specific dataset (representing any test spectrum) and the residue of interest.

The volume structural analysis is performed as the chemical shift one, varying only the option “b” of Fig. 5.17 and without selecting any atom in the “d” and “e” options. If the reference spectrum is simultaneously compared to all the four test spectra, the dialog will show four bars for each residue. In this manner the user can simultaneously detect which datasets (test spectra) differ mostly with respect to the reference case and in which residues. This task can be done graphically positioning the cursor over the bars of interest (see Fig. 5.19).

In Fig. 5.20 another example of the chemical shift structural analysis is represented. In this case the user has decided to compare the combined chemical shift [Schumann et al., 2007] of the reference spectrum with all the test cases. As previously explained the plot contains four bars for each residue if the option “c” in Fig. 5.17 is set to all. In this case the user has required a linear interpolation on those bar (checking the option “h”) thus one bar for each residue will be obtained. They represent the absolute slope of every linear interpolation (one for each residue). The mean and the double of the standard deviation are computed and plotted on the dialog as a blue and a green line respectively. The residues whose absolute slope is greater than the double of the standard deviation are easily recognizable.

They are automatically listed in another dialog. The list of such residues will contain their name if the reference spectrum was previously assigned otherwise only the peak name will be provided.

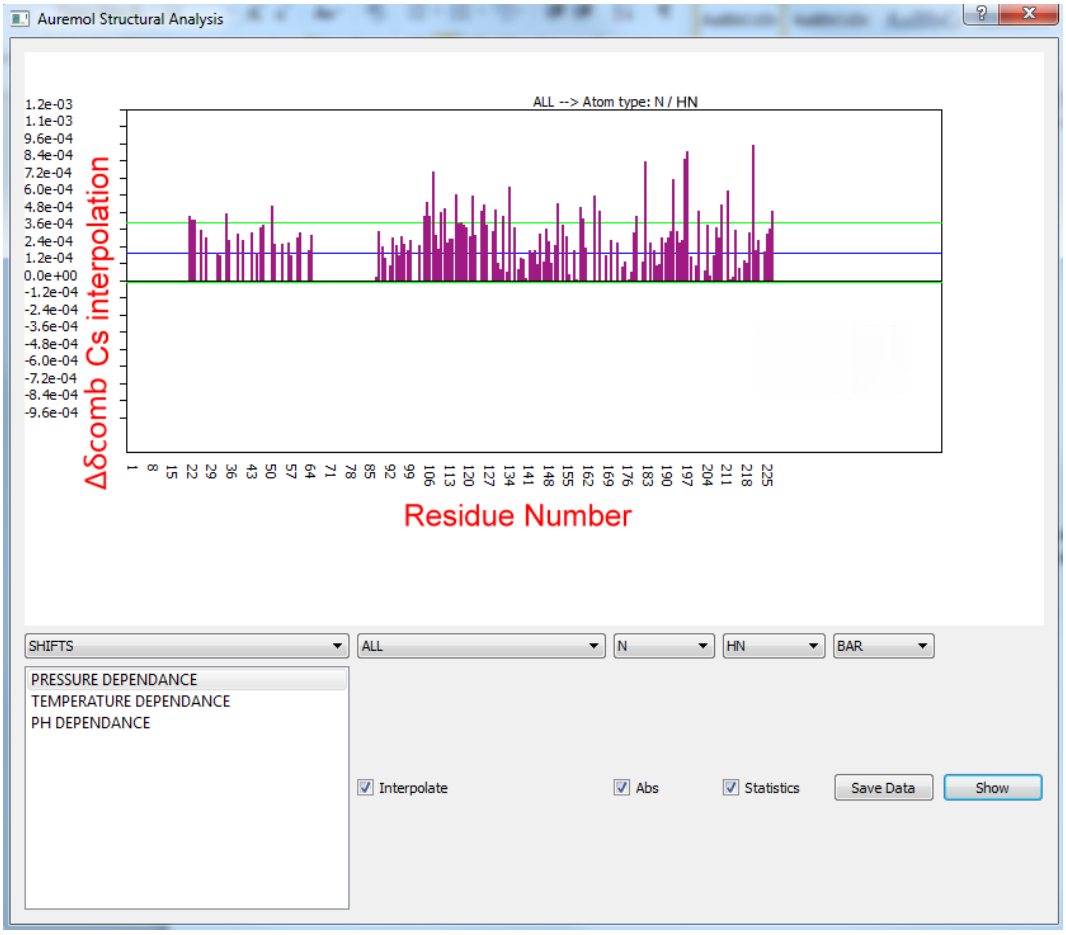


Fig. 5.20 The combined chemical shift dialog of the AUREMOL Structural Analysis of the human prion protein (*hu*PrP^{C}) with xenon binding: the dialog shows the combined [Schumann et al., 2007] chemical shift variations (N-H^{N}) with linear interpolation of the shift differences in absolute values between the reference and every test spectrum. Due to the linear interpolation the four bars obtained from the comparison are expressed by a unique one representing the slope computed from the linear interpolation of those four values. The mean and the double of the standard deviation of the dataset are plotted as a blue and a green line respectively.

The residues overcoming the double of the standard deviation correspond to those one having a low probability (smaller than 0.5), thus the structural variations reported in Fig. 5.14 and in Fig. 5.15 are detected in both cases, namely by the Bayesian probability and by the histogram analysis.

In Appendix D the list of the residues whose chemical shift and volume change is larger than the double of the standard deviation is reported. It is a sub-class of the list reported in Appendix C which describes the residues whose feature probability is lower than 0.6.

5.2.2.4 General results of the quality control

Beyond these detailed results and the visual investigation of the signals in the spectra, the user can obtain some general results including matching and mismatching ratios.

Once the whole quality test analysis has been performed an additional window appears showing to the user the main results, as depicted in Fig. 5.21.

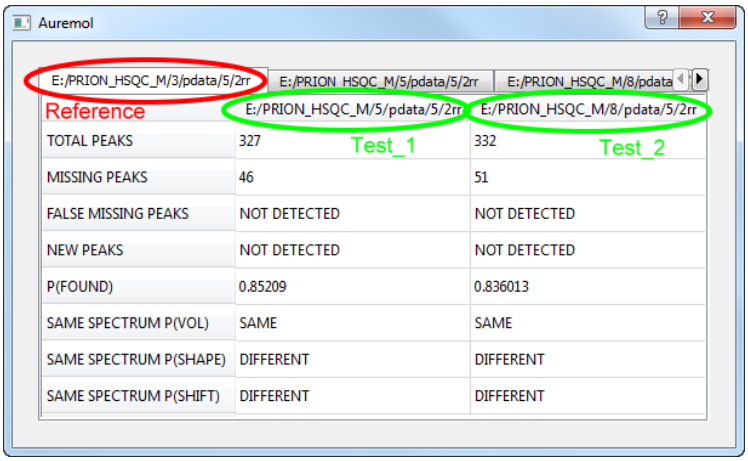


Fig. 5.21 Screenshot of the dialog containing the final results obtained from the human prion protein (*hu*PrP^{C}) with xenon binding: the dialog shows the comparison among all spectra of the human prion protein bound with xenon. The first tab widget (red ellipse) identifies the reference spectrum (without xenon) that is compared with all the other measurements. In particular, 46 peaks are missing comparing the reference spectrum with the first test case (including the not assigned ones). The rows named NEW PEAKS and FALSE MISSING PEAKS (or ambiguous) are used when comparing NOESY or TOCSY spectra. In the described comparison the matching ratio between the reference and the first test spectrum is about 85 %.

(P(FOUND)) The last three rows of the window show the results of the KS-Test (using the volume, the shape and the shift features). The quality test correctly rejects the hypothesis that the test spectra are the same respect to the reference one when the test is applied on the shape and on the shift.

The reference spectrum (experiment number three) is compared with all the other test spectra. As shown in Fig 5.21a list of results is obtained.
This list contains:

- The number of total picked peaks of the test cases (327 in the first and 332 in the second one).
- The number of encountered missing peaks between the reference and the test spectrum (46 and 51 in the reference spectrum compared with the first and with the second test case respectively). These reference peaks are those ones that have not been associated with any other peak in the compared test spectra.
- The number of false missing or ambiguous peaks (see paragraph 3.4.4.2). This information is computed only for NOESY, COSY and TOCSY spectra.
- The number of new peaks (see paragraph 3.4.4.2). It is possible to know this information only for NOESY, TOCSY and COSY spectra.
- The probability P(FOUND) that the compared spectra are the same based on the number of matching peaks (see eq. 3.27). The matching ratio between the high probability peaks of the first test and those ones of the reference spectrum is 85%, while it decreases to 83% when the second test case is compared with the reference spectrum.
- The result of the KS-test to determine if the compared spectra are the same or not based on different features. In accordance to the volume comparison the structures of the first two test cases possess the same conformation of the reference spectrum, while using the shape and the shift this assertion is correctly denied.

In particular the term P(FOUND) represents the signal matching ratio. Since HSQC spectra are compared, it also corresponds to the structural matching fraction, computed as

described in eq. 3.27. It defines the number of signals that overcome the Bayesian threshold of 0.6 with respect to the total number of peaks in every considered spectrum. The more the matching ratio approaches the one the more reference and test spectra are identical.

In addition, all the values computed from the quality test analysis are stored in an .xml file called "*qdata.xml*". This has been done for practical reasons allowing the user to close and reopen either the AUREMOL application or the investigated spectra without losing the results obtained from the quality test analysis.

5.2.3 Quality control of the high pressure and temperature datasets

The two dataset of HSQC-TROSY spectra (with varying pressure and temperature respectively) are loaded separately in the main interface setting all the parameters as reported in Fig. 5.3.

The high pressure dataset is made up of one reference (at 0.1 MPa and 293 K) and six test spectra. The volume of this dataset needs to be normalized since the number of scans of the test experiment at 50 MPa is 96, while all the others (reference and test cases) have been acquired using 48 scans. The receiver gain, the spectral width and the offset are unchanged among the spectra. The experiment acquired at 50 MPa has a different *NC_proc* with respect to the other spectra, compensating the different *NS*.

The temperature dataset encompasses one reference spectrum (at 293 K and 200 MPa) and four test spectra. The volume normalization is also required in this case since the *NS* of the test experiment at 313 K is 128, while 48 scans have been applied in all the other cases. The parameter *NC_proc* of this experiment is different with respect to the other spectra, whereas the *RG*, SW and OFF are unchanged in the whole dataset.

