Accurate proton-proton distance calculation and error estimation from NMR data for automated protein structure determination in AUREMOL

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

vorgelegt von

Jochen Markus Trenner aus Regensburg

2/2006

Promotionsgesuch eingereicht am:	30.01.2006
Die Arbeit wurde angeleitet von:	Prof. Dr. Dr. Hans Robert Kalbitzer

Prüfungsausschuß

Vorsitzender:	Prof. Dr. Günter Hauska
Erstgutachter:	Prof. Dr. Dr. Hans Robert Kalbitzer
Zweitgutachter:	Prof. Dr. Eike Brunner
Drittprüfer:	Prof. Dr. Reinhard Sterner

Summary

In this dissertation the development of the NMR spectrum analysis program AUREMOL is continued. The goal of AUREMOL is to provide routines for an automatic protein structure determination from a minimum of experimental NMR data with a minimum of user intervention. General improvements are the establishment of a IUPAC compliant nomenclature for atom names in the different AUREMOL modules, the addition of a new strips tool that aids the user during a manual sequential assignment task and a correct implementation of the support for arbitrary motional models and finite relaxation delay in the spectrum simulation module RELAX. The present threshold based peak picking routine is enhanced to adapt the threshold locally based on a local noise estimate. This procedure leads to significantly less artifacts in the produced signal list. In addition the new adaptive routine has been combined with the Bayesian signal analysis in AUREMOL and the resulting automatically determined signal list for an experimental 2D NOESY spectrum shows very good artifact suppression.

The segmentation based integration routine in AUREMOL is extended to produce an error estimate for every signal integral by calculating a noise and overlap contribution. Testing the proposed integration error estimation approach on a simulated noised dataset reveals that the error estimates of 52% of the integrated signals explain the originally simulated data. For the remaining signals the errors are underestimated because of undetected overlap, where two or more overlapping signals do not have distinct extrema.

As the simulation based *de novo* assignment routine in AUREMOL shows poor results when no *a priori* partial assignment is present, the pseudo energy function used in the threshold accepting optimization is supplemented with a second contribution describing the global matching quality. This results in a performance increase from 10% correctly assigned signals using the original pseudo energy function to 100% correctly assigned signals when using the additional term on comparable synthetic datasets of HPr H15A (*S. carnosus*) and HPr WT (*S. aureus*) starting without *a priori* partial assignment.

In the main part of this thesis a new distance calculation module called REFINE is presented that is based on the relaxation matrix based spectrum simualtion in RELAX. The calculated distance restraints are used for structure calculation in a molecular dynamics calculation. A starting relaxation matrix calculated from a trial structure is iteratively refined to best describe the experimental NOE data. During the process, a distance error estimate is produced from the errors in the experimental peak integrals. That the algorithm basically works is proven on a perfect synthetic dataset of the 66 amino acid protein TmCSP (*Thermotoga maritima*), where 100% correct distances are obtained for the original structure used for the simulation as input for REFINE. For an extended strand as input 86.9% of the calculated distances exhibit an error below 20%, so in both cases REFINE yields better results than the approximately 70% resulting distances with an error below 20% from the ISPA approach. The robustness against noise in the dataset is evaluated on a newly integrated artificial dataset of the 88 amino acid protein HPr (*S. carnosus*) with varying levels of additive Gaussian noise. The resulting distances from REFINE show a less pronounced, linear dependence on the noise level compared to ISPA. Between 5%-16% more distances with an error below 20% are calculated by REFINE using an extended strand and for the correctly folded structure as input the advantage for REFINE rises to 25%-37% compared to the ISPA results.

For the application of REFINE to unassigned NOESY spectra, a combination of the NOESY assignment module KNOWNOE and REFINE is proposed and the user defined KNOWNOE scaling parameter is calculated using REFINE. By the alternating application of KNOWNOE assignment and REFINE distance calculation, a protocol for the automatic structure determination from unassigned NOESY spectra is developed. For an experimental spectrum of HPr (*S. carnosus*) it is shown that this procedure is capable of determining the correct fold after nine iterations, additionally using predicted angle restraints. This qualitatively proves that this approach is valid. In a quantitative analysis the application of the same protocol to experimental 2D and 3D data of the 86 amino acid protein RalGDS-RBD yields structures of comparable quality (R-factor: 0.322) to the published X-ray structure (R-factor: 0.325). Here as additional restraints predicted dihedral angles and experimental H-bonds are used.

To include simulation/modelling errors in the distance error estimate of REFINE, a parameter variation approach is evaluated. From the repeated application of REFINE to an experimental 2D dataset of HPr (*S. aureus*) with 25% known assignment distance distributions are obtained through the normally distributed variation of the experimental peak integrals as well as the backbone and sidechain order parameters. The standard deviations of these distributions are interpreted as distance error bounds of the restraints. Initially an extended strand is used as trial structure in REFINE. The structure calculated from these restraints is then used as trial structure for a second REFINE run. Compared to the structures obtained using an automatic ISPA approach (R-factor: 0.37), the structures calculated from both REFINE restraint sets show better R-factors (first run: 0.35, second run: 0.33).

Table of contents

1. Introduction

The term ‘proteomics’ has been coined in analogy to the genomics projects of recent years [1]. It reflects the move away from the mere identification of all the genes present in a given organism in genomics towards the elucidation of the function of all the related proteins in proteomics. Knowledge about proteins and their function that is closely related to their 3D structure is steadily gaining interest in scientific research, especially in biology, medicine and pharmacy, but also e.g. in the direction of materials sciences when considering interesting protein based materials like spider silk [2;3].

In structural proteomics the goal is to determine and characterize protein structures. Next to *X-ray crystallography*, where diffraction patterns of protein crystals are analyzed to obtain structural information, sophisticated *Nuclear Magnetic Resonance* – NMR – techniques have been developed that allow protein structure determination [4] of proteins in solution under near-physiological conditions.

1.1 Proteins

In the genome of every organism, the genes encode a large number of essential macromolecules – proteins – as a series of base pairs [5]. In the cell the gene sequences are translated into amino acid chains, where the amino acids are connected via peptide bonds:

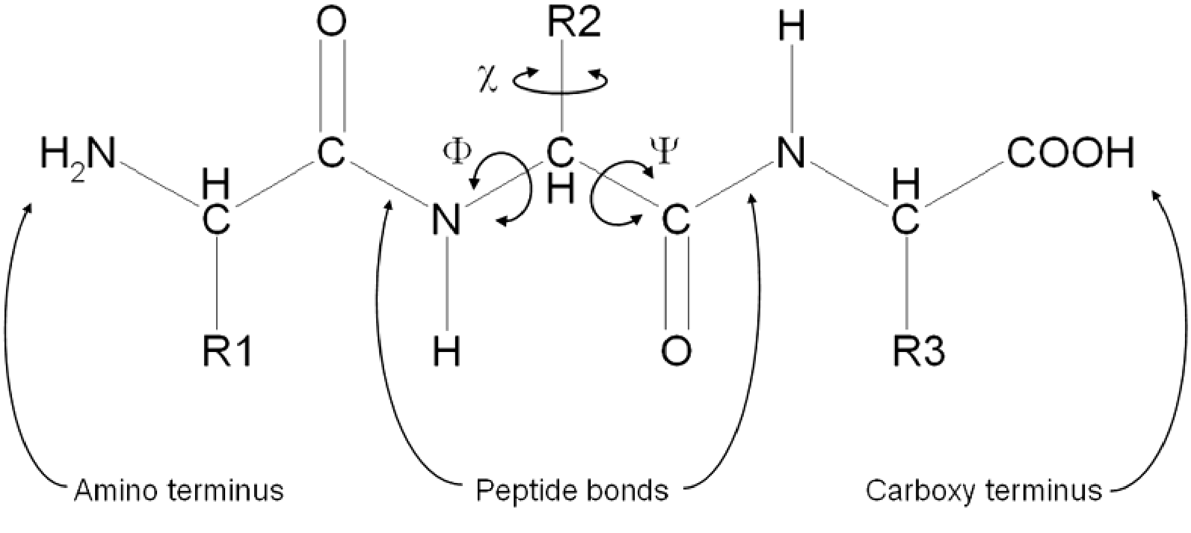

Figure 1.1 Three amino acids forming a tri-peptide

The resulting peptide or protein strands then usually adopt a specific three-dimensional shape, the so called tertiary structure (Fig. 1.2), that is essential for their functioning by rotating the chain segments around the backbone torsion angles φ and ψ. Common sub-motifs in folded structures include the secondary structure α–helices and β–sheets that are defined through characteristic dihedral φ/ψ-angle combinations and stabilized by hydrogen bonds. To get a more vivid impression of the three dimensional arrangement of a protein backbone a *ribbon display* is often used, where the secondary structure elements are highlighted:

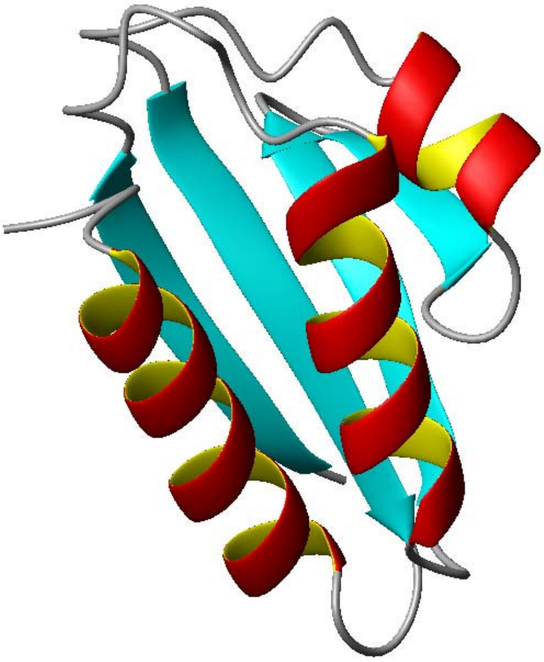

Figure 1.2 Ribbon plot of HPr from *S. carnosus* [6] showing secondary structure elements, *α*-helices (red/yellow) and *β*-sheets (cyan)[*)].

Exactly how protein folding is achieved in the cell is still subject to investigation, the process can differ from protein to protein. Some fold all by themselves, driven mostly by entropic effects [8-10] when establishing the hydrophobic core, others need helper molecules, *chaperones*, to acquire the correct structure [11]. Folding of smaller proteins takes place on a micro- to millisecond timescale [12] but can take significantly more time, especially for larger and more complicated proteins where the formation of different domains and the interaction with chaperones comes into play.

In, around and outside a cell, proteins fulfill various tasks. E.g. ligand proteins can exhibit regulatory and signal transduction functions when they dock on to corresponding receptor

[*)] All 3D molecule plots in this thesis were prepared with MOLMOL [7].

proteins, effector proteins influence numerous metabolic pathways, membrane proteins stabilize cell walls, etc [13].

Errors in a protein's fold can result in partial or complete loss of its functionality that often leads to pathological symptoms. Prominent examples for diseases related to protein misfolding and subsequent build up of plaques are BSE, Creutzfeld-Jakob and Alzheimer's disease [14-18].

Against this background it is clear that the interest of medical and pharmaceutical sciences in protein structures reflects the need to obtain as much information about the onset of protein related diseases as possible, since by the availability of this knowledge the development of practical cures can be greatly enhanced.

1.2 NMR in protein structure determination

In protein NMR, the sample is exposed to a strong static magnetic field $\boldsymbol{B_0}$ in the z-direction. The protons exhibit a spin $\boldsymbol{I}$ of ½ and carry a magnetic dipole moment given by $\mu = \gamma\hbar \cdot \boldsymbol{I}$[*)].

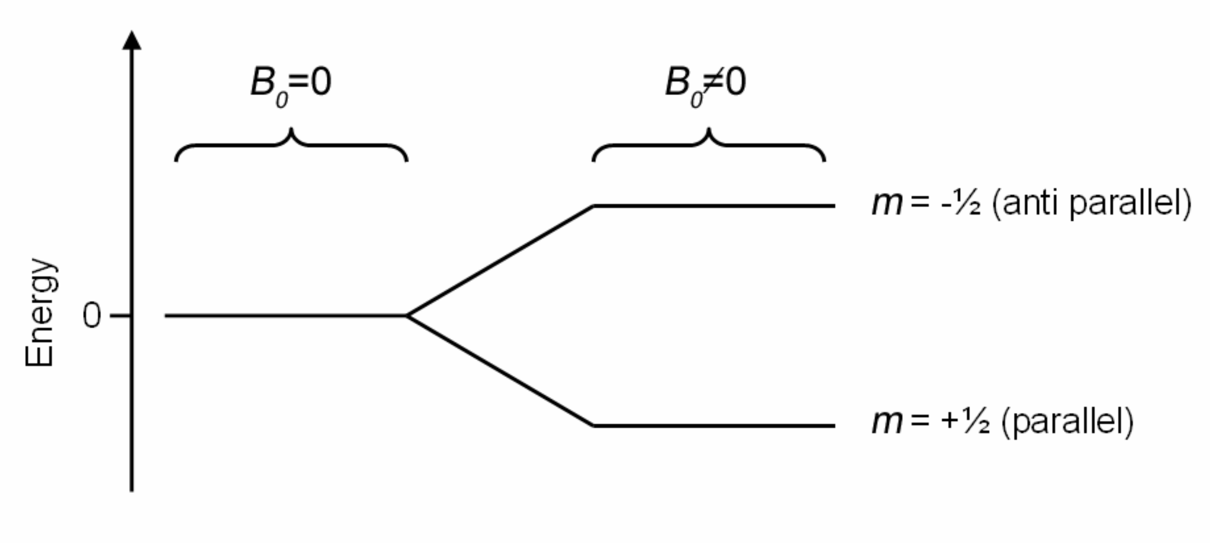

Figure 1.3 Energy levels of a nucleus of spin quantum number $\boldsymbol{I}$ = ½ in a magnetic field $\boldsymbol{B_0}$.

Coupling to the field leads to a splitting of the energy into two energy levels $E_{\pm 1/2} = -\gamma\hbar \boldsymbol{I_z B_0}$ for the spin states *±½* and the population of the states $N_{\pm 1/2}$ in thermal equilibrium follows the *Boltzmann statistics*:

[*)] γ: Gyromagnetic ratio, $\hbar = h/2\pi$: Reduced Planck constant

$$\frac{N_{-1/2}}{N_{+1/2}} = \exp(-\frac{\gamma\hbar B_o}{k_B T}) = \exp(-\frac{\Delta E}{k_B T}). \quad (1.1)$$

Here *ΔE* denotes the energy difference of the two states, *T* the temperature and k_B the *Boltzmann* constant.

Under the influence of $\boldsymbol{B_0}$ the proton spins precess around the z-direction with the *Larmor frequency*

$$\nu_0 = \frac{\gamma}{2\pi} B_0 = \frac{\Delta E}{h}. \quad (1.2)$$

In a protein, the different proton spins experience a slightly different external field due to the shielding effect of the electrons in their chemical environment. Because of that, also the Larmor frequencies vary slightly, which is known as *chemical shift*.

By radiating a hf-pulse at the resonance frequency ν_0, spin flips can be induced. During relaxation back to the equilibrium, radiation – the sum of the damped oscillations of all resonating spins – is emitted. This *Free Induction Decay* – FID – is recorded during t_2 in NMR experiments.

The central NMR experiment in protein structure determination is the *Nuclear Overhauser Effect Spectroscopy* – NOESY. It allows to measure signals that are related to the distances between neighboring protons, since in this experiment magnetization is transferred between spatially close spins during the mixing time τ_m:

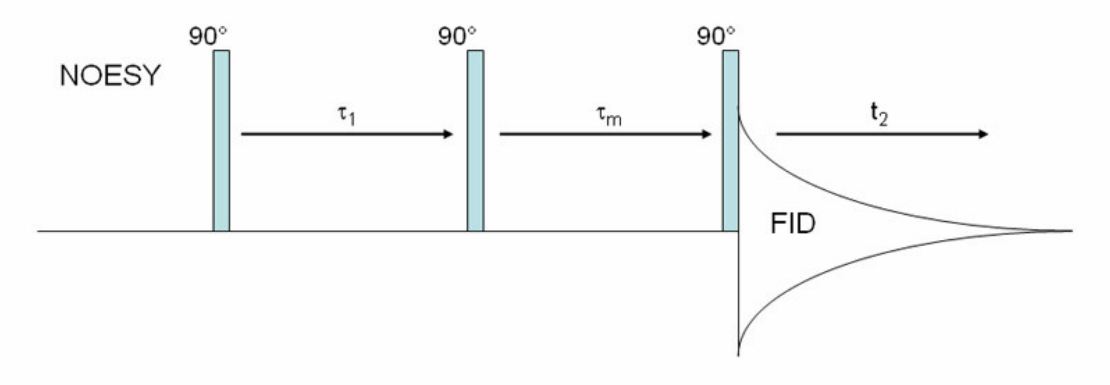

Figure 1.4 Pulse sequence for a 2D NOESY experiment with evolution time τ_1 and mixing time τ_m. After the final pulse, the FID is acquired during t_2.

In a 2D H/H-NOESY experiment, after an initial 90° pulse the net magnetization can evolve in the x/y plane for an evolution time τ_1. After τ_1 and another 90° pulse the mixing time τ_m

allows magnetization transfer through space via dipolar interaction until the final 90° pulse rotates the net z-magnetization to the x/y plane for detection and recording of *n* datapoints. Iterating this pulse sequence and at the same time incrementing τ_1 in steps of $\tau_{1,inc}$ for *m* times results in a 2D dataset in the indirect dimension of t_1 (the time increments) and the direct dimension of t_2 (the recording time) for this experiment. According to the *Nyquist theorem*, in order to be able to correctly sample a signal of frequency f_{max}, the sampling frequency f_s has to be at least two times f_{max}. So the sampling interval $\tau_{1,inc}$ is given by

$$\tau_{1,inc} = \frac{1}{f_s} = \frac{1}{2 \cdot f_{max}} \tag{1.3}$$

When using quadrature detection, the increments $\tau_{1,inc}$ can be calculated for the desired spectral width *SW* in [Hz]:

$$\tau_{1,inc} = \frac{1}{SW} \tag{1.4}$$

Higher dimensional NOESY experiments like 3D or 4D-NOESY-HSQC (*Heteronuclear Single Quantum Coherence* – HSQC) require additional pulses and evolution periods and also isotope labelled protein samples containing ^{13}C and/or ^{15}N nuclei with spin ½. The increased expenditure of time and money for these experiments is rewarded by considerably less signal overlap in the resulting higher dimensional spectral datasets.

For evaluation the time domain data is *Fourier transformed* to the frequency domain. Further processing includes phase correction, baseline correction, filtering, etc [4], before analysis is carried out.

1.3 Automated protein structure determination from NMR data

Several approaches to the problem of automated protein structure determination from NMR data exist. They can be divided into classic bottom-up strategies that make use of a large base of experimental data and top-down approaches that try to minimize the experimental effort and concentrate on the structure part.

For the bottom-up class of programs a strong focus is put on obtaining a complete sequential assignment. To this end a vast array of assignment programs applicable to different kinds of multi-dimensional spectra has been developed. Obviously the inherent drawback in this approach is the experimental effort involved in obtaining the required spectra.

In a top-down approach the focus shifts to structure itself. As a starting point a homology modelled structure can be used or in the extreme case even an extended strand. Using this structure and as much additional information as possible, like predicted chemical shifts and backbone torsion angles, the assignment and the structure itself are refined in an iterative process. This is the concept of AUREMOL described below.

The current state of the available methods for automatic NMR structure determination is discussed in section two, especially for those parts also of relevance to the top-down approach maintained in this thesis. All developed extensions and improvements to AUREMOL are presented together with the results in section three and discussed in section four.

1.4 AUREMOL

In 1999 a cooperation between NMR spectrometer manufacturer Bruker BioSpin and the Biophysics Department of the University of Regensburg was started to develop a software package – AUREMOL [19] – based on AURELIA [20] for the automated determination of protein structures from NMR data. Bruker BioSpin provides the Amix™ Viewer as a framework and the routines required for analysis, evaluation and automation are contributed by the Biophysics Department.

The concept for automation is a molecule centered top-down approach where all of the available *a priori* information is used to eliminate as many free parameters as possible and reduce the amount of experimental data to a minimum. Along the way, any need for expert intervention is also to be reduced to a minimum.

The goal is to deliver an extensive tool for protein structure determination from NMR data, that speeds up and automates this process as outlined in the following scheme:

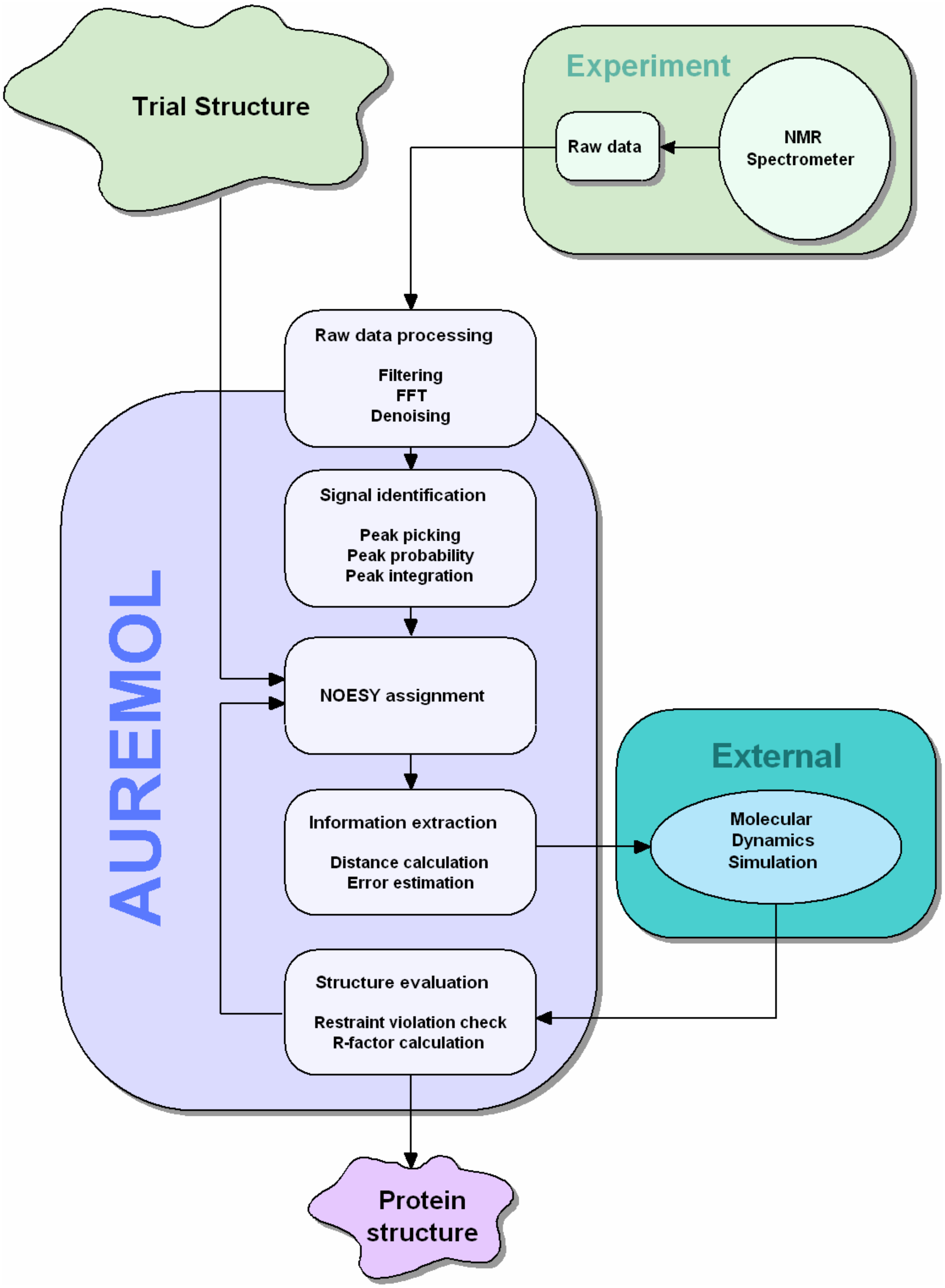

Figure 1.5 Typical structure determination workflow using AUREMOL.

AUREMOL currently relies on external software only for raw data processing and restrained molecular dynamics calculations. It offers routines and utilities for the data analysis and structure validation steps in one NMR program with an emphasis on NOESY evaluation.

Advanced modules of AUREMOL include the spectrum simulation module RELAX [21-23], the NOESY assignment tool KNOWNOE [24], the structure evaluation module RFAC [25] and as the main part of this thesis the distance calculation module REFINE.

Trial versions can be obtained from the AUREMOL website and require a trial license from Bruker BioSpin[*)].

1.5 Goals of this study

Since it is desirable to solve protein structures quickly, automation is essential as the manual approach is very time consuming. In protein structure determination from NMR data using a top-down approach as sketched in Fig. 1.5, the individual steps are not *a priori* suited for automation. So the aim of this thesis is first to find starting points for the development of a largely automated structure determination procedure among the currently available routines in AUREMOL. The second goal is to implement the necessary changes in a way that the results produced using the now integrated automatic apporach yield an increase in overall quality and efficiency as well as reproducibility, a very important factor in automation.

General improvements to AUREMOL are the complete migration to a consistent nomenclature throughout the program as described in section 3.1 and the incorporation of finite relaxation delays in the spectrum simulation presented in section 3.2.

Concerning signal identification, the current implementations of peak picking and peak integration are reviewed in sections 2.1 and 2.2. The available peak picking routine in AUREMOL is extended to be usable for an automated approach as described in section 3.3 and in section 3.4 the integration module is complemented by a simultaneous error estimation to extract the maximum of information included in the experimental data.

In signal assignment, the main problem is how to obtain a correct sequential assignment with minimal effort. The current state of affairs is presented in section 2.3 and two possible routes for automation are encountered: The strip route uses heteronuclear 3D experiments to obtain the sequential assignment. For this approach an interactive strip tool has been implemented in AUREMOL that allows the manual as well as automatic definition of spectrum strips as

[*)] AUREMOL website: www.auremol.de
For licensing information please contact the Bruker BioSpin GmbH license department: license@bruker.de

described in section 3.5. Using the strip information, the sequential assignment can then be determined semi-automatically in AUREMOL. The alternative route uses spectrum simulation and pattern matching for a *de novo* assignment. Since the available AUREMOL routine showed only limited success, the existing pseudo energy function used in the optimization was supplemented with a global pseudo energy term as shown in section 3.6, so that the best match can be reached more easily.

The central point of this thesis is the accurate extraction of distance information from the experimental data. To this end the relaxation matrix formalism is used for distance calculation. It requires a sequential assignment and is the basis for the spectrum simulation used in the iterative REFINE algorithm presented in section 3.7. Errors in the experimental data are automatically accounted for and together with an optional parameter variation a distance error estimate is produced. In the evaluation section 3.8 the performance of REFINE is first evaluated for a simulated dataset. Then, in combination with KNOWNOE, the capabilities for automated structure determination are evaluated on real world data.

By eliminating the scaling factor required by the NOE assignment module KNOWNOE using REFINE, a useful combination of these two modules regarding automated structure determination is created as outlined in section 3.9.

In section four, the success of the AUREMOL improvements developed for this thesis is discussed in respect to the potential for automation, the REFINE performance and the error estimation benefits.

2. Current state of the field

In this section an overview of the available methods used in protein structure determination from NMR is presented. References to existing implementations are given, for a near exhaustive list refer to [19].

2.1 Peak picking

Peak picking in NMR is the task of identifying the location of potential signals in a spectral dataset. The most basic variant of peak picking is the manual identification of peak maxima. Here, the greatest advantage is at the same time the greatest disadvantage: An expert is needed. So although the results may be very good, the process takes a lot of time and is biased.

For automation three main approaches exist: Thresholding, shape based and Bayesian signal identification. In the widely used threshold based peak picking approach [26-30] possible signals are identified by analyzing the intensity distribution in a spectrum. Above and/or below a threshold intensity level all local maxima and/or minima in the dataset are considered as signals. Having an expert manually choose an appropriate intensity is the crucial part here since too low values result in large numbers of noise spikes being added to the list of potential signals whereas too high values neglect weak signals:

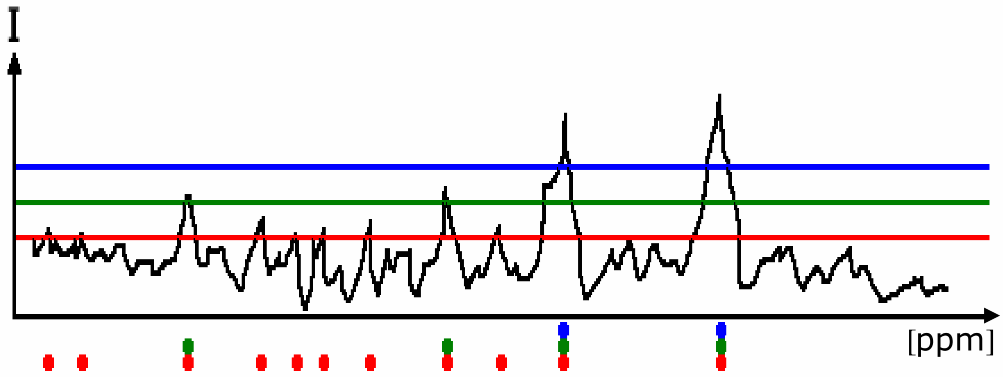

Figure 2.1 Schematic example of peak picking results for different thresholds (only positive signals considered). A high threshold (blue line) yields only the two strongest signals, medium (green) four signals and low (red) eleven signals

Advantages of this method are its speed and ease of use, disadvantages the free threshold parameter and the assumption of a negligible baseline distortion, noise and artifacts. In the program ATNOS [31] thresholding in NOESY spectra is supplemented by considering the local baseline and noise as well as symmetries, chemical shift and structure information to obtain better results.

Incorporating line shape information can also improve the peak picking performance, especially the discrimination between signal and artifact [32]. In STELLA [33] a database containing user defined signal and noise/artifact peak shapes is used. Here the results are of course dependent on the reference sets. CAPP [34] uses fitting of ellipses at different levels of a peak for signal identification in multidimensional spectra. Local peak shape symmetries are used in AUTOPSY [35] together with the shape information of well resolved cross peaks and local noise levels for peak picking. Still, for these approaches spectral regions of strong overlap remain a problem.

Bayes driven analysis of the peak data [36;37] can also be used for signal discrimination. The algorithm is trained on reference sets for noise and signal peaks where distinguishing local features are stored and then applied to the experimental spectrum. Additionally global features like symmetries can be considered. The success of this method depends mainly on the choice of the training peaks, that are usually selected manually.

2.2 Peak integration

The most suitable approaches to peak integration for a fast and automatic workflow are peak fitting and segmentation algorithms.

Fitting user defined reference peaks to the experimental spectrum [30;38] can yield very good results even for overlapping signals, however it is dependent on the choice of the reference peaks.

In the segmentation approach the peak coordinates returned from the peak picking routine serve as seeds for a region growing algorithm [39] that calculates the signal integral. Around each seed an integration box depending on the expected linewidth is created. All local maxima inside this box are considered as additional seeds. If the neighboring datapoints of the main and additional seeds fulfill the growing conditions, they become new seeds and are added to the peak integral if they belong to the main peak. After iterating until the growing conditions cannot be met anymore, or only seeds are left, the integral is calculated.

Both presented approaches are subject to errors because of baseline distortions, overlap and noise as well as limited resolution.

2.3 Peak assignment

Manually assigning the peaks in a spectrum to the corresponding interacting spins in the sample protein is one of the most time consuming tasks in NMR structure determination. A set of appropriate 2D and 3D NMR experiments is usually required to complete this sequential assignment. Manual tools for the alignment of spectrum *strips* (see below) from 3D triple resonance experiments exist and speed up this process.

Programs that allow manual or semi-automatic sequential assignment include ANSIG [40;41], AURELIA/AUREMOL, EASY/XEASY [30], FELIX [42], GIFA [43;44], PIPP [34], SPARKY [45].

Automated bottom-up approaches to sequential assignment like AUTOASSIGN [46], CONTRAST [47], GARANT [48] or PASTA [49] rely on triple resonance data that is first consistency checked, then the signals are grouped to spin systems and identified using statistical analyses of the chemical shifts. After that fragments of sequentially neigboring residues are determined which are finally linked to the primary sequence. For these steps deterministic methods like exhaustive searching or optimization methods like simulated annealing or threshold accepting are employed. The results naturally depend on the quality of the input data.

For the automated assignment of NOESY spectra also several methods have been proposed. The main problem here lies in the occurrence of assignment ambiguities due to massive signal overlap. SANE [50] presents a semi-automatic approach relying on user interaction during the analysis, ARIA [51;52], CANDID [53] and NOAH are automatic approaches. ARIA uses molecular dynamics directly with ambiguous distance restraints. NOAH uses distance geometry calculations together with restraint violation analysis and CANDID combines methods from ARIA and NOAH with molecular dynamics and network anchoring. In the approaches above a sequential assignment, the quality of which strongly influences the resulting NOE assignment, is required. CLOUDS [54;55] is an example of a NOESY assignment approach that does not need a sequential assignment.

In the following, the state of the different available assignment routines in AUREMOL is presented.

