ELSEVIER

Progress in Nuclear Magnetic Resonance Spectroscopy 44 (2004) 33–96

PROGRESS IN NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

www.elsevier.com/locate/pnmrs

Automated structure determination of proteins by NMR spectroscopy

Wolfram Gronwald, Hans Robert Kalbitzer*

Institut für Biophysik und Physikalische Biochemie, Universität Regensburg, 93040 Regensburg, Germany

Received 19 September 2003

Contents

1. Introduction . . . 33
   1.1. Automated structure determination methods in the postgenomics area . . . 34
2. Automated structure determination by solution NMR spectroscopy . . . 35
   2.1. General aspects of high through-put structure determination . . . 35
   2.2. Fundamental steps in automated NMR structure determination . . . 35
      2.2.1. Target selection for NMR-spectroscopy . . . 36
      2.2.2. High through-put protein production . . . 36
      2.2.3. Optimized strategies for spectra recording . . . 37
      2.2.4. Automated NMR data evaluation and image analysis . . . 37
      2.2.5. Structure validation . . . 61
      2.2.6. Data deposition . . . 66
3. Classical bottom-up computer-aided structure determination . . . 66
   3.1. General strategies . . . 67
   3.2. Programs and program packages . . . 72
4. Automated top-down NMR-structure determination . . . 77
   4.1. General strategies . . . 77
   4.2. Molecule-centered approach (AUREMOL) . . . 79
      4.2.1. Overview . . . 79
      4.2.2. Databases and data structures . . . 80
      4.2.3. Preprocessing of experimental NMR spectra . . . 81
      4.2.4. Simulation of nD-NMR spectra . . . 81
      4.2.5. Structure based assignment and iterative structure determination . . . 84
      4.2.6. Knowledge driven assignment of NOESY spectra . . . 85
      4.2.7. Structure calculation and validation . . . 87
5. Conclusions and outlook . . . 91
References . . . 91

Keywords: NMR; Automation; Spectra assignment; Spectra simulation; Structure determination; Molecule centered

1. Introduction

A knowledge of the three-dimensional structures of biological macromolecules is the key for understanding molecular processes occurring in living systems. For a rather long time, the availability of suitable objects such as folded proteins was the main limitation for the application of the two most important experimental methods for structural determination at atomic resolution, X-ray crystallography and NMR spectroscopy. This has been changed in the last few years where a dramatic methodological progress has been made in biosciences. Especially, new developments in molecular biology and genetics allow the investigation of

* Corresponding author. Tel.: +49-941-943-2594; fax: +49-941-943-2479.

E-mail address: hans-robert.kalbitzer@biologie.uni-regensburg.de (H.R. Kalbitzer).

doi:10.1016/j.pnmrs.2003.12.002

previously inaccessible information such as the complete genome of whole organisms. In turn, structural biology has to adapt to these new developments.

1.1. Automated structure determination methods in the postgenomics area

The most prominent example for the new genomic area was the successful effort to decode the human genome. With the experimental methods created for solving this central problem, subsequently the elucidation of small genomes is now routine work and the number of available DNA sequences and hence of protein sequences has increased almost exponentially. However, to fully make use of the genomic information, it is necessary to know the three-dimensional structures of the encoded proteins. The spatial structures allow one to understand the function and course of biological processes on a molecular level, to establish previously unknown evolutionary relationships between large protein sequence families, and to investigate intermolecular interactions on an atomic scale. The last point is of particular importance to pharmaceutical research.

In contrast to the large number of available protein sequences only about 21,500 protein structures have been solved so far (date: 31.07.03). In addition, a large number of these deposited structures stem from identical or highly homologous proteins. The gap between the number of solved structures and the number of known protein sequences is huge and will continue to widen in the future since today the complete elucidation of whole genomes is essentially automated and thus almost routine work which can be performed in a few months. As an example, the protein database SWISS-Prot contains 141681 131945 proteins sequences (date: 14.01.03), the number of sequences deposited increases much faster than that of the structures deposited in the protein database (which is probably much smaller than the number of protein sequences solved since many are not accessible in public databases) (Fig. 1).

The two major methods for structure elucidation of large biomolecules are X-ray crystallography and solution NMR spectroscopy. Of the two methods, X-ray crystallography is older and more mature and allowed as early as 1958 determination of the first three-dimensional structure of a protein (myoglobin) [1]. In contrast, the first protein NMR structures, that of bovine pancreatic trypsin inhibitor, were solved almost 30 years later [2]. Both structural methods have their specific advantages and disadvantages so that they complement each other in many aspects. The main advantage of X-ray crystallography is that virtually no size limit exists for the system under investigation; on the other hand, only crystallizable systems can be analyzed. While NMR spectroscopy has the advantage that analysis is performed in solution under nearly physiological conditions and dynamic properties can be studied in detail.

As long as the computational methods are not sufficiently well-developed to predict an unknown structure for a particular protein sequence with high accuracy and reliability at atomic resolution, experimental methods for structure determination will play a dominant role in structural biology. These methods need to be optimized for higher efficiency to keep pace with the rapid increase of genetic information available. The only practical solution to this problem is a complete or almost complete automation of the experimental structure determination process. For X-ray crystallography, a rapid automated structure determination will be straightforward when the problem of automated protein production and crystallization is finally solved and when, for example, synchrotron radiation and anomalous scattering of seleno-methionine enriched samples are used. However, there are classes of proteins which will probably not be crystallizable such as proteins which are only partly folded or which exist as multiple fast exchanging conformers. The alternative method, solution NMR

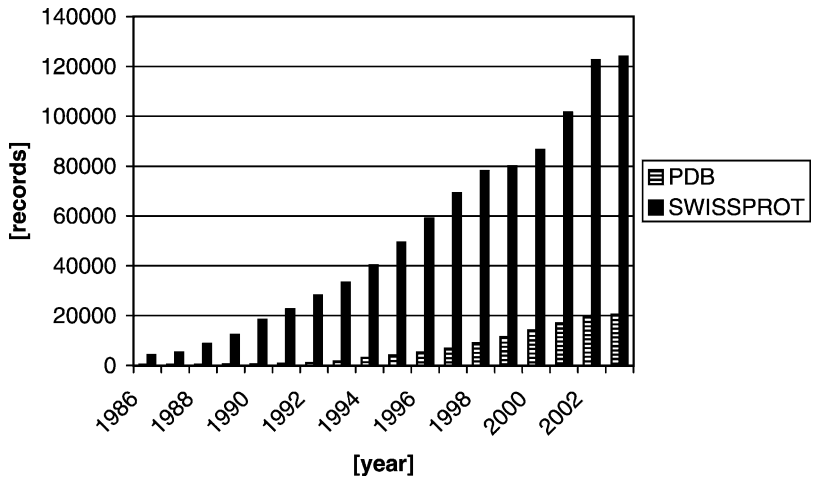

Fig. 1. Protein sequences and structures deposited in databases. (Black) Number of sequences deposited in SWISS-Prot. (Stripes) Number of three-dimensional structures deposited in the protein database PDB (Rütgers, formerly Brookhaven). Entries are listed until mid 2003.

spectroscopy will probably only begin to play a major role in structural genomics when automation decreases drastically the time necessary for a structure determination and allows medium to high-throughput. In the following we will try to describe the new developments which are required for this aim and discuss and review the progress that has been already made in this field. We will also discuss more specifically the project AUREMOL (to be published) developed at the University of Regensburg in cooperation with a major manufacturer of NMR instruments which is aimed to solve the problem of automated NMR structure determination.

2. Automated structure determination by solution NMR spectroscopy

NMR structural determination of small well-behaved proteins (well soluble, globular and uniquely folded) is nowadays a manageable scientific problem which leads at the end to a safe solution. However, an expert must be involved and needs several months for completing the structure. This is in general not acceptable in proteomics research where a large number of structures will have to be solved essentially automatically.

2.1. General aspects of high through-put structure determination

High through-put NMR structure determination has much in common with high through-put crystallography (Table 1), some steps in this process are virtually identical for the two methods. Automated structure determination implies that essentially all steps can be fulfilled with one fundamental strategy and that no experts are required for solving unexpected specific problems or for devising new strategies. This implies that always only a subset of all existing proteins is amenable to automated structure determination because there are always limits in the methodology which exclude some proteins from automated structure determination.

Table 1
Comparison of high through-put structure determination by NMR and X-ray crystallography

Step #	NMR-spectroscopy	X-ray crystallography
1	Target selection	
2	Protein expression	
3	Protein purification	
4	Isotope labeling	Seleno cysteine incorporation
5	Buffer optimization	Crystallization
6	Data recording	
7	Data evaluation	
8	Structure calculation and refinement	
9	Structure deposition	

In target selection for NMR spectroscopy solubility, size, and lack of significant unspecific aggregation are the main determinants. In X-ray diffraction, size does not play a role but usually only well soluble, non-aggregating proteins crystallize properly. A uniquely folded state is usually required for crystallization, whereas NMR spectroscopy can also deal with proteins which are partly unstructured.

Robotized high through-put methods for protein expression are now under development and are required for the two methods equally. Once the protein is available, sample preparation is usually easy for NMR but often provides a bottle-neck in X-ray diffraction since crystals are required. Here, major efforts are being made to automate crystallization procedures. Data collection is easy to automate for both methods. Although the total recording time of a minimal NMR data set probably can be substantially reduced, it will be difficult to decrease it to the short time required by crystallography when synchrotron radiation is used. Data evaluation is the true bottleneck in NMR spectroscopy and major efforts should be made to solve this problem. In X-ray spectroscopy this aspect is almost routine and a first structure can be obtained within a few hours after recording the data. Structure calculation has many similarities in the two methods; however, a main drawback in this regard is the fact that in NMR spectroscopy the spectra cannot be simulated satisfactorily from the structure alone. In contrast, the calculation of diffraction patterns from structures is simple and can be performed exactly. Structure validation such as determining the stereochemical quality of the results follows similar routes in the two methods.

In conclusion, we have to keep in mind that a selection and definition of the class of proteins and their properties is mandatory in NMR spectroscopy when we want to create a working method for automated structure elucidation. This also means that we have to apply the whole set of existing specialized methods and possibly have to generate new methods if a particular protein is not a member of the class of proteins actually solvable by automated methods. Here again, an expert is required. However, in the long term the number of special cases will decrease when the methods have been refined.

In the context of structural genomics, the type of automated NMR structure determination really required has to be distinguished from already existing partly optimized ‘automated’ methods and has to compete with those commonly used in X-ray crystallography which is much easier to automate.

2.2. Fundamental steps in automated NMR structure determination

Automated structure determination in solution can be separated in the main steps: (1) target (protein) selection, (2) protein production and isotope labeling, (3) data

recording, (4) data evaluation, (5) structure validation, and (6) submission of the structure to a database. For highest performance, these main steps should be performed sequentially since repeating some of the main steps causes unnecessary delays in the structure determination. However, in practice it often happens that the general procedure has to be restarted. When working with protein domains it is only after the first structural information (step 4) is available that educated guesses may help to find the optimal length of the construct or to define mutants with better spectroscopic properties (step 1). In other cases, insufficient data during the data evaluation (step 4) may require the recording of new spectra (step 3) or even the production of protein with different isotope labeling.

2.2.1. Target selection for NMR-spectroscopy

Suitable target selection is one of the most important factors determining the success of a structural project. It depends largely on the specific goal one has. In the case of traditional structural genomic projects, the typical goal is an even coverage of the fold space but other goals are also in the focus of newer structural genomics programs. The motivation behind the search for new folds is the hope that with a complete set of folds all proteins can be modeled by homology modeling techniques. Currently, it is expected that between 1500 and 5000 distinct stable folds exist [3]. Based on sequence similarity, proteins are grouped in sequence families and one tries to solve the three-dimensional structure of at least one member of each family with either X-ray crystallography or NMR spectroscopy [4]. In this regard, it is also wise to screen the possible candidates for properties allowing a rapid structure elucidation such as good solubility (>1 mM), sufficient stability under conditions typically used for NMR spectroscopy (>1 week at 298 K), negligible unspecific aggregation, a unique fold, and in the case of NMR spectroscopy limited size (<25 kDa). This screening also includes the definition of the optimum domain borders in the case of large proteins and still needs to be done mainly experimentally since safe prediction of these properties is not yet possible. When automated methods for NMR structural determination are being applied, the last aspect is most important since the obtainable spectral quality and completeness of the data determines the success of the approach.

Other goals for high through-put structural determination of proteins, which are at least as interesting as the fold recognition projects, are the elucidation of an almost complete set of structures coded in a small viral or bacterial genome, or the investigation of the protein structures of important classes of proteins independent of the species they originate from. Here, for non-membrane-bound proteins a screen of selected proteins from several different species increased the output from typical $\sim$50% soluble proteins to more than 90% [5]. In summary, in the field of target selection substantial methodological development of bioinformatical tools and experimental screening methods will be required in the future.

2.2.2. High through-put protein production

Establishing the automated production of proteins is mainly necessary for two different reasons: (1) experimental optimizations of protein properties such as solubility and minimum aggregation tendency require the simple and fast production of protein varieties. (2) High through-put NMR methods are dependent on mass production of proteins. For task (1), in principle, only low quantities of protein have to be produced; for task (2), protein has to be produced in mg quantities and usually has to be isotope enriched for NMR spectroscopy. This implies that different techniques may be optimal for fulfilling these tasks.

Automated production of expression constructs for genes without introns should be straightforward, while for expression constructs of intron containing genes full-length cDNA clones are required. Libraries of full length cDNA clones are currently developed and also tools are available for finding a suitable cDNA library for a specific task, e.g. (http://cgap.nci.nih.gov/Tissues/Tissues/LibraryFinder). For automation purposes, it will probably be necessary to attach at least one affinity tag to the protein and to isotopically enrich by growing the bacteria in isotope enriched minimal media or special commercial full media. Proteins with disulphide bonds and proteins that require glycosylation or other post-translational modifications are often difficult if not impossible to obtain from expression in *E. coli*. In these cases, yeast expression systems such as *Pichia pastoris* can be used [6]. Baculoviral systems [7] in insect cells or mammalian cell cultures are only used when the target protein cannot be expressed in other systems since mass production of isotope enriched proteins is extremely expensive here.

Cell-free expression systems [8] have a very large potential for automated production of proteins at small or intermediate scale. They have the principal advantage that interference of a toxic target protein with the cell metabolism cannot occur and that the environment during the protein expression can easily be manipulated by addition of molecular components such as protease inhibitors or chaperons [9]. High yields of proteins can be obtained under favorable conditions in the combined protein transcription translation assay [10–15] and optimized as described in Refs. [16–19]. Up to 8 mg protein/ml translation assay can be expressed with this method [20]. When isotope enrichment is necessary it can be as cost effective as expression in *E. coli*. When amino acid type specific labeling is necessary in vitro translation is extremely powerful since virtually all the labeled compounds are introduced in the target protein. Also site-specific labeling is possible by using the amber stop codon [21]. However, this is not yet a routine method applicable in automation.

