New Methods for Automated NMR Data Analysis and Protein Structure Determination

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich

Chemische und Pharmazeutische Wissenschaften (FB 14)

der Johann Wolfgang Goethe-Universität

in Frankfurt am Main

von

Dancea Felician

aus Aiud, Rumänien

Frankfurt 2004

(DF1)

vom Fachbereich Chemische und Pharmazeutische Wissenschaften (FB 14) der Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan:	Prof. Dr. Harald Schwalbe
Gutachter:	Prof. Dr. Heinz Rüterjans
	PD Dr. Ulrich Günther
Datum der Disputation:	10.03.2005

Acknowledgements

In the first place I wish to thank my scientific supervisors Prof. Dr. Heinz Rüterjans and PD Dr. Ulrich Günther for all their scientific support and for giving me the chance to pursuit this challenging work.

I would like to acknowledge the contributions of several people who have helped me to carry out this work: Dr. Frank Löhr for NMR experiments, PD Dr. Oliver Klimmek for protein sample preparations, Dr. Hans Wienk for insightful discussions and suggestions, Dr. Michael Nilges for help with ARIA-related computations, Dr. Yi-Jan Lin for help with the Sud project and Nikola Trbovic for the Wavepca collaboration.

Special thanks to Prof. Dr. Volker Dötsch for his kind support.

Many thanks to all former and present BPC members: Joana Kleinhaus, Dr. Gary Yalloway, Mitch Maestre, Tanja Mittag, Lucia Muresanu, Alexander Koglin, Veronica Noskova, Dr. Vladimir Rogov, PD Dr. Christian Lücke, Dr. Stefania Pfeiffer-Marek, Dr. Christian Wolf, Dr. Marco Betz, Bernd Weyrauch, Michael Reese, Dr. Wesley McGinn-Straus, Horng Ou, Dr. Dirk Beilke, Dr. Ulrich Schieborr, Dr. Helmut Hanssum, Birgit Schäfer, Juliana Winkler, Christina Fischer, Dr. Frank Bernhard, Dr. Vicky Katsemi, Dr. Raed Aljazzar and Dr. Kaushik Sengupta, for the excellent atmosphere in the working group. It has been a great place where ideas were shared and generated. Special thanks to our secretary, Ms. Sigrid Fachinger, for her great help with the official paper work.

I would like to acknowledge the financial support from Deutsche Forschungsgemeinschaft (SFB472) and from the Center of Biomolecular Magnetic Resonance (BMRZ) at J. W. Goethe-University of Frankfurt.

Abbreviations

NMR	nuclear magnetic resonace
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser enhancement and exchange spectroscopy
TROSY	transverse relaxation spectroscopy
HSQC	heteronuclear single quantum coherence
ADR	ambiguous distance restraints
RDC	residual dipolar coupling
Sud	polysulfide-sulfur transferase (formerly Sulphide Dehydrogenase) protein
Str	sulfur transferase protein
hsp90	heat shock protein 90
RMSD	root mean squared deviation
rms	root mean squared
1D, 2D, 3D	one-, two-, three-dimensional
DWT	discrete wavelet transform
MRA	multiresolution analysis
PCA	principal component analysis
pc*i*	principal component *i*
SVD	singular value decomposition
SA-MD	simulated annealing with molecular dynamics
SA-TAD	simulated annealing with torsion angle dynamics
ARIA	ambiguous restraints for iterative assignment
CYANA	combined assignment and dynamics algorithm for NMR applications
CNS	crystallography and NMR system
CPU	central processing unit

Units

Da	Dalton
Hz	Hertz
K	Kelvin
M	$mol \cdot l^{-1}$
l	liter
s	second
T	Tesla
cal	gram calorie

Contents

1 Introduction **8**

2 Theoretical concepts **15**

2.1 NMR spectroscopy 15

2.1.1 Nuclei in magnetic fields 15

2.1.2 Density matrix formalism 18

2.1.3 Product operator formalism 18

2.2 NMR data for protein structure calculation 19

2.2.1 Nuclear Overhauser effects 19

2.2.2 Residual dipolar couplings 21

2.2.3 Scalar couplings 23

2.2.4 Hydrogen bonds 23

2.2.5 Chemical shifts 24

2.3 Structure calculation algorithms 25

2.3.1 Simulated annealing with molecular dynamics 25

2.3.2 Iterative NOE assignment and structure calculation 27

2.4 Numerical analysis algorithms 34

2.4.1 Multiresolution analysis and wavelet series expansion 34

2.4.2 Discrete wavelet transform 38

2.4.3 Wavelet de-noising 39
2.4.4 Translation invariant wavelet transform 41
2.4.5 Principal component analysis 42

3 Experimental procedures **46**

3.1 NMR sample preparation for Sud protein 46
3.2 NMR sample preparation for Sud-Str complex 47
3.3 NMR experiments 48

4 Data analysis methods **51**

4.1 Structural data preparation for Sud protein 51
4.2 Sud protein structure calculation 52
4.3 Consistency check of the NOESY peak lists 55
4.4 Wavelet de-noising of the multidimensional NMR spectra 57
4.5 Automated peak picking and peak integration of the multidimensional NMR spectra 59
4.6 NMR chemical shift mapping 61
4.7 Multivariate analysis of the NMR screening data 62

5 Results and Discussion **67**

5.1 Sud protein 67
5.1.1 Solution structure of Sud protein 67
5.1.2 Chemical shift mapping of the polysulfide binding 75
5.1.3 Chemical shift mapping of the Sud-Str interaction 79
5.2 Automated protein structure determination using wavelet de-noised NOESY spectra 82
5.2.1 Optimal wavelet based de-noising scheme 82
5.2.2 NOESY peak list validation 85

5.2.3 Iterative NOE assignment and structure calculations using wavelet de-noised spectra 86

5.3 Wavelet de-noising for NMR screening 92

6 Zusammenfassung **98**

7 CURRICULUM VITAE **104**

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is a well established method for the determination of solution structures of biological macromolecules. NMR plays an important role in structural genomics which is driven by the need to supplement protein sequences by structural and functional information (Staunton et al., 2003). The efficiency of protein NMR structure determination has recently improved because many of the time-consuming interactive steps carried out by a spectroscopist during the process of spectral analysis can now be accomplished by automated, computational approaches (Moseley and Montelione, 1999).

Recent advances in automation of protein NMR structure determination were the product of a series of computational algorithms which link the iterative assignment of NOESY spectra with structure calculations (Mumenthaler and Braun, 1995; Mumenthaler et al., 1997; Nilges et al., 1997; Montelione et al., 2000; Savarin et al., 2001; Herrmann et al., 2002a). While new types of constraints such as residual dipolar couplings (Tjandra and Bax, 1997), orientational information from heteronuclear relaxation in anisotropically tumbling molecules (Tjandra et al., 1997a), or restraints obtained in the presence of paramagnetic centers in a protein (Banci et al., 1997) have facilitated protein structure determination, distance information from NOESY spectra remains an important basis for NMR structure elucidation. Peak picking in NOESY spectra has been a time consuming process, mainly due to spectral overlap and because NOESY spectra are often obscured by noise and spectral artifacts. Therefore, automation of the peak picking process requires reliable filters to select the relevant

signals.

An initial implementation of a program which combines NOESY peak picking with automated structure determination by using intermediate protein structures as a guide for the interpretation of the NOESY spectra has recently been described (Herrmann et al., 2002b). In this thesis a different approach to automated peak picking, employing wavelet transforms for spectral de-noising, was evaluated. The core of this new procedure is the generation of incremental peak lists by applying different wavelet de-noising schemes with complementary features. In the first stage of iterative NOE assignment and structure calculations, a peak list containing only the most reliable peaks is used, while a wavelet de-noising scheme with modest noise suppression and large number of signals is employed in the later stages, when the previously determined structural models can be utilized to filter the NOESY peak list. In addition, the peak list generated by automated peak picking on wavelet de-noised spectra is subject to a consistency check based on symmetries in, and between heteronuclear-edited NOESY spectra, and on the fact that the NOE signals are usually part of a network of connectivities between adjacent spin systems. Automated peak picking is further combined with a robust numerical scheme for peak integration of multi-dimensional NMR spectra using an object-related growing algorithm which can cope with severe spectral overlap without any assumptions on peak shapes. These algorithms were implemented in the context of the ARIA software for automated NOE assignment and structure determination (Linge et al., 2003) and were validated using the high-resolution structure of the polysulfide-sulfur transferase protein (Sud) from *Wolinella succinogenes*, which has been previously elucidated by manual interactive peak picking.

Wavelet transforms became a popular tool in analytical chemistry during the late eighties and, since then, about 400 papers and several books have been published (Shao et al., 2003). Wavelet transforms were employed for signal processing in different fields of analytical chemistry including high-performance liquid chromatography (Collantes et al., 1997), capillary electrophoresis (Perrin et al., 2001), ultraviolet-visible

spectroscopy (Xiaoquan et al., 2004), infrared spectroscopy (Chen et al., 2004), Raman spectroscopy (Ehrentreich and Summchen, 2001), photoacoustic spectroscopy (Shao et al., 1999), atomic emission spectroscopy (Ma and Zhang, 2003), X-ray diffraction (Main and Wilson, 2000), and analytical image processing (Sorzano et al., 2004). They have been utilized to solve certain problems in quantum chemistry and chemical physics (Fischer and Defranceschi, 1998) as well. Recent applications of wavelet transforms to the high-resolution biomolecular NMR spectroscopy show potential applications in data processing, in particular for the suppression of the water signal (Günther et al., 2002), signal de-noising (Cancino-De-Greiff et al., 2002) and data compression (Cobas et al., 2004).

One of the most important applications of the wavelet transform is noise suppression. Compared to many other algorithms used to reduce spectral noise, wavelet de-noising is exceptionally stable and computationally efficient. For optimal de-noising, noise reduction must be achieved while preserving the fine structure of the signals. The result depends predominately on three variables: the wavelet base function (*e.g.* Symmlet, Daubechies, Coiflet), the wavelet transform (*e.g.* periodic orthogonal, translation invariant) and the thresholding procedure (*e.g.* soft, hard). In this work the most relevant de-noising variables were optimized for multidimensional NOESY spectra of isotopically labeled proteins.

Another emerging application of the wavelet transform is the combination of the exploratory data analysis algorithms (such as principal component analysis, partial least squares, canonical variables or artificial neuronal networks) with the multiresolution analysis offered by the wavelet representation of the analytical signals (Bakshi, 1998; Teppola and Minkkinen, 2000; Laakso et al., 2001). This approach can be particularly useful to analyze NMR screening data where a large number of spectra need to be compared for changes and similarities. Typical applications are ligand screening employing two-dimensional NMR spectra and metabolomics using one-dimensional NMR spectra. An automated comparison tool requires a robust exploratory data anal-

ysis algorithm, which is insensitive to insignificant spectral variations caused by small pH and concentration changes.

NMR spectroscopy has become an important technique in screening for protein inhibitors (Shuker et al., 1996). Both NMR spectra of isotopically labeled proteins and the spectra of the inhibitors can be used for ligand screening (Shuker et al., 1996; Stockmann and Dalvit, 2002). A great variety of NMR methods, including transferred NOEs (Meyer et al., 1997; Vogtherr and Peters, 2000), saturation transfer difference (STD) experiments (Mayer and Meyer, 2001, 1999, 2000), ePHOGSY(Bertini et al., 1997), diffusion editing (Lin et al., 1997) or NOE pumping (Chen and Shapiro, 2000, 1998), are used. Most of these techniques can be exploited for rational drug design as well as for screening of large numbers of inhibitors. NMR is now also used as a screening technique in metabonomics/metabolomics where biological samples such as bio-fluids or tissue extracts are subject of investigations (Lindon et al., 2000). NMR screening using predominantly one-dimensional spectra of body fluids has been employed to study toxicity and gene function (Nicholson et al., 2002). Similarly, NMR has been used to screen fruit juices (Belton et al., 1998; Vercauteren and Rutledge, 1996) or beer (Duarte et al., 2002) as a measure of quality control.

Principal component analysis (PCA) (Wold et al., 1987) is the most commonly used pattern recognition method for analyzing the NMR screening data. A series of filters are applied to the experimental data to obtain suitable descriptors for PCA (to cluster similar data and obtain good separation between clusters) which minimize the weight of small chemical shift variations and optimize computational efficiency. The most common filter is 'bucketing' where adjacent points are summed to a 'bucket' (Ross et al., 2000). Bucketing eliminates artifacts caused by small chemical shift and intensity variations and improves computational efficiency by a significant compression of the original spectroscopic data. Bucketing also causes artifacts when peaks experience small chemical shift perturbations at the border between buckets or due to cancellations within a bucket when different points add and subtract equal or similar intensities. To

overcome some inherent disadvantages of the bucketing procedure the effect of wavelet de-noising on multivariate analysis has been explored using an experimental data set of [^{15}N,^{1}H]-HSQC spectra of proteins with different ligands present. The combination of wavelet de-noising and PCA proved to be most efficient when PCA is applied directly in wavelet space. The new algorithm combines the advantages of wavelet data representation with the data visualization and clustering obtained by PCA.

In addition to the methodological part comprising of new software tools for efficient NMR data analysis and protein structure determination, the thesis presents the *de novo* solution structure determination of the periplasmic polysulfide-sulfur transferase protein (Sud) from *Wolinella succinogenes.* Sud is induced in the anaerobic gram-negative bacterium *Wolinella succinogenes* upon growing with formate and polysulfide as catabolic substrates (Klimmek et al., 1991; Kreis-Kleinschmidt et al., 1995) and serves as a polysulfide binding and transferase protein (Klimmek et al., 1998) allowing rapid polysulfide-sulfur reduction at low polysulfide concentrations. The Sud protein comprises of two identical subunits of about 15 kDa and does not contain prosthetic groups or heavy metal ions. Each monomer contains a single cysteine residue, which was found to be essential for the protein function. *In vitro*, it appears that each cysteine covalently binds up to 10 polysulfide-sulfur (S_n^{2-}) atoms, when incubated in a polysulfide solution (Klimmek et al., 1999). Sud is thought to transfer polysulfide-sulfur to the catalytic molybdenum ion located at the periplasmic active site of the membrane protein polysulfide reductase (Prisner et al., 2003). The polysulfide-sulfur transfer from Sud to polysulfide reductase probably occurs in a complex of the two proteins, when a sulfur atom is reductively cleaved from the polysulfide chain (Klimmek et al., 1999).

A BLAST search (Altschul et al., 1997) indicated that Sud shows little primary sequence homology to other proteins with known three-dimensional structure (Figure 1.1). The homologous partners are sulfurtransferase or rhodanese enzymes that catalyze the transfer of a sulfur atom from suitable donors to nucleophilic acceptors (*e.g.* from thiosulfate to cyanide). Their three-dimensional structures display a typical α/β

```
Sud      1    ADMGEKFDATFKA...QVKAAKADMVMLSPKDAYKLLQENPDITLID.   44
GlpE     11   ............................DAHQKLQEK.EAVLVD.   25
Rhobov   154  ATLNRSLLKTYEQVLENLESKRFQLVDSRAQGRYLGTQPEPDAVGLDS  201
RhdA     133  APAGGPVALSLHD...EPTASR.DYLLGRLGAA........DLAIWD.  167
                                                          *
Sud      45   ..VRDPDE......LKAMGK..PDVKNYKHMS..........RGKLEP  72
GlpE     26   ..IRDPQS......F.AMGH..A.VQAF.HLT..........NDTLGA  50
Rhobov   202  GHIRGSVN......MPFMNF..LTEDGFEK.S..........PEELRA  230
RhdA     168  ..ARSPQEYRGEKVLAAKGGHIPGAVNFEWTAAMDPSRALRIRTDIAG  213
                 *
Sud      73   LLAKSGLDPEKPVVVFCKTAARAALAGKTLREYGFKTIYNSEGGMDKW  120
GlpE     51   FMRDNDFD..TPVMVMCYHGNSSKGAAQYLLQQGYDVVYSIDGGFEAW  96
Rhobov   231  MFEAKKVDLTKPLIATCRKGVTA.......................... 253
RhdA     214  RLEELGITPDKEIVTHCQTHHRSGLTYLIAKALGYPRVKGYAGSWGEW  261
                           *
```

Figure 1.1. Multiple amino acid sequence alignment of rhodanese-like proteins with known three-dimensional structure: Sud, *Escherichia coli* GlpE (Spallarossa et al., 2001), bovine liver rhodanese (Ploegman et al., 1978) and *Azotobacter vinelandii* rhodanese (Bordo et al., 2000). The primary sequence of Sud was taken as reference and the pair alignments were obtained with BLAST. The active-site loop residues are shown in red and the additional charged residues in the active site of Sud in blue. Invariant residues are marked with asterisks.

topology and have a similar environment in the active site, primarily with respect to the main-chain conformation of the Cys-located active-site loop. The highest primary sequence similarity (30% identity) is observed between Sud and *Escherichia coli* GlpE, a protein that has been proposed to have the prototype structure for the ubiquitous single-domain rhodanese module (Spallarossa et al., 2001). The amino acid composition and the location of charged residues in the active-site region best matches the rhodanese of *Azobacter vinelandii* (Bordo et al., 2000), despite a lower sequence homology (23% identical residues).

The solution structure of the homodimeric Sud protein has been determined using heteronuclear multi-dimensional NMR techniques. The structure is based on NOE-derived distance restraints, backbone hydrogen bonds and torsion angle restraints as well as residual dipolar coupling restraints for a refinement of the relative orientation of the monomer units. Within the NMR spectra of homodimers, all symmetry-related nuclei have equivalent magnetic environments and therefore are degenerated in chem-

ical shift. This simplifies the resonance assignment (only half of the nuclei have to be assigned), but complicates the NOESY assignment and structure calculations considerably, mainly because it is not possible to distinguish *a priori* between the intra-monomer, inter-monomer and co-monomer (mixed) NOE cross peaks. There are two possibilities to overcome the symmetry degeneracy problem of the NOESY data: the use of asymmetric labeling experiments in order to separate between intra- and intermolecular NOEs and/or special structure calculation methods which can incorporate the inherent ambiguity of the NOE derived distance restraints. The NMR structure of the Sud homodimer was calculated using the symmetry-ADR method (O'Donoghue and Nilges, 1999) in combination with data from asymmetric labeling experiments (Ferentz et al., 1997; Melacini, 2000).

Recently, a second polysulfide-sulfur transferase protein (Str, 40 kDa) with a five-fold higher native concentration compared to Sud has been identified in the bacterial periplasm of *Wolinella succinogenes*. The two proteins are thought to form a polysulfide-sulfur harvesting complex in the sense that Str collects and delivers the aqueous polysulfide to Sud, which in turn mediates the sulfur transfer to the catalytic molybdenum ion located at the periplasmic active site of the membrane protein polysulfide reductase. The primary sequence of the Str protein contains seven cysteine residues, from which one is likely to be the polysulfide-sulfur binding site. Chemical shift mapping by NMR spectroscopy (Zuiderweg, 2002; Clarkson and Campbell, 2003) was used to examine the interaction between Sud and Str. The [^{15}N,^{1}H]-TROSY spectra of the Sud protein were compared in the absence and in the presence of the polysulfide-sulfur and Str protein. The results provide further insights into the mechanism of the polysulfide-sulfur binding and transfer within the bacterial periplasm.

2 Theoretical concepts

2.1 NMR spectroscopy

2.1.1 Nuclei in magnetic fields

NMR spectroscopy relies on the quantum effects induced by an external magnetic field to the magnetic moment of the atomic nuclei. The nuclear magnetic moment μ is defined by:

$$\mu = \gamma \mathbf{I}; \quad \left|\mu^2\right| = \mu \cdot \mu = \gamma\hbar[I(I+1)]; \quad \mu_z = \gamma I_z = \gamma\hbar m, \tag{2.1}$$

where γ is the gyromagnetic ratio, $\mathbf{I}$ is the nuclear spin angular momentum, an intrinsically quantum mechanical property without a classical analog, I_z is the z component of $\mathbf{I}$, m is the nuclear magnetic quantum number and $\hbar$ is the Planck's constant divided by 2π. The magnetic quantum number $m = (-I, -I+1, ..., I-1, I)$ has $2I+1$ discrete values, where I is the nuclear spin angular quantum number. Atomic nuclei can be divided in three classes: nuclei with odd mass number have half-integral quantum spin numbers, nuclei with even mass number and even charge number have quantum spin numbers equal to zero (inactive for NMR) and nuclei with an even mass number and an odd charge number have integral quantum spin numbers.

The orientation of the nuclear magnetic moment vector μ is quantified because the magnitude of the vector is constant and the z component has a set of discrete values (Egn. 2.1). For an isolated spin in the absence of external fields the spin angular

momentum does not have a preferred orientation and therefore the quantum states corresponding to the $2I+1$ values of m have equal energy (degenerate quantum states). An external magnetic field raises the degeneration (Zeeman effect) and the spin states of the nucleus have energies given by:

$$E = -\mu \cdot \mathbf{B}, \tag{2.2}$$

where $\mathbf{B}$ is the magnetic field vector. In an NMR spectrometer the static external magnetic field is along the z-axis of the laboratory coordinate frame and Egn. 2.2 reduces to: $E = -\gamma I_z B_0 = -\gamma \hbar m B_0$, where B_0 is the static magnetic field strength. The selection rule governing magnetic dipole transitions between Zeeman states is $\triangle m = \pm 1$. Consequently, the photon energies required to excite a transition between m and $m+1$ Zeeman states is: $\triangle E = \gamma \hbar B_0$. Thus, in a NMR spectrometer at a constant magnetic field B_0, one can record the absorbance of radiation as a function of frequency with resonances at:

$$\nu_0 = \frac{\omega_0}{2\pi} = \frac{\gamma B_0}{2\pi}, \tag{2.3}$$

where ω_0 is the Larmor frequency of different nuclei.

High-resolution NMR focuses predominantly on $I = \frac{1}{2}$ nuclei because nuclear spins with a higher quantum number possess electric quadrapole moments that lead to fast relaxation and broad spectral lines. These nuclei have only two spin states and two energy levels are obtained by application of an external magnetic field. The spin state with $m = \frac{1}{2}$ is referred to as the α state and the state with $m = -\frac{1}{2}$ is referred to as the β state. At equilibrium, the energy states are unequally populated because lower-energy orientation of the magnetic dipole vector is more probable. The relative population of a state is given by the Boltzmann distribution:

$$\frac{N_\alpha}{N_\alpha + N_\beta} = exp\left(\frac{-E_\alpha}{k_B T}\right) \Big/ \left[exp\left(\frac{-E_\alpha}{k_B T}\right) + exp\left(\frac{-E_\beta}{k_B T}\right)\right]. \tag{2.4}$$

The population difference is on the order of 1 in 10^5 for ^{1}H spins in an 11.7 T magnetic field, which explains much of why is desirable to use powerful magnetic fields in NMR spectroscopy.

Three of the four most abundant elements in biological materials, hydrogen carbon and nitrogen, have naturally occurring isotopes with $\frac{1}{2}$ quantum spin number, and are therefore suitable for high-resolution biomolecular NMR. The proton (^{1}H) has the highest natural abundance (0.98%) and the highest sensitivity due to its large gyromagnetic ratio (26.7519 [$10^7 \cdot$rad$\cdot$T$^{-1}\cdot$s^{-1}]). Because of the low natural abundance and gyromagnetic ratios of ^{15}N and ^{13}C (0.37% and 1.11%, -2.7126 and 6.7238 [$10^7 \cdot$rad$\cdot$T$^{-1}\cdot$s^{-1}], respectively) the NMR experiments with this nuclei require isotope enrichment which is routinely achieved by overexpression of proteins in isotope-labeled media.

The observed resonances differ slightly from the frequencies calculated with Eqn. 2.3 as an influence of the local environment of the nuclei. The change of the Larmor frequency caused by changes in the local environment of individual nuclei is referred to as *chemical shift*. It stems from secondary magnetic fields induced by the motion of electrons in the external magnetic field. The net magnetic field at the location of a nucleus depends on the static magnetic field and the secondary fields. Chemical shift dispersion provides the spectral resolution which is exploited to study chemical structure, molecular conformations and the solvent environment of molecules.

NMR spectra of molecules in liquid solution show a splitting of resonances into multiplets. The fine structure of the spectral lines cannot be explained by direct dipolar interactions between the nuclear magnetic dipole moments because the dipolar coupling is averaged to zero by isotropic tumbling of the molecule in solution. This splitting appears due to spin-spin interaction mediated by the electrons which form the chem-

ical bonds connecting the nuclei. This type of interaction is commonly referred to as *scalar coupling*. Its strength is measured by the scalar coupling constant ${}^{n}J_{ab}$ for two nuclei a and b separated by n covalent bonds.

2.1.2 Density matrix formalism

Theoretical analysis of a modern NMR experiment requires calculation of the signal observed following a sequence of radio-frequency pulses and delays. The full description involves a quantum statistical representation known as the density matrix formalism (Abragam, 1967). The initial state of the system is described by a equilibrium density operator $\sigma(0)$. Evolution of the density operator during the sequence of pulses and delays $\sigma(t)$ is given by the Liouville-von Neumann equation:

$$\frac{\mathrm{d}\sigma(t)}{\mathrm{d}t} = -i\left[\mathcal{H}, \sigma(t)\right]. \tag{2.5}$$

The Hamiltonian $\mathcal{H}$ includes the Zeeman, scalar coupling and radio-frequency pulse terms that govern the evolution of the density operator.

2.1.3 Product operator formalism

Although the density operator theory provides a rigorous description of the evolution of nuclear spin system during a NMR experiment, the required matrix calculation becomes prohibitive as the number of spins increases. Furthermore, it is difficult to get a direct interpretation of the density operator evolution and therefore the formalism lacks the qualitative insight into the NMR experiment.

A simplified formalism referred to as the product operator formalism was developed to describe the system of weekly coupled nuclear spins. In the weak coupling regime a simple set of operators is sufficient to describe the magnetization transfer pathways during the NMR experiment (Sørensen et al., 1983). The product operator formalism gives the spectroscopist an intuitive idea for the time evolution of the density operator

while retaining much or the rigor of the full density matrix approach. It provides simple rules for the chemical shift and the scalar coupling evolution during periods of free precession and for the applied radio-frequency pulses.