In the two investigated datasets, the reference and the test spectra were all previously assigned.

5.2.3.1 Quality control detailed results

As previously described general and detailed results are available after performing the quality routine.

In Fig. 5.22 the detailed analysis on the high pressure dataset of the residue Thr199 has been zoomed out. This dialog appears after positioning the cursor over the reference peak of interest and pushing the keyboard character “q”.

The peak Thr199 of the reference spectrum has been automatically associated with Thr199 peaks of the first four test spectra, while it has not been possible to detect any connection with the last two test cases (the peak disappears as reported in Fig. 5.2 part *a*). The shift, volume, shape, time cross-correlation and line width differences among Thr199 and the other test peaks is very large implying several low probability results (as reported in the red colored rows containing the probability of the shape, of the line width, of the correlation, of the volume and of the shift). The VOL OLD and VOL NEW rows reveal that the test volumes do not need to be normalized (see paragraph 3.4.2.3.2). It is evident that there is a remarkable volume decrement from the reference to the last test cases, while the chemical shift reveals large variations along both directions.

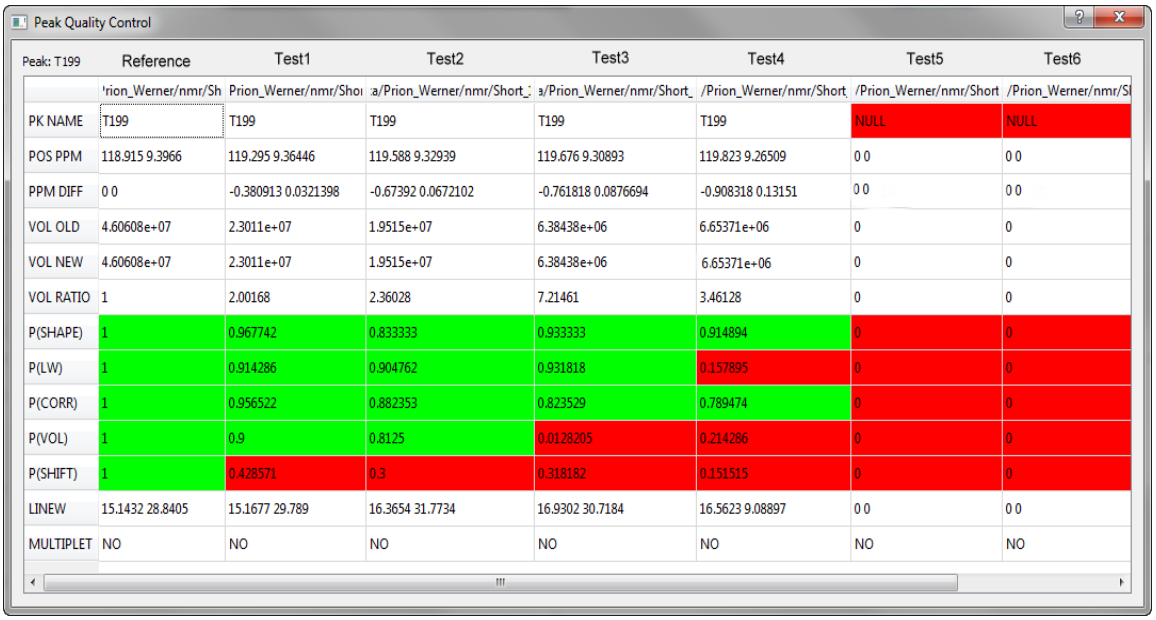


Fig. 5.22 The peak quality control dialog of the reference residue Thr199 (0.1 MPa and 293 K) showing variations of features (high pressure dataset of the human prion protein (*hu*PrP^{C})): the dialog shows red-colored boxes where the probability value of the investigated feature is smaller than 0.5 (as the line width probability of the fourth test case, the volume probability of the third and the fourth test spectra and the chemical shift probability of all the test spectra). The reference peak Thr199 disappears in the last two test cases thus it is not associated to any peak.

As shown in Fig. 5.2 the reference peak Thr199 disappears in the last two test cases thus it cannot be associated to any other peak in those spectra.

The detailed analysis has also been performed on the temperature dataset and it is reported in Fig. 5.23. The dialog has been obtained after pushing the character “q” in the keyboard and after positioning the cursor over the Gly123 reference peak.

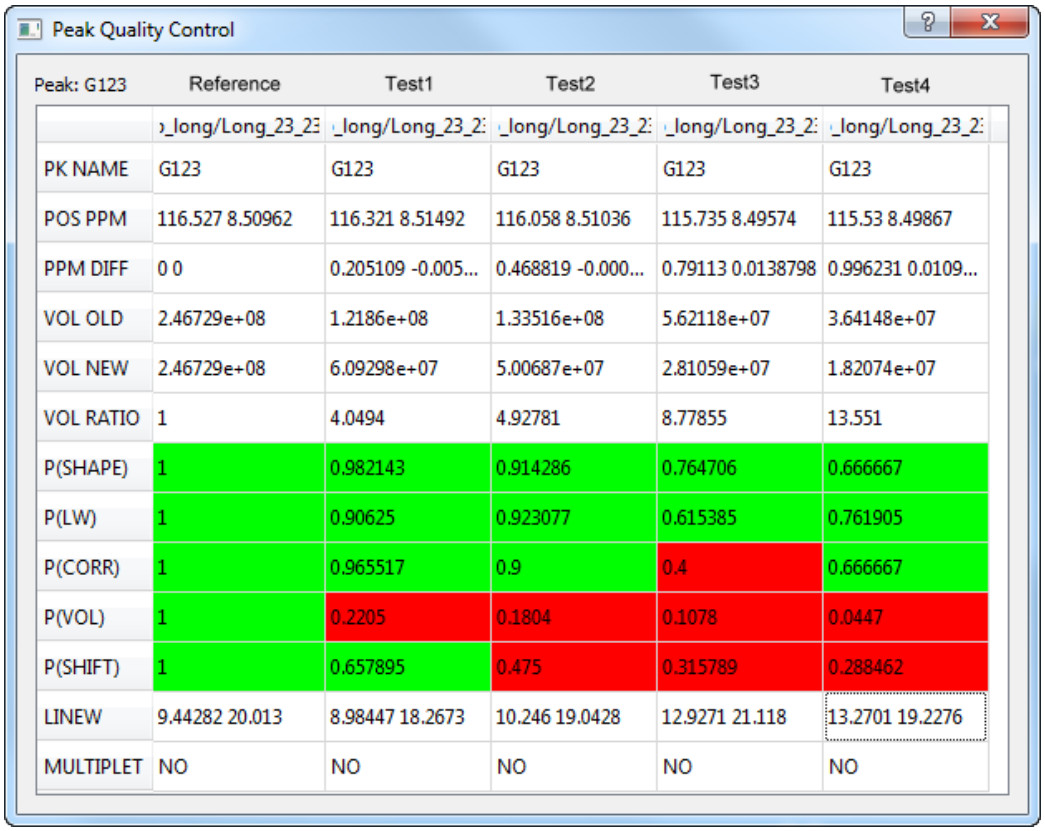


Fig. 5.23 The peak quality control dialog of the reference residue Gly123 (293 K and 200 MPa) showing variations of features (temperature variation dataset of the human prion protein (*hu*PrP^{C})): the dialog shows red-colored boxes where the probability value of the investigated feature is smaller than 0.5 (as the cross-correlation probability of the third test case, the volume probability of all the test spectra and the chemical shift probability of the last three test spectra). Increasing the temperature, the reference peak Gly123 experiences a very strong chemical shift variation and a remarkable volume decrement.

5.2.3.2 Bayesian feature analysis

Using the option “Show analyzed peaks” from the “Quality test” menu (see Fig. 4.1) it is possible to obtain graphical results as the appearance of colored boxes around the peaks having a low probability for the specific requested feature. In particular, if the option “All” in Fig. 5.8 is selected it is possible to see such boxes surrounding those peaks whose probability is lower than a selected value (e.g. 0.5) for all the features simultaneously. In Fig. 5.24 this analysis has been applied on the high pressure dataset and the following reference residues has shown a probability lower than 0.6: Met129, Ser132, Met134,

Ile138, Tyr145, Tyr150, Asn153, Val161, Tyr163, Asn173, Val176, His177, Ile182, Thr193, Asp202, Val203, Glu211, Met213, Ile215, Gln217, Glu219 and Glu221 (22 residues). In Fig. 5.25 the superimposition of all the features is reported. Blue, green, yellow, black and red boxes surround respectively peaks whose volume, line width, chemical shift, correlation and shape probabilities are lower than 0.5.

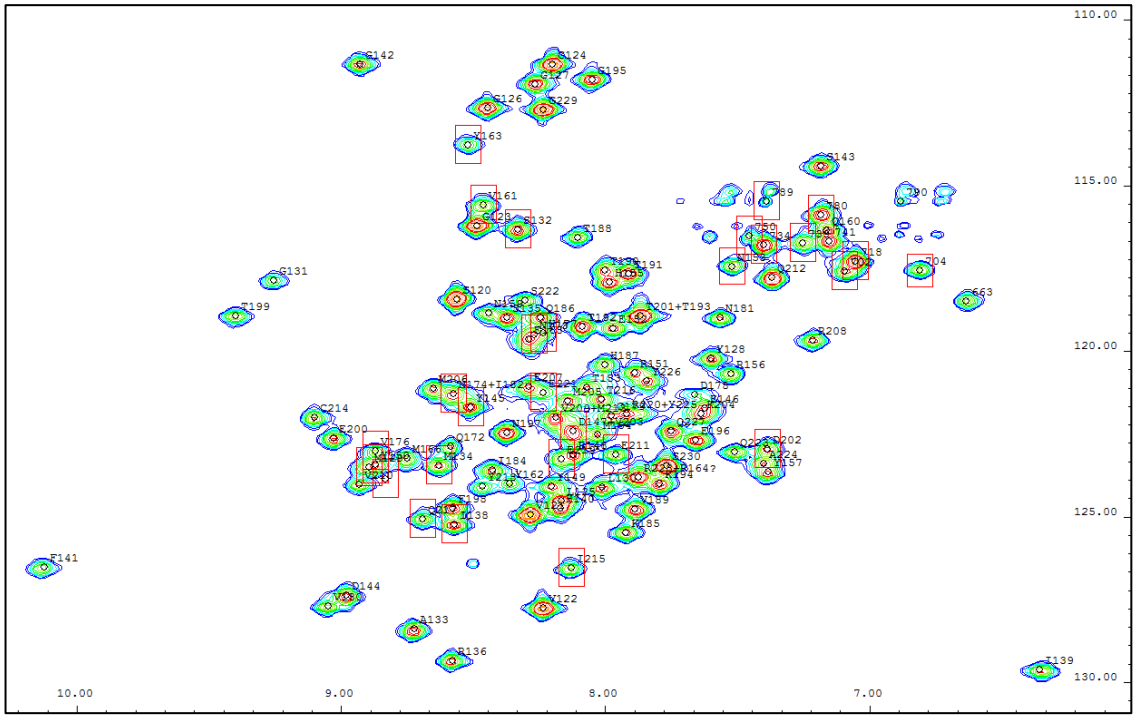


Fig. 5.24 Identification of the simultaneous variation of all the features of the reference spectrum of the human prion protein (*hu*PrP^{C}) (high pressure dataset) by means of the feature selection routine: the red boxes are showing the peaks whose probability is smaller than 0.6 for all the features simultaneously: Met129, Ser132, Met134, Ile138, Tyr145, Tyr150, Asn153, Val161, Tyr163, Asn173, Val176, His177, Ile182, Thr193, Asp202, Val203, Glu211, Met213, Ile215, Gln217, Glu219 and Glu221.