Specific isotope labeling can simplify the assignment procedure. An example is the fast identification of certain

amino acid types from simple 2D $^{1}H-^{15}N$ HSQC spectra by selective amino acid labeling [22]. However, for each amino acid type a separate sample is required. Since high throughput NMR structure determination aims to minimize the spectrometer time required, these specific labeling schemes are not useful for smaller proteins but may be helpful when the structure of larger proteins has to be solved automatically. Here, also the intein method for regio-specific isotope enrichment is promising [23–27]. However, in contrast to the initial expectations in our experience it is far away from being a generally applicable routine method.

2.2.3. Optimized strategies for spectra recording

The number of pulse sequences published which are meant to improve NMR structure determination of proteins increases steadily with time. A good overview of the pulse-sequences currently used for the structure determination of biological macromolecules in solution is given in Ref. [28]. However, in the context of automated structure determination only a small number of experiments is necessary. When defining a minimal set of NMR-experiments, it is obvious that the experiments which contain the necessary structural information are indispensable. That is actually at least one experiment relying on dipolar couplings (NOEs or residual dipolar couplings), although in the long-term chemical shift information together with molecular modeling techniques may be sufficient [29]. Actually, it is not yet settled what the minimal set of experiments is and it is obvious that this will depend on the software approach used. An additional important parameter is the complexity of the problem which mainly depends on the size of the protein under consideration and its spectral dispersion. Both, experiments and programs have to be optimized simultaneously with respect to the problem encountered. As an example isotope enrichment with ^{13}C seems not to be necessary for small proteins but is probably mandatory for larger proteins.

Higher dimensional experiments are, in principle, useful for automated data evaluation since the main problem in automation remains ambiguity. However, they increase the minimum spectrometer time required. Here, reduced dimensionality 3D and 4D triple resonance experiments may be useful [30–32]. Using these experiments, it is possible to reduce by one the number of dimensions compared to the corresponding conventional experiment. It is based on a projection technique where the chemical shifts of the projected dimension are encoded as an in-phase doublet splitting.

To facilitate semi-automatic assignments using these experiments the program SPSCAN [30] was developed and this includes a peak picking routine adapted to the observed peak patterns and allows the mutual interconversion of frequencies detected in conventional and reduced dimensionality spectra, respectively. Using a best first method, a search for adjacent spin-systems is performed to help the user in the interactive sequential assignment process.

Recently, the so-called GFT approach [33] was developed by the same group to reduce the required amount of NMR time. In this approach a joint sampling of several indirect dimensions is applied leading to so-called chemical shift multiplets, where the individual chemical shift values can be obtained from a suitable combination of the various multiplet components.

With a new approach described by Frydman et al. [34], it is in principal possible to acquire multidimensional spectra with a single scan allowing a drastically reduction in measurement time. The key of this method is a position dependent evolution of the indirect dimension(s) using pulsed field gradients. However, in practice due to the limited signal to noise ratio of this approach usually more than one scan will most probably be necessary.

Experiments that are selective to the amino acid type can be used to resolve ambiguities in the assignment process. A set of two-dimensional triple resonance $^{1}H-^{15}N$ correlation experiments is presented to achieve this goal [35–37]. They are based on incorporation of the MUSIC [38] pulse sequence elements in triple resonance experiments. MUSIC basically accomplishes an in-phase magnetization transfer for either XH_2 or XH_3 groups, while for other multiplicities this transfer will be suppressed (X can be either ^{13}C or ^{15}N).

Two-dimensional versions of CBCACONH experiments can also be used to select for different amino acid types. The experiments are based on the existence or absence of the $^{13}C\beta-^{13}C\gamma$ coupling in a certain residue type. Therefore, these experiments are selective for groups of residue types and not for one specific type [39]. This approach was also adapted to the use of deuterated samples [40]. By incorporating phase labeling techniques into standard triple resonance experiments used for sequential assignment it is possible to obtain information about the type of residue [41] in addition to connectivity information [42].

2.2.4. Automated NMR data evaluation and image analysis

Besides the optimization of the protein production, automated NMR data evaluation has the highest potential for substantially reducing the total time needed in automated NMR structure determination. Independent of the specific strategy used in the automated NMR structure determination, the analysis of the multidimensional NMR data comprises the following steps that can be viewed as a special problem of image analysis. Image analysis is usually characterized by three different stages of operations: (1) data processing including improvement of image quality and feature enhancement, (2) pattern recognition and classification of objects, and (3) interpretation of objects and classes of objects.

After recording a set of multidimensional spectra, the proper processing of that data (i.e. improvement of image quality and feature enhancement) is the first critical step, since all subsequent operations are based on the information obtainable from the processed spectra. Optimal processing

of the data is especially important in automated data evaluation since computer programs are usually not as good as human experts in distinguishing artifacts from meaningful signals. Although the full data analysis process could be performed, at least in principle, using the complete NMR data matrices, the computational efficiency is significantly increased when the resonance peaks are recognized and isolated from the noise and artifacts (separation of the relevant objects from the background). This separation leads to a large reduction in the size of the data matrices that must be handled by the computer since consecutive operations can now be performed exclusively on these objects, which can often be sufficiently described by a few parameters such as spectral position (or chemical shift), peak intensity, integrated area or volume and line shape. In the next step, the spectral peaks must be assigned to multiplet and spin system patterns (classification of objects). Finally, these partial solutions must be combined to form a consistent solution which contains the complete (or nearly complete) resonance assignment and an exhaustive interpretation of all structure relevant information, e.g. observed NOEs, *J*-couplings, residual dipolar couplings, hydrogen bonds, secondary chemical shifts, and relaxation data (i.e. interpretation of objects and classes of objects). In the context of this article, the last part of the interpretation of the objects would then include the calculation of three-dimensional structure of the protein.

2.2.4.1. Data processing. Careful processing of raw NMR data is very important since the processed data determines the quality of the results of peak recognition and other data reduction procedures. Moreover, after the data reduction step is performed, all information not recognized as part of the spectra is lost from all subsequent analysis procedures. Several multidimensional NMR data processing packages have been developed in the past. AZARA [43], DELTA [44], FELIX [45], GIFA [46], NMRLAB [47], NMRPipe [48], NMR Toolkit (Hoch, 1985), NMRZ [49] (New Methods Research Inc., Syracuse, NY), Pronto [50], PROSA [51], TRIAD [52], TRITON (Boelens, unpublished), VNMR [53], and XWINNMR [54].

Enhancement of spectral quality. Usually, a single method of image enhancement which is optimal in every respect does not exist, although each method has advantages. In computer aided spectral evaluations, as in manual evaluation, it can be useful to compare the results for spectra processed in various ways. This approach is best exemplified by the time-domain filtering process discussed below. The enhancement of the spectral quality always depends on (often not obvious) additional knowledge about the system, such as the expected line widths of the signals or the frequency distribution of the noise. It improves as more information is available and is used for this purpose. A simple example is time domain filtering of NMR data before Fourier transformation. The same procedures can also be performed in the frequency domain by convolution of the Fourier transformed data with the Fourier transform of the filter function. Although the two methods are fundamentally equivalent, time domain filtering is computationally much more efficient, and hence is usually preferred. In practical applications, one usually starts in the time domain, applies time domain image enhancement methods, Fourier transforms the data, and then continues with frequency domain methods. However, this sequence is not the only conceivable one because it is possible to jump between time and frequency domain at will with the aid of forward and inverse Fourier transformations without information loss (apart from usually insignificant rounding errors).

Time domain filtering. Appropriate time domain filtering of the data is one of the most important steps performed prior to Fourier transformation. The key assumption used in these filtering methods is that resonance signals, noise, and artifacts have different time-constants so that their contribution to the total detection signal varies during the acquisition period. Accordingly, a reduction in the intensity of the initial part of the time domain signal decreases contributions from component signals which slowly vary in the frequency domain, such as baseline rolls and tails of resonance signals. A reduction in the intensity of the final segments of time domain signal decreases the intensity of rapidly varying components such as instrumental noise and as a consequence enhances the signal-to-noise ratio but also increases the line width (line broadening). These effects are discussed in detail by DeLikatny et al. [55] for the sine bell function. The choice of the filter function depends on the type of acquired spectra and the kind of information desired. An important example is the computer-aided extraction of *J*-coupling constants from the separations between resonance peaks. In this case, it is necessary to obtain the smallest possible line width, even at the cost of decreased signal-to-noise ratios so that the individual multiplet components are clearly resolved. On the other hand, peak-picking, multiplet recognition, and pattern recognition are controlled by the signal-to-noise ratio, therefore a slight line broadening is usually acceptable. In TOCSY and NOESY spectra of macromolecules, the multiplet structure of the in-phase components is only barely resolved and a maximum signal-to-noise ratio is usually required to detect even weak signals, so that in general it is possible to adjust the window functions to larger line widths. For practical purposes, the Lorentzian-to-Gaussian transformation is well-suited for such applications. When a good estimate for a line width is available, a single parameter then defines the resulting filtered line width [56–58]. For any predetermined line-broadening, the optimal suppression of truncation errors can be obtained by use of the so-called Dolph-Chebycheff window. However, due to its complexity this window is normally not used, but it is useful for evaluating the efficacy of other filter functions. Maximum signal-to-noise ratio is achieved by applying a matched filter function prior to Fourier transformation. The matched filter is equal to

the envelope function of the time domain signal. In an ideal solution experiment, the signal can be described as the sum of exponentially decaying sinusoids. Therefore, if sufficient data has been recorded to minimize truncation artifacts, e.g. in the acquisition domain optimal sensitivity can be obtained by applying a matched exponential filter function [59].

Time domain manipulations for ridge suppression. The first row and column of the time domain data matrix must, in accordance with the initial delay, be scaled for proper integration during the FFT. The first FID must be multiplied by $C_1 = D_1/2\Delta t_1$ [60], where D_1 is the smallest t_1 variable delay and Δt_1 is the sampling interval. Analogously, the first point of every FID, i.e. the first column of the time domain data matrix, must be multiplied by $C_2 = D_2/2\Delta t_2$ when D_2 is the delay between the last pulse and the start of acquisition and Δt_2 is the sampling interval in t_2. However, the intensity of the first point is also influenced by the dead time of the receiver and the response of the analog filters which strongly attenuate the signal. Therefore, in practice, application of a modified multiplication factor, C_2' (e.g. $C_2' = 6.6$) is recommended [60]. Alternatively, scaling of the first row can be omitted if D_1 is chosen to equal 1/2 Δt_1 ($C_1 = 1$) which has the additional advantage that the spectrum is easier to phase [61].

Oversampling. Oversampling of the NMR spectra was first proposed by Delsuc and Lallemand [62] as a mean to improve the detectability of very weak signals, for removing folding artifacts and for improving the baseline [63]. Using oversampling the demands placed on the analog audio filters being used are considerably reduced. It is simply the recording of a spectrum with time increments Δt_1 which are smaller than required by the Nyquist theorem at a given spectral width $\Delta\nu$

$$\Delta t_1 = \frac{1}{2\Delta\nu} \tag{2.1}$$

It can be performed with any NMR spectrometer and the degree of oversampling possible depends only on the speed of the analog-to-digital (AD) converter. However, the size of the time domain data to be stored is proportional to the degree of oversampling and can be very large. A simple (and in principle optimal way) would consist of a fast forward Fourier transformation followed by a fast backward Fourier transformation of the data of the spectral range of interest only.

A faster way to reduce the data size of the oversampled data consists of the digital frequency filtering of the time domain data before storage on the disc. To filter out frequencies above a certain frequency from the time domain signal, the signal must be convoluted with the Fourier transform of the rectangular function, a sinc function. After digital filtering decimation of the data is used which eliminates each nth data point, where n is the degree to which the data have been oversampled [64]. The corresponding program is implemented in hardware of commercially available spectrometers directly after the AD-converter. Since this time domain filtering is not perfect artifacts at the edges of the spectra are usually observed. They can be reduced by a subsequent baseline correction. Also it has been shown that linear prediction algorithms benefit from oversampled data [65].

Frequency domain filtering. Frequency domain filtering has the advantage that it is performed on the Fourier transformed data so that various filters can be rapidly tested with the same frequency domain data. Examples of such filters include the polynomial filters where each point, x_i, is replaced by x_i'

$$x_i' = \sum_{n=-N}^{N} a_n x_{i+n} \tag{2.2}$$

where a_n ($n = -N, \ldots, N$) is a series of coefficients defining the filter [66]. The most common of these filters is the moving average filter (defined by $a_n = 1/2(N+1)$) which leads to a smoothing of the spectrum. Experience shows that time domain filtering gives superior results and is preferred for the preparation of spectra to be used for automated pattern recognition.

Base plane correction in the frequency domain. A flat base plane is not only important for the correct integration of multidimensional NMR spectra, where base plane variation can dominate the integral, but also for peak recognition where a threshold must be defined in order to sort resonance peaks from noise spikes. The fundamental assumption used in this process is that the base plane is flat in the absence of signals and that the slopes of resonance peaks are greater than those of base plane artifacts. Published base plane correction methods differ in the functions used for approximating baseline artifacts and in the way regions where the ideal baseline should be zero are defined. Those regions which contain no cross peaks can either be defined by the user [67–69] or identified automatically by the program [70–72]. When the user defines the regions where the base plane should be flat, external information can be incorporated, e.g. evidence derived from other experiments, such as well-resolved 1D spectra. Incorporation of this additional information potentially leads to improved results. However, the methods which are more convenient for the user are those which automatically identify base plane points. At least for spectra with similar signal-to-noise ratios, line widths, and spectral resolution, these automated routines work well. However, a few general parameters must be adapted to the experiment (i.e. external knowledge must be incorporated into the program). The simplest base plane correction method fits the baseline of each row to a cubic Lagrange polynomial where only three reference columns which contain no signals, are defined [68]. After correction of all the rows, the same method is applied to the corresponding columns. A similar method is implemented in the program,

XWINNMR, where the baseline points are automatically identified and the baseline is fitted to a polynomial of up to sixth order.

Better results are obtained using the spline method [67], where an arbitrary number ($n > 4$) of cross-peak free rows and columns can be defined. The spline function then approximates the base plane between two neighboring points using a cubic polynomial function. A simple variation of this is the sectionally linear interpolation method [69]. Here, the base plane is approximated by short sections of straight lines. This method has the advantage of being computationally very fast and avoids over correction which often results from cubic interpolation.