2.3.1 Manual assignment using strips

In an ideal case, a single 3D heteronuclear *triple resonance* HNCA spectrum would be sufficient to determine the sequential assignment of a protein, i.e. the assignment of a HN chemical shift value to a certain residue in the protein sequence. Cross signals between the alpha carbon atom and the amide proton of residue *i*, CA_i / HN_i, and also between the sequentially preceding alpha carbon atom *i*-1 and the amide proton of residue *i*, CA_{i-1} / HN_i, will show in a narrow spectral region around a given HN_i resonance, a so-called *strip*. Using this information from all HN resonances, the sequential assignment can be deduced. However, limited resolution and signal overlap can lead to dead ends and ambiguities during the assignment process, so usually different spectra are required to obtain the complete sequential assignment of a protein. The most commonly used triple resonance experiments for sequential assignment are listed in the following table:

Experiment	Observed resonances
HNCA	CA i, CA i-1
HN(CO)CA	CA i-1
CBCANH	CA i, CA i-1, CB i, CB i-1
CBCA(CO)NH	CA i-1, CB i-1

Table 2.1 Useful triple resonance NMR experiments for the determination of the sequential assignment. At a given HN resonance signals from the current amino acid i and/or the amino acid preceding in the sequence i-1 are detected.

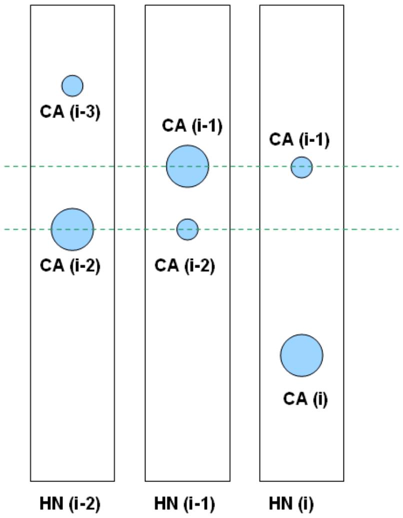

Figure 2.2 The HNCA strips are sorted according to this scheme to obtain the sequential assignment.

By sorting the spectrum strips, the sequential assignment can be obtained.

2.3.2 Automated sequential assignment using strips

In AUREMOL, a list of pseudo residues with the experimental chemical shift information is the starting point for a Monte Carlo approach to automatic sequential assignment. Building random sequences and evaluating a pseudo energy derived from the chemical shift match between neighboring residues and random coil chemical shift values allows to optimize the sequence to best fit the experimental data. For this a variant of a threshold accepting algorithm with additional bouncing [56] is used. However, the input data still has to be prepared manually.

2.3.3 Automated sequential assignment using NOE spectra and spectrum simulation

This method is currently being integrated into AUREMOL and has been outlined in a previous PhD thesis [57]. It requires a trial structure of the protein under investigation and uses among other things the line shape information of simulated and experimental signals to determine the peak assignment in a simulated annealing like procedure by varying the chemical shift assignments. Current changes include an extended pseudo energy function and restrained chemical shift variation depending on the respective proton, as described in section 3.6.

As this method is only limited by the available types of simulation it is a fairly universal approach to the assignment problem and shold be applicable to various kinds of NMR spectra.

2.3.4 Automated NOE assignment using grid search

In case a NOE assignment of a given spectrum is known, it can be transferred to other NOE spectra of the same protein that have been recorded under different experimental conditions. This is especially useful for series of experiments with e.g. varying temperature or pH, where for each step signals can slightly shift or for the adaption of averaged shifts obtained from a number of different spectra to one specific spectrum.

For appliances like these, the PEAK ASSIGN module in AUREMOL uses a grid search algorithm that maps the known signal assignments to the eventually shifted signals in the spectra from the test series [58].

2.3.5 Probability driven automatic NOE assignment using Bayes' Theorem

The goal of KNOWNOE [59] is to calculate the most probable solution for ambiguous assignments by taking statistical information from known structures and the individual NOE volumes into account. To this end, starting from known protein structures determined by X-ray diffraction or NMR, local databases containing distance probability distributions for protons between different combinations of amino acids have been created.

This information is then used together with the sequential assignment information in a conditional probability approach to determine the most probable assignment of a given NOE signal. From a set of 326 NMR protein structures statistical tables for all assignment probabilities relevant to interproton NOE cross peaks were derived in the form of *Volume Probability Distributions* – VPDs in the current version. A known signal integral and distance have to be specified for calibration. The probability for a given peak with volume V_0 to belong to assignment class C_i is calculated using Bayes' theorem:

$$P(C_i \mid V_0) = \frac{P(C_i)P(V_0 \mid C_i)}{\sum_i P(C_i)P(V_0 \mid C_i)} \qquad (2.1)$$

In an iterative process the most probable assignments are determined. Additionally, mutual information concerning other NOE contacts of a certain proton spin is considered in a way similar to the network anchoring proposed by [53].

2.4 Relaxation matrix analysis

Describing the magnetization transfer during the mixing time of NOESY spectra mathematically is the aim of relaxation matrix analysis [60-62] used in RELAX [21-23] and other programs for spectrum simulation including BIRDER [63] and CORMA [64] (2D), SPIRIT [65] (3D NOESY-HSQC) and RELAX (2D & 3D) which mainly differ in the number of available motional models and consideration of finite relaxation delay, anisotropy and chemical exchange effects. As a starting point the time dependent Hamiltonian for the dipolar interaction of a system with two spins *i*, *j* is considered:

$$\hat{\mathbf{H}}_D(t) = -\gamma_1\gamma_2 \frac{\hbar\mu_0}{4\pi} \cdot \sqrt{\frac{48\pi}{10}} \sum_{m=-2}^{2} \frac{\mathrm{Y}_{2m}^{*}(\Phi_{ij}^{B_0}(t))}{r^3(t)} \cdot \hat{\mathbf{T}}_m(\hat{\mathbf{I}}, \hat{\mathbf{J}}) \qquad (2.2)$$

with the gyromagnetic ratios $\gamma_{1/2}$ of the participating nuclei, the induction constant μ_0 and the spherical harmonic function Y_{2m}^* depending on the steradian $\Phi_{ij}^{B_0}$ of the spin *i* and spin *j* connecting vector and the external magnetic field B_0. The tensor operators $\hat{T}_m$ induce *m*-quantum transitions via the spin operators $\hat{\boldsymbol{I}}$ and $\hat{\boldsymbol{J}}$ of the investigated spins,

$$\begin{aligned}\hat{\mathbf{T}}_0 &= \sqrt{\frac{1}{6}}\cdot(3\hat{\mathbf{I}}_z\hat{\mathbf{J}}_z - \hat{\mathbf{I}}\hat{\mathbf{J}}),\\ \hat{\mathbf{T}}_{\pm1} &= \mp\frac{1}{2}\cdot(\hat{\mathbf{I}}_z\hat{\mathbf{J}}_\pm + \hat{\mathbf{I}}_z\hat{\mathbf{J}}_\mp),\\ \hat{\mathbf{T}}_{\pm2} &= \frac{1}{2}\cdot\hat{\mathbf{I}}_\pm\hat{\mathbf{J}}_\pm, \end{aligned} \tag{2.3}$$

and the creation/annihilation operators $\hat{\boldsymbol{I}}_\pm$ and $\hat{\boldsymbol{J}}_\pm$ are defined as:

$$\hat{\mathbf{X}}_\pm := \hat{\mathbf{X}}_x \pm i\hat{\mathbf{X}}_y, \quad \hat{\mathbf{X}} \in \{\hat{\mathbf{I}}, \hat{\mathbf{J}}\} \tag{2.4}$$

Looking at the associated energy scheme, one can see the different possible transitions:

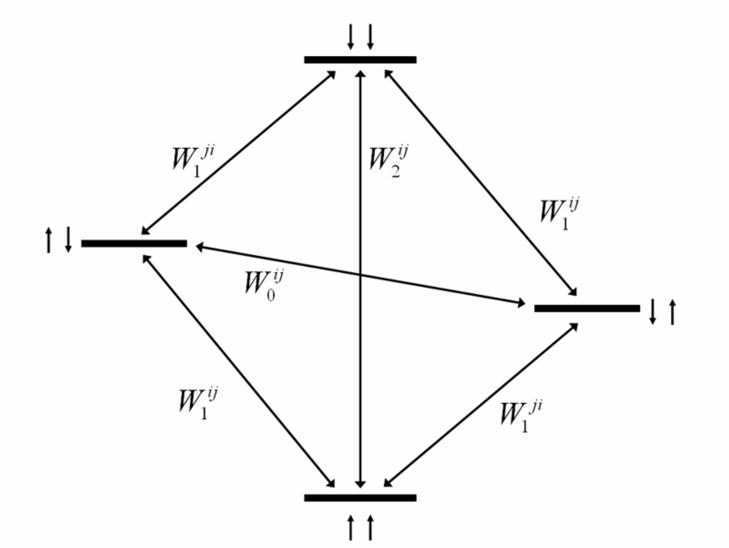

Figure 2.3 Transition probabilities W_n^{ij} in a two spin ½ system with spins i and j

Counting the participating transitions, the changes in longitudinal magnetization differences of *N* spins over time are described by the Solomon equations [60], with ***ΔM*** the *N×1* column vector containing each magnetization difference $M_{zi}(t)$-M_0 in respect to the equilibrium magnetization M_0 (chemical exchange is neglected):

$$\frac{d}{dt}\mathbf{\Delta M}(t) = -\mathbf{R}\cdot\mathbf{\Delta M}(t) \tag{2.5}$$

R denotes the *N*×*N* relaxation matrix. Grouping magnetically equivalent spins and introducing the diagonal matrix ***N*** containing the numbers n_i of group members so that ***R*** = ***NR***´, the solution is

$$\mathbf{\Delta M}(t) = \mathbf{\Delta M}(0)\cdot\exp(-t\cdot\mathbf{N}\cdot\mathbf{R}'). \tag{2.6}$$

The NOE cross peak integrals are then given by

$$A_{ij}(t) = \alpha\cdot M_{z,j}(0)\cdot\left[\exp(-t\cdot\mathbf{N}\cdot\mathbf{R}')\right]_{ij} \tag{2.7}$$

with an arbitrary scaling factor α and $M_{z,j}(0) = n_j/\alpha$ for fully relaxed spectra, hardly ever recorded. In reality the magnetization at the beginning of the mixing time depends on the recovery time t_r between the FID recordings and the next pulse sequence:

$$M_{z,j}(t_r) = \frac{1}{\alpha}\cdot\sum_k\left[\mathbf{1}-\exp(-t_r\cdot\mathbf{N}\cdot\mathbf{R}')\right]_{jk}\cdot n_k, \tag{2.8}$$

leading to the final NOE volume matrix

$$A_{ij}(t_r,t) = \left[\exp(-t\cdot\mathbf{N}\cdot\mathbf{R}')\right]_{ij}\cdot\sum_k\left[\mathbf{1}-\exp(-t_\mathrm{r}\cdot\mathbf{N}\cdot\mathbf{R}')\right]_{jk}\cdot n_k \tag{2.9}$$

for not fully relaxed spectra.

Using the transition probabilities[*)] [21;66]:

$$W_0^{ij} = q\cdot \mathrm{J}_{ij}^0(\omega_i-\omega_j), \tag{2.10}$$
$$W_1^{ij} = \frac{3}{2}q\cdot \mathrm{J}_{ij}^1(\omega_i),$$
$$W_2^{ij} = 6q\cdot \mathrm{J}_{ij}^2(\omega_i+\omega_j), \tag{2.11}$$

the auto and cross relaxation rates of the symmetrized relaxation matrix ***R***´can be calculated:

[*)] q denotes the dipolar interaction constant, $\mathrm{q} = \frac{2\pi}{5}\gamma_i^2\gamma_j^2\hbar^2\left(\frac{\mu_0}{4\pi}\right)^2 \approx 7.1584\cdot10^{-69}\mathrm{m}^6\mathrm{s}^{-2}$

$$R'_{ii} = q \cdot (1 - \frac{1}{n_i}) \cdot \left[3\,\mathrm{J}^1_{ii}(\omega) + 12\,\mathrm{J}^2_{ii}(2\omega)\right] + q \sum_{j \neq i} \frac{n_j}{n_i} \cdot \left[\mathrm{J}^0_{ij}(0) + 3\,\mathrm{J}^1_{ij}(\omega) + 6\,\mathrm{J}^2_{ij}(2\omega)\right],$$

$$R'_{ij} = q \cdot \left[6\,\mathrm{J}^2_{ij}(2\omega) - \mathrm{J}^0_{ij}(0)\right] \tag{2.12}$$

The spectral density functions J^n_{ij} model intramolecular motions of the spins as well as the overall motion of the molecule. They have the general form

$$\mathrm{J}^n_{ij}(\omega) = \int_0^\infty \frac{\mathrm{Y}_{2n}(\Phi^{B_0}_{ij}(t)) \cdot \mathrm{Y}^*_{2n}(\Phi^{B_0}_{ij}(t+\tau))}{r^3_{ij}(t) \cdot r^3_{ij}(t+\tau)} \cdot e^{-i\omega\tau} d\tau \tag{2.13}$$

which represents the Fourier transformed time correlation function of the relative motion of the spins *i* and *j*. By choosing appropriate spectral densities for different parts of the molecule, the rates can be calculated taking into account specific molecular properties to increase the accuracy.

As an example, the following spectral density function describes a rigid molecule, underlying *Brownian* motion and tumbling isotropically in solution with a correlation time τ_c:

$$\mathrm{J}^n_{ij}(\omega) = \frac{1}{4\pi r^6_{ij}} \cdot \frac{\tau_c}{1 + \omega^2 \tau_c^2}. \tag{2.14}$$

Since the assumption of a rigid protein is not realistic, an alternative is offered by the model-free approach of Lipari and Szabo [67]:

$$\mathrm{J}^n_{ij}(\omega) = \frac{1}{4\pi r^6_{ij}} \cdot \left(\frac{S^2 \cdot \tau_c}{1 + \omega^2 \tau_c^2} + \frac{(1 - S^2) \cdot \tau}{1 + \omega^2 \tau^2} \right), \quad \frac{1}{\tau} = \frac{1}{\tau_c} + \frac{1}{\tau_e}. \tag{2.15}$$

Here, the motions of the spins are described using an additional internal correlation time τ_e and an order parameter S^2 that ranges from 0 – free motion to 1 – rigid. A simplified version of the model-free approach that is also used later on in REFINE assumes $\tau_e << \tau_c$ and thus neglects the second term in eq. (2.15). It is given by:

$$\mathrm{J}^n_{ij}(\omega) = \frac{1}{4\pi r^6_{ij}} \cdot \frac{S^2 \cdot \tau_c}{1 + \omega^2 \tau_c^2}. \tag{2.16}$$

Currently three more spectral density functions are available in RELAX for the description of fast methyl proton rotation and slower ring proton flips.

Using the full relaxation matrix analysis approach for NOE simulation allows to capture various higher order effects that occur in complicated systems with more than two spins, like e.g. so called spin diffusion effects, where magnetization is transferred via an intermediate spin to a third spin. These effects can enhance or attenuate the NOE cross peak integrals.

2.4.1 Simulating a NOESY spectrum with RELAX

With a list of expected signals, their chemical shift values, a trial structure and motional models assigned to the atoms in the protein under investigation, a spectrum can be simulated using the above formalism. From this input data the relaxation rates are calculated and by exponentiation of the relaxation matrix the NOE integrals are obtained for the desired experimental parameters. Numerically, the matrix exponential function can be computed by a number of different methods [68]. In RELAX matrix diagonalization is used and the exponential is calculated via the following similarity transform:

$$\mathbf{V} = \exp(-\tau_{\mathrm{m}}\mathbf{R}) = \mathbf{A} \cdot \exp(-\tau_{\mathrm{m}}\mathbf{B}) \cdot \mathbf{A}^{\mathrm{T}} \tag{2.17}$$

Here $\boldsymbol{A}$ is the matrix of eigenvectors of $\boldsymbol{R}$, and $\boldsymbol{B}$ is the diagonal eigenvalue matrix of $\boldsymbol{R}$. The resulting symmetric matrix $\boldsymbol{V}$ represents the spectrum for the mixing time τ_m.

Then an artificial dataset is created from $\boldsymbol{V}$ where the simulated signals are transformed to actual peaks with *Gaussian* or *Lorentzian* line shape (currently these two line shapes are supported by the RELAX spectrum backcalculation module in AUREMOL) in the frequency domain. The resulting simulated spectrum corresponds to a Fourier transformed and processed experimental spectrum and allows qualitative comparison. An absolute comparison of simulated and experimental peak integrals is not possible at this point, since the maximum starting magnetizations of all the spins have been set to $M_z(0) = 1$ for the simulation, without loss of generality, while in the experiment the signal intensities depend on several factors, primarily the sample concentration.

2.4.2 Comparing experiment and simulation – global scaling

To achieve direct comparability between experimental and simulated peak integrals, some kind of calibration has to be performed. For this thesis, the following volume scaling procedure based on a maximum likelihood estimation as proposed in [69] was chosen:

$$\alpha = \frac{\sum A_{ij}^{ex} \cdot A_{ij}^{sim}}{\sum \left(A_{ij}^{sim}\right)^2}. \tag{2.18}$$

After α is calculated, the simulated NOE integrals, corrected for a finite relaxation delay τ_r, are now directly comparable to the experiment and given by

$$V_{ij}(\tau_m, \tau_r) = \alpha \cdot \left[\exp(-\tau_m \cdot \mathbf{R})\right]_{ij} \cdot \sum_k \left[\mathbf{1} - \exp(-\tau_r \cdot \mathbf{R})\right]_{jk} . \tag{2.19}$$

This is the central formula used for the simulation of NOE signals in the REFINE distance calculation algorithm.

2.5 Distance calculation

From a known relaxation matrix ***R*** – calculated using known spectral densities – a distance d_{ij} between two protons *i*, *j* can be extracted from a cross relaxation rate R_{ij} using the above formalism, since

$$d_{ij} \propto R_{ij}^{-\frac{1}{6}} . \tag{2.20}$$

Several approaches exist for the acquisition of the relaxation matrix ***R*** from experimental data including IRMA [70], where NOE signals simulated from a trial structure are replaced by the experimental signals. From the resulting NOE matrix the relaxation matrix is backcalculated. This process is repeated for several different mixing times and from the averaged relaxation matrix distances are calculated. Kim and Reid [71] scale the resulting distances according to the difference in simulated and experimental NOE value. In both approaches structure calculation is used after each step to obtain a new trial structure for NOE simulation which is computationally very expensive, as is the backtransformation of the NOE matrix.

MARDIGRAS [72] abandons the structure calculation and instead iterates the cycle NOE matrix calculation – merging with experimental data – backtransformation, until the difference between simulation and experiment has become minimal. Again, the two transformations per cycle take a lot of computation time, especially for large systems. NO2DI [73] rescales the distances used in the rate calculation according to sixth root ratio of the simulated and experimental NOEs in an iterative fashion and thus saves one transformation step per iteration, however has to keep the distance matrix in memory and recalculate the relaxation rates. The REFINE procedure presented below (→ 3.7) directly rescales the relaxation rates until experiment and simulation match best. That way only relaxation rate rescaling and matrix exponentiation have to be performed during one iteration and no distance matrix is required.

In contrast to the relaxation matrix approach, the so called *Isolated Spin Pair Approximation* – ISPA – assumes linearity between NOE cross peak volumes $\boldsymbol{V}_{ij}$ and cross relaxation rates R_{ij}. So the distance is calculated directly from a cross peak integral as

$$d_{ij} = \alpha \cdot V_{ij}^{-\frac{1}{6}} \tag{2.21}$$

when a calibration factor α has been determined from a known distance and the corresponding signal integral. As indirect magnetization transfer effects are completely neglected by this method, it can yield accurate results only for very short mixing times where the approximation holds.

A method for the automated calculation of distances from NOE signals based on ISPA has been described by [74], where the NOE signals are divided into three classes:

	NOEs	Distance formula
Class I	H^{α}, H^{N} backbone-backbone; intraresidual, sequential and medium-range signals with one contributing proton H^{β}	$d = \sqrt[6]{A/V}$
Class II	All other signals excluding methyl groups	$d = \sqrt[4]{B/V}$
Class III	All signals involving methyl groups	$d = \sqrt[4]{C/V}$

Table 2.2 NOE classes and corresponding distance calculation formula according to [74]

The scaling factor *A* is calculated from the signals belonging to Class I under the assumption that the average distance here is 3.4 Å. Then the factors *B* and *C* are calculated from

$$B = \frac{A}{d_{\min}^2},$$
$$C = \frac{B}{2}, \qquad (2.22)$$

where d_{min} = 2.4 Å and the resulting distance constraints are restrained to the range 2.4-5.5 Å. This approach is based on the essentially very similar spatial distribution of protons in different globular proteins and assumes that in the different NOESY spectra the range of observable distances is comparable. Therefore good results can mainly be expected for globular proteins, although even here a dependence on secondary structure is probable, for example in the case of a globular protein containing only beta sheets.

2.6 Distance error estimation

Regardless of the way the distances are calculated, there is always a certain amount of error that has to be attributed to a certain distance mainly because of the following reasons:

- Errors in the experimental peak integral data due to noise, artifacts, baseline distortion, overlap
- Incorrect signal assignment
- Insufficient model parameters in the spectral density functions, insufficient consideration of e.g. chemical exchange effects in spectrum simulation (for the relaxation matrix approach in REFINE)
- Numerical errors (round-off errors, etc.)

Errors in the experimental data directly influence the distance determination process. In REFINE this is handled by introducing a new peak integral error estimate to the signal segmentation routine as described below (see section 3.4) and using this information during the calculation.

The occurrance of incorrect assignments can at least be quantified by the use of automated assignment procedures like KNOWNOE, where a minimum probability value can be specified that allows an estimate of the number of wrongly assigned signals, yet not which signals these are. An indication of a wrong assignment can be derived from the results of a *molecular dynamics* calculation, when a given restraint is strongly violated whereas the neighboring restraints seem all right. In that case, the distance calculation and MD should be repeated with the suspicious signal excluded or reassigned.

To get a measure of the errors inherent in the spectrum simulation part due to modelling of proton mobility, a statistical approach can be used in REFINE, where the modelling parameters are varied throughout a number of calculations and thus a distribution for each distance restraint can be obtained.

As numerical errors are assumed to have a minor influence on the resulting distances compared to the error sources described above, they are neglected in REFINE.

2.7 Dihedral angle prediction

Angle restraints for the φ and ψ backbone angles can be predicted from the protein sequence and chemical shifts for example by the program TALOS [75]. There, a combined chemical shift/structure fragment database is searched for matching residue triplets and from the result the backbone angles for the central residue are derived.

Since the φ/ψ angle distribution is especially important for secondary structure elements and the prediction is computationally not very expensive, it is a useful source of additional information to be used in structure calculation and fits well into the top-down approach of AUREMOL.

2.8 Structure calculation

Starting from an initial structure model, often an *extended strand* for proteins, a hypothetical structure is calculated using restraints derived from NMR experiments. Since distance geometry methods are not widely used anymore, the focus will in the following be on molecular dynamics simulations.

Molecular dynamics – MD – simulations calculate the motion of all atoms that make up a considered molecule in a heat bath under the influence of different forces. These include models of *Van-der-Waals*, bond stretching, *Coulomb* and bond torsion forces as well as pseudo forces derived from e.g. distance or angle restraints. In the beginning, a random momentum is applied to each individual atom in the molecule and the trajectories resulting from their evolution of the motion under the influence of the various restraining potentials is recorded. The goal is to minimize the total energy of the molecule under consideration.

Of special importance for the sake of protein structure determination are the distance restraints that can be derived from NOE data since they drive the virtual folding during the dynamics simulation. The total pseudo energy of the simulated system serves as a measure for the quality of the structure.

Cartesian dynamics simulations allow free motion of the atoms only restrained by the present force fields. Thus bond stretching, torsion and compression as well as bond angle changes are possible. By keeping the bond lengths and bond angles fixed, *torsion agle dynamics* simulations reduce the degrees of freedom as the bond lengths are kept fixed and only rotations around bonds are allowed which significantly reduces computation times.

Advanced molecular dynamics software packages allow the addition of solvent molecule layers around a protein and carry out the calculations in a solvent environment, which is clearly more realistic than a protein in total vacuum. This increases computation times as the dynamics for the solvent molecules have to be calculated as well, but leads to more refined structure hypotheses [76].

For the calculations in this thesis the program packages CNS [77] and XPLOR-NIH [78] have been used, other well known alternatives are e.g. DYANA/CYANA [79] or GROMACS [80].

Generally, in structure determination simulated annealing MD protocols are used that are divided into a heating phase and one or more cooling steps (e.g. standard protocol of CNS).

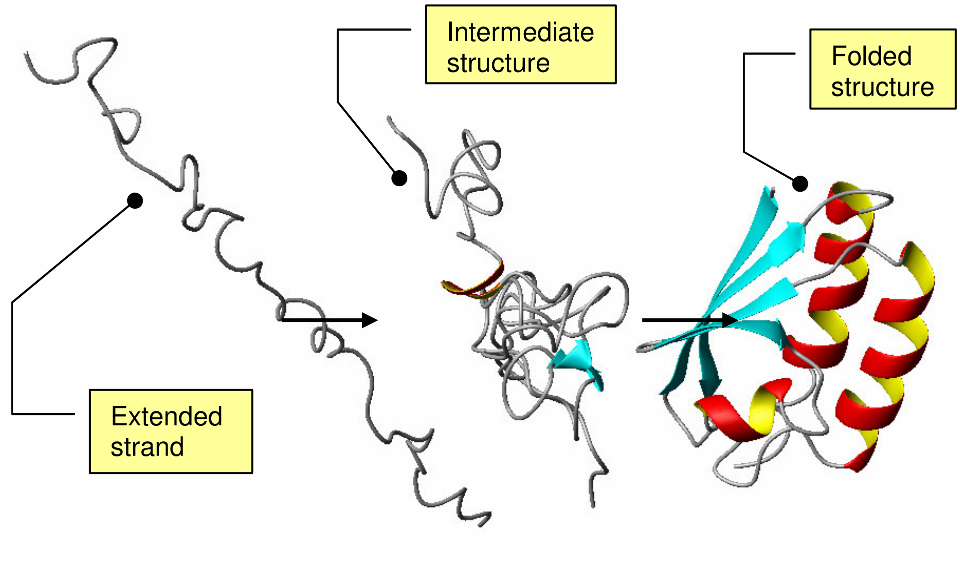

Figure 2.4 Simulated folding driven by distance restraint pseudo potentials in a molecular dynamics simulation run

For a given set of restraints, the simulation is repeated *n* times with different random seeds, leading to *n* different structure models. From these, the *m* models with the least overall pseudo energies and often the least restraint violations are chosen as a resulting ensemble of most probable structures.

In case of a known protein structure, MD simulations can be used to study the internal dynamics of a protein for certain environmental conditions (variable/constant temperature, pH, etc.). That way e.g. order parameters can be estimated.

An alternative structure calculation approach is *ab initio* structure prediction as in ROSETTA [81;82], where structures are assembled from fragments of homologous known protein structures. Including sparse NMR data in the calculations results in moderate to high resolution structures [83], depending on the quality of the fragment library.

2.9 Structure evaluation

Next to simple *rmsd* calculations a variety of statistical analyses, e.g. Ramachandran plots for φ and ψ angle distribution, side-chain χ angle orientation, etc., is available in a variety of programs like PROCHECK-NMR [84], WHATCHECK [85], PROSAII [86], etc.

For judging the quality and plausibility of the determined structures in AUREMOL, in analogy to X-ray an NMR R-factor has been introduced [25] and incorporated. Here simulated signals and experimental signals together with the experimental peak probabilities $p_{ex,i}$ provide a measure to judge structural quality. Experimental signals for which a corresponding simulated signal exists belong to set *A*, the remaining signals to set *U*. For the evaluation of the structures calculated in this thesis the following two R-factors have been used:

Global R-factor

$$R_5(\alpha) = \sqrt{\frac{\sum\limits_{i\in A}\left(V_{ex,i}^{\alpha} - sf_{\alpha}\cdot V_{sim,i}^{\alpha}\right)^2 \cdot p_{ex,i}^2 + \sum\limits_{i\in U}\left(V_{ex,i}^{\alpha} - sf_{\alpha}\cdot V_{noise}^{\alpha}\right)^2 \cdot p_{ex,i}^2}{\sum\limits_{i\in A} V_{ex,i}^{2\alpha}\cdot p_{ex,i}^2 + \sum\limits_{i\in U}\left(V_{ex,i}^{\alpha} - sf_{\alpha}\cdot V_{noise}^{\alpha}\right)^2 \cdot p_{ex,i}^2}} \tag{2.23}$$

R-factor using only assigned signals

$$R_3(\alpha) = \sqrt{\frac{\sum\limits_{i\in A}\left(V_{ex,i}^{\alpha} - sf_{\alpha}\cdot V_{sim,i}^{\alpha}\right)^2 \cdot p_{ex,i}^2}{\sum\limits_{i\in A} V_{ex,i}^{2\alpha}\cdot p_{ex,i}^2}} \tag{2.24}$$

The scaling factor is given by

$$sf_{\alpha} = \frac{\sum_{i \in A} \left(V_{ex,i} \cdot V_{sim,i} \right)^{\alpha}}{\sum_{i \in A} V_{sim,i}^{2\alpha}} \ , \ \alpha = -\frac{1}{6} \, . \tag{2.25}$$

For better agreement between experiment and simulation lower R-factors are obtained. In section 3.9.3.3, R_3 is also used as distance dependent R-factor, which is calculated separately for different distance ranges and only from the respective signal subset in A.

3. Theoretical considerations and results

3.1 Implementing the IUPAC nomenclature in AUREMOL

In an attempt to keep a consistent atom naming standard throughout the AUREMOL routines, the nomenclature proposed by IUPAC [87] has been introduced and replaces the formerly confusing and inconsistent atom naming in the different AUREMOL modules.
As a consequence, all input data, like for example structure files from the protein databank, is required to be in IUPAC format. For format conversion of input files several tools are now available within AUREMOL.

3.2 Improved spectrum simulation

The original implementation of RELAX suffered from a programming error that prevented it from correctly using different spectral densities for all the atoms in a protein. This has been fixed and the increase in accuracy is now available for REFINE. Additionally, spectra can now be simulated taking a finite relaxation delay time τ_r into account, so a simulated NOE integral V_{ij} for a mixing time τ_m is given by:

$$V_{ij}(\tau_m, \tau_\mathrm{r}) = \alpha \cdot \sum_{j,k} \left[\exp(-\tau_m \cdot \mathbf{R})\right]_{ij} \cdot \left[\mathbf{1} - \exp(-\tau_\mathrm{r} \cdot \mathbf{R})\right]_{jk} . \tag{3.1}$$

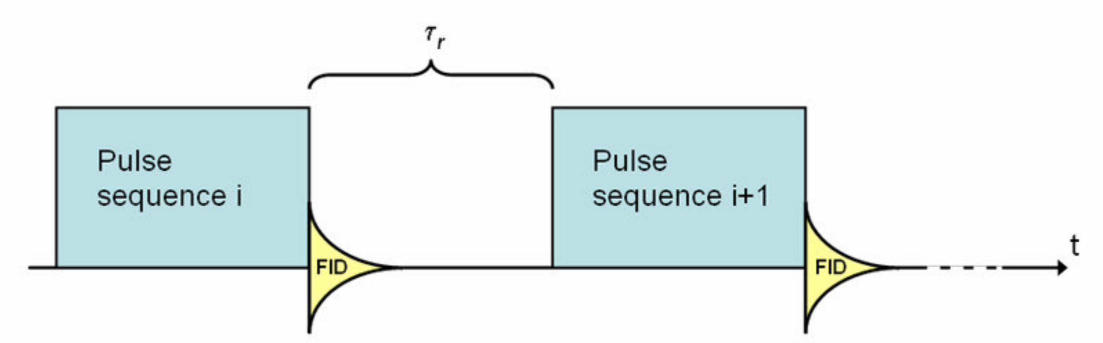

Figure 3.1 Definition of the relaxation delay time τ_{r}.

This leads to an asymmetric peak intensity distribution in the spectrum for shorter relaxation delay times, as observed in experimental data, since the nuclear spins cannot relax to a (near) equilibrium state before the next pulse sequence starts.

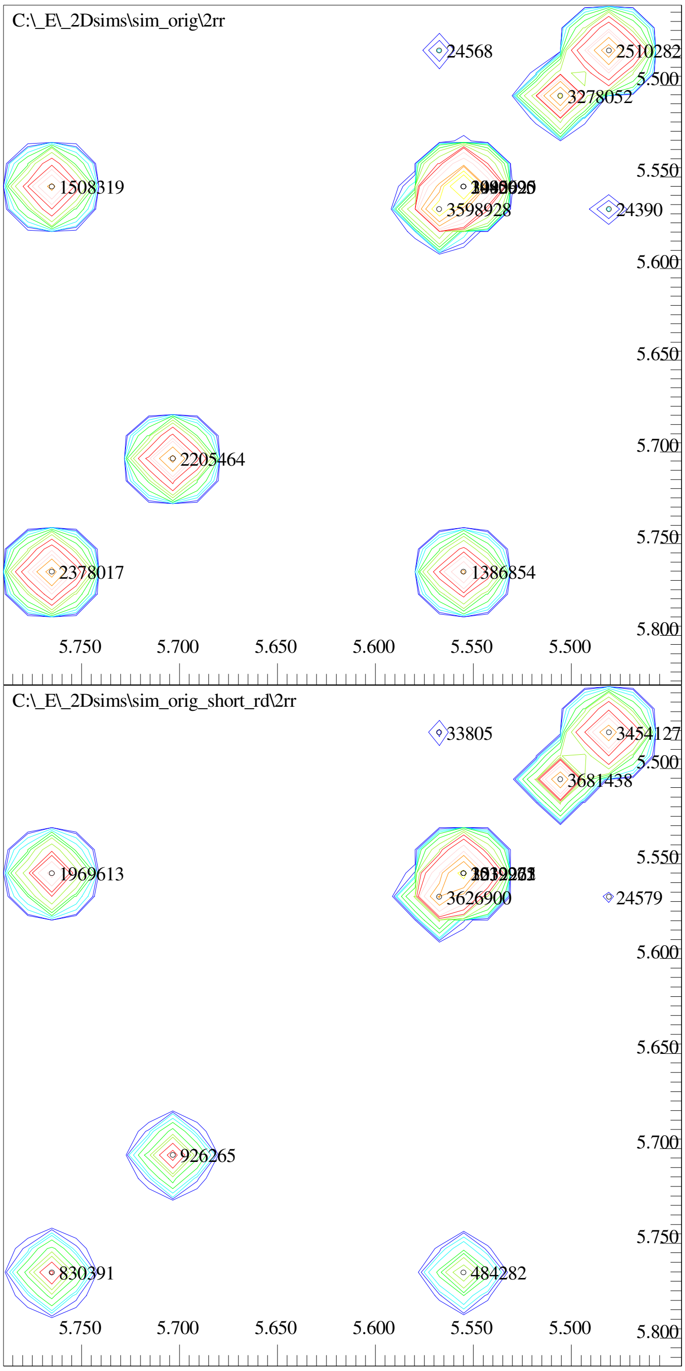

Figure 3.2 Only slightly asymmetric intensity distribution of symmetric cross peaks for a realistic relaxation delay time of 1.54 s in the simulation of a 2D NOESY spectrum for HPr from *S. aureus* (top), strong asymmetry for very short relaxation delay time of 100 ms (bottom).