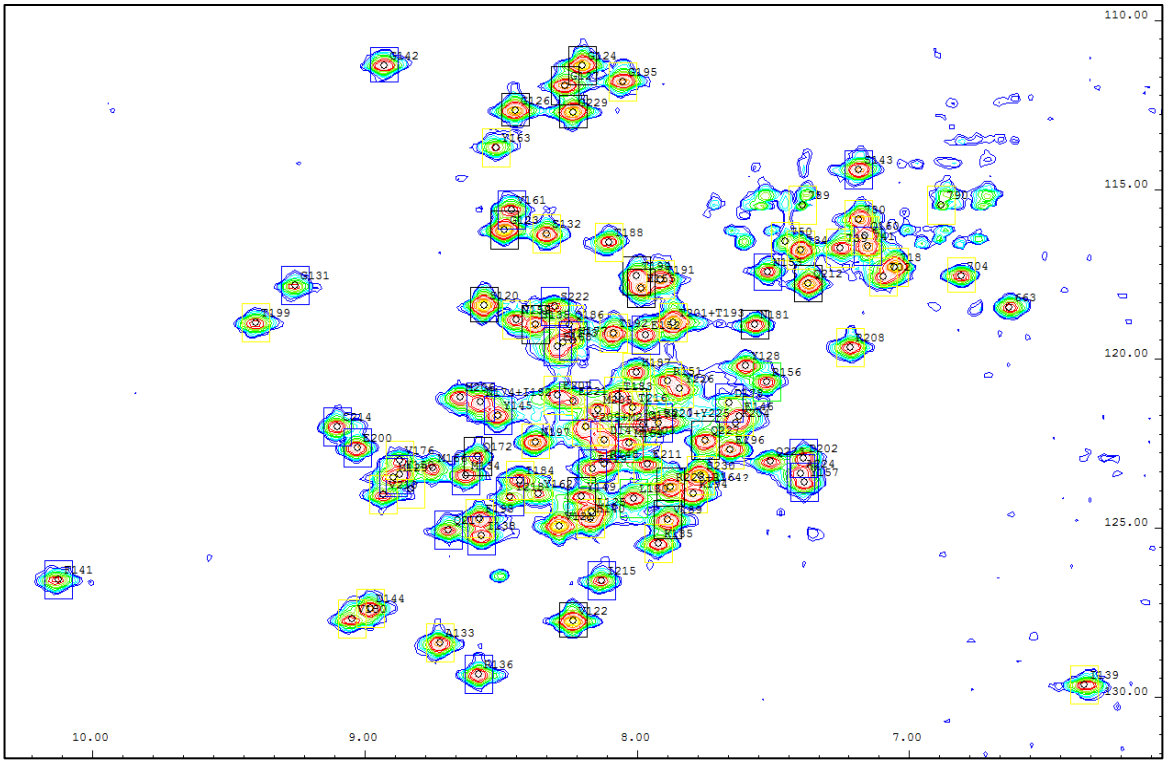


Fig. 5.25 Identification of all the peak features with a probability smaller than 0.5 in the reference spectrum of the human prion protein (*hu*PrP^{C}) (high pressure dataset): peaks of the reference spectrum are highlighted with different colors with respect to the investigated feature. Highlighted are peaks having a volume probability (blue boxes), a line width probability (green boxes), a chemical shift probability (yellow boxes), a cross-correlation probability (black boxes) and a shape-similarity probability (red boxes) smaller than 0.5.

In Appendix E the list of the residues whose feature probabilities are lower than 0.6 is reported. The names of the residues verifying this condition (in at least one of the test spectra) are reported according to the evaluated feature. In particular, 22 residues have a probability lower than 0.6 for all the possible features (red highlighted in Appendix E).

5.2.3.2.1 Peaks that have not been associated among the spectra (missing signals)

Some of the picked peaks of the reference spectrum disappear increasing the pressure, thus they become missing peaks. As shown in Fig. 4.1 those not associated or missing peaks can be visualized by the user via the option “Show missing peaks” from the “Quality test” menu. The missing signals detected in the reference spectrum appear surrounded by red boxes. The missing peaks of the high pressure dataset of the reference spectrum are red highlighted in Fig. 5.26: Gly142, Tyr163, Val161, Ser143, Gly131, Thr199, Cys214, Glu200, Asn174, Ile182, Asp178, Tyr128, Arg156, Lys204, Asp202, Ala224, Tyr157, Met166, Val176, Val210, Phe141, Ile139, R136 and Asp144.

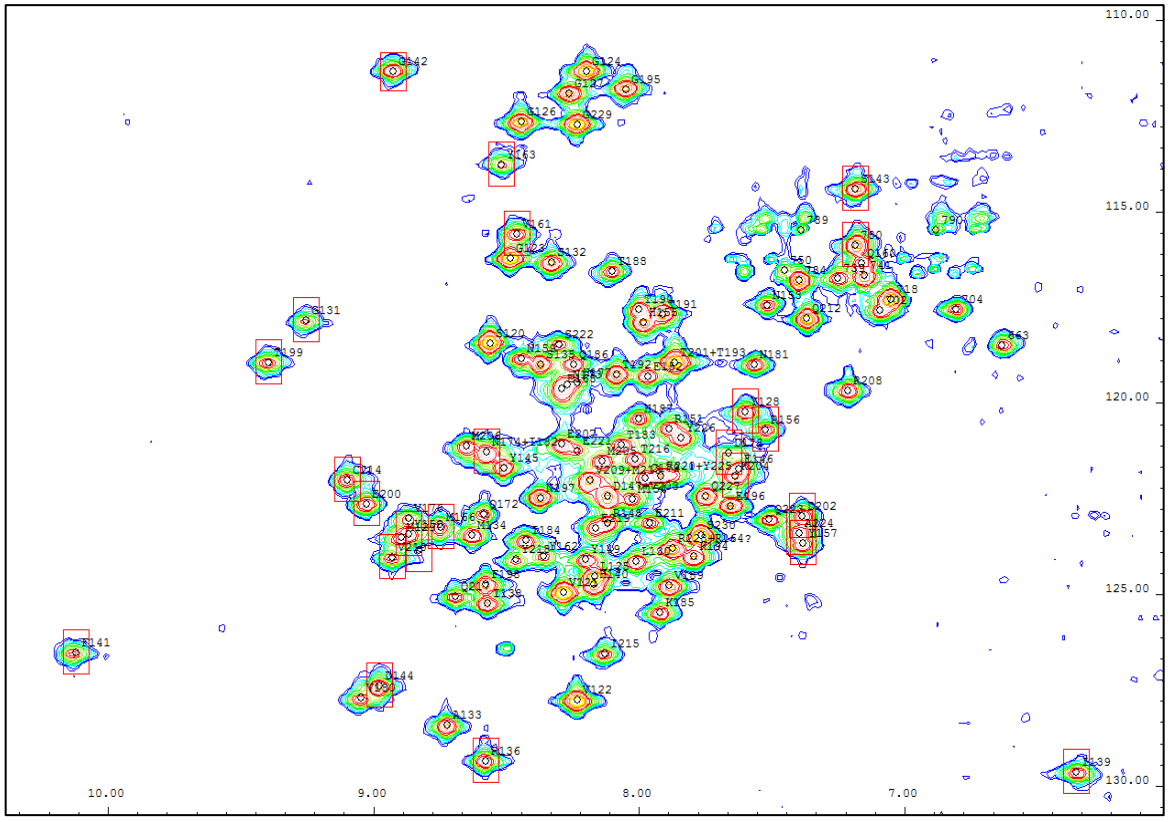


Fig. 5.26 Identification of disappearing (missing) peaks between the reference spectrum and all the test cases (high pressure dataset of the human prion protein (*hu*PrP^{C})): missing peaks of the reference spectrum are highlighted with red boxes: Gly142, Tyr163, Val161, Ser143, Gly131, Thr199, Cys214,

Glu200, Asn174, Ile182, Asp178, Tyr128, Arg156, Lys204, Asp202, Ala224, Tyr157, Met166, Val176, Val210, Phe141, Ile139, R136 and Asp144.