2.2 NMR data for protein structure calculation

2.2.1 Nuclear Overhauser effects

The primary source of information for protein NMR structure determination is given by a dense network of distance restraints derived from nuclear Overhauser effects (NOEs) between neighboring hydrogen atoms in the protein. The nuclear Overhauser effect reflects the magnetization transfer between spins coupled by the dipole-dipole interaction in a molecule that undergoes Brownian motion in a liquid (see Neuhaus and Williamson, 1989 for a thorough discussion). In principle, all hydrogens atoms of a protein form a network of spins coupled by dipole-dipole interactions. Magnetization can be transfered from one spin to another not only directly but also by spin diffusion, *i.e.* indirectly via other spins in the vicinity. The Solomon's equations (Solomon, 1955) provide a semi-classical description of multiple interacting spins. The spin-lattice relaxation is described by the rates of spin transitions between energy levels. For the simplest approximation of an isolated homonuclear spin pair (IS, $\gamma_I = \gamma_S = \gamma$) and assuming an isotropic tumbling rigid body model for the inter-proton vector, the cross-relaxation rate predicted by the Solomon's equations is:

$$\sigma_{IS}^{NOE} = \frac{\hbar\mu_0\gamma\tau_c}{40\pi^2 r_{IS}^6}\left(-1+\frac{6}{1+4\omega_0\tau_c^2}\right). \tag{2.6}$$

The intensity of a NOE, *i.e.* the volume V_{IS}^{NOE} of the corresponding cross peak of the NOESY spectrum is proportional to the cross-relaxation rate and therefore it is proportional with the inverse of the sixth power of the distance between the two interacting spins:

$$V_{IS}^{NOE} = r_{IS}^{-6} f(\tau_c). \tag{2.7}$$

The isolated pair spin approximation is valid only for very short mixing-times of the NOESY experiment. However very short mixing times are impractical because the cross peak intensities have low signal-to-noise ratios. For longer mixing times the NOE volumes are no longer proportional to the cross-relaxation rates of the isolated spin pair because the magnetization is transfered between spins in multiple steps via spin diffusion. Also, the intramolecular mobility and chemical exchange are additional factors not taken into account by Egn. 2.6. As a consequence, Eqn. 2.7 cannot be used to determine precise proton-proton distances. Instead, as an extension of Eqn. 2.7, NOEs are usually treated as upper (U) and lower (L) bounds on interatomic distances rather than precise distance restraints:

$$\begin{aligned} U &= \left(\frac{d_{ref}^{-6}}{V_{ref}} V\right)^{-1/6} + \Delta^{+} \\ L &= \left(\frac{d_{ref}^{-6}}{V_{ref}} V\right)^{-1/6} - \Delta^{-} \end{aligned}, \tag{2.8}$$

where $\Delta^{+/-}$ are error estimates and d_{ref} and V_{ref} are reference distance and volumes, respectively. There are several possible choices for the reference distances: fixed distances defined by the covalent geometry, distances derived from the distribution of the backbone-backbone distances or average distances from all distances smaller than a cutoff in an ensemble of model structures (Nilges and O'Donoghue, 1998). Additionally, spin diffusion can be taken into account by a relaxation matrix approach based on the simulation of the NOE spectrum from the intermediate model structures to derive correction factors for Eqn. 2.8 (Linge et al., 2004).

2.2.2 Residual dipolar couplings

For a macromolecule in liquid solution which experiences restricted orientational sampling, due to the presence of a liquid crystal or due to the paramagnetic properties of the molecule, strong first order interactions such as chemical shift anisotropy of dipolar coupling are no longer averaged to zero as in the case of an isotropic solution (Saupe and Englert, 1963; Gayathri et al., 1982). While partial alignment will affect any first order phenomenon, the most important application of non-isotropic averaging is the measurement of residual dipolar couplings (Tolman et al., 1995; Tjandra and Bax, 1997). The intrinsic strength of the dipolar coupling interaction allows measurable effects under conditions of week alignment where the solution properties necessary for high resolution NMR can be retained.

The Hamiltonian of the static dipolar interaction between two spins (I,S) in magnetic field depends on the angle of the internuclear vector relative to the magnetic field. In solution NMR, the measured dipolar coupling is described by the time and ensemble average of the dipolar Hamiltonian over all sampled orientations. For isotropic tumbling the average reduces to zero. Under the condition of partial alignment, where a preferential orientation of the molecule relative to the static magnetic field exists, the average is a convolution of the restricted motion of the molecule and the orientation of the vector with respect to the molecule. Assuming a fixed molecular shape the resultant residual dipolar coupling can be expressed in the terms of the orientation (θ,ϕ) of the internuclear vector relative to the alignment tensor attached to the molecule:

$$D^{IS}(\theta,\phi) = D_a^{IS}\left\{\left(3\cos^2\theta - 1\right) + \frac{3}{2}R(\sin^2\theta\cos 2\phi)\right\}, \tag{2.9}$$

where D_a^{IS} and R are the axial component and the rhombicity of the alignment tensor, respectively. For a given fixed distance $I-S$ (*e.g.* $N-HN$), the extreme values D^{IS} correspond to the orientation of the $I-S$ vectors closest to the z ($\theta = 0^{\circ}$) and y ($\theta = 90^{\circ}$and $\phi = 90^{\circ}$) axes of the alignment tensor. If the $I-S$ vectors are distributed

uniformly and isotropically, a histogram describing the probability of finding values of D^{IS} between these two extreme will have the same shape as the CSA (chemical shift anisotropy) powder pattern. The following overdetermined system of equations links the axial component and rhombicity values of the alignment tensor to the singularities of the D^{IS} histogram (Clore et al., 1998b):

$$\begin{aligned} D_{zz}^{IS} &= 2D_a^{IS} \\ D_{yy}^{IS} &= -D_a^{IS}(1+1.5R) \\ D_{xx}^{IS} &= -D_a^{IS}(1-1.5R) \end{aligned} \quad , \tag{2.10}$$

where D_{zz}^{IS}, D_{yy}^{IS} are the average high and low extreme values and D_{xx}^{IS} is the most populated average value of the residual dipolar couplings histogram, respectively. With two unknowns (D_a^{IS}, R) and three observables (D_{xx}^{IS} , D_{yy}^{IS}, D_{zz}^{IS}), the values of axial component and rhombicity of the alignment tensor can be calculated by nonlinear least-squares optimization.

As a consequence of Egn. 2.9 the residual dipolar couplings (RDCs) provide angular restraints between all $I-S$ vectors and the alignment tensor frame. This represents useful long range geometric information because the (I,S) atoms can be far away in the space. Residual coupling restraints can be incorporated into the structure calculation protocols by minimizing the difference between the observed and back-calculated values (Tjandra et al., 1997b). Although the size of the alignment tensor (rhombicity and axial component) can be derived from the distribution of the experimental dipolar couplings, its orientation with respect to the coordinate system of the molecule is unknown at the beginning of structure determination. This may cause convergence problems in the structure calculation process. As an alternative, the dipolar couplings can be translated into intervector projection angle restraints, which are independent of the orientation of the alignment tensor with respect to the molecule (Meiler et al., 2000).

In the case where the residual dipolar coupling histogram is sparse and the deter-

mination of the most populated value D_{xx}^{IS} ambiguous, the extreme values D_{zz}^{IS} and D_{yy}^{IS} can be used to estimate the alignment tensor components. The initial estimation should be iteratively corrected in several rounds of structure calculations based on the observation that maximum residual coupling value D_{zz}^{IS} can be underestimated with up to 15-20% (Clore et al., 1998a).

2.2.3 Scalar couplings

The scalar coupling constants between atoms separated by three covalent bonds ^3J are related to the enclosed torsion angle θ by the Karplus equation (Karplus, 1963):

$$^3J(\theta) = A\cos\theta + B\cos\theta + C, \tag{2.11}$$

where A, B and C are empiric parameters which must be optimized for various types of couplings and residues based on the best fit between the measured ^3J values and the corresponding value calculated with Egn. 2.11 for known protein structures.

In contrast to NOEs and RCDs, scalar couplings give geometrical information only for the local conformation of the polypeptide chain (confined to angles between 3 neighboring atoms). However, they are extremely important for an accurate definition of the local conformation of the backbone (ϕ and ψ angles), to obtain stereospecific assignments for the stereotopic protons (usually β protons) and to detect torsion angles that occurs in multiple states (usually χ^1).

2.2.4 Hydrogen bonds

The slow hydrogen exchange in proteins is often caused by the fixation of the amide protons in hydrogen bonds. The acceptor oxygen atom is frequently identified by a careful analysis of the NOE connectivities between the neighboring protons belonging to regular secondary structure elements. More reliable experimental proof is obtained from the small hydrogen bond scalar couplings ($^{3h}J_{NCO}$) recorded in predeuterated pro-

teins between the hydrogen bond donating amides and the accepting carbonyl groups (Wang et al., 1999). The scalar coupling confirms the overlap between the electronic orbitals of the atoms involved and unambiguously defines the pairs of atoms forming the hydrogen bond.

The hydrogen bonds are used as distance restraints for structure calculations, typically by restraining the acceptor-hydrogen distance to 1.8-2.1 Å and the acceptor-donor distance to 2.7-3.0 Å. As tight medium or long range restraints their impact on structure determination is considerable. Restraints for architectural hydrogen bonds in secondary structures enhance the regularity of the secondary structure elements.

2.2.5 Chemical shifts

Chemical shifts are very sensitive probes for the chemical environment of nuclear spins. However, since quantitative correlation between chemical shifts and protein structures has been difficult (Williamson and Asakura, 1997), empirical approaches which attempt to link the chemical shift information to the protein architecture using databases of high resolution protein structures were developed.

TALOS is a commonly used computer program for empirical prediction of ϕ and ψ backbone torsion angles using a combination of five chemical shifts (Hα, Cα, Cβ, CO, N) and the protein amino acid sequence (Cornilescu et al., 1999). The program uses the chemical shifts of three consecutive residues (*i.e.* 15 chemical shifts) to make predictions for the central residue in the triplet by searching the protein database for a similar combination of chemical shift values. The search is evaluated by computing a similarity factor based on the sum of square differences between the chemical shifts in the target protein and the database entries. In addition, the similarity score includes a qualitative residue-type term to bias towards similar sequences. TALOS database contains 20 protein structures for which both a high-resolution X-ray crystal structure and almost complete NMR resonance assignments were available. The program

searches the database for the best 10 matches to a given triplet in the target protein and makes a prediction if 9 out the 10 pairs of ϕ and ψ angles fall in the same region of the Ramachandran map. The average and the standard deviation of these ϕ and ψ values provide an empirical estimation of the backbone torsion angles. A torsion angle prediction is considered unambiguous when its standard deviation does not exceed 45^{o}.

2.3 Structure calculation algorithms

2.3.1 Simulated annealing with molecular dynamics

NMR spectroscopy is not a 'microscope with atomic resolution' which produces an image of a protein but rather a technique which provides a wealth of indirect structural information from which the protein three-dimensional structure can be obtained. The calculation of a protein structure from NMR data represents a minimization problem for a target function which measures the agreement between a conformation and a given set of experimental restraints (NOEs, RDCs, *J*-couplings and hydrogen bonds). Owing to the complexity of the problem (a protein typically consists of several thousand atoms) an exhaustive search of the allowed configurations is not feasible. Instead, a variety of non-linear optimization techniques adapted to this specific minimization problem have been developed: the metric matrix distance geometry approach, the variable target function method and simulated annealing in conjunction with molecular dynamics in Cartesian or torsion angle spaces (see Güntert, 1998 for a thorough review).

The most efficient minimization algorithm for NMR structure calculations is simulated annealing combined with molecular dynamics (SA-MD). In Cartesian coordinates, the SA-MD minimization consists of finding the numerical solution of the Newton's equation of motion:

$$m_i \frac{\mathrm{d}^2 \overrightarrow{r_i}}{\mathrm{d}t^2} = -\nabla_i E_{hybrid}, \tag{2.12}$$

where $\overrightarrow{r_i}$ and m_i are the position vectors and the masses of atoms forming the molecular system, and E_{hybrid} is the hybrid target function of the minimization problem (the potential energy of the system). The hybrid target function contains contributions from both experimental data and *a priori* knowledge of local architecture defined by the covalent geometry (the force field):

$$\begin{aligned} E_{hybrid} \;=\; & w_b \sum_{bonds} E_b + w_a \sum_{angles} E_a + w_i \sum_{improper} E_i + w_{nb} \sum_{non-bonded} E_{nb} \\ & + w_{dr} \sum_{distance\,restraints} E_{dr} + w_{ar} \sum_{angle\,restraints} E_{ar}, \end{aligned} \tag{2.13}$$

where w_b, w_a, w_i, w_{nb}, w_{dr} and w_{ar} are the weighting factors of the force field and experimentally derived geometric constraints. The principles of classical mechanics ensure a convergent trajectory of the molecular system towards its minimum potential energy and therefore solving Egn. 2.12 is equivalent with minimizing the pseudo-energy target function. The molecular dynamics minimization is coupled with simulated annealing (heating and slow cooling of the molecular system) to provide the kinetic energy necessary to cross barriers of the potential surface, thereby reducing the problem of becoming trapped in local minima. During the high temperature stage an approximate structure is calculated and as the temperature decreases the model is gradually refined. Because the temperature, *i.e.* kinetic energy, determines the maximal height of the energy barrier that can be overcome in the molecular dynamics simulation, the simulated annealing schedule is important to avoid local minima. Consequently, complex protocols of MD-SA have been designed for efficient protein NMR structure calculations (Güntert, 1998; Nilges and O'Donoghue, 1998).

The principles of molecular dynamics can be applied in torsion angle space using

torsion angles instead of Cartesian coordinates as degrees of freedom (the Newton equation is replaced by the Lagrange equation). Molecular dynamics in torsion angle space (torsion angle dynamics, TAD) has two main advantages: it reduces the degrees of freedom and fixes the covalent geometry (the high force constants used to maintain the covalent geometry in Cartesian dynamics lead to high vibrational frequencies and consequently longer time steps for the numerical integration). The simulated annealing with torsion angle dynamics (SA-TAD) provides at present the most efficient way to calculate NMR structures of macromolecules.

2.3.2 Iterative NOE assignment and structure calculation

Protein structure determination has been a driving force for NMR spectroscopy. The flow of actions for protein NMR structure determination includes: sample preparation, NMR experiments, spectrum calculation, peak picking, chemical shift assignment, NOE assignment and collection of other conformational restrains, structure calculation and structure refinement (Wüthrich, 1986). Despite several new computational approaches to circumvent the chemical shift and NOE assignment (Kraulis, 1994; Atkinson and Saudek, 2002; Grishaev and Llinas, 2004), up to now all *de novo* protein NMR structure determinations were conducted following the 'standard' procedure. However, the iterative NOE assignment and structure calculations based on chemical shift information has proven to be more accessible to automatization (Güntert, 2003).

One of the most time-consuming steps in NMR structure determination is the interpretation of the NOESY spectrum, *i.e.* the NOE assignment, where pairs of hydrogen atoms that correspond to the experimental NOESY cross peaks are identified based on the previous sequence specific resonance assignments. The number of NOEs that can be assigned based on the chemical shift information alone is restricted by the accuracy of the NOESY cross peak positions and chemical shift values. Because of the limited accuracy of chemical shift values and peak positions many NOEs cannot be attributed

to a single proton pair but have an ambiguous assignment comprising of several proton pairs (often referred to as dispersion degeneracy). In addition, a poor chemical shift dispersion (determined by the similar local environment of different protons, *e.g.* α-helical regions) and a high number of NOEs (determined by the protein size) may increase the problem complexity. In general it is impossible to assign all the NOESY peaks unambiguously based on chemical shifts, not even in very small proteins (Mumenthaler et al., 1997). For manual NOE assignment an iterative process has typically been used, in which preliminary protein structures calculated from a fraction of the NOE derived distance restraints help to reduce the ambiguity of the remaining cross peak assignments. Automated NOE assignment and structure calculation approaches follow the same general scheme, although without manual intervention. They all have three main features in common: a method to deal with the inherent chemical shift ambiguity of the NOE data, a noise filter for spurious NOESY cross peaks and an assignment filter which gradually reduces the dispersion degeneracy of the NOE signals. Two of the most commonly used computer programs for interactive NOE assignment and structure calculations are CYANA[1] (Güntert, 2004) and ARIA[2] (Habeck et al., 2004). Since ARIA was chosen for protein NMR structure calculations presented in this thesis, the following paragraphs will describe the ARIA approach.

The cornerstone of the ARIA algorithm is the concept of ambiguous distance restraints (ADRs) introduced for handling ambiguities in chemical shift based NOE assignment (Nilges, 1995). For ambiguous distance restraints every NOESY cross peak is treated as a superposition of the signals corresponding to different assignments allowed by the frequency tolerances. The ADR is defined by an effective distance $\overline{D}$, which contains contributions from distances between all pairs of protons which are possible assignments:

[1] Combined assignment and dYnamics Algorithm for NMR Applications.
[2] Ambiguous Restraints for Iterative Assignment.

$$V_{NOE} = \sum_{i=1}^{N} V_i = k \sum_{i=1}^{N} d_i^{-6} \Rightarrow \overline{D} = \left(\sum_{i=1}^{N} d_i^{-6} \right)^{-1/6}, \tag{2.14}$$

where V_{NOE} is the NOESY cross peak volume of the given NOE, N is the number of assignment possibilities within chemical shift tolerances, V_i is the cross peak volumes of the hypothetic assignment i, and k is the NOE calibration constant in the isolated spin pair approximation (see Eqn. 2.7). Because the effective distance $\overline{D}$ (also referred to as 'd^{-6} summed distance') is always shorter than each of the individual distances d_i, an ambiguous distance restraint is never misinterpreted by including incorrect assignment possibilities as long as the correct assignment is present. In addition, the ambiguous distance restraints allow a straightforward modality to include the additional symmetry degeneracy present in symmetric oligomers where symmetry related protons have identical chemical shifts:

$$\overline{D} = \left(\sum_{j=1}^{M} \sum_{i=1}^{N} d_i^{-6} \right)^{-1/6}, \tag{2.15}$$

where M is the number of monomeric units forming the symmetric oligomer and N is the number of assignment possibilities allowed by the frequency tolerances. During structure calculations, the effective distances can be restrained in a similar way as distances between protons by using a 'flat-bottom' harmonic potential with an asymptotic region for large violations where the function becomes linear:

$$E_{NOE}(\overline{D}) = w_{NOE} \begin{cases} (\overline{D} - L)^2, & \overline{D} < L \\ 0, & L \leq \overline{D} \leq U \\ (\overline{D} - U)^2, & U < \overline{D} < U + A \\ \alpha + \beta(\overline{D} - U)^{-1} & \\ \quad + \gamma(\overline{D} - U), & \overline{D} \geq U + A \end{cases}, \tag{2.16}$$

where w_{NOE} is the NOE potential weight within the overall target function, U and L are the upper and lower limits defined by Eqn. 2.8, A is a parameter that determines the

cutoff distance where the potential switches from harmonic to asymptotic behavior, γ is the slope of the asymptotic potential, and the coefficients α and β are defined such that the potential is continuous and differentiable at $\overline{D} = U + A$. The asymptotic-linear potential allows large violations transiently, thus allowing the structure to escape from local minima.

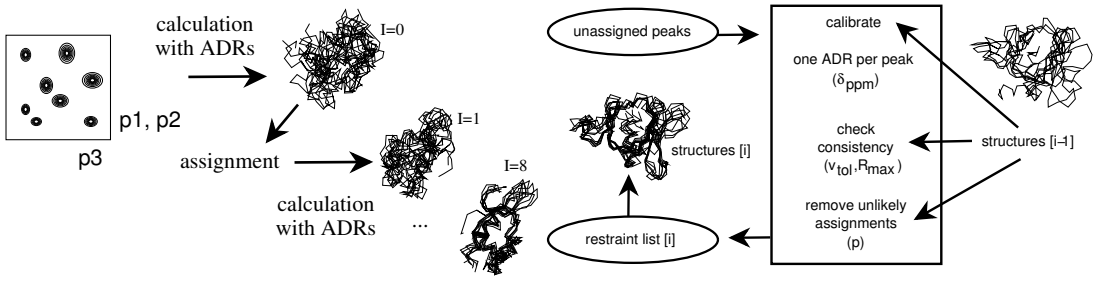

Figure 2.1. Schematic representation of the ARIA algorithm. The NOESY cross peaks are iteratively assigned in nine cycles of structure calculations. Iteration zero defines a structural model using ambiguous distance restraints (ADRs). In each later iteration the original NOE data is calibrated, assigned, filtered against the previous models and an improved generation of structures are calculated.

The ARIA approach comprises of the following steps successively repeated along nine iterations of coupled NOE assignment and structure calculations (Figure 2.1):

1. Read the full lists of NOEs and chemical shift assignments. In every iteration the program uses the original, unassigned NOE lists.

2. For each NOE all possible assignments allowed by the frequency tolerance δ_{ppm} are used to define an ambiguous distance restraint (ADR).

3. Select the S_{conv} lowest energy structures from the previous iteration.

4. Convert the peak volumes into distance restraints by automatic calibration.

5. Extract the average distance d_{aver} for each assignment possibility based on the S_{conv} structures.

6. Apply a structural consistency check filter based on d_{aver} to remove the artificial peaks. The noise filter is tolerant in early iterations and becomes more stringent towards the end of the iterative procedure.

7. Discard the unlikely assignment possibilities based on an average distance cutoff which is gradually decreased over the nine iterations.

8. Calculate the new generation of structures (typically 20) with the CNS[3] program (Brünger et al., 1998) using the updated set of distance restraints.

The conversion of NOESY cross peak volumes to distance restraints is achieved by estimating a reference distance (d_{ref}) and a reference volume (V_{ref}, see 2.8). d_{ref} is computed as the $< d^6 >^{-1/6}$ average over all values for which the target distance is smaller than a cutoff of 6 Å in the previously calculated structures and V_{ref} is given by the arithmetic average over all corresponding volumes. Due to ensemble averaging, the reference distances and volumes do not change much from iteration to iteration.

The noise filter is based on the structural consistency hypothesis which discards NOEs violated with more than a threshold μ in more than a fraction R_{max} (typically 0.5) of the converged structures S_{conv}. To allow violations caused by the insufficient convergence of the structure calculation algorithm, μ has to be gradually decreased during the iterative procedure. It is usually set to values between 10 Å in the first iteration and 0.1 Å in the last iteration. The robustness of the method against noisy restraints becomes particularly important if the structure determination is attempted from automatically picked peak lists. Some noise can be rejected in a trivial way, they fall outside the frequency tolerance (δ_{ppm}) of any assigned resonance. NOESY spectra de-noising prior to automated peak picking may also improve the peak list quality. However, if the resonance assignment is not complete, even the most carefully prepared peak list will contain peaks that cannot be correctly assigned.

[3] Crystallography and NMR System.

Although ADRs which contain wrong together with at least one correct assignment are compatible with the correct structure, it is important to reduce the ambiguity of the NOE assignments as much as possible because the additional assignment possibilities 'dilute' the structural information and make it more difficult to the structure calculation algorithm to converge to the correct structure. Therefore a structural based assignment filter is applied in each iteration with the aim of discarding the assignment possibilities which are incompatible with the previous generation of three-dimensional structures. This filter is based on the relative contributions (c_n) of different assignment possibilities to the peak volume:

$$c_n = \frac{d_{aver,n}^{-6}}{\sum_{i=1}^{N} d_{aver,i}^{-6}}, \tag{2.17}$$

where $d_{aver,i}$ is the average distance of the i^{th} assignment possibility in the previously calculated structures, and N the number of contributions to the given ADR allowed by the frequency tolerance. To obtain a partial assignment the relative contributions are ordered by decreasing size and the smallest contributions are discarded such that: $\sum_{i=1}^{N_p} c_i > p$, where p is the assignment cutoff and N_p the number of contributions necessary to account for a fraction of the peak volumes larger than p. The parameter p is decreased form cycle to cycle and takes values between 1.0 and 0.8. For a hypothetical NOE with two assignment possibilities and with the shorter of the two distances of 2.5 Å, a value $p = 0.999$ will exclude a second distance of 7.9 Å, a value of $p = 0.95$ a second distance of 4.1 Å and a value $p = 0.8$ a second distance of 3.3 Å. If the shorter distance is 4 Å, the corresponding minimal excluded distances are 12.6, 6.6 and 5.2 Å, respectively.

The ground iteration of ARIA (iteration zero) is the key moment where structures are calculated based on the ambiguous distance restraints defined by the frequency tolerances alone. To ensure that the calculation will converge either additional unambiguous structural restraints (manually defined) or carefully prepared NOE (low noise

content) and chemical shift lists (tight tolerances) has to be supplied. In the first case the program uses the unambiguous structural information to build the initial model, whereas in the former case the tight chemical shift tolerances provide a reasonable low ambiguity level for the NOE data. From this respect, the approach used in CYANA has a principial advantage by adding two new features which reduce the NOE potential hypersurface and significantly improve the convergence rate. These two features are as follows: *(I)* a pre-filtering of the NOE assignment list based on the concepts of 'network anchoring' which requires that any given NOE should be part of a self-consistent subset of NOEs, and a 'symmetry mapping' which is based on the fact that NOESY spectra are symmetric with respect to their diagonal, the presence of the symmetry related partner being a criterion to chose between different assignment possibilities; *(II)* a restraint combination, which aims to minimize the impact of wrong assignments on the expense of the temporary loss of information. Despite these conceptual improvements CYANA performs well only for almost complete chemical shift assignments lists (about 90%) and clean NOESY cross peak lists (Jee and Güntert, 2003). The drawbacks of this method are the result of the excessive filtering of the NOESY lists against chemical shift and peak list oriented criteria. A compromise between peak list filtering and the completeness of the chemical shift assignment has to be found for each particular structure calculation project. Therefore, a method capable of switching on and off the peak list oriented filtering and using alternative de-noising strategies coupled with efficient automated peak picking leads to a step forward for the full automatization of the NOE assignment process. This approach has been pursued in this thesis by combining the wavelet de-noising of NOESY spectra with automated peak picking, peak integration and consistency check of the NOEs list in the frame of the ARIA program.