This further increases the simulation accuracy and subsequently improves the performance of REFINE on real world data with inherent relaxation delay.

3.3 Adaptive peak picking

The already existing fixed threshold routine in AUREMOL exhibits the inherent drawbacks of this method, including insensitivity to noise and artifact peaks and is not suited for automation since a user defined threshold is required. Therefore it is extended by an automatic adaptive threshold determination procedure. It relies on a local noise estimate [35] to facilitate local thresholding, which eliminates the need of user input, aids automation and increases reproducibility. Additionally this approach is coupled to the Bayesian signal probability module [36;37] in AUREMOL by automatically creating reference sets for signal and noise/artifact peaks.

As a first step for every row *n* in each dimension *m* of the spectral data, according to [35] the intensity variance σ_i^2 in a sliding window of 5% of the spectrum resolution around every datapoint is calculated and the minimal variance is chosen as least noise estimate σ_n^2.

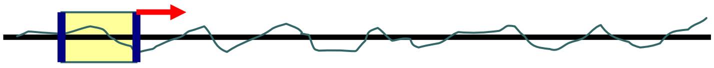

Figure 3.3 The local variance is calculated in a window around each datapoint

With the global minimum variance of the dataset σ_{min}^2 the additional relative noise level $\sigma'^2_{x_n,i}$ for a given row *i* is given by

$$\sigma'^2_i = \sigma_i^2 - \sigma_{min}^2 \tag{3.2}$$

So the minimum noise level *N* for a datatpoint with the coordinates $x_1, \ldots, x_n$ of a dataset with dimension *dim* = *n* is given as

$$N_{x_1,\ldots,x_n} = \sqrt{\sum_{i=1}^{n} \sigma'^2_{x_i} + \sigma_{min}^2} = \sqrt{\sigma_{x_1}^2 + \ldots + \sigma_{x_n}^2 - (n-1)\cdot\sigma_{\min}^2} \tag{3.3}$$

with σ_{min}^2 the global minimum variance of the dataset. Generally baseline errors are neglected in this procedure, so baseline corrected data is required.

To account for strong noise in the neighboring rows, the *m* next neighbor (NN) noise levels can be regarded with a ratio of *a/b*:

$$N_{loc(x,y,...)} = \frac{1}{a+m\cdot b}\left(a\cdot N_{(x,y,...)} + \sum_{i=1}^{m} b\cdot N_i^{NN}\right) \tag{3.4}$$

Here the ratio of actual noise level at the datapoint to that of its next neighbors *a* : *b* has been fixed to 2 : 1 in the 2D case and 3 : 1 in the 3D case and thus the calculation corresponds to the application of a simplified Gaussian smoothing filter kernel:

0.27	1.21	0.27
1.21	2.0	1.21
0.27	1.21	0.27

0	1	0
1	2	1
0	1	0

0.13	1.37	0.13
1.37	3.0	1.37
0.13	1.37	0.13

0	1	0
1	3	1
0	1	0

Figure 3.4 Top: 2D Gaussian filter kernel values for a standard deviation of 1.0 (left) and corresponding simplified integer kernel (right). Bottom: Central cross section of 3D filter kernel for a standard deviation of 0.8 (left) and corresponding simplified integer kernel (right).

This measure for local noise can be used for a local threshold determination. An intensity maximum (and analogously minimum) I_i is interpreted as a potential signal if the following two conditions are met:

$$\begin{aligned} \bar{I}^{NN}_{loc,i} &> \alpha\cdot N_{loc,i} \\ I_i &> \bar{I}^{NN}_{loc,i} + \beta\cdot N_{loc,i} \end{aligned} \tag{3.5}$$

$\bar{I}^{NN}_{loc,i}$ denotes the smoothed average of the local intensity and its next neighbors (2D: simplified kernel for standard deviation of 0.8, 3D: simplified kernel for standard deviation of 0.6). For now, α and β are empirical constants (see table 3.1) determined from the application of the procedure to a range of test datasets, in the future these constants should be determined implicitly from the spectrum under consideration. So according to the conditions in eq. (3.5) a real signal must not have an absolute smoothed central intensity level below a certain multiple α of the smoothed local noise level, meaning that the local average intensity generally has to exceed the local noise for a real signal. Additionally the actual central intensity value must be greater than the smoothed central intensity plus a multiple β of the smoothed local noise level, implying that a certain signal fall-off from the center is required for a real signal to be recognized.

	2D	**3D**
$\bar{I}^{NN}_{loc,i}$	$= \dfrac{3 \cdot I_{loc} + \sum_{i=1}^{4} I_i^{NN}}{7}$	$= \dfrac{4 \cdot I_{loc} + \sum_{i=1}^{6} I_i^{NN}}{10}$
α	3.87	2.06
β	0.26	1.29

Table 3.1 Calculation of local intensity average and empirically determined α and β values for 2D and 3D data.

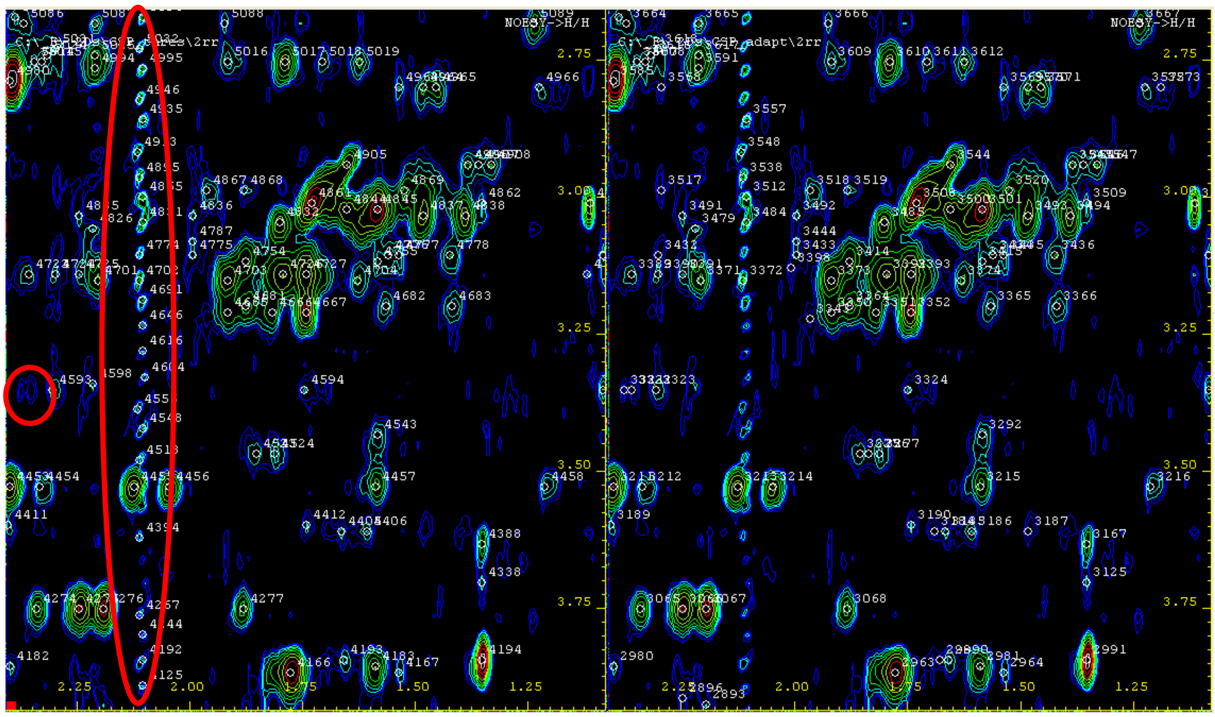

Figure 3.5 Comparison of threshold peakpicking (left) vs. adaptive peak picking (right) in an experimental 2D NOESY spectrum of TmCSP. For a threshold of 15000 the simple thresholding routine classifies neraly all artifacts of the vertical noise streak to the left as signals, whereas in the adaptive routine no threshold has to be specified, most artifacts of the streak are discarded and weak signals discarded by the simple thresholding procedure are identified.

As the figure above shows, this approach works very well on local noise streaks. When looking at the water artifact region, the number of wrongly picked artifact signals is reduced, however still a number of artifact signals are picked:

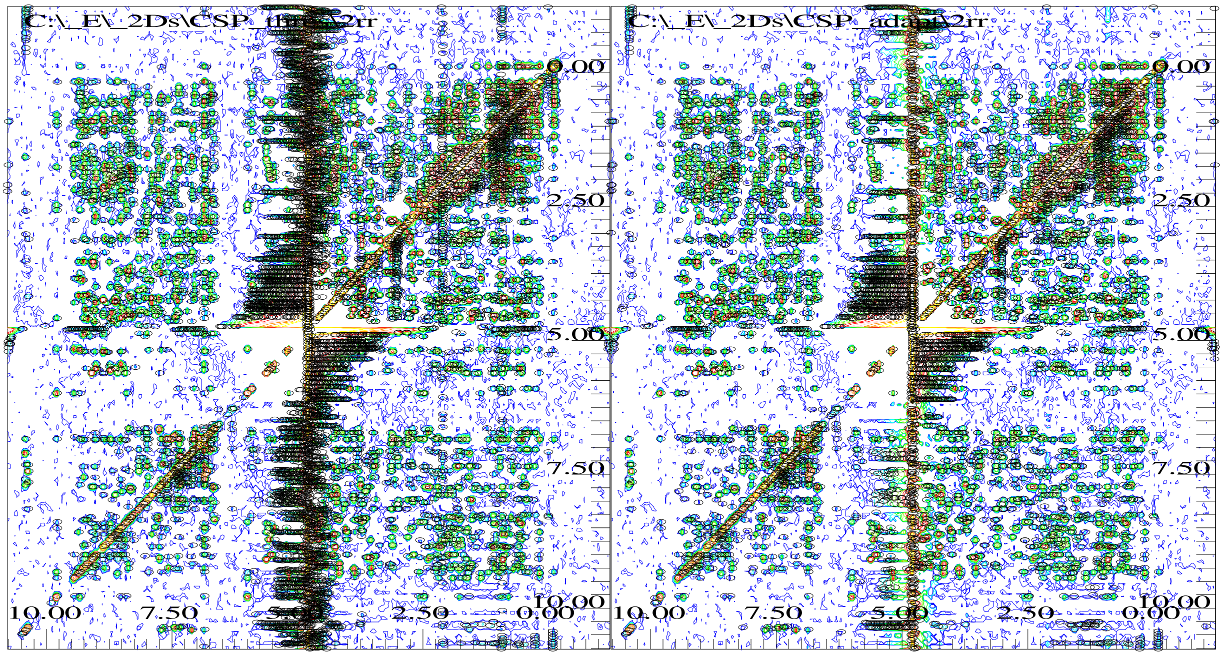

Figure 3.6 Comparison of signals identified (black circles) by threshold peakpicking (left, 7792 signals) vs. adaptive peak picking (right, 5559 signals) in an experimental 2D NOESY spectrum of TmCSP. The number of wrongly picked signals in the central vertical water artifact streak is strongly reduced by the adaptive routine.

Currently, this part of the adaptive peak picking algorithm is implemented for 2D and 3D NMR spectra. The extension to higher dimensional datasets is straight forward.

Since in AUREMOL also a Bayesian peak analysis is available that allows to discriminate between signal and artifact peaks based on user defined trainig sets, the idea was to integrate this approach into the adaptive peak picking routine to further reduce the amount of picked artifact signals. So the task is to automatically generate the training sets for the Bayes algorithm during peak picking. To this end for 2D NOESY data a simple signal pre-classification is employed:

Classification by adaptive thresholding	**Pre-classification for Bayes analysis**
Artifact/noise	**Artifact/noise**
Signal	If symmetric cross peak exists: → **Signal** Otherwise: → **Unknown**

Table 3.2 Signal pre-classification scheme for further Bayesian peak analysis.

The final training sets are then determined by a local peak shape analysis of the artifact/noise and signal class members. This is accomplished by examining the intensity distribution in a grid around a given signal maximum:

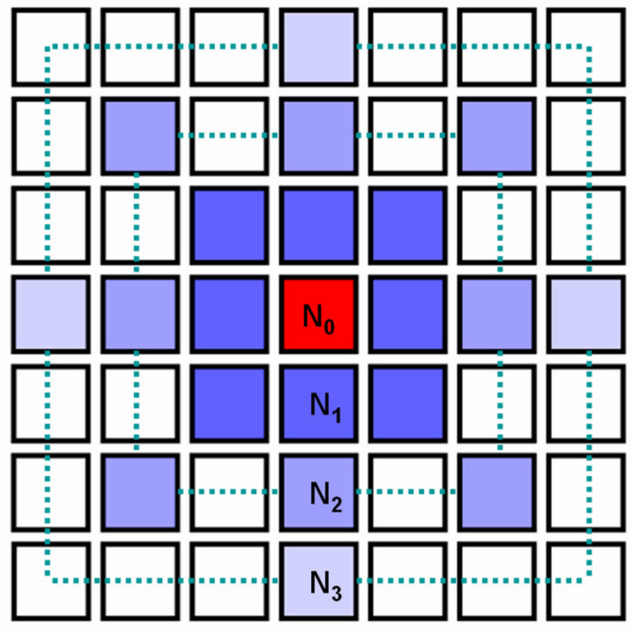

Figure 3.7 The intensities of sample datapoints (blue) with neighborhood orders of N_1 to N_3 around a signal maximum N_0 (red) are evaluated. Note that the spacing of the sample points can be adjusted for different resolutions and should be chosen for the grid to cover the expected mean peak area.

Considering the slope and local symmetry of each peak pre-classified as signal or artifact/noise, a score value *S* is generated from the intensity values *I* (see table 3.3) where higher score values indicate a more well defined, more symmetric signal.

Falling slope between closest neighbors in $N_0 \ldots N_3$, $I_{N_{n+1}} < I_{N_n}$:	Add $\frac{I_{N_{n+1}}}{I_{N_n}} - 0.5$ to score S .
Symmetry in N_1 and N_2 regions:	Add $\sum_{i=1}^{2}\sum_{n=1}^{8}\left(-\frac{\left\lvert\bar{I}_{N_i} - I_{N_i,n}\right\rvert}{\bar{I}_{N_i}}\right)$ to *S*.

Table 3.3 Per peak score calculation. Top: Positive score for falling slopes > 0.5. Bottom: Negative score penalty for symmetry deviations in N_1 and N_2.

So in the next step the 100 highest scoring peaks from the signal class are tranferred to the signal training set and the 50 lowest scoring peaks from the artifact/noise class are transferred to the artifact/noise training set. From here the original Bayes routine in AUREMOL takes over and calculates the probabilities for every picked peak to actually be a signal. In the final step, all peaks with a probability below a desired user defined threshold are deleted:

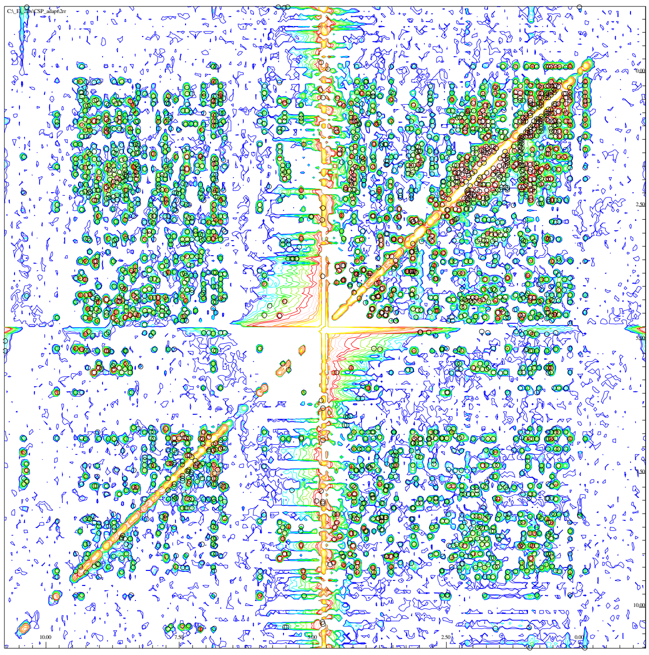

Figure 3.8 In this example, after the Bayesian signal analysis with automatically generated training sets all signals with a probability below 0.20 have been deleted, resulting in 2535 signals (black circles).

Setting	Value
Spectrum resolution	W1: 1024, W2: 2048
Peak picking mode	Only positive peaks
W2 steps for local sampling	2
W1 steps for local sampling	1
Iteration count for segmentation	10
Max. integration witdh	25 Hz
Remove solvent peaks	•
Remove diagonal peaks	•
Remove artifact streaks	•
Global symmetry matching	•
Probability threshold	0.20

Table 3.4 Parameter settings used to produce the result in figure 3.7. The W2 and W1 steps parameters define the grid spacing (see Fig. 3.7) in data points.

As figure 3.8 shows, most artifact signals can be removed using this fully automatic approach. The most important remaining parameter is the probability threshold. For higher threshold values, the probability for removing real signals increases.

The generalization of this approach to arbitrary spectra is a work in progress.

3.4 Integration error estimation

The optimization target for the REFINE distance calculation algorithm described below is the experimental peak integral data. Therefore it is important to get a realistic estimate of the contained error. In the following approach, parallel to the integral calculation by the segmentation approach implemented in AUREMOL, the error contributions from noise and peak overlap are now calculated in the extended segmentation routine:

For a signal integral V_0 the estimate consists of a noise error contribution Δ_{noise} and an overlap error contribution $\Delta_{overlap}$. The intensity maximum of the signal under consideration is the main seed, all other local maxima inside the integration box are additional seeds. For each signal the integration box is calculated from an assumed Lorentzian line shape with a given expected linewidth ν_k and a segmentation level of at maximum 10% of the maximum signal intensity; the integration will carry on until the signal intensity drops below 10% of the central peak intensity. Summing up all n datapoints I_i that fulfill the growing condition for the main seed yields the peak integral V_0:

$$V_0 = \sum_{i=1}^{n} I_i \tag{3.6}$$

Analogously, the relative noise error contribution Δ_{noise} is calculated from the local noise levels at the n datapoints i belonging to the integral:

$$\Delta_{noise} = t \cdot \frac{\sqrt{\sum_{i=1}^{n} N_{loc,i}^{2}}}{V_0} \tag{3.7}$$

Here, t is the *t-test* value for one sample and the user specified confidence level of the noise error estimate (see eq. 3.3). The result represents a minimum volume error contributed by the local noise in the dataset, weighted for the confidence level.

For the estimation of the relative overlap error $\Delta_{overlap}$, currently as a worst case estimate a *Lorentzian* line shape peak with the user defined linewidth ν_k is assumed at the position of each additional seed. The contributions at the center of the integration box, i.e. at the main seed with the intensity I_0, are summed up and divided by I_0:

$$\Delta_{overlap} = \frac{1}{I_0} \sum_{\substack{n\neq 0 \\ \text{seeds in} \\ \text{integration} \\ \text{box}}} I_n \cdot \prod_{k=1}^{\text{dim}} \frac{\nu_k^2}{\nu_k^2 + \left(d\nu_{k,n}\right)^2} \tag{3.8}$$

This overlap error term is an estimate of the influence of neighboring peaks on the volume integral of the main peak.

The overall absolute error estimate V_{err} for a signal integral V_0 is then given by the sum of both contributions:

$$V_{err} = \left(\Delta_{noise} + \Delta_{overlap}\right) \cdot V_0 \tag{3.9}$$

To test this error estimation approach, it was applied to a simulated 2D NOESY dataset:

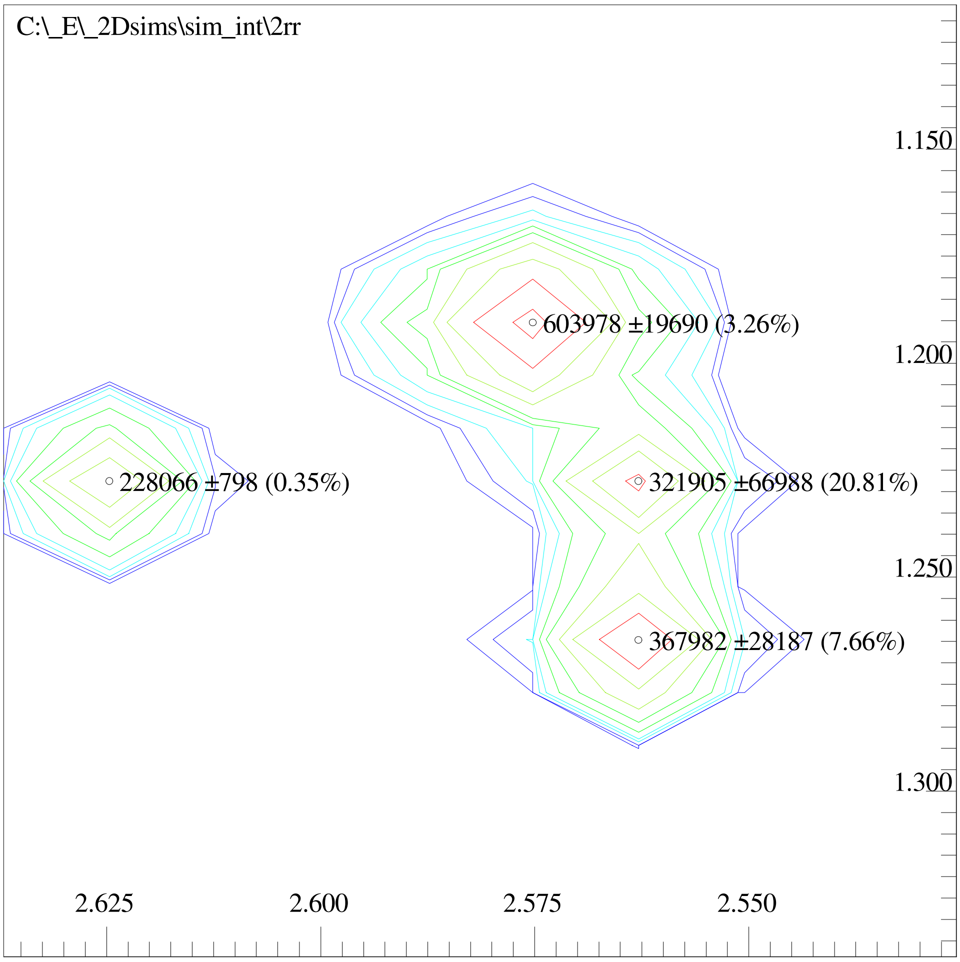

Figure 3.9 Overlap error estimation results on a simulated noiseless 2D NOESY dataset. For the rather isolated peak to the left only a very small contribution of 0.35% is calculated, while for the overlapping cluster of three peaks the overlap error estimate ranges from 3.26% (top) and 7.66% (bottom) for the stronger signals overlapping directly with only one peak to 20.81% (middle) for the weakest signal that directly overlaps with the two neighboring resonances.

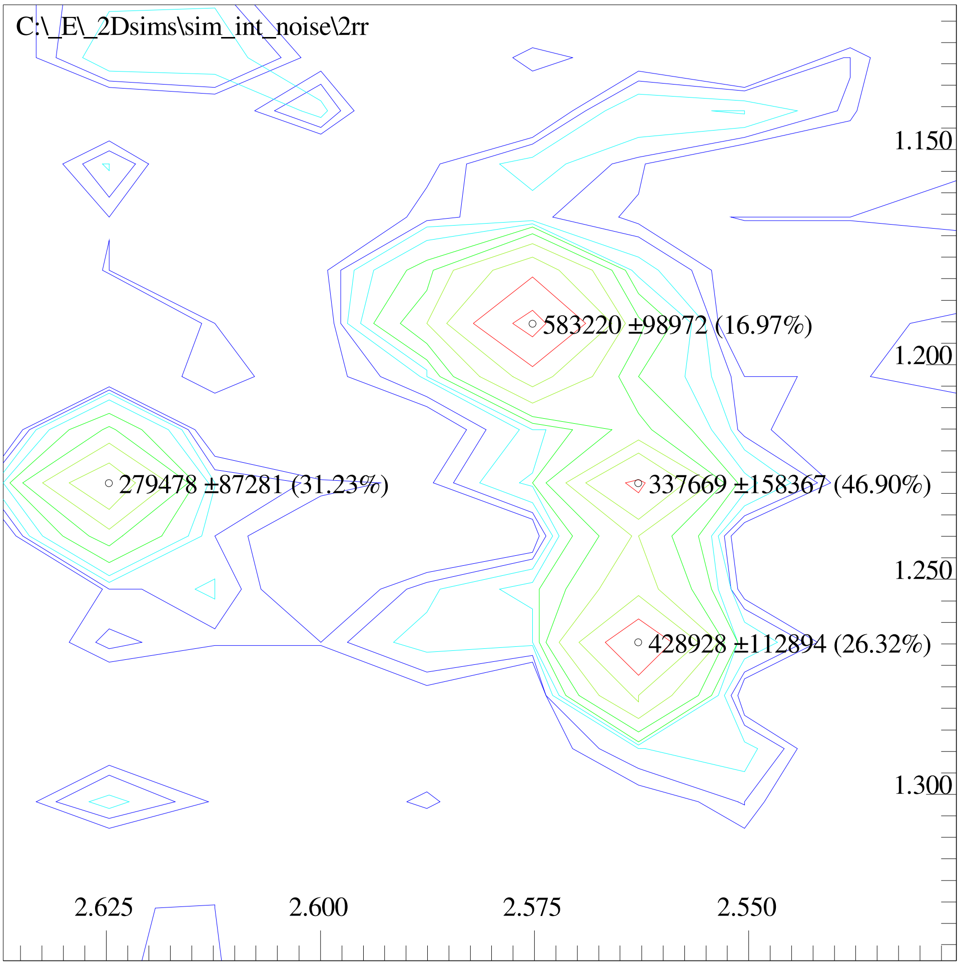

Figure 3.10 The addition of Gaussian noise with a standard deviation of 10% of the mean intensity of the spectrum raises the error estimates as now a noise contribution to the error can be calculated. The lower intensity signal integrals are expectedly influenced stronger by noise.

As the example figures above show, this error estimation approach generally produces comprehensible results. The integration parameters used in this example are shown in table 3.5:

Lowest segmentation level	0.1
Number of iterations	10
Max. W1 expected half line width at half height	10 Hz
Max. W2 expected half line width at half height	10 Hz
Confidence level	99.95 %

Table 3.5 Integration parameters used in the example shown in Figs. 3.9 and 3.10.

Comparing the calculated integrals from the noised spectrum to the simulated values shows that 52% of the signals are within the error bounds. This low number indicates that there have to be a number of cases where the presented approach reaches its limit. The main problem is undetected overlap:

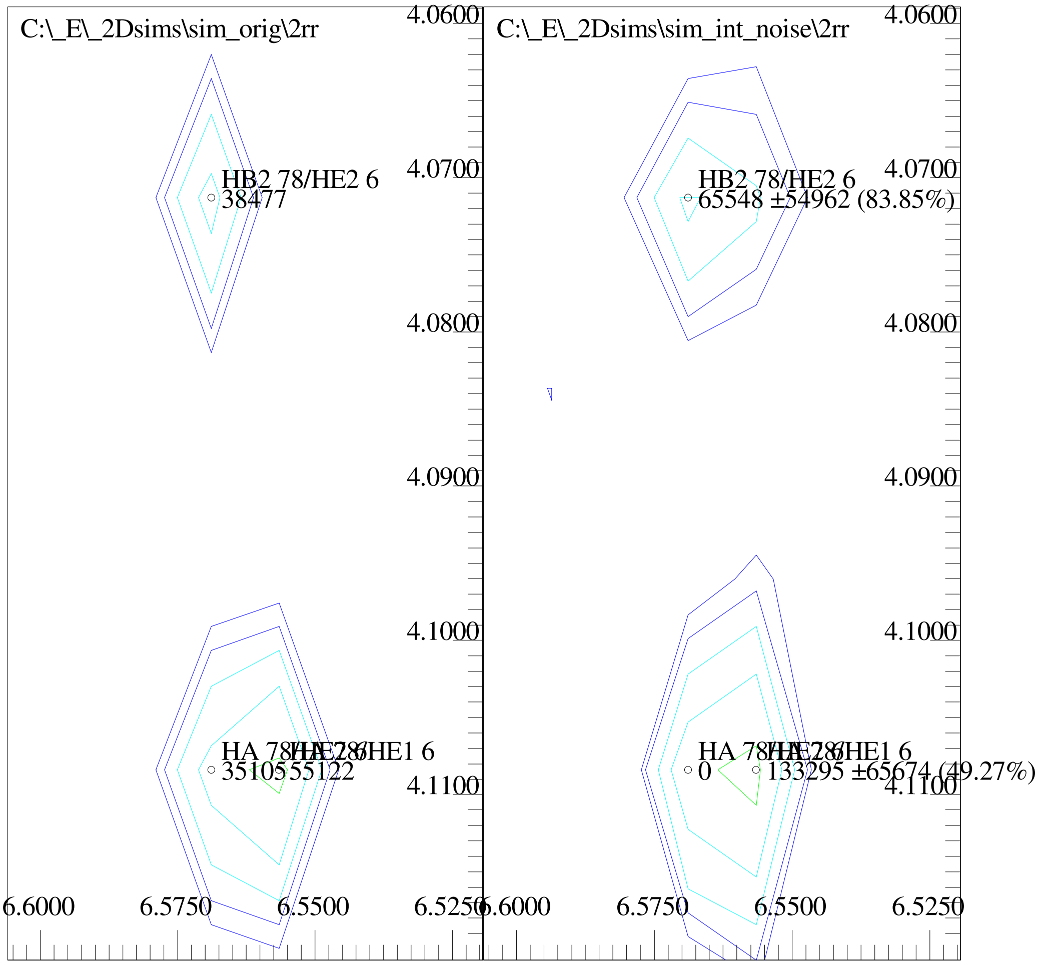

Figure 3.11 Three signals are shown from the simulated dataset (left) and the noised and integrated dataset (right). For the single signal in the upper half, the integration routine yields a too large integral (65548 vs. 38477 in the simulation) due to the added noise, however the error estimate of 83.84% covers the original value. In the lower half, two overlapping signals were simulated. Since at the position of the left hand peak there is no local maximum, the segmentation routine cannot integrate this signal. As a consequence, only the right hand signal is integrated and the intensity values of both signals are attributed to the right hand signal. As the overlap error estimation also relies on a local maximum to be present, the overlap error contribution cannot produce a result for the left hand peak and thus the overall error estimate is too low.

As the effect exemplified in Fig. 3.11 is inherent to the segmentation approach, there is no possibility to remedy this problem, unless additional information regarding for example the positions of overlapping peaks is available. For real world data this information is not easy to obtain.

3.5 The strips tool

For the analysis of triple resonance spectra as used in the sequential assignment process a strip tool has been incorporated in AUREMOL. It allows to collect the strip data, where usually one strip contains the corresponding information for one spin system, from a number of different 3D spectra and align them in slots for the actual manual assignment. Before strips can be used for obtaining a sequential assignment, they first have to be defined. Instead of doing this manually, strips from a 3D spectrum can now be automatically defined according to the corresponding peaks in a 2D HSQC spectrum. The peak coordinates of the 2D spectrum are used as input for the strip tool to cut the corresponding strips from the 3D dataset:

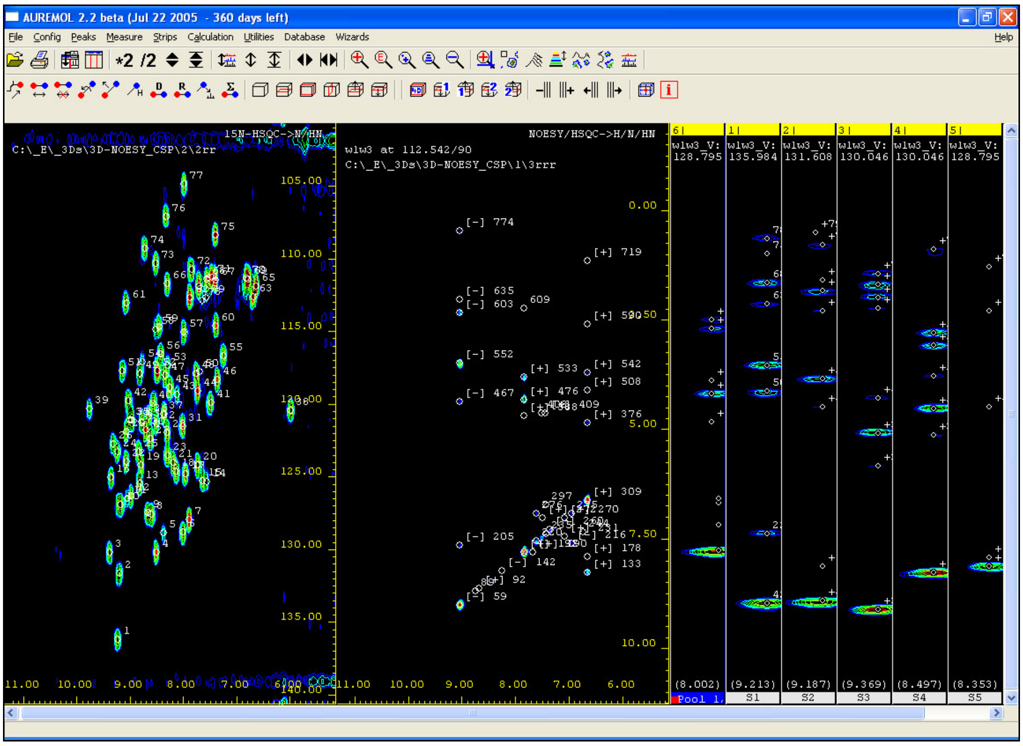

Figure 3.12 The 2D HSQC spectrum (left) provides the coordinates for the strips that are automatically defined in the 3D NOESY spectrum (middle) and collected in the strip window (right) for further manual analysis.