The missing peaks of the temperature dataset are red highlighted in Fig. 5.27:

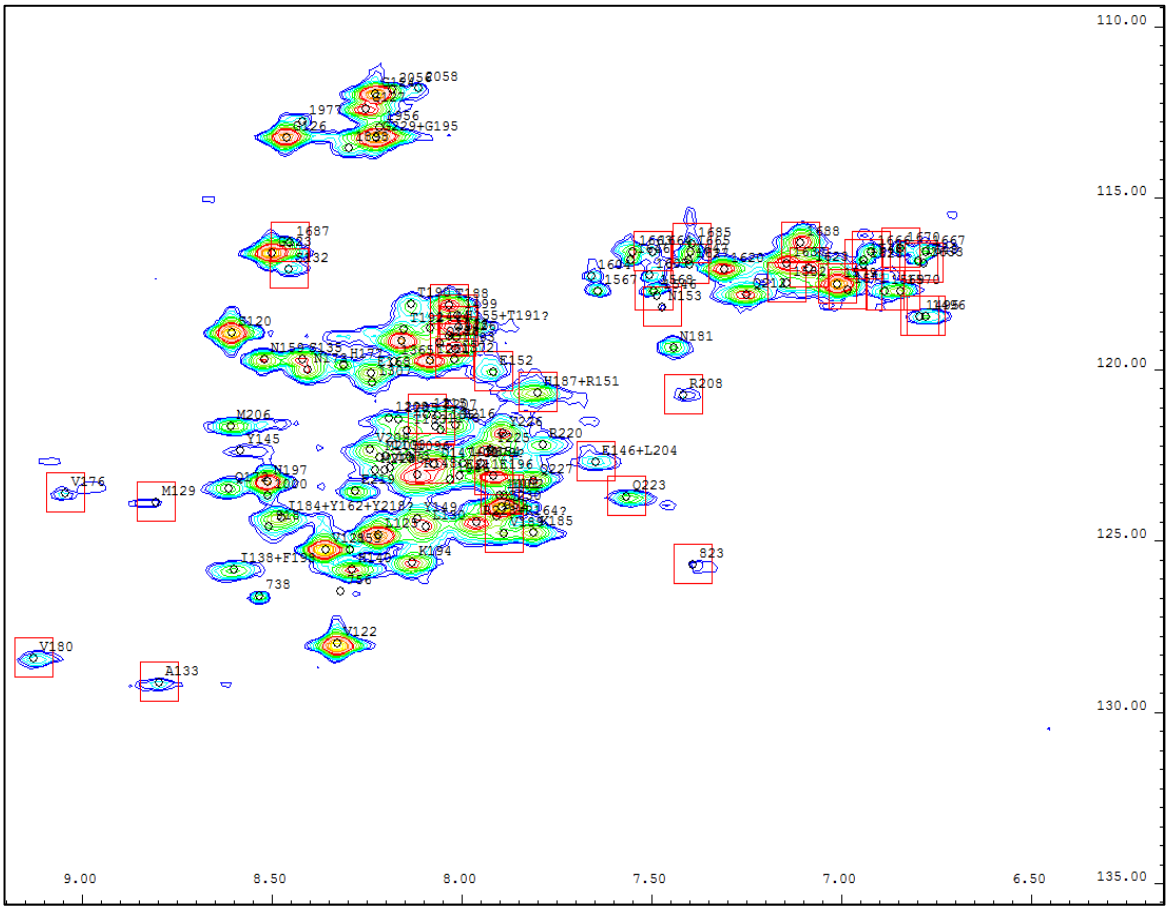


Fig. 5.27 Identification of disappearing (missing) peaks between the reference spectrum and all the test cases (temperature dataset of the human prion protein (*hu*PrP^{C})): missing peaks of the reference spectrum are highlighted with red boxes: Met129, Ser132, Ala133, Phe146, Asp147, Arg151, Glu152, Asn153, His155, Val176, Cys179, Val180, His187, Val189, Thr191, Thr193, Glu196, Leu204, Arg208 and Gln223.

5.2.3.3 Structural analysis by means of histograms

The structural analysis has been applied on the high pressure dataset whose combined chemical shift [Schumann et al., 2007] mapping is reported in Fig. 5.28.

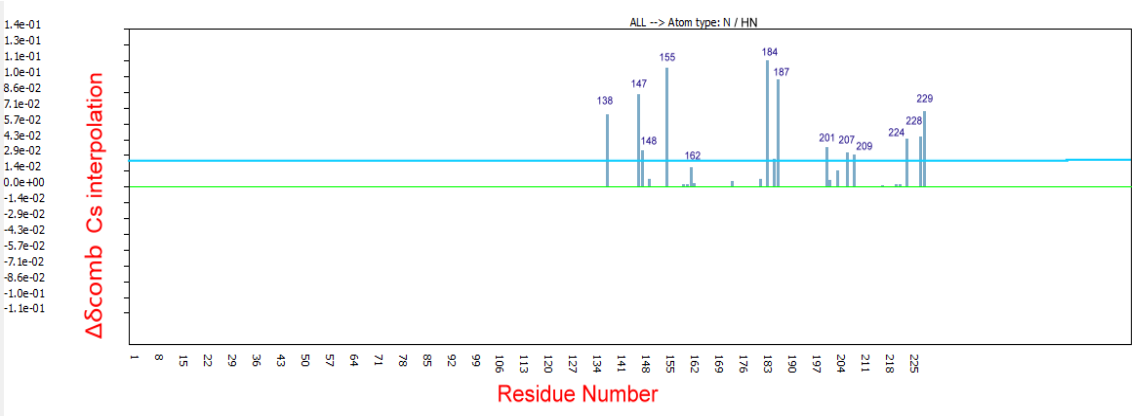


Fig. 5.28 The combined chemical shift dialog of the AUREMOL Structural Analysis of high pressure human prion protein: the dialog shows the combined [Schumann et al., 2007] chemical shift variations ($N\text{-}H^N$) with linear interpolation of the shift differences in absolute values between the reference and every test spectrum. Due to the linear interpolation the four bars obtained from the comparison are expressed by a unique one representing the slope computed from the linear interpolation of those four values. The double of the standard deviation is plotted as a cyan line. Only the residues with the strongest shift change are visible.

The bars represent the absolute slope of every linear interpolation (one for each residue). The double of the standard deviation is computed and plotted on the dialog as a cyan line. The residues whose absolute slope is greater than the double of the standard deviation are listed in another dialog.

The residues overcoming the double of the standard deviation represent a part of the residues having a low probability (smaller than 0.5), thus they are the most involved in the structural change (see Appendix E).

5.2.3.4 Quality control general results

As explained in paragraph 5.2.2.4 the user can obtain some general results including matching and mismatching ratios. The windows containing those results are reported in Fig.5.29 (part *a*) and in Fig.5.29 (part *b*) for the high pressure and the temperature case respectively.

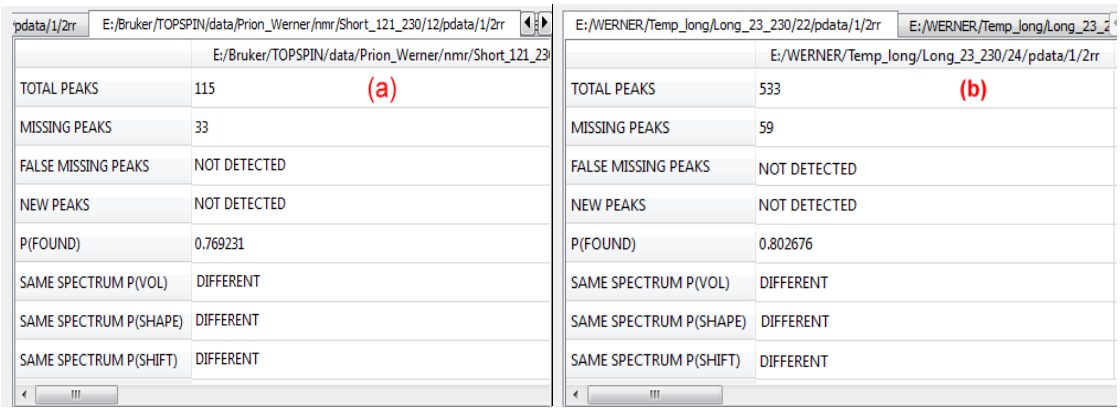


Fig. 5.29 Screenshot of the dialog containing the final results: the dialogs show the comparison between the reference and the test cases. (a) high pressure comparison with the test case at 50 MPa: total number of picked peaks in the test case, 115; missing peaks in the reference spectrum compared with this test case, 33; matching ratio, 76%; KS-test, not the same spectra based on the volume, shape and shift. (b) temperature comparison with the test case at 303 K: total number of picked peaks in the test case, 533; missing peaks in the reference spectrum compared with this test case, 59; matching ratio, 80%; KS- test, not the same spectra based on the volume, shape and shift.

The high pressure reference spectrum (experiment number five) is compared with all the other test spectra and the matching results of the comparison with the high pressure test case at 50 MPa are available at the end of the quality computation. As shown in Fig. 5.29 (part a), the total number of picked peaks in the test case is 115. The number of missing peaks in the reference spectrum compared with the first test case is 33. The matching ratio between the spectra corresponds to 76%. The KS-test reveals that they are not the same spectra based on the volume, shape and shift. The temperature reference spectrum (at 293

K) is compared with all the other test spectra and the final matching results with the temperature test case at 303 K are shown in the part b of Fig. 5.29. In particular, 533 peaks have been picked in the test case; 59 reference peaks are missing. The matching ratio is 80%. The compared spectra are not identical in accordance to the KS-test based on the volume, shape and shift features.

6

Conclusions and Discussions

6.1 General considerations

The NMR spectroscopy has the advantage to provide the most complete characterization of molecular structures in solution. The complex task of manually collecting and interpreting vicinity relationships changes (through space and bonds) in order to extract structural modifications of proteins is time demanding and requires skilled personal. For this reason, a wider and completely new approach has been described and developed in this work.

Similar methods are known in literature and they are generally applicable to one-dimensional NMR spectra (either ^{1}H or ^{13}C experiment type) to solve subsets of the vast quality control procedure. In addition, these techniques are typically usable for small molecules and not for the more complex protein structures [Golotvin et al., 2006; Thiele et al., 2011].

The new automated AUREMOL-QTA package has been developed for controlling the quality of a protein in a set of investigated spectra in order to automatically infer conformational changes from spectral modifications. The AUREMOL-QTA package eases the whole quality control procedure through a multi-dimensional data investigation.

The power of the developed method relies on a hybrid analysis (time-frequency domain) performed on the user provided data considering the presence of cross peaks showing a multiplet structure.

Noteworthy, several efforts have been done in this project to extend the standard quality control (usually performed for structural confirmation) to a more complex analysis induced by perturbations of the external conditions (e.g. temperature, pressure, pH and ligand binding). In particular, it is applicable to solve present-day center of interest topics as

protein-protein docking or protein-DNA binding. It surely facilitates the spectroscopist and the pharmaceutical industry through a totally automated structural computation.

As demonstrated in chapter four, the developed method has been successfully tested when dealing with partially or completely unassigned reference and test spectra and when comparing folded proteins with partially or totally denatured ones. In addition, the method has demonstrated a high reliability when comparing spectra after strong perturbations of the external conditions (see chapter five).

The AUREMOL-QTA package has been developed encompassing hundreds of computational methods (corresponding to a source code of many thousands of lines) and has been logically divided in four main stages:

1. The pre-processing
2. The low-level
3. The mid-level
4. The high-level

In the next paragraph each of these levels are discussed separately.