In the case where the baseline points are not defined by the user, the performance of the baseline correction program critically depends upon the quality of the automated identification of those points. Two published procedures are both based on the assumption that small stretches of baseline can be fitted by a straight line [72], or have a first derivative significantly smaller than regions containing peaks [70]. In an another procedure, the standard deviation of the signal intensities in a small window of data points is used to decide if a particular data point belongs to the baseline [71]. After selection of specific baseline points, the programs then calculate from these points fragments of a smoothed spectrum connected by straight lines to approximate the baseline [71] or a fifth order polynomial [70] or a sum of cosine and sine functions with various amplitudes and frequencies [72] is used for baseline approximation. The approximation used in the latter method, e.g. the program FLAT, appears to be somewhat more appropriate than others since most base plane distortions originate from intensity fluctuations in the first points of the FID. These intensity changes can be viewed as an additional time-domain signal consisting of a few non-zero points superimposed onto the unperturbed FID. Fourier transformation of such a signal results in a sum of cosine and sine functions which are therefore well-suited for approximating the resulting baseline artifacts.

Removal of spectral artifacts. A typical artifact which often dominates spectra recorded using older instruments is t_1-ridges. A simple method for their attenuation, the mean row subtraction, was devised by Klevit [73]. In this method, the user defines a region, usually several rows near the border of the spectrum, where no cross peaks, but t_1-ridges are present. The mean of these rows is calculated and then subtracted from all other rows. This method initially was devised for absolute-value spectra, but was later generalized to include phase-sensitive spectra [74–76].

Phase-distortions which originate from a delayed acquisition, unavoidable in experiments using soft pulses, can be corrected by a computational projection back to time zero (backward prediction, see also linear prediction) [77,78]. The oscillatory components of ridges that originate from truncation effects can be effectively removed by a frequency domain filter developed to suppress periodic features [79].

The signal of the physiological solvent, H_2O, is by far the most intense feature in ^{1}H NMR spectroscopy of biological macromolecules and causes spectral artifacts even where strongly attenuated by pre-saturation or selective excitation. The dispersive tails of the water resonance can be largely removed from spectra by fitting these tails to a hyperbolic function which is then removed from the data. After interactively defining cross-peak-free data points on the water tail, the hyperbolic function is fitted to these points. The fit is further improved by including a linear and a constant term to account for baseline variations arising from other sources [80]. A computationally simpler method, similar to the diagonal peak suppression method for phase-sensitive COSY spectra [81], makes use of the fact that the water resonance is usually positioned at the center of the spectrum (i.e. at $\omega = 0$). Therefore, its time domain signal is a non-modulated exponential. The contribution of the water signal is reconstructed by filtering out the oscillatory parts of the FIDs and then subtracting those parts from the original FID [82]. The dispersive tails of the water resonance can also be suppressed by phasing the water signal in absorption mode, zeroing the relatively small absorption signal in the frequency domain data, discarding the imaginary part and regenerating the signal from the processed real part via a Hilbert transformation. After phase-correction of the spectrum, the water signal is largely suppressed [83]. Another possibility is to calculate the second derivatives of the FIDs and to Fourier transform these which also suppresses signals at $\omega = 0$, such as the solvent peak [84]. Application of the Karhunen–Loeve transformation to multidimensional data can be used for the removal of the strongest signals, which are usually the solvent resonances, from the data matrix [85]. Similar results can be obtained from a principal component analysis of the frequency domain data [86] or a linear prediction of the time domain data (see below) and removal of very strong singular values (signals) [87].

Another possibility for water peak suppression is the wavelet transformation, which allows one to decompose a signal in terms of elementary contributions called wavelets. By discarding components corresponding to low frequencies before data reconstruction the water signal can be suppressed (assuming that the water is located in the middle of the spectrum) [88–90]. Independent component analysis appears to be a promising ansatz for the reduction of base plane artifacts since no spectrum dependent parameters have to be adjusted [91,92] (Fig. 2). However, details of its application are still under development.

Symmetry enhancement. Inherent symmetries in multidimensional spectra provide redundant information useful for discriminating resonance signals from noise and artifacts. Under ideal conditions, i.e. sufficient digital resolution, identical filtering and equal digital resolution in each dimension, and proper phasing, absorption peaks have C_4 symmetry. This local symmetry is useful for enhancing moderately isolated peaks which show no

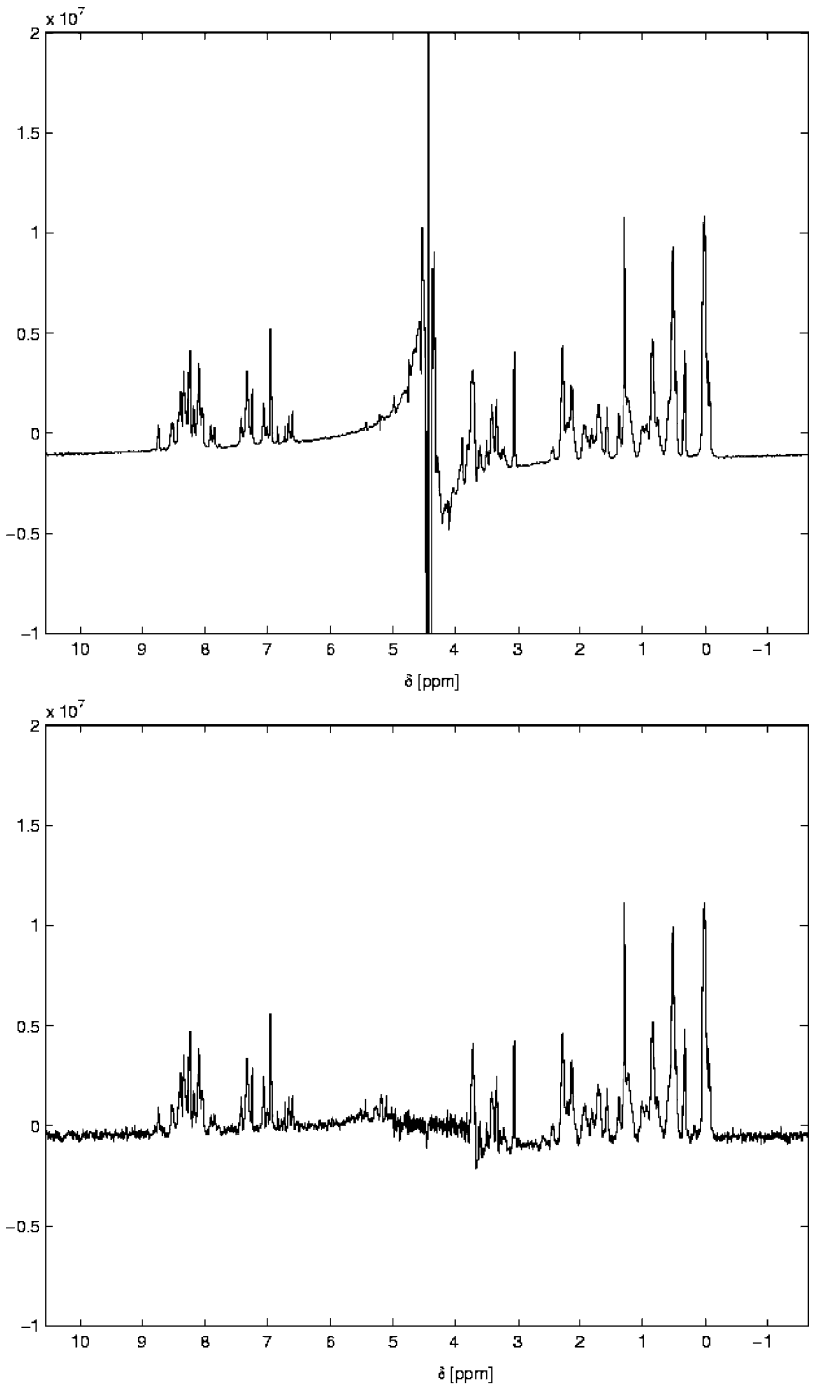

Fig. 2. Artifact reduction by independent component analysis. (Top) First trace of the experimentally determined ^{1}H 500 MHz 2D-NOESY spectrum of the 24 residue peptide P11 measured in 90%/10% H_2O/D_2O (v/v). For water suppression presaturation was applied during the relaxation delay and during the mixing time. (Bottom) Reconstructed P11 spectrum with the water artifact removed using independent component analysis.

overlap with other signals [93]. A more powerful method of enhancing pertinent information is the exploitation of global symmetries such as cross-peak symmetries about the main diagonal present in many homonuclear 2D spectra or multidimensional spectral planes. A more general way of using global symmetry information consists of comparing areas of symmetry-related peaks at positions (i,j) and (j,i), calculating a 'match factor', m, which is a measure of the symmetry, and then modifying the original intensities $I(i,j)$ and $I(j,i)$ to $I'(i,j)$ and $I'(j,i)$ in accordance with the match factor [74,94,95].

The match factor can be defined in a way that it adopts a value of 1, if the two related signals possess the same shape and intensity and is negligible if the shapes do not correlate. In fact, all published symmetrization procedures are special cases of this general formalism. All symmetry enhancement

procedures work best when the symmetry of the raw data is optimized. This means that, when applying symmetry enhancement algorithms, identical zero-filling and filter functions should be used for each dimension. Other artifact reduction methods should be applied before the symmetry enhancement procedures. As for conventional time-domain filtering, symmetry enhancement methods must be carefully adapted in accordance with the information desired. This is especially true for homonuclear NOESY spectra which are asymmetric by definition when a finite relaxation delay is used. In this case symmetrization procedures can cause loss of information.

Linear prediction and related methods. In high-resolution NMR the frequency domain line shapes are closely approximated by a Lorentzian function which corresponds to a cosine-modulated exponential in the time domain. This property is useful in peak fitting procedures applied to experimental data as discussed by Gesmar and Abildgaard [96]. These fitting procedures can be used either to reduce artifacts such as truncation wiggles or to fit and describe all cross peaks in a spectrum. In the latter case, these fitting procedures simultaneously represent peak recognition methods.

Linear prediction [65,96–107] is a method for directly obtaining resonance frequencies and relaxation rates from time domain signals, which are a superposition of exponentials, by solving the characteristic polynomial. Phases and intensities, however, must be calculated iteratively using a least square procedure. In the presence of noise the total number of exponentials assumed must be greater than the number of cross peaks expected. In the one-dimensional case, depending on the algorithm used, the number of operations is roughly proportional to mn^2, where n is the number of unknowns and $m(m > n)$ is the number of data points used for the prediction. Thus, the computational complexity increases rapidly as the number of resonances to be observed increases.

Most of the methods proposed for simplifying linear predictions are based on a reduction in dimensionality. A simple method consists of limiting predictions to only a part of the n-dimensional data matrix. In typical applications, the FIDs are first Fourier transformed as a function of the time variable of the acquisition dimension (that is, t_2 in the 2D NMR spectra and t_3 in 3D NMR spectra) since the number of data points, m, and the number of resulting resonance frequencies, n, to be considered is usually rather large. The columns of the data matrix obtained are then analyzed using linear prediction methods [104,108–112]. The speed of the prediction process is significantly improved since the number of data points, m, in the remaining direction is usually small (especially in 3D and 4D spectra), and only a limited number, n, of the resonances contribute to the signal and so must be considered.

In macromolecular multidimensional NMR, linear prediction is most often used in the indirect dimensions of 3D and 4D data sets, where the experimentally obtainable resolution is usually rather limited, since it avoids truncation errors and leads to an increase in resolution. With increasing computer power, it has become feasible to use two-dimensional linear prediction approaches for data that are severely truncated in both dimensions, e.g. planes of 3D and 4D spectra [106]. Spectral distortions which arise from delayed acquisition and non-linearities of the receiver can also be corrected by replacing the first points of the FID by applying a backward prediction. When only those frequencies in a restricted spectral window are of interest, the prediction can be accelerated using the LP-ZOOM [100,105] or the VAPRO method [113]. In cases where the signal-to-noise ratio is low a priori knowledge about the expected frequency intervals of the damped sinusoids can be used to obtain reliable predictions [114]. Other line fitting methods which do not necessarily rely on the assumption of combined exponential functions are the HSVD and LPSVD methods [97,115–117]. Alternatively, fitting of the data can be performed in the frequency domain [118].

Maximum entropy reconstructions and related methods. The maximum entropy method (MEM) [119] has attracted considerable interest as an alternative to Fourier transformation [120–157]. The principle behind maximum entropy reconstructions involves finding all spectra which are consistent with the experimental data (as tested, for example, by the χ^2 test) and identifying the spectrum that has the minimum information content, or equivalently, the maximum entropy.

The following advantages of that approach are purported to include that the information content of the spectrum is used in an optimal and unbiased way, free from any a priori assumptions. Also, since one starts the calculation from a uniform or random distribution of frequencies, the 'true' solution is selected by the entropy, a notion which suggests an absolute physical measure. Finally, compared to Fourier transformation, a simultaneous enhancement in both sensitivity and resolution seems to result from MEM reconstructions. In contrast to linear prediction methods, no assumptions about line shape are usually applied. Therefore, MEM can be applied to spectra with components having unknown line shapes. Furthermore, the computing time needed for a MEM reconstruction does not depend on the number of resonance lines, but only on the size of the data set. Although 'entropy' in physics and information theory are well-defined terms, this is not true for its application to complex-valued NMR data. The Shannon entropy, S, can be applied in a straightforward manner only for real positive functions as they occur in standard image reconstructions. The entropy, S, is defined as

$$S = -\sum_{i=1}^{M} p_i \ln p_i \tag{2.3}$$

where p_i is the probability of the state i and M is the total number of such states. For real-valued NMR spectra, p_i can

be simply replaced by the normalized intensities x_i/b at position i in the spectrum [121–123,138]. The normalization constant, b, can be defined as the sum of all intensifies x_i. For complex-valued NMR data, the treatment of probabilities, p_i, as real quantities incorporated into the definition of entropy cannot be simply defined using the intensities. Therefore, several different notions of entropy in this context have been proposed [125,131,132,145,146,148,150,151].

Although typical MEM reconstructed NMR spectra seem to have much better signal-to-noise ratios with simultaneous resolution enhancement compared to conventionally processed spectra, a closer inspection shows that this is only partially true [138,147,150,151]. In simple cases, the noise-suppression in MEM reconstructed spectra is only cosmetic when compared to the corresponding Fourier transform reconstruction. Moreover, equivalent results can be obtained by using a non-linear plotting scale on the vertical coordinate and by applying a threshold to the data (i.e. setting points with intensities lower than a given threshold to zero) [126,147,150]. Similar resolution enhancement can be obtained by appropriate filtering of the data. In one application, $^{13}C\alpha-{}^{13}C\beta$ splittings in protein triple resonance spectra are eliminated by deconvolution with MEM reconstruction [157]. An advantage is that in cases were broad and narrow lines occur, MEM leads to an optimal representation in one spectrum, whereas conventional processing would require two different transformations [126,147]. However, this is not surprising, because in the traditional MEM processing no additional information is incorporated during the data evaluation.