At first, all spectrum strips are collected in the so called strip pool, the left-most part of the strip window display. From there, a strip can be assigned to a free slot and help lines can be set to assist when browsing the strip pool for other matching strips (Fig. 3.13).

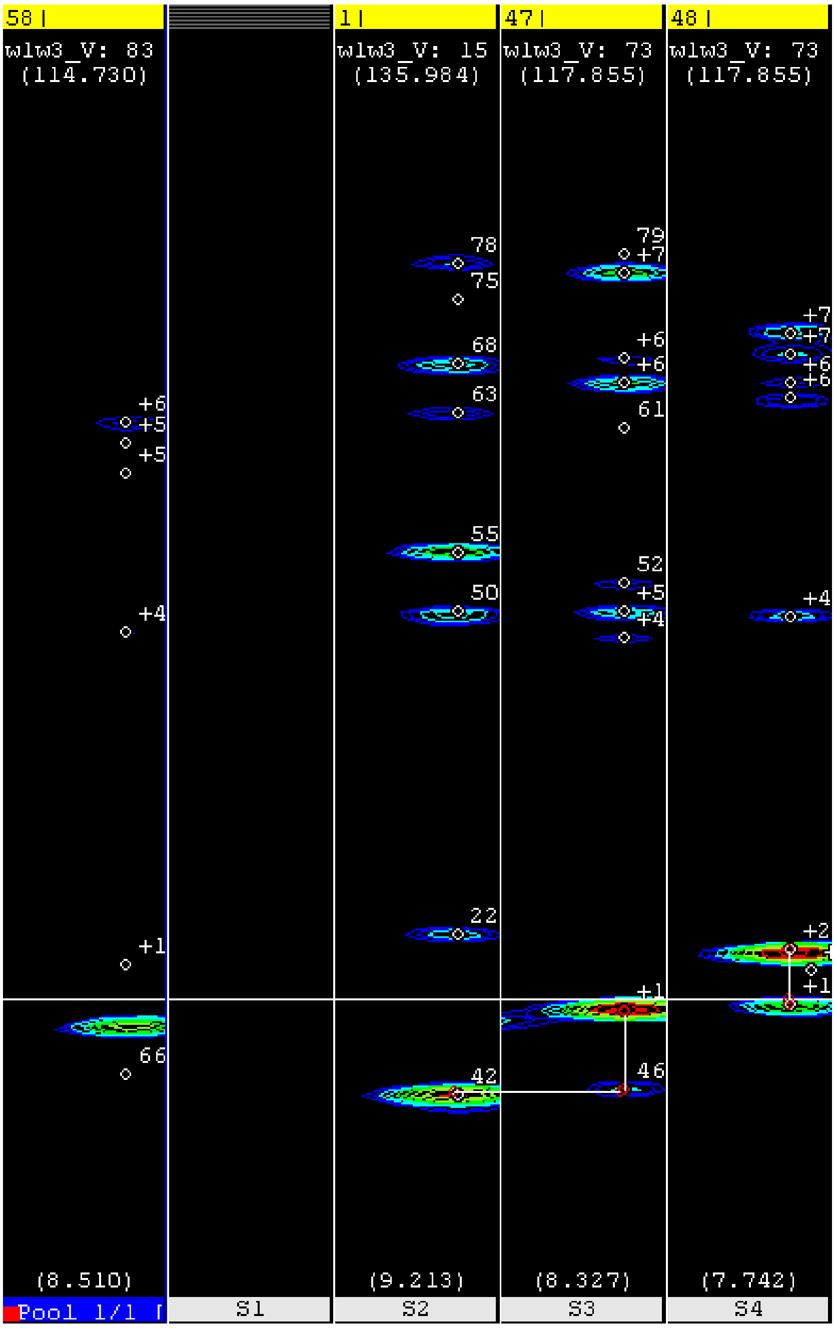

Figure 3.13 Strips from the pool (left) can be placed into free slots S1,…, Sn and arranged in the correct sequence. The help line display is activated.

If additional strip data from other spectra is available, the slots are divided to take up the additional strips. That way the results from two or more experiments can be used simultaneously during the assignment which greatly helps to resolve ambiguities.

The actual task of determining the sequential assignment from the strips within this tool still has to be carried out manually, although for incomplete or ambiguous shift data the automatic sequential assignment tool in AUREMOL can be employed.

3.6 Simulation based *de novo* assignment

The TWOSTEP algorithm for automatic sequential assignment described in [57] showed mixed results when used without a priori available partial assignment even for artificial data and was very time consuming. Furthermore the first step of the algorithm has recently been replaced by the PEAK ASSIGN module (Kirchhöfer et al., to be published), so the second step is used as a starting point for the new ASSIGN routine outlined in the following paragraph:

The basic algorithm based on threshold accepting for the optimization of a pseudo energy function calculated iteratively for varying assignments is kept. By restraining the possibilities of chemical shift assignment to the known distributions for each proton-amino acid combination[*)] the search space is considerably reduced and computation time is generally decreased.

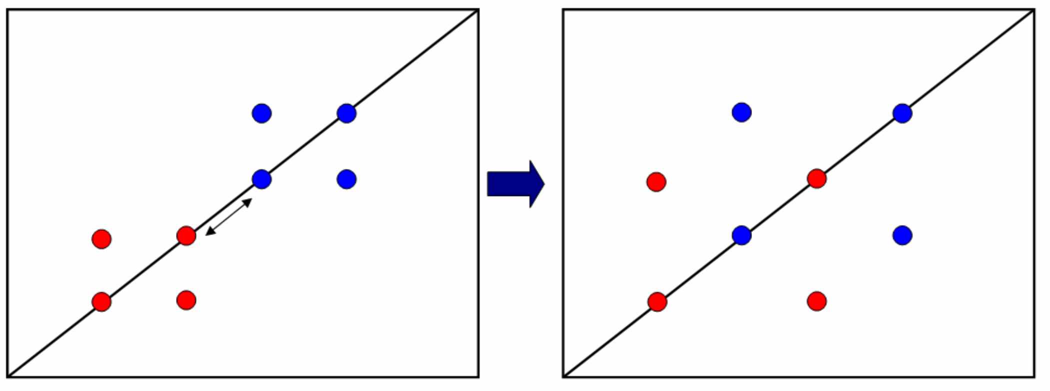

Figure 3.14 Effect of exchanging the chemical shift assignment of two spins on the intensity distribution of the simulated 2D NOESY spectrum

To improve the convergence of the improved implementation, an extended pseudo-energy function E_{tot} is used.

$$E_{tot} = E_{peak} + E_{grid} \tag{3.10}$$

Now both a local energy value E_{peak} and a global matching value E_{grid} are considered whereas in the original approach only E_{peak} was used.

Both of the pseudo energy contributions are calculated from the intensity values I^{exp} and I^{sim} for the n peaks in the experimental and simulated dataset, the index zero indicating the central intensity value of a signal.

[*)] Chemical shift infomation taken from the Biological Magnetic Resonance Data Bank (www.bmrb.wisc.edu)

$$E_{peak} = n - \sum_{n} \frac{\sum\limits_{\substack{z,\\ \text{all pixels}\\ \text{of peak}}} I_z^{exp} \cdot I_z^{sim}}{\sqrt{\sum\limits_{\substack{z,\\ \text{all pixels}\\ \text{of peak}}} I_z^{exp^2} \cdot \sum\limits_{\substack{z,\\ \text{all pixels}\\ \text{of peak}}} I_z^{sim^2}}} \qquad (3.11)$$

E_{peak} is a measure for peak shape similarity, but the exclusive use of this term for the optimization could not prevent the clustering of a large number of assignments on only a few resonances when starting from 0% known assignments.

Introducing E_{grid} as additional term to the optimization pseudo energy function is expected to alleviate this problem, since it is a measure for the global distribution of the peak maxima and serves as a penalty function when an improper global distribution due to clustering occurs. It is calculated analogously to E_{peak}, with the difference that here the global match of the central peak intesities I_0 is considered:

$$E_{grid} = n \cdot \left(1 - \frac{\sum\limits_{\substack{\text{all non-}\\ \text{diagonal peaks}}} I_0^{exp} \cdot I_0^{sim}}{\sqrt{\sum\limits_{\substack{\text{all non-}\\ \text{diagonal peaks}}} I_0^{exp^2} \cdot \sum\limits_{\substack{\text{all non-}\\ \text{diagonal peaks}}} I_0^{sim^2}}} \right) \qquad (3.12)$$

With theses additions, the ASSIGN algorithm was applied to a simulated 2D NOESY dataset of the wild type HPr *(S. aureus)* protein consisting of 88 amino acids to compare the performance (see Fig. 3.15) of the new routine with previous results on a similar protein (the HPr *S. aureus*, H15A mutant). A total of 6348 peaks originating from 520 individual proton spins for a cutoff value of 0.5 nm were the input for the improved optimization. Using a stepsize of 8000 iterations and a cooling factor of 0.99 the calculation yielded a 100% correct assignment starting from 0% known assignments after < 4.5 million iteration steps.

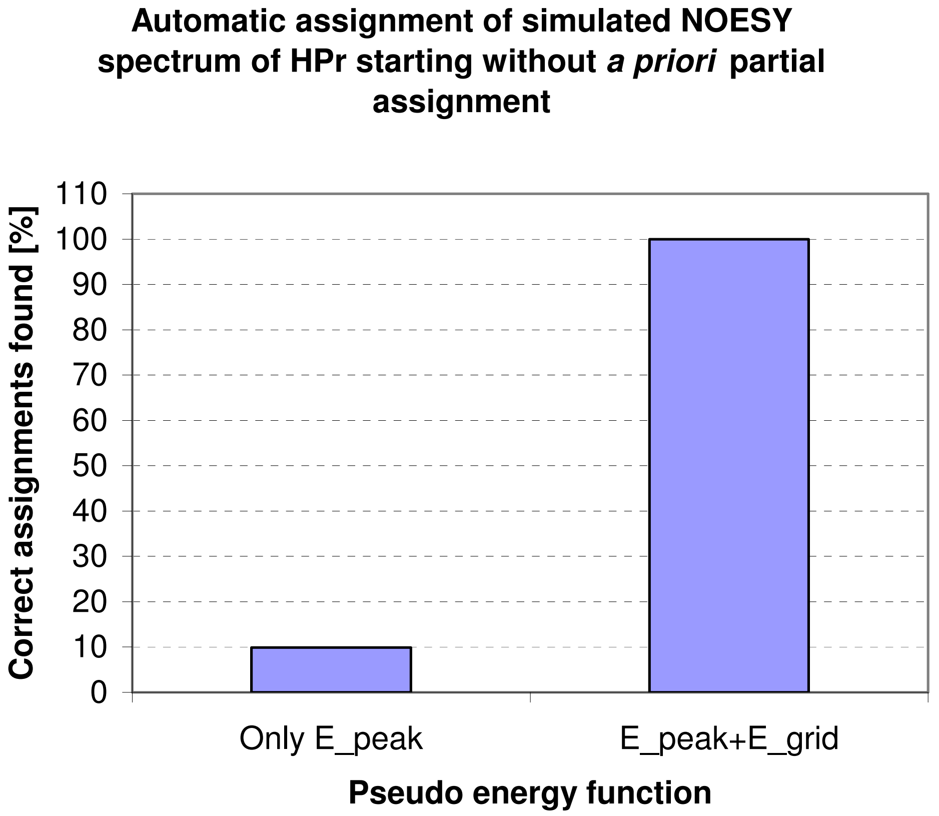

Figure 3.15 Assignment results from the previous version [57] (left) compared to the new version (right) using the additional pseudo energy term in the optimization starting without *a priori* partial assignment.

The unmodified routine yielded <10% correct assignments starting from 0% known assignments, although due to a very low simulation cutoff value of 0.2 nm a smaller number of 2122 peaks had to be assigned. In comparison, the introduction of the additional pseudo energy term in the new routine results in a fairly large improvement. The transfer of these results to real world spectral data is a work in progress.

3.7 The REFINE algorithm - Overview

Distance calculation is a core component in automatic structure determination. Since high accuracy is desirable, the relaxation matrix approach implemented in RELAX is used for spectrum simulation and higher order spin difusion effects are accounted for. On this basis the purpose of REFINE (see Fig. 3.16) is to allow the accurate automatic determination of proton-proton distances from experimental NOESY data in a fast, reliable and reproducible way without strongly depending on the structural input. REFINE requires a structure ensemble or a single trial structure, the integrals of assigned experimental signals and the corresponding simulation parameters for spectrum backcalculation as input.

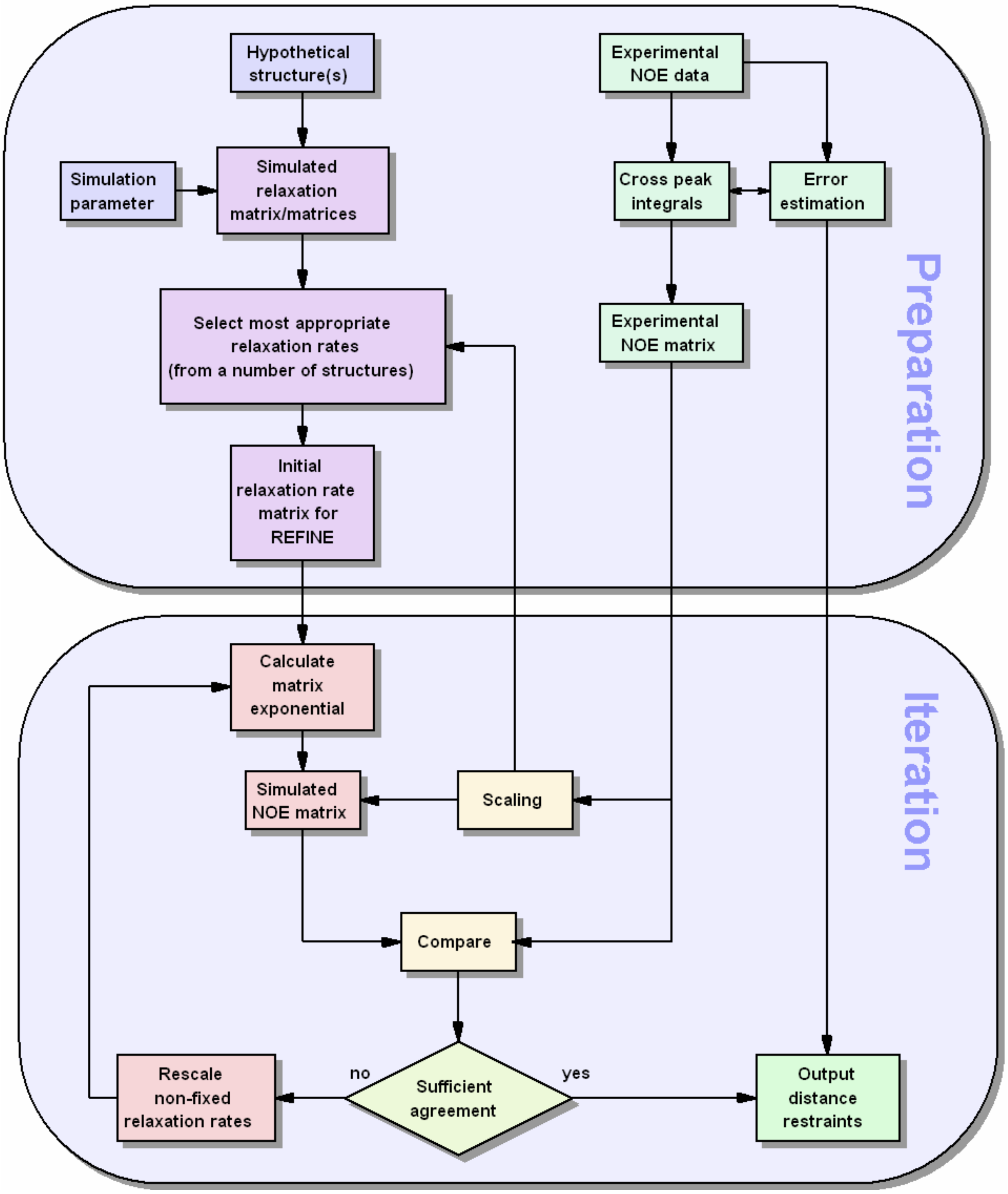
Hypothetical structure(s)
Experimental NOE data
Simulation parameter
Simulated relaxation matrix/matrices
Cross peak integrals
Error estimation
Select most appropriate relaxation rates (from a number of structures)
Experimental NOE matrix
Preparation
Initial relaxation rate matrix for REFINE
Calculate matrix exponential
Iteration
Simulated NOE matrix
Scaling
Compare
no
yes
Rescale non-fixed relaxation rates
Sufficient agreement
Output distance restraints

Figure 3.16 Flowchart of the REFINE algorithm. *Blue, violet*: Generation of starting relaxation matrix. *Green*: Experimental data, output. *Red and yellow shades*: Iterative refinement of the relaxation matrix. In the preparation section (top) a starting relaxation matrix is calculated from one or more trial structures. This starting matrix is then iteratively refined (bottom) until the NOEs simulated from this relaxation matrix best match the experimental NOEs.

At first, NOE integrals are simulated for the trial structure(s). Therefore the relaxation matrix for every trial structure is determined by calculating the relaxation rates of interacting proton spins depending on the proton-proton distance in the trial structure and the specified spectral density function according to eqs. (2.12)-(2.16). For distances corresponding to intra-residual proton-proton contacts of known distance (e.g. between CH protons in aromatic rings) standard values are used and kept fixed throughout the calculations. Note that in the REFINE algorithm for every signal present in the experimental data a simulated signal is calculated; for any given experimental signal whose corresponding protons in a trial structure are too far apart to be considered in a normal RELAX backcalculation (i.e. their distance in the structure is larger than the RELAX cutoff value), the simulation is forced by overriding the RELAX cutoff and calculating the respective relaxation rate as stated above. Then for every trial structure the NOE matrix is calculated using eq. (3.1) and the global scaling procedure described in eq. (2.18) is applied to allow a direct comparison between the observed experimental and simulated signals. In the case of more than one trial structure, the simulated signal $V_{sim,i,n}$ with the least deviation $\left|V_{sim,i,n} - V_{exp,i}\right|$ from the experimental value $V_{exp,i}$ is chosen from the n simulations for all experimental signals i and the corresponding relaxation rate is transferred to a quasi-hybrid relaxation matrix. Rates corresponding to signals not observed in the experiment are simply averaged and also entered into this matrix. At this point the starting relaxation matrix for the iterative part of REFINE is obtained.

The second half of the algorithm consists of a loop (see lower half of Fig. 3.16): For each iteration the matrix exponential function of the relaxation matrix is numerically evaluated yielding a simulated NOE matrix (see eq. 3.1). The simulated peak integrals are compared to the corresponding experimental data with the help of the global scaling factor (eq. 2.18). Based on this comparison, in contrast to other approaches a direct cross relaxation rate scaling according to eq. (3.14) is applied to all rates (excluding rates corresponding to fixed interproton distances, see above) whose corresponding simulated NOEs do not match the experiment. The advantage of this direct rate scaling approach is that the memory load and calculation times are reduced, since no distance matrix is needed and per iteration only one Eigenvalue decomposition (see eq. 2.17) of the matrix is necessary. Then the algorithm returns to the loop start.

In the final step of REFINE, when sufficient agreement between simulation and experiment has been reached (see section 3.7.3), the distance information is extracted from the relaxation rates (see section 3.7.4) and written to a CNS/XPLOR/XPLOR-NIH [78] compliant restraints file for MD calculations. In case a rate could not be satisfactorily determined during the process, the respective distance is estimated from the experimental NOE volume using the simple ISPA approach with an ISPA scaling factor calculated from the successfully determined restraints and their corresponding experimental peak integrals.

The distance error for each calculated distance is estimated from the maximum and minimum relaxation rates (eq. 3.15), which are calculated from the volume error during relaxation rate rescaling, or optionally from a variation of the experimental peak integrals and also of the order parameters used in the model-free spectral density classes (section 3.7.5).

3.7.1 Preprocessing of input data from NOE experiments

From a previously recorded, Fourier transformed and processed – phased, baseline corrected and optionally filtered – NOESY spectrum a list of potential signals is generated. After that the signals need to be assigned and integrated to be used as target peak integrals in REFINE. Since in 2D NOESY spectra usually two cross peaks exist for a given combination of spins, the input data has to be preprocessed by REFINE to get non-ambiguous matching data. In this preliminary step, the experimental peak integral with the least integration error is chosen from the two peaks corresponding to the same spin system, or in case of equal errors the stronger signal. Additionally, the peak position in the spectrum is stored to allow correct comparison of peak integrals in spectra with non-symmetric intensity distribution originating from finite relaxation delay.

3.7.2 NOE cross peak assignment

For the REFINE algorithm to work, experimental NOE signals that are known to be related to a certain proton-proton contact in the protein are essential. It is up to the user to choose the means to get a reasonable assignment. For the sakes of time efficiency and complete automation it is of course desirable to leave this task to more or less automatic routines. This is why for the synthetic test case described below the KNOWNOE module of AUREMOL was employed with the available sequential assignment and the final solution structure as input. For the experimental datasets KNOWNOE as well as the grid search based PEAK ASSIGN routine were used for automatic assignment.

3.7.3 Relaxation rate scaling

Depending on the matching quality of two corresponding NOE volumes in simulation and experiment, a cross relaxation rate scaling with subsequent correction of the auto relaxation rates is applied. In a first order approximation a NOE integral is proportional to the corresponding relaxation rate, so the volume ratio of an experimental and a simulated NOE can be used for scaling the corresponding rate. Using the following iteration, the non-fixed cross-relaxation rates R_{ij} corresponding to experimental signals are rescaled in each iteration step:

$$R_{ij,n+1} = a_n \cdot R_{ij,n} \tag{3.13}$$

If at iteration n for a given experimental signal integral V_{exp} the corresponding simulated integral value $V_{sim,n}$ is smaller, the ratio $a_n = V_{exp} : V_{sim,n}$ is greater than one, leading to a relaxation rate increase, $R_{ij,n+1} > R_{ij,n}$. This in turn is reflected in a larger calculated integral value $V_{sim,n+1} > V_{sim,n}$ in the next iteration, so that $1 < a_{n+1} < a_n$. From this it follows that a_n approaches unity and $V_{sim,n}$ approaches V_{exp} for $n \rightarrow \infty$. So convergence can be achieved using this iterative scheme and the same can be shown in analogy for $V_{sim,n} > V_{exp}$.

The originally employed direct volume ratio scaling in REFINE with the scaling factor a_n restrained to a range from 0.5 to 1.5 as proposed in [69] often leads to weak convergence. It enforces very strong rescaling even for not too different peak integral values as a linear dependence between peak integral and relaxation rate is assumed. This can lead to single rates dominating their local environment, preventing further convergence or even inducing divergence of the algorithm. Therefore a new approach is proposed that keeps the rate changes per iteration small.

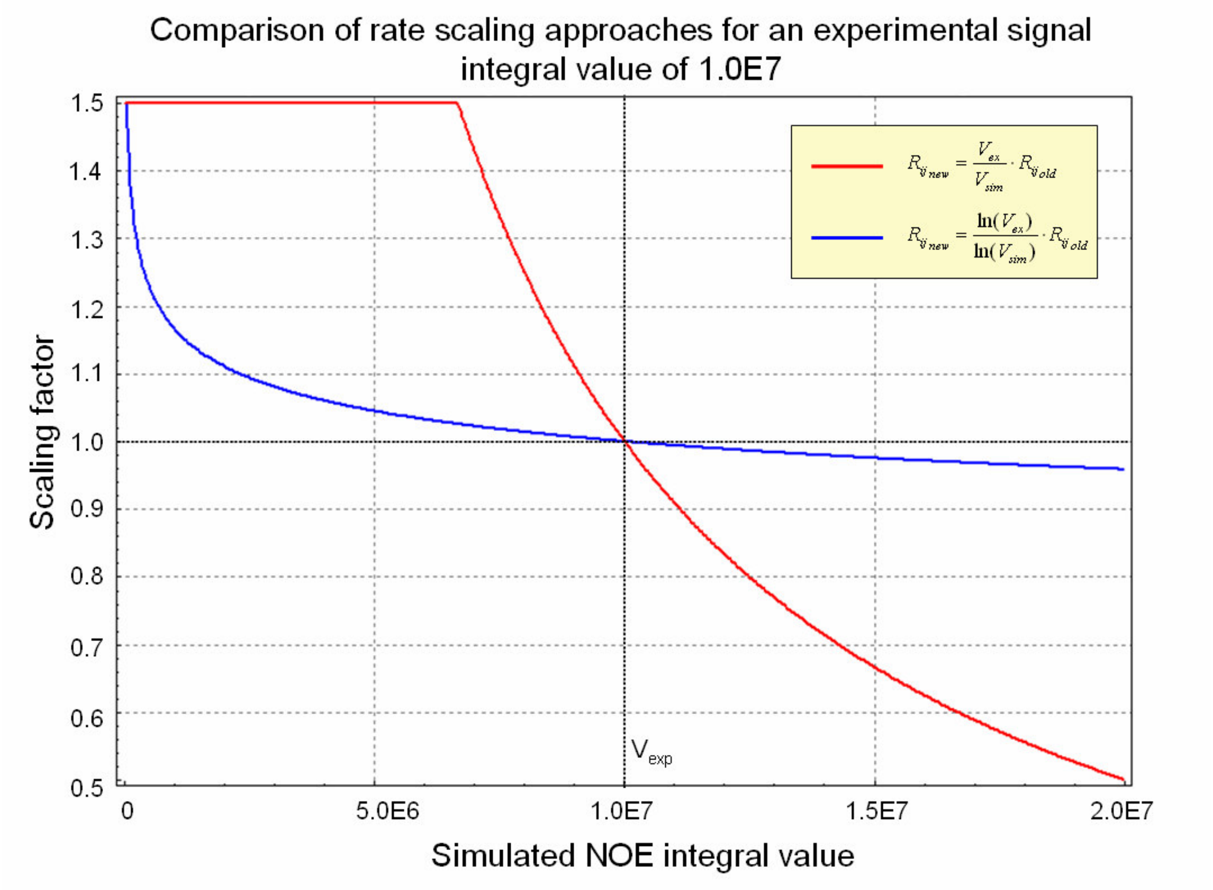

Figure 3.17 Plot of rate scaling factors for a given experimental NOE integral value of $1.0{\cdot}10^7$ and simulated integral values between $1.0{\cdot}10^0$ and $2.0{\cdot}10^7$. The old approach (red) produces large scaling factors that have to be clipped to a maximum value of 1.5 for smaller simulated NOEs. It generally introduces much more change to a given relaxation rate per iteration than the newly proposed logarithmic ratio scaling approach (blue).

As shown in Fig. 3.17, logarithmic ratio scaling has the advantage of having much less impact on a given rate than direct volume scaling, so now the respective cross relaxation rate is rescaled according to

$$R_{ij\,new} = \frac{\ln(V_{ex})}{\ln(V_{sim})} \cdot R_{ij\,old} \tag{3.14}$$

for every experimental peak, leading to an only slightly modified relaxation matrix at the end of an iteration step. This is especially important for cases when a rate is supposed to be of a large value due to the experimental data, but the starting rate is very small due to a large distance of the corresponding protons in the trial structure. In these cases logarithmic scaling results in much better convergence of the algorithm, as less change in the relaxation matrix is introduced per iteration. Also the dependence on a well defined starting structure is reduced.

If no rescaling is necessary – that is if the rescaling factor is within the interval *1* ± *p*, a user specified rate tolerance percentage for all signals – the current simulated volumes sufficiently describe the experimental data. Otherwise the algorithm loops back to the exponentiation of the relaxation matrix and the matching procedure, until sufficient agreement is reached.

For the distance error estimation from the signal integral errors the minimum and maximum relaxation rates are determined using the integral error V_{err} introduced in eq. (3.9):

$$R_{ij\,new}^{\min/\max} = \frac{\ln((1 \pm p) \cdot (V_{ex} \pm V_{err}))}{\ln(V_{sim})} \cdot R_{ij\,old} \tag{3.15}$$

3.7.4 Distance calculation

In the last step of the REFINE algorithm, since the motional models are known, the interproton distances d_{ij} can be calculated from the converged relaxation rates R_{ij} according to

$$d_{ij} = \left(\frac{R_{ij}}{f_{ij}(\tau, \omega)} \right)^{-\frac{1}{6}}. \tag{3.16}$$

with $f_{ij}(\tau, \omega)$ the distance independent contribution of the spectral density function to the relaxation rate R_{ij} (see eq. (2.12)-(2.16)). In case a given rate did not converge the respective distance is not discarded, but estimated from the associated experimental peak integral value, since an approximate restraint can still contribute to the structure calculation. Here ISPA is used according to eq. (2.21) and the scaling factor α is determined from the *n* experimental peak integrals V_i for which REFINE distances d_i were calculated:

$$\alpha = \frac{\sum_{i=1}^{n} d_i \cdot \sqrt[6]{V_i}}{n} \tag{3.17}$$

3.7.5 Distance error estimation

To provide information about the accuracy of the calculated distances reasonable error bounds have to be determined. For a single run of REFINE the experimental peak integral errors are taken and converted to minimum and maximum distances d_{min} and d_{max}, calculated from the maximum and minimum relaxation rates according to eqs. (3.15), (3.16). In an error estimate for clipped and non-converged rates the corresponding average volume contribution V_e to the experimental signal integral V_{exp} for an expected absolute distance error d_{e0} at a user defined distance d_0 can be considered:

$$\begin{aligned}
&V_0 = \alpha \cdot \frac{1}{d_0^{\;6}},\; V_{e1} = V_0 - \alpha \cdot \frac{1}{(d_0 + d_{e0})^6},\; V_{e2} = \alpha \cdot \frac{1}{(d_0 - d_{e0})^6} - V_0 \\
&V_e = \frac{V_{e1} + V_{e2}}{2} \\
&d_{\max} = \sqrt[6]{\frac{\alpha}{V_{exp} - V_e}},\; d_{\min} = \sqrt[6]{\frac{\alpha}{V_{exp} + V_e}}
\end{aligned} \tag{3.18}$$

In a more general approach, the distance errors arising from the errors in the experimental peak integral data are determined during repeated runs of REFINE where the target integrals are varied within their error bounds. From the resulting distance distributions for each NOE contact an error measure can be derived, e.g. the standard deviation of the distribution.

Analogously, the errors inherent in the spectrum simulation because of using an insufficient number of different spectral density classes for spins with different mobilities and/or only estimated order parameters S^2 in the model-free approach of Lipari and Szabo [67] for the spectral density functions can be quantified by a statistical approach. By repeating the REFINE calculation n times with a normally distributed random variation of the order parameters used, a distribution of distances for each restraint is obtained from which the mean distance d_{mean} and the standard deviation are calculated.

The target peak integrals as well as the order parameters are varied using Gaussian distributions with the widths specified by the volume errors and the expected parameter variation. That way the combined lower/upper distance error estimate, containing contributions from experiment and simulation, is obtained from the standard deviation σ_d of the resulting distance distributions,

$$\Delta d_{tot,\min/\max} = a \cdot \sigma_d\,, \tag{3.19}$$

from which lower and upper distance bounds are then calculated as

$$\begin{aligned} d_{\min} &= d_{mean} - \Delta d_{tot,\min} \\ d_{\max} &= d_{mean} + \Delta d_{tot,\max} \end{aligned} \qquad (3.20)$$

and written to the output file together with the mean distances. The weighting factor *a* could for example be determined for a desired confidence level, but has been set to $a = 1$ for the following calculations, i.e. a single standard deviation has been used as lower/upper error estimate.

3.8 Using REFINE to eliminate KNOWNOE scaling parameters

As complete automation of structure determination is the goal, it is mandatory to reduce the total number of required user defined parameters as much as possible. In KNOWNOE the general requirement of a known distance and corresponding experimental peak integral for internal ISPA scaling has been eliminated by integrating the REFINE module into the workflow (Fig. 3.18).