2.4 Numerical analysis algorithms

2.4.1 Multiresolution analysis and wavelet series expansion

The multiresolution analysis (MRA) as introduced by Mallat (Mallat, 1989a,b, 1998) provides a general framework to construct wavelet bases suitable to describe functions at different resolution levels. MRA is a sequence of nested spaces $\{V_j\}_{j\in\mathbb{Z}}$ which approximate the space $\mathcal{L}_2(\mathbb{R})$ of all square integrable functions[4] with increasing resolution. The first step is to define a scaling function (father wavelet) ϕ in such a way that the family $\{\phi_{0,k} = \phi(x-k),\, k \in \mathbb{Z}\}$ forms an orthonormal base for the reference space V_0. Except for the Haar wavelet basis[5] for which ϕ is the characteristic function of the interval $[0,1)$, the scaling function is chosen to satisfy certain continuity, smoothness and tail requirements. The functions of V_0 can be written as: $f(x) = \sum_k c_k \phi(x-k)$.

Starting from V_0 linear spaces can be defined:

$$\begin{aligned} V_1 &= \{g(x) = f(2x) \;:\; f \in V_0\} \\ &\ldots \\ V_j &= \{g(x) = f(2^j x) \;:\; f \in V_0\}. \end{aligned} \tag{2.18}$$

The set $\{\phi_{1,k},\, k \in \mathbb{Z}\}$ is an orthonormal basis in V_1 with $\phi_{1,k}(x) = \sqrt{2}\phi(2x-k)$. Analogously, the basis functions of V_j are $\phi_{j,k} = 2^{j/2}\phi(2^j x - k)$. In this way $\phi_{j,k}$ generates a sequence of spaces $\{V_j,\, j \in \mathbb{Z}^+\}$ which are nested:

$$\begin{aligned} V_0 \subset V_1 \subset \ldots \subset V_j \subset \ldots \\ V_j \subset V_{j+1},\, j \in \mathbb{Z}^+. \end{aligned} \tag{2.19}$$

If in addition every square integrable function can be approximated by functions in

[4] $\mathcal{L}_2(\mathbb{R})$ is the space of complex valued functions f on $\mathbb{R}$ with a finite norm: $||f||_2 = \left(\int_\infty^\infty |f(x)|^2 dx\right) < \infty$.

[5] The oldest wavelet basis introduced by the Hungarian mathematician Alfred Haar in 1909. For the Haar wavelet basis the scaling function is: $\phi(x) = 1(0 \le x < 1)$.

$\bigcup_{j\geq 0} V_j$ than $\{V_j,\, j \in \mathbb{Z}^+\}$ is a MRA[6].

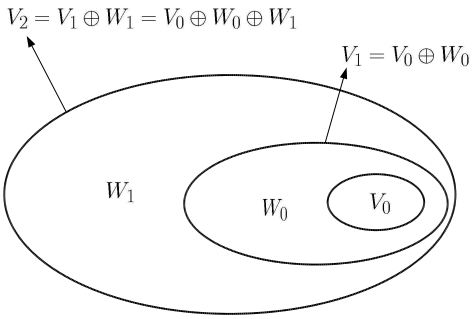

Figure 2.2. Sequence of nested spaces V_0, V_1, V_3 and their orthogonal complements W_0 and W_1.

The nested spaces V_0 and V_1 define W_0 as the orthogonal complement of V_0 in V_1: $V_1 = V_0 \bigoplus W_0$ (Figure 2.2). Because $V_0 \subset V_1$, any function in V_0 can be written as a linear combination of base functions $\phi(2x-k)$ from V_1 and in particular:

$$\phi_{0,k}(x) = \sum_k h(k)\sqrt{2}\phi(2x-k), \tag{2.20}$$

where the coefficients $h(k)$ are defined by the scalar product $\left\langle \phi(x), \sqrt{2}\phi(2x-k) \right\rangle$. Analogously, one can define a set of functions $\psi_{0,k}$ for the orthogonal complement W_0 as:

$$\psi_{0,k}(x) = \sum_k (-1)^k h(-k+1)\sqrt{2}\phi(2x-k). \tag{2.21}$$

It can be shown that $\{\psi_{0,k}(x) = \psi(x-k),\, k \in \mathbb{Z}\}$ forms an orthonormal basis for W_0. The same process can be repeated for higher values of j (Figure 2.2). This leads to

[6] $\bigcup_{j\geq 0} V_j$ is dense in $L_2(\mathbb{R})$.

consecutive summation of subspaces:

$$\begin{aligned} V_{j+1} &= V_j \oplus W_j \\ &= V_j \oplus W_{j-1} \oplus W_j \\ &= V_0 \oplus W_0 \oplus W_1 \oplus \ldots \oplus W_j \\ &= V_0 \oplus \bigoplus_{l=0}^{j} W_l. \end{aligned} \tag{2.22}$$

Owing to the similarity property of MRA, $\{\psi_{j,k},\, k \in \mathbb{Z}\}$ is an orthonormal basis in W_j, where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. Since the sum of nested spaces spans the space of square integrable functions: $\mathcal{L}_2(\mathbb{R}) = V_0 \oplus \bigoplus_{l=0}^{j} W_l$, the family $\{\psi_{j,k},\, k \in \mathbb{Z}\}$ is a basis for $\mathcal{L}_2(\mathbb{R})$. For any given function $f \in \mathcal{L}_2(\mathbb{R})$, one can find j such that $f_j \in V_j$ approximates f up to a preassigned precision in terms of $\mathcal{L}_2$ closeness. If $w_i \in W_i$ and $v_i \in V_i$, then:

$$f_j = v_j + w_j = v_{j-k,k\leq j} + \sum_{i=1}^{k,k\leq j} w_{j-i} = v_0 + \sum_{i=1}^{j} w_{j-i}, \tag{2.23}$$

which gives the wavelet decomposition of f. The properties and the functional form of the wavelet base functions $\{\psi_{j,k},\, k \in \mathbb{Z}\}$ are determined by the properties of the chosen father wavelet ϕ.

In summary, starting from a father wavelet (scaling function) ϕ an orthonormal mother wavelet ψ is obtained. Dyadic dilatations (2^j) yield nested subspaces which form a MRA. The base functions ψ_{jk} are derived by additional translations (k):

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k). \tag{2.24}$$

The wavelet base functions have compact support, i.e. the wavelet is zero outside a finite interval $\left[k \cdot 2^{-j}, (k+1) \cdot 2^{-j}\right)$ and form an orthonormal basis for $\mathcal{L}^2(\mathbb{R})$. Therefore any square integrable function $f(x) \in \mathcal{L}^2$ can be represented as a series of $\psi_{j,k}$

with the corresponding scaling function $\phi_{0,k}$:

$$f(x) = \sum_k \alpha_{0,k}\phi_{0,k}(x) + \sum_j \sum_k \beta_{j,k}\psi_{j,k}(x), \tag{2.25}$$

where the scaling $\alpha_{0,k}$ and the wavelet $\beta_{j,k}$ coefficients are defined by:

$$\alpha_{0,k} = \int_0^1 f(x)\phi_{0,k}(x)dx, \quad \beta_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)dx. \tag{2.26}$$

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

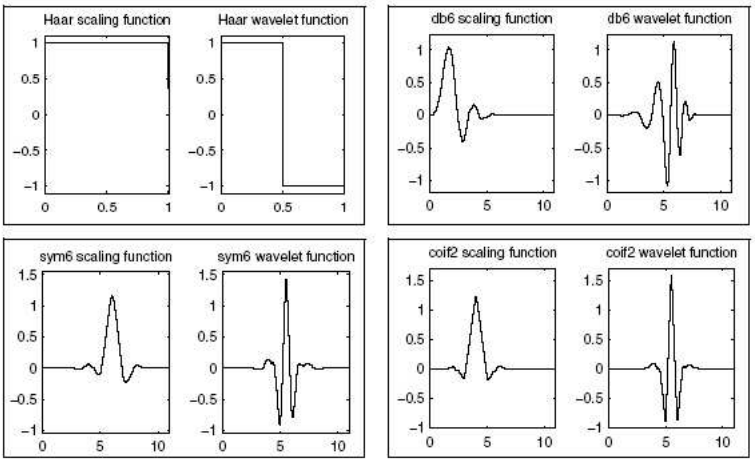

Figure 2.3. Graphic representation of Haar, Daubechies 6, Symmlet 6 and Coiflet 6 wavelets.

Daubechies wavelets (D) , Coiflets (C) and Symmlets (S) are the most commonly used wavelets which fulfill these basic requirements without the discontinuity of the Haar wavelet (Figure 2.3). They were originally designed to represent smooth functions with a sparse set of coefficients.

2.4.2 Discrete wavelet transform

Practical applications are usually involving discretely sampled rather than continuous functions. The extension of the wavelet series expansion to discretely sampled functions leads to the discrete wavelet transform (DWT), which can represented in a matrix form as:

$$d = \mathbf{W} f, \tag{2.27}$$

were $f = \{f_1, f_2, ..., f_N\}'$ is the original signal represented as a column vector of $N = 2^n$ discrete data points, d is a $N \times 1$ vector comprising both the discrete scaling coefficients, $\alpha_{0,k}$, and the discrete wavelet coefficients, $\beta_{j,k}$. $\mathbf{W}$ is a $N \times N$ orthogonal transformation matrix defined by the chosen orthonormal wavelet basis.

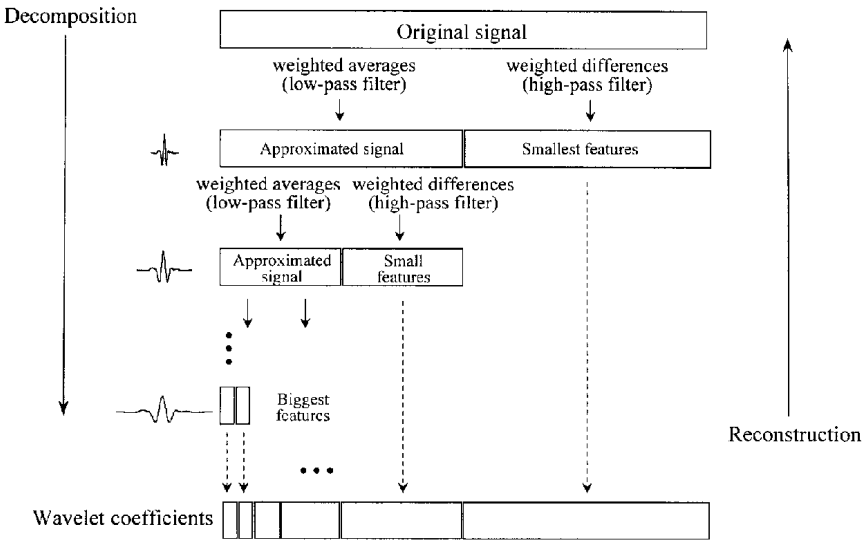

Figure 2.4. Schematic representation of Mallat's pyramidal algorithm for fast discrete wavelet transform.

A computationally efficient algorithm for fast discrete wavelet transform is the Mallat's pyramidal wavelet decomposition and reconstruction (Mallat, 1989b). The connection between the fast discrete wavelet transform and MRA is described by the operator representation of the quadrature mirror filters, known as the low-pass (L) and

the high-pass (H) filters which are specifically defined by the chosen orthonormal wavelet basis. If $f^{(n)}$ is the original signal (of 2^n data points), at each stage the wavelet decomposition moves to a coarser approximation, i.e $f^{(n-1)} = Lf^{(n)}$and $d = Hf^{(n)}$, where $d^{(n-1)}$ is the detail lost by approximating $f^{(n)}$ by the averaged $f^{(n-1)}$. In this way the discrete wavelet decomposition of $f^{(n)}$ is represented as another sequence of length 2^n, where the coarser approximation, $f^{(n-1)}$, has only half of the original signal length. This procedure can be continued until one approximation coefficient remains (Figure 2.4). Thus the DWT (the equivalent of Eqn. 2.27) can be summarized as:

$$\begin{aligned} f \quad \rightarrow \quad & (Hf, HLf, HL^2f, ..., HL^jf, ..., HL^{n-1}f, H^nf) \\ & = (d^{(n-1)}, d^{(n-2)}, ..., d^j, ..., d^1, d^0, f^0), \end{aligned} \tag{2.28}$$

where the 'detail' sequences d^j contain the wavelet coefficients ($\beta_{j,k}$). The original signal can be reconstructed from the wavelet coefficients by reversing the filter operations.

2.4.3 Wavelet de-noising

Wavelet de-noising is based on the property of wavelets to represent signals with a set of coefficients which have desirable statistical properties in the suppression of noise (Daubechies, 1992). A substantial reduction of the noise level is achieved by applying a wavelet transform followed by a suppression of noise-related wavelet coefficients and backward wavelet transform (Figure 2.5). The most widely used methods to suppress noise-related coefficients are global hard- and soft-thresholding of the wavelet coefficients (Donoho and Johnstone, 1994; Donoho and Johstone, 1995). In hard-thresholding all coefficients below a threshold λ are zeroed (keep or kill), while in the soft-thresholding, in addition, all the other coefficients are also shrunk towards zero by subtracting λ (shrink or kill):

$$\beta_{jk,\text{hard}} = \begin{cases} \beta_{jk} & \text{if} \quad |\beta_{jk}| > \lambda \\ 0 & \text{if} \quad |\beta_{jk}| \leq \lambda \end{cases}, \qquad \beta_{jk,\text{soft}} = \begin{cases} \beta_{jk} + \lambda & \text{if} \quad \beta_{jk} < -\lambda \\ \beta_{jk} - \lambda & \text{if} \quad \beta_{jk} > \lambda \\ 0 & \text{if} \quad |\beta_{jk}| \leq \lambda \end{cases}. \tag{2.29}$$

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

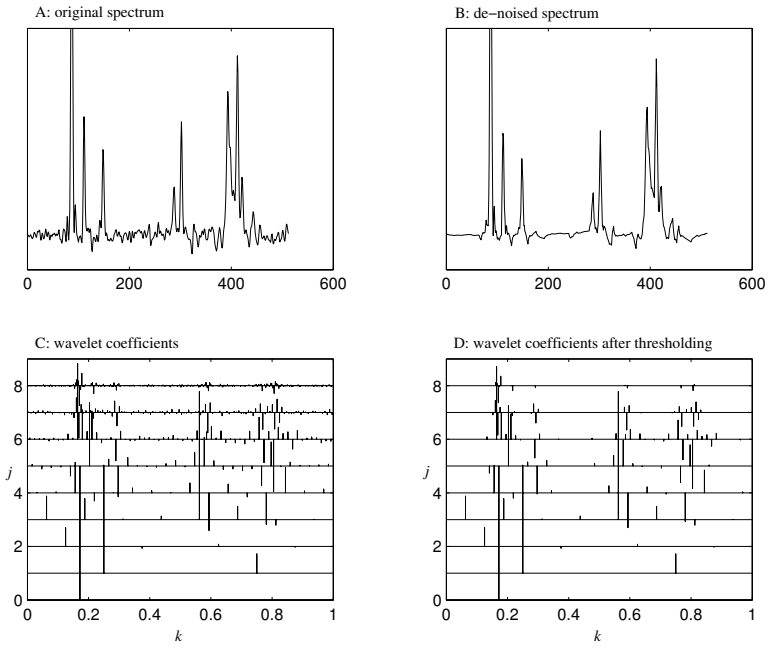

Figure 2.5. Schematic representation of wavelet de-noising*:* the DWT decomposes the original signal in wavelet coefficients (k) at different dyadic levels (j). Noise related coefficients are eliminated by thresholding and the spectrum is reconstructed by an inverse wavelet transform.

The first dyadic levels ($j = \{1,2,3,...\}$ in Eqn. 2.28) represent the low frequency

components of the signal, i.e. baseline and peak shape features. Therefore the suppression of the wavelet coefficients belonging to these levels is usually not desirable. For this reason a low-frequency cutoff J is applied to coefficient thresholding.

In addition to wavelet thresholding which has a smoothing effect on the spectra and suppresses noise it is also possible to apply a multiresolution analysis (MRA). This is based on the idea that a function or a signal can be approximated at different dilatation levels. MRA has previously been exploited for solvent suppression in NMR spectra (Günther et al., 2002). In a MRA only a subset of the resolution levels are used to restore the signal. This concept is useful to suppress low frequency components of the signal leading to baseline correction or high frequency signal components for smoothing/de-nosing.

2.4.4 Translation invariant wavelet transform

Wavelet suppression using hard- or soft-thresholding causes artifacts in the vicinity of the discontinuities introduced by suppressing individual coefficients (Gibbs phenomena). These artifacts can be attributed to the lack of translation invariance of the wavelet base. A simple method to average the translation dependence is 'cycle spinning' where data is shifted, de-noised and un-shifted. Subsequently the results for different shifts are averaged. A translation invariant (TI) transformation algorithm (Coifman and Donoho, 1995) was designed for fast cycle-spinning over all N points of the spectrum. In conjunction with de-noising, the TI wavelet transform has significant advantages, particularly when sharp signals in an NMR spectrum cause pronounced Gibbs artifacts.

2.4.5 Principal component analysis

Principal component analysis (PCA) is a linear transformation which can be used to visualize similarities and differences in large data sets. PCA can be applied to large data sets to detect similar groups of data, outliers and trends of changes within groups of data (Jackson, 1991; Wold et al., 1987). A PCA describes the variation in data with a minimum set of variables. Variables are often dependent on to each other and PCA reveals the latent variables which describe the underlying structure in the data.

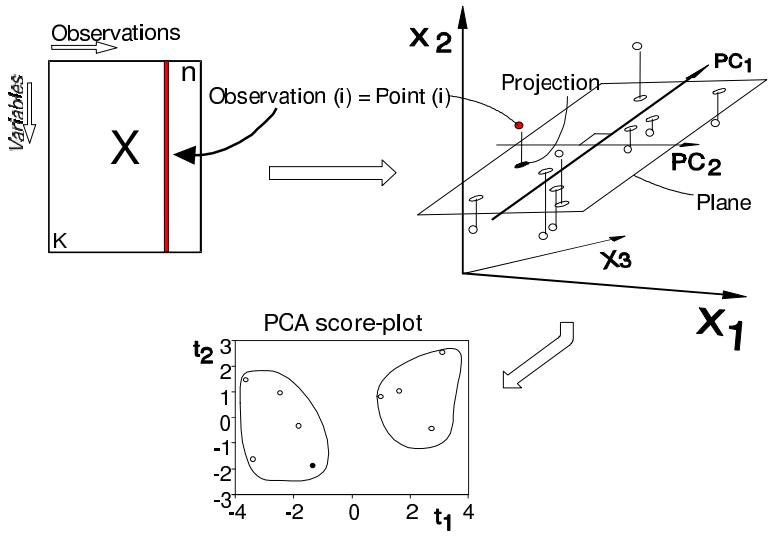

Figure 2.6. The three-dimensional representation of the data projection using principal components.

Figure 2.6 shows the principles of PCA in three dimensions. The data matrix X contains n objects described by k variables ($k = 3$). Each observation is represented by one point in the k-dimensional space, the matrix X forming a swarm of points in this space. The PCA is equivalent to a least squares fitting of orthogonal lines in the k-dimensional variables space. The first principal component is the line which best approximate the data, the second principal component improves the approximation of X as much as possible and so forth. The first two principal components are orthogonal to each other and form a plane in the space of X, a two-dimensional window into the k-dimensional space. The original points are projected onto this plane.

If X is the data matrix which contains n NMR spectra in columns with k frequencies in rows[7]

$$X = \begin{matrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{matrix}, \tag{2.30}$$

one can estimate a reference centered covariance matrix of X as:

$$C_X \equiv (X - X_r)(X - X_r)^T, \tag{2.31}$$

where X_r is the reference data matrix which represents a reference spectrum. The reference data set in spectra of proteins recorded for ligand screening is most commonly the spectrum of the free protein in absence of ligand. Alternatively, the mean over all spectra may be used.

For fully uncorrelated NMR spectra, all off-diagonal elements of the C_x matrix are zero and the diagonal elements represent the variances of the individual rows. For correlated data the off-diagonal elements are the covariances between the different spectra. The visualization of noisy data represents a high-dimensionality problem because noisy NMR spectra are never fully correlated. Multivariate statistics helps to visualize different degrees of correlation between noisy data.

To reduce the dimensionality of a series of NMR spectra contained in the matrix X one has to find a linear transformation $Y = MX$ to a new set of variables with a diagonal covariance matrix C_Y (so that each of its elements is uncorrelated). The covariance matrices of X and Y are related by:

$$C_X = M^T C_Y M. \tag{2.32}$$

Because C_Y is a diagonal and M is an orthonormal matrix, the columns of M^T are the

[7] In the MATLAB implementation this matrix is transposed. The two-dimensional NMR spectra ([^{1}H,^{15}N]-HSQC) were represented as vectors obtained by a concatenation of the ^{15}N strips.

eigenvectors of C_X while the diagonal elements of C_Y are the corresponding eigenvalues. If there are linear combinations among the elements of the original data matrix X then some of the eigenvalues in C_y will vanish. For highly correlated data the values of the Eigenvalues in Cy will be small. For screening data these small Eigenvalues represent spectra which are similar to the reference spectrum.

Using the covariance matrix of a set of experiments to find a transformation to a new set of uncorrelated variables is called *Principal Component Analysis* (PCA). PCA starts with the covariance matrix of all the original data and then eliminates the insignificant components.

Eqn. 2.32 is an eigendecomposition problem which can be solved with the *Singular Value Decomposition* (SVD) algorithm for the general case of a non-square eigenvectors matrix. SVD (Golub and van Loan, 1996) decomposes a real matrix A according to:

$$A = UWV^T, \tag{2.33}$$

where U and V are orthonormal matrices, *i.e.* $UU^T = I$ and $VV^T = I$ where I is the identity matrix. W is a square diagonal matrix containing the singular values of A. Proceeding from the singular value decomposition of the matrix A one can demonstrate that:

$$AA^T = UW^2U^T. \tag{2.34}$$

The SVD can be used to obtain a PCA of a covariance matrix. For a reference centered matrix $A = X - X_r$, matrix U in Eqn. 2.34 represents the orthonormal eigenvector matrix M^T and W^2 represents the corresponding eigenvalues C_y. Therefore, applying SVD on the reference centered matrix $A = X - X_r$ yields the transformation matrix M (also referred to as PCA loadings) and the diagonal eigenvalues of the covariance matrix C_x which represent the variances of the principal component representation. The

principal components (also referred to as PCA scores) of X are easily computed from: $Y = MX$.

3 Experimental procedures

3.1 NMR sample preparation for Sud protein

The expression and purification of the Sud protein was carried out using a similar protocol to that previously described (Klimmek et al., 1998) including a C-terminal His-tag of six residues. Uniform ^{15}N and $^{15}N/^{13}C$ labeling was performed by growing bacteria on isotope enriched minimal medium using ^{15}N ammonium chloride (Martek) and $^{13}C_3$ enriched glycerol (Martek) as main nitrogen and carbon sources. For protein samples labeled with $^{2}H/^{15}N/^{13}C$, the bacteria were grown on Celtone@-dCN (Martek, deuteration degree: 97 %). NMR samples of purified protein (0.6 - 1.2 mM dimer) were prepared in 50 mM sodium phosphate at pH 7.6, 1 mM polysulfide (S_n^{2-}), 13 mM sulfide, and 5% (v/v) D_2O. The protein was loaded with sulfur before transferring to the buffer solution described above. In order to exclude oxygen from the NMR probes, the sample tubes were flushed with nitrogen while filling and tightly sealed afterwards. Under these conditions the protein remains loaded with sulfur during the NMR experiments.

The asymmetrically labeled samples used for the measurement of the inter-monomer NOEs were prepared from unlabeled and $^{2}H/^{15}N(^{13}C)$-labeled Sud-His_6 dimers mixed in equal amounts at a very low concentration (each species 10 mM) in an anaerobic buffer containing 50 mM potassium phosphate and 10% (v/v) glycerol at pH 8.0. To induce the monomerization of the isolated dimers 0.02% (w/v) sodium dodecylsulfate was added. The mixture was stirred for 48 h at room temperature under anaerobic

conditions. To initialize the dimerization and to recover the protein the whole mixture was applied to a 10 ml Ni-nitrilotriacetic agarose (Qiagen) column equilibrated with 50 mM potassium phosphate and 10% (v/v) glycerol at pH 8.0. The column was extensively rinsed with the same buffer (0.5 l) for removing the SDS, and then the protein was eluted with this buffer containing 0.2 M imidazole. The eluted Sud protein was concentrated up to 30 g/l by pressure dialysis using a 10 kDa filter and the imidazole was removed by repeated dilution and concentration (five times) of the protein with a buffer containing 50 mM potassium phosphate at pH 7.65.

For residual dipolar coupling measurements, the isotropic sample contained 0.55 mM dimeric protein in 50 mM sodium phosphate at pH 7.6, 1 mM polysulfide (S_n^{2-}), 13 mM sulfide, and 10% (v/v) D_2O, while the anisotropic one had 0.48 mM protein in the same buffer plus the alignment medium C8E5/n-octanol (Rückert and Otting, 2000). The molar ratio of C8E5 to n-octanol was 0.87 and the C8E5/water ratio was 6% (w/v).

3.2 NMR sample preparation for Sud-Str complex

Three different samples were prepared for the Sud-Str complex (1:1 Sud monomer, Str protein): (*I*) in the absence of the polysulfide-sulfur substrate, (*II*) only the Str protein was loaded with polysulfide before complex formation and (*III*) both proteins were loaded with polysulfide. In the case of the second sample the Str protein was fully loaded with polysulfide and dialyzed over night in a potassium phosphate buffer to remove the loosely bound sulfur atoms. A second 24 hours dialysis did not produce any further polysulfide-sulfur removal suggesting a stable ligand bound form of the protein. Afterwards the Str protein was mixed with the ligand-free form of the Sud protein in a polysulfide-free buffer. These conditions guarantee that any polysulfide-sulfur attached to the catalytic cysteines of Sud protein must be a result of the transferase activity of the Str protein. The third sample was prepared using the polysulfide loaded forms of

the both proteins mixed in an anaerobic buffer containing an excess of polysulfide and sulfide.

3.3 NMR experiments

Unless stated otherwise, the NMR data was acquired at 300 K using Bruker DMX-600 and DRX-800 NMR spectrometers equipped with xyz-gradient ^{1}H, ^{15}N, ^{13}C triple resonance probe heads. The sensitivity and resolution of the triple resonance experiments was improved employing the TROSY technology (Pervushin et al., 1997; Salzmann et al., 1999). The software packages XWINNMR, AURELIA (Bruker Analytische Messtechnik GmbH, Karlsruhe), NMRLab (Günther et al., 2000) and NMRPipe (Delaglio et al., 1995) were used for data processing and data analysis. ^{1}H chemical shifts were referenced to internal DSS (2,2-dimethyl-2-silapentane-5-sulfonate sodium salt) while the ^{15}N and ^{13}C chemical shifts were calibrated indirectly using the appropriate gyromagnetic ratios (Wishart et al., 1995).