6.2 State of the art and future developments of the pre-processing stage

The pre-processing stage (par. 3.4.1) encompasses the automatic identification of peak classes (noise, baseline, solvent and resonances of interest), spectra standardization (e.g. spectral width, offset, number of scans, receiver gain, intensity scaling factor and general spectral shift management) and multiple referencing (more than one spectrum used as a reference candidate). It includes the new developed multi-dimensional methods for automatically suppressing solvents (AUREMOL-SSA) and for automatically correcting baseline (AUREMOL-ALS) distortions [Malloni et al., 2010; De Sanctis et al., 2011]. The user has to take particular care avoiding different water suppression techniques that could affect the quality analysis since each applied procedure would reveal or hide some resonances of interest. The most reliable comparison would be performed if the reference

and the test spectra had been treated with the same solvent suppression and baseline correction methods easing the mutual signal association task among the spectra.

Furthermore, the application of the same window functions to the investigated spectra along all measured spectral directions is particularly useful for facilitating the comparison. In the case that the reference and the test spectra have been processed with different window functions, it would be particularly useful to automatically detect and invert the effects of any filtering function. The produced time domain signals (that are free from window filtering effects) could be compared in every domain (both time and frequency domain) avoiding misinterpretation of peak features. However, some filters are theoretically not invertible (sine). In this case, the time domain data have to be re-processed, a task that is possible in AUREMOL since an nD-Fourier transformation was introduced by us. It would be useful to introduce into the AUREMOL software package additional window functions (to the already developed ones) as the trapezoidal, the sine squared and the traficante [Traficante & Nemeth, 1987].

6.3 State of the art and future developments of the low-level analysis

The most computational time demanding stage is the low-level (par. 3.4.2). This module associates peaks among spectra collecting all information (par. 3.4.2.1) related to every peak in each spectrum including an additional backtracking algorithm (par. 3.4.2.3.2). This latter is required in order to compare and to optimize all the possible peak connections between different spectra.

Each single modification of every evaluated peak is considered as an important feature. For example, it takes into account volume changes, line width variations, shape behavior, splitting of the peaks, changes of relative positions in all measured directions, the symmetrical properties in case of a NOESY or TOCSY experiment, the relative peak intensities and the cross-correlations (par. 3.4.2.5) in the time domain (through the newly developed AUREMOL-FFT). The core of the peak association algorithm has been developed considering the whole peak pattern, the peak surrounding, the shape similarity criterion, the number of peaks in the pattern, their shift variations and their normalized volumes (see par. 3.4.2.3.2.2).

Particular care has been taken while computing the line width information (par. 3.4.2.4). This calculation has been optimized for peak singlet and multiplet taking into account typical problems related to the peak overlapping.

If the user provides the three-dimensional structure of the investigated protein (.pdb file) the routine automatically back-calculates the desired spectrum according to all other experimental parameters using it as an additional reference spectrum. From this simulated spectrum all the important information related to the chemical shifts and the J couplings (performing an automatic search of peak multiplets) are used and matched with the experimental spectra.

As described in the paragraph 3.4.2.2, the multiplet research analysis (MSA) is performed in three main steps (the masking, the smoothing and the peak maxima adaption). In particular, the filtering procedure is performed through the Savitzky-Golay smoothing filter of the data. In case of antiphase absorptive cross-peaks (COSY experiment) the method could not smooth the data in a correct manner requiring a modification of the algorithm. In such case, it would be more appropriate to calculate the line width differences between the experimental and the simulated peaks using this information in order to adapt the simulation as much as possible to the experimental data. For example, each back-calculated cross peak could be adapted to the corresponding experimental reference peak leading to additional strong information for multiplet recognition.

In order to obtain a reliable peak association between different spectra the neighborhood information (par. 3.4.2.3.1) of each peak has been analyzed through a method that builds a neighborhood distance list (NLST). This list has been used in order to instantly have a list of neighbors for each considered peak and to have the maximal multidimensional distance between the considered peak and its neighborhood.

The automatically computed volume scaling factor plays an important role in the comparison of spectra. This volume scaling factor (see eq. 3.2 and eq. 3.14) is strongly related to the peaks that have been picked during the peak picking step. It is optimized avoiding critical situations where the volume scaling factor is computed using solvent peaks. This is avoided through the AUREMOL-SSA solvent removal method and through the computation of this score that has to be found in the first third part of the histogram

(see Fig. 3.21 and par. 3.4.2.3.2.1) containing volume ratios. In another extreme case, where the number of noise peaks is notably higher than the number of true peaks, the calculation of the volume scaling factor could not work properly. Also this limitation has been overcome using the previously developed peak-picking routine that avoids such effect [Antz et al., 1995].

Noteworthy, additional features could be included into the AUREMOL-QTA routine. For example, in the hybrid analysis, the information coming from the time-domain data (obtained through the inverse Fourier transformation) could be investigated with other methods as the variation of the Mutual Information [Cover et al., 1991] quantity, the Wavelets analysis [Norman, 1953] and the Hilbert-Huang transform [Huang et al., 1998]. Some tests have been conducted and the latter could be helpful to trace the separation among overlapping signals. In case of multidimensional NMR spectra, an exhaustive hybrid analysis could be conducted where all the 2^{DIM} domains are considered and exploited to collect as much information as possible. A mixed-domain analysis would be particularly effective in order to extract cross-knowledge.

6.4 State of art and future developments of the mid-level stage

In order to reveal data structures and to draw general conclusions, an interpretation of the computed results (stored as scores in the previous low-level) is needed. This is done through the mid-level stage (par. 3.4.3). The basic idea is to analyze all values from a score-like system (low-level) to a more interesting probability-like one (mid-level). The Bayesian analysis performed during the mid-level stage of this project computes several probabilities for each considered peak association. In particular, every peak connection is identified by a set of five probabilities (e.g. volume, line width, chemical shift variation, cross-correlation in the time domain and shape) and not by a unique probability that encompasses all the previously described features. This single probability could be used to have a unique measure of the peaks and consequent matching of spectra.

During this project, the Kolmogorov-Smirnov test (par. 3.4.3.5) has been used with the aim to univocally declare, in the error limit of five percent if the compared spectra are the same (e.g. measured spectra are representing the same protein). This test has been chosen

for its non-parametric nature but it would be possible to compare it with a set of different tests like the Mann-Whitney [Wilcoxon, 1945; Mann & Whitney, 1947] (U-test), the Friedman's test [Friedman, 1937] and the parametric (requiring the normality assumption) two-tailed t-test [Student, 1908]. This multiple statistical comparison increases the reliability of the null hypothesis statement. In case of statistical divergence the AUREMOL-QTA user should be informed with a set of warning messages.

6.5 State of the art and future developments of the high-level analysis: investigating structural changes

The high-level stage (par. 3.4.4) facilitates the spectroscopic analysis and it is used to obtain structural overviews from the considered dataset of spectra. After computing spectral matching ratios and statistical probabilities (mid-level stage), it is possible to define if the structures represented by the investigated spectra are identical or similar and in which percentage. The high-level stage has been developed in order to express spectral properties into molecular features.

The high-level stage has been differently developed depending on the type of considered experiment. In particular, the routine follows two different approaches:

1. dealing with HSQC-type spectra, it evaluates the average variation of the user selected feature (e.g. volume and chemical shift) of all the resonances (residues). In addition, it identifies those ones whose changes exceed the double of the standard deviation computed over the entire dataset generating a list of most changed residues from a set of histograms.
2. dealing with NOESY and TOCSY spectra, before creating the list of most altered resonances, it evaluates the global symmetry properties of the spectrum. The additional residue pattern recognition of the peaks classified as new, missing and ambiguous is performed in order to verify if the typical pattern of a certain residue has moved somewhere else in the spectrum, even with a remarkable volume variation.

During this project many comparisons have been performed among different classes of spectra (reference or test), both assigned (partially of totally) and unassigned showing a good reliability.

Typically, the user might furnish an assigned reference spectrum in order to verify how much and where exactly it differs with respect to any other possible unassigned test dataset yielding meaningful results.

In addition, the proposed routine consents a comparison of totally unassigned reference and test spectra. In this case the quantitative statement coming from the Bayesian analysis is still reliable but the feature mapping (by means of histograms) cannot be interpreted as a structural modification due to the lack of a direct correspondence signals-residues. Therefore, the method produces ppm positions instead of residue lists.

When dealing with HSQC-type spectra the eq. 3.27 can be used since almost each signal in the spectra corresponds to a residue in the structure. The matching and mismatching ratios of the spectral signals thus represent altered and unaltered fraction of the three-dimensional structure.

If NOESY, TOCSY and COSY spectra are considered, the eq. 3.27 (a matching factor of all the previously described features simultaneously) is no more applicable and it is modified as described in eq. 3.31. This latter represents a matching or a mismatching ratio between all the peaks in the reference and in the test spectra, but it does not directly correspond to a fraction of structural variations. Supposing that the reference dataset is a NOESY or a TOCSY spectrum (i.e. the native folded and the H15A mutant HPr from *Staphylococcus aureus*), without previous assignment, the matching ratio of patterns could be computed as follows:

$$F_{HP} = \frac{PA_{HP}}{PA} \tag{6.1}$$

where PA represents the total number of patterns and PA_{HP} defines the numbers of patterns where more than the 80% of their peaks belongs to the high probability class. In this manner it would be possible to determine the amount of altered and unaltered three-dimensional structure without any assignment. Obviously, if the spectra have been

acquired on the same protein with very similar external conditions, the unaltered and the altered structural fractions must approach respectively the one and the zero.

Comparing NOESY or TOCSY reference spectra previously assigned the amount of structural modification could be determined more precisely. The matching ratio of patterns could be computed in the following manner:

$$F_{HP}= \frac{1}{R}\frac{\sum_{k=1}^{R}\sum_{i,j} P_{HP_i}(PA_j)}{\sum_{k=1}^{R}\sum_{i,j} P_i(PA_j)} \tag{6.2}$$

In eq. 6.2 the term *k* defines the total number of residues of the investigated protein, *j* represents the number of patterns of each residue and *i* is the number peaks encountered along each pattern. In particular, P_{HP_i} represents the number of high probability peaks along a specific pattern, PA_j are the considered patterns for each residue and P_i are all the peaks encountered along the analyzed residue patterns.