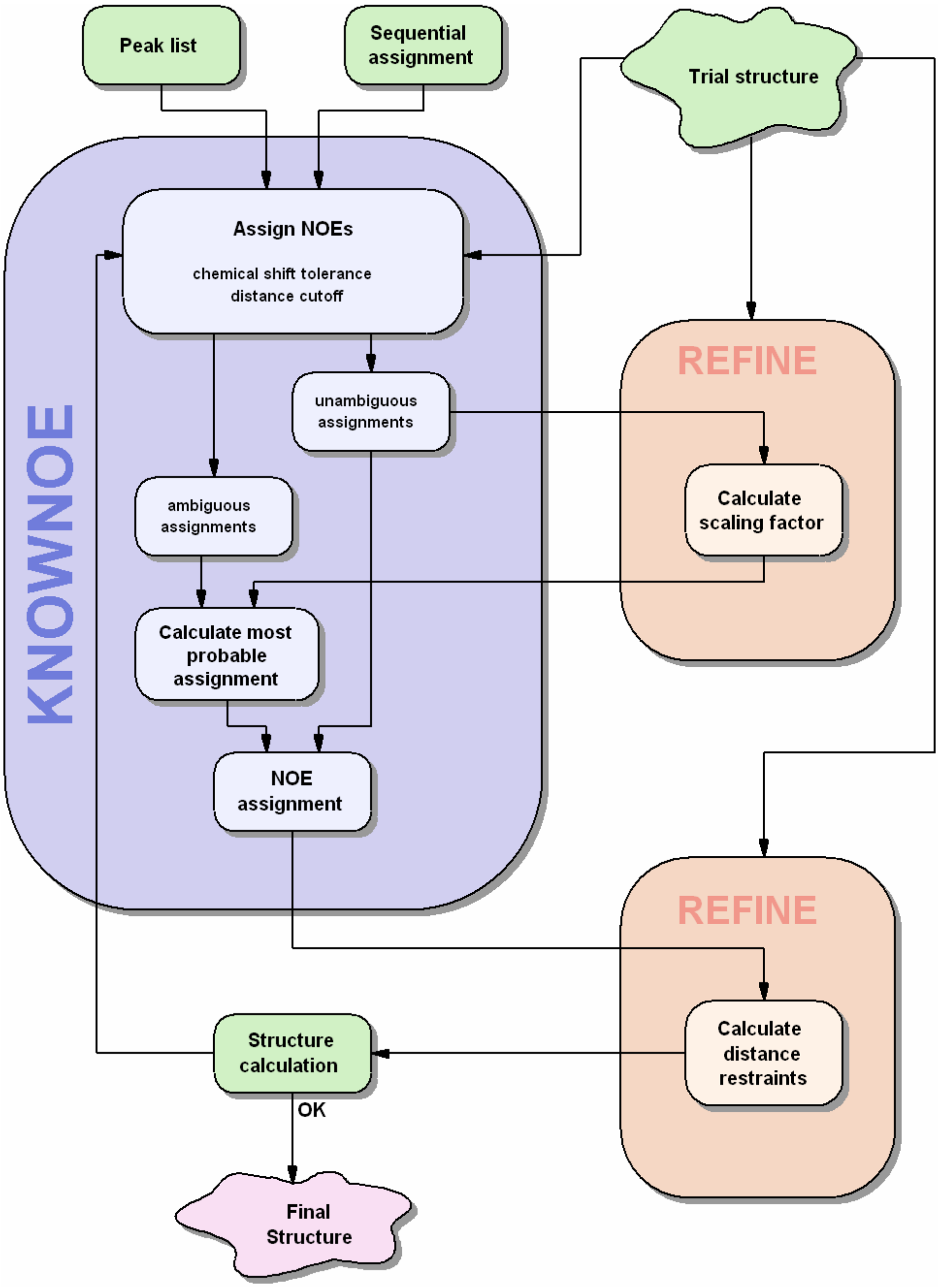

Figure 3.18 Interaction of REFINE and KNOWNOE in the structure determination workflow. During the NOE assignment phase in KNOWNOE, REFINE is used for the automatic calculation of an ISPA scaling factor required by KNOWNOE (top). After the assignment has been determined, it is used by REFINE for distance calculation (bottom).

After the first calculation step, REFINE is called to determine an ISPA scaling factor from the unambiguously assigned cross peaks according to eq. (3.17). This factor is then used in the following stages of KNOWNOE where structural information is evaluated to resolve assignment ambiguities. In the following section this KNOWNOE modification will be used in combination with REFINE to automatically determine protein structures from unassigned NOESY spectra.

3.9 Evaluation of the REFINE algorithm

First of all the REFINE algorithm has to be tested whether it works as expected. Therefore it has been applied to an ideal synthetic test spectrum. When taking the original structure used for the simulation of the dataset as REFINE input, the resulting distances must not deviate from the actual distances in the structure, otherwise this would indicate an error in REFINE. To be able to judge the performance of REFINE in a more realistic situation, a simulated dataset for a well defined trial structure had to be prepared. This artificial data was then treated like a real world dataset and used as input for the REFINE distance calculation. As the trial structure used for the creation of the test data set is known, the results can be easily compared by pairwise RMSD value calculation between the newly calculated structures based on the distances obtained by REFINE and the trial structure.

REFINE has also been applied to experimental data to evaluate the accuracy and robustness of the algorithm for real world data containing noise, artifacts and baseline distortions. Here, the quality of the resulting structures is judged by NMR R-factors, which are a good measure for the agreement between structures and experimental data.

3.9.1 Ideal synthetic test case

The published structure of TmCSP was used for spectrum simulation. Together with this structure REFINE was applied to the resulting dataset and the resulting distances as well as distances obtained using ISPA were compared to the distances to the original structure used for the simulation. Additionally, and extended strand was used as trial structure. The results in Fig. 3.19 show that the REFINE algorithm works as expected. In fact the distances obtained using the original structure were exact, with the exception of two distances that were about 5% off the original value. The cause of this slight flaw was a programming error in the internal treatment of pseudo atoms and has been fixed.

Larmor frequency	600.13 MHz
Cutoff radius	0.5 nm
Mixing time	250 ms
Relaxation delay	1.31 s
Line shape	Gaussian
Points in W1	1024
Points in W2	1024
Linebroadening	5 Hz (both dimensions)
J-correction of T2	activated
J-splitting	disabled
Max. iterations	100
Rate tolerance	5%

Table 3.6 Parameters used for the 2D NOESY spectrum simulation of TmCSP.

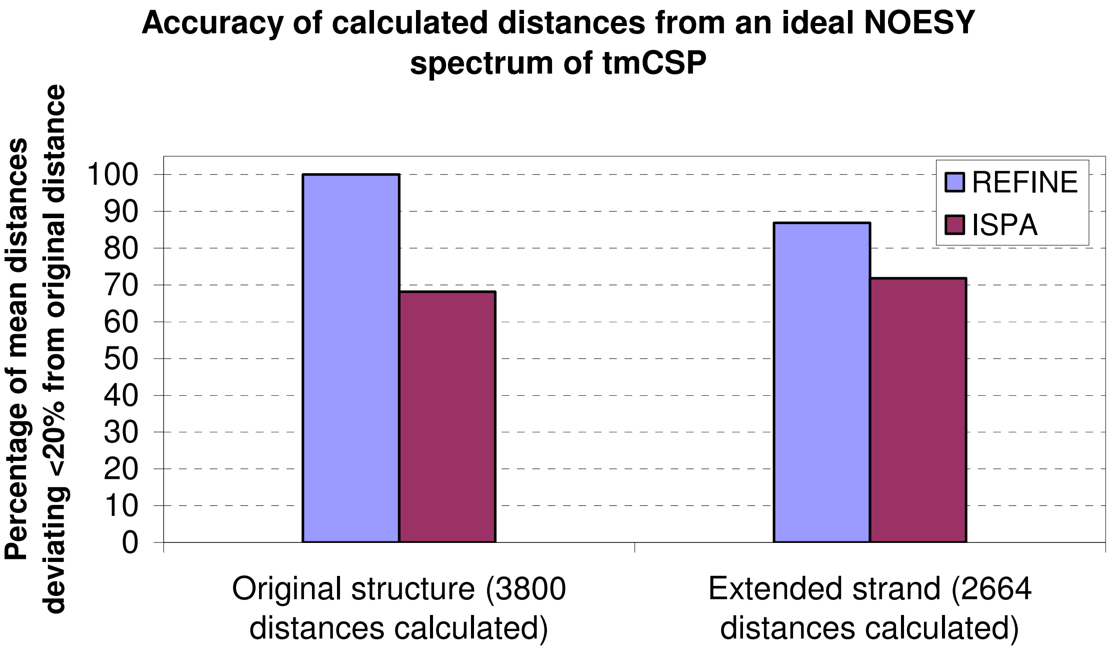

Figure 3.19 Number of calculated mean distances in relation to the total number of calculated distances with an error in respect to the original structure used for simulating the dataset.

When using an extended strand as trial structure, REFINE also manages to produce >85% of all calculated distances with an error of <20% in respect to the distances in the original structure. This indicates that structure determination in REFINE is possible starting from an extended strand and implies that the iterative application with the structure obtained from the produced distances and molecular dynamics used as input for the following run should lead to an increasingly better structure hypothesis.

3.9.2 Artificial test dataset

Since the solution structure of HPr from *Staphylococcus carnosus* had already been solved by our group [6], an experimental 2D NOESY spectrum of this protein was available for testing. As the usual artifacts and noise were present in the data, the idea was to use simulated datasets from a well defined trial structure of this protein and add different levels of noise in a reproducable way. From that point on the artificial datasets were treated like real world experimental spectra.

3.9.2.1 Preparation of the target structure for spectrum simulation

Previous to the simulation, the existing structure ensemble of HPr (Fig. 3.20) from the PDB database was run through a water refinement procedure following the standard protocol described by [76] in order to provide a structure with minimum inherent tension as a starting point. Here, a molecular dynamics simulation in explicit solvent – water in the present case – is performed along with a structure quality analysis by WHATCHECK [85] before and after the process. For the resulting structure ensemble, especially the RMS Z-scores improved considerably (table 3.7).

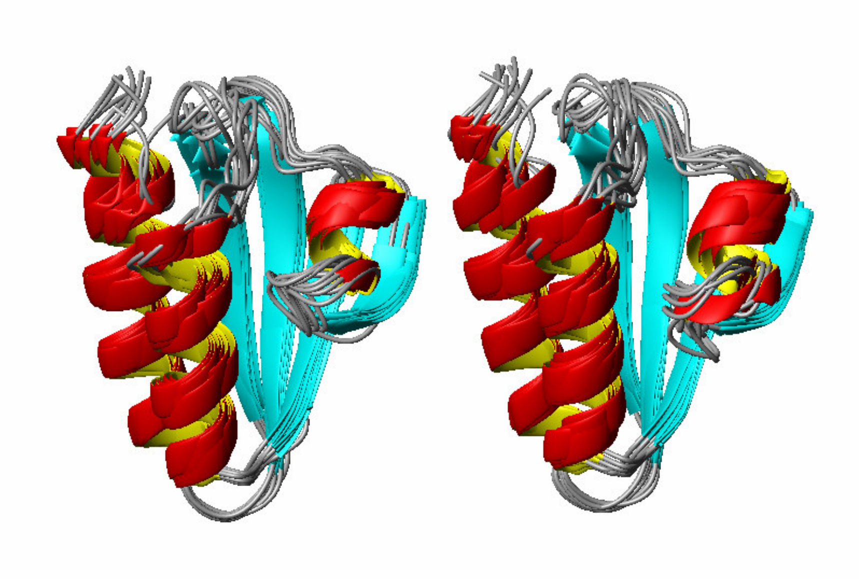

Figure 3.20 The bundle of ten HPr (*S. carnosus*) structures before (left) and after (right) the water refinement calculation

RMS Z-score	Before water refinement	After water refinement
Bond lengths	0.35 ± 0.01	0.99 ± 0.04
Bond angles	0.46 ± 0.01	0.96 ± 0.03
Omega angle restraints	0.19 ± 0.01	0.88 ± 0.05
Side chain planarity	0.10 ± 0.02	1.00 ± 0.16
Improper dihedral distribution	0.31 ± 0.01	0.97 ± 0.04
Inside/outside distribution	1.00 ± 0.02	0.98 ± 0.02

Table 3.7 Z-scores calculated by the WHATCHECK structure analysis program.

The values close to one indicate that the structures have settled in a more relaxed configuration after the water refinement, although no dramatic changes in the overall fold are observed.

From the ten calculated structures the one with the least overall energy is chosen as the starting point for the backcalculation.

3.9.2.2 Spectrum simulation using RELAX

From the water-refined trial structures, a spectrum was simulated by RELAX with the following parameters according to the experimental settings used for the original recording of the dataset:

Larmor frequency	800.13 MHz
Cutoff radius	0.5 nm
Mixing time	150 ms
Relaxation delay	2.37 s
Line shape	Gaussian
Points in W1	1024
Points in W2	8192
Linebroadening	5 Hz (both dimensions)
J-correction of T2	activated
J-splitting	disabled

Table 3.8 Parameters used for the 2D NOESY spectrum simulation of HPr from *S. carnosus*

To simulate the available experimental 2D dataset as closely as possible, the simulation parameters have been chosen to match the experimental settings. The calculation of peak splitting due to J-coupling has been neglected in the simulation, since the peak shape is not relevant for REFINE, as only peak integrals are considered.

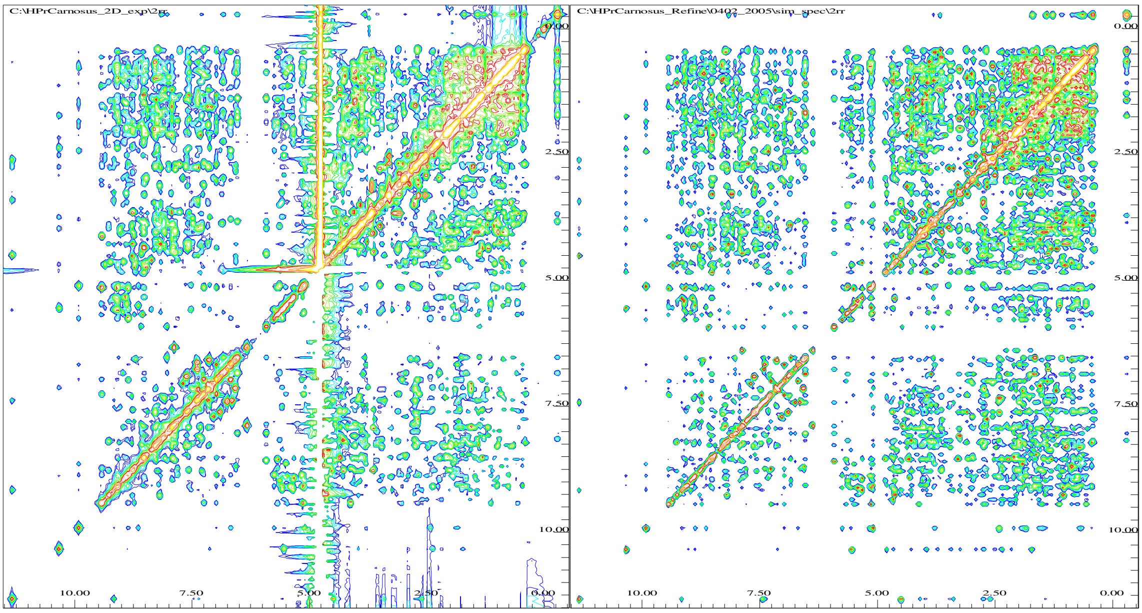

Figure 3.21 Experimental 2D NOESY spectrum of HPr from *S. carnosus* (left) and simulated spectrum from water-refined structure model (right)

After the simulation the automatically created peaklist was discarded and an automated peak picking was performed by the adaptive routine (without automatic Bayes training set generation) described above, so as to treat the simulated dataset as an experimental dataset. The resulting list of 5973 signals was automatically assigned using KNOWNOE which yielded 5500 assignments. Note that KNOWNOE assigns only one of two corresponding symmetric cross peaks and REFINE uses only one of two corresponding signals, so 2750 signals were used in the REFINE run.

3.9.2.3 Addition of Gaussian white noise

By adding different levels of noise, the robustness of REFINE in respect to the S/N ratio of the data can be judged. Four additional copies were produced from the initially prepared artificial dataset and peaklist. That way the number of signals and the assignment were the same, regardless of the present noise. Some signal locations shifted due to the noise, so for increasing noise levels less signals could be integrated.

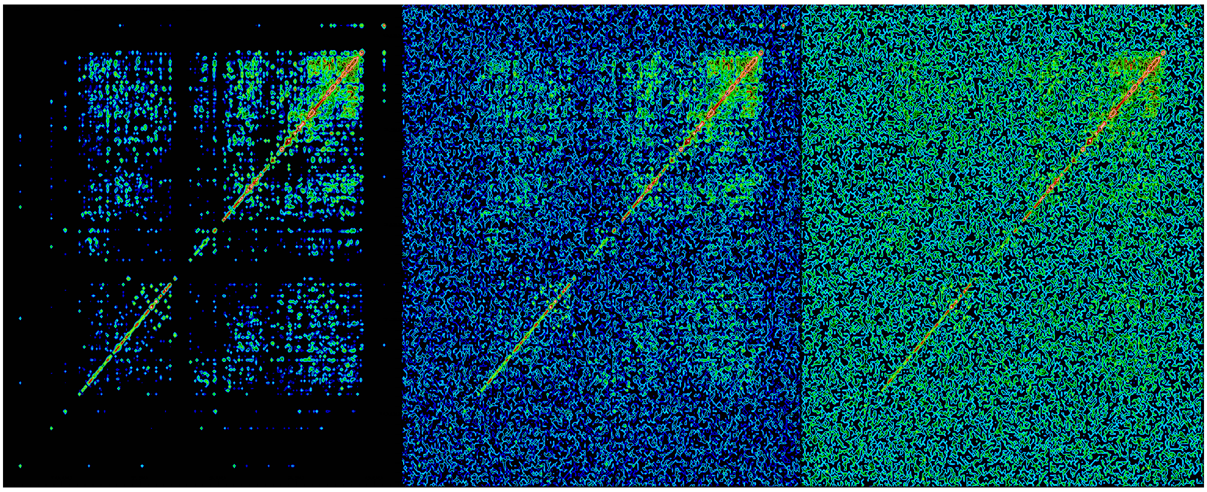

Figure 3.22 Simulated 2D NOESY spectra of HPr from *S. carnosus* with different levels of noise added (*left: 0% middle: 50% right: 100%*). For [ppm]-scale see Fig. 3.21 (right).

The amount of noise added was set in relation to the signal of the HN 38/HA 38 NOESY reference cross peak corresponding to a distance of 2.71Å. So 100% noise translate to a noise standard deviation equal to the maximum intensity of 276442 of this reference peak.

3.9.2.4 Preparation of peak integral data

The five datasets with added noise ranging from 0%-100% in steps of 25% had the signals integrated by the extended segmentation routine (→ 3.2). As expected, the number of integrable peaks decreases with the amount of noise added because of the slight shifting of signal maxima in the dataset.

HPR *S. carnosus* artificial test dataset	**0% noise[*)]**	**25% noise[*)]**	**50% noise[*)]**	**75% noise[*)]**	**100% noise[*)]**
unique signals integrated	2715	1998	1582	1289	1112
integrated volume of ref. signal[*)]	737609	655205	828513	750201	1841755
alpha scaling from ref. signal[*)] for ISPA	25,76	25,25	26,26	25,83	30

[*)] 100% noise ≘ noise stdev = 276442 (max. intensity of HPr *S. carnosus* HN 38/HA 38 NOESY reference cross peak corresponding to a distance of 2.71Å)

Table 3.9 Number of integrable peaks depending on the amount of noise present

3.9.2.5 Trial structures used for testing

Two different structure models have been used as trial input for the distance calculation in the HPr (*S. carnosus*) test case: The original structure used for the creation of the artificial dataset and a completely unfolded extended strand structure.

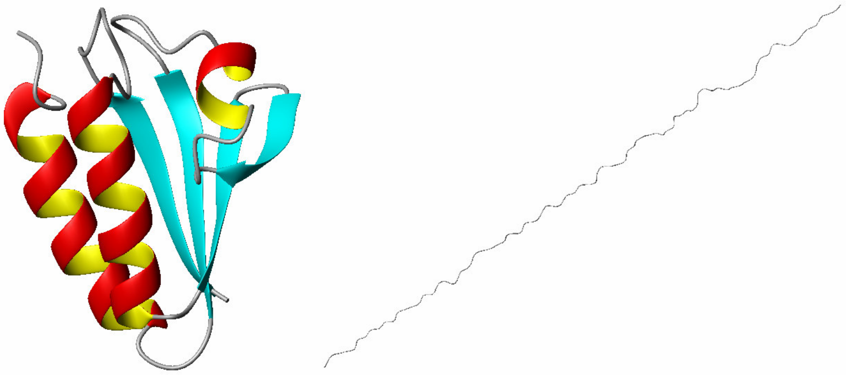

Figure 3.23 Original HPr (*S.carnosus*) structure used also for the simulation of the test dataset (*left*) and extended strand structure (*right, not to scale*)

The distances calculated for these two structures allow to judge the performance of REFINE on the one hand for already well defined structure hypotheses and on the other hand for completely undefined structure hypotheses.

HPR *S.carnosus* artificial test dataset	**0% noise*)**	**25% noise*)**	**50% noise*)**	**75% noise*)**	**100% noise*)**
refined distances ext**)	2873	2140	1715	1398	1221
refined distances best**)	2939	2186	1755	1423	1248

*) 100% noise ≘ noise stdev = 276442 (max. intensity of HPr *S. carnosus* HN 38/HA 38 NOESY reference cross peak corresponding to a distance of 2.71Å)

**) in REFINE all magnetically equivalent spins possessing the same chemical shift yield individual distance restraints if they are within the cutoff radius

Table 3.10 Number of calculated restraints depending on noise level

3.9.2.6 Accuracy of calculated proton proton distances

The distance restraints calculated by REFINE, as well as the ISPA restraints calculated in parallel for comparison show an expected decrease in accuracy with increasing noise level (Fig. 3.24). For the ISPA restraints this decrease is more pronounced and the REFINE restraints are generally more accurate; even the results from the extended strand input structure yield significantly more restraints with a deviation from the original distances of the structure used for the dataset generation below 20%.

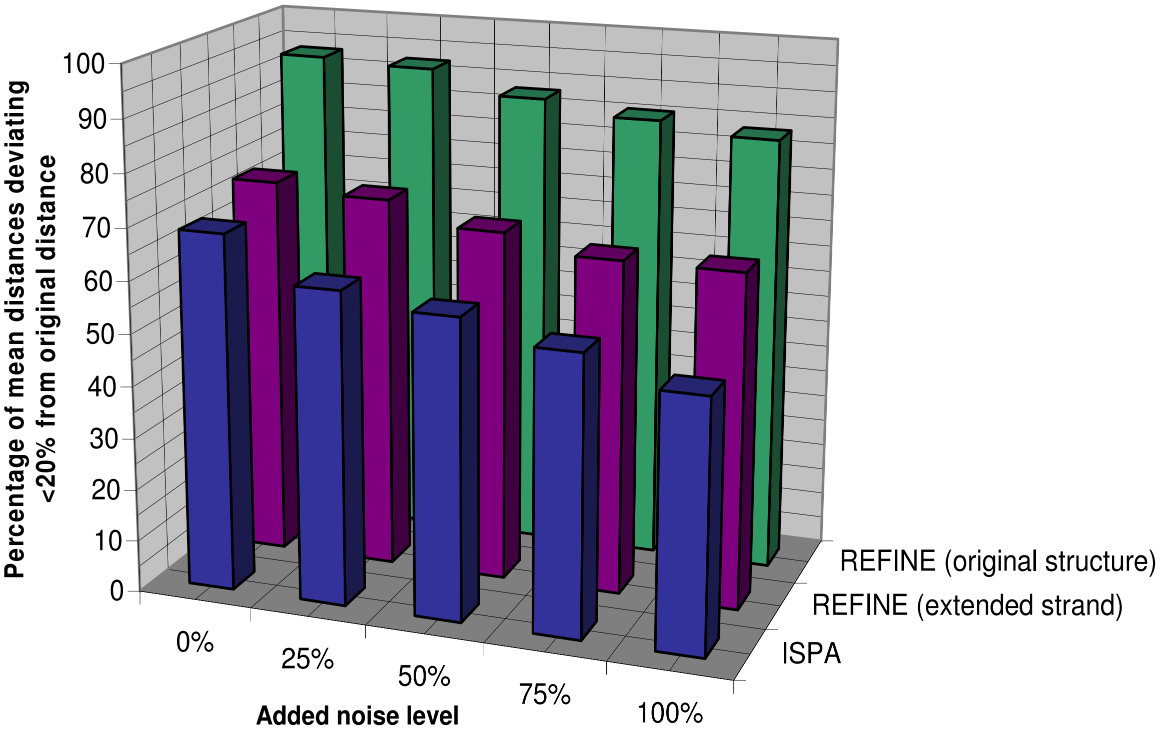

Figure 3.24 Number of calculated mean distances in relation to the total number of calculated distances by REFINE with error below twenty percent, depending on the noise level and the starting structure, compared to ISPA calculated distances. The distance deviations were calculated in respect to the distances in the original structure used for the simulation.

The fact that during the automatic peak picking procedure several overlapping peaks can be detected as one single signal explains why the REFINE calculation using the original structure as input does not yield 100% correct distances.

3.9.2.7 Error estimates

The error estimates derived from noise and integration errors alone are minimum error estimates and cover in this synthetic test case at least one third of the distances calculated from the extended strand structure and at least more than 64% of the distances calculated from the original structure.

HPR *S. carnosus* artificial test dataset	**0% noise*)**	**25% noise*)**	**50% noise*)**	**75% noise*)**	**100% noise*)**
% within error range**) ext*)	46,26	39,63	37,26	35,19	35,63
% within error range**) best*)	72,88	71,04	67,64	66,41	64,26

*) 100% noise ≙ noise stdev = 276442 (max. intensity of HPr *S. carnosus* HN 38/HA 38 NOESY reference cross peak corresponding to a distance of 2.71Å)

**) error range calculated automatically from volume integration error (minimum error estimation) + 5% rate tolerance, no user defined error, no target integral variation

Table 3.11 Number of calculated distances within estimated error range with respect to the corresponding distances in the original structure. Here ISPA results are not shown because of the arbitrary nature of possible error estimates.

Note that for this test run no variation of the experimental signal integrals was used for error estimation.

3.9.2.8 Resulting structures from molecular dynamics calculations

For each of the cases with varying noise levels restrained molecular dynamics calculations were performed by CNS using the REFINE restraints. From the 200 structures obtained for each case, the ten best in terms of total energy were selected for evaluation without further water refinement. Calculating pairwise RMSD values between the newly calculated structures and the lowest energy original structure after water refinement allows to judge the accuracy of the REFINE restraints.

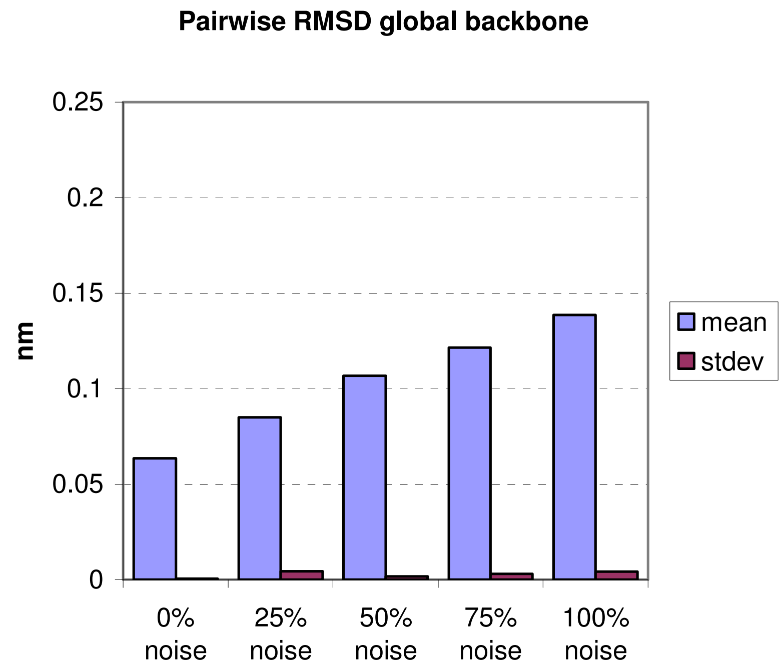

Figure 3.25 Averaged pairwise RMSD values (backbone atoms) between the ten lowest energy structures determined from the MD and the water refined structure with the lowest overall energy used for the simulation of the test data set. The restraints were calculated by REFINE starting from an extended strand trial structure for different noise levels in the peak data.

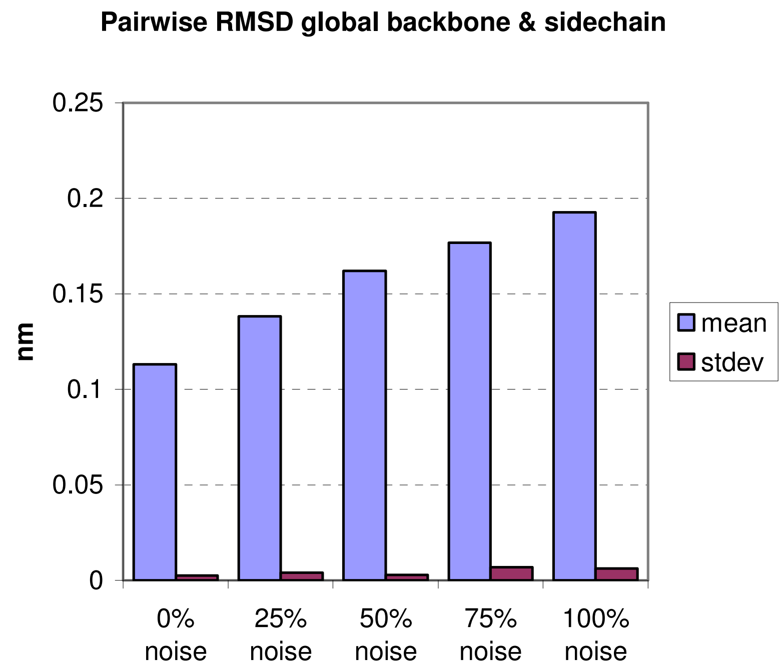

Figure 3.26 Averaged pairwise RMSD values (backbone & sidechain atoms) between the ten lowest energy structures determined from the MD and the water refined structure with the lowest overall energy used for the simulation of the test data set. The restraints were calculated by REFINE starting from an extended strand trial structure for different noise levels in the peak data.

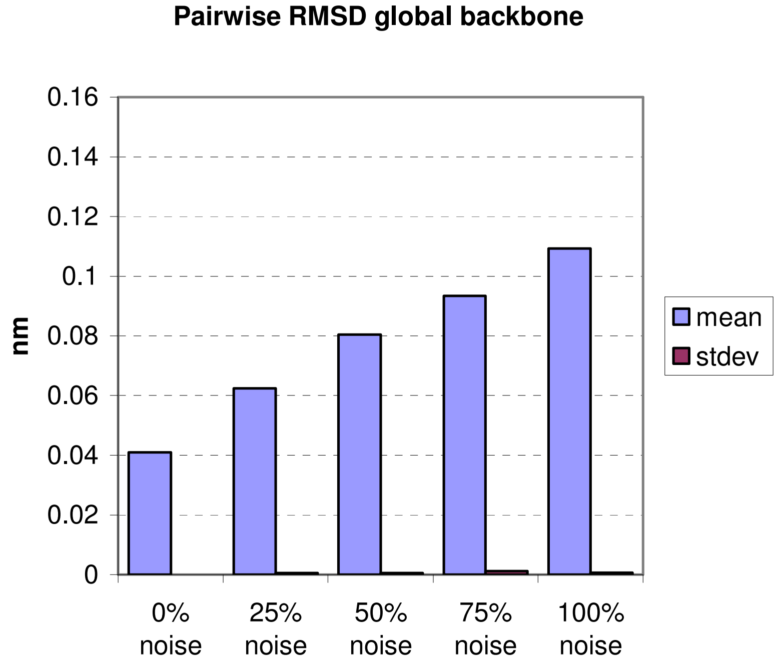

Figure 3.27 Averaged pairwise RMSD values (backbone atoms) between the ten lowest energy structures determined from the MD and the water refined structure with the lowest overall energy used for the simulation of the test data set. The restraints were calculated by REFINE starting from the original structure for different noise levels in the peak data.

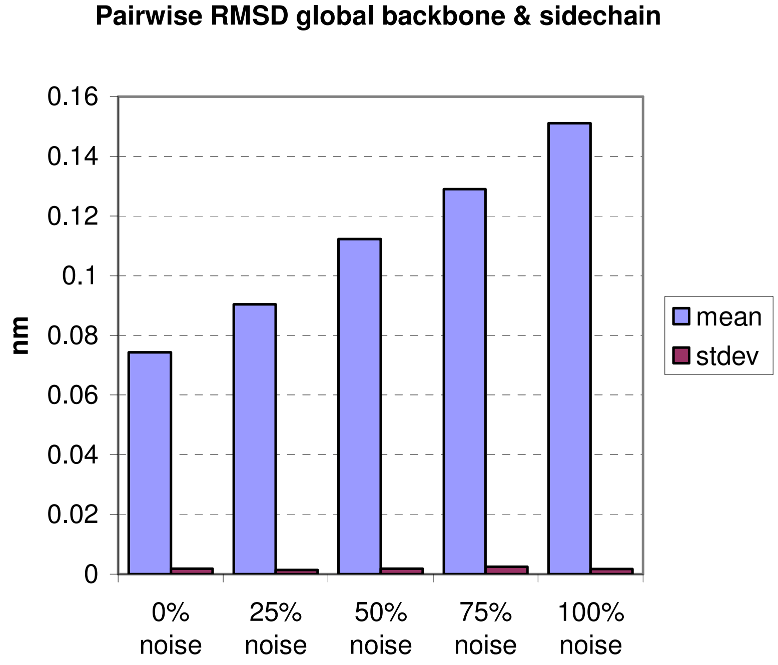

Figure 3.28 Averaged pairwise RMSD values (backbone & sidechain atoms) between the ten lowest energy structures determined from the MD and the water refined structure with the lowest overall energy used for the simulation of the test data set. The restraints were calculated by REFINE starting from the original structure for different noise levels in the peak data.

A linear dependence of the mean RMSD values for the increasing noise in the data set and the associated decreasing number of restraints is observed for both trial structures.