The ^{13}C side chain assignments for $^{2}H/^{13}C/^{15}N$-labeled Sud were based on 3D CC(CO)NH (Farmer and Venters, 1995) and CC(CA)NH (Löhr and Rüterjans, 2002) experiments with ^{13}C spin-lock times of 21 ms and 17 ms, respectively. The $^{1}H_{\alpha}$ chemical shifts were obtained from a 3D HCACO experiment and the ^{1}H side chain resonances were assigned using 3D H(C)CH-COSY and H(C)CH-TOCSY experiments with a ^{13}C spin-lock time of 17 ms on a uniformly $^{15}N/^{13}C$ labeled protein. The aromatic proton resonances were obtained via a 2D NOESY with a mixing time of 70 ms, a 2D TOCSY with 44 ms ^{1}H spin-lock time on an unlabeled sample in D_2O and a 3D ^{13}C-separated NOESY HSQC experiment with a mixing time of 70 ms employing a constant-time [^{13}C, ^{1}H]-TROSY evolution period (Pervushin et al., 1998) optimized for aromatic carbons on a $^{15}N/^{13}C$ labeled sample in H_2O.

Stereospecific assignments for the isopropyl groups of Val and Leu residues were determined using a biosynthetic approach (Neri et al., 1989) based on the ^{13}C-^{13}C

one-bond couplings observed in 2D ^{13}C-HSQC and 2D constant-time ^{13}C-HSQC experiments on a 10% ^{13}C-labeled sample. The NOE assignments and distance restraints for NH-NH correlations were obtained from a 4D ^{15}N/^{15}N-separated NOESY spectrum (Venters et al., 1995; Grzesiek et al., 1995) recorded with a mixing time of 300 ms on uniformly ^{1}H/^{15}N labeled protein in H_2O. Additional NOE data was collected from a 3D ^{13}C-separated NOESY-HSQC with a mixing time of 80 ms using a uniformly ^{13}C/^{15}N labeled protein in D_2O, from a 3D ^{15}N-separated NOESY-HSQC with a mixing time of 75 ms recorded with a uniformly ^{15}N labeled protein in H_2O, from a 3D constant-time methyl ^{13}C-separated NOESY-HSQC with a mixing time of 100 ms on a uniformly labeled ^{13}C/^{15}N protein in H_2O, and from a 2D NOESY with a mixing time of 70 ms on an unlabeled sample in D_2O.

To determine NOEs across the dimer interface a 3D ^{15}N-separated NOESY-HSQC experiment with a mixing time of 120 ms on a heterodimer sample containing a mixture of ^{2}H/^{15}N-labeled and unlabeled monomers (Ferentz et al., 1997) and a 4D constant-time *J*-resolved ^{13}C-separated NOESY experiment (Melacini, 2000) with a mixing time of 150 ms on a sample containing a mixture of ^{13}C-labeled and unlabeled monomers were recorded. The first experiment yields inter-monomer NOEs between the amide protons of the ^{2}H/^{15}N-labeled and the carbon-bound protons of the unlabeled species. The second allows the separation of inter- and intra-molecular NOEs along the *J*-resolved dimension in which the intra-molecular NOEs between ^{13}C-bound protons appear at $\pm J_{CH}/2$ Hz, while inter-molecular NOEs between ^{13}C- and ^{12}C-bound protons appear at zero-frequency offset because they are not *J*-modulated.

Slowly exchanging amide protons were identified by recording 2D [^{1}H,^{15}N]-HSQC experiments, one day and five days after transferring the protein into a D_2O solution. This information combined with the strong HN-H_α connectivities within and between different β-strands were used to identify the β-strands and the hydrogen bonds between β-strands. The amide protons located in the internal five-stranded parallel β-sheet remain unexchanged five days after the addition of D_2O. The backbone amide protons

involved in hydrogen bonding were also measured using a $^{h3}J_{NCO}$ TROSY-NHCO experiment (Wang et al., 1999).

A generalized version of the [^{15}N-^{1}H]-TROSY experiment (Pervushin et al., 1997; Andersson et al., 1998; Lerche et al., 1999) was used for the measurement of $^{1}J_{HN}$ and $(^{1}J_{HN}+D^{HN})$ couplings on a non-oriented and oriented sample, respectively. The residual dipolar coupling values (RDC,D^{HN}) were calculated from the coupling differences between the couplings $(J_{HN}+D^{HN})$ and $^{1}J_{HN}$ scalar couplings.

All NMR measurements for the Sud-Str complex were carried out on a Bruker DRX500 spectrometer equipped with a 5 mm triple-resonance gradient probe, operating at 293 K. 2D sensitivity-enhanced [^{15}N,^{1}H]-TROSY experiments were recorded with 512 and 256 complex data points in the ^{1}H and ^{15}N dimensions, respectively. The spectral widths were set to 7002 Hz (^{1}H) and 2532 Hz (^{15}N). After quadratic sine apodization and linear prediction up to twice the original size, FIDs were zero-filled and subjected to Fast Fourier Transformation in both dimensions. After stripping the high-field half of the spectrum a data matrix of 512x512 real points was recovered. Protons were referenced to internal DSS, and the ^{15}N dimension was calibrated indirectly with respect to the proton chemical shift. Processing and subsequent analysis of the spectra were performed with the use of NMRLab.

4 Data analysis methods

4.1 Structural data preparation for Sud protein

NOESY cross peaks were picked manually using the AURELIA program. The peak volumes were determined using an automated routine of AURELIA (Geyer et al., 1995) and converted into distance restraints using the symmetry-ADR protocol (Nilges, 1993) which accounts for the ambiguity in the NOEs arising from dispersion and symmetry degeneracy. The experimental unambiguous inter-monomer NOEs and the NH-NH NOE assignments derived from the 4D ^{15}N/^{15}N-separated NOESY experiment were used as 6 Å upper- and 2 Å lower-bounds, respectively. The hydrogen bond restraints were defined as 1.8-2.3 Å for the H-O distance and 2.8-3.3 Å for the N-O distance.

The TALOS program was used to predict the backbone torsion angle intervals from the amino acid sequence and chemical shift information. The tolerance of ϕ and ψ angles was set to $\pm 2 \cdot SD$ (standard deviation) for all the dihedral angle constraints. Only the unambiguous torsion angle predictions (*i.e.* consistent values for the ϕ and ψ angles within the used database of protein structures) were taken into account, covering 66% of the residues for which the chemical shift information was available.

The alignment tensor parameters (axial component and rhombicity) of the oriented sample were determined from the 'powder pattern' distribution of the residual dipolar couplings values (Figure 4.1). The RDC histogram singularities (high, low and the most populated values) determine an overdetermined system of linear equations (see

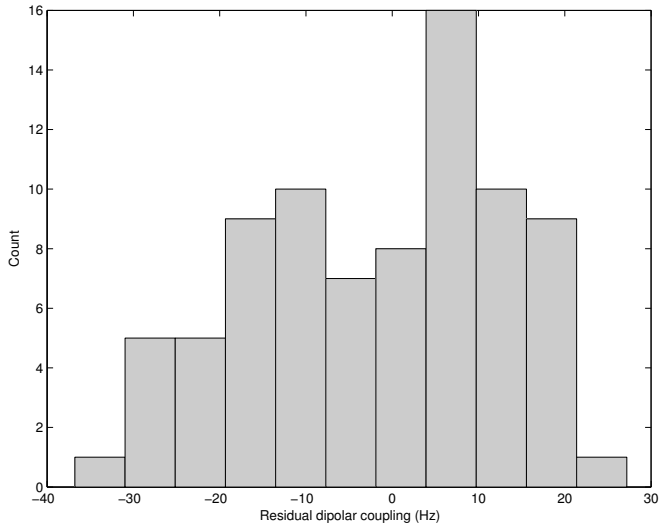

Figure 4.1. Histogram of the (RDC, D^{HN}) values obtained for the Sud protein.

Eqn. 2.10) which yields three different solutions for the alignment tensor components when pairs of equations are solved separately. For a dense RDC data set, the three sets of possible solutions should be consistent with standard deviations smaller than 10-20% of the average values. The further non-linear least squares optimization provides the axial component and rhombicity values for structure calculations.

4.2 Sud protein structure calculation

NMR studies of homodimers are problematic due to difficulty in distinguishing between intra-, inter- and co-monomer (mixed) NOE correlations. To overcome the intrinsic symmetry degeneracy of NOE data the symmetry-ADR protocols of ARIA (O'Donoghue and Nilges, 1999; Nilges and O'Donoghue, 1998) were used for iterative NOE assignment and structure calculations. Symmetry-ADRs describe the ambiguity of NOE peaks arising from both symmetry and dispersion degeneracy by computing a d^{-6} summed distance over all pairs of protons that are possible assignments for a particular cross-peak, including the intra- and inter-monomer contributions. To enforce the two-fold symmetry, the conformational target function contains two special

pseudo-energy terms: a non-crystallographic symmetry (NCS) restraint (Brünger et al., 1998) and a distance symmetry (DSYM) restraint potential (Nilges, 1993). The former serves to minimize the atomic r.m.s. deviation between the two monomers, thus making the two monomers identical, while the latter forces the two monomers into a symmetrical arrangement.

The NOE assignment was performed in nine iterations using the ARIA scheme (Linge et al., 2001), where a generation of structures is used for the NOE analysis (calibration, partial assignment and noise removal) of the following one. In each iteration 50 structures were calculated (60 in the last iteration) and the best 30% of these models were used for the refinement of the next ones. Starting conformers were constructed using hydrogen bonds and manually assigned NOEs from the 4D ^{15}N/^{15}N-separated NOESY and from the NOESY experiments on the asymmetrically labeled dimers, in conjunction with the TALOS prediction for dihedral angles. RDC data was introduced in the third iteration with a low weighting factor. After complete iterative NOE assignments, the weight of the target function RDC term was increased within several turns of structure calculations until the observed and back-calculated values of the residual dipolar couplings agreed within experimental error.

A collection of MATLAB routines were developed (see Section 4.3) to examine the consistency between the NOE assignment table and the 3D ^{15}N- and ^{13}C-edited NOESY spectra by looking at the cross peaks symmetry within the spectra. In the absence of spin diffusion, every HN-HN NOE should appear twice in a ^{15}N-resolved NOESY spectrum. The same type of symmetry should be present for Hα-Hα pairs within ^{13}C- and for HN-Hα pairs between ^{15}N- and ^{13}C-resolved NOESY spectra. All NOE assignments which were not supported by other cross peaks between the corresponding residues and which had no symmetry partners were verified manually.

Structures were calculated using a simulated annealing protocol comprising of four stages: (*I*) a high temperature torsion angle dynamics phase at 10000 K (2200 MD steps), (*II*) a torsion angle cooling stage from 10000 to 2000 K (2200 steps), (*III*)

a Cartesian dynamics cooling stage from 2000 to 1000 K (20000 steps) and (*IV*) a second Cartesian dynamics cooling stage from 1000 to 50 K (18000 steps). All non-stereospecifically assigned prochiral groups, except the manually assigned isopropyl groups, were treated with a floating chirality approach (Folmer et al., 1997).

The polysulfide-binding Sud structure was calculated by assuming a Cys residue containing five polysulfide-sulfur atoms. The polysulfide tail topology was derived from structural information available from the X-ray studies of the polysulfide containing organic complexes (Steudel et al., 1995) which exhibit a helical geometry for the sulfur chain. The following force field parameters were used: S-S bond length of 0.203 Å, C-S-S angle of 103.7998° and S-S-S-S dihedrals with a zero phase shift and a multiplicity of 2 (defines minima at $\pm 90°$ for the dihedral angle pseudo-energy function). Since there is no structural information available for the polysulfide from NMR, the orientation of the sulfur chain has been modeled from steric considerations.

The final structures were calculated with an *ab initio* simulated annealing protocol starting from a random monomer structure with good local geometry. The dimer was generated by a duplication of the monomer unit followed by a 180° rotation around one of its internal axes and a 60 Å translation in the same dimension. Therefore the starting monomer orientation has a two-fold symmetry, which is completely unbiased due to the explicit inter- and intra-monomer NOE assignments performed in an iterative manner. To account for the electrostatic interactions between the side-chains of the ionic residues (44 per monomer not including the His-tag) and to prevent the unrealistic packing that might result from a simple repulsive representation of the non-bonding energy term, the final structures were refined in explicit solvent (water) using a full force field for electrostatic and van der Waals interactions (Linge and Nilges, 1999). The symmetry restraint terms were deliberately left out at this final stage to allow small deviations from the ideal two-fold symmetry. Out of 100 calculated structures the 10 structures with lowest conformational energy were refined in water.

4.3 Consistency check of the NOESY peak lists

The NOE assignment process can be substantially facilitated by checking the consistency of the NOESY peak lists using sequence-specific resonance assignments. The checking procedure is based on the following two considerations: (*I*) a NOESY cross peak is usually part of a network of connections between pairs of spin systems (network anchoring), (*II*) NOESY spectra have an intrinsic symmetry (symmetry mapping). In ^{15}N-edited NOESY spectra, symmetry mapping selects pairs of NH-NH signals, whereas between ^{15}N- and ^{13}C-edited NOESY spectra HN-C$^{\alpha}$H pairs are identified (see Figure 4.2). A similar scheme was originally introduced to discriminate between multiple NOE assignments (Herrmann et al., 2002a) and later used for NOESY cross peak validation (Herrmann et al., 2002b).

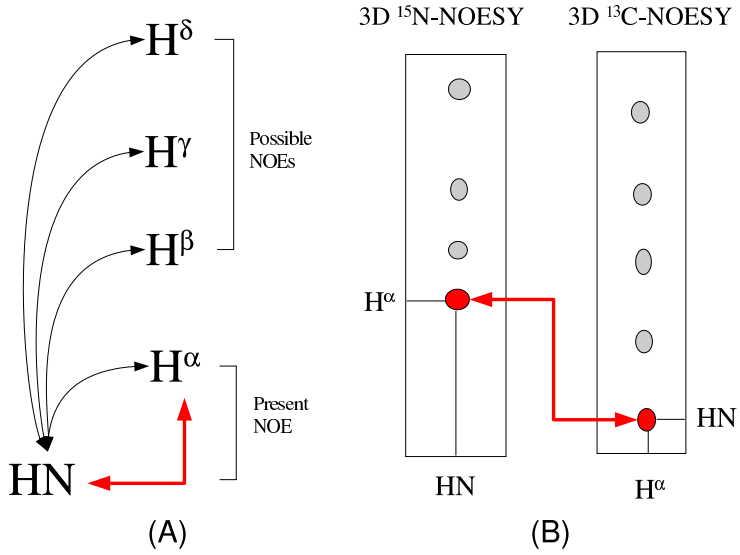

Figure 4.2. Schematic representation of network anchoring (A) and symmetry mapping (B).

The consistency check was implemented within the NMRLab software package in three different functional forms: for NOESY cross peak validation, as a method to select NOE assignments and alternatively, as a method to verify NOE assignment tables obtained by other means. The routine uses ^{15}N- and ^{13}C-edited NOESY peak lists and

the corresponding chemical shift lists (both XEASY and CNS formats are accepted). The individual assignment possibilities of the NOESY peaks allowed by the frequency tolerance are subject to a two-pass filtering which yields a zero-or-one score as follows: (*I*) the network anchoring score is positive if at least a second non-diagonal peak between to the same pair of residues was found, (*II*) the symmetry mapping score is positive if a symmetric partner exists and if this is also anchored in its own network of NOE contacts. The last condition was introduced to minimize the amount of erroneous symmetry partners owing to the residual noise or missing chemical shifts.

The validation filter identifies the 'lonely' NOESY cross peaks which do not belong to any possible network of NOE contacts and do not have any symmetry partner. To discriminate between the different assignment possibilities of a NOESY cross peak the conditions are more restrictive: an assignment is made only if it anchors the peak in a network and if it allows a symmetry related partner. Based on this selection criterion the unambiguous assignments which form a self-consistent network of NOE contacts and possess symmetry related partners are determined. The output consists of two files for each NOESY spectrum. The first file contains a list of all cross peaks which are scored for at least one of the filter criteria and can therefore be considered as reliable. The second file contains a list of the validated NOEs together with the unambiguous assignments obtained using the network/symmetry search.

The same type of filtering is used to verify the NOE assignment tables against the original spectroscopic data. For each entry of the NOE assignment table the network anchoring and the symmetry mapping are examined using the original NOESY peak lists and the assignments of ^{1}H-, ^{13}C- and ^{15}N-nuclei. The software provides an annotated list of NOE assignments which shows the filter criteria which were fulfilled for each entry. All the long-range NOE assignments which are not anchored in a network of contacts or do not have a symmetry related partner should be regarded with caution and manually checked.

The consistency check filters of the NOESY cross peak lists were also embedded

into the ARIA program (Linge et al., 2003). They represent additional non-structural filters which can be utilized during the iterative process of NOE assignment and structure calculations. A supplementary field was added to ARIA interface which allows to select between including the full NOESY data, only the validated NOEs and/or the unambiguous assignments provided by network anchoring and symmetry mapping at any stage of a structure calculation project using ARIA program.

4.4 Wavelet de-noising of the multidimensional NMR spectra

The basis of noise suppression by wavelet shrinkage was described in Section 2.4.3. It is achieved by performing a wavelet transform and applying a threshold to the wavelet coefficients. In the simplest approach, a one-dimensional (1D) wavelet transform can be applied to each 1D strip of the multidimensional NMR spectra. Alternatively, two-dimensional (2D) wavelet transforms can be used to de-noise 2D slices of the NMR spectra. The NMRLab wavelet de-noising routines are based on the WAVELAB8.02 wavelet toolbox[1] of MATLAB (Buckheit and Donoho, 1995).

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

(*I*) For each ^{1}H-^{1}H-slice of the NOESY spectrum a statistical de-noising factor $dfactor = \sigma^{raw}/\sigma^{wav}$ was calculated using the noise standard deviation σ of the baseline regions (see Section 4.5).

(*II*) The effect of the wavelet shrinkage on the fine structure of the NMR signals was quantified by a fine structure score which compares the reference peak volumes (V_{ref})

[1] http://www-stat.stanford.edu/~wavelab/

with the corresponding volumes after wavelet de-noising (V_{wav}):

$$fscore = 1 - \text{mean}\left(\frac{\left|V_{ref} - V_{wav}\right|}{V_{ref}}\right) \quad (8).$$

Peaks were picked using the automated procedures described in Section 4.5 in the case of the wavelet de-noised spectra and manually for the reference peak list. Peak volumes were obtained by the integration algorithm which is described in the next section.

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

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

With the exception of the de-noising factor which is always larger than one, these scores have values between zero and one where a value of one represents the ideal case of a fully de-noising peak list without any distortion of the real signals. A negative *dscore* indicates truncation artefacts (causing additional local extrema) introduced by the wavelet transform.

4.5 Automated peak picking and peak integration of the multidimensional NMR spectra

A robust numerical procedure for automated peak picking and peak integration of the multidimensional NMR spectra was developed and integrated into NMRLab. The peak picking procedure consists of four distinct steps which will be described in detail for a paradigmatic 2D data set.

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

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

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

(*II*) In a second step, the spectrum was segmented into regions of points with the absolute value of the intensity larger than the local noise levels $bnoise(P)$ (Figure 4.3, blue crosses). Because the standard deviation of the signal after de-noising is not a suitable descriptor for local noise levels, the $bnoise(P)$ values obtained for the raw spectra were also used for the segmentation of the de-noised spectra.

(*III*) The local extrema (maxima or minima, depending on the peak signs) were determined by a grid search using the sparse matrix obtained after segmentation. In the present implementation the width of the grid cell can be adjusted by the user according to the digital resolution of the data set; in this work the smallest possible grid cell size of 3×3 points was used. A peak list containing the coordinates of all the local extrema

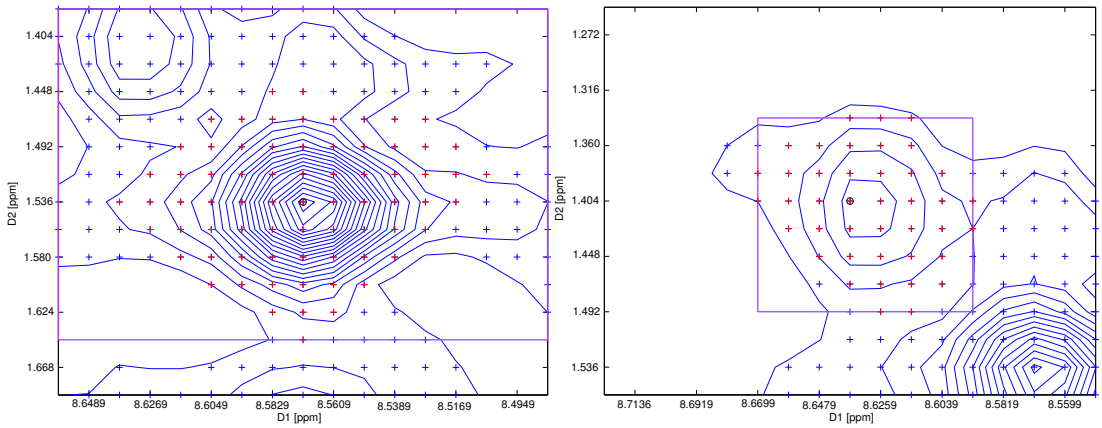

Figure 4.3. Two examples for peak integration in the presence of the spectral overlap. The plots represent the initial integration boxes, blue crosses depicts the digits with an intensity larger than the estimated local noise levels, magenta squares are the refined integration boxes and the red crosses the actual points which are found to be part of the peak subject to integration.

above the local noise levels is obtained.

(*IV*) An algorithm for *digital peak integration* which can separate overlapping signals (even if those have very different intensities and widths) was designed. This algorithm first defines an initial integration box around each local maximum[2] (Figure 4.3: full boxes). Its rational size is determined by a rectangular local minimum search starting from the central maximum along the Cartesian dimensions of the spectrum which stops either if the background noise level is reached or if a local minimum is encountered (Figure 4.3: magenta boxes). Within the refined rectangular integration box the peak shape is resolved by an object-related growing algorithm around the local maximum (Figure 4.3: red crosses) which iteratively adds one square shell centered on the central maximum (Figure 4.4: continuous line) until the end of the integration box is reached in each dimension. A point of the new shell (Figure 4.4: point 7) is added to the peak if the first order neighbor (Figure 4.4: point 2) has a higher intensity in the previous layer and if the second order neighbors have intensities above the local noise levels (points 1 and 3 in Figure 4.4). For corners (Figure 4.4: point 5) the condition is

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

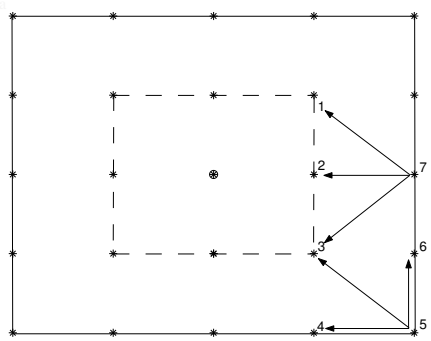

Figure 4.4. Object-related growing algorithm used for the peak integration: the dashed and solid lines represent the first and second shells which define the peak limits around the central local maximum, respectively. Point 7 is considered to be a part of the peak if the first order neighbor (point 2) has a higher intensity and if the two second order neighbors (points 1 and 2) have intensities above the noise threshold. For corners a variation of this definition is used: the first order neighbor of point 5 is point 3 and the second order neighbors are points 6 and 4.

slightly different, the first order neighbor is considered to be the edge of the previous shell (Figure 4.4: point 3) and the second order neighbors are located within the same layer (points 4 and 6 in Figure 4.4).

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

4.6 NMR chemical shift mapping

NMR is a powerful tool for assessing protein-ligand and protein-protein interactions. Information about these interactions can be obtained on various levels by using different experimental techniques (Zuiderweg, 2002; Clarkson and Campbell, 2003). Chemical shift perturbation mapping is the most widely used method to identify protein-substrate interfaces. The method is based on monitoring chemical shifts in the

[^{15}N,^{1}H]-HSQC(TROSY) spectrum of a protein when the unlabeled partner is added. The key parameter is the chemical shifts of the backbone amides which are very sensitive for local geometric and electrostatic changes. The interaction causes changes in the magnetic environment of the N and HN atoms on the protein-substrate interface and, hence, affects the chemical shifts of the nuclei in this area. Both surface and non-surface residues can be affected by secondary effects to regions under the protein surface. In some cases when a large part of the protein changes conformation and many chemical shifts are affected, chemical shift perturbations may not provide the information of the binding interface.

In this thesis chemical shift mapping has been used to examine the interaction between the Sud protein and its functional substrate, the polysulfide-sulfur ligand, as well as between Sud and Str proteins. [^{15}N,^{1}H]-TROSY spectra of the Sud protein were compared in the absence and in the presence of the polysulfide-sulfur and Str protein. Changes of backbone amide proton (^{1}HN) and nitrogen (^{15}N) resonances of Sud protein were determined and weighed according to the formula: $\sqrt{(^{1}HN)^2 + (^{15}N/6.5)^2}$ (Mulder et al., 1999). The weighted chemical shift changes were normalized to a maximum of 100% for each data set and the individual values were color-coded and mapped on the Sud structure plots.

4.7 Multivariate analysis of the NMR screening data

NMR spectroscopy is now commonly used in screening of pharmaceutical libraries for protein inhibitors. When series of 2D [^{15}N,^{1}H]-HSQC(TROSY) spectra of ^{15}N-labeled proteins are used to detect ligand binding, the high sensitivity of the ^{1}HN and ^{15}N chemical shifts of the protein backbone for small geometric or electrostatic changes induced by ligand binding is exploited. To analyze large numbers of spectra for changes and similarities efficient pattern recognition methods such as principal component analysis (PCA) are frequently used. Principal components are linear com-

binations of the original data which help to visualize similarities in an ensemble of spectra. Since all principal components are orthogonal and ordered with respect to maximum variance between the samples, the largest two or three principal components provide an excellent representation of variability within a set of data (see Section 2.4.5). The first two or three principal components of the spectra with little variance as compared to the reference spectrum cluster around the reference, while outliers of the main cluster represent hits in ligand screening.