The situation changes when varying amounts of supplementary information is included, leading in the extreme case to the more general framework of Bayesian analysis of which maximum entropy analysis is only a special case [127,129,140,149]. A simple example is the reconstruction of strongly truncated data, where, in contrast to zero-filling followed by Fourier transformation, the information that the FID is not simply zero after time t, is an inherent part of the MEM reconstruction, therefore, truncation ripples can be avoided as in the case of linear predictions [122,123,126,131,133,145].

Bayesian statistics together with Metropolis Monte Carlo simulations are used to determine parameters like coupling constants from time domain data [158]. It is based on the comparison between a model and real data. Time domain data are modeled as a linear combination of exponentially damped sinusoids. The parameter space of the model is searched using Metropolis Monte Carlo simulations and hereby employing Bayesian statistics to determine the probability of the current set of parameters.

Non-linear sampling of the data promises a somewhat better signal-to-noise ratio than equidistant sampling of the data according to the Nyquist theorem. The spectrum from these non-linearly sampled data cannot be reconstructed by the FFT, and so the MEM method appears to be best-suited in such cases. Exponential sampling was used in the first applications of non-linear sampling [134], and this was later generalized to more variable sampling schemes [136,152–154]. However, similar results can be obtained with conventional processing in conjunction with application of the CLEAN algorithm used in astronomy [137,138]. Line width information can also be included in MEM reconstructions [130,141,144]. MEM can also be used for suppressing zero-quantum peaks in NOESY spectra [136] and removing baseline artifacts from acoustic ringing or pulse breakthrough [133,143]. Another reconstruction method related to MEM is the maximum likelihood deconvolution method [159–161]. Using a least squares procedure 'maximum likelihood' minimizes the variance between the measured FID and the parameterized data model. However, the entropy is not maximized in these approaches. In the ChiFit [161] method, data are modeled by a linear combination of exponentially damped sinusoids. It was applied to enhance the resolution in the indirect dimensions of 3D and 4D NMR spectra. Other reconstruction methods related to MEM are the constrained iterative spectral convolution [142] and the parametric estimation using simulated annealing [162]. In general, all of these methods give results that are similar to MEM reconstructions; the selection of the optimal method depends on the problem under consideration (and on the availability of the corresponding software).

In a recent article, a detailed comparison of linear-prediction extrapolation and maximum-entropy reconstruction was performed on simulated and experimental data [163]. It was concluded that in most cases maximum-entropy is superior to linear-prediction although linear-prediction is much more widely used. This is especially true for the accuracy of peak positions and the introduction of false peaks. In addition, the ability of maximum-entropy to accommodate non-linearly sampled data can provide significant improvements in both sensitivity and resolution for short data sets, compared to linear sampling.

Filter diagonalization. The multidimensional filter diagonalization method [164,165] offers an alternative to standard Fourier transformation of multidimensional spectra. The aim of the method is to obtain good resolution in the indirect dimensions even when only a limited number of points have been sampled. One can say that by using this method sensitivity is converted into resolution. Basically, it is an efficient way to fit the time domain data to a sum of multidimensional exponentially damped sinusoids. The fitting is done locally over small overlapping regions in frequency. In addition, the different spectral dimensions are not independent of each other in this method. This means that information in one dimension provides improvement in the quality of the overall fit and therefore, for example, the resolution in another dimension can be increased.

Data filling. Data filling is a method for increasing the resolution in the indirect dimensions of symmetric 2D spectra [166]. In comparison to linear prediction and MEMs, it is computationally less demanding. The method

utilizes redundant information in the direct dimension to predict time domain data points in the indirect dimension to increase the overall resolution.

Three way decomposition. The program MUNIN [167] for the automated analysis of three-dimensional NMR spectra is based on the concept of three-way decomposition. In the MUNIN approach, a spectrum is decomposed into a sum of components, where each component is represented as the product of three (one for each dimension) one-dimensional shapes. Each component then generally represents a single peak or a group of peaks. Exemptions are, for example, E. COSY data where several components are required for a single cross peak. An important feature of MUNIN is that the method can be applied to frequency-domain or time-domain data or to a mixture of both. Therefore, uniform sampling of the time domain data as required for FFT is not necessary in the MUNIN approach. In the given example of a 3D $^{1}H-^{15}N$ NOESY–HSQC spectrum peak picking and peak integration of the resulting one-dimensional ^{1}H shapes should be much easier than in the corresponding conventional processed spectrum. In addition, the method can be used for a substantial data compression.

2.2.4.2. Peak and multiplet recognition. Since in general a set of spectra is used in any automated structure determination process, it is important that all spectra have been referenced properly, for heteronuclei it is advisable to use an indirect referencing scheme [168–170]. Although it is theoretically possible to simultaneously recognize all multiplet and spin patterns in a set of multidimensional data by fitting the data with a general model function characterized by a suitable number of parameters, in practice this approach is only possible for very simple systems involving only a few variables. The more economic way is to try to quickly reach a level of abstraction that can reduce the size of the data set to be handled by the computer. The simplest objects in NMR spectra for such an abstraction are the resonance peaks which must be separated from the background. If the spectral resolution is sufficient to resolve single multiplet components, those components can then be combined to form multiplets. Finally, the multiplets have to be assigned to complete spin systems.

Since the first program was published for performing pattern recognition in two-dimensional NMR spectra of polypeptides [171], this general strategy has been used almost exclusively. In practical applications, the signal-to-noise and the signal-to-artifact ratio is usually not sufficient for unambiguously observing all theoretically expected cross peaks. Therefore, cross peaks may be missed and noise or artifact signals may be recognized as true cross peaks. With this incomplete and erroneous information, only partial solutions are possible. In our experience, it is extremely important to be able to control any step in the assignment procedure removing incorrect hypotheses and including correct hypotheses whenever possible. Specifically, after peak picking, one should remove the peaks which are obviously artifacts (i.e. 'clean the peak list') and, after multiplet recognition, one should control the identified multiplets, remove incorrect hypotheses or correct these hypotheses, if possible. It is clear that in routine work all of these tasks should be performed in a fully automated fashion. However for more complicated problems (e.g. involving very large proteins), interactive routines are required which allow the expert to introduce general knowledge or to create new, problem-adapted strategies.

Peak picking. Peak picking in multidimensional spectra is a straightforward procedure since a maximum is defined by the property that all adjacent data points have a lower intensity, and conversely, a minimum is defined by the property that the adjacent points have a greater intensity. However, since resonance peaks must be distinguished from the large number of noise peaks, additional criteria must be defined which allow this classification. Approaches to automated peak picking can usually be divided into three types: (1) threshold-based methods, (2) peak-shape-based methods, and (3) Bayesian approaches.

(1) The simplest and most widely used criterion is the intensity threshold criterion, that is, only peaks with absolute intensities above a specific threshold are recognized as resonance peaks [171–176]. Since the reliability of automatic assignment procedures improves when a minimum number of 'false' peaks must be considered, optimal reduction of the number of noise and artifact peaks has proved to be beneficial. A simple method for significantly reducing the number of noise and artifact peaks is the exclusion of areas from the peak search where no meaningful resonances can be expected. Such spectral areas include regions outside the spectral range of the molecule under investigation and spectral regions where resonance peaks cannot be separated from artifact peaks (e.g. near the water t_1-ridge). In programs such as AURELIA and AUREMOL, these spectral regions can be defined interactively by the user. Improved results for the automated NOE signal identification in 2D and 3D NOESY spectra can be obtained with ATNOS [177] by including local baseline corrections, evaluation of local noise amplitudes, spectrum specific threshold values, symmetry criteria and incorporation of chemical shift and preliminary structural information.

(2) Additional information can be derived from the line shape itself. With a segmentation procedure, the n-dimensional line widths can be determined and peaks with very small line widths (i.e. noise spikes) or very large line widths (ridges and baseline rolls) can be automatically removed [178]. A more involved method of eliminating noise and artifacts used in STELLA [179] 'learns' from user-defined real and artifact peaks the typical line shapes of both of them and stores them in an internal database. STELLA compares the automatically picked signals with the database shapes by calculating the cosine between the vectors representing the line shapes of the picked

and database shapes. The position of the peak maximum is in this approach refined by a polynomial interpolation of its surrounding points.

CAPP [180] uses peak shapes to discriminate between noise and artifacts as the STELLA algorithm. However, CAPP does not require the user to select a set of real and artifact peaks interactively, but is based on the calculation of ellipses which best fit the contour lines. To discriminate between signals and artifacts, the calculated ellipses are evaluated by several criteria: the peak contours must be approximated sufficiently well at different levels by the ellipses, the radii of the ellipses (line widths) and the ratio of the radii (line shape) must be within user-defined limits. The position of the peak maximum is defined by the average of the centers of the ellipses used for peak recognition. The CAPP approach is applicable to up to four-dimensional data.

Also GIFA [46] uses a peak-shape based signal filter to discriminate between real signals and noise and artifacts.

Important features of AUTOPSY [181] are a routine for local noise level calculation, symmetry considerations of peak shapes and the use of peak shapes obtained from well-resolved cross peaks to resolve spectral overlap. To save computer memory, the spectrum in use is segmented into connected regions of data points above the noise level before the actual analysis.

Stoven et al. [175] define an additional criterion that the slope of a putative peak should exceed a given threshold (which also helps to separate resonance peaks from ridges). The line shape data can be used to even greater advantage by fitting the data to theoretical line shapes. However, in the frequency domain this is complicated because after filtering of the data in the time domain, in general, there is no simple analytical expression available to describe the resulting line shapes. In addition, frequency-domain fitting of data is extremely time-consuming, therefore, linear prediction methods, which also give peak intensities and coordinates, are probably superior, although also very time-consuming and in most applications unnecessary. However, the determination of the exact peak position is mainly hampered by the low digital resolution of multidimensional NMR spectra and is only improved insignificantly by the methods just described.

(3) A Bayesian approach coupled to a multivariate linear discriminant analysis of the data [182] can be used as a generally applicable method for the automated classification of multidimensional NMR peaks. The analysis relies on the assumption that different signal classes have different distributions of specific properties such as line shapes, line widths, and intensities (Fig. 3). In addition, a non-local feature is included that takes the similarities of peak shapes in symmetry related positions into account. The calculated probabilities for the different signal class memberships are realistic and reliable with a high efficiency of discriminating between peaks that are true NOE signals and those that are not [183]. A Bayesian method was also reported for the recognition of baseline artifacts [184].

Cluster analysis and multiplet recognition. In crowded spectra which are typical for homonuclear 2D spectra of proteins it is extremely difficult to analyze the many overlapping cross peaks and cross peak multiplets all at once. Here, clustering of the cross peaks which may be part of the same class and are located in close neighborhood can reduce the complexity of the problem. Typical examples are multiplet structures arising from *J*-coupling or residual dipolar coupling. Most programs for multiplet analysis were originally developed for the analysis of *J*-coupling patterns. However, residual dipolar couplings are nowadays routinely used for the structure determination of biological macromolecules. As in standard *J*-coupling experiments such as COSY that are used to obtain dihedral angle information, signals are split by the residual dipolar coupling. In typical heteronuclear applications, dipolar coupling leads only to a change of the magnitude

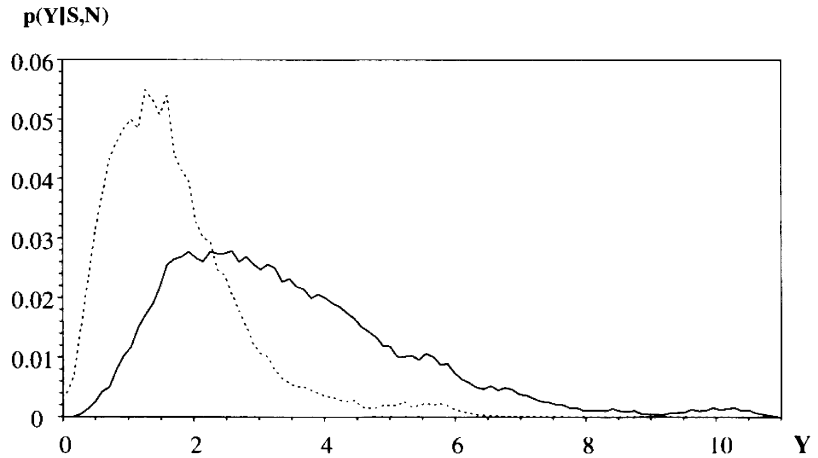

Fig. 3. Separation of signals and artifacts by Bayesian analysis. Probability distributions of the reduced variable *Y* for peaks belonging of the class of true signals (—) and for peaks belonging to the class of noise and artifacts (– –). The reduced variable *Y* is a linear combination of statistically independent peak properties. Figure adapted from Ref. [182].

of the already observable *J*-coupling induced splitting. Here, a typical example is the splitting of the amide nitrogen resonances by the *J*-coupling which is modified by the small additional residual dipolar coupling. However, additional multiplet structures can also be induced by residual couplings which is especially evident in homonuclear COSY or TOCSY-type spectra [185,186]. As a consequence methods previously proposed for the automated analysis of clusters and multiplets from COSY type spectra can be used with minor modifications for the automated analysis of residual dipolar couplings.

The digital resolution in 2D and sometimes in 3D NMR spectra recorded with modern instruments is usually sufficient to resolve the cross-peak multiplets. These multiplet structures can be accentuated by resolution enhancement methods or greatly suppressed by the application of filter functions that expand the line widths. Especially for COSY-type spectra containing anti-phase peaks, extensive broadening of the multiplet components is not advisable since the components with different phases will partially cancel each other.

Numerous different procedures for multiplet recognition have been proposed. The simplest approach assumes that all cross peaks which are separated by less than $2J_{max}$, where J_{max} is the largest expected coupling constant, are part of one cross-peak multiplet in a given spectral region [171,187]. This algorithm represents a simplified version of a cluster analysis which is the first step in most of the multiplet recognition procedures. Peaks which could possibly be part of a multiplet are combined in a cluster which then can be analyzed separately by a multiplet recognition algorithm. Clusters are usually defined using the assumption that peaks which are separated by a distance less than a given threshold value are members of a cluster. The maximum separation of sub-peaks in a cluster is limited by the magnitude of the coupling constants and the kind of multiplet structures detected by the *n*-dimensional experiment; only peaks of the same cluster can be part of one multiplet [188,189]. In very crowded regions of the spectrum these clusters tend to become very large. In such cases, it is more economical to split these large clusters into smaller sub clusters by automatic iteration of cluster parameters such as intensity differences of the peaks [189]. Multiplet recognition starts first with the smaller sub clusters. Peaks in these clusters which are not part of recognized multiplets will form new reorganized clusters which are then analyzed in a consecutive manner. Cluster reorganization is continued until all peaks are analyzed, that is, are either assigned to a multiplet or recognized as noise or artifacts [189].