The influence of the trial structure on the resulting distance restraints is shown by the comparison of the mean RMSD values:

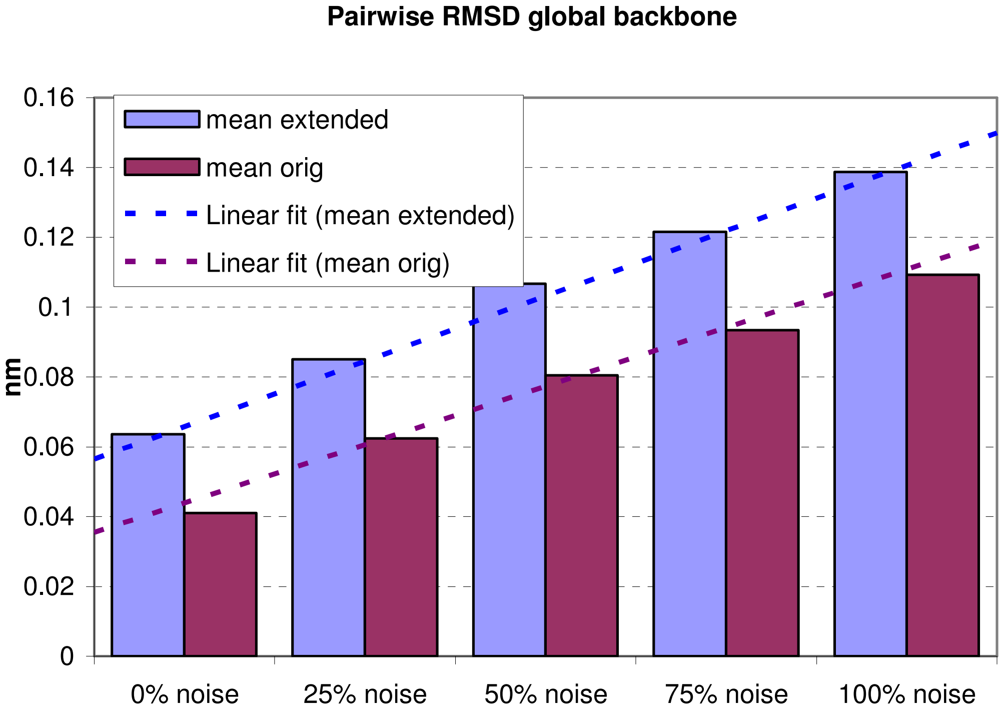

Figure 3.29 Comparison of the averaged mean pairwise RMSD values (backbone) for an extended strand (blue) and the original structure (red) as input trial structure for REFINE.

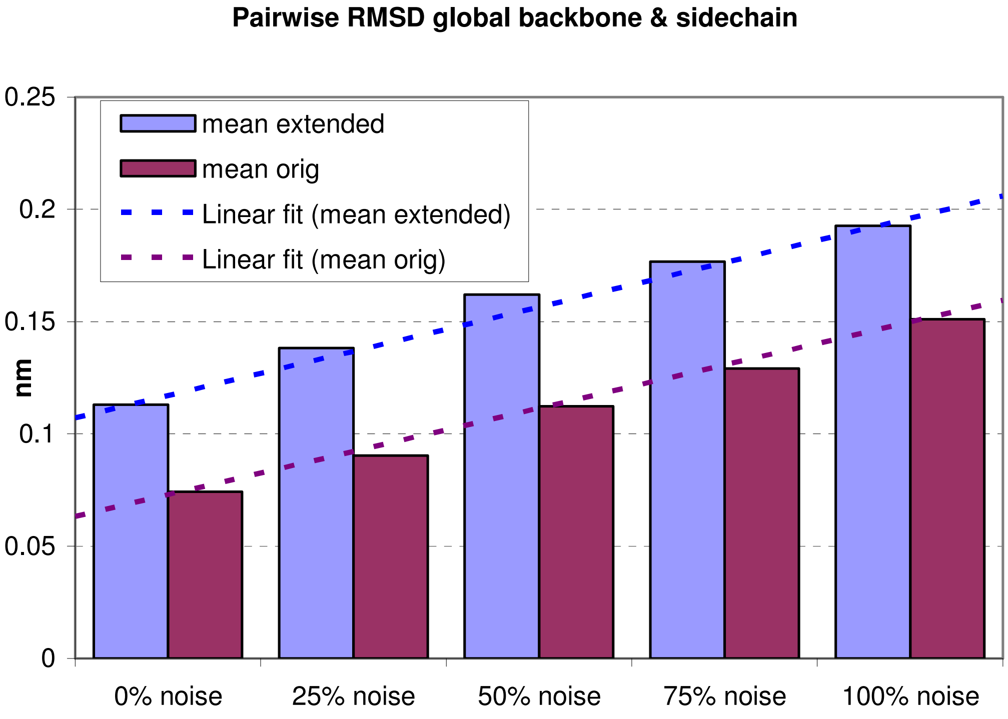

Figure 3.30 Comparison of the averaged mean pairwise RMSD values (backbone & sidechain atoms) for an extended strand (blue) and the original structure (red) as input trial structure for REFINE.

As the spatial environment of the protons in the original structure is closer to the conditions in the folded protein than in the extended strand, the results from the relaxation matrix analysis become more accurate and the resulting structures deviate less from the original structure. However, also for the extended strand trial structure fairly accurate distances and structures are calculated.

3.9.3 Results for experimental datasets

3.9.3.1 2D NOESY of HPr (*S. carnosus*) with unknown NOESY assignment

To find out qualitatively whether it is a viable way to persue an iterative automated structure determination approach using REFINE distance calculation combined with KNOWNOE signal assignment, the procedure outlined in Fig. 3.31 has been applied to experimental data.

An existing 2D NOESY dataset of the HPr protein from *S. carnosus* together with the sequential assignment from [6] was used.

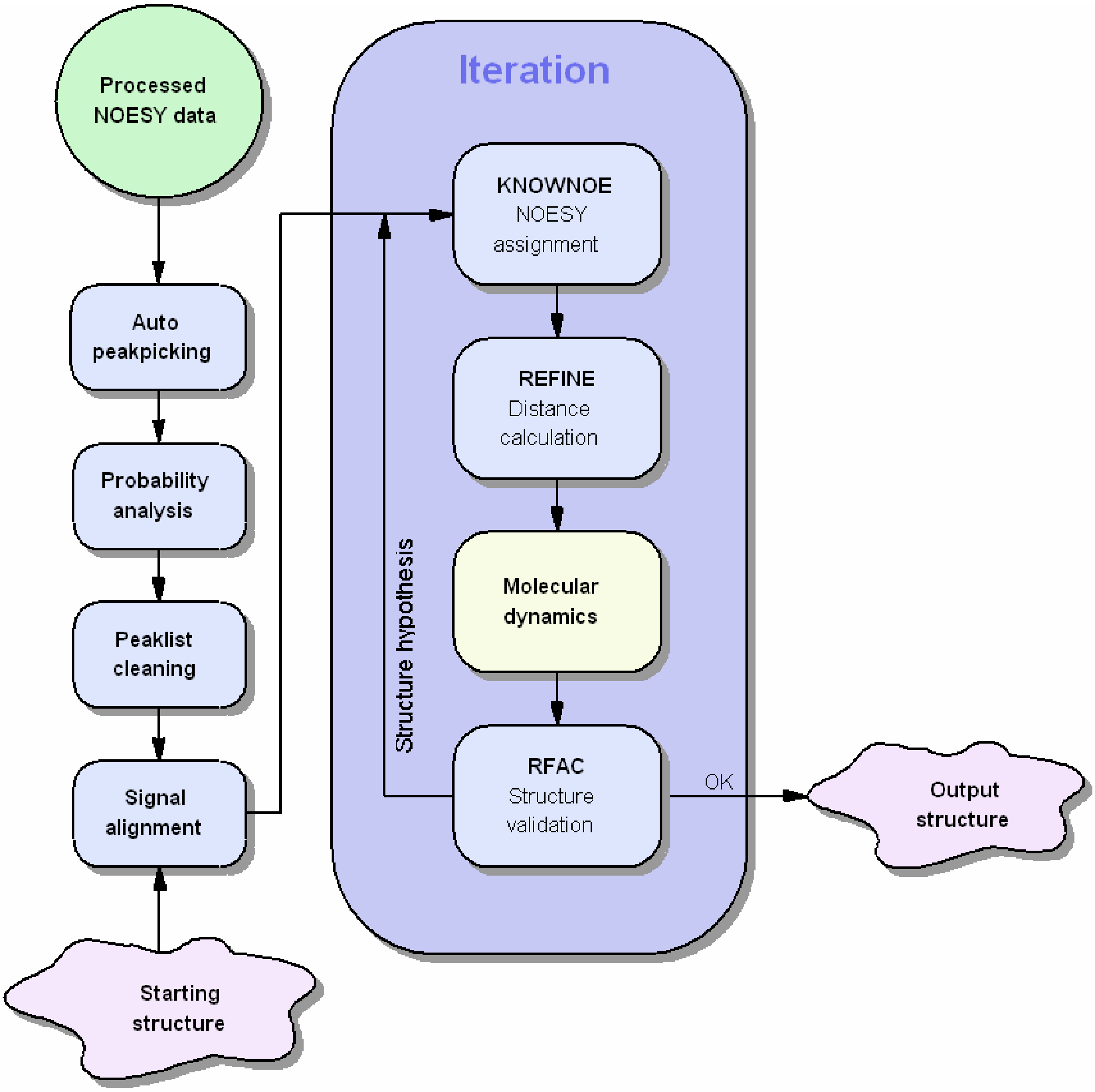

Figure 3.31 Workflow for the automated structure determination with unknown NOESY assignment from experimental NOESY data using KNOWNOE and REFINE as incorporated in AUREMOL.

Since the spectrum contained strong artifact signals (Fig. 3.32), peakpicking was performed by the adaptive routine (without automatic Bayes training set generation) described above to get a signal list. It succeeded in reducing the initial number of wrongly picked signals, especially in the center region of the water artifact, but still a large number of artifact signals remained in the list (Fig. 3.33) of 8116 potential signals.

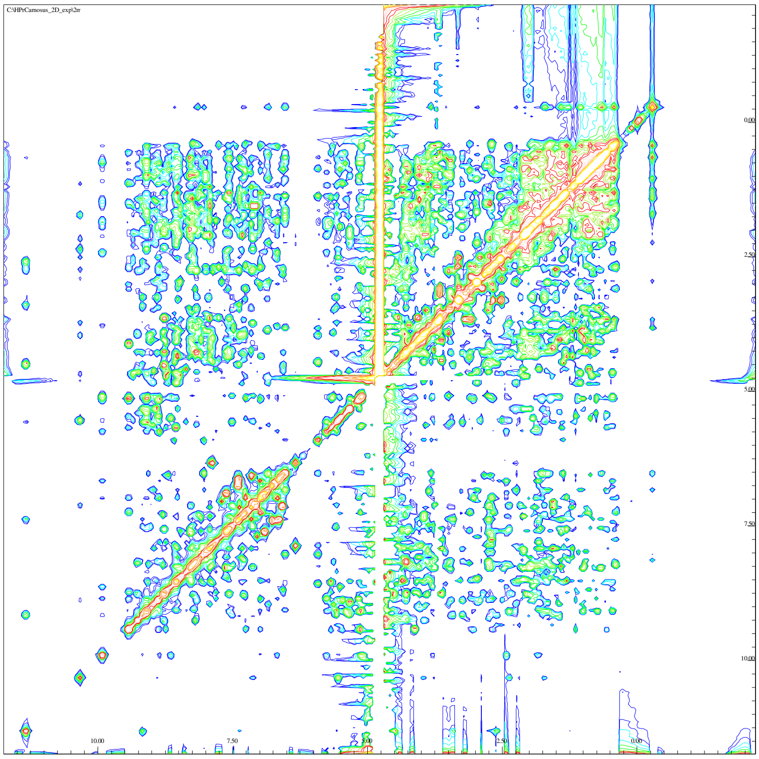

Figure 3.32 Experimental NOESY data of HPr (*S. carnosus*) containing noise and artifacts

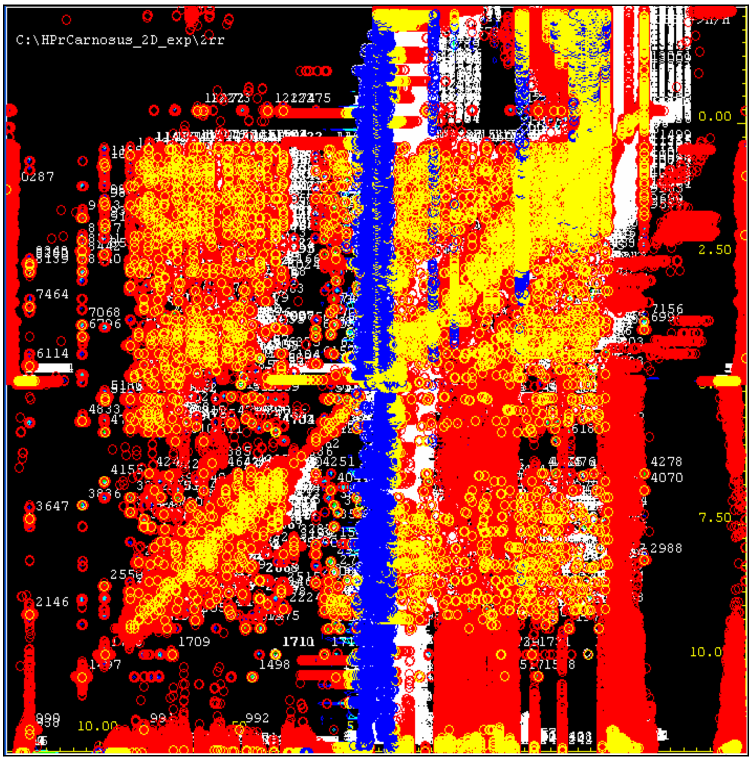

Figure 3.33 In the adaptive peak picking routine the local noise levels in the spectrum are used to calculate a local threshold. Datapoints generally passing the picking criteria are colored in red, yellow indicates signals that are kept as they coincide with the acceptable (red) datapoints, while blue marks signals that are discarded as one or both of the picking criteria are violated at their position.

Using the peak probability tool [36;37] of AUREMOL, the *Bayesian* probabilities for each signal to be artifact or not were calculated using the default values and the manually chosen signal and noise areas shown in Fig. 3.34. Keeping all signals with a probability value of at least 0.8 reduced the number of potential signals to 2665.

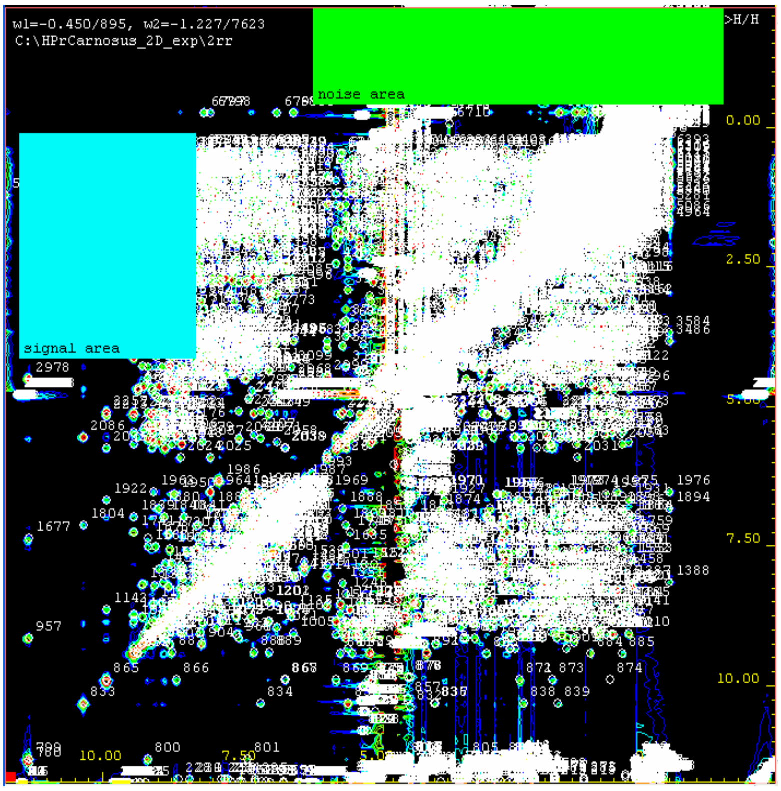

Figure 3.34 For the calculation of the Bayesian peak probabilities a noise (*green*) and a signal area (*cyan*) have to be specified.

Finally removing all peaks outside of the chemical shift bounds given by the diagonal signals yielded a list of 2598 signals (Fig. 3.35).

Starting with an extended strand structure the known sequential assignment was adjusted to the signal list using the AUREMOL tool PEAK ASSIGN (Kirchhöfer et al., to be published) that assigns signals based on a grid search procedure. Then the peak integrals including error estimates were calculated. The first NOESY assignment was produced by KNOWNOE, which generally assigns only one of two corresponding symmetric cross peaks. From the resulting assigned peaks the first distance restraints were calculated by REFINE.

With these distance restraints and additional TALOS restraints, a MD run with CNS was performed. From the resulting ensemble of 200 structures the ten with the least overall energy were chosen as solution.

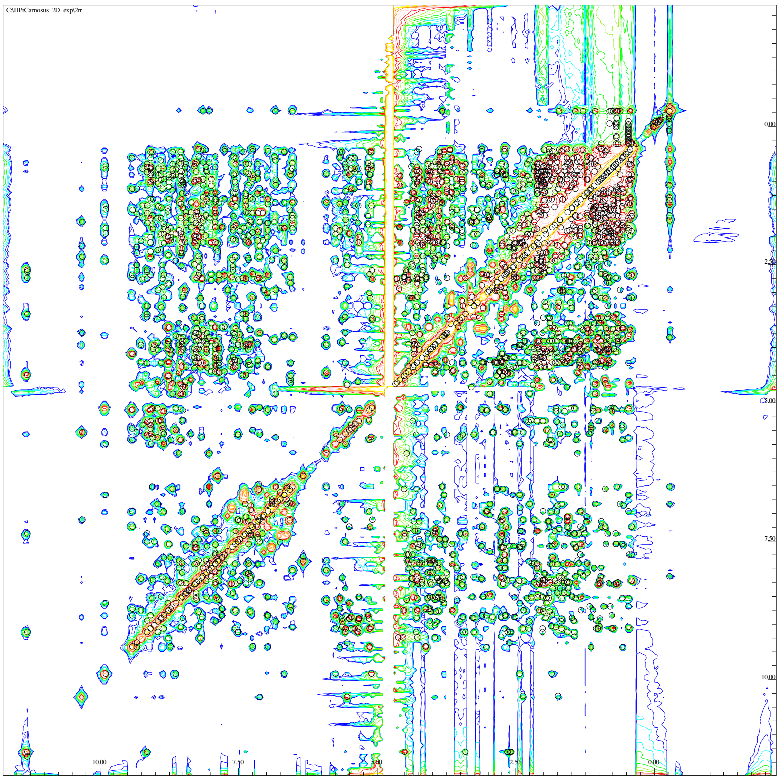

Figure 3.35 Spectrum of HPr with remaining signals (black circles) used for structure determination.

The top ten energy ensemble was then used again as input for KNOWNOE to generate an improved NOE assignment.

Iteration	Number of assigned signals	ppm limit F1	ppm limit F2	lower prob. limit	distance limit [nm]	mutual inf. prob. limit	'All' flag
1	739	0.015	0.015	0.95	100.00	0.20	-
2	985	0.015	0.015	0.95	1.50	0.10	-
3	953	0.015	0.015	0.95	1.25	0.10	-
4	996	0.015	0.015	0.95	1.00	0.05	-
5	1099	0.015	0.015	0.95	0.75	0.01	-
6	1256	0.020	0.020	0.95	0.75	0.01	-
7	1410	0.020	0.020	0.95	0.75	0.01	•
8	958	0.020	0.020	0.95	0.75	0.01	-
9	1384	0.020	0.020	0.95	0.75	0.01	•

Table 3.12 KNOWNOE settings for the different iterations. The corresponding structures are shown in Fig. 3.36.

Again distance restraints were determined using REFINE and another 200 structures were calculated using CNS and the best ten in terms of total energy were selected. Iterating this process for a total of nine times reveals a gradual convergence of the results towards a structure hypothesis with the correct fold:

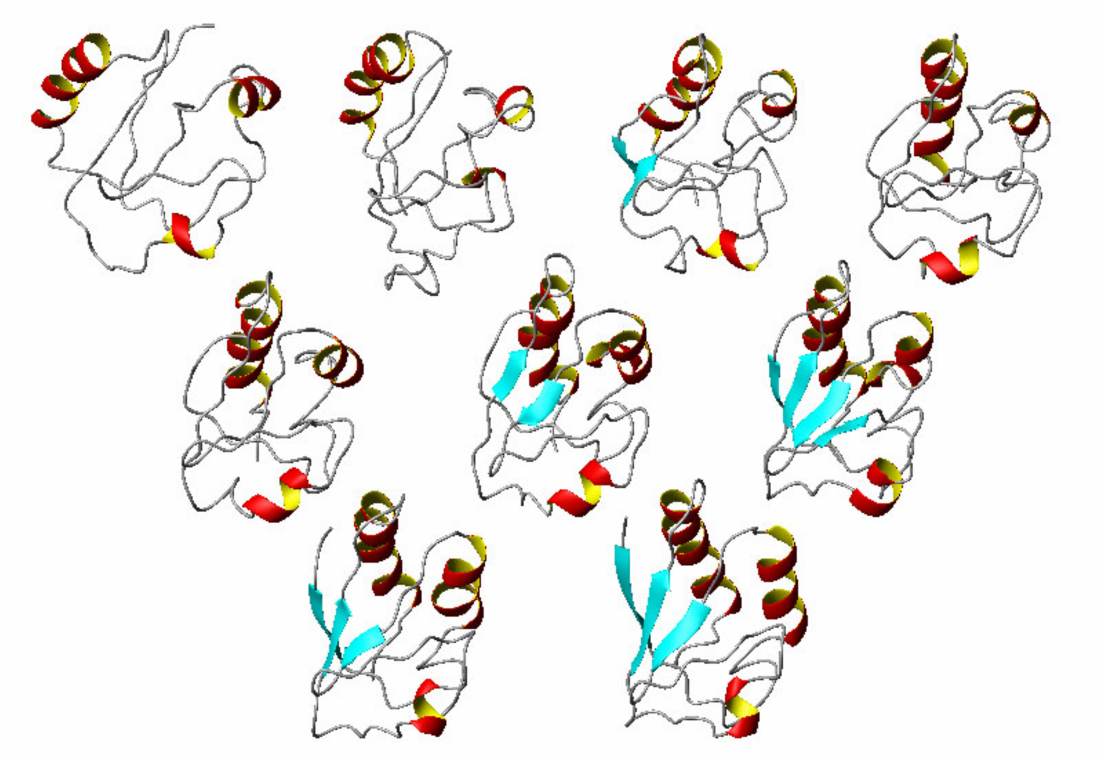

Figure 3.36 Lowest overall energy structure after each of the nine iterations. (top-left: structure after iteration one, bottom-right: structure after iteration nine)

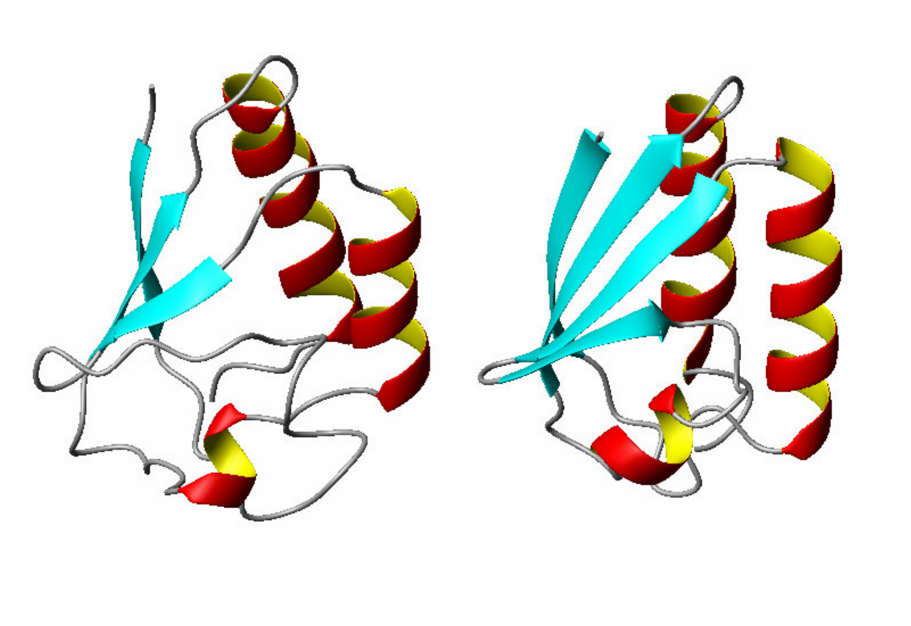

Figure 3.37 Structure calculated from a single 2D NOESY spectrum after nine iterations of KNOWNOE/REFINE (left) using only REFINE restraints and TALOS predicted angle restraints, lowest energy water refined structure from PDB ensemble (right) originally solved using NOE, dihedral angle and H-bond restraints.

Comparing the calculated structures to the structure from the PDB ensemble with the lowest overall energy after water refinement shows that the correct global fold has been found after iteration six and most secondary structure elements are present and located at the correct position after iteration nine.

This result indicates that the combination of KNOWNOE and REFINE is generally capable of solving structures automatically from a limited amount of available, automatically pre-processed NOESY data without a priori available structural information. Further extending the iteration count is expected to yield even more accurate structure hypotheses.

3.9.3.2 2D & 3D NOESY data of RalGDS-RBD with unknown NOESY assignment

The same procedure as above has been employed for the determination of the structure of the Ras binding domain – RBD – of RalGDS. Here a 800 MHz 2D NOESY and a 600 MHz 3D ^{15}N-NOESY-HSQC spectrum were available as experimental input for REFINE as well as 104 calculated TALOS and 52 experimental H-bond restraints. The spectra were automatically assigned in KNOWNOE using an extended strand as starting trial structure and subsequently the respective best structure in terms of energy of the previous iteration.

Iteration	assign limits [ppm]					lower prob. limit	dist. limit [nm]	mutual information probability limit	use mutual inform.	assign all peaks	ass. to master list	input structure
	F1 (2D)	F2 (2D)	F1 (3D)	F2 (3D)	F3 (3D)							
1	0.02	0.02	-	-	-	0.95	100	0.2	●	-	●	ext. strand
2	0.01	0.01	-	-	-	0.95	1.5	0.1	●	-	●	best from previous it.
3	0.015	0.015	-	-	-	0.95	1.25	0.1	●	-	●	best from previous it.
4	0.015	0.015	-	-	-	0.95	1	0.05	●	-	●	best from previous it.
5	0.015	0.015	-	-	-	0.95	0.75	0.01	●	-	●	best from previous it.
6	0.015	0.015	0.1	0.5	0.02	0.95	0.75	0.01	●	●	●	best from previous it.

Table 3.13 KNOWNOE settings for the different iterations. The assign limits per dimension are given in ppm, the lower probability limit, distance limit and mutual information settings are set to increase the number of correctly assigned signals with increasing iteration count.

After a distance restraint set was calculated a molecular dynamics run was performed yielding 100 structures. These were analyzed in AUREMOL to remove violated restraints and the cleaned restraint list was used in a second MD run that yielded the final structures for the given iteration. In the first six iterations only the 2D NOESY data was assigned by KNOWNOE and used for distance restraint calculation in REFINE.

Iteration	assigned signals 2D	assigned signals 3D	average RMSD to mean [nm]
1	483	-	0.22
2	964	-	0.12
3	877	-	0.13
4	884	-	0.11
5	965	-	0.22
6	1442	392	0.06

Table 3.14 Number of assigned signals and C^{α} RMSD values in respect to the mean structure for the different iterations.

After seven iterations the calculated structures had reached good agreement with the known X-ray structure [88]:

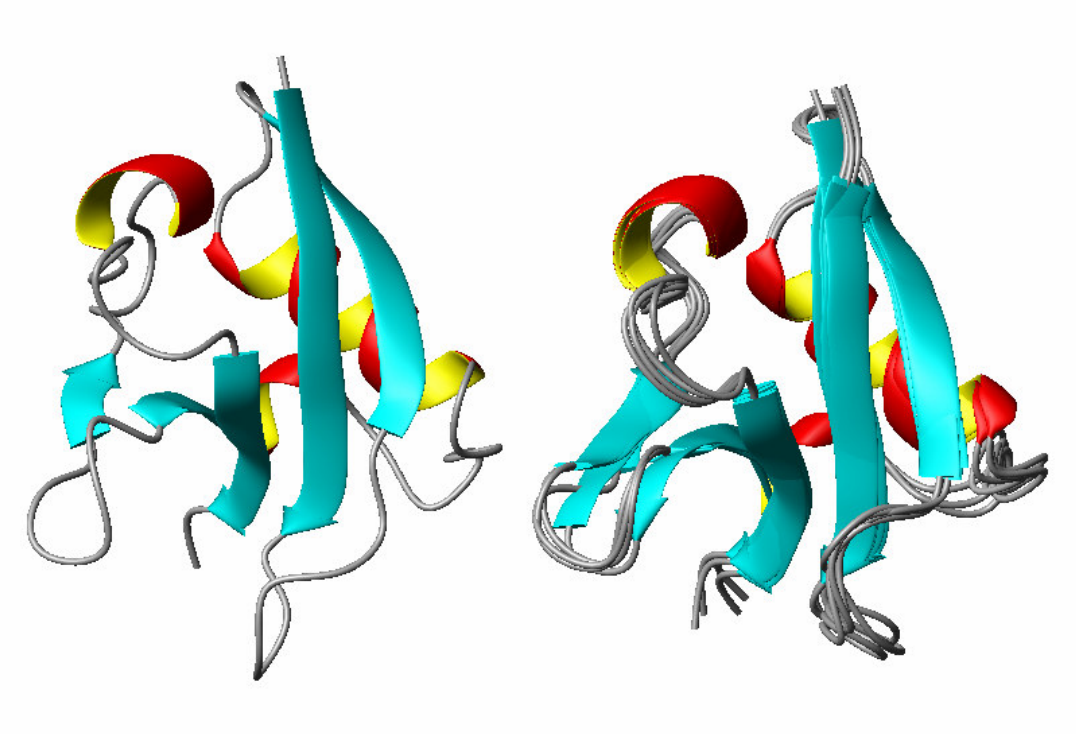

Figure 3.38 X-ray structure of RalGDS-RBD (left) compared to the four lowest energy water refined structures after seven iterations (right).

	X-ray	KNOWNOE / REFINE
R-factor	0.325	0.322
Pairwise RMSD [nm]	0.19	

Table 3.15 R-factors and pairwise RMSD for X-ray and the calculated structures after seven iterations

The results in table 3.15 show that the combined iterative application of KNOWNOE and REFINE to one 2D and one 3D NOESY dataset yields structures quantitatively comparable to X-ray.

3.9.3.3 2D NOESY data of HPr (*S. aureus*) with known NOESY assignment

A 2D NOESY spectrum with known NOESY assignment was the starting point for the following REFINE testcase with the emphasis on evaluating the statistical error estimation approach in comparison to an automatic ISPA approach. The data was prepared using adaptive peak picking with subsequent (non-automatic) Bayesian probability analysis and additional manual peak list optimization. Since the available assignment is complete, the distance information contained in the spectrum is expected to overdetermine the structure in the molecular dynamics simulation, thus limiting both the ability to judge the effect of the distance error estimation and the benefits of accurate distance determination. So to emulate an early stage of the protein structure determination process where only a small amount of distance information is available, about 75% of the signals were randomly deleted. The resulting signal list contained 315 peaks, of which 271 were automatically assigned by the PEAK ASSIGN module using grid search.

Distances were calculated using REFINE and the statistical error estimation approach. To this end, the motional properties of the protein were modelled in the following way, assuming an in general increased mobility of sidechain protons relative to backbone protons:

Proton type	Spectral density class
HA, HN	Lipari (Backbone[*)], S^2= 0.85 ± 0.15)
Methyl protons	Fast_jump
Ring protons	Slow_jump
Other protons	Lipari (Sidechain[†)], S^2 = 0.65 ± 0.25)

Table 3.16 Spectral density classes and order parameters assigned to different protons

	HA, HN	Methyl	Ring	Other
HA, HN	Lipari_backbone	Fast_jump	Slow_jump	Lipari_sidechain
Methyl	Fast_jump	Fast_jump	Fast_jump	Fast_jump
Ring	Slow_jump	Fast_jump	Slow_jump	Slow_jump
Other	Lipari_sidechain	Fast_jump	Slow_jump	Lipari_sidechain

Table 3.17 The spectral density classes defined in table 3.16 are assigned to the different proton-proton combinations as shown here.

For every proton-proton combination described via a Lipari/Szabo model-free spectral density function, a normally distributed variation of the order parameter was carried out; also every target peak integral was varied normally distributed within the error bounds determined during signal integration. For the calculation a number of 50 independent variations of the order parameters for the backbone as well as the sidechain spectral densities was chosen, as for the target integrals. That way, a total number of 2500 distance restraint files were generated, from which distributions consisting of 2500 (if the distance results converged for each run, which was the case for all but ten restraints; among these ten restraints the minimum number of converged distances was 1250) distances were created (Fig. 3.39). Then the respective mean value and standard deviation for each restraint were chosen as distance estimate and error bound for the final MD restraint file.