In case of a previously assigned reference spectrum, the result of eq. 6.2 is definitively more reliable with respect to eq. 6.1, since the patterns of the residues are well-defined.

As reported in par. 3.4.4.3, the histograms may contain the slope of every linear interpolation computed on the differences of a specific user selected feature. The dialog reported in Fig. 5.17 could be modified (option h) including additional interpolation types (e.g. quadratic and polynomial ones). This enhancement would facilitate the time demanding investigation of non-linear behavior of the feature of interest without using external programs for such purposes.

The lack of bars in the histograms is due either to missing association of peaks between the reference and the test spectra or to the invariability of the considered feature through some specific peaks in the spectra. A colored highlighted histogram would provide meaningful information to distinguish the previously described cases.

A very interesting application of the quality algorithm is the identification and quantification of spectral changes induced by binding of small molecules (drugs).

Ligand binding affects the chemical environment of some residues. The exchange process can be slow, intermediate or fast involving different spectral variations. Slow exchange processes provide intensity reduction of the cross peak and the observation of a

new signal in a different position. The computed slope (obtained analyzing the signal from the reference to the test spectra) could be used to automatically identify the new arising signal conducting an inverse interpolation from the last test spectrum to the reference one (e.g. in the HSQC experiments pressurized with xenon). Intermediate exchange determines line broadening, thus a consequent decrease of the signal intensity that is not accompanied by the appearance of a new cross peak. The analysis of the line width variation could be conducted by means of histograms. In case of fast exchange the cross peak is observed as split in two signals (the free and the bounded states) with a chemical shift variation. The one-to-one peak association between the reference and test spectra could be modified in order to obtain a one-to-many peak connection considering that the volume of the split peaks must be conserved with respect to the original reference signal.

7

Appendices

7.1 The Levenberg-Marquardt algorithm

The Levenberg-Marquardt [Levenberg, 1944] method has been chosen as the standard optimization routine. In order to optimize the fitting of Gaussian functions the first and the second derivative has been computed as following:

$$f(x) = \sum_{i=1}^{K} A_i e^{\left[-\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right]} \tag{7.1}$$

It represents the Gaussian function describing a typical peak's shape where A_i corresponds to the amplitude of the signal, μ_i is the ppm position of the peak center and σ_i is the width. In addition, *K* is the number of Gaussians used to perform the fitting (i.e. two if it is a peak doublet).

$$f'(x) = \frac{2A\left[-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right]e^{\left[-\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right]}}{\sqrt{2}\sigma} \tag{7.2}$$

$$f''(x) = \frac{2A\left[-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right]^2 e^{\left[-\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right]}}{\sqrt{2}\sigma} \tag{7.3}$$

The full width at half maxim (FWHM) of the Gaussian function is computed as following

$$\Delta v_{\frac{1}{2}} = \sigma_i \, 2 \left[2log \left(\frac{1}{0.5} \right) \right]^{0.5} \tag{7.4}$$

In case of Lorentzian line shapes the following optimization has been used:

$$f(x) = \sum_{i=1}^{K} A_i \frac{\sigma^2}{(x-\mu)+\sigma^2} \tag{7.5}$$

$$f'(x) = \frac{2A\sigma^2(x-\mu)}{((x-\mu)^2+\sigma^2)^2} \tag{7.6}$$

$$f''(x) = \frac{2A\sigma(x-\mu)^2}{((x-\mu)^2+\sigma^2)^2} \tag{7.7}$$

7.2 The Welch's t-test

The Welch's t-test [Welch, 1938] is a modification of the Student's t-test. The former is used in cases where the variances of the samples are possibly not equal. The statistic *t* is computed as following:

$$t = \frac{sample\ mean_1 - sample\ mean_2}{\sqrt{\frac{variance_1}{sample\ size_1} + \frac{variance_2}{sample\ size_2}}} \tag{7.8}$$

and the parameter representing the degrees of freedom is obtained as following:

$$df = \frac{\left(\frac{variance_1}{sample\ size_1} + \frac{variance_2}{sample\ size_2} \right)^2}{\frac{\left(\frac{variance_1}{sample\ size_1} \right)^2}{sample\ size_1 - 1} + \frac{\left(\frac{variance_2}{sample\ size_2} \right)^2}{sample\ size_2 - 1}} \tag{7.9}$$

7.3 Appendix C

FEATURE	shape	cross-correlation	line width	chemical shift variation	volume
Residue/ Experiment	**K23 (5/8/11/14)**	**K23 (5/11/14)**	**K23 (8/14)**	**K23 (11/14)**	**K23 (5/11)**
				R25 (11)	
			K27 (14)		
				S36 (11)	
	Y49 (5/8/11/14)	Y49 (8/11/14)		Y49 (14)	
				Q52 (11)	
	G53 (5/11/14)	G53 (8/11/14)			
					G56 (14)
				T95 (8)	
			H96 (14)		
		T107 (11/14)		T107 (11/14)	
				N108 (14)	
	M109 (11/14)	**M109 (11/14)**	**M109 (5/8/11/14)**	**M109 (11/14)**	**M109 (5/8/11/14)**
			K110 (5/14)	K110 (11)	K110 (5)
				A113 (11)	
				A117 (11/14)	
				G119 (11)	
				G123 (11/14)	
				G126 (11/14)	
			G127 (5/8/11/14)	G127 (11/14)	

				L130 (11)	
			G131 (5)	G131 (5/8/11/14)	
			M134 (11)		
			R136 (5/14)	R136 (14)	
			I138 (14)	I138 (11)	
				I139 (14)	
	H140 (5/8/11/14)	H140 (8/11/14)	H140 (5/11/14)		
					F141 (11)
	Y145 (5/8)	Y145 (8)	Y145 (5/14)	Y145 (11)	
				E146 (8)	E146 (5)
	R148 (5/8/11/14)	R148 (8/11/14)	R148 (8/14)		R148 (11)
	Y14 9(14)		Y149 (14)		
			Y150 (14)	Y150 (11/14)	Y150 (14)
			R151 (14)	R151 (11)	
			N153 (14)	N153 (14)	
			R156 (14)	R 156 (11)	R156 (11)
			N159 (5)		
	V161 (5/8/11/14)	**V161 (5/8/11/14)**	**V161 (11/14)**	**V161 (5/8/11/14)**	**V161 (14)**
				Y162 (11)	
			Y163 (5/8/11/14)		
			M166 (11)	M166 (11/14)	M166 (11)
				S170 (5)	
			Q172 (14)		
			N174		N174

			(5/8/11/14)		(5/14)
			V176 (8/14)	V176 (5/8/14)	V176 (14)
	H177 (8/14)	H177 (8/11/14)			H177 (14)
		D178 (11)	D178 (5/8/11/14)	D178 (8)	D178 (5/11/14)
	C179 (5/8/11/14)	C179 (5/8/11/14)	C179 (5/8/11/14)		
			V180 (14)	V180 (8)	
			N181 (14)	N181 (11/14)	
			T183 (5/14)	T183 (11)	
	I184 (8)	I184 (8/11)	I184 (14)	I184 (8/11/14)	
	Q186 (5/8/11/14)	Q186 (5/8/11/14)	Q186 (11)		
			T188 (14)	T188 (11)	
			T191 (14)		
				T192 (11)	
			K194 (5/11/14)	K194 (5/8/11/14)	
			E196 (5)	E196 (11)	
	N197 (14)	N197 (14)			
			F198 (5)	F198 (11/14)	
	T199 (14)		T199 (11/14)	T199 (8/14)	T199 (11)
		V203 (11)	V203 (5/8/11/14)	V203 (5/8/11/14)	V203 (5/11/14)
			M206 (5/14)	M206 (11)	
	E207 (5/11/14)	E207 (11/14)			
		V209 (14)	V209 (5)	V209 (5/14)	
	V210 (14)	**V210 (14)**	**V210 (5/8/14)**	**V210 (14)**	**V210 (14)**
	E211 (5/8)			E211 (11)	

	M213 (5/8/11/14)	**M213 (5/8/11/14)**	**M213 (5/8/11/14)**	**M213 (11/14)**	**M213 (5/11)**
			C214 (5)	C214 (5/11)	
			I215 (5/11/14)		I215 (14)
			T216 (5/14)	T216 (8/11)	
			Q 217 (5/14)		
			E219 (5/8/11/14)		
			R220 (5/8/11/14)	R220 (5/11)	R220 (14)
			E221 (5/8/11/14)		E221 (5/8/11)
	S222 (14)	**S222 (14)**	**S222 (14)**	**S222 (11/14)**	**S222 (14)**
			Q223 (8/14)	Q223 (11)	Q223 (8)
			A224 (11/14)		A224 (11/14)
				G229 (11/14)	
	20 residues	20 residues	53 residues	55 residues	26 residues

7.4 Appendix D

FEATURE	chemical shift variation	volume variation
RESIDUE	K23	K23
	S36	
	Q52	
		G56
	T107	
	N108	
	M109	M109
		K110
	A113	
	A117	
	G123	
	G126	
	G127	
	G131	
	G136	
		F141
	E146	E146
		Y150
	R151	
	N153	
	V161	V161
	M166	M166
		N174
		V176
		H177
	D178	D178
	N181	
	I184	
	K194	
	F198	
		V203
	V210	V210
	E211	
	M213	M213
		I215
		R220
		E221

	S222	**S222**
		Q223
		A224
	G229	
	28 residues	22 residues

7.5 Appendix E

FEATURE	chemical shift variation	volume variation	line width	cross-correlation	shape
RESIDUE				A120	
				V122	
				G123	
			G124	G124	
				G126	
				G127	G127
	M129	**M129**	**M129**	**M129**	**M129**
			L130		
		G131	G131		
	S132	**S132**	**S132**	**S132**	**S132**
	A133	A133	A133		
	M134	**M134**	**M134**	**M134**	**M134**
			S135	S135	
		R136	R136		
	I138	**I138**	**I138**	**I138**	**I138**
	I139	I139	I139		
	H140			H140	H140
		F141	F141		
		G142	G142		
		S143	S143		
		A144	A144		
	Y145	**Y145**	**Y145**	**Y145**	**Y145**
	E146	E146	E146	E146	
			D147	D147	D147
	R148	R148		R148	
			Y149	Y149	Y149
	Y150	**Y150**	**Y150**	**Y150**	**Y150**
	R151			R151	R151
		E152		E152	E152
	N153	**N153**	**N153**	**N153**	**N153**
					M154
				H155	
			R156		
		Y157	Y157	Y157	Y157
	N159			N159	
	Q160	Q160		Q160	Q160