PCA is computationally expensive even if only few principal components are calculated. Therefore, several data manipulations are usually applied prior to PCA, both to reduce the data size and to minimize artifacts. Simple thresholding helps to eliminate noise related alterations between the spectra. In addition, in a procedure called 'bucketing' (Ross et al., 2000), adjacent data points (in the case of two-dimensional spectra a rectangular subsection of the spectrum) are added to one 'bucket' thus reducing the amount of data points. This procedure is broadly used in common NMR screening software. Bucketing helps to eliminate artifacts by averaging small chemical shift perturbations arising from small variations in pH or other sample conditions and reduces the size of the data. The 'bucket' descriptors which are subsequently used in PCA maintain much of the information of the spectrum although details which are only available at high resolution may be lost. Bucketing is also prone to introduce artifacts when peaks experience small chemical shift perturbations at the border between buckets. In this case a large change may be detected for a small effect. Another artifact may arise from cancellations in the bucket when different points which contribute to one bucket add and subtract equal or similar intensities. In this case no or only a small overall effect is left in the bucket.

With higher resolution offered by increasing field strengths of NMR spectrometers PCA should preserve the full information available in the spectra. The present study shows how wavelet de-noising and multiresolution analysis can be efficiently combined with PCA to analyze large series of NMR data. To demonstrate the advantage

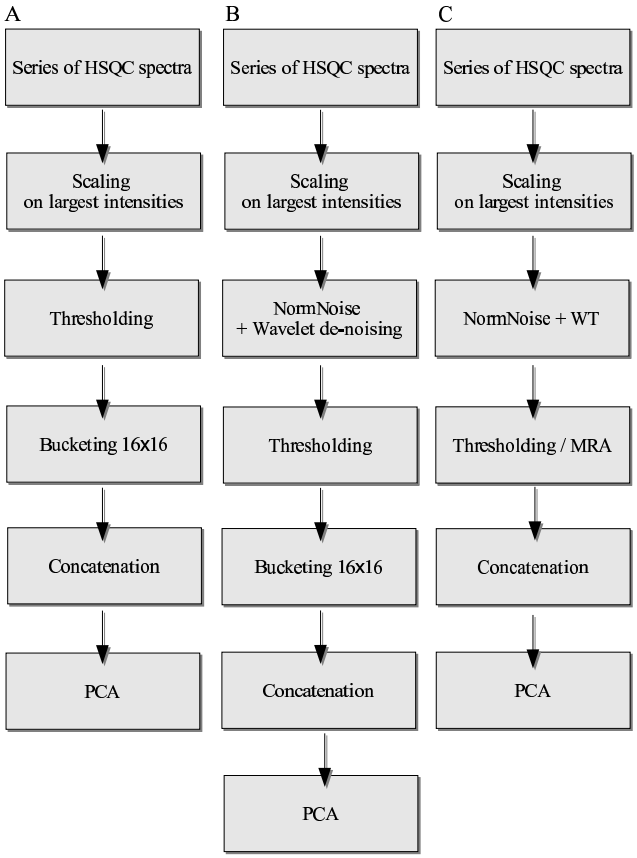

Figure 4.5. Schematic representation of different PCA data reduction schemes: (A) bucketing scheme, (B) bucketing on wavelet de-noised data, (C) PCA on wavelet coefficients.

of a wavelet filter as compared to the plain bucketing approach three different pre-processing schemes prior to PCA analysis were compared (Figure 4.5). The test data set was comprising of 101 [^{15}N,^{1}H]-HSQC spectra of hsp90 protein recorded in the presence of different ligands on a Bruker DMX600 spectometer. 1024 complex data points were recorded in the direct dimension and 128 increments were recorded for each spectrum. All spectra were processed in NMRLab with a quadratic sine apodization prior to the Fast Fourier Transform in both dimensions and two-dimensional automated phase correction. After stripping the high-field half of the spectrum and removing some lines of the spectrum without any signals a data matrix of $512 \times 512 \times 101$ real points was recovered for subsequent multivariate analysis.

The 101 HSQC spectra were scaled using the mean of the largest peaks with mini-

mum variability within the data set. Subsequently, different protocols were employed for the three analysis schemes shown in Figure 4.5. For scheme A a threshold value of 20% of the largest point in each spectrum was applied prior to adding data points in 16×16 point bucket cells. Subsequently, common baseline regions of all spectra with zero intensity after thresholding were removed and the two-dimensional HSQC spectra were concatenated into one-dimensional objects prior to PCA analysis.

In scheme B a wavelet de-nosing step was added before thresholding and bucketing using a one-dimensional discrete wavelet transform in both dimensions. The overall process of wavelet de-noising consists of four stages: (*I*) *data scaling* with respect to the average noise level estimated by the median absolute deviation of the wavelet coefficients on the first dyadic level (σ, see Section 2.4.3), (*II*) *a discrete wavelet transform* using a Symmlet 8 quadrature mirror filter and a low-frequency cutoff of 4, (*III*) *a global soft-thresholding* of the wavelet coefficients applying the universal threshold $\lambda = \sqrt{2 \cdot \log N}$ (where N is total number of data points) and (*IV*) *an inverse discrete wavelet transform* which returns a matrix with the same size as the input data matrix. The subsequent steps (thresholding, bucketing, concatenation, removal of common zeroes and PCA) were identical with scheme A.

In scheme C the thresholding and bucketing steps of schemes A and B were circumvented by using the sparse representation of soft-thresholded wavelet coefficients for subsequent concatenation, removal of common zeroes and PCA. The scaled data was subject to a wavelet transform and soft-thresholding with identical parameters as in scheme B. The wavelet coefficient thresholding has a triple effect: it reduces the stochastic component of the spectra (de-nosing), minimizes insignificant spectral perturbations (smoothing) and decreases the size of the data matrix (compression). This approach is more efficient than the 'blind' bucketing procedure in the sense that wavelet decomposition is a clever bucketing, precisely adapted to the nature of the analyzed NMR data. Additional multiresolution analysis (MRA) was applied by suppressing the four low-frequency dyadic levels for a selective filtering of the baseline-related

artifacts.

Automated analysis of the PCA clusters was used to evaluate the PCA result. The clustering algorithm was based on the hierarchical clustering analysis (Johnson, 1967), where objects are linked together based on the network of Euclidean distances between pairs of objects. In a first step binary clusters of objects in close proximity were formed. As objects were paired into binary clusters, the newly formed clusters were grouped into larger clusters until a hierarchical tree was formed. Finally, the hierarchical tree was divided into clusters of objects by detecting the natural groupings in the cluster tree. To allow the formation of a main cluster surrounded by several outliers, a tight clustering threshold was used to cut the hierarchical tree. The principal components were scaled with respects to the largest distance in the data set and a clustering factor was estimated as the mean scaled distance between the PCA outliers and the closest neighbor belonging to the main cluster. Compression factors were calculated as the ratio between the number of elements of the original data matrix and the number of elements of the pre-processed matrix used for the subsequent PCA. The de-noising factor describes the relative number of points eliminated by thresholding. In schemes A/B it is calculated as the quotient of the number of elements of the data matrix zeroed by thresholding and the total number of elements of the original matrix. For scheme C the quotient of the number of wavelet coefficients eliminated by thresholding and the total number of elements of the original matrix was used.

5 Results and Discussion

5.1 Sud protein

5.1.1 Solution structure of Sud protein

Using the backbone chemical shifts of Sud protein (Lin et al., 2000) as a starting point, the majority of side chain ^{1}H and ^{13}C resonances were assigned by a combination of 3D CC(CO)NH, CC(CA)NH, H(C)CH-COSY and H(C)CH-TOCSY experiments. Approximatively 74% of the resonances were assigned. Stereospecific assignments of nearly all isopropyl groups of Val and Leu residues were obtained, with a single exception: Leu73.

A total number of 8 (16 considering the symmetric related ones) inter-monomer NOEs were unambiguously assigned using asymmetric labeling experiments. Due to the low concentration of the asymmetrically labeled dimers in the NMR sample only few inter-monomer NOEs could be determined. Based on these experiments, the contact regions between the two monomers were found to involve mainly the residues F7, D8, T10, F11 of one monomer and A75, Y105 of the other monomer. From 9532 experimental NOESY peaks (2D homonuclear and 3D ^{15}N- and ^{13}C-edited NOESYs) 1095 ambiguous and 3758 unambiguous non-redundant NOE derived distance restraints were obtained, including 86 inter-monomer and 142 co-monomer NOEs. This structural data together with 340 TALOS-derived ϕ and ψ angle constraints, 402 backbone-backbone and backbone-side chain NH-NH distance restraints

Table 5.1. Structural statistics for the 10 lowest energy simulated annealed and water refined structures of Sud homodimer.

Unambiguous NOE distance restraints	4182
Sequential ($\|i-j\|=1$)	1115
Medium range ($1<\|i-j\|\leq 4$)	605
Long range ($\|i-j\|>4$)	830
Inter-monomer	102
Ambiguos NOE distance restraints	1095
Co-monomer	142
H-bond restraints	22
Dihedral angle restraints	340
RDC restraints	162
[a]R.m.s. deviation from distance restraints (Å)	0.014±0.001
[a]R.m.s. deviation from angle restraints (deg.)	0.509±0.016
[a]R.m.s. deviation from RDC restraints (Hz)	1.043±0.039
[a]R.m.s. deviation from covalent geometry	
Bond lengths (Å)	0.0034±0.0001
Angles (deg.)	0.507±0.0154
Impropers (deg.)	1.335±0.057
[b]Ramachandran plot	
Most allowed regions	86.3 %
Disallowed regions	0.6 %
[c]RMSD of the NMR ensemble (Å)	
Monomer core backbone	0.81±0.18
Dimer backbone	0.96±0.20

[a]Evaluated by CNS/ARIA.
[b]Calculated with PROCHECK-NMR (Laskowski et al., 1996).
[c]Mean global backbone RMSD calculated with MOLMOL(Koradi et al., 1996). The flexible loop between residues 89-94 was not used for RMSD calculations.

derived from the 4D ^{15}N/^{15}N-separated NOESY, 22 hydrogen bonds between the β-strands, 16 experimental inter-monomer NOEs (asymmetric labeling) and 162 amide residual dipolar couplings were used for the final structure calculations (see Table 5.1 for a detailed structural statistics).

The consistency check of the NOE assignment table (see Section 4.3) obtained with ARIA for the Sud protein revealed 10% of the 4639 distance restraints originating from 3D heteronuclear NOESY experiments which did not fulfill any of the filter criteria. Further inspection showed that none of these 'questionable' NOE restraints were in-

volving long range contacts (between more than five residues in the primary sequence), about half of them were clearly wrong assignments and the rest could not be interpreted unambiguously. The short-range restraints define the local geometry of the secondary structure elements. Although they may not be essential for the global fold definition, wrong entries always cause structural inconsistencies and distortions in the local geometry reflected by elevated values of the target function. The removal of these 10% 'questionable' short-range restraints led to a improved structure calculation. While the RMSD between the 10 best structures was unchanged within the error limits the total conformational energy dropped by 14%. The detailed results of this analysis are summarized in Table 5.3.

Table 5.3. Consistency check of the NOE assignment data of the Sud protein.

NOE assignments (3D NOESY)	Symmetry mapped	Spin system anchored	Symmetry mapped or spin system anchored
4639	1365	3862	4145
		Target function [kcal/mol]	Mean bb RMSD [Å]
	All NOEs	386±5	0.95±0.21
	Symm. OR Anch.	334±7	0.96±0.20

Figure 5.1 shows the backbone superposition of the energetically best 10 models of the Sud structure calculated with and without residual dipolar coupling restraints. The relative orientation of the two monomers was significantly improved by the residual dipolar coupling restraints, the RDC-refined dimer structure showing a more compact form compared to the RDC-free one. The mean backbone RMSD of the NMR ensembles drops from 1.64 to 0.96 Å for the RDC-refined conformers considering the whole dimer, although the monomer cores (without the α-helical N-terminus) were equally well defined having a value of about 0.8 Å in both cases. For the RMSD calculations the segment between residues 89 and 94, which is poorly defined due to the lack of experimental data, was not considered. Comparing the monomer structures in the two

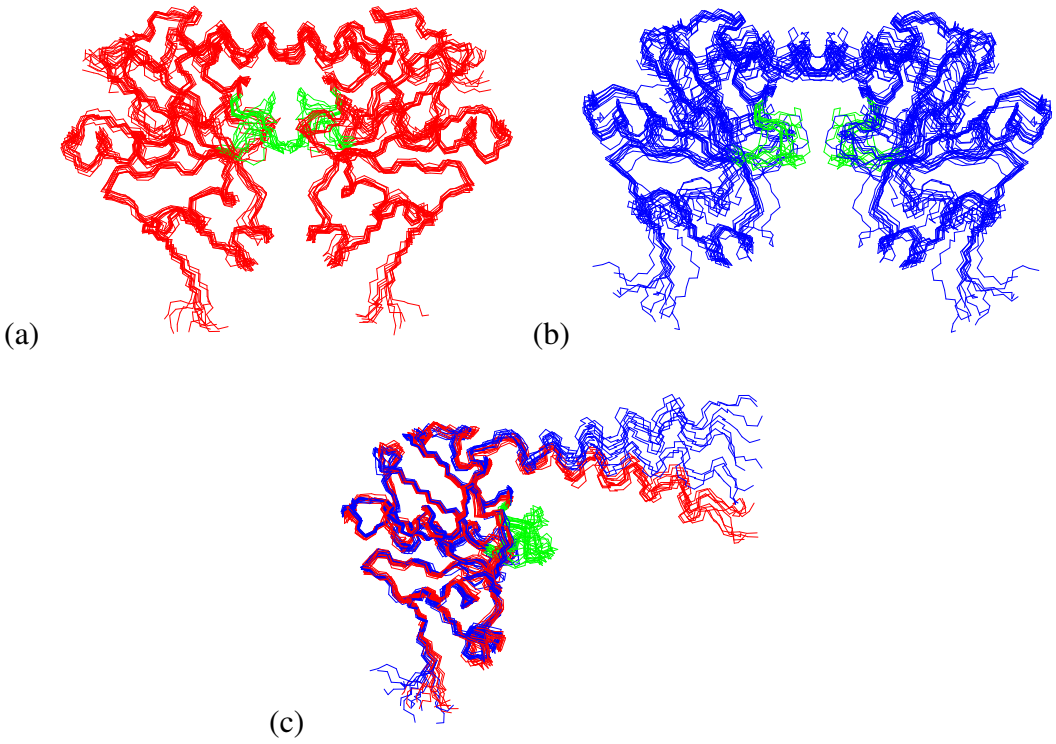

Figure 5.1. (a) Backbone plot of the best 10 RDC-refined structures of the Sud dimer. (b) Backbone plot of the RDC-free NMR ensemble (10 models) of Sud. (c) Backbone representation of the monomer cores (residues: 25-130) superposition for the RDC-refined (red) and RDC-free (blue) structures of the Sud protein. The flexible loop (residues: 89-94) is colored in green.

ensembles, the core regions are nearly superimposable, the main difference resulting from the relative orientation of the monomer units dictated by different positioning of the N-terminal α-helix. The secondary structure elements are also similar. The result illustrates how the residual dipolar coupling restraints improve the definition of the relative orientation of the monomer units within the homodimer.

Figure 5.2 shows the ribbon representations of the solution structure of the Sud homodimer. The monomer unit has an α/β topology with six α-helices (helix I to VI) packed against a central core of five parallel β-strands (βA, βB, βD, βE and βF) and a lateral two-stranded antiparallel β-sheet (βC and βG) which may be involved in the structural stabilization of the C-terminal segment. For the minimized average structure, the β-strands were formed by the residues 23-25 (βA), 41-44 (βB), 56-57

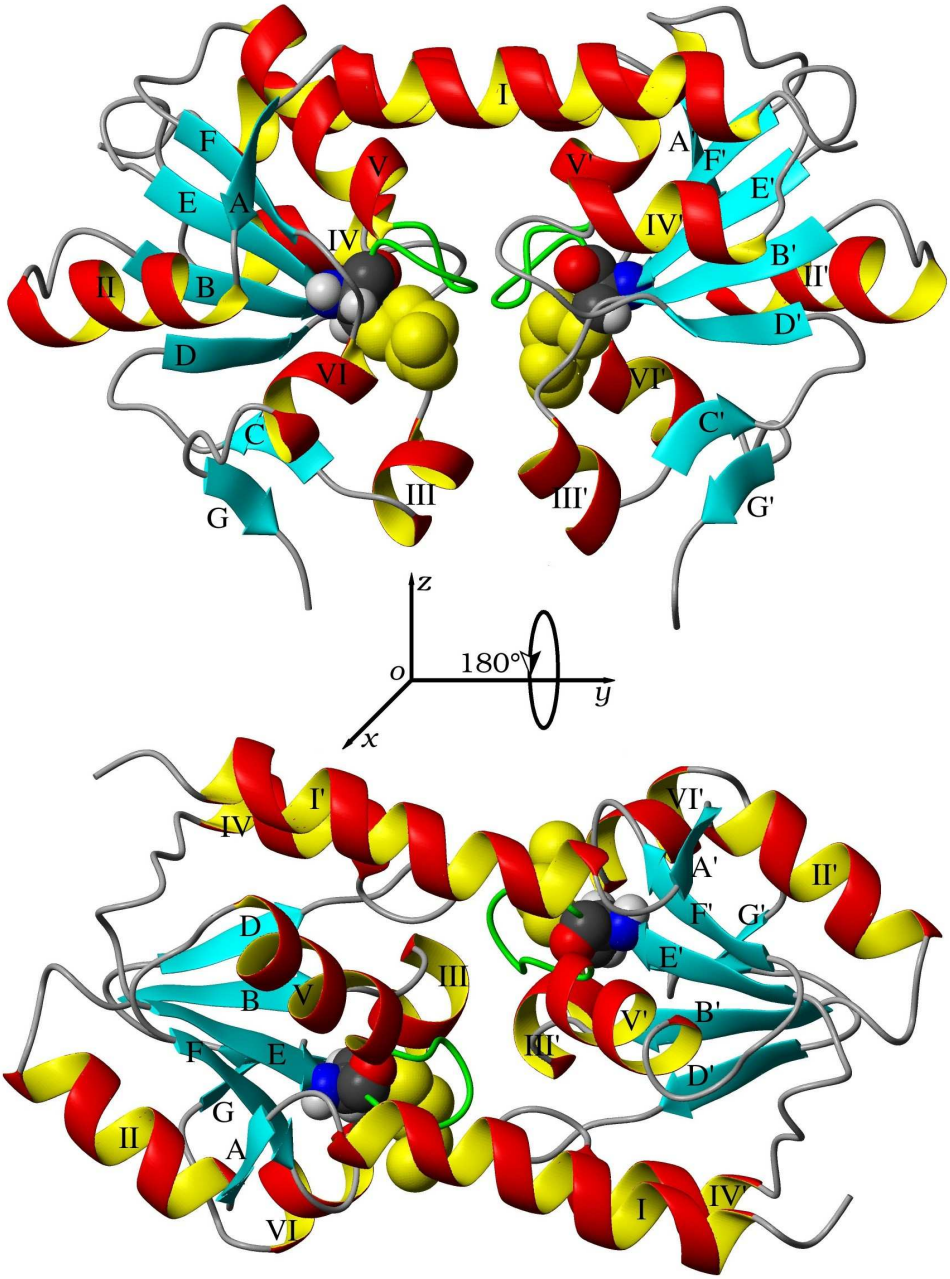

Figure 5.2. Ribbon representations of the Sud dimer. The catalytic cysteines (with a 5-atoms long polysulfide chain attached) are depicted using a CPK model. The two-fold symmetry axes are the OZ and OX axes for the top and bottom structures, respectively. The active-site loop (residues 89-94) is colored in green.

(βC), 62-64 (βD), 85-88 (βE), 110-113 (βF), and 127-128 (βG) within each subunit, while the helices are observed for the sequences 4-20 (I), 27-36 (II), 48-53 (III), 71-77 (IV), 95-103 (V) and 117-122 (VI). The polysulfide chains bound to the cysteine residues are pointing in opposite directions, to the outside of the protein. The distance between the two Sδ atoms of cysteine residues is ranging between 16.7 and 18.9 Å. The N-terminal helices (I and I') of the two monomers are oriented parallel to each other, but rather distant in space, and therefore an interaction between them is unlikely. Helix I of one monomer unit interacts with helices IV' and V' of the second unit to form a three-helix bundle stabilizing the dimeric structure.

Based on the inter-monomer NOEs assigned with ARIA, the residues participating in the interaction between the two monomers are mainly F7, F11 and V15, located at the hydrophobic side of helix I with partners L74', A75', L79', L97' and Y105' in the opposite unit. Hence, mainly hydrophobic interactions are involved in stabilizing the dimeric structure. For example, the aromatic side-chains of F7 and F11 interact with Y105' and the methyl groups of L74', while F11 and V15 interact with L97'. Due to the interactions with aromatic rings, the Hδ resonances of L74 and L79 are shifted to higher field by values between 0.1 and 0.2 ppm. A salt-bridge in this region between K12 (helix I) and E71' (helix IV') may form, since the side chains approach each other by less than 5 Å.

Sud serves as a polysulfide-sulfur binding and transfer protein, transferring the aqueous polysulfide to the active site of polysulfide reductase, which is exposed to the periplasmic side of the cytoplasmic membrane of *W. succinogenes*. The two identical subunits of Sud, each with a single cysteine residue covalently bind two polysulfide chains with up to 10 sulfur atoms. Using site-directed mutagenesis it was shown that the cysteine residues are essential for the sulfur binding and transferase activity of Sud (Klimmek et al., 1999).

The active-site environment of Sud resembles that of rhodanese of *A. vinelandii* (RhdA), in both cases the catalytic cysteine residues being located at the bottom of

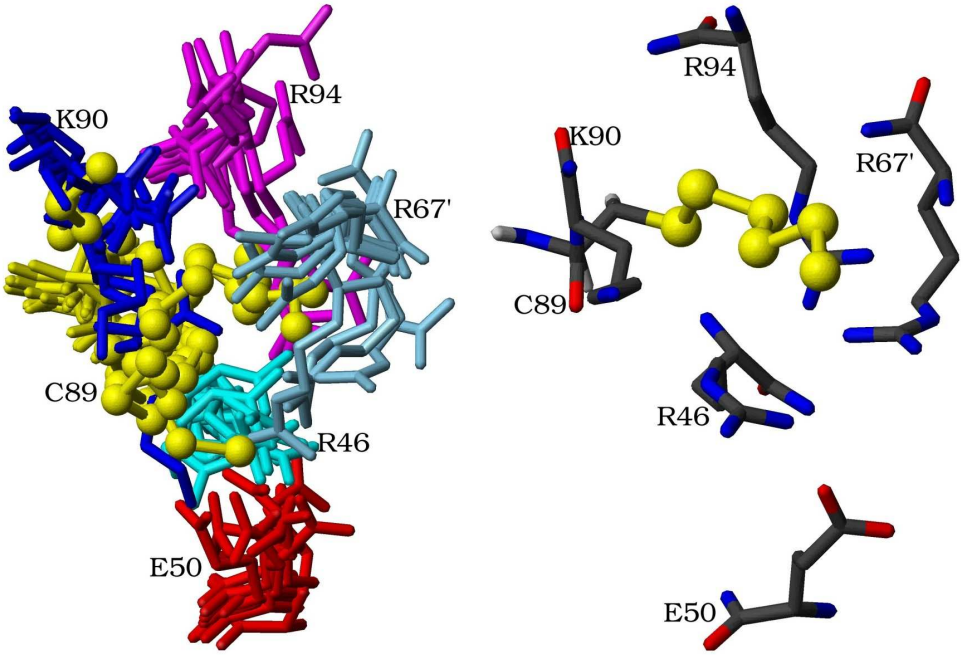

Figure 5.3. The positively charged polysulfide binding pocket. Due to the lack of resonance assignments the side chains conformation is poorly defined. The left side of the plot represents a superposition of the NMR ensemble (10 models) of the Sud dimer structure, while the right side represents the minimized average conformer of the ensemble.

shallow round pockets close to the inter-domain boundary, at the beginning of a loop with a cradle-like conformation that is connecting a central β-strand with an interface α-helix. RhdA is a covalently bound multidomain protein consisting of two similar but not identical α/β domains, respectively the N-terminal and the C-terminal domain. Unlike Sud which has a polysulfide binding site in each monomer unit, RhdA has a single functional cysteine residue located in the C-terminal domain. Side chains in the intermonomer contact area around the active site of Sud indicate that the dimer formation is required for the protein function. The active-site cysteine is the first residue of the 89-94 loop, connecting the βE strand to helix V (Figure 5.2). Due to the lack of chemical shift assignments (both backbone and side chain resonances were unassignable, excepting HN-N, Hα-Cα of C89 and Hα-Cα of T91) this segment is poorly defined, which may be a result of multiple conformations induced by the polysulfide mobility. In a similar way resonances of R46 and R67, both located near the sulfur tail, were not

assigned presumably to such a conformational heterogeneity and mobility.

The Sud structure reveals a positively charged binding pocket for the negative polysulfide-sulfur chain formed by the residues R46, R67' (adjacent monomer unit) K90 and R94 (Figure 5.3). An electrostatic binding pocket that partially covers the Sud-polysulfide tail is consistent with previous MALDI mass spectroscopy investigations which indicated a much lower dissociation constant for the first two sulfur atoms of the polysulfide chain (Klimmek et al., 1999). The positively charged side chains of R46, R67', K90 and R94 of Sud interact with and stabilize the first two S-S bonds of the negatively charged polysulfide, while the negatively charged side chain of E50 interacts with R46 (Figure 5.3). The mutation of any of the above-mentioned residues leads to a loss of the sulfur-transferase activity (data not shown). The amino acid sequence alignment of RhdA and Sud indicates that R46, E50, R67, C89 and R94 (Sud numbering is used) are conserved residues (Figure 1.1).

The active-site loop surrounding the catalytic cysteine (which is preserved in all rhodanese enzymes) appears to be flexible for the Sud protein as evidenced by the missing chemical shift assignments of the related residues (residues 89-94). The polysulfide tail extends out of a positively charged binding pocket (residues R46, R67', K90 and R94), where Sud may contact the polysulfide reductase.