[*)] Correlation time τ_c = 5.62ns and backbone order parameter adopted from measurements on HPr from S.carnosus [89]

[†)] Sidechain order parameter taken from XPLOR manual [90]

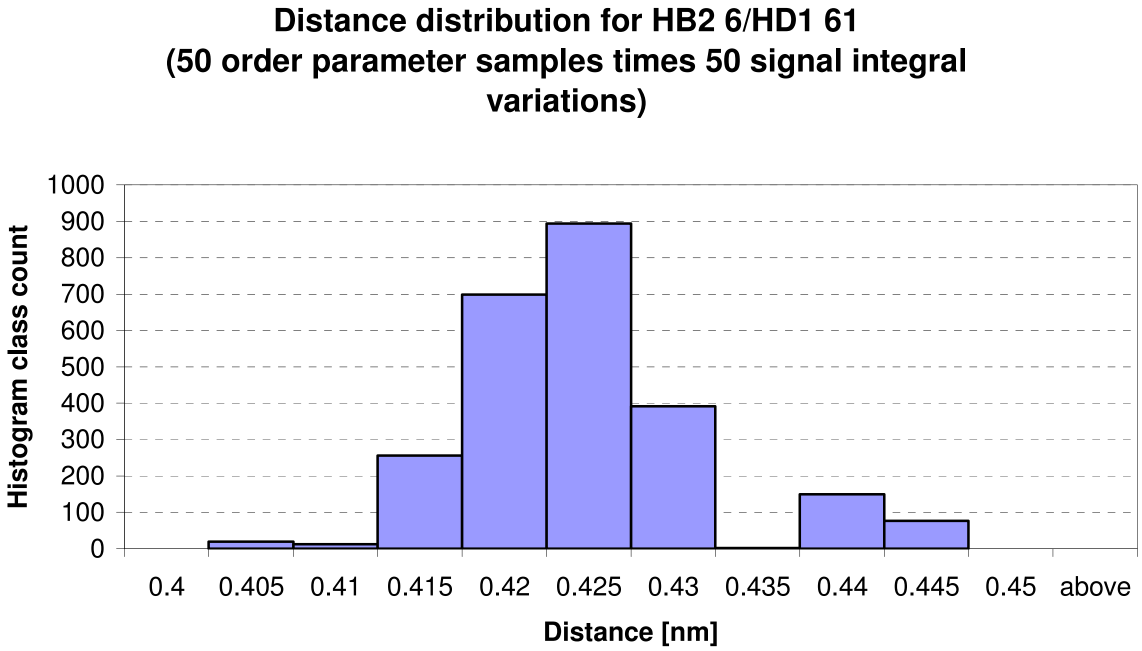

Figure 3.39 Example of a calculated distance distribution for the NOE contact HB2 6 / HD1 61 obtained using 50 independent variations of the backbone and sidechain order parameters times 50 variations of the corresponding experimental peak integral.

Contrary to other approaches to distance error estimation in NMR structure determination, where for example 20% of the actual distance are used for the definition of upper and lower error bounds across-the-board and thus are directly proportional to the distance, the errors determined from the distance distributions show no predominant dependence of the actual calculated distances:

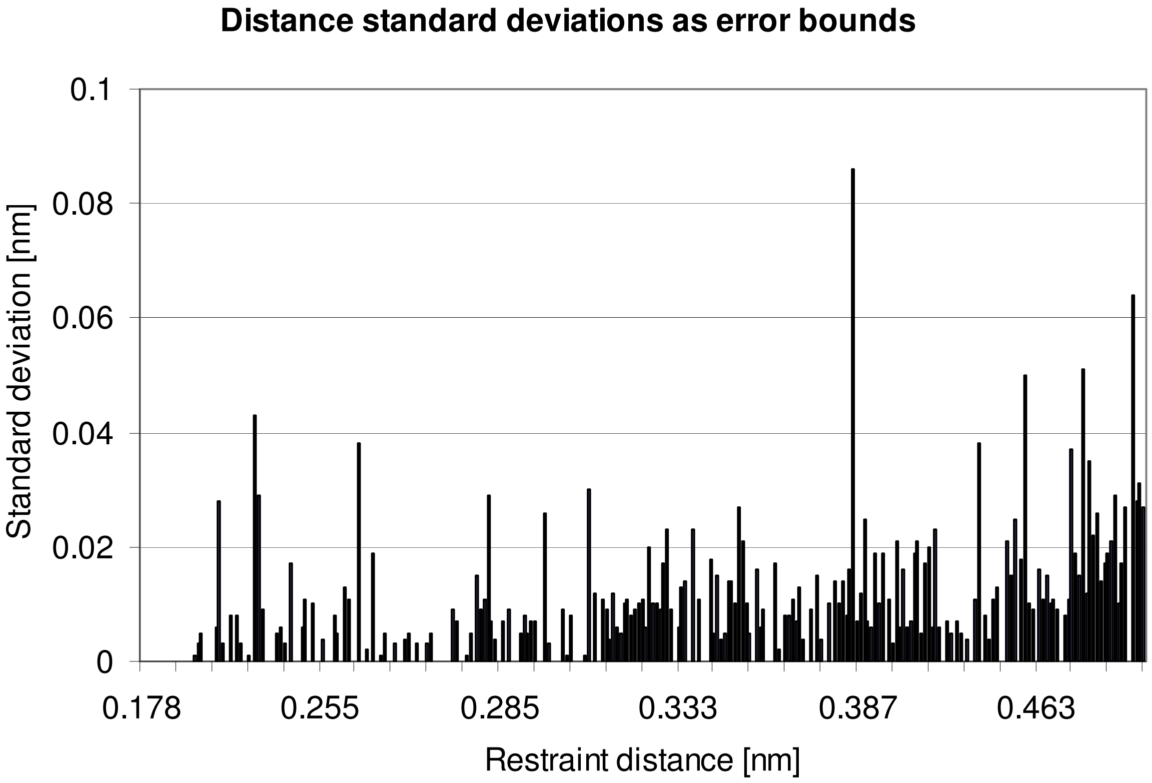

Figure 3.40 The standard deviation error bounds calculated from the distance distributions show no direct dependence on the restraint distance.

Taking a closer look at the individual distributions for given restraints reveals that the calculated distances are not continuously distributed as expected, which can on the one hand be explained by the under-sampling of the normal distribution using only 50 values for variation. On the other hand possible correlation effects due to the simplified modelling and variation scheme are suspected to have an additional influence on the smoothness of the distribution as for example the left hand part of the distribution shown in the following figure stems from order parameter combinations that describe an overall very rigid molecule with order parameter values close to one for the backbone and sidechain.

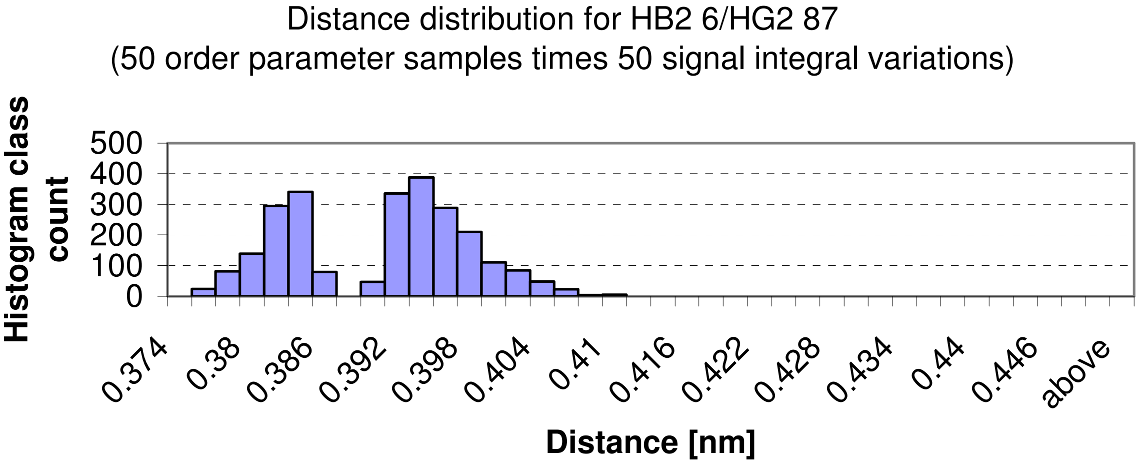

Figure 3.41 Calculated distance distribution for the long range NOE contact HB2 6 / HG2 87 with 50 variations per backbone and side-chain order parameter times 50 signal integral variations. The splitting of the distribution is caused mainly by the neighboring HB3 6 side-chain proton (see Fig. 3.42 below) that influences the relaxation rate between the other two protons (one side-chain proton and one averaged methyl group pseudo-proton) depending on the under-sampled mobility variation and by the under-sampled target integral variation.

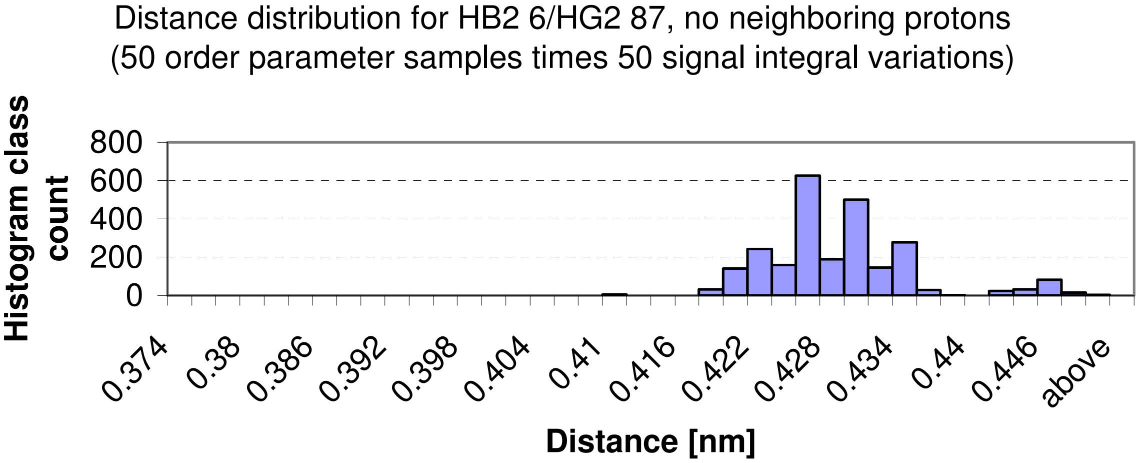

Figure 3.42 Same as Fig. 3.41 except the neighboring protons for HB2 6 and HG2 87 have been removed in the calculation. The result is an expected shift to larger distances and a far less pronounced splitting of the distribution, originating from variation undersampling.

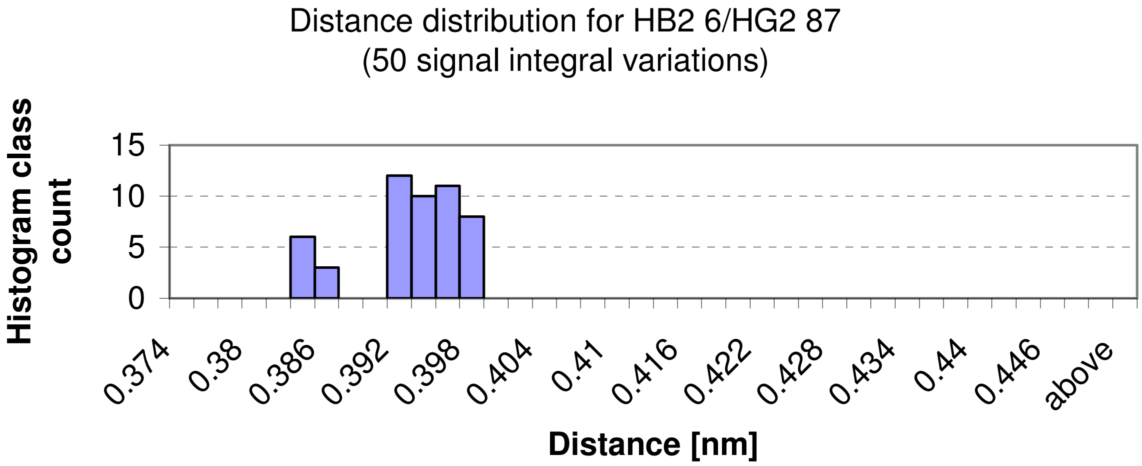

Figure 3.43 Using only 50 target integral variations leads to a distribution resembling Fig. 3.41. For a higher sample count a more uniform distribution is expected.

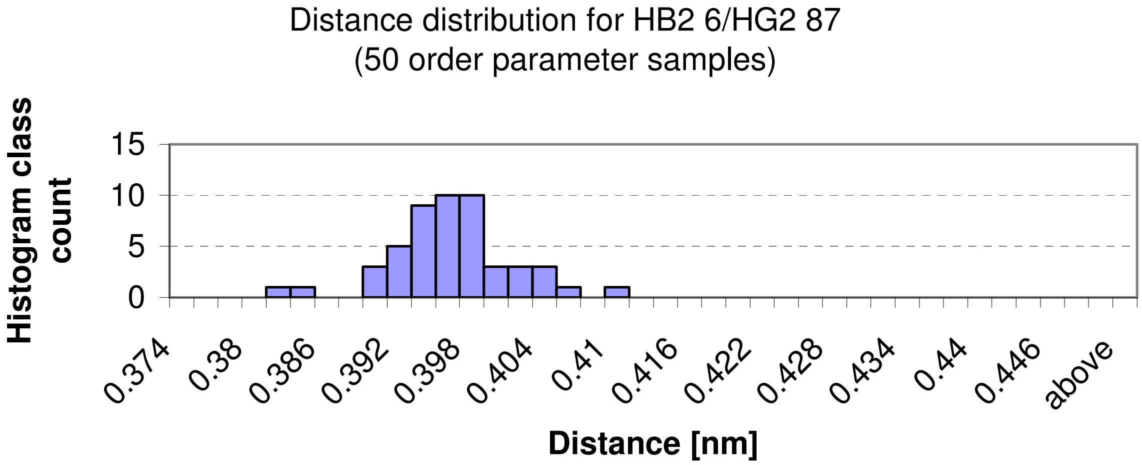

Figure 3.44 The distance distribution obtained using 50 order parameter variations, also not very uniform due to undersampling of the parameter variation.

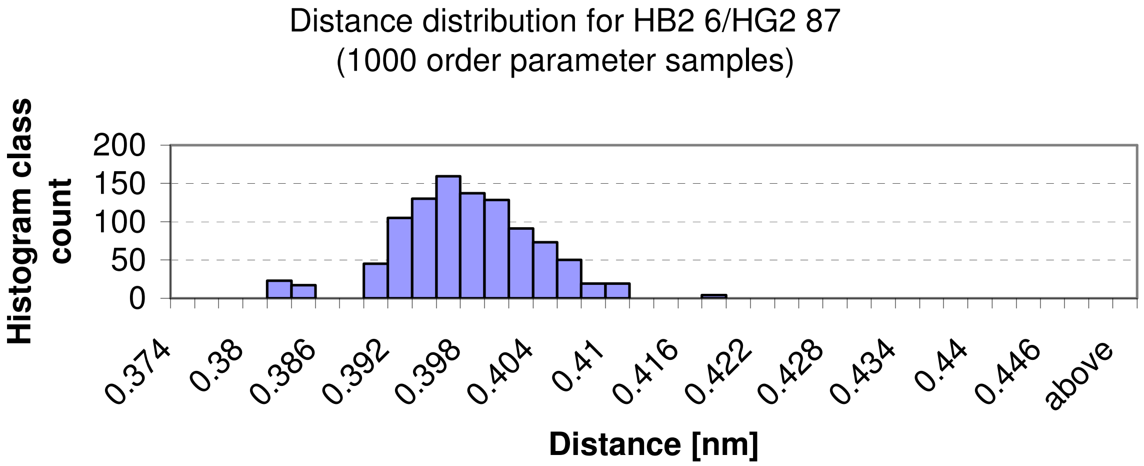

Figure 3.45 Calculated distance distribution for the long range NOE contact HB2 6 / HG2 87 with 1000 variations per backbone and side-chain order parameter and fixed target integrals.

As the example for NOE contact HB2 6 / HG 87 in Fig. 3.41 shows, for this NOE contact a rather interesting bimodal distance distribution is obtained using only 50 order parameter and 50 volume variation samples. Repeating the same calculation excluding neighboring protons of HB6 6 yields a more uniform distribution (Fig. 3.42), which demonstrates the influence of neighboring proton spins on the observed relaxation rate. A look at 50 variations of either target peak integrals or order parameters alone (Figs. 3.43, 3.44) reveals that the corresponding distance distributions generally lack uniformity, so for a combination of these variations also no uniformity can be expected. Increasing the sample count for the order parameter variation to 1000 samples results in a more continuous distribution of distances (Fig. 3.45), as the influence of order parameter combinations that describe an overall extremely rigid or mobile molecule is decreased.

Investigating the actual order parameter distribution used for the calculations reveals that the desired normal distribution for the variation is quite poorly reproduced:

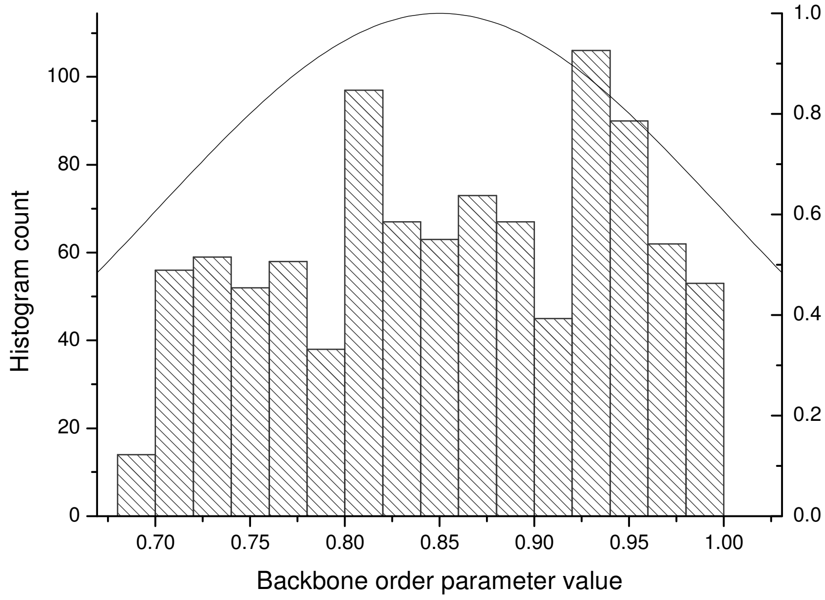

Figure 3.46 Theoretical backbone order parameter distribution (black line) vs. distribution obtained from 1000 samples.

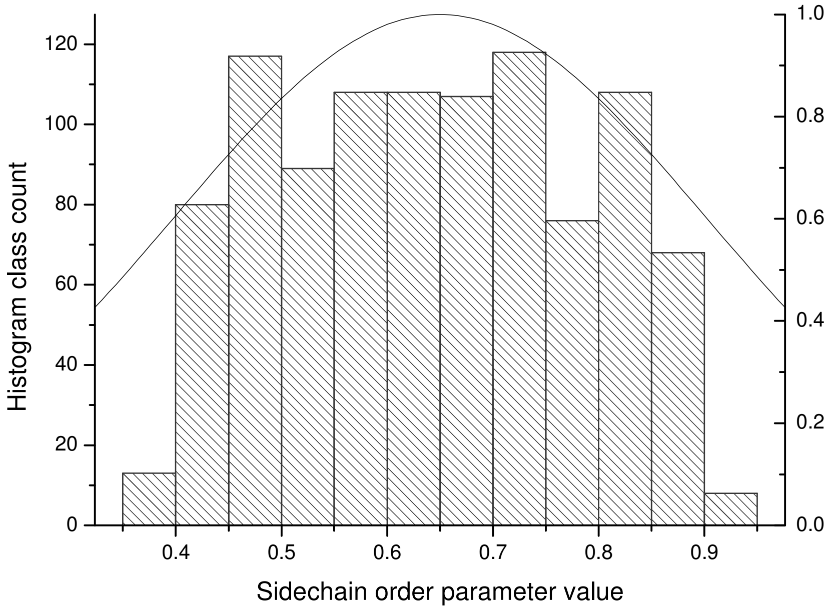

Figure 3.47 Theoretical sidechain order parameter distribution (black line) vs. distribution obtained from 1000 samples.

This is the result of undersampling together with a very large standard deviation in comparison to the desired parameter range (ratio 1:2). A general flaw in the random number generator can be ruled out, since in test runs a Gaussian distribution of random numbers was obtained.

In the following, the description of the distance error in terms of the standard deviation of the distances obtained from the undersampled distributions is used. To evaluate the structure quality that can be achieved using this approach to restraint generation as compared to ISPA_auto (see section 2.5) with automatic scaling as described in table 2.2, first molecular dynamics calculations were performed. CNS was used together with distance restraints from REFINE and ISPA respectively plus 23 additional experimental H-bond [91] and 118 calculated TALOS angle restraints. From each ensemble of 200 structures the 10 best in terms of overall energy were selected for the R-factor analysis in AUREMOL. To find out whether the iterative application of REFINE can further improve the results, the structure obtained after the first iteration was used as the trial structure for a second iteration. The results show overall better R-factor values for the REFINE structures across the distances, especially at the lower and higher end of the scale (Fig. 3.48).

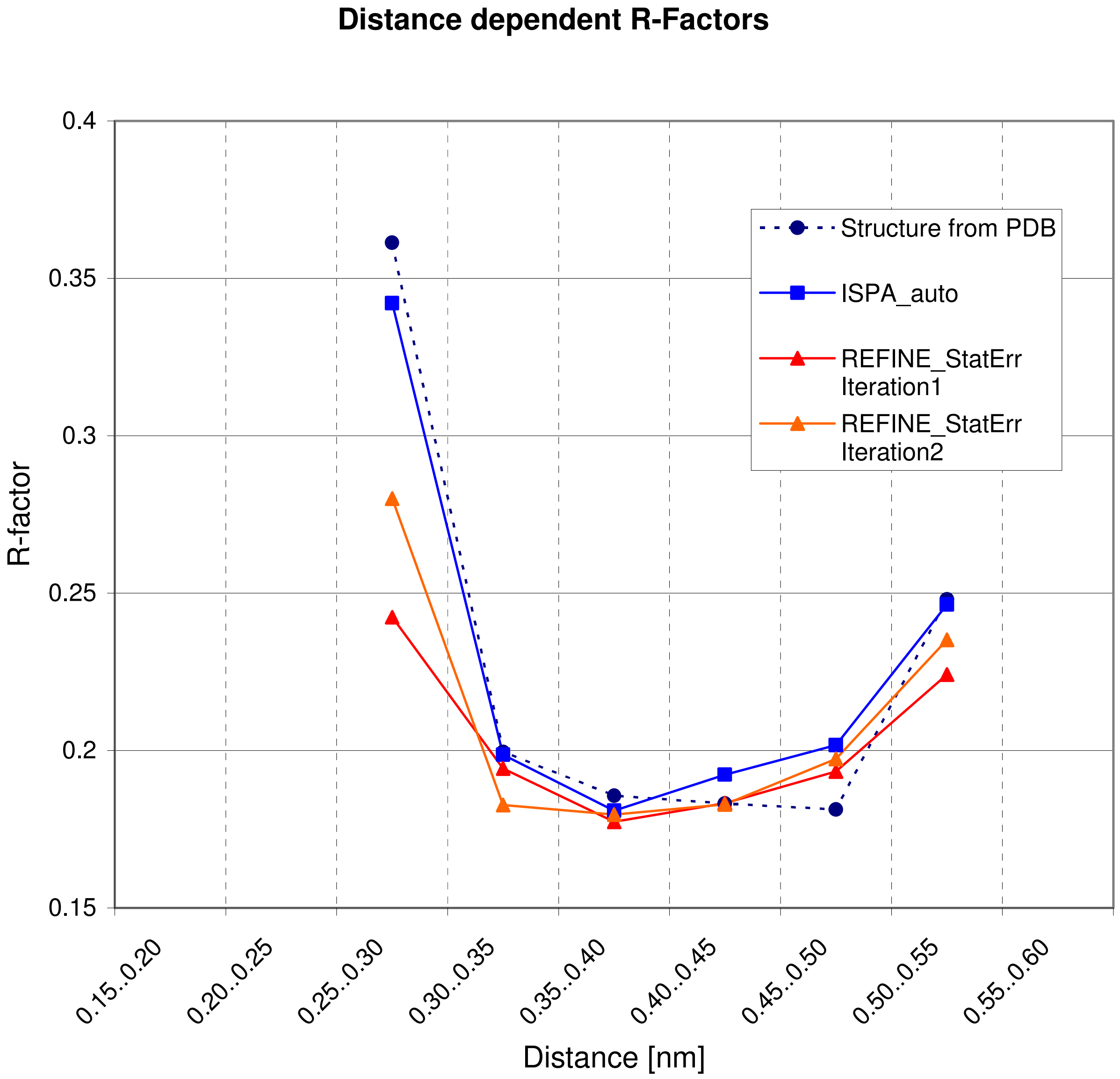

Figure 3.48 R-factor distribution depending on the distance in the structures calculated from ISPA_auto (see table 2.2) restraints (squares) and REFINE restraints (triangles) from the two iterations compared to the structure from the PDB (HPr from *S. carnosus*, PDB ID: 1KA5) (dashed). For R-factor calculations the R-factor R_3 as described in section 2.9 was used on the fully assigned spectrum. Data points are only shown for classes containing at least ten entries. In the calculation the mean order parameter values (table 3.16) were used in the simulation.

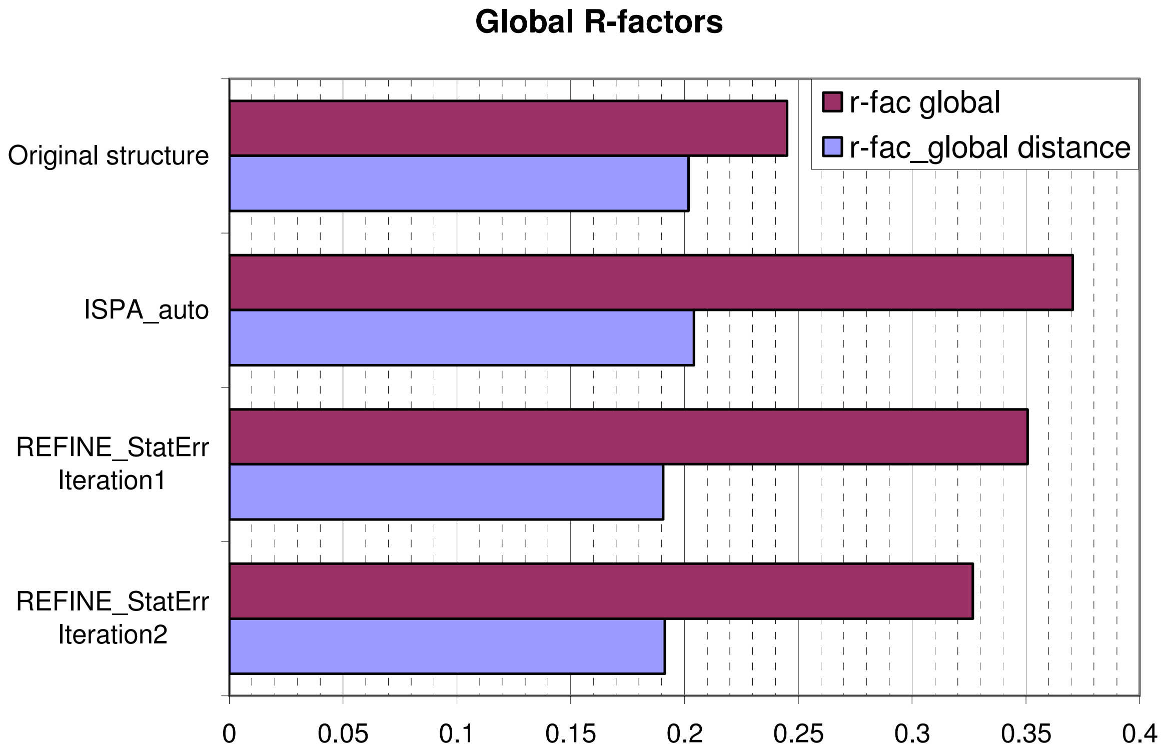

Figure 3.49 Global R-factor comparison between the original structure of HPr from *S. carnosus* from the PDB (top) and the resulting top ten structures resulting from MD with ISPA (middle) and from the first and second iteration REFINE (bottom) restraints. In the first REFINE iteration, an extended strand was used as trial structure and the resulting structures were used as trial structure input for the second REFINE distance calculation. For the calculation of global R-factors (r-fac global) shown in red R_5 (see section 2.9) was used while for the calculation of the distance related global R-factors (r-fac_global distance) R_3 was employed, using only one distance class for all distances. All R-factors were calculated in respect to the fully assigned experimental spectrum using the mean order parameter values (table 3.16) in the simulation.

A comparison of the global R-factors for the structures obtained from MD using ISPA_auto and REFINE restraints and the structure from the PDB shows that the structure from the database scores the lowest global R-factor (Fig. 3.49). This is expected, since this structure was solved using a much larger number of NOE restraints from 2D and 3D experiments. The REFINE structure scores the second best global R-factor, which indicates that the restraints generated by REFINE from only 271 NOEs describe the experimental spectral data in a better way than the restraints generated by the automatic ISPA approach from the same number of NOEs. Most notably the R-factor improved considerably for the structures obtained from MD calculations using the restraints generated by a second REFINE run (Iteration2). Here the best structures obtained after the first REFINE run (Iteration1) were used as starting structures.

The results for the global distance related R-factor summarize the findings from Fig 3.48 in a single R-factor, where REFINE produces the overall best value.

4. Discussion

4.1 General AUREMOL enhancements

Throughout AUREMOL the IUPAC-proposed atom nomenclature has been introduced. By that, previously inconsistent naming requirements in different AUREMOL modules have been standardised which improves the overall usability of the program. At the interface to external applications requiring proprietary atom naming conversion routines are provided for straightforward data exchange.

With the new strips tool AUREMOL can make the task of manually determining a sequential assignment more comfortable. To this end, several spectra can be used simultaneously during the process, help lines can be superimposed on the spectrum/strip display and optionally the relevant strip data can be prepared automatically from triple resonance data according to the resonances observed in a corresponding HSQC spectrum.

The RELAX spectrum simulation in AUREMOL now allows arbitrary combinations of motional models and finite relaxation delays, which further increases simulation accuracy. In the current state RELAX is a highly flexible program for spectrum simulation, although chemical exchange effects are only emulated by occupancy values for certain atoms. As soon as chemical exchange effects are fully accounted for in the simulation, RELAX will probably be the most complete NOESY simulation software as it combines the advantages of other spectrum simulation software like BIRDER [63], CORMA [64], SPIRIT [65], etc. in a single program.

4.2 Adaptive peak picking

The presented adaptive peak picking routine makes use of a local noise estimate for a local threshold determination. For a signal to be recognized, its smoothed intensity value must exceed the local noise and additionally an average intensity fall-off from the center is required. From a number of test datasets calibration constants have been determined empirically so that the presented results clearly show a reduction in erroneously picked artifact signals. Also, weak signals are not a priori omitted as they are using the fixed threshold routine in AUREMOL for a too large threshold. However, it cannot be expected that with the currently used calibration values equally good results can be achieved on arbitrary spectral data. So the best means to make this approach universal would be the implicit determination of the calibration constants from the dataset under consideration. To this end for example an analysis of global signal versus noise area or an intensity histogram analysis could be employed.

In an attempt to further increase the ability to discriminate between signal and artifact for peak picking, an automatic training set generation for the Bayesian peak probability analysis in AUREMOL has been tested. Together with the signal fall-off, global and local symmetries serve as a measure for pre-classifying the signal list obtained from the adaptive peak picking routine into potential signals and potential artifacts. A subset of these pre-classified signals is then used as training set for the Bayes algorithm. In the final step, signals with a calculated probability below a specified threshold can be deleted. That way nearly artifact free signal lists can be produced. It has shown however, that not only artifacts are removed by this procedure. The reason for this lies most probably in the pre-classification step, which favors isolated symmetric peaks as true signals in its current implementation and thus biases the Bayes algorithm towards symmetric signals. A possible solution would be to introduce separate classes for isolated and overlapping signals to the Bayes algorithm and extend the pre-classification. With regard to automatic structure determination it could be promising to start with a nearly artifact free signal lists for an automatic NOE assignment in KNOWNOE. Subsequently distance restraints are obtained using REFINE. The resulting structure is used in the follwing steps together with an increasingly less restricted signal list for further KNOWNOE/REFINE cycles. That way the influence of wrong assignments on the structure determination process could be minimized.

4.3 Integration error estimation

In experimental peak integral data the sources for errors include noise, overlap, baseline distortion, artifacts and the limited digital resolution. The presented error estimation procedure focuses on noise and overlap. While the noise contribution can easily be calculated, for the overlap error estimation information about the locations of the overlapping peaks is required. The results of the calculated integrals and the corresponding error bounds on a simulated noised dataset in comparison to the originally simulated integrals show that only 52% of the integrals have sufficient error estimates, while for the remaining 48% of the integrals the errors are underestimated.

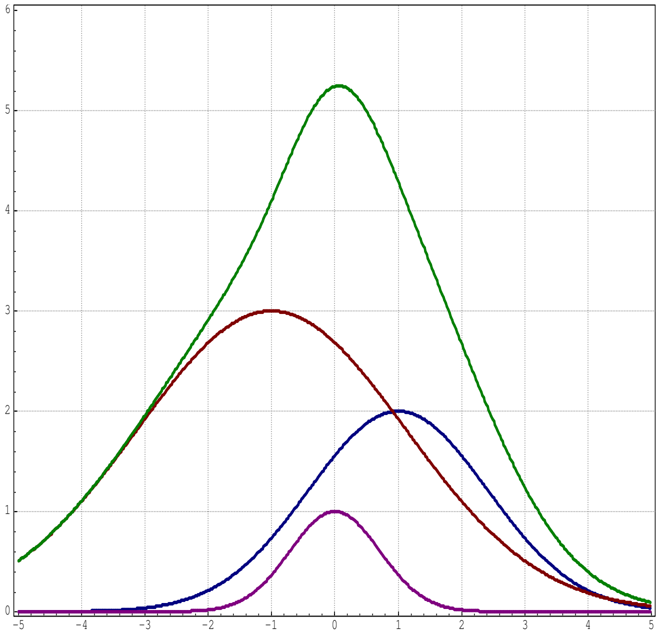

Figure 4.1 The effect of peak overlap on signal integration and error estimation is outlined in this graph. The weak purple signal is overlapped by the stronger red and blue signals. The observed intensity distribution in a spectrum is represented by the green curve. At the position where the purple signal is expected, there is a signal, although it is way too strong. In addition, the segmentation routine cannot find local maxima corresponding the red and blue overlapping peaks in the green distribution. As a result the integral of the purple signal is too large, the corresponding overlap error estimate is zero and the red and blue signal cannot be integrated.