	V161	**V161**	**V161**	**V161**	**V161**
	Y162	Y162	Y162	Y162	
	Y163	**Y163**	**Y163**	**Y163**	**Y163**
				R164	
		E168		E168	E168
				Q172	
	N173	**N173**	**N173**	**N173**	**N173**
		N174	N174	N174	N174
	V176	**V176**	**V176**	**V176**	**V176**
	H177	**H177**	**H177**	**H177**	**H177**
		D178	D178	D178	D178
	C179			C179	C179
	V180		V180		
				N181	
	I182	**I182**	**I182**	**I182**	**I182**
	T183			T183	T183
	I184	I184	I184		
	K185			K185	K185
	Q186			Q186	Q186
	H187			H187	
	T188				
			V189	V189	V189
			T190	T190	
	T191		T191		
	T192	T192		T192	T192
	T193	**T193**	**T193**	**T193**	**T193**
	K194			K194	K194
	G195			G195	G195
	E196			E196	
	N197			N197	
	F198	F198		F198	
	T199	T199	T199		
		E200			
	T201				
	D202	**D202**	**D202**	**D202**	**D202**
	V203	**V203**	**V203**	**V203**	**V203**
	M205	M205	M205	M205	
	E207		E207	E207	
	R208	R208			
			V209	V209	
	V210	V210	V210	V210	

	E211	**E211**	**E211**	**E211**	**E211**
				Q212	
	M213	**M213**	**M213**	**M213**	**M213**
		C214			
	I215	**I215**	**I215**	**I215**	**I215**
				T216	
	Q217	**Q217**	**Q217**	**Q217**	**Q217**
		Y218	Y218	Y218	Y218
	E219	**E219**	**E219**	**E219**	**E219**
	R220			R220	
	E221	**E221**	**E221**	**E221**	**E221**
		S222		S222	
		A224		A224	A224
		Y225		Y225	Y225
	Y226			Y226	Y226
				Q227	
				R228	
				G229	
	S230			S230	
	55 residues	52 residues	52 residues	76 residues	46 residues

Bibliography

[Alder et al., 1957] Alder, B. J. and Wainwright, T. E., Phase transition for a hard sphere system, *J. Chem. Phys.* **27** 1208 (1957).

[Antz et al., 1995] Antz,C., Neidig,K.-P., & Kalbitzer,H.R. 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 (1995).

[Arnold et al., 2002] Arnold MR, Kremer W, Ludemann HD, Kalbitzer HR: 1H-NMRparameters of common amino acid residues measured in aqueous solutions of the linear tetrapeptides Gly-Gly-X-Ala at pressures between 0.1 and 200 MPa. *Biophys Chem* **96**:129-140 (2002).

[Arnold et al., 2003] Arnold MR, Kalbitzer HR, Kremer W. High sensitivity sapphire cells for high pressure NMR spectroscopy on proteins. *J Magn Reson* **161**:127-131 (2003).

[Aue et al., 1976] Aue, W. P.; Bartholdi, E. Ernst, R. R. Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *Journal of Chemical Physics*, Volume **64**, Issue 5, pp. 2229-2246 (1976).

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

[Berendsen et al., 1984] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. "Molecular-Dynamics with Coupling to an External Bath". *Journal of Chemical Physics* 81 (8): 3684–3690 (1984).

[Berendsen et al., 1995]Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: a message-passing parallel molecular dynamics implementation. *Comp. Phys. Commun.* 1995;**91**:43–56 (1995).

[Bourne et al., HR 1990] Bourne HR, Sanders DA, McCormick F. The GTPase superfamily: a conserved switch for diverse cell functions. *Nature* **348** 125-32 (1990).

[Bragg, 1907] Bragg WH."The nature of Röntgen rays". *Transactions of the Royal Society of Science of Australia* **31**: 94 (1907).

[Briggs & Henson, 1995] W. L. Briggs & V. E. Henson: *The DFT: An Owner's Manual for the Discrete Fourier Transform*, SIAM (1995).

[Brünger et al., 1998] Brünger A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse–Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr D Biol Crystallogr*. **54**, 905-921 (1998).

[Cavanagh et al., 1996] Cavanagh,J., Fairbrother,W.J., Palmer III,A.G., & Skelton,N.J. Protein NMR Spectroscopy Principles and Practice. *Academic Press Inc.*, San Diego (1996).

[Chothia et al., 1986] Chothia C and Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO* **J5**:823–6 (1986).

[Clore et al., 1998] Clore, G.M., Gronenborn, A.M. & Tjandra, N. Direct refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. *J. Magn. Reson.* **131**, 159-162 (1998).

[Cornell et al., 1995] Cornell WD, Cleplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**:5179–5197 (1995).

[Cornfield, 1967] Cornfield, *J. Rev. Int. Statist. Inst.,* **35**, 34-49 (1967).

[Cornfield, 1969] Cornfield, *J. Biometrics*, **25**, 643-657 (1969).

[Cover et al., 1991] Cover, T.M. and Thomas, J.A. Elements of information theory. *John Wiley & Sons*, New York, NY (1991).

[Day & Garcia, 2008] Day, R., and Garcìa, A. E. Water penetration in the low and high pressure native states of ubiquitin. *Proteins: Struct., Funct., Bioinformatics.* **70**, 1175–1184 (2008).

[De Laplace et al., 1951] De Laplace P. S., A Philosophical Essay on Probabilities, (transl.) Dover, New York (1951).

[De Sanctis et al., 2011] S. De Sanctis, W.M. Malloni, W.Kremer, A.M. Tomé, E.W.Lang, K.P.Neidig, H.R. Kalbitzer, Singular spectrum analysis for solvent artifact removal from one-dimensional NMR spectra. *J.Magn. Reson.* **210** (2): 177-183 (2011).

[Delsuc, 1988] Delsuc M. A., "Spectral Representation of 2D NMR Spectra by Hypercomplex Numbers" *JMR* **77**, II 19-124 (1988).

[Dembowski et al., 1994] Dembowski, N. J., and E. R. Kantrowitz. The use of alanine scanning mutagenesis to determine the role of the N-terminus of the regulatory chain in the heterotropic mechanism of Escherichia coli aspartate transcarbamoylase. *Protein Eng.* **7**:673-679 (1994).

[Fischer, 1903] Fischer E."Synthese von Polypeptiden". *Berichte der deutschen chemischen Gesellschaft* **36** (3): 2982–2992 (1903).

[Friedman, 1937] Friedman, M. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance". *Journal of the American Statistical Association* (1937).

[Gengying et al., 1999] Gengying L. and Haibin X. Digital quadrature detection in nuclear magnetic resonance spectroscopy. *Rev. Sci. Instrum.* **70**, 1511 (1999).

[Geyer et al., 1995] Geyer,M., Neidig,K.-P., & Kalbitzer,H.R. Automated Peak Integration in Multidimensional NMR Spectra by an Optimized Iterative Segmentation Procedure. *J. Magn Reson.* **B 109**, 31-38 (1995).

[Gibson et al., 1960] Gibson, J. B., Goland, A. N., Milgram, M., and Vineyard, G. H., Dynamics of radiation damage, *Phys. Rev.* **120** 1229 (1960).

[Glaser & Kalbitzer, 1987] Glaser, S. & Kalbitzer, H. R. Automated Recognition and Assessment of Cross Peaks in Two-Dimensional NMR Spectra of Macromolecules. *J. Magn. Reson.* **74** , 430-436 (1987).

[Golotvin et al., 2006] Golotvin SS, Vodopianov E, Lefebvre BA, Williams AJ, Spitzer TD. Automated structure verification based on 1H NMR prediction. *Magn. Reson. Chem.* **44**: 524 (2006).

[Görler et al., 1997] Görler,A. & Kalbitzer,H.R. Relax, a Flexible Program for the Back Calculation of NOESY Spectra Based on Complete-Relaxation-Matrix Formalism. *J. Magn Reson.* **124**, 177-188 (1997).

[Görler et al, 1999a] Görler,A., Gronwald,W., Neidig,K.P., & Kalbitzer,H.R. Computer assisted assignment of 13C or 15N edited 3D-NOESY-HSQC spectra using back calculated and experimental spectra. *J. Magn Reson.* **137(1)**, 39-45 (1999).

[Görler et al., 1999b] Görler,A., Hengstenberg,W., Kravanja,M., Beneicke,W., Maurer,T., & Kalbitzer,H.R. Solution Structure of the Histidine-Containing Phosphocarrier Protein from Staphylococcus carnosus. *Appl. Magn. Reson.* **17**, 465-480 (1999).

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

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

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

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

[Güntert, 1998] Güntert, P. Structure calculation of biological macromolecules from NMR data. *Q. Rev. Biophys*. **31**, 145–237 (1998).

[Havel, 1991] Havel TF."An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by NMR spectroscopy" *Prog. Biophys. molec. Biol*. **56**:43-78 (1991).

[Huang et al., 1998] Huang, N. E., Shen, Z., Long, S. R., Wu, M. L. Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. and Liu, H. H. The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proceedings of the Royal Society London*, **454**, 903–995 (1998).

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

[Jorgensen et al., 1983] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 1983; **79**:926–935 (1983).

[Kachel et al., 2006] Observation of intermediate states of the human prion protein by high pressure NMR spectroscopy. Kachel N, Kremer W, Zahn R and Kalbitzer H.R. *BMC Structural Biology* **6**:16 (2006).

[Kalchhauser & Robien, 1984] Kalchhauser, H. and Robien, W. CSEARCH: A computer program for identification of organic compounds and fully automated assignments of carbon-13 nuclear magnetic resonance spectra, *J. Chem. Inf. Comput. Sci*. **25(2)**, 103–108 (1984).

[Karplus, 1959] Karplus, Martin. "Contact Electron-Spin Coupling of Nuclear Magnetic Moments". *J. Chem. Phys*. **30 (1)**: 11–15 (1959).

[Kauppinen & Partanen, 2002] Kauppinen J., Partanen J. Fourier Transforms in Spectroscopy (2002).

[Kauppinen et al., 2002] Kauppinen J., Partanen J. Fourier Transforms in Spectroscopy (2002).

[Keepers & James, 1984] Keepers, J.W. & James,T.L. A Theoretical Study of Distance Determination from NMR. Two-Dimensional Nuclear Overhauser Effect Spectra. *J. Magn. Reson.* **57**, 404-426 (1984).