Multiplet analysis programs operate based on predefined models of multiplet structure. Information on multiplet structure can be introduced into the program in several possible ways [190].

(1) The program can learn about multiplet structure variations by analyzing a set of typical multiplets defined by the user [74,189,191]. Introduction of these user-defined multiplets into the program can be performed by marking representative multiplets in an experimental spectrum or by independently entering multiplet data in a more or less abstract form.

(2) The user can define multiplet templates with fixed geometry. The program then searches the spectrum for a match to these templates.

(3) The spectrum can be analyzed for the occurrence of some general features of specific multiplets such as distinctive local symmetry [172,174,176,192–199].

(4) Multiplets can be completely modeled by fitting their patterns to model equations that describe the complete spin system and the evolution of the magnetization under the experimental conditions [188].

The first method is probably the most flexible since multiplets of any form can be defined. Continued learning of multiplet features in a training set and subsequent recognition of those features in experimental spectra can be performed elegantly using neural networks [191]. A computationally more efficient procedure extracts representative features from multiplet patterns defined by the user in an interactive fashion using a graphics display terminal. The experimental spectra are then analyzed using these representative multiplet features [74,189]. The method performs well even where multiplet overlap and artifacts are present in spectra.

Method 2 is computationally very efficient but is only useful for small, rigid molecules. In protein spectra the number of possible multiplets is enormous, since the *J*-coupling constants often are not fixed and, therefore, a continuous set of multiplets is expected. However, the procedure can be extended to a special case of method (1) where the computer creates a multitude of templates of a basic pattern by varying the coupling constants [173, 200–202]. The local symmetry of cross-peak multiplets is the fundamental feature most frequently used for multiplet recognition [172,174,176,188,192–198,203,204]. The local symmetry of a cross-peak multiplet depends on the type of multidimensional NMR experiment used to generate a spectrum. Although this cross-peak symmetry is ideally conserved only in the weak-coupling limit, in practical cases, the effect of intermediate couplings can usually be neglected. The computationally most straightforward application of local symmetries involves the computerized checking of a group of peaks in a given cluster to determine whether the required symmetry relationships are fulfilled. There are two main criteria used: (i) the geometrical factor, that is, are cross peaks present at the correct resonance positions, and (ii) do those cross peaks have the correct intensities and signs [172,203]. Some variance in peak positions must be allowed because of shifts due to digitalization (at least $\pm$ 1 data point) and overlap with other multiplets. The same is true of cross-peak intensities; in the worst case cross peaks of opposite sign may cancel completely. Therefore, during multiplet recognition it must

be remembered that a few cross peaks in any given multiplet are likely to be missing. The symmetry of candidate multiplets must be measured in some fashion. Cross-peak symmetry can be calculated from the peak intensities [196] or an area surrounding the peaks [172,193] (e.g. a rectangular box surrounding the peaks at half height) [172].

It is not absolutely necessary that peak picking be performed before multiplet recognition. Alternatively, *symmetry filters* can be applied to the complete two-dimensional spectrum or to spectral areas identified by cluster analysis [174,188,192–197]. Ideally, these filters generate a peak in the center of a multiplet when the form and couplings of that multiplet agree exactly with a multiplet in the test set. The program then suppresses all other combinations. This peak can be recognized later by a peak-picking algorithm which also simultaneously provides information on the multiplet. However, the performance of these filters is not very satisfactory for those cases where line widths are of the order of the peak separations and where multiplets overlap with other signals or artifacts. The theory of symmetry recognition by application of symmetry operations to higher dimensions has been described by Shen and Poulsen [198].

Peak and multiplet integration. The basis for macromolecular structure determination in solution is still given by distance information from NOE data. As a consequence, automated routines for automated NOE integration are required. Accurate integration of spectral cross peaks demands a reliable definition of the cross-peak area. However, such a definition is always a compromise between requirements that the integration area be as large as possible so that a complete integration is obtained, but also, as small as possible to reduce the inclusion of area arising from artifacts associated with baseline rolls and tails of other peaks. The simplest method of defining peak integration areas is interactively using a graphics display terminal, e.g. in the Bruker program, XWINNMR, where the user defines a box around the peak in which all pixels are summed. However, this manual integration procedure is not useful for a protein spectrum where several thousand integrals must be determined. One way of defining the integration areas automatically makes use of the observation that the slope of a peak decreases monotonically with the distance to the peak center, at which point it approximates zero. This feature can be used to define a rectangular integration area by determining the points where the slope is smaller than a predetermined value in a row and a column through the peak maximum [205]. A more elaborate procedure determines the peak contour at which this condition is met [175].

A similar approach defines the peak integration area using an iterative 'region-growing' algorithm [172,178,206], which recognizes all data points that are part of a given cross-peak, the integration can be performed based on a user-defined threshold level (Fig. 4). If not zero this threshold should be defined relative to the maximum value of the peak since otherwise the relative volumes are not directly proportional to the strength of interaction. This automatic integration procedure works surprisingly well even for overlapping peaks as long as the peak maxima are separately visible and therefore recognizable by the peak picking procedure.

In a different approach the peaks are fitted by a set of reference peaks defined by the user [176,207]. This approach is probably best suited in cases where peaks strongly overlap; however, it demands a careful selection of the reference peaks by the user.

A completely different solution of the integration problem is provided by the LP-related methods when those methods are used to completely predict the two-dimensional data set. From the peak intensities and the decay constants, the integrals of the resonance peaks can be calculated directly. When comparing various integration procedures, one must always keep in mind that many problems other than imperfections in the integration procedure can easily lead to very large errors.

A common source of such errors is baseline variations and an insufficient digital resolution [208]. However, even an error in volumes of a factor of 2 (+100 or −50%), which is usually much larger than the error produced by integration routine per se, only leads to a distance error of −11 and +12%, respectively. Therefore, convenience and reliability of the integration procedure is of prime importance, especially in three- or four-dimensional experiments, where manual integration is not practical.

2.2.4.3. Assignment of resonance lines. Very different approaches have been published in the literature for this stage of the automated structure determination process. In this section we will summarize the methods in use for spin system recognition and sequential resonance assignments that are necessary steps in most schemes proposed for automated structure determination in solution. Although, we will see later that it is also possible to obtain a structure without these steps.

For the methods described in this section usually four separate steps are necessary, that can vary in the order they are applied and sometimes several steps are performed simultaneously [190]. These steps are (a) grouping of resonances from one or more spectra to spin systems, (b) association of spin systems with amino acid types, (c) linking of spin systems to smaller or longer fragments, and (d) mapping of fragments obtained from step (c) to the primary sequence.

The initially proposed procedures for the assignment of resonance lines were developed for homonuclear NMR data, but could be used analogously to the heteronuclear data predominantly used nowadays. The classical method for sequential assignment consists of the following steps. First, identification of all resonances which are part of the elementary building blocks of the molecule, e.g. amino acids in the case of proteins, and then the assignment of

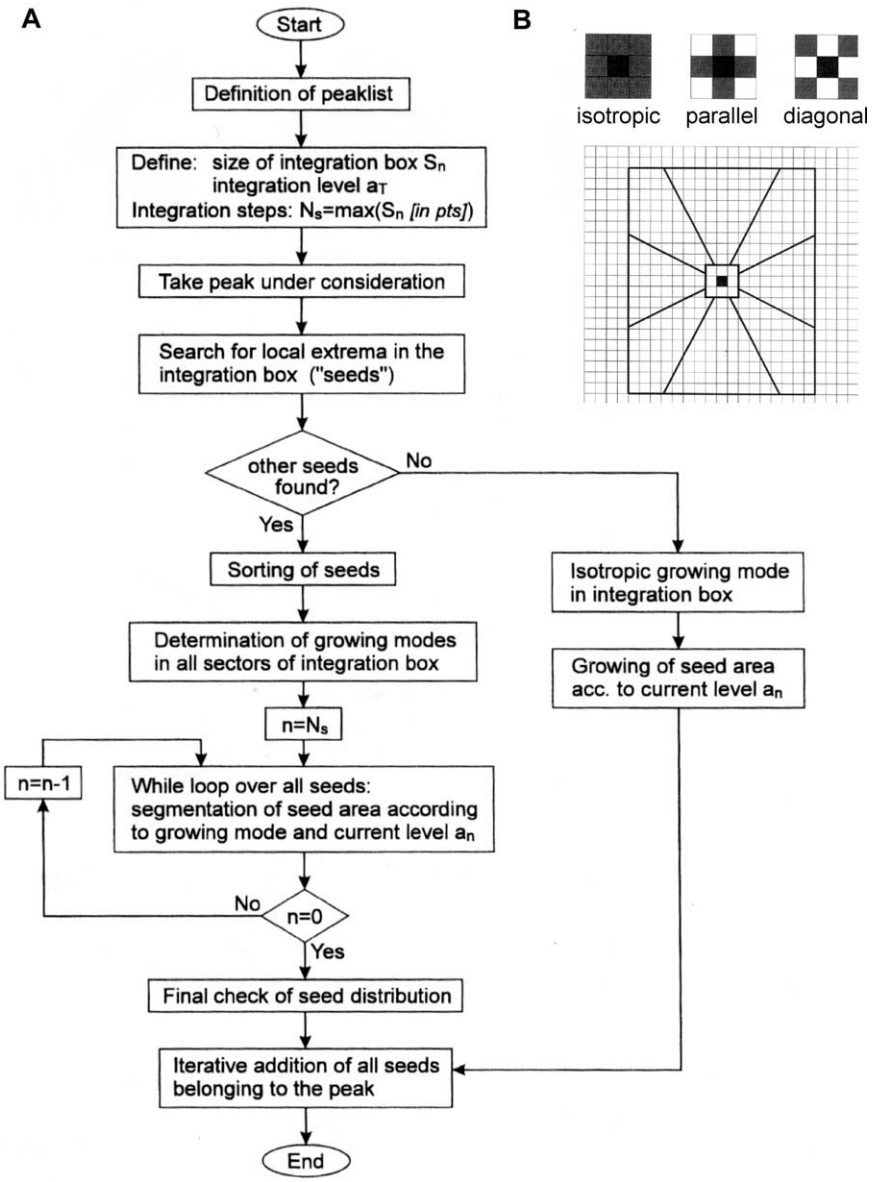

Fig. 4. Iterative cross-peak integration by a region-growing algorithm. (A) Schematic representation of the integration procedure in multidimensional NMR spectra. (B) Growing conditions and sector arrangement in two-dimensional NMR spectra. (A) Three different conditions for the segmentation of neighboring data points are defined: 'isotropic growth', 'parallel growth', and 'diagonal growth'. (B) Definition of the sectors around a central peak in a rectangular integration box. Each sector will be combined with a separate growing mode according to the actual arrangement of the seeds. The inner sector around the peak is set to isotropic growth. Figure adapted from Ref. [206].

these spin systems to specific positions in the sequence via the use of sequential NOEs [209] between, for example, directly neighboring amino acids. This approach was realized in the CLAIRE [210] program that is designed for the automated assignment of smaller proteins from two homonuclear 2D TOCSY and NOESY spectra. It involves tools for pattern (spin system) generation, for mapping of these patterns to residue types using standard chemical shift information and tools for linking individual patterns using NOE connectivities to obtain the sequential assignment. This approach basically tries to automate the manual assignment approach originally suggested by Wüthrich [211].

The other extreme is the main chain directed (MCD) approach to the spectral assignments, a method for the automated backbone and partial side-chain assignment of peptides and small proteins [212–214]. It is based on homonuclear 2D *J*-correlated and 2D NOESY spectra, but has similarities with the now most used method for sequential assignment via heteronuclear 3D-NMR spectroscopy. The MCD method concentrates mainly on protons from the protein backbone. Residue fragments containing the NH, Hα and Hβ protons are generated from COSY and TOCSY spectra while the fragments are linked by NOE connectivities. In this approach typical NOE patterns for the various secondary structure elements are

employed. Complete spin systems of the side-chains are identified later, a task that is usually easily performed for smaller proteins since the type of spin system (amino acid residue) is known from the primary structure.

In practice, most manual and automated assignment procedures use a mixture of these two methods, since complete assignments are usually obtained by an iterative process where main chain and side-chain information is used in repetitive cycles. This is even more evident in multidimensional heteronuclear NMR spectroscopy where sequence and spin system information is present in the same spectrum.

A problem common to all automated assignment procedures is missing peaks (multiplets), which must be expected even for spectra with excellent signal-to-noise ratios, and non-perfect peak and multiplet recognition when peaks are near to the diagonal of the spectrum (chemical shift degeneration) or are in very crowded regions of the spectrum. Great difficulties and erroneous assignments also result from artifact peaks which should be removed from the spectrum as far as possible in the preceding steps.

Chemical shift prediction. It is obvious that it would be of considerable advantage for automated assignment processes to have a precise prediction of the expected ^{1}H, ^{13}C, and ^{15}N chemical shifts. This is especially true for the association of spin systems with amino acid types and for mapping fragments to the primary sequence. Another application of predicted chemical shifts is the completion of side-chain assignments based on NOESY spectra (see in the structure calculation paragraph the section concerning ambiguous NOE restraints). Chemical shifts can either be predicted from a statistical database (knowledge based chemical shift prediction) or from a physical model for chemical shifts including quantum chemical computations. Of course, even in the latter case empirical data such as random-coil shifts usually enter the calculation and present, principally, statistical data.

Systematic, empirical relationships between chemical shift homology and sequence (structural) homology have been established in the past few years [215]. Often the similarity of the ^{1}H chemical shifts of a protein to those of a homologous one are used as an indicator of similar global folds. With complete or nearly complete chemical shift assignments for a large number of proteins now deposited in the BioMagResBank [216] it seems logical to use all this prior knowledge for predicting chemical shifts for a new protein. The program ORB [217] predicts ^{1}H, ^{13}C, and ^{15}N chemical shifts of previously unassigned proteins. The program makes use of the information contained in a chemical shift database of previously assigned proteins supplemented by a statistically derived averaged chemical shift database in which the shifts are categorized according to their residue, atom, and secondary structure type [218]. The prediction process starts with a multiple sequence alignment of all previously assigned proteins with the unassigned query protein. ORB uses the sequence and secondary structure alignment program XALIGN [219] for this task. The prediction algorithm in ORB is based on a scoring of the known shifts for each sequence. The scores determine how much weight one particular shift is given in the prediction process.