In Fig. 4.1 this problem is explained. In the test dataset there is very strong overlap, of 4716 simulated signals 3537 (75%) could be integrated. Consequently for the remaining integrals intensities from certain signals can be attributed to another signal and the overlap error contribution cannot be calculated, leading to wrong integral values with underestimated errors. Although this is not a very satisfying situation, a solution to this problem is not easy. In case information about the peak positions is available, peak fitting or a modified segmentation routine could be used for integration and error estimation, but for spectra of unknown proteins this can hardly be expected.

The contributions of the remaining error sources are not considered in the current implementation, although the influence of minor artifact streaks could to some extent be accounted for by using a smoothed noise estimate that takes neighboring rows in the dataset into account. Baseline errors can (and should) be minimized by a thorough processing with baseline correction of the spectra. The influence of digital resolution on the integrals is hard to judge and strongly depends on the sample and experiment, however its importance is expected to be minor compared to signal overlap.

4.4 Automatic *de novo* assignment

Already in [57] it was discussed that a pseudo energy function based purely on local peak match might not be an ideal optimization target. This could be confirmed, as by the introduction of an additional global pseudo energy term the assignment performance could greatly be enhanced. On an artificial dataset similar to the test case in [57] the number of correctly assigned signals starting without *a priori* partial assignment was increased from 10% to 100% using the new pseudo energy function. In a recent application to a real world spectrum the preliminary results (not shown in this thesis) indicate that the current extended pseudo energy function still is not sufficient, since false assignment possibilities exist with an overall lower pseudo energy than for the correct configuration. One reason for this can be dicrepancies between experiment and simulation due to neglected chemical exchange, an erroneous trial structure or insufficient modelling of the protein mobilities in the simulation. By including chemical exchange effects, using an ensemble of trial structures and a variational approach for the simulation parameters these problems shold become controllable. Also the problem of peak overlap comes into play. In regions of high overlap the assignment problem exhibits degenerate behaviour and several assignment combinations can locally describe the experiment equally well. So a way has to be found to favor the correct assignment in these cases. One possibility for that is another contribution to the pseudo energy defined by the match of whole spectrum strips. In the end for the algorithm to be reliable, a pseudo energy function description has to be found where the global minimum corresponds to the correct assignment. The work on this assignment routine is currently part of a different thesis. When this problem is solved for experimental data, the complete automation of structure determination in AUREMOL is made possible.

4.5 REFINE distance calculation

The REFINE approach using direct cross relaxation rate scaling compares very well with the approach in NO2DI [73], where instead of the rates the distances corresponding to experimental NOEs are rescaled by the ratio of the first order distance estimates calculated from the simulated and experimental NOE values; from the new distances again rates are calculated for the next NOE simulation step. In REFINE no distance matrix is needed and only the auto-relaxation rates are re-calculated as the cross-relaxation rates are manipulated directly. This leads to reduced memory use and shorter calculation times in REFINE. The same is true when comparing REFINE to approaches using rate backcalculation from NOE data as for example in MARDIGRAS [72], where per iteration step first the experimental data for non-fixed contacts is merged to the simulated NOE set, which basically amounts to a direct NOE scaling. Then for this hybrid NOE matrix the rate matrix is calculated by evaluating the matrix logarithm and following that a new NOE matrix is calculated for the next iteration. So in MARDIGRAS two computationally costly transformations (matrix-exponential and -logarithm) have to be calculated whereas in REFINE only one transformation (matrix-exponential) per iteration is necessary. So although by the introduction of the logarithmic ratio scaling in REFINE a slight increase in the required iteration count for convergence has been noticed, the overall execution times per run are generally in the range of minutes for a single trial structure calculated on an average PC.

Since the relaxation matrix approach used for distance calculation in REFINE generally promises more accurate results than ISPA because higher order spin diffusion effects are accounted for, this claim has been verified for the REFINE algorithm on artificial and experimental data.

4.5.1 Artificial test data

The ability of the REFINE algorithm to find a relaxation matrix that describes experimental data best has been verified on a simulated dataset of TmCSP. Using the calculated NOEs together with the structure used for the simulation as REFINE input yields 100% correct distances after the first iteration, as can be expected since for the simulation as well as by REFINE the RELAX backcalculation is used. Switching to an extended strand as input for REFINE still yields more than 85% of all distances with an error below 20% as opposed to ISPA, where around 70% of all distances were calculated with an error below 20%. So regardless of the trial structure used in REFINE the calculated distances are more accurate than those obtained using ISPA. To examine REFINE in a more realistic test case, a simulated spectrum from a well defined structure of HPr *(S. carnosus*) has been provided with different levels of additive Gaussian noise and treated like experimental data, i.e. peak picked and integrated. Compared to the ISPA approach generally a larger number of distances with a lower deviation from the optimum values is produced by REFINE, regardless of the noise level:

Noise level	0%	25%	50%	75%	100%
ISPA	69%	60%	58%	54%	48%
REFINE (extended strand)	73%	72%	67%	64%	64%
REFINE (original structure)	93%	92%	88%	85%	83%

Table 4.1 Percentage of calculated distances deviating less than 20% from the distance in the original structure used for simulating the dataset. With increasing noise, the distance accuracy degrades. ISPA is affected more strongly as the calibration integral changes with the noise.

As some overlapping signals are lost due to the application of automatic peak picking, REFINE cannot reproduce the excellent results of the ideal testcase described above, but still between 64% and 93% of the distances are calculated with a deviation from the original distances below 20%, depending on the trial structure and the noise level. For the folded trial structure an overall larger number of accurate distances is calculated. From this situation the conclusion can be drawn that the iterative application of REFINE, where the structures resulting from the previous run are used as input for the following distance determination step, leads to distance restraints that are progressively becoming more accurate.

Regarding the robustness against noise in the spectral data, the quality of the determined distances generally degrades with increasing amounts of additive noise. For REFINE this degradation is less pronounced and still a larger amount of accurate distances is produced than by using ISPA. The reason for that is the influence of the noise especially on the single reference peak integral for the ISPA scaling, which is of no relevance for REFINE, since here a global scaling procedure is applied.

The distance error estimate derived from the integration error estimate alone is a step into the right direction for automation. However, the estimates generally are too strict, as they only accout for a minimum of about one third of the distance deviations for the case of the extended strand trial structure and about two thirds for the folded input structure (for the worst case with 75%-100% added noise). This outcome is not unexpected, since here only the errors in the experimental input are considered and errors due to the internal modelling for the relaxation matrix calculations are omitted. Additionally the limitations of the integration error estimate due to overlap apply here as discussed above. Judging from that result, the optional distance error estimation using experimental input data and simulation parameter variation will be a valuable alternative for the application of REFINE to experimental data.

4.5.2 Experimental data with unknown NOESY assignment – combining KNOWNOE and REFINE

When the sequential assignment of a protein is available but a corresponding NOESY dataset is not yet assigned, an iterative approach combining automatic NOESY assignment by KNOWNOE and automatic distance calculation by REFINE has proven to yield feasible structure hypotheses with minimal effort by the user. It is important to allow generous distance limits for the initial assignment, as the trial structure at that time is still far from the correct fold. With increasing iteration count, as the structures from each previous run gradually approach the folded state, the limits should be more and more restrained to achieve an increasingly accurate NOESY assignment The example of HPr from *S. carnosus* shows that the protocol qualitatively works and that the correct fold of a protein can be obtained that way from a single automatically peak picked and initially unassigned 2D NOESY dataset and TALOS predicted angle restraints after nine cycles of KNOWNOE and REFINE.

For RalGDS-RBD, where additional 3D NOESY data and 52 experimentally determined H-bonds were available, the application of this procedure led to a structure that was quantitatively comparable to X-ray results after six iterations, as shown by the global R-factor values of 0.322 for the structures obtained using REFINE and 0.325 for the X-ray structure.

So in case no *a priori* information about the correct NOESY assignment is known, but the sequential assignment is available, this approach can yield accurate structures in a short amount of time (a few days, including the time for the molecular dynamics calculations) and with a minimum of expert intervention. One has to keep in mind, that KNOWNOE uses a statistical approach, so a perfect NOESY assignment cannot be expected. The errors introduced by incorrect assignments can lead to warped structures, but the effect can be minimized by the removal of violated restraints from the restraint list followed by a repeated run of the restrained molecular dynamics calculation as it was done in these two test cases. It is important to note that by the automatic KNOWNOE scaling factor calculation using REFINE an additional source of error for the NOESY assignment has been eliminated and the reproducibility of the assignments is improved.

4.2.3 Experimental data with known assignment

On the experimental 2D NOESY dataset of HPr from *S. aureus* the performance of REFINE using the variational error estimation approach for a low amount of experimental information has been evaluated. For about 25% randomly selected signals of the fully assigned spectrum distance restraint distributions have been calculated by target peak integral and order parameter variation. A distance restraint list consisting of the mean distances and their respective standard deviations obtained from these distributions as error bounds was generated. One striking result is that the influence of the atom mobilities described by the order parameter S^2 can be substantial, so if these mobilities are unknown, the variational approach is essential for an accurate distance error estimate. Secondly, the calculated distance errors are not directly proportional to the actual distance values as is assumed in other approaches to restraint generation like the widely used application of three distance classes for ISPA restraints. For a restrained molecular dynamics calculation the distance restraint list obtained from the distance distributions was used together with 118 TALOS angle restraints and 23 experimental H-bond restraints. The structures resulting from the REFINE restraints showed a better agreement with the experiment as demonstrated by overall lower global R-factors (0.35) when compared to structures calculated using restraints that had been determined using an automatic ISPA approach (0.37). Even though the dataset was recorded using a short miximg time of 80ms where the ISPA approach should be applicable, compared to ISPA REFINE produces distance restraints that lead to structures with a 5% better global R-factor using an extended strand as input. The most important reason for the better performance here lies in the distance error bounds in REFINE that are calculated individually for each distance from the uncertainty in the experimental data and the modelling parameters. As the calculated distance distributions have shown to require a very large number of statistical samples to produce a smooth distribution, the current approach should be improved. Instead of the presented random sampling approach to parameter variation a systematic variation of the parameters within the desired distribution is proposed. Alternatively all permutations of the uniformly sampled parameter ranges could be used in the calculation to obtain the spread of the resulting distances as a basis for the error estimation.

The quality of the structures resulting from the REFINE restraints could be improved even more by performing a second calculation step, where the best ten structures obtained in the first REFINE run were used as starting structures. Here after the MD simulation structures with a 13% improved global R-factor of 0.33 were obtained. This demonstrates that the influence of neighboring spins on the relaxation rate of two spatially close protons is properly accounted for during the execution of the REFINE algorithm as the experimental data is better explained by the structure obtained from the first REFINE run than by an extended strand. Judging from this also a refinement of available, poorly defined structures should be possible.

Generally an even larger advantage for REFINE over ISPA can be expected in case data recorded for longer mixing times is analyzed.

4.6 Benefits from distance error estimation

Introducing a self contained error estimation concept to protein structure determination offers a number of advantages. First of all, distance restraints can effectively be attributed more or less influence on the restrained molecular dynamics calculation depending on the quality of the calculated distances. Distances that are poorly defined because of a large error in the experimental peak integral or because of insufficiently accurate modelling parameters are provided with a larger error estimate and are that way less likely to have a negative effect on the whole structure during MD. That way REFINE is contrasting the usual ISPA approaches where the error bounds depend predominantly on the corresponding distances.

Secondly, the reproducibility of the structure determination process is greatly enhanced by adding the automatic error estimation to the already very much automated process. This is an essential prerequisite for the complete automation of structure determination from NMR. It allows to monitor and quantify the progress of structure determination, e.g. because of an increasingly complete assignment.

As a third point the variational approach to error estimation is a universal approach. It can easily be extended to include additional sources of error. The produced distance distributions have up to now only been used by means of mean values and standard deviations. Here also more refined approaches can be implemented that take the actual shape of a given distribution into account for molecular dynamics.

4.7 Potential for automation

AUREMOL now offers a wide range of routines for the automation of typical tasks for structure determination from NMR data. Starting from adaptive peak picking and the extended integration routine to distance calculation and structure evaluation, the different modules work with a minimum of user intervention and consistent IUPAC compliant atom naming. The introduction of an automatic and individual error estimation to the workflow is a novelty and allows to reproducably judge the results of the structure determination process.

It is clear that the experimental peak integrals used for REFINE always contain errors, simply because of the limited digital resolution, peak overlap, artifacts, noise and baseline distortions. This fact is now accounted for by the new integration routine that additionally calculates an error estimate for the noise and overlap contributions. Subsequently the errors introduced to the distance calculation by imperfect experimental data and insufficient modelling parameters like unknown order parameters are reflected in the error estimates for the distances calculated by REFINE. In the following molecular dynamics the uncertainty of distances directly affects the resulting structures. That way the errors are carried throughout the whole process from integration to the molecular dynamics calculation and are quantified in the final structure bundle instead of getting lost on the way and resulting in probably erroneous structures because of error bounds simply proportional to the distances.

So in case the sequential assignment of a protein is known and at least one 2D NOESY dataset is available, the procedures described in sections 3.7-3.9 allow already an automatic structure determination. The results presented indicate that the quality to expect from this approach is at least on par with the results from X-ray methods.

A combination of the presented modules to form a fully automatic structure determination routine where only spectral data and a starting trial structure are used as input is the logical next step. Here the missing link right now would be a flexible and reliable automatic sequential assignment module. The current state of the ASSIGN module already allows the complete automatic assignment of synthetic 2D NOESY data using a homology modelled structure as input. As soon as this functionality has been successfully transferred to experimental data, where the preprocessing of the dataset for the algorithm and the correct modelling of the pseudo energy function for the optimization have shown to play a crucial part, the vision of full automation is within close reach.

5. Summary & Outlook

In this thesis a method for automated and accurate distance determination from protein NMR data is presented. Based on the relaxation matrix formalism the REFINE algorithm has been developed that uses the simulation of NOESY data in an iterative approach to fit the simulated to the experimental data for distance information extraction. The addition of an error estimation for the integration of the experimental data and the modelling parameters in the simulation allows to calculate individual distance error estimates. Especially the influence of the order parameter choice on the calculated distances has shown to be substantial. As the results for artificial and real world experimental data show, the procedure is capable of accurate distance restraint determination, which allows the calculation of high-quality structures even from a limited amount of experimental data. Furthermore the whole process is largely carried out automatically with a minimum of user intervention. For more accurate results, in the future the current implementation of the integration error estimation should be improved to account for strongly overlapping signals and the parameter variation in the presented approach could be replaced by a systematic sampling of an expected parameter distribution.

Although the results were produced only for 2D and 3D datasets, this is no general limitation, the application to higher dimensional data is possible. Since higher order spin diffusion effects are automatically accounted for, the application to experimental data obtained using longer mixing times is expected to further increase the advantage of REFINE over ISPA approaches.

Considering the vision of a fully automatic procedure for the determination of protein structures from NMR data there remains only one missing link, which is the automatic *de novo* resonance line assignment. The state of this module as presented in this thesis shows very good results on artificial data by using again spectrum simulation. If these results can be transferred to experimental spectra the final gap would be closed. In that respect it might be helpful to include a chemical shift prediction routine in the future.

Appendix

A. Abbreviations and terms

AUREMOL	NMR software project, framework of this dissertation
FFT	**F**ast **F**ourier **T**ransform
Homology modelling	Modelling of a protein structure using known structures of homologous proteins
HPr	**H**istidine containing phosphocarrier **Pr**otein
HSQC	**H**eteronuclear **S**ingle **Q**uantum **C**oherence
ISPA	**I**solated **S**pin **P**air **A**pproximation
IUPAC	**I**nternational **U**nion of **P**ure and **A**pplied **C**hemistry
KNOWNOE	NOE assignment module in AUREMOL
(r)MD	(**r**estrained) **M**olecular **D**ynamics
NMR	**N**uclear **M**agnetic **R**esonance
NOE	**N**uclear **O**verhauser **E**ffect
NOESY	**N**uclear **O**verhauser **E**ffect **S**pectroscop**Y**
PEAK ASSIGN	General assignment module in AUREMOL
RalGDS	**Ral G**uanine nucleotide **D**issociation **S**timulator
REFINE	Distance calculation module in AUREMOL
RELAX	Spectrum simulation module in AUREMOL
Ribbon plot	Type of protein display highlighting backbone configuration and secondary structure elements
RMS(D)	**R**oot **M**ean **S**quare (**D**eviation)
S/N ratio	**S**ignal to **N**oise ratio
S. aureus, S. carnosus	Gram-positive bacteria of the genus **S**taphylococcus
TmCSP	**C**old **S**hock **P**rotein from the hyperthermophilic organism **T**hermotoga **M**aritima

B. Programming environment

The AUREMOL software is being developed in compliance with the ANSI C programming language standard. For programming, the integrated development environment – IDE – Visual Studio 6.0 by Microsoft has been used, together with the DevPartner Studio 7.1 package by Compuware for software quality control.

The development workstation used was a Dell Optiplex GX260 running Windows XP Professional by Microsoft.

The Linux port is built on the Red Hat Linux Workstation V 4.0 distribution.

Molecular dynamics calculations were carried out on the Linux compute cluster of the computing center of the University of Regensburg.

C. Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the PhD program ”Nonlinearity and Nonequilibrium in Condensed Matter“ (transl.: Graduiertenkolleg „Nichtlinearität und Nichtgleichgewicht in kondensierter Materie“), GRK 638.
Further funding was provided by the SPINE program (Structural Proteomics In Europe) of the European Union.

I would like to thank

- My Ph.D. supervisor Prof. Dr. Dr. Hans Robert Kalbitzer for his support and near inexhaustible supply of new ideas often discussed during the work on this dissertation.
- PD Dr. Wolfram Gronwald for the close and productive collaboration on AUREMOL, especially his application of the combined KNOWNOE/REFINE approach to the RalGDS-RBD dataset and for proof-reading previous drafts of this dissertation.
- Dr. Klaus-Peter Neidig for his support concerning questions on the Amix Viewer.
- Dr. Bernhard Ganslmeier for giving me an excellent hands-on introduction to AUREMOL programming at the beginning of my work.
- Konrad Brunner for taking over the source code maintenance task, for very productive source code merging sessions and together with Kumaran Baskaran, Barbara Domogalla, Alexander Fink and Thorsten Graf for the practicals on probability in Bavarian card games, the variability in Bavarian dialects and chaos in classical mechanics involving a table and an elastic sphere.
- All members of the Biophysics institute for the great atmosphere.
- The members of the Graduiertenkolleg “Nonlinearity and Nonequilibrium in Condensed Matter” for providing interesting insights into different fields.
- All students that worked with us for a practical training semester, especially Markus Graf, Sebastian Helbig and Alexander Ross for implementing the automatic training set generation for the adaptive peak picking routine.
- My wife Michaela for her excellent proof-reading skills and for having a lot of patience.

D. References

[1] Tyers,M. & Mann,M. (2003) From genomics to proteomics. *Nature* **422,** 193-197.

[2] Oroudjev,E., Soares,J., Arcdiacono,S., Thompson,J.B., Fossey,S.A., & Hansma,H.G. (2002) Segmented nanofibers of spider dragline silk: atomic force microscopy and single-molecule force spectroscopy. *Proc. Natl. Acad. Sci. U. S. A* **99 Suppl 2,** 6460-6465.

[3] Rising,A., Nimmervoll,H., Grip,S., Fernandez-Arias,A., Storckenfeldt,E., Knight,D.P., Vollrath,F., & Engstrom,W. (2005) Spider silk proteins -- mechanical property and gene sequence. *Zoolog. Sci.* **22,** 273-281.

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

[5] Stryer,L. (1995) *Biochemistry*, 4th edn. W.H. Freeman & Company.

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

[7] Koradi,R., Billeter,M., & Wüthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14,** 51-55.

[8] Baldwin,R.L. (1989) How does protein folding get started? *Trends Biochem. Sci.* **14,** 291-294.

[9] Gutin,A.M., Abkevich,V.I., & Shakhnovich,E.I. (1995) Is burst hydrophobic collapse necessary for protein folding? *Biochemistry* **34,** 3066-3076.

[10] Wei,J. & Hendershot,L.M. (1996) Protein folding and assembly in the endoplasmic reticulum. *EXS* **77,** 41-55.

[11] Borges,J.C. & Ramos,C.H. (2005) Protein folding assisted by chaperones. *Protein Pept. Lett.* **12,** 257-261.

[12] Sadqi,M., Lapidus,L.J., & Munoz,V. (2003) How fast is protein hydrophobic collapse? *Proc. Natl. Acad. Sci. U. S. A* **100,** 12117-12122.

[13] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, & Peter Walter (2002) *Molecular Biology of the Cell*, 4th edn. Garland.

[14] Dobson,C.M. (2003) Protein folding and misfolding. *Nature* **426,** 884-890.

[15] Selkoe,D.J. (2003) Folding proteins in fatal ways. *Nature* **426,** 900-904.

[16] Stefani,M. (2004) Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. *Biochim. Biophys. Acta* **1739,** 5-25.

[17] Ramos,C.H. & Ferreira,S.T. (2005) Protein folding, misfolding and aggregation: evolving concepts and conformational diseases. *Protein Pept. Lett.* **12,** 213-222.

[18] Jeyashekar,N.S., Sadana,A., & Vo-Dinh,T. (2005) Protein amyloidose misfolding: mechanisms, detection, and pathological implications. *Methods Mol. Biol.* **300,** 417-435.

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

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

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

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

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

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

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

[26] Neidig,K.-P., Bodenmüller,H., & Kalbitzer,H.R. (1984) Computer aided evaluation of two-dimensional NMR spectra of proteins. *Biochem. Biophys. Res. Comm.* **125,** 1143-1150.

[27] Glaser,S. & Kalbitzer,H.R. (1987) Automated Recognition and Assessment of Cross Peaks in Two-Dimensional NMR Spectra of Macromolecules. *J. Magn. Reson.* **74,** 450-463.

[28] Novic,M., Eggenberger,U., & Bodenhausen,G. (1988) Similarities between Self-Convolution and Symmetry Mapping of Multiplets in Two_dimensional NMR Spectra. *J. Magn. Reson.* **77,** 394-400.

[29] Pfändler,P. & Bodenhausen,G. (1988) Analysis of Multiplets in Two-Dimensional NMR Spectra by Topological Classification: Applications to Vinblastine and Cyclosporin A. *Magn. Reson. Chem.* **26,** 888-894.

[30] Eccles,C., Guntert,P., Billeter,M., & Wuthrich,K. (1991) Efficient analysis of protein 2D NMR spectra using the software package EASY. *J. Biomol. NMR* **1,** 111-130.

[31] Herrmann,T., Güntert,P., & Wüthrich,K. (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J. Biomol. NMR* **24,** 171-189.

[32] Neidig,K.-P. & Kalbitzer,H.R. (1990) Improved Representation of Two-Dimensional NMR Spectra by Local Rescaling. *J. Magn. Reson.* **88,** 155-160.

[33] Kleywegt,G.J., Boelens,R., & Kaptein,R. (1990) A Versatile Approach toward the Partially Automatic Recognition of Cross Peaks in 2D ^{1}H NMR Spectra. *J. Magn Reson.* **88,** 601-608.

[34] Garrett,D.S., Powers,R., Gronenborn,A.M., & Clore,G.M. (1991) A Common Sense Approach to Peak Picking in Two-, Three-, and Four-Dimensional Spectra Using Automatic Computer Analysis of Contour Diagrams. *J. Magn Reson.* **95,** 214-220.

[35] Koradi,R., Billeter,M., Engeli,M., Guntert,P., & Wuthrich,K. (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J. Magn Reson.* **135,** 288-297.

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

[37] Schulte,A.C., Gorler,A., Antz,C., Neidig,K.P., & Kalbitzer,H.R. (1997) Use of global symmetries in automated signal class recognition by a bayesian method. *J. Magn Reson.* **129,** 165-172.

[38] Denk,W., Baumann,R., & Wagner,G. (1986) Quantitative Evaluation of Cross-Peak Intensities by Projection of Two-Dimensional NOE Spectra on a Linear Space Spanned by a Set of Reference Resonance Lines. *J. Magn. Reson.* **67,** 386-390.

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

[40] Kraulis,P.J. (1989) ANSIG: A Program for the Assignment of Protein ^{1}H 2D NMR Spectra by Interactive Computer Graphics. *J. Magn. Reson.* **84,** 627-633.

[41] Helgstrand,M., Kraulis,P., Allard,P., & Hard,T. (2000) Ansig for Windows: an interactive computer program for semiautomatic assignment of protein NMR spectra. *J. Biomol. NMR* **18,** 329-336.

[42] FELIX. 2003. San Diego CA, Accelrys Inc. (Computer Program)

[43] Pons,J.L., Malliavin,T.E., & Delsuc,M.A. (1996) Gifa V. 4: A complete package of NMR data set processing. *J. Biomol. NMR* **8,** 445-452.

[44] Malliavin,T.E., Pons,J.L., & Delsuc,M.A. (1998) An NMR assignment module implemented in the Gifa NMR processing program. *Bioinformatics* **14,** 624-631.

[45] Goddard,T.D. & Kneller,D.G. SPARKY 3. 2003. San Francisco CA, University of California. (Computer Program)

[46] Zimmerman,D., Kulikowski,C., Wang,L., Lyons,B., & Montelione,G.T. (1994) Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. *J. Biomol. NMR* **4,** 241-256.

[47] Olson,J.B., Jr. & Markley,J.L. (1994) Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances: a demonstration of the connectivity tracing assignment tools (CONTRAST) software package. *J. Biomol. NMR* **4,** 385-410.

[48] Bartels,C., Güntert,P., Billeter,M., & Wüthrich,K. (1997) GARANT-A General Algorithm for Resonance Assignment of Mutidimensional Nuclear Magnetic Resonance Spectra. *J. Comput. Chem.* **18,** 139-149.

[49] Leutner,M., Gschwind,R.M., Liermann,J., Schwarz,C., Gemmecker,G., & Kessler,H. (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *J. Biomol. NMR* **11,** 31-43.

[50] Duggan,B.M., Legge,G.B., Dyson,H.J., & Wright,P.E. (2001) SANE (Structure Assisted NOE Evaluation): an automated model-based approach for NOE assignment. *J. Biomol. NMR* **19,** 321-329.

[51] Linge,J.P., O'Donoghue,S.I., & Nilges,M. (2001) Automated assignment of ambiguous nuclear overhauser effects with ARIA. *Methods Enzymol.* **339,** 71-90.

[52] Linge,J.P., Habeck,M., Rieping,W., & Nilges,M. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19,** 315-316.

[53] Herrmann,T., Guntert,P., & Wuthrich,K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319,** 209-227.

[54] Grishaev,A. & Llinas,M. (2002) CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proc. Natl. Acad. Sci. U. S. A* **99,** 6707-6712.

[55] Grishaev,A. & Llinas,M. (2002) Protein structure elucidation from NMR proton densities. *Proc. Natl. Acad. Sci. U. S. A* **99,** 6713-6718.

[56] Schneider,J.J. Effiziente parallelisierbare physikalische Optimierungsverfahren. 1999. Regensburg, University of Regensburg, Germany. (Dissertation)

[57] Ganslmeier,B. AUREMOL - Softwareprojekt zur automatischen Auswertung von multidimensionalen NMR-Spektren. 2002. Regensburg, University of Regensburg, Germany. (Dissertation)

[58] Kirchhöfer,R. Computergestützte Analyse von NMR-Spektren. 2005. Regensburg, University of Regensburg, Germany. (Dissertation)

[59] Moussa,S. NMR Spectroscopy of Polypeptides and a New Statistical Method for the Assignment of Nuclear Overhauser Effect Signals. 2001. Regensburg, University of Regensburg, Germany. (Dissertation)

[60] Solomon,I. (1955) Relaxation Processes in a System of Two Spins. *Phys. Rev.* **99,** 559-565.

[61] Macura,S. & Ernst,R.R. (1980) Elucidation of cross relaxation in liquids by two-dimensional N.M.R. spectroscopy. *Mol. Phys.* **41,** 95-117.

[62] Luginbühl,P. & Wüthrich,K. (2002) Semi-classical nuclear spin relaxation theory revisited for use with biological macromolecules. *Prog. NMR Spectrosc.* **40,** 199-247.

[63] Zhu,L. & Reid,B.R. (1995) An Improved NOESY Simulation Program for Partially Relaxed Spectra: Birder. *J. Magn. Reson. B* **106,** 227-235.

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

[65] Zhu,L., Dyson,H.J., & Wright,P.E. (1998) A NOESY-HSQC simulation program, SPIRIT. *J. Biomol. NMR* **11,** 17-29.

[66] Abragam,A. (1961) *The Principles of Nuclear Magnetism.* Clarendon Press, Oxford.

[67] Lipari,G. & Szabo,A. (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity. *J. Am. Chem. Soc.* **104,** 4546-4559.

[68] Moler,C. & Van Loan,C. (2003) Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review* **45,** 3-49.

[69] Görler,A. Quantitative Auswertung von NOESY-Spektren unter Berücksichtigung der Spindiffusion. 1994. Heidelberg, Max-Planck-Institut für medizinische Forschung, Heidelberg, Germany. (Thesis)

[70] Boelens,R., Koning,M.G., van der Marel,G.A., van Boom,J.H., & Kaptein,R. (1989) Iterative Procedure for Structure Determination from Proton-Proton NOEs Using a Full Relaxation Matrix Approach. Application to a DNA Octamer. *J. Magn. Reson.* **82,** 290-380.

[71] Kim,S.-G. & Reid,B.R. (1992) Automated NMR Structure Refinement via NOE Peak Volumes. Application to a Dodecamer DNA Duplex. *J. Magn. Reson.* **100,** 382-390.

[72] Borgias,B.A. & James,T.L. (1990) MARDIGRAS-A Procedure for Matrix Analysis of Relaxation for Discerning Geometry of an Aqueous Structure. *J. Magn. Reson.* **87,** 475-487.

[73] van de Ven,F.J.M., Blommers,M.J.J., Schouten,R.E., & Hilbers,C.W. (1991) Calculation of Interproton Distances from NOE Intensities. A Relaxation Matrix Approach without Requirement of a Molecular Model. *J. Magn. Reson.* **94,** 140-151.

[74] Mumenthaler,C., Guntert,P., Braun,W., & Wuthrich,K. (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J. Biomol. NMR* **10,** 351-362.

[75] Cornilescu,G., Delagio,F., & Bax,A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13,** 289-302.

[76] Linge,J.P., Williams,M.A., Spronk,C.A., Bonvin,A.M., & Nilges,M. (2003) Refinement of protein structures in explicit solvent. *Proteins* **50,** 496-506.

[77] Brünger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grossekunstleve,R.W., Jiang,J.-S., Kuszewski,J., Nilges,M., Pannu,N.S., Read,R.J., Rice,L.M., Simonson,T., & Warren,G.L. (1998) Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Cryst.* **D54,** 905-921.

[78] Schwieters,C.D., Kuszewski,J., Tjandra,N.L., & Clore,G.M. (2003) The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160,** 65-73.

[79] Güntert,P., Mumenthaler,C., & Wüthrich,K. (1997) Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA. *J. Mol. Biol.* **273,** 283-298.

[80] Van Der,S.D., Lindahl,E., Hess,B., Groenhof,G., Mark,A.E., & Berendsen,H.J. (2005) GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26,** 1701-1718.

[81] Simons,K.T., Kooperberg,C., Huang,E., & Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **268,** 209-225.

[82] Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C., & Baker,D. (1999) Improved Recognition of Native-Like protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins. *Proteins* **34,** 82-95.

[83] Bowers,P.M., Strauss,C.E.M., & Baker,D. (2000) De novo protein structure determination using sparse NMR data. *J. Biomol. NMR* **18,** 311-318.

[84] Laskowski,R.A., Rullmann,J.A.C., MacArthur,M.W., Kaptein,R., & Thornton,J.M. (1996) AQUA and PROCHECK-NMR Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8,** 477-486.

[85] Hooft,R.W.W., Vriend,G., Sander,C., & Abola,E.E. (2003) Errors in protein structure. *Nature* **381,** 272.

[86] Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17,** 355-362.

[87] Markley,J.L., Bax,A., Arata,Y., Hilbers,C.W., Kaptein,R., Sykes,B.D., Wright,P.E., & Wüthrich,K. (1998) Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. *Pure & Appl. Chem.* **70,** 117-142.

[88] Huang,L., Weng,X., Hofer,F., Martin,G.S., & Kim,S.H. (1997) Three-dimensional structure of the Ras-interacting domain of RalGDS. *Nat. Struct. Biol.* **4,** 609-615.

[89] Schubel,U. Untersuchungen der Dynamik des HPr-Proteins von Staphylococcus carnosus mit Hilfe von heteronuklearen Kernoverhausereffekt- und Relaxationszeitmessung. 2000. Regensburg, University of Regensburg, Germany. (Thesis)

[90] Brünger,A.T. XPLOR Manual Version 3.1. 1993. New Haven, Yale University Press. (Computer Program)

[91] Meier,S. Kernresonanzspektroskopie am Histidine-Containing Phosphocarrier Protein aus Staphylococcus aureus: Strukturbestimmung und biologische Relevanz. 2000. Regensburg, University of Regensburg, Germany. (Thesis)