[Kitchen et al., 2004] Kitchen D.B., Decornez H., Furr J.R., Bajorath *J. Nat. Rev. Drug. Discov.* **3**, 935-949 (2004).

[Kolmogorov, 1933] Kolmogorov, A. "Sulla determinazione empirica di una legge di distribuzione" *G. Inst. Ital. Attuari*, **4**, 83 (1933).

[Kremer et al., 2004] Kremer W, Arnold MR, Kachel N, Kalbitzer HR. The use of high-sensitivity sapphire cells in high pressure NMR spectroscopy and its application to proteins. *Spectroscopy* **18**:271-278 (2004).

[Kuwata et al., 2002] Kuwata K, Li H, Yamada H, Legname G, Prusiner SB, Akasaka K, James TL. Locally disordered conformer of the hamster prion protein: a crucial intermediate to PrPSc? *Biochemistry* **41**:12277-12283 (2002).

[Levenberg, 1944] Levenberg, K. "A Method for the Solution of Certain Non-Linear Problems in Least Squares". *The Quarterly of Applied Mathematics* **2**: 164–168 (1944).

[Lindahl et al., 2001] Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Mod. 2001* **7**: 306–317 (2001).

[Malloni et al., 2010] W.M.Malloni, S.De Sanctis, A.M. Tomé, E.W.Lang, C.E.Munte, K.Stadlthanner, K.P. Neidig, H.R.Kalbitzer, Automated solvent artifact removal and base plane correction from multidimensional NMR protein spectra by AUREMOL-SSA, *J. Biomol. NMR* **47(2)**, pp. 110-111 (2010).

[Mann & Whitney, 1947] Mann, H. B.; Whitney, D. R. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". *Annals of Mathematical Statistics* **18 (1)**: 50–60 (1947).

[Massey, 1951] Massey, F. J. Jr. The Kolmogorov-Smirnov test of goodness of fit. Journal of the American Statistical Association, Vol. **46**. The table of critical values of D is found on p. 70 (1951).

[Maurer et al., 2004] Maurer T, Meier S, Kachel N, Munte CE, Hasenbein S, Koch B, Hengstenberg W, Kalbitzer HR 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 (2004).

[Miyamoto & Kollman, 1992]Miyamoto S, Kollman PA. SETTLE: an analytical version of the SHAKE and RATTLE algorithms for rigid water models. *J. Comp. Chem.* **13**:952–962 (1992).

[Möglich et al., 2005] Möglich A., Weinfurtner,D., Gronwald,W., Maurer,T. and Kalbitzer,H.R. PERMOL: restraint-based protein homology modeling using DYANA or CNS. *Bioinformatics,* **21**, 2110–2111 (2005).

[Moitessier et al., 2008] Moitessier N., Englebienne P., Lee D., Lawandi J., Corbeil C.R. Br. *J. Pharmacol*. **153**, S7-S26 (2008).

[Molkentin, 2007] Molkentin D. The Art of Building Qt Applications (2007).

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

[Neal et al., 2003] Neal, S., Nip, A. M., Zhang, H., and Wishart, D. S. Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *Journal of Biomolecular NMR*, **26(3)**:215–240 (2003).

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

[Norman, 1953] Ricker, Norman "WAVELET CONTRACTION, WAVELET EXPANSION, AND THE CONTROL OF SEISMIC RESOLUTION". *Geophysics* **18 (4)**, (1953).

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

[Paul & Steinwedel 1953] Paul W., Steinwedel H. "Ein neues Massenspektrometer ohne Magnetfeld". *Zeitschrift für Naturforschung A* **8 (7)**: 448–450 (1953).

[Pervushin et al., 1997] Pervushin, K., Riek, R., Wider, G.& Wüthrich, K. Attenuated T-2 relaxation by mutual cancellationof dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. natn Acad. Sci.* USA **94**, 12366-12371 (1997).

[Prusiner, 1998] Prusiner SB: Prions. *Proc. Natl. Acad. Sci.* USA Vol. **95**, pp. 13363–13383 (1998).

[Purcell et al., 1946] E.M. Purcell, H.C. Torrey, and R.V. Pound. "Resonance Absorption by Nuclear Magnetic Moments in a Solid." *Phys. Rev.* **69** 37 (1946).

[Rabi et al., 1938] I.I. Rabi, J.R. Zacharias, S. Millman, P. Kusch. "A New Method of Measuring Nuclear Magnetic Moment". *Physical Review* **53 (4)**: 318. (1938).

[Renugopalakrishnan et al., 1991] Renugopalakrishnan, V., Carey, P. R., Smith, I. C. P., Huans, S. G., and Storer Proteins: Structure, Dynamics and Design. *Springer*, 1 edition (1991).

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

[Ross et al., 2000] Ross A, Schlotterbeck G, Klaus W, Senn H. Automation of NMR measurements and data evaluation for systematically screening interactions of small molecules with target proteins. *J. Biomol. NMR* **16**, 139–146 (2000).

[Rossè et al., 2002] Rossè G, Neidig P, Schröder H: Automated structure verification of small molecules libraries using 1D and 2D NMR techniques. *Methods Mol. Biol.* **201**:123-139 (2002).

[Savitzky & Golay, 1964] Savitzky, A.; Golay, M.J.E. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures". *Analytical Chemistry* **36(8)**: 1627–1639 (1964).

[Schröder & Neidig, 1999] Schröder, H. and Neidig P. AutoDROP, a new method of automated structure verification in combinatorial chemistry. *Bruker Report* **147**, 18–21 (1999).

[Schulte et al., 1997] Use of global symmetries in automated signal class recognition by a bayesian method. Schulte AC, Gorler A, Antz C, Neidig KP, Kalbitzer HR. *J. Magn. Reson.* **129(2)**:165-72 (1997).

[Schumann et al., 2007] Frank H. Schumann, Hubert Riepl, Till Maurer, Wolfram Gronwald, Klaus-Peter Neidig and Hans Robert Kalbitzer. Combined chemical shift changes and amino acid specific chemical shift mapping of protein–protein interactions. *J. Biomol. NMR* **39(4)**:275-89 (2007).

[Schwarzinger et al., 2000] Schwarzinger, S., Kroon, G. J. A., Foss, T. R., Wright, P. E., and Dyson, H. J. Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView. *J. Biomol. NMR* **18**, 43–48 (2000).

[Schwarzinger et al., 2001] Schwarzinger, S.; Kroon, G. J. A.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J. Sequence-Dependent Correction of Random Coil NMR Chemical Shifts. *J. Am. Chem. Soc.* **123**, 2970 (2001).

[Scott, 2010] Scott's rule "Scott's rule". *Computational Statistics* **2(4)** (2010).

[Sepetov & Issakova, 1999] Sepetov, N. and Issakova, O. Analytical characterization of synthetic organic libraries. *Comb. Chem. Technol*. **169**–203 (1999).

[Sorin et al., 2005] Sorin EJ, Pande VS. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys J.* **88**:2472–2493 (2005).

[Spoerner et al., 2010] Spoerner, M., Hozsa, C., Poetzl, J., Reiss, K., Ganser, P., Geyer, M. & Kalbitzer, H.R. Conformational states of human rat sarcoma (Ras) protein complexed with its natural ligand GTP and their role for effector interaction and GTP hydrolysis. *J. Biol. Chem.* **285**: 39768-39778 (2010).

[Student, 1908] Student. The probable error of a mean. *Biometrika* **6(1)**: 1-25 (1908).

[Thiele et al., 2011] Thiele H., McLeod G., Niemitz M. And Kühn T. Structure verification of small molecules using mass spectrometry and NMR spectroscopy. *Monatsh. für Chemie* **142**:717-730 (2011).

[Traficante & Nemeth, 1987] Traficante D., Nemeth G. A., A new and improved apodization function for resolution enhancement in NMR spectroscopy, *JMR*, vol. **71**, 1 February 1987, 237-245 (1987).

[Trenner PhD, 2006] Trenner, J.M. Accurate proton-proton distance calculation and error estimation from NMR data for automated protein structure determination in AUREMOL. *Ph.D. Thesis*. University of Regensburg, Germany (2006).

[Viana et al., 2000] Viana, R., V. Monedero, V. Dossonnet, C. Vadeboncoeur, G. Pe´rez-Martinez, and J. Deutscher. Enzyme I and HPr from Lactobacillus casei: their role in sugar transport, carbon catabolite repression and inducer exclusion. *Mol. Microbiol*. **36**:570–584 (2000).

[Wagner et al., 1983] Wagner G., Pardi A. and Wüthrich K: Hydrogen bond length and 1H NMR chemical shifts in proteins. *J. Am. Chem. Soc.* **105**, 5948-5949 (1983).

[Wang et al., 2001] Wang B., Fleischer U., Hinton JF., Pulay P. Accurate prediction of proton chemical shifts. *Journal of Computational Chemistry*, 1887~1895 (2001).

[Welch, 1938] The significance of the difference between two means when the population variances are unequal. *Biometrika* **29** 350–362 (1938).

[Wilcoxon, 1945] Wilcoxon, F. "Individual comparisons by ranking methods". *Biometrics Bulletin* **1(6)**: 80–83 (1945).

[Wishart et al., 1995] DS Wishart, CG Bigam, A Holm, RS Hodges and BD Sykes. "1H, 13C and 15N random coil chemical shifts of the common amino acids: I. investigations of nearest neighbor effects." *Journal of Biomolecular NMR* **5** : 67-81 (1995).

[Wüthrich, 1986] Wüthrich K. NMR of Proteins and nucleic acids. *Wiley interscience* (1986).

[Wüthrich, 1990] Wüthrich K. "Protein structure determination in solution by NMR spectroscopy". *J. Biol. Chem.* **265(36)**: 22059–62. (1990).

[Zahn et al., 2000] NMR solution structure of the human prion protein. Zahn R, Liu A, Lührs T, Riek R, von Schroetter C, López García F, Billeter M, Calzolai L, Wider G, Wüthrich K. *Proc. Natl. Acad. Sci.* U S A **97(1)**:145-50 (2000).

[Zonst, 1995] Zonst E. Understanding the FFT: A Tutorial on the Algorithm & Software for Laymen, Students, *Technicians & Working Engineers* (1995).

Erklärung

Hiermit erkläre ich, das ich die vorliegende Arbeit selbständig angefertigt, und keine Hilfsmittel, außer den angegebenen, benutzt habe.

Regensburg, 21-11-2011

Wilhelm Massimiliano Malloni