In some regards similar to ORB is the program SHIFTY [220] that make use of the information content present in the BioMagResBank, but uses only the most homologous protein for which shifts are available for its predictions. Standard sequence alignment techniques are used to compare the sequence of the trial protein with the sequences of previously assigned proteins contained in the BioMagResBank. The chemical shifts of the most homologous previously assigned protein are then used for chemical shift prediction of the trial protein. In addition the algorithm adjusts for differences in the primary sequences of the two proteins.

A large empirical database of $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts together with the corresponding three-dimensional X-ray structures was constructed to investigate the effects of backbone geometry, side-chain geometry, hydrogen bonding, ring currents, and sequence on chemical shifts [221]. It was found that contributions from backbone and side-chain geometry, as well as hydrogen bonds have significant effects on $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts, while the other factors can be neglected in most cases. The results from this study should be a useful tool for the refinement of protein structures employing $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts. Predictions were made using the program TANSO.

When chemical shifts from homologous proteins are not available, but three-dimensional structural information is, this information can be used for chemical shift predictions. Semi-empirical methods have been developed for the prediction of proton chemical shifts [222–225] as well as ab initio methods for the prediction of ^{1}H, ^{13}C, ^{15}N, and ^{19}F chemical shifts [226–230]. Note that the paper by Ando et al. [230] concerns peptides and polypeptides in the solid state. Another approach was developed by Xu et al. Their program SHIFTS [231] predicts ^{15}N and ^{13}C shifts from a known structure using previously calculated shifts from a database. This database of chemical shift patterns for 1335 tripeptides is calculated using density functional calculations. The backbone angles of the tripeptides are limited to the regions of regular secondary structure. Therefore, predictions made by SHIFTS are limited to these regions. In addition to backbone dihedral angles, the program takes side-chain dihedral angles of the trial residue and the preceding residue and estimated hydrogen bond effects into account. The program PROSHIFT uses neural networks trained on solved structures together with the corresponding experimentally determined chemical shifts to predict chemical shifts for a new protein when three-dimensional structural information is available [232].

2.2.4.4. Extraction of structurally relevant information. The next step after the resonance line assignment has been finished is usually to apply this knowledge to extract

Table 2
General ambiguity of NOE assignments in Csp from *T. maritima*

N_{ab}	D/ppm	Number of NOE cross peaks with ambiguous assignments								
		correctly folded[b]			extended[c]			without structure[d]		
		A	B	C	A	B	C	A	B	C
1	0.01	548	462	614	268	242	364	8556	7500	11830
	0.02	170	136	176	80	70	100	2256	1836	2850
	0.03	54	40	50	18	16	22	650	518	720
2	0.01	790	654	1100	420	368	560	16926	13758	22670
	0.02	390	302	456	176	152	244	7776	6348	9864
	0.03	196	148	186	86	76	102	3224	2528	3542
3	0.01	282	202	424	148	108	280	6138	4620	10114
	0.02	190	150	326	122	96	200	3360	2814	6624
	0.03	62	50	152	38	30	94	1664	1496	3584
4	0.01	876	710	682	478	398	388	19908	14484	15100
	0.02	564	484	596	302	266	308	11472	9216	12102
	0.03	336	256	290	180	138	150	6642	5206	6060
5	0.01	176	122	158	94	76	74	4278	3156	3202
	0.02	118	68	84	48	32	56	2784	1650	2106
	0.03	82	64	94	34	28	50	1352	1040	1826
6	0.01	506	382	392	280	222	250	12516	8196	9694
	0.02	418	352	538	220	192	320	8454	6646	12648
	0.03	260	246	456	138	130	250	5320	4458	9558
7	0.01	66	38	22	32	24	12	1302	840	390
	0.02	74	58	78	32	30	30	1536	1116	1380
	0.03	28	26	54	20	20	30	832	568	1078
8	0.01	534	360	196	286	200	136	14628	8898	4120
	0.02	378	294	306	222	172	178	9768	7386	6996
	0.03	406	292	264	202	152	140	7652	5928	5108
>8	0.01	2472	1046	388	1454	674	248	62820	24858	9190
	0.02	3948	2132	1416	2258	1302	876	99666	49298	31740
	0.03	4826	2854	2430	2744	1722	1474	119736	64568	54834
–[e]	–[e]	6250	3976	3976	3460	2312	2312	147072	86310	86310

The numbers of signals for which at least one possible assignment was found are shown. Signals were generated using published data. The resonance assignments of *Tm*Csp were taken from Ref. [395]. The number of ambiguous NOEs was calculated for three different sets A, B, C. N_{ab} specifies the number of assignment possibilities for a given cross peak. Set A considers all possible NOE cross peaks and for the assignment process it is assumed that the correct stereospecific assignments of the resonances are not known. In comparison to set A all NOEs corresponding to side-chain to side-chain contacts are excluded in set B. In set C the same NOE cross peaks as in set B are considered, however in this case it is assumed that the stereospecific assignment is known (as given in the assignment table). In the application of the automated assignment described, D is the maximum separation of the chemical shift δ_{ij} of a candidate assignment from the experimental cross peak, that is two resonances can be separated by 2D and correspond to one cross peaks. For the calculations two resonances δ_{ij} and δ_{ml} are assumed as not distinguishable if $|\delta_{ij} - \delta_{lm}| \leq 2D$.

[a] Distance cutoff of 0.5 nm for possible assignments, the final structure of *Tm*Csp was used as test structure.

[b] Distance cutoff of 0.5 nm for possible assignments, an extended strand of *Tm*Csp was used as test structure.

[c] No distance cutoff was assumed.

[d] Total number of cross peaks considered.

structure relevant information. This is still predominantly distance information obtained from NOESY spectra, supplemented with additional information obtained from *J*-couplings, residual dipolar couplings, chemical shifts, and hydrogen and/or disulfide bonds.

Computer aided assignment of NOESY spectra. The assignment of the NOE-cross-peaks in two-, three-, and four-dimensional spectra is a tedious and error-prone process simply because the number of cross peaks is very large. Consequently, it appears reasonable to automate this part of the spectra evaluation. One major problem in manual and automated evaluation of NOE data of proteins is the chemical shift degeneracy in the NMR spectra. Therefore, for a high percentage of the NOESY signals no unambiguous assignments can be obtained based on chemical shifts alone. An example is shown in Table 2 for the small coldshock protein Csp from *Thermotoga maritima* [233]. It shows the very high degree of ambiguity in a 2D-NOESY spectrum since in principle any combination of proton chemical shifts of the protein represents a NOESY cross peak (in most cases a very weak signal). One important aspect in this regard is the set of restrictions applied to the assignment of the individual cross peaks. These can include tolerance values describing how close the shifts of the sequential assignment table match a particular NOESY spectrum in the various dimensions and how well a tentative assignment agrees with the current structure model. Too loose

restrictions lead to an unmanageable number of assignment possibilities, while too tight restrictions may prevent the correct assignment from being considered.

Use of ambiguous NOESY restraints. The problem of calculating three-dimensional structures with ambiguous restraints was approached by a number of groups. ARIA [234–236] is using an iterative combination of resonance assignments and structure calculations. All possible assignments are listed for each peak that are compatible with the resonance assignment using a fixed chemical shift tolerance value. In ARIA the restraints from ambiguous assignments are included as an r^{-6} weighted sum D_s in the NOE target function. A typical NOE-potential V_{NOE} is then given by

$$V_{NOE} = k_{NOE}|(D_s - D_{NOE})|^{\alpha} \tag{2.4}$$

with

$$D_s = \left(\sum_{i=1}^{M} D_i^{-6}\right)^{-1/6}$$

for $(D_s - D_{NOE}) > U$; for $(D_{NOE} - D_s) > L$ and

$V_{NOE} = 0$ otherwise

with the NOE force constant k_{NOE}, the number M of all assignment possibilities contributing to a specific cross peak, the corresponding distances D_i in the trial structure, the distance D_{NOE} determined from the experimental cross-peak volume, and an exponent α (typically 2). The upper and lower errors U and L, respectively, define a range outside which V_{NOE} becomes effective. Inside this range V_{NOE} adopts a value of 0. The equation shown above is often modified so that when the discrepancy between the summed distances and D_{NOE} exceeds the upper limit U plus a certain cutoff value, then V_{NOE} increases only linearly until it approaches a maximum value for large violations. This is done to prevent single cross peaks from dominating the whole structure determination process and to make optimization numerically more stable.

After the following structure calculation the ambiguous assignment possibilities will be judged, based on the fit of the corresponding restraints to the obtained structures. For each assignment possibility k to the ambiguous NOE, the minimum distance D^k_{min} in the ensemble of converged structures is determined and from this minimal distance the contribution C^k to a cross peak is calculated by

$$C^k = \frac{(D^k_{min})^{-6}}{\sum_{i=1}^{M} (D^i_{min})^{-6}} \tag{2.5}$$

In the next iteration only contributions that exceed a certain threshold are further considered. When all but one contribution are excluded the NOE is unambiguously assigned.

ARIA [234,237] can also be applied to solve ambiguous disulfide connectivities provided that a sufficiently high density of NOE or other restraints is present. In this approach the ambiguous disulphide bonds are treated as ambiguous NOE contacts.

When a preliminary structure but only partial side-chain assignments are available the method can be extended to make sequence-specific assignments of the side-chain protons [238]. For this approach it is necessary to have approximate chemical shift estimates for the protons in question. Several methods are available for obtaining these chemical shift predictions, e.g. use of random coil shifts or shifts from a previously assigned close homolog. Since in the latter case the approximate three-dimensional structure is known in addition semi-empirical or ab initio methods can be used. Ambiguous assignments are obtained for the set of unassigned signals using the predicted chemical shifts together with error ranges reflecting the expected accuracy of the chemical shift predictions. In contrast to the original ARIA protocol, ARIA is here used to refine existing structures. After refinement, the assignment possibilities that correspond to large distances in the resulting structures are discarded. In the case where additional cross peaks can be unambiguously assigned by this procedure these cross peaks are used to obtain new sequential assignments.

Recently, ambiguous restraints have also been successfully used in the HADDOCK approach for the docking of protein–protein complexes. Here for both proteins, the residues involved in complex formation must be identified using, e.g. chemical shift perturbation data originating from NMR titration experiments. However, knowledge about the exact pairwise interactions is not required, since the experimental information is coded as ambiguous restraints between the interacting residues [239].

The SANE method for automated NOE assignment is similar to the ARIA program. However, while ARIA is interfaced with X-PLOR/CNS, the program SANE is compatible with the MD programs DYANA and AMBER [240].

The automated NOE assignment approach from Savarin et al. [241] also uses ambiguous distances constraints in an iterative NOE assignment/structure determination process that is similar to ARIA. Assignments are considered based on chemical shift tolerance values and a distance cutoff in the trial structures. In the case of ambiguity a maximum number of assignment possibilities contributing to one cross peak and their relative volume contributions to this signal are considered. Calculations are usually started employing small chemical shift tolerance values and a small distance cutoff. In later iterations, these values are increased to allow the assignment of all signals. The aim of this procedure is to obtain intermediate structures presenting only a few violations.

Filtering of NOESY restraints by violation analysis. The self-correcting distance geometry based NOAH/DIAMOD approach [242–245] is an iterative method using a combination of automated NOE assignments, structure calculations, and violation analysis to obtain the final solution structures. As in ARIA for each peak all possible

assignments are listed that are compatible with the resonance assignment using a fixed chemical shift tolerance value. For each assignment possibility, a restraint is created if the number of assignment possibilities does not exceed a user specified value (typically two to four) and put into the list of test assignments. After structure calculations candidate assignments are judged by the violations in the set of obtained structures. If a peak has only one possible assignment and is not heavily violated it is transferred to the list of unambiguous assignments. If a peak is ambiguous three different cases have to be distinguished: (1) If more than one assignment are equally compatible with the structures, the peak with all its compatible assignments will be transferred to the list of ambiguous peaks. (2) If one assignment is substantially more compatible with the obtained structures than the second best assignment it will be transferred to the list of unambiguous assignments. (3) If none of the proposed assignments are compatible with the structure, the peak is put back to the pool of unassigned peaks. Peaks from the unambiguous list fall back to the unassigned peak list if their assignment has become incompatible with the rest of the assignments. For the next iteration of structure calculations new test assignments are created as described above and also by taking into account their compatibility with the structures of the previous round. In subsequent structure calculations signals from the unambiguous assignment list are weighted five times stronger than peaks from the ambiguous and test assignment lists. Similar to the approach described above the program NOAH has also been implemented in the distance geometry program DIANA [246]. The main difference between ARIA and NOAH is that ARIA tries to avoid assignment errors by using the sum of properly weighted ambiguous distance restraints for ambiguous assignments, while NOAH purposely uses incorrect restraints and hopes to identify them by violation analysis.

CANDID [247] is an iterative approach for automated NOE cross-peak assignment and automatic 3D protein structure generation. It combines features from NOAH and ARIA, such as the use of ambiguous constraints and the use of filters based on the three-dimensional trial structures. To deal with noise and artifacts, which is of special importance in the first iteration where no trial structure is available CANDID includes tools for network anchoring and constraint combination. It is interfaced with the molecular dynamics algorithm DYANA.

AutoStructure (Huang et al., in preparation) is an expert system that uses rules like the ones applied by a human expert for restraint generation from experimental spectra. Structures are then obtained in an iterative approach using the program DYANA. Here also violation analysis is used.

Calculation of assignment probabilities for ambiguous NOESY restraints. The program KNOWNOE [233] presents a novel, knowledge based approach to the problem of automated NOE assignment. KNOWNOE is devised to work directly with the experimental spectra without interference of an expert. Besides making use of routines already implemented in the new program AUREMOL, it contains as a central part a knowledge driven Bayesian algorithm for solving ambiguities in the NOE assignments. KNOWNOE will be explained in more detail in the second part of this review.

Use of back-calculated spectra to obtain NOESY restraints. NOESY spectra back-calculated from a single trial structure or a set of trial structures offer various possibilities for computer assisted assignment procedures by automatically comparing them with their corresponding experimental counter parts to automatically assign the experimental signals. However, this procedure has to be applied with care in cases where strong deviations between the trial structures and the real structure can be assumed. Another application is the calculation of NMR *R*-factors that is based on the comparison of experimental and simulated intensities to judge how well the trial structure fits the experimental data. The same method can also be used to distinguish between different structural models. An example is given in Fig. 5 where the back-calculation of a NOESY spectrum demonstrates that the structure of HPr from *E. faecalis* as obtained in single crystals using X-ray diffraction is not the dominant structure in solution [248].