The coordinates of the Sud protein together with the NMR structural data have been deposited in RCSB Protein Data Bank under the PDB ID code 1QXN.

5.1.2 Chemical shift mapping of the polysulfide binding

The Sud dimer is a polysulfide-sulfur binding and transferase protein. Each momomer unit contains a catalytic cysteine residue which covalentely binds the polysulfide substrate. Dialysis and MALDI experiments indicate that Sud binds two sulfur atoms with a low dissociation constant and seven more sulfur atoms with a higher dissociation constant, in addition to the sulfur atom linked to cysteine (Klimmek et al., 1999). Despite of the different apparent dissociation constants, it is likely that all the bound sulfur atoms form a common chain that is covalently linked to the cysteine residue. The structural work carried out using the substrate bound form of the Sud protein shows a positively charged binding pocket that covers partially the polysulfide ligand, which explains the uneven distribution of the dissociation constant values. The S-S bonds of the sulfur chain are stabilized by the covalent linkage to cysteine and by the environment of the binding pocket, both having a decreasing power of influence as the distance from the cysteine residue increase. The substrate binding was further investigated by a comparison between the [^{15}N,^{1}H]-TROSY spectra of the Sud protein in the presence and in the absence of the polysulfide ligand. With this approach, every amide group which shows a chemical shift perturbation yields structural information about the region affected by the polysulfide binding. In addition, the comparison provides the backbone amide chemical shift assignments of the polysulfide free form of the Sud protein. An excess of polysulfide was added to ensure that the protein was fully loaded with sulfur.

The comparison between the spectra of free and polysulfide loaded forms of the Sud protein revealed a large spectral variation (Figure 5.4, red/blue spectra). Besides chemical shift perturbations some additional peaks were found for the substrate free form of the Sud protein which may be a result of the slow-intermediate exchange induced by the polysulfide binding (see Figure 5.4, marked peaks). This view is supported by the previous results where resonances belonging to the active-site loop surrounding the

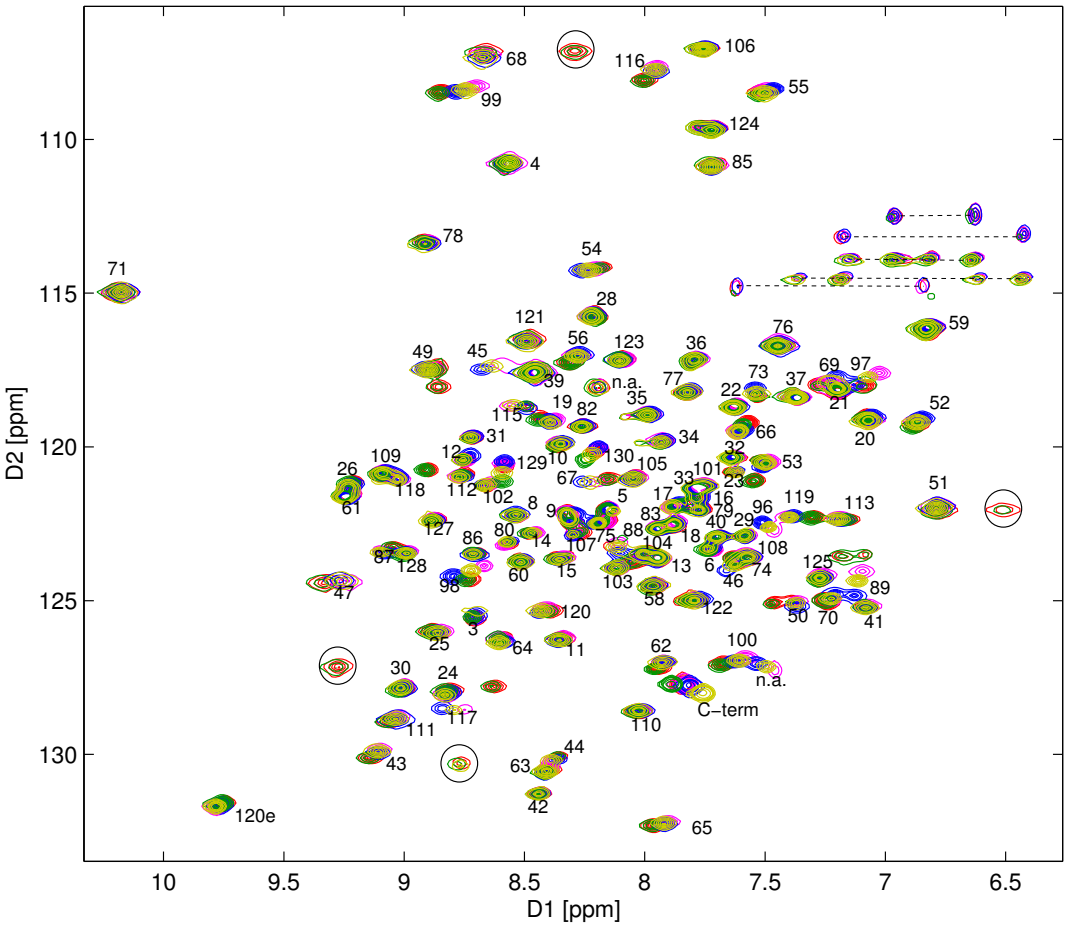

Figure 5.4. [^{15}N, ^{1}H]-TROSY spectra of the Sud protein. Blue depicts the spectrum obtained for the polysulfide bound form of the Sud protein, red the polysulfide free form of the protein, green the Sud-Str complex in the absence of the polysulfide substrate, magenta the Sud-Str complex where only the Str protein was loaded with polysulfide before complex formation and yellow the Sud-Str complex where both proteins were fully loaded with polysulfide. The small circles mark peaks which are disappearing upon polysulfide binding. The backbone assignments were indicated by the corresponding residue numbers.

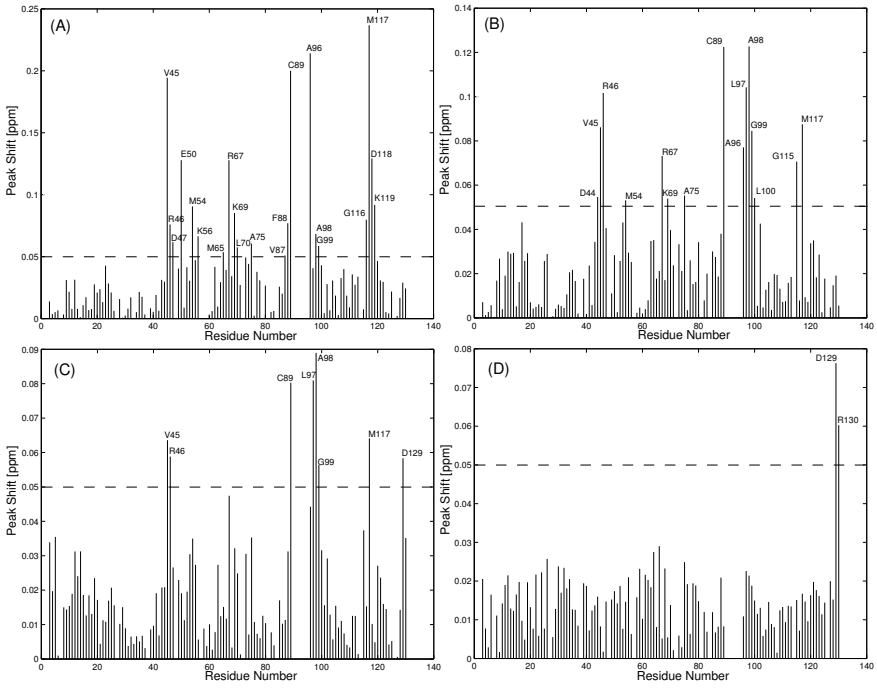

Figure 5.5. Overview of the weighted chemical shift changes for the backbone amide ^{1}HN and ^{15}N resonances of the Sud protein. Panel (A) displays the chemical shift changes upon the polysulfide binding, panel (B) the changes caused by the Sud-Str interaction when only the Str protein was polysulfide loaded before complex formation, panel (C) the changes induced by the Sud-Str interaction for a complex were both partners were fully loaded with their functional substrate and panel (D) the spectral variation for the Sud-Str complex in the absence of the polysulfide-sulfur.

catalytic cysteine (residues 89-94) could not be observed, presumably due to the multiple conformations induced by the polysulfide-sulfur chain mobility to the neighboring residues.

A large number of chemical shifts showed significant perturbations: 21 backbone amides out of 112 were shifted with more than 0.05 ppm (Figure 5.5, panel A). The residues with the largest chemical shift changes (>0.1 ppm) were: V45, E50, R67', C89, A96, M117 and D118, all in the near vicinity or part of the polysulfide sulfur binding pocket (ball-and-sticks in Figure 5.6). Figure 5.6 also depicts the color coded distribution of the chemical shift perturbations (from gray to red) plotted on the ribbon representation on the Sud protein structure. Substrate binding affects chemical shifts in a large region surrounding the active site of the protein. Because the polysulfide

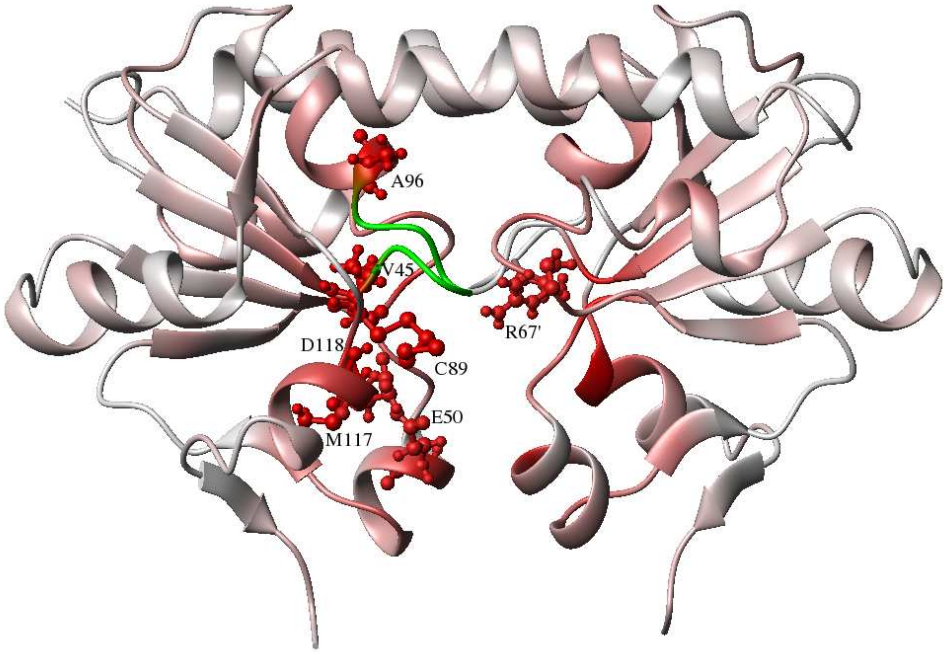

Figure 5.6. The ribbon representation of the Sud protein structure color coded (from grey to red) according to the weighted backbone amide chemical shift changes induced by the polysulfide binding as depicted in the panel (A) of Figure 5.5. The most affected residues are drawn as ball-and-sticks and the active-side loop is shown in green. The prime symbol (*e.g.* R67') indicates a residues belonging to the second monomer unit.

binding site is situated at the bottom of a shallow pocket formed in the dimer interface, polysulfide binding affects both non-surface residues and residues belonging to the opposite monomer unit. Furthermore, the large number of residues belonging to the dimer interface which are affected by the substrate binding suggests a change of the native dimer fold.

The chemical shift changes induced by the polysulfide ligand may indicate certain conformational rearrangements for the Sud protein upon substrate binding, where both the local geometry of the polysulfide binding site and the dimer interface are affected. The conformational changes and the internal dynamics induced by the polysulfide binding could be the trigger of the subsequent polysulfide-sulfur transfer.

5.1.3 Chemical shift mapping of the Sud-Str interaction

The Str protein is a second polysulfide-sulfur binding and transferase protein found in the bacterial periplasm of the *W. succinogenes*. It is a 40 kDa protein and its primary sequence contains seven cysteine residues. The amino acid sequence alignment with other sulfur transferase enzymes (rhodanese-like proteins) suggests a covalently bound two domains protein, with one catalytic cysteine for polysulfide binding. The native concentration of the Str protein in the bacterial periplasm is approximatively five times higher than of the Sud protein. Therefore, the two proteins are thought to form a polysulfide harvesting complex in which Str collects and delivers the aqueous polysulfide to Sud, which in turn mediates the sulfur transfer to the catalytic molybdenum ion located at the periplasmic active site of the membrane protein polysulfide reductase. The transferase interaction between the two proteins was assessed by observing the chemical shift perturbation induced in the [^{15}N,^{1}H]-TROSY spectra of Sud protein by the complex encounter and polysulfide-sulfur transfer.

[^{15}N,^{1}H]-TROSY spectra of Sud in complex with the Str protein (1:1) were recorded in the presence and in the absence of the polysulfide ligand (Section 3.2: samples *I*, *II* and *III*). The Sud-Str complex in the absence of the polysulfide-sulfur produced a spectrum virtually identical with the one of the substrate free form of the Sud protein (Figure 5.4, green/red spectra). Only two residues located in the C-terminal part (D129 and R130) showed significant chemical shift perturbations, most likely caused by transient interactions of the His-tag attached to Sud rather than by a direct protein-protein interaction (Figure 5.5, panel D).

For the second sample of the Sud-Str complex (ligand bound Str and ligand free Sud) the [^{15}N,^{1}H]-TROSY spectrum was significantly different compared to the spectrum originating from the polysulfide free form of the Sud protein and similar to the spectrum of the polysulfide loaded Sud protein (Figure 5.4, spectra magenta/blue). This result shows that the polysulfide-sulfur is transferred between the interaction partners

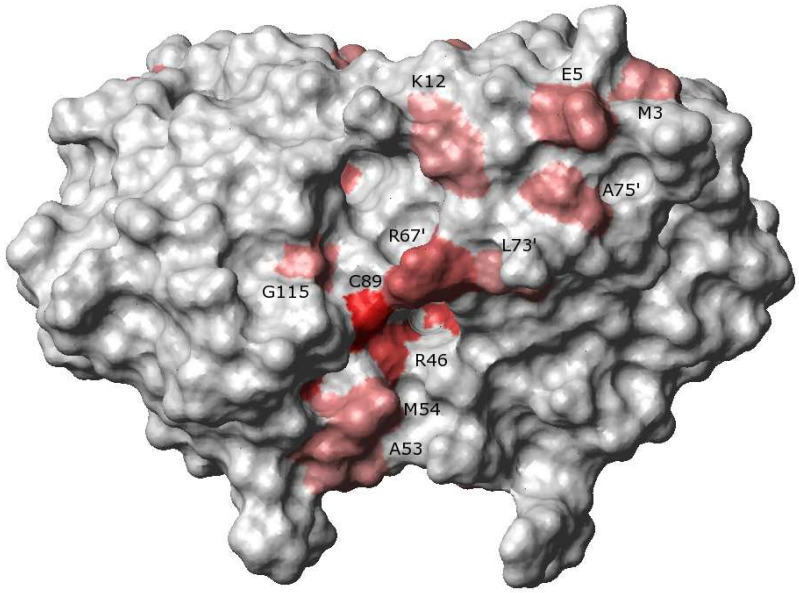

Figure 5.7. Graphic representation of the accessible surface of the Sud protein. The surface is color coded (from grey to red) according to the weighted chemical shift changes induced by the interaction between Sud and Str proteins as presented in panel (C) of Figure 5.5. The peak shifts smaller than a cutoff value of 0.03 ppm were not considered for color mapping. The prime symbol (*e.g.* R67') indicates a residues belonging to the second monomer unit.

upon complex encounter. The residues which show large chemical shift perturbations are virtually the same as those affected by polysulfide binding (Figure 5.5, panels B and A). The most affected residues are part of the polysulfide binding pocket (R45, R67, C89), direct neighbors (G115, M117) and most of the helix V (residues 96-99) which is formed as a continuation of the active-site loop. Based on this spectra comparison it is difficult to distinguish between effects related to protein-protein interaction and effects caused by polysulfide binding. Furthermore, the analysis is complicated by the presumably short length of the sulfur chain attached to the Sud protein in the process of polysulfide transfer. To answer this question a third experiment was performed using a Sud-Str complex where both proteins were fully loaded with polysulfide before the addition to the buffer solution containing an excess of polysulfide-sulfur. In this case the previous chemical shift changes were reproduced at a lower scale (Figure 5.4, yellow/blue spectra and Figure 5.5, panel C). Most affected residues belong to the active

site (R46, R67, C89) and its close vicinity (V45, L97, A98, G99, M117). In Figure 5.7 the largest shifts (> 0.03 ppm) are color coded and mapped on accessible surface of the Sud dimer. Four residues closely surrounding the catalytic cysteine of the Sud protein: R46, M54, R67' and G115, and three additional residues: A53, L73' and K12, form a potential contact surface of the Sud-Str complex.

Chemical shift perturbation mapping was used to probe the interaction between Sud and the second sulfur binding protein (Str) involved in the electron transfer chain catalyzing the polysulfide respiration in *W. succinogenes*. The polysulfide-sulfur transfer between Sud and Str protein was confirmed and a possible protein-protein interface is proposed. In the absence of the polysulfide substrate no interaction between the Sud and the Str protein could be observed, implying a transferase mode of action whereby the two proteins encounter each other and allows the polysulfide-sulfur transfer only when the suitable driving force is present.

5.2 Automated protein structure determination using wavelet de-noised NOESY spectra

5.2.1 Optimal wavelet based de-noising scheme

Spectral noise impairs significantly the automation of peak picking. An efficient tool for spectral de-nosing should be designed in such a way that noise suppression does not affect the fine structure of the signals of interest. For wavelet based de-nosing there are three variables that can be optimized: the wavelet base function, the wavelet transform type and the thresholding procedure. In addition, the decomposition depth (low-frequency cutoff) may affect the fine structure of overlapping peaks. In this work an efficient wavelet based de-nosing scheme for the multidimensional NOESY spectra of isotopically labeled proteins was developed. The effect of various de-noising schemes on the completeness and accuracy of the automatically picked NOESY peak list has been subject to a detailed investigation.

Different schemes for wavelet de-noising were evaluated and compared. These included one-dimensional (1D) and two-dimensional (2D) discrete wavelet transforms (DWT), where each was evaluated for several mother wavelets (Symmlet 5, 8, 10; Daubechies 4, 20; Coiflet 1, 5 and Haar), different de-noising schemes (hard, soft, TI hard and TI soft) and various low-frequency cutoffs ($J = 2 - 5$). De-noising was always applied to the ^{1}HN-^{1}H planes of the NOESY test spectrum and in the case of the 1D DWT the effect of the order in which the two dimensions were de-noised (^{1}H/^{1}HN or ^{1}HN/^{1}H) was examined.

384 different de-noising protocols were tested using the scores described in Section 4.4 for a two-dimensional ^{1}H-^{1}H cross section of the 3D ^{15}N-edited NOESY of Sud protein with significant spectral overlap (Figure 5.8). The Haar wavelet scored low regardless of the shrinkage scheme (methods: 8, 16, 24 and 32 in Figure 5.8). It suffers from the fact that its basis is not continuous which makes it less suitable for a sparse

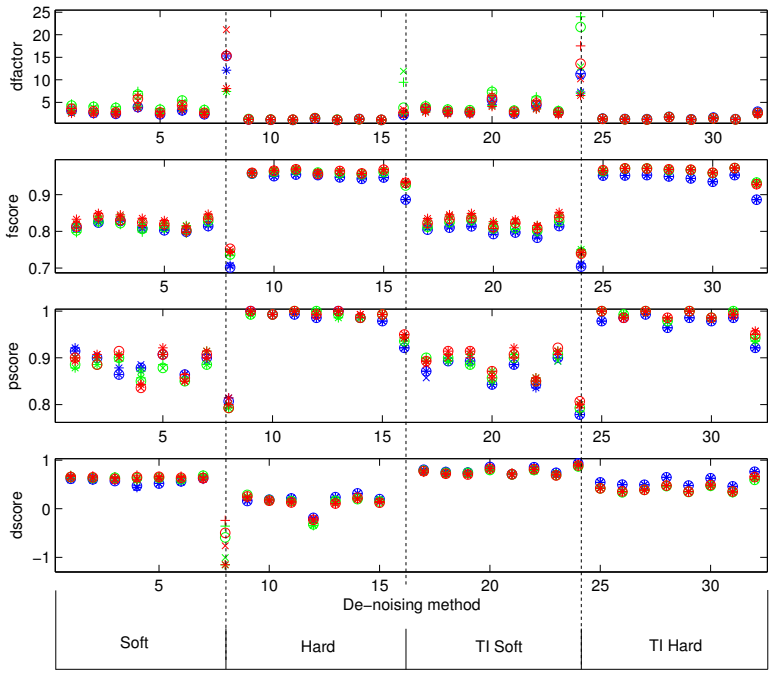

Figure 5.8. Scores for 384 de-noising protocols using a test plane of a 3D ^{15}N-edited NOESY spectrum of the Sud protein. The four sections separated by dashed lines correspond to soft-, hard-, TI soft- and TI hard-thresholding. For each section, the following wavelet bases were used: S5, S8, S10, D2, D20, C1, C5 and Haar. Red and green colors represents the 1D DWT de-noising order ^{1}H/^{1}HN and^{1}HN/^{1}H, respectively whereas blue represents the 2D DWT. The low-frequency cutoffs are represented by the symbols $+$ ($J = 2$), $\circ$ ($J = 3$), $\times$ ($J = 4$) and $*$ ($J = 5$).

representation of smooth functions. Wavelets with better smoothing properties were designed to minimize number of wavelet coefficients for smooth functions. Symmlets and Daubechies wavelets represent a good compromise between noise reduction and the preservation of the fine structure (methods: 1-5, 9-13, 17-21 and 25-29 in Figure 5.8). Compared to the 1D DWT the 2D decomposition is computationally more efficient, however the overall scores were inferior (Figure 5.8: blue spots). The decomposition order for the 1D DWT within the 2D data matrix has little influence although slightly better scores were obtained when the incremented proton dimension was de-noised first (^{1}H/^{1}HN, Figure 5.8: red spots). Soft-thresholding yields the best possible noise suppression (large $dfactor$ and $dscore$) at the expense of fine structure (low $fscore$) and completeness of the peak list (low $pscore$). In contrast, hard-thresholding

Table 5.5. Scores for different de-noising procedures applied on the ^{15}N-edited NOESY of Sud protein.

De-noising Method	*dfactor*	1D DWT *fscore*	*pscore*	*dscore*
S5 Soft	3.254±0.298	0.850	0.812	0.522
S5 TI Soft	2.850±0.262	0.857	0.828	0.576
D4 Soft	5.398±0.651	0.848	0.805	0.525
D4 TI Soft	4.277±0.491	0.849	0.799	0.611
S5 Hard	1.242±0.056	0.973	0.945	0.134
S5 TI Hard	1.335±0.068	0.975	0.943	0.264
D4 Hard	1.621±0.180	0.968	0.942	-0.401
D4 TI Hard	1.800±0.194	0.978	0.933	0.232

De-noising Method	*dfactor*	2D DWT *fscore*	*pscore*	*dscore*
S5 Soft	2.673±0.260	0.829	0.813	0.517
S5 TI Soft	2.943±0.298	0.829	0.806	0.609
D4 Soft	4.045±0.556	0.832	0.798	0.532
D4 TI Soft	4.378±0.533	0.823	0.774	0.655
S5 Hard	1.232±0.056	0.960	0.916	0.183
S5 TI Hard	1.400±0.083	0.957	0.923	0.354
D4 Hard	1.473±0.130	0.960	0.932	-0.447
D4 TI Hard	1.807±0.155	0.964	0.903	0.426

preserves the fine structure at a modest gain of signal-to-noise. TI de-noising proved superior in all scores because it eliminates truncation artifacts and averages residual noise (de-noising methods 16-32 in Figure 5.8). The low-frequency cutoff (J) was not an essential parameter for hard-threshold based de-noising scoress owing to the much larger wavelet coefficients of the peaks (intense singularities) compared to the baseline areas. For soft de-noising lower values of J yield smoothing because all the wavelet coefficients are shrunk regardless of their absolute value. As a general result, for digital signals with less than 2500 data points, a low-frequency cutoff of three ($J = 3$) yields a good compromise between signal-to-noise and resolution.

For a more accurate analysis, the S5 (Symmlet 5) and D4 (Daubechies 4) wavelet de-nosing protocols were applied on the full 3D spectrum (Table 5.5). The low-

frequency cutoff J was set to a value of 3 for all de-noising schemes. The previous result obtained for 2D slices was qualitatively confirmed for the 3D spectrum: depending on the desired result there were two possible de-noising strategies which yield either strong de-noising or high preservation of fine structure, respectively. Soft-thresholding leads to a high signal-to-noise ratio ($dfactor = 2.7 - 5.4$) but suppresses the low intensity signals ($pscore = 0.77 - 0.83$) whereas hard-thresholding preserves the fine structure ($pscore = 0.92 - 0.94$) on the expense of the signal-to-noise gain ($dfactor = 1.2 - 1.8$). Best results were obtained when the 1D DWT was used in combination with the TI de-noising. For soft-thresholding Daubechies wavelets gave the best scores while hard-thresholding gave the best scores for the Symmlet basis. A better compromise could not be found even with more sophisticated thresholding schemes.

5.2.2 NOESY peak list validation

By incorporating the validation filters based on *network anchoring* and *symmetry mapping* (Section 4.3) all de-noising scores were further improved with a minimal loss of real peaks. This is reflected by a larger de-noising score (*dscore*) and minimally smaller peak picking scores (*pscore*) (see Table 5.8). Limitations for this validation scheme are excessively noisy peak lists (when the large number of noise-related peaks are obscuring the relevant signals), incomplete assignment tables, shifted peaks or tight frequency tolerances. Furthermore, unique contacts between amino acids of different structural elements of proteins with high information content may get lost. When the validation filters are applied without prior wavelet de-noising the quality scores indicate that 4% of the real peaks were eliminated while 70% of the noisy entries were removed. However, by combining de-noising and consistency assessment up to 90% of the residual noise can be removed while only 2% additional peaks are eliminated.