Using back-calculated spectra it is also possible to obtain accurate distance information from the corresponding experimental spectra. This allows the replacement of distance constraints in three-dimensional structure calculations by including the difference between simulated and experimental intensities as a pseudo energy. The comparison of experimental and simulated spectra should also permit the extraction of motional parameters of the protein of interest. During the last few years several programs have been developed that allow the simulation of multidimensional NOESY spectra using the full relaxation matrix approach CORMA [249], BCKCALC [250], IRMA [251], MORASS [252], MARDIGRAS [253], DINOSAUR [254], MIDGE [255], NO2DI [256], a program by Kim and Reid [257], X-PLOR [258], BIRDER [259], RELAX [260,261] and, SPRIT [262].

The first of these calculations were mostly performed for theoretical reasons [249] and to test the influence of spin diffusion on NOESY signal intensity [263]. The main differences between the various approaches are the treatment of internal motions, if non-isotropic tumbling can be considered, and if the effects of finite relaxation delays can be simulated. As an example RELAX allows the simultaneous application of different motional models describing the internal and overall motion of the molecule under investigation for individual spin pairs or groups of spins. It will be described here in more detail.

RELAX [260,261], a program for the back-calculation of NOESY spectra based on complete relaxation matrix formalism, is part of AURELIA [264], an improved version has been implemented in AUREMOL. RELAX allows the simulation of ^{1}H 2D NOESY spectra

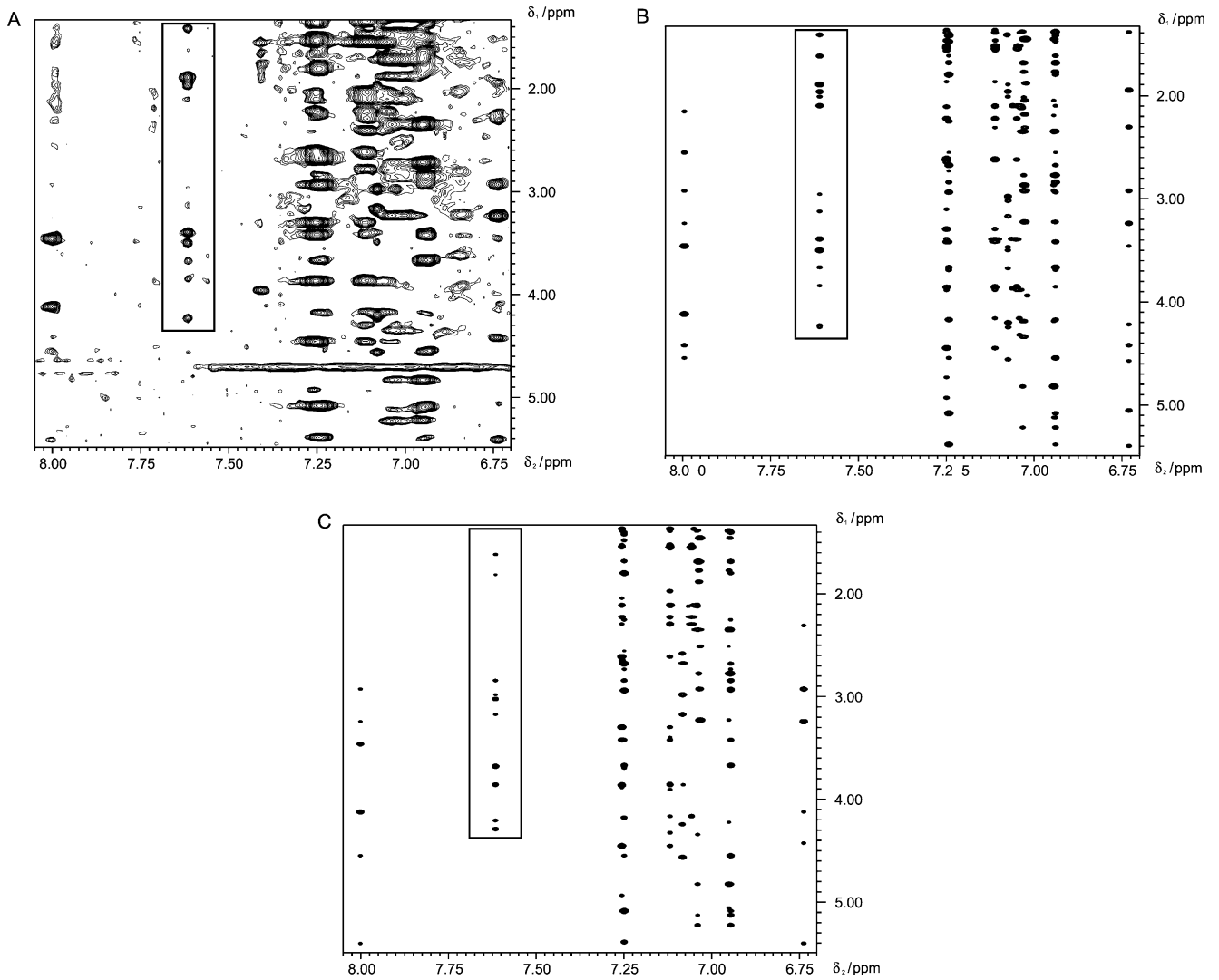

Fig. 5. Comparison of ^{1}H 800 MHz NOESY spectra of the histidine-containing phosphocarrier protein (HPr) from *E. faecalis* to verify differences between the X-ray and NMR structure. Structural differences were observed especially for the region of the active site (His 15). In the displayed spectra this region is surrounded by a rectangle. (A) Experimentally determined spectrum. (B) Spectrum back-calculated from one of the final set of NMR structures. (C) Spectrum back-calculated from the X-ray structure. Figure adapted from Ref. [248].

and *IS* ($I = {}^{1}$H, $S = {}^{13}$C or ^{15}N) NOESY–HSQC spectra. The *IS* NOESY–HSQC experiment is basically a concatenation of a homonuclear ^{1}H-NOESY and a heteronuclear *IS* HSQC-experiment.

For a NOESY experiment the evolution of the deviation of longitudinal magnetization from thermal equilibrium ΔM_z is described by the generalized Solomon equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta M_z(t) = -\mathbf{D}\Delta M_z(t) \tag{2.6}$$

The dynamics matrix **D** that governs the time evolution of the cross-peak intensities in a 2D-NOESY experiment is given by

$$\mathbf{D} = \mathbf{R} + \mathbf{K} \tag{2.7}$$

K is the kinetic matrix that describes chemical and/or conformational exchange [265], while **R** is the relaxation matrix [56,266,267]. If the effects of chemical exchange are neglected, as in the current version of RELAX, the solution of Eq. (2.6) simplifies to

$$\Delta M_z(t) = \Delta M_z(0)\exp(-t\cdot\mathbf{R}) \tag{2.8}$$

wherein $\Delta M_z(0)$ is the deviation of the longitudinal magnetization from thermal equilibrium at time zero. However, fully relaxed spectra are hardly ever recorded. An improved signal-to-noise ratio can be obtained if more experiments are accumulated with a shortened relaxation delay t_d. In this case the longitudinal magnetization M_{zj} of a nucleus j recovers only partly during the recovery time t_r. Sometimes purge pulses are used prior to t_d to enhance spectral quality. In this case it can be assumed that $M_z(0) = 0$ at the beginning of t_d and the magnetization

recovers only during t_d ($t_r = t_d$). In the absence of purge pulses under ideal conditions $M_z(0) = 0$ is fulfilled after the 90° detection pulse. Therefore, there is in this case an additional contribution from the acquisition time t_{ac} to the total recovery time t_r ($t_r = t_d + t_{ac}$). The z magnetization $M_z(t_d)$ at the beginning of the mixing time ($t_1, t_2 = 0$) can be written as

$$M_{z,j}(t_r) = \frac{1}{\alpha}\sum_k [1 - \exp(-t_r R)]_{jk} \tag{2.9}$$

where k runs over all nuclei in the relaxation matrix and α is an arbitrary scaling factor.

For dipolar homo- or heteronuclear relaxation and spin $I = 1/2$ the rates of auto-relaxation R_{ii} and the cross-relaxation R_{ij} between two spins i and j are given by

$$R_{ii} = \sum_{j\neq i} q_{ij}[J^0_{ij}(\omega_i - \omega_j) + 3J^1_{ij}(\omega_i) + 6J^2_{ij}(\omega_i + \omega_j)] \tag{2.10}$$

and

$$R_{ij} = q_{ij}[6J^2_{ij}(\omega_i + \omega_j) - J^0_{ij}(\omega_i - \omega_j)] \tag{2.11}$$

with J^n_{ij} ($n = 0, 1, 2$) being the spectral densities for n-quantum transitions characterizing the motion of a vector connecting spin i and j relative to the $\mathbf{B}_0$-field. The dipolar interaction constants q_{ij} are given by

$$q_{ij} = (1/10)\gamma_i^2\gamma_j^2 h^2(\mu_0/4\pi)^2 \tag{2.12}$$

where γ_i and γ_j are the gyromagnetic ratios of spin i and j, respectively.

Within RELAX, the following motional models can be used to describe internal and external motions of the molecule:

Rigid in cases where isotropic overall tumbling of a rigid molecule is assumed:

$$J^n_{ij}(\omega) = \frac{1}{r^6_{ij}}\left(\frac{\tau_c}{1 + \omega^2\tau_c^2}\right) \tag{2.13}$$

Slow Jump for the description of internal movements which are slow relatively to the time scale of the overall tumbling as it is the case for aromatic ring-flips. It is also often referred to as r^{-6} averaging:

$$J^n_{ij}(\omega) = \frac{1}{N_i N_j}\left(\frac{\tau_c}{1 + \omega^2\tau_c^2}\right)\sum_{\mu=1}^{N_i}\sum_{\nu=1}^{N_j}\frac{1}{r^6_{i_\mu j_\nu}} \tag{2.14}$$

Fast Jump where the correlation for the internal jumps between equilibrium positions is much smaller than that of the whole molecule as it is found for fast rotating methyl groups:

$$J^n_{ij}(\omega) = \frac{1}{2N_i^2N_j^2}\left(\frac{\tau_c}{1 + \omega^2\tau_c^2}\right)\sum_{\mu,\upsilon=1}^{N_i}\sum_{o,\pi=1}^{N_j}\frac{1}{(r^5_{i_\mu j_\nu} r^5_{i_o j_\pi})} \times (3(\vec{r_{i_\mu j_\nu}}\vec{r_{i_o j_\pi}}) - r^2_{i_\mu j_\nu} r^2_{i_o j_\pi}) \tag{2.15}$$

Average 3 represents the r^{-3} averaging for the description of fast motions:

$$J^n_{ij}(\omega) = \frac{1}{N_i^2N_j^2}\left(\frac{\tau_c}{1 + \omega^2\tau_c^2}\right)\left|\sum_{\mu=1}^{N_i}\sum_{\nu=1}^{N_j}\frac{1}{r^3_{i_\mu j_\nu}}\right|^2 \tag{2.16}$$

Lipari model free approach for internal motions not easily to describe by a simple motional model. In this approach the motions of the molecule are assumed to be a superposition of a slow overall motion with correlation time τ_c and rapid internal motions with correlation times $\tau_{1,ij}$ and generalized order parameter S_{ij} with $\tau^{-1}_{e,ij} = \tau^{-1}_c + \tau^{-1}_{1,ij}$:

$$J^n_{ij}(\omega) = \frac{1}{r^6_{ij}}\left(\frac{S^2_{ij}\tau_c}{1 + \omega^2\tau_c^2} + (1 - S^2_{ij})\left(\frac{\tau_{e,ij}}{1 + \omega^2\tau^2_{e,ij}}\right)\right) \tag{2.17}$$

Lipari 1 a simplified version of the above definition that is sufficient in most cases when $\tau_{1,ij} \ll \tau_c$:

$$J^n_{ij}(\omega) = \frac{1}{r^6_{ij}}\left(S^2_{ij}\frac{\tau_c}{1 + \omega^2\tau_c^2}\right) \tag{2.18}$$

By a suitable combination of the spectral densities presented above, it is possible to set up a detailed model for the internal and overall motions of the molecule. Such a model describes these motions as a superposition of a slow overall rotational diffusion, with a rotational correlation time τ_c and fast internal motions, that may vary from spin pair to spin pair in the molecule. Usually, an isotropic overall diffusion is assumed for the overall rotational reorientations.

However, that is only true for spherical molecules or only a good approximation if the anisotropy of rotational diffusion of the molecule is relatively small. For molecules that undergo anisotropic rotational diffusion described by the diffusion coefficients $D_\parallel$ and $D_\perp$ for rotation around the main the transverse axes, a practical limit where the anisotropy can be neglected is given by $D_\parallel/D_\perp \leq 1.3$ [268]. It turns out that for many proteins the isotropic approximation does not apply. As a consequence in RELAX an automatic optional anisotropy correction is included which models the molecule as an ellipsoid. The necessary parameters are automatically calculated from the given structure [260]. However, no internal motions are considered in the current implementation of RELAX when the anisotropy correction is turned on.

The theory to combine internal motion with anisotropic reorientation exists [268], and introducing this should improve the simulation of NOEs substantially if a large anisotropy exists.

In 3D NOESY–HSQC spectra, the NOESY-parts of the 3D pulse sequences differ slightly from the standard homonuclear NOESY pulse sequence, since it is advantageous to decouple the heteronuclei during the evolution period t_1. This is often done by an additional 180° pulse on the S-spin at the midpoint of the t_1-evolution.

Alternatively, the *S*-spin can be decoupled during the evolution time by broadband decoupling; this means that the initial value of the *S*-magnetization is then zero in the ideal case. Finally, the influence of the *S*-spin can be completely removed from the NOESY part by additional broadband decoupling during the NOE mixing time. In the following we consider the standard case, where a 180° pulse on *S* is used during t_1 and both the *I*- and *S*-magnetizations are inverted at the beginning of the mixing time.

In NOESY–HSQC spectra, ideally the HSQC-part of the 3D-pulse sequence reflects only the polarization transfer produced by the NOESY-part of the sequence. However, the signal amplitude is modified by several factors during the HSQC-sequence. For the quantitative analysis only factors are important which change the relative intensities of the NOE-signals. Two main sources are responsible for such changes, the differences in the indirect *IS*-spin coupling constants which are essential for the INEPT-transfer, and the differences in individual transversal relaxation times. In principle, the *IS* coupling constants could be determined independently and the transversal relaxation times could be calculated for every individual *IS* spin pair from the three-dimensional structure and the motional parameters.