Table 5.8. Quality scores after NOESY peak list validation using the network anchoring and symmetry mapping filters.

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

*automatically picked peaks using the original spectrum.

5.2.3 Iterative NOE assignment and structure calculations using wavelet de-noised spectra

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

This strategy was tested using the experimental NOESY data of the Sud dimer. To simplify the assignment procedure the NOE assignment and structure calculations

were carried out only for the monomer unit (residues 20-130). The N-terminal α-helix was not considered since its positioning is essentially determined by the dimer fold. The monomer reference structure was recalculated using only the intra-monomer distance constraints originating from the ^{15}N,^{13}C and methyl-^{13}C edited-NOESY spectra.

The incremental peak lists obtained after wavelet de-noising were implemented in a three stage protocol of iterative NOE assignment and structure calculations with ARIA. *The first stage* of structure calculation started with the 'cleanest' NOESY peak list (*I*) and five iterations in ARIA. In this stage 2117 NOEs were collected from the three heteronuclear NOESY spectra. Besides validation of NOESY peaks, the network anchoring and symmetry mapping filters produced 562 unambiguous NOE assignments. This unambiguous assignment list was verified using the reference model and only 22 entries were found to be misinterpreted. The coupled NOE assignment and structure calculation protocol followed the standard ARIA scheme (Linge et al., 2001) of the first five iterations. To take the best possible advantage of the clean but incomplete peak list (*I*) and to minimize the amount of peaks that may be incompatible with the transient three-dimensional models owing to underestimated upper limits, the *qmove* flag of the violation analysis module in ARIA was used throughout these initial five iterations[1]. In each iteration 30 structures were calculated and the 10 models with lowest energy were used to interpret the spectra in the following cycle. The ambiguity cutoff[2] was gradually decreased from 1 to 0.98. At this stage a bundle of conformers with a mean backbone RMSD of 4.68±1.08 Å between the best 10 models was obtained. The RMSD between the average structure and the reference model was 2.64 Å (Figure5.9, panel B).

In the *second stage* these models were used as a starting point for a new cycle of four ARIA iterations using the peak list (*II*) and after the anchoring/symmetry based

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

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

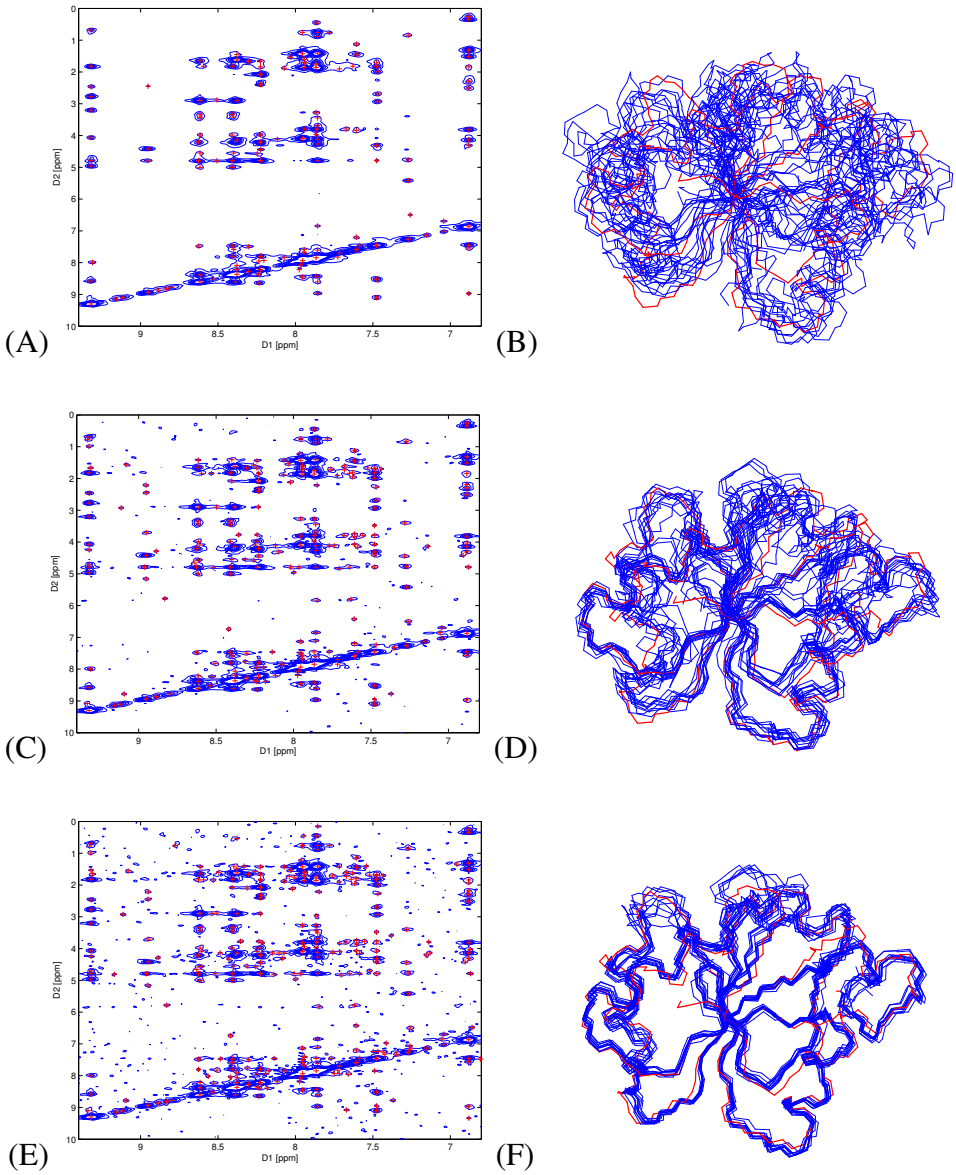

Figure 5.9. (A), (C) and (E) represent a 2D slice of the 3D ^{15}N-edited NOESY spectrum of the Sud protein from *W. succinogens*: (A) after 1D-DWT-D4-TI-Soft de-nosing, (C) after 1D-DWT-S5-TI-Hard de-noising and (E) the original cross section. Red crosses depict the automatically picked peaks for each spectrum. (B), (D) and (F) show backbone plots of the reference structure (red) together with the 10 best conformers (blue) obtained in subsequent stages of automated NOE assignment and structure calculation using NOESY spectra (A), (C) and (E), respectively.

validation (2615 NOEs). The protocol was identical with the one employed in the first stage but no initial assignments were imposed. In this way all assignment possibilities were reassessed based on the previous structural models. After four iterations a bundle of conformers with a mean backbone RMSD of 2.00±0.36 Å and a deviation between the average and the reference structure of 1.72 Å was achieved (Figure 5.9, panel D). Despite a high ambiguity cutoff for the NOE assignments (0.98) which allows a large number of ambiguous distance restraints the calculation converged to a reasonably well-defined model. The sparseness of the cross peak list in this stage does not represent a drastic limitation because NOESY based structure calculations are tolerant with respect to the data incompleteness (Jee and Güntert, 2003).

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

The final list of NOE derived distance restraints (1923 non-redundant restraints) were subject to an ultimate ARIA structure calculation (100 structures) leading to a bundle of the 10 best conformers with a mean backbone RMSD of 0.85±0.2 Å (Figure 5.9, panel F). An identical structure calculation protocol was applied to the distance restraints previously obtained by an interactive approach with manual peak picking. The automated and manual scheme gave similar target functions and almost identical RMSD values. The backbone RMSD between the mean structures of the two bundles (automated versus manual) was 1.06 Å. Table 5.10 presents the structural statistics summary of the three stage automated NOE assignment and structure calculation compared to the corresponding values for the interactive manual approach.

Table 5.10. Structural statistics for the three stages of automated NOE assignment and structure calculation; comparison with the result of the interactive manual approach.

	Stage 1	Stage 2	Stage 3	Manual
NOE cross peaks	2117	2615	3507	2700
NOE distance restraints[a]	1615	1965	1923	1896
Target function [kcal/mol]	2215.1±417.3	944.3±309.1	132.9±7.0	110.6±3.4
backbone RMSD [Å][b]	4.68±1.08	2.00±0.36	0.85±0.20	0.84±0.10
	2.64	1.72	1.06	

[a] unambiguous and ambiguous distance restraints (ADR).
[b] first row denotes the mean backbone RMSD of energetically best 10 models, the second row the RMSD between the ensemble average structure and the reference model. For all RMSD calculations only residues 21-89 and 95-129 were considered.

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

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

concept in ARIA.

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

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

The advantages of the de-noising strategy will be most significant for somewhat noisy NOESY spectra. The required amount of processing to obtain de-noised spectra is very limited, in fact commonly used DWT algorithms are faster than the Fast Fourier Transformation (Mallat, 1989b). Post-processing and peak picking require minimal added computational time to obtain peak lists for different stages of the procedure. The combined software tools provide de-noising, peak picking and integration with export modules to different file formats. Therefore this software should be commonly applicable in conjunction with different programs for combined NOESY assignment and structure calculation. The symmetry and network anchoring filters were directly incorporated into the ARIA program, a software broadly used for iterative NOE

assignment and structure calculations.

5.3 Wavelet de-noising for NMR screening

Principal component analysis (PCA) is a commonly used algorithm for multivariate analysis of NMR screening data. PCA substantially reduces the complexity of data in which a large number of variables are interrelated. For large series of NMR spectra obtained for ligand binding, PCA is employed to visually group spectra with a similar response to ligand binding. The correct classification of the NMR screening data by PCA is a notoriously difficult problem owing to the noise and baseline distortions and to the small spurious shifts caused by pH changes upon addition of the ligand solution. The approach described here uses the noise filtering, baseline correction and data compression ability of wavelet transforms to address this problem. Different schemes for pre-processing the NMR screening data prior to PCA have been compared (Section 4.7: schemes A, B and C). Scheme A involves the standard bucketing approach, scheme B is a hybrid of wavelet de-nosing and bucketing while scheme C propose a combination of PCA with the wavelet coefficient thresholding and multiresolution analysis. The novel concept is to apply the PCA in the multiresolution space of wavelet coefficients, which allows for a selective filtering of noise- and baseline-related artifacts.

In scheme A (Figure 4.5) a standard bucketing approach with 16×16 buckets was used to reduce the size of the data. The result of this bucketing procedure is presented in the first panel of Figure 5.10 which shows a plot of the first three principal components (pc1, pc2 and pc3). A cluster between 0 and 100 on pc1 and pc2 and -40 and 40 on pc3 (blue '+') represents spectra with little change compared to the reference. Positive hits in the screening appear with negative values in pc2 (green '+'). In addition, spectra 42 and 28 (red '+') appear with large values in pc3. The corresponding spectra for both cases show few effects compared to the reference. Figure 5.10D shows

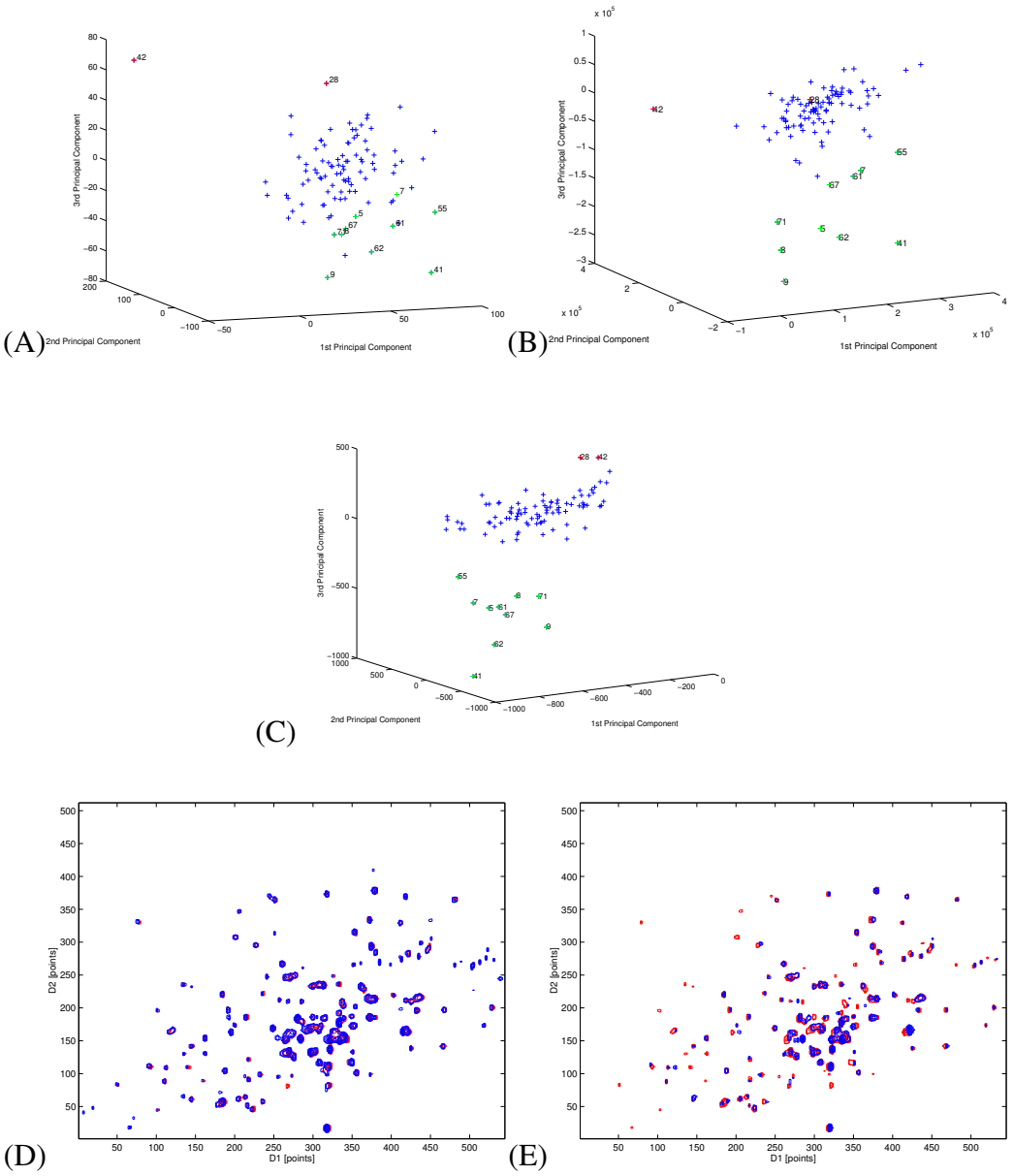

Figure 5.10. (A) The first three principal component obtained using scheme A with 16×16 bucketing prior to PCA for a set of 101 HSQC spectra of hsp90 recorded with different ligands. Each '+' represents one HSQC spectrum. (B) Principal components obtained using scheme B with 16×16 bucketing on wavelet de-noised data using a Symmlet 8 quadrature mirror filter and soft-threshold de-noising prior to PCA. (C) Principal components obtained employing scheme C using a wavelet transform with a Symmlet 8 quadrature mirror filter and MRA including dyadic levels $j = [5,8]$ prior to PCA. (D) spectrum 42 (blue) superimposed to the reference spectrum (red, without ligand) showing few chemical shift changes. (E) spectrum 41 (blue) superimposed to the reference spectrum (red, without ligand) showing significant chemical shift changes.

the HSQC spectrum of the complex form (protein with ligand, blue) superimposed on the reference spectrum (without ligand, red) for the false hit 42. In contrast, Figure 5.10E shows an example for a positive hit with various small chemical shift changes compared to the reference.

In scheme B (Figure 4.5) a wavelet de-nosing step was applied prior to thresholding and bucketing (see Section 4.7 for details of the procedure). Here wavelet shrinkage is used for de-noising and smoothing of the spectra but not for data compression. Combined de-noising/bucketing depicted in Figure 5.10B exhibits improved clustering in spectra of protein with non-binding ligands. In addition, some hits are clearly separated (green '+'). Spectrum 42 appears again as a false hit (red '+') whereas spectrum 28 joins the cluster around the reference. The improvement for spectrum 28 can be explained by the smoothing effect of wavelet thresholding on the spectrum.

In scheme C (Figure 4.5) PCA was applied directly to the wavelet coefficients. Since the wavelet transform is a unitary transformation (Eqn. 2.27) eigenvalues of the wavelet coefficients are equivalent to the eigenvalues of the original data. Therefore a PCA analysis performed directly on the wavelet coefficients conveys the multivariate properties as if it was applied on the original data. In addition, the soft-thresholding of the wavelet coefficients eliminates the stochastic component of the spectra (de-noising), minimizes insignificant spectral perturbations (smoothing) and decreases the size of the data matrix (compression). Figure 5.10C shows good clustering for the first three principal components obtained by applying the PCA on the sparse matrix of thresholded wavelet coefficients. In this case spectra 42 and 28 appear on the edge of the cluster around the reference.

For a quantitative comparison of the three different schemes, a compression factor, a de-noising factor, a clustering factor and the CPU time for the PCA were evaluated (Table 5.10, see Section 4.7 for details). Compression, de-nosing and clustering factors are better for scheme B compared to scheme A owing to the reduced noise in spectra. For scheme C the de-noising factor is better than for schemes A and B owing to the

larger number of zeroes in the thresholded matrix. However, the compression factor is lower because the number of common zeroes between all spectra is much lower for de-noising in wavelet space than for de-noising of the actual spectra. The separation of outliers from the cluster representing spectra of protein with non-binding ligands is greatly improved in scheme C. This leads to a higher clustering factor compared to schemes A and B. The lower compression rate of scheme C leads to increased CPU time.

Table 5.12. Compression factors, de-noising factors, clustering factors and elapsed CPU time[1] obtained for the three different schemes (A, B and C) of multivariate analysis of the NMR screening data using a set of 101 HSQC spectra of hsp90 protein recorded in the presence of different ligands.

Scheme	Compression factor	De-noising factor	Clustering factor	CPU time [s]
A	1008	0.858	0.106	0.58
B	1231	0.861	0.198	0.41
C	11	0.903	0.346	15.25

[1]CPU time required for the PCA of the pre-processed data on a 1.5 GHz AMD processor.

Further analysis of the false hit 42 in the bucketing schemes showed that it is not the outcome of the local noise dissimilarities but rather an artifact resulting from peak shifts in the vicinity of the bucket borders. Figure 5.11A shows an example of signals in spectrum 42 which cause border artifacts in bucketing. When several effects of this kind are accumulated in one spectrum relatively large principal components will be observed. This cumulative effect can be shown by computing the sum of the squared differences between the reference and each spectrum at bucket borders. Figure 5.11B shows that this function has a clear maximum for spectrum 42 due to an accumulation of bucket border artifacts. The effect of artifacts on bucket borders has also been confirmed using simulated spectra (not shown).

While PCA has become a standard technique for data reduction and visualization of large data sets, the preparation of NMR data for PCA remains difficult. Filters ap-

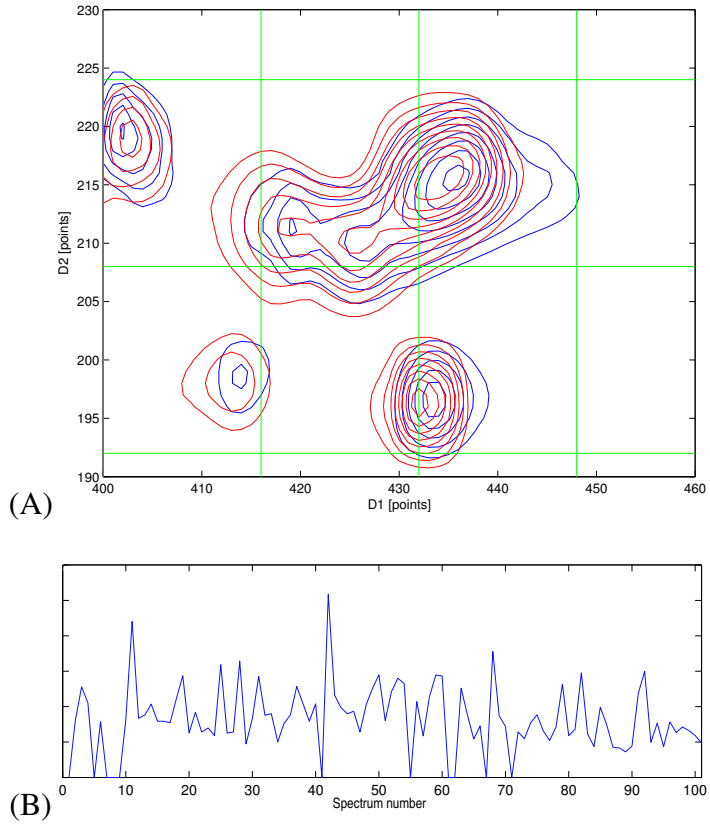

Figure 5.11. (A) Example of peaks which cause a border effect in the bucketing schemes. Blue corresponds to spectrum 42 and red to the reference. The green grid represents the bucket borders. (B) Summed squared differences between each of the 101 spectra and the reference spectrum calculated at the bucket borders. Real hits (spectra 5, 7, 8, 9, 41, 55, 61, 62, 67 and 71) were not included. For a better visualization all the spectra identified as true hits were excluded from this analysis (zeroes in the plot).

plied prior to PCA should reduce the size of data to improve computational efficiency and minimize the sensitivity towards small irrelevant shifts in the NMR data. This has typically been achieved employing bucketing as a simple and highly efficient filter. Unfortunately bucketing may introduce artifacts when peaks move on borders between buckets and in the case of spectra with large variations of the background noise levels. The addition of spectral points into one bucket causes a modest reduction of noise depending on the size of the bucket. However, large buckets would be required to achieve a noticeable noise reduction. Applying a threshold to the experimental data is a frequently used alternative. Nevertheless, sharp thresholds tend to distort the buck-

ets leading to more severe artifacts. In addition, with increasing resolution of spectra at higher magnetic fields with proton frequencies of up to 900 MHz, typical bucketing schemes reduce the effective resolution substantially. For this reason more subtle methods of smoothing and noise suppression are required.

In this study it has been shown that wavelet coefficient thresholding is a suitable alternative with desirable properties for subsequent PCA analysis. Two different schemes which combine the wavelet transforms to suppress noise related coefficients with the PCA analysis have been tested. When PCA is applied to spectra after wavelet de-noising subsequent bucketing is still required to reduce the size of the data (scheme B). This scheme showed improved clustering owing to the reduced noise contribution to buckets. It also eliminates the noise-related artifacts observed for spectrum 28, but not the bucketing artifacts observed for spectrum 42. Further improved clustering was achieved when PCA was directly applied to the wavelet coefficients (scheme C). This scheme eliminates noise-related (spectrum 28) and bucketing artifacts (spectrum 42) efficiently. The scheme offers a modest and scalable smoothing for one- or two-dimensional NMR data. The result can be optimized by selecting threshold levels in wavelet space and suitable levels to be suppressed in MRA. The effect of MRA will be more pronounced for data sets with strong baseline distortions typical for one-dimensional spectra.

Although the formation of data buckets is computationally less demanding than calculating the wavelet transformation, the additional computational effort seems justified considering the preservation in fine structure and the reduction in artefacts that can be achieved. The computing time of the lifting scheme used to obtain the wavelet coefficients is proportional to the number of data points N of the data set and therefore by a factor of $\log(N)$ faster than the Fast Fourier Transformation. In scheme C where PCA is performed in wavelet space no inverse transform is required. Once data is represented in wavelet space different thresholding or MRA schemes can be applied rapidly.

6 Zusammenfassung

Die Ermittlung von Proteinstukturen mittels NMR-Spektroskopie ist ein komplexer Prozess, wobei die Resonanzfrequenzen und die Signalintensitäten den Atomen des Proteins zugeordnet werden. Zur Bestimmung der räumlichen Proteinstruktur sind folgende Schritte erforderlich: die Präparation der Probe und $^{15}N/^{13}C$ Isotopenanreicherung, Durchführung der NMR Experimente, Prozessierung der Spektren, Bestimmung der Signalresonanzen ('Peak-picking'), Zuordnung der chemischen Verschiebungen, Zuordnung der NOESY-Spektren und das Sammeln von konformationellen Strukturparametern, Strukturrechnung und Strukturverfeinerung. Aktuelle Methoden zur automatischen Strukturrechnung nutzen eine Reihe von Computeralgorithmen, welche Zuordnungen der NOESY-Spektren und die Strukturrechnung durch einen iterativen Prozess verbinden. Obwohl neue Arten von Strukturparametern wie dipolare Kopplungen, Orientierungsinformationen aus kreuzkorrelierten Relaxationsraten oder Strukturinformationen, die sich in Gegenwart paramagnetischer Zentren in Proteinen ergeben, wichtige Neuerungen für die Proteinstrukturrechnung darstellen, sind die Abstandsinformationen aus NOESY-Spektren weiterhin die wichtigste Basis für die NMR-Strukturbestimmung.

Der hohe zeitliche Aufwand des 'peak-picking' in NOESY-Spektren ist hauptsächlich bedingt durch spektrale Überlagerung, Rauschsignale und Artefakte in NOESY-Spektren. Daher werden für das effizientere automatische 'Peak-picking' zuverlässige Filter benötigt, um die relevanten Signale auszuwählen. In der vorliegenden Arbeit wird ein neuer Algorithmus für die automatische Proteinstrukturrech-

nung beschrieben, der automatisches 'Peak-picking' von NOESY-Spektren beinhaltet, die mit Hilfe von Wavelets entrauscht wurden. Der kritische Punkt dieses Algorithmus ist die Erzeugung inkrementeller Peaklisten aus NOESY-Spektren, die mit verschiedenen auf Wavelets basierenden Entrauschungsprozeduren prozessiert wurden. Mit Hilfe entrauschter NOESY-Spektren erhält man Signallisten mit verschiedenen Konfidenzbereichen, die in unterschiedlichen Schritten der kombinierten NOE-Zuordnung/Strukturrechnung eingesetzt werden. Das erste Strukturmodell beruht auf stark entrauschten Spektren, die die konservativste Signalliste mit als weitgehend sicher anzunehmenden Signalen ergeben. In späteren Stadien werden Signallisten aus weniger stark entrauschten Spektren mit einer größeren Anzahl von Signalen verwendet. Die Auswirkung der verschiedenen Entrauschungsprozeduren auf Vollständigkeit und Richtigkeit der NOESY Peaklisten wurde im Detail untersucht. Durch die Kombination von Wavelet-Entrauschung mit einem neuen Algorithmus zur Integration der Signale in Verbindung mit zusätzlichen Filtern, die die Konsistenz der Peakliste prüfen ('Network-anchoring' der Spinsysteme und Symmetrisierung der Peakliste), wird eine schnelle Konvergenz der automatischen Strukturrechnung erreicht. Der neue Algorithmus wurde in ARIA integriert, einem weit verbreiteten Computerprogramm für die automatische NOE-Zuordnung und Strukturrechnung. Der Algorithmus wurde an der Monomereinheit der Polysulfid-Schwefel-Transferase (Sud) aus *Wolinella succinogenes* verifiziert, deren hochaufgelöste Lösungsstruktur vorher auf konventionelle Weise bestimmt wurde.

Neben der Möglichkeit zur Bestimmung von Proteinlösungsstrukturen bietet sich die NMR-Spektroskopie auch als wirkungsvolles Werkzeug zur Untersuchung von Protein-Ligand- und Protein-Protein-Wechselwirkungen an. Sowohl NMR Spektren von isotopenmarkierten Proteinen, als auch die Spektren von Liganden können für das 'Screening" nach Inhibitoren benutzt werden. Im ersten Fall wird die Sensitivität der ^{1}H- und ^{15}N-chemischen Verschiebungen des Proteinrückgrats auf kleine geometrische oder elektrostatische Veränderungen bei der Ligandbindung als Indikator

benutzt. Als 'Screening'-Verfahren, bei denen Ligandensignale beobachtet werden, stehen verschiedene Methoden zur Verfügung: Transfer-NOEs, Sättigungstransferdifferenzexperimente (STD, 'saturation transfer difference'), ePHOGSY, diffusionseditierte und NOE-basierende Methoden. Die meisten dieser Techniken können zum rationalen Design von inhibitorischen Verbindungen verwendet werden. Für die Evaluierung von Untersuchungen mit einer großen Anzahl von Inhibitoren werden effiziente Verfahren zur Mustererkennung wie etwa die PCA ('Principal Component Analysis') verwendet. Sie eignet sich zur Visualisierung von Ähnlichkeiten bzw. Unterschieden von Spektren, die mit verschiedenen Inhibitoren aufgenommen wurden. Die experimentellen Daten werden zuvor mit einer Serie von Filtern bearbeitet, die u.a. Artefakte reduzieren, die auf nur kleinen Änderungen der chemischen Verschiebungen beruhen. Der am weitesten verbreitete Filter ist das sogenannte 'bucketing', bei welchem benachbarte Punkte zu einen 'bucket' aufsummiert werden. Um typische Nachteile der 'bucketing'-Prozedur zu vermeiden, wurde in der vorliegenden Arbeit der Effekt der Wavelet-Entrauschung zur Vorbereitung der NMR-Daten für PCA am Beispiel vorhandener Serien von HSQC-Spektren von Proteinen mit verschiedenen Liganden untersucht. Die Kombination von Wavelet-Entrauschung und PCA ist am effizientesten, wenn PCA direkt auf die Wavelet-Koeffizienten angewandt wird. Durch die Abgrenzung ('thresholding') der Wavelet-Koeffizienten in einer Multiskalenanalyse wird eine komprimierte Darstellung der Daten erreicht, welche Rauschartefakte minimiert. Die Kompression ist anders als beim 'bucketing' keine 'blinde' Kompression, sondern an die Eigenschaften der Daten angepasst. Der neue Algorithmus kombiniert die Vorteile einer Datenrepresentation im Wavelet-Raum mit einer Datenvisualisierung durch PCA. In der vorliegenden Arbeit wird gezeigt, dass PCA im Wavelet-Raum ein optimiertes 'clustering' erlaubt und dabei typische Artefakte eliminiert werden.

Darüberhinaus beschreibt die vorliegende Arbeit eine *de novo* Strukturbestimmung der periplasmatischen Polysulfid-Schwefel-Transferase (Sud) aus dem anaer-

oben gram-negativen Bakterium *Wolinella succinogenes*. Das Sud-Protein ist ein polysulfidbindendes und transferierendes Enzym, das bei niedriger Polysulfidkonzentration eine schnelle Polysulfid-Schwefel-Reduktion katalysiert. Sud ist ein 30 kDa schweres Homodimer, welches keine prosthetischen Gruppen oder schwere Metallionen enthält. Jedes Monomer enhält ein Cystein, welches kovalent bis zu zehn Polysulfid-Schwefel (S_n^{2-}) Ionen bindet. Es wird vermutet, dass Sud die Polysulfidkette auf ein katalytischen Molybdän-Ion transferiert, welches sich im aktiven Zentrum des membranständigen Enzyms Polysulfid-Reduktase (Psr) auf dessen dem Periplasma zugewandten Seite befindet. Dabei wird eine reduktive Spaltung der Kette katalysiert.

Die Lösungsstruktur des Homodimeres Sud wurde mit Hilfe heteronuklearer, mehrdimensionaler NMR-Techniken bestimmt. Die Struktur beruht auf von NOESY-Spektren abgeleiteten Distanzbeschränkungen, Rückgratwasserstoffbindungen und Torsionswinkeln, sowie auf residuellen dipolaren Kopplungen, die für die Verfeinerung der Struktur und für die relative Orientierung der Monomereinheiten wichtig waren. In den NMR Spektren der Homodimere haben alle symmetrieverwandte Kerne äquivalente magnetische Umgebungen, weshalb ihre chemischen Verschiebungen entartet sind. Die symmetrische Entartung vereinfacht das Problem der Resonanzzuordnung, da nur die Hälfte der Kerne zugeordnet werden müssen. Die NOESY-Zuordnung und die Strukturrechnung werden dadurch erschwert, dass es nicht möglich ist, zwischen den Intra-Monomer-, Inter-Monomer- und Co-Monomer- (gemischten) NOESY-Signalen zu unterscheiden. Um das Problem der Symmetrie-Entartung der NOESY-Daten zu lösen, stehen zwei Möglichkeiten zur Verfügung: (*I*) asymmetrische Markierungs-Experimente, um die intra- von den intermolekularen NOESY-Signalen zu unterscheiden, (*II*) spezielle Methoden der Strukturrechnung, die mit mehrdeutigen Distanzbeschränkungen arbeiten können. Die in dieser Arbeit vorgestellte Struktur wurde mit Hilfe der Symmetrie-ADR- ('Ambigous Distance Restraints') Methode in Kombination mit Daten von asymetrisch isotopenmarkierten Dimeren berechnet. Die Koordinaten des Sud-Dimers zusammen mit den NMR-basierten Strukturdaten wur-

den in der RCSB-Proteindatenbank[1] unter der PDB-Nummer 1QXN abgelegt.

Das Sud-Protein zeigt nur wenig Homologie zur Primärsequenz anderer Proteine mit ähnlicher Funktion und bekannter dreidimensionaler Struktur. Bekannte Proteine sind die Schwefeltransferase oder das Rhodanese-Enzym[2], welche beide den Transfer von einem Schwefelatom eines passenden Donors auf den nukleophilen Akzeptor (z.B von Thiosulfat auf Cyanid) katalysieren. Die dreidimensionalen Strukturen dieser Proteine zeigen eine typische α/β Topologie und haben eine ähnliche Umgebung im aktiven Zentrum bezüglich der Konformation des Proteinrückgrades. Die Schleife im aktiven Zentrum umgibt das katalytische Cystein, welches in allen Rhodanese-Enzymen vorhanden ist, und scheint im Sud-Protein flexibel zu sein (fehlende Resonanzzuordnung der Aminosäuren 89-94). Das Polysulfidende ragt aus einer positiv geladenen Bindungstasche heraus (Reste: R46, R67, K90, R94), wo Sud wahrscheinlich in Kontakt mit der Polysulfidreduktase tritt. Das strukturelle Ergebnis wurde durch Mutageneseexperimente bestätigt. In diesen Experimenten konnte gezeigt werden, dass alle Aminosäurereste im aktiven Zentrum essentiell für die Schwefeltransferase-Aktivität des Sud-Proteins sind. Die Substratbindung wurde früher durch den Vergleich von [^{15}N,^{1}H]-TROSY-HSQC-Spektren des Sud-Proteins in An- und Abwesenheit des Polysulfidliganden untersucht. Bei der Substratbindung scheint sich die lokale Geometrie der Polysulfidbindungsstelle und der Dimerschnittstelle zu verändern. Die konformationellen Änderungen und die langsame Dynamik, hervorgerufen durch die Ligandbindung können die weitere Polysulfid-Schwefel-Aktivität auslösen.

Ein zweites Polysulfid-Schwefeltransferaseprotein (Str, 40 kDa) mit einer fünffach höheren nativen Konzentration im Vergleich zu Sud wurde im Bakterienperiplasma von *Wolinella succinogenes* entdeckt. Es wird angenommen, dass beide Protein einen Polysulfid-Schwefel-Komplex bilden, wobei Str wässriges Polysulfid sammelt und an Sud abgibt, welches den Schwefeltransfer zum katalytischen Molybdän-Ion auf das

[1] http://www.rcsb.org/

[2] *e.g.* GlpE Protein aus *Escherichia coli*, Rhodanese Protein aus *Azobacter vinelandii* und Rinderleber Rhodanese Protein.

aktive Zentrum der dem Periplasma zugewandten Seite der Polysulfidreduktase durchführt. Änderungen chemischer Verschiebungen in [^{15}N,^{1}H]-TROSY-HSQC-Spektren zeigen, dass ein Polysulfid-Schwefeltransfer zwischen Str und Sud stattfindet. Eine mögliche Protein-Protein-Wechselwirkungsfläche konnte bestimmt werden. In der Abwesenheit des Polysulfidsubstrates wurden keine Wechselwirkungen zwischen Sud und Str beobachtet, was die Vermutung bestätigt, dass beide Proteine nur dann miteinander wechselwirken und den Polysulfid-Schwefeltransfer ermöglichen, wenn als treibende Kraft Polysulfid präsent ist.

7 CURRICULUM VITAE

Name: Felician Dancea

Date of Birth: April 25, 1975

Nationality: Romanian

Education and Qualifications

2000-present: Dissertation to achieve a Ph.D. in Biochemistry.

Institute of Biophysical Chemistry, J. W. Goethe-University, Frankfurt am Main.

Advisers: Prof. Dr. Heinz Rüterjans, PD Dr. Ulrich Günther.

1994-2000: M.Sc. in Physics, Physics Engineer.

Faculty of Physics, "Babes-Bolyai" University, Cluj-Napoca, Romania.

Major specialization: Technological Physics (Physics Engineering).

Additional specializations: Biophysics, Physics of Nuclear Radiations.

Scientific Activities

1. International Max Planck Research School courses, winter semester 2001-2002, J. W. Goethe-University, Frankfurt am Main.

2. EMBO Course on Structure Determination of Biological Macromolecules by Solution NMR, September 2001, EMBL Heidelberg, Germany.

3. Research stage at the Structural Bioinformatics Unit, June 2001, Pasteur Institute, Paris, France.

4. Research work in the Laboratory of Nuclear Physics of Gent University, March-June 2000, Belgium.

Workshops and Conferences

1. Spine NMR software workshop, May 2004, Regensburg, Germany.
   Oral presentation: "NMRLab - Advanced NMR data processing in MATLAB".
   Demo presentation of NMRLab software package.

2. 5th ENC, Experimental Magnetic Resonance Conference, April 2004, Asilomar, CA, USA.
   Poster presentations:
   (I) "NMR structure of the polysulfide-sulfur transferase protein (Sud) from *Wolinella succinogenes* and interactions with a second polysulfide-sulfur transferase protein (Str)"
   (II) "Improved automatic structure determination using wavelet de-noised NMR spectra".

3. 7th User Meeting of the European Large Scale Facilities for NMR, November 2003, Oosterbeek, Netherlands.
   Poster presentations:
   (I) "Automatic peak picking using wavelet de-noised NMR spectra"
   (II) "Using wavelet de-noised NMR spectra in screening".

4. Mini-symposium for foreign students working for their Ph.D. thesis in biology, biological chemistry and medicine at the J. W. Goethe-University of Frankfurt, October 2002, Frankfurt am Main.
   Oral presentation: "NMR structure determination of the Sud dimer from *Wolinella succinogenes*".

Publications

1. Lin Y.J.[*], Dancea F.[*], Löhr F., Klimmek O., PfeifferMarek S., Nilges M., Wienk H., Kröger A., Rüterjans H. (2003) Solution structure of the 30 kDa polysulfide-sulfur transferase homodimer from *Wolinella succinogenes*. *Biochemistry*, **43**, 141-824.

2. Dancea F. and Günther, U. Automated protein NMR structure determination using wavelet de-noised NOESY spectra. *Submitted*.

3. Trbovic N.[**], Dancea F.[**], Langer T., Günther, U. Using wavelet de-noised NMR spectra in NMR screening. *Submitted*.

4. Dancea F. , Löhr F., Klimmek O., Rüterjans H. NMR study of the interaction between two polysulfide-sulfur binding proteins from *Wolinella succinogenes*. *In preparation*.

5. Cosma C., Dancea F., Jurcut T. and Ristoiu D. (2001) Determination of ^{222}Rn emanation fraction and diffusion coefficient in concrete using accumulation chambers and the influence of humidity and radium distribution. *App. Rad. and Isotopes*, **54**, 467-473.

6. Dancea F. , Poffijn A., Cosma C. (2001) The influence of the relative humidity on the detection efficiency of RADIM 2P radon monitoring devices. *Studia-Physica*, **XLVI**, 47-52.

*L.Y.J. and D.F. made equal contributions to this publication.
**T.N. and D.F. made equal contributions to this publication.

Bibliography

Abragam, A. (1967) *Principles of nuclear magnetism*. Oxford University Press, New York.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Andersson, P., Annila, A. and Otting, G. (1998) *J. Magn. Reson.*, **133**, 364–367.

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

Atkinson, R. and Saudek, V. (2002) *FEBS Lett.*, **510**, 1–4.

Bakshi, B. R. (1998) *AIChE J.*, 1596–1610.

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

Belton, P., Colquhoun, I., Kemsley, E., Delgadillo, I., Roma, P., Dennis, M., Sharman, M., Holmes, E., Nicholson, J. and Spraul, M. (1998) *Food Chem.*, **61**, 207–213.

Bertini, I., Dalvit, C., Huber, J., Luchinat, C. and Piccioli, M. (1997) *FEBS Lett.*, **415**, 45–48.

Bordo, D., Deriu, D., Colnaghi, R., Carpen, A., Pagani, S. and Bolognesi, M. (2000) *J. Mol. Biol.*, **298**, 691–704.

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. and Warren, G. L. (1998) *Acta Crystallogr. D Biol. Crystallogr.*, **54 (Pt 5)**, 905–921.

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

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

Chen, A. and Shapiro, M. J. (1998) *J. Am. Chem. Soc.*, **120**, 10258–10259.

Chen, A. and Shapiro, M. J. (2000) *J. Am. Chem. Soc.*, **122**, 414–415.

Chen, D., Hu, B., Shao, X. and Su, Q. (2004) *Anal. Bioanal. Chem.*, **379**, 143–148.

Clarkson, J. and Campbell, I. (2003) *Biochem. Soc. Trans.*, **31**, 1006–1009.

Clore, G., Gronenborn, A. and Tjandra, N. (1998a) *J. Magn. Reson.*, **131**, 159–162.

Clore, G. M., Gronenborn, A. M. and Bax, A. (1998b) *J. Magn. Reson.*, **133**, 216–221.

Cobas, J., Tahoces, P., Martin-Pastor, M., Penedo, M. and Javier Sardina, F. (2004) *J. Magn. Reson.*, **168**, 288–295.

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

Collantes, E., Duta, R., Welsh, W., Zielinski, W. and Brower, J. (1997) *Anal. Chem.*, **69**, 1392–1397.

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

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

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

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

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

Duarte, I., Barros, A., Belton, P., Righelato, R., Spraul, M., Humpfer, E. and Gil, A. (2002) *J. Agric. Food Chem.*, **50**, 2475–2481.

Ehrentreich, F. and Summchen, L. (2001) *Anal. Chem.*, **73**, 4364–73.

Farmer, B. T. and Venters, R. A. (1995) *J. Am. Chem. Soc.*, **117**, 4187–4188.

Ferentz, A. E., Opperman, T., Walker, G. C. and Wagner, G. (1997) *Nat. Struct. Biol.*, **4**, 979–983.

Fischer, P. and Defranceschi, M. (1998) *J. Numer. Anal.*, **35**, 1–12.

Folmer, R. H., Hilbers, C. W., Konings, R. N. and Nilges, M. (1997) *J. Biomol. NMR*, **9**, 245–258.

Gayathri, C., Bothner-By, A. A., van Zijl, P. C. M. and Maclean, C. (1982) *C. Chem. Phys. Lett.*, **87**, 192–196.

Geyer, M., Neidig, K. P. and Kalbitzer, H. R. (1995) *J. Mag. Res.*, **B 109**, 31–38.

Golub, G. and van Loan, C. (1996) *Matrix Computations, Third Edition*. The Johns Hopkins University Press, London.

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

Grishaev, A. and Llinas, M. (2004) *J. Biomol. NMR*, **28**, 1–10.

Grzesiek, S., Wingfield, P., Stahl, S., Kaufman, J. D. and Bax, A. (1995) *J. Am. Chem. Soc.*, **117**, 9594–9595.

Güntert, P. (1998) *Q. Rev. of Biophys.*, **31**, 145–237.

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

Güntert, P. (2004) *Methods Mol. Biol.*, **278**, 353–378.

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

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

Habeck, M., Rieping, W., Linge, J. and Nilges, M. (2004) *Methods Mol. Biol.*, **278**, 379–402.

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

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

Jackson, J. E. (1991) *A user's Guide to Principal Components*. Wiley, New York.

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

Johnson, S. C. (1967) *Psychometrika*, **2**, 241–254.

Karplus, M. (1963) *J. Amer. Chem. Soc.*, 2870–2871.

Klimmek, O., Kreis, V., Klein, C., Simon, J., Wittershagen, A. and Kröger, A. (1998) *Eur. J. Biochem.*, **253**, 263–269.

Klimmek, O., Kröger, A., Steudel, R. and Holdt, G. (1991) *Arch. Microbiol.*, **155**, 177–182.

Klimmek, O., Stein, T., Pisa, R., Simon, J. and Kröger, A. (1999) *Eur. J. Biochem.*, **263**, 79–84.

Koradi, R., Billeter, M., Engeli, M., Güntert, P. and Wüthrich, K. (1998) *J. Magn. Reson.*, **135**, 288–97.

Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graph.*, **14**, 51–55.

Kraulis, P. (1994) *J. Mol. Biol.*, **243**, 696–718.

Kreis-Kleinschmidt, V., Fahrenholz, F., Kojro, E. and Kröger, A. (1995) *Eur. J. Biochem.*, **227**, 137–142.

Laakso, J., Juhola, M., Surakka, V., Aula, A. and Partala, T. (2001) *Medinfo*, **10**, 489–492.

Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. and Thornton, J. M. (1996) *J. Biomol. NMR*, **8**, 477–486.

Lerche, M. H., Meissner, A., Poulsen, F. M. and Sørensen, O. W. (1999) *J. Magn. Reson.*, **140**, 259–263.

Lin, M., Shapiro, M. J. and Wareing, J. R. (1997) *J. Am. Chem. Soc.*, **119**, 5249 – 5250.

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

Lindon, J., Nicholson, J., Holmes, E. and Everett, J. (2000) *Concepts Magn. Reson.*, **12**, 289–320.

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

Linge, J., Habeck, M., Rieping, W. and Nilges, M. (2004) *J. Magn. Reson.*, **167**, 334–342.

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

Linge, J. P. and Nilges, M. (1999) *J. Biomol. NMR*, **13**, 51–59.

Löhr, F. and Rüterjans, H. (2002) *J. Magn. Reson.*, **156**, 10–18.

Ma, X. G. and Zhang, Z. X. (2003) *Anal. Chim. Acta*, **485**, 233–239.

Main, P. and Wilson, J. (2000) *Acta Crystallogr. D Biol. Crystallogr.*, **56 (Pt 5)**, 618–624.

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

Mallat, S. (1989b) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.

Mallat, S. (1998) *A wavelet tour of signal processing.*. Academic Press.

Mayer, M. and Meyer, B. (1999) *Angew. Chem. Int. Ed.*, **35**, 1784–1788.

Mayer, M. and Meyer, B. (2000) *J. Med. Chem.*, **43**, 2093–2099.

Mayer, M. and Meyer, B. (2001) *J. Am. Chem. Soc.*, **123**, 6108–6117.

Meiler, J., Blomberg, N., Nilges, M. and Griesinger, C. (2000) *J. Biomol. NMR*, **16**, 245–252.

Melacini, G. (2000) *J. Am. Chem. Soc.*, **122**, 9735–9738.

Meyer, B., Weimar, T. and Peters, T. (1997) *Eur. J. Biochem.*, **246**, 705–709.

Montelione, G., Zheng, D., Huang, Y., Gunsalus, K. and Szyperski, T. (2000) *Nat. Struct. Biol.*, **7**, 982–985.

Moseley, H. and Montelione, G. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Mulder, F., Schipper, D., Bott, R. and Boelens, R. (1999) *J. Mol. Biol.*, **292**, 111–123.

Mumenthaler, C. and Braun, W. (1995) *J. Mol. Biol.*, **254**, 465–80.

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

Neri, D., Szyperski, T., Otting, G., Senn, H. and Wüthrich, K. (1989) *Biochemistry*, **28**, 7510–7516.

Neuhaus, D. and Williamson, M. P. (1989) *The Nuclear Overhausr Effect in Structural and Conformational Analysis*. New York: VCH.

Nicholson, J., Connelly, J., Lindon, J. and Holmes, E. (2002) *Nat. Rev. Drug. Discov.*, **1**, 153–161.

Nilges, M. (1993) *Proteins*, **17**, 297–309.

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

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

Nilges, M. and O'Donoghue, S. I. (1998) *Prog. in NMR Spect.*, **32**, 107–139.

O'Donoghue, S. I. and Nilges, M. (1999) *Biol. Mag. Res.*, **17**, 131–161.

Perrin, C., Walczak, B. and Massart, D. (2001) *Anal. Chem.*, **73**, 4903–4917.

Pervushin, K., Riek, R., Wider, G. and Wüthrich, K. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 12366–12371.

Pervushin, K., Riek, R., Wider, G. and Wüthrich, K. (1998) *J. Am. Chem. Soc.*, **120**, 6394–6400.

Ploegman, J. H., Drent, G., Kalk, K. H., Hol, W. G., Heinrikson, R. L., Keim, P., Weng, L. and Russell, J. (1978) *Nature*, **273**, 124–129.

Prisner, T., Lyubenova, S., Atabay, Y., MacMillan, F., Kröger, A. and Klimmek, O. (2003) *J. Biol. Inorg. Chem.*, **8**, 419–426.

Ross, A., Schlotterbeck, G., Klaus, W. and Senn, H. (2000) *J. Biomol. NMR*, **16**, 139–146.

Rückert, M. and Otting, G. (2000) *J. Am. Chem. Soc.*, **122**, 7793–7797.

Salzmann, M., Wider, G., Pervushin, K., Senn, H. and Wuthrich, K. (1999) *J. Am. Chem. Soc.*, **121**, 844–848.

Saupe, A. and Englert, G. (1963) *Phys. Rev. Lett.*, **11**, 462–465.

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

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

Shao, X., Li, W., Chen, G. and Su, Q. (1999) *J. Anal. Chem.*, 215–218.

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

Shuker, S., Hajduk, P., Meadows, R. and Fesik, S. (1996) *Science*, **274**, 1531–1534.

Solomon, I. (1955) *Phys. Rev.*, **99**, 559–565.

Sørensen, O. W., Eich, G. W., Levitt, M. H., Bodenhausen, G. and Ernst, R. R. (1983) *Prog. NMR Spectrosc.*, **16**, 163–192.

Sorzano, C. O., Jonic, S., El-Bez, C., Carazo, J. M., De Carlo, S., Thevenaz, P. and Unser, M. (2004) *J. Struct. Biol.*, **146**, 381–392.

Spallarossa, A., Donahue, J. L., Larson, T. J., Bolognesi, M. and Bordo, D. (2001) *Structure*, **9**, 1117–1125.

Staunton, D., Owen, J. and Campbell, I. (2003) *Acc. Chem. Res.*, **36**, 207–214.

Steudel, R., Pridohl, M., Buschmann, J. and Luger, P. (1995) *Chem. Ber.*, **128**, 725–728.

Stockmann, B. and Dalvit, C. (2002) *Progr. Nucl. Magn. Reson. Spectroscopy*, **41**, 187–231.

Teppola, P. and Minkkinen, P. (2000) *J. Chemometrics*, 383–399.

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

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

Tjandra, N., Omichinski, J., Gronenborn, A., Clore, G. and Bax, A. (1997b) *Nat. Struct. Biol.*, **4**, 732–738.

Tolman, J., Flanagan, J., Kennedy, M. and Prestegard, J. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 9279–83.

Venters, R. A., Metzler, W. J., Spicer, L. D., Mueller, L. and Farmer, B. T. (1995) *J. Am. Chem. Soc.*, **117**, 9592–9593.

Vercauteren, J. and Rutledge, D. (1996) *Food Chem.*, **57**, 441–450.

Vogtherr, M. and Peters, P. (2000) *J. Am. Chem. Soc.*, **122**, 6093–6099.

Wang, Y., Jacob, J., Cordier, F., Wingfield, P., Stahl, S., Lee-Huang, S., Torchia, D., Grzesiek, S. and Bax, A. (1999) *J. Biomol. NMR*, **14**, 181–184.

Williamson, M. and Asakura, T. (1997) *Methods. Mol. Biol.*, **60**, 53–69.

Wishart, D. S., Bigam, C. G., Yao, J., Abildgaard, F., Dyson, H. J., Oldfield, E., Markley, J. L. and Sykes, B. D. (1995) *J. Biomol. NMR*, **6**, 135–140.

Wold, S., Esbensen, K. and Geladi, P. (1987) *Chemometrics Intell. Labs. Syst.*, 735–743.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. Wiley, New York.

Xiaoquan, L., Hongde, L., Zhonghua, X. and Qiang, Z. (2004) *J. Chem. Inf. Comput. Sci.*, **44**, 1228–1237.

Zuiderweg, E. (2002) *Biochemistry*, **41**, 1–7.