The cross-peak volume V_{ijk} of two protons i and j obtained at the end of the pulse sequence is given by

$$V_{ijk} = cI_{ij}f_{jk} \tag{2.19}$$

with the NOESY intensity after the NOESY mixing time I_{ij} which is finally transferred from atom j to the directly bonded heteronucleus k by the HSQC sequence, a general scaling factor c and f_{jk} the transfer efficiency for the INEPT transfer from atom j to k or back to atom j. As a first approximation the transfer factor f_{jk} can be expressed by [262]

$$f_{jk} = (\sin(\pi J_{jk}\tau)\mathrm{e}^{-(R_2^{I_j}+\frac{1}{2}R_1^{S_k})\tau})^n \tag{2.20}$$

with J_{jk} the coupling constant between the proton j and the heteronucleus k, τ the INEPT mixing time (typically $1/2J$), n the number of INEPT periods, $R_2^{I_j}$ the relevant transversal relaxation rate of the directly bonded proton j, and $R_1^{S_k}$ the longitudinal relaxation rate of the corresponding heteronucleus.

The individual transfer factors f_{jk} can also be obtained experimentally by recording a two-dimensional data set under identical conditions but with a zero NOESY-mixing time and a constant value of $t_1 = 0$. This method works fine under almost all conditions.

In both RELAX and CORMA several jump models are available to treat methyl group rotation, and for considering internal motions the Lipari Szabo 'model free' approach can be used. In addition, it is possible to specify separate effective correlation times for each interaction. In the case where an X-ray structure is available an auxiliary program in CORMA allows to provide the automatic calculation of atomic diffusion times from the crystallographic B-factors. Solvent and/or chemical exchange are modeled by an exchange matrix, and calculations can be performed for an ensemble of structures where it is possible to specify the relative contribution of each ensemble member, and occupancy values can be specified separately for each atom which might be useful for example for partially deuterated samples.

In the program DINOSAUR [254] internal motions like methyl group rotation, aromatic ring flip and fast local motions can be considered. Within X-PLOR [258] internal motions can be described by the model-free approach from Lipari and Szabo, and the rotation of methyl groups and aromatic ring-flips are modeled by distance averaging.

In IRMA [251] the molecule is assumed to be isotropically tumbling and internal motions such as aromatic ring-flips and methyl group rotation are allowed for. As in SPIRIT [262], the molecule is generally assumed to be rigid, however, slow and fast internal motions, e.g. rotation of methyl groups are considered. MORASS [252] assumes a rigid molecule with fast rotating methyl groups. The molecule is treated as rigid in BIRDER [259], BCKCALC [250], MIDGE [255], and NO2DI [256].

In BIRDER anisotropic tumbling for non-spherical molecules is incorporated, and differential external relaxation rates for protons in different chemical environments are possible. For anisotropic tumbling the molecule is considered as a rigid symmetric top with diffusion coefficients $D_\parallel$ and $D_\perp$ parallel and perpendicular to the symmetry axis, respectively. However, the user has to provide the ratio of $D_\parallel/D_\perp$ to the program that can be calculated from the hydrodynamic theory of Tirado and de la Torre [269] using for example their program HYDRONMR. Also the differential external relaxation rates that adjust for differences between the simulated and experimental data have to be provided by the user.

In SPIRIT as in BIRDER, additional manual input is required for anisotropic tumbling and for the use of differential external relaxation rates.

In experimental NOESY spectra z-magnetization usually recovers only partially between scans. As a consequence it is important for a realistic simulation to allow the back-calculation with finite relaxation delays, as it is done in the programs BIRDER, RELAX, and SPIRIT on the basis of Eq. (2.8).

In NMR the three-dimensional conformation of a molecule is usually described by an ensemble of structures. As a consequence in RELAX, CORMA and SPIRIT a single trial structure or a set of trial structures can be used as input. To simulate effects like exchange with solvent or partial deuteration it is possible in programs such as RELAX and X-PLOR to define separate occupancies for the various atoms.

The programs RELAX and SPIRIT allow the simulation of ^{15}N or ^{13}C edited 3D NOESY spectra. Different transfer efficiencies of the INEPT and reverse INEPT steps can be calculated directly by SPIRIT or can be taken from

the corresponding 2D HSQC spectrum as implemented in both programs. In SPIRIT it is possible to simulate the different effects on CH, CH_2, and CH_3 groups caused by sensitivity enhancement schemes used in pulse sequences. The program CORCEMA [265] may be viewed as an extension of CORMA, it is aimed at the investigation of interacting systems such as ligand-enzyme complexes that undergo multistate conformational exchange. In contrast to most other programs that use matrix diagonalization to calculate NOE signals BCKCALC [250] applies numerical integration of the Bloch equations. In this approach a cross-relaxation rate scaling parameter and a so-called 'Z leakage' parameter are used for an empirical spectral density description.

Deriving accurate distance information from NOESY spectra. One important application of NOE back-calculation approaches is to derive accurate distance information from NOESY spectra as is done in IRMA [251], MARDIGRAS [253], MORASS [252], NO2DI [256], MIDGE [255], and a program by Kim and Reid [257]. In some approaches an initial relaxation matrix is iteratively refined until the difference between experimental and simulated intensities is minimal, e.g. in MARDIGRAS [253], MORASS [252], NO2DI [256], and MIDGE [255].

Relaxation matrix calculations can be applied iteratively with three-dimensional structure calculations, e.g. in IRMA [251] and a program by Kim and Reid [257] but these are quite time-consuming approaches. In IRMA the program starts with an initial structure model. For this model the relaxation matrix is set up and the corresponding NOE matrix is calculated. If possible theoretical off-diagonal elements are now replaced by the corresponding experimental values. This combined NOE matrix is back-transformed to obtain a relaxation matrix where the off-diagonal elements now include spin diffusion effects. A set of these matrices for a series of mixing times is averaged and from this averaged matrix improved distance restraints (in the first cycle to obtain distance restraints at all) are obtained, which in turn can be used to obtain a refined structural model.

The program by Kim and Reid [257], which is similar to IRMA, combines relaxation matrix analysis and structure calculations for refinement purposes. It is based on the program BCKCALC. In each cycle corresponding experimental and simulated volumes are compared with each other and in cases with large differences between simulated and experimental volumes the corresponding distance constraints are scaled accordingly. With the refined distance restraints new structures are calculated which allow an improved NOE simulation.

MARDIGRAS is based on the program CORMA described above. In contrast to IRMA, for example, new structure calculations are not required in each iteration step. Using CORMA and a starting model a theoretical NOE matrix is calculated and merged with the corresponding experimental NOE matrix. An improved relaxation matrix is back-calculated from the embedded NOE matrix to allow the calculation of a refined theoretical NOE matrix. The procedure is repeated until the error between simulated and experimental NOEs reaches a minimum. Distances are then calculated from the final cross-relaxation rates. MARDIGRAS has been extended to the determination of distances from ROESY spectra using the CARNIVAL algorithm [270]. Experimental cross-peak intensities can be symmetrized with the program SYMM in order to use partially relaxed experimental spectra [271].

In the program NO2DI [256] no knowledge about the three-dimensional structure of the molecule is required. Therefore, the starting distances in the relaxation matrix calculations can be far off from the real distances. In an iterative procedure these distances will be scaled by the sixth root ratio of calculated and experimental volumes to obtain an improved set of distances which will be used for the next round relaxation matrix calculations. Also the program MIDGE [255] requires no starting structural model for its calculations.

Recently, we have included the REFINE algorithm (to be published) within RELAX to obtain accurate distance information from NOE data. All features of RELAX, e.g. all motional models are also available in REFINE. The implemented iterative algorithm is in some regards similar to NO2DI. However, in contrast to NO2DI a distance matrix is not required in REFINE. Instead the already available relaxation matrix is iteratively refined. The rates σ_{ij} of step $n+1$ are calculated from the rates of the previous one, $\sigma_{ij}(n)$ by

$$\sigma_{ij}(n+1) = \sigma_{ij}(n)\frac{\ln cA_{ij}(\exp)}{\ln A_{ij}(n,\mathrm{sim})} \tag{2.21}$$

with an arbitrary scaling factor c to take into account unknown experimental and instrumental factors, the experimental cross-peak volumes $A_{ij}(\exp)$, and the corresponding simulated volumes $A_{ij}(n,\mathrm{sim})$ of step n. After the auto relaxation rates have been adjusted, new NOEs are calculated from the refined relaxation matrix and the next iteration step is performed. After convergence, distances are obtained from the refined relaxation matrix (manuscript in preparation).

In the programs DINOSAUR [254] and X-PLOR [258] relaxation matrix calculations are used for structural refinement by directly minimizing the difference between observed and simulated NOE intensities. Therefore, it is not necessary to convert the experimental NOEs into distances. Calculations within X-PLOR [258] can be restricted to a subset of cross peaks to save time, and the occupancy of separate atomic sites can be specified, e.g. to account for solvent exchange processes and to take partial deuteration into account. In order to save time X-PLOR allows restriction of the calculations to a subset of cross peaks. For the molecular dynamics calculations performed in DINOSAUR the program is interfaced with the GROMOS force field [272].

Secondary structure determination from chemical shifts. Three-dimensional structure calculations are usually not solely based on NOE derived distance restraints but are supplemented with additional information. Even before the extraction of exact structural information for, e.g. the assignment of the NOESY spectra it is helpful to have an idea about the secondary structure of the molecule. A statistical analysis of more than 70 proteins has revealed a strong relationship between secondary chemical shifts and a proteins secondary structure [218]. As a consequence after a completed resonance line assignment the secondary structure of the molecule can be deduced from chemical shifts reasonably well. However, it should be noted that in most cases for a precise determination of the secondary structure elements additional information such as NOE pattern information is required.

Automated procedures for secondary structure identification have been developed by several groups [273–275]. In the chemical shift index method CSI [273,274] Cα, Cβ, C′, and Hα shifts are compared with random coil values and a set of rules is applied for secondary structure determination. In the probabilistic method from Wang and Jardetzky the experimentally derived N, Cα, C′, Cβ, Hα, and HN chemical shifts from the trial protein structure are automatically compared with a set of database values. The program uses its own database derived from a statistical analysis of 36 distinct proteins. PSICSI [276] combines information from chemical shifts and the primary sequence to obtain secondary structure predictions using neural networks. The program was trained on 92 proteins for which chemical shifts as well as secondary structure and tertiary structure information was available. In this approach sequence information can be used as a substitute for sparse NMR data.

Automated determination of backbone and side-chain dihedral angles. Since chemical shifts can be used for secondary structure determination it is obvious that they can also be used for the prediction of dihedral angles. The following three methods are based on the use of databases constructed from known structures. In the approach described by Beger and Bolton [277] experimentally observed HN, N, Cα, Hα, and Cβ chemical shifts are empirically correlated with the corresponding observed ϕ and ψ dihedral angles in 49 known X-ray and NMR structures. This allows the chemical shift based prediction of backbone dihedral angles for proteins of unknown tertiary structure.

TALOS [278] uses a slightly different set of chemical shifts (Cα, Cβ, CO, Hα, and N chemical shifts). The TALOS approach makes use of a database constructed from 20 proteins for which high resolution X-ray structures and chemical shift tables are available. In comparison to the program by Beger and Bolton local sequence information is also included in the prediction process. TALOS searches its database for tripeptides with chemical shift and residue type homology to the query sequence. The backbone dihedral angles of the central residues of the 10 best matches are then averaged to obtain the predicted values.

Using Cβ and Cγ chemical shifts it is also possible to differentiate between *cis* and *trans* peptide bonds that are preceding prolines [279]. In the program POP predictions are made in a probabilistic fashion based on a database generated from 1033 prolines for which Cβ and Cγ chemical shifts and three-dimensional structural information are available.

In the following methods dihedral angles are more directly calculated from experimental data other than chemical shifts. Most of these programs employ one- or multidimensional grid-search methods. MULDER [280] and a program by Kloiber et al. [281] employ one-dimensional methods, ANGLESEARCH [282], FOUND [283], and HYPER [284] use multidimensional methods.

The multidimensional grid search algorithm in HYPER is hierarchical in the sense that constraints that greatly limit the conformational space are searched first. Experimental input includes usually distance and/or scalar coupling restraints. An exception is the program by Kloiber et al. [281] that uses five cross-correlated NMR spin relaxation rates for the backbone nuclei and the $^3J_{C'-C'}$ scalar-coupling constant to determine backbone dihedral angles. In addition to experimental restraints FOUND takes steric, and stereochemical restraints into account as well. All the programs provide torsion-angle restraints, and more specifically backbone ϕ, ψ, and side-chain χ_1 dihedral angles together with the stereospecific assignments of the Hβs are automatically determined within HYPER. In ANGLESEARCH χ_2 angles are also determined.

The program by Kloiber et al. [281] specially aims at the automated determination of backbone dihedral angles. In FOUND the results include allowed torsion-angle ranges, and stereospecific assignments for diastereotopic substituents. A special feature of FOUND is that it can be applied to contiguous nucleic acid or protein fragments of arbitrary length. FOUND has been incorporated in the structure determination package DYANA.

2.2.4.5. Structure calculation. In most cases it is still not feasible to sample the conformational space of a biological macromolecule in an exhaustive way for all conformations in agreement with the experimental restraints. Traditionally, the main methods in use are distance geometry (DG), restrained molecular dynamics (rMD), and simulated annealing (SA) although pure distance geometry methods are only rarely used now.

In pure distance geometry, a matrix of distances between all atoms that is consistent with the input is created from the experimental restraints together with the covalent structure of the molecule. This set of distances from *n*-dimensional space is projected into three-dimensional Cartesian coordinate space in Metric Matrix Distance Geometry Methods, e.g. DISGEO [2]. The advantage of this method is that

a direct method for solving the structure exists, that it is computationally very fast, and that no initial structure model is required. A disadvantage of distance geometry is that distances are not sufficient to define the chirality of a structure. However, in the case where local or global mirror images occur these can be rejected at an early stage since the chiralities of the amino acids and helices are known. Structures obtained by distance geometry usually contain violations in bond lengths and bond angles and closer distances than van der Waals interatomic distances. As a consequence the structures obtained by distance geometry should be refined with, for example, restrained molecular dynamics techniques.

Another possibility is to work in dihedral angle space where bond lengths and bond angles are kept fixed and only the torsion angles are allowed to be varied, e.g. DISMAN [2] and DIANA [285]. In this approach the conformation of the protein can be calculated from a set of distance constraints by minimizing a target function that is zero if all distances are compatible with the constraints.

In the FANTOM (Soman et al., unpublished) approach energy minimizations and Monte Carlo simulations can be performed in torsion-angle space.

Several programs for restrained molecular dynamics have been developed in the last years for protein structure calculation, e.g. X-PLOR [258], GROMOS [272], DL_POLY [286], DYANA [287], CNS [288], NAMD2 [289], CYANA [247], AMBER [290], X-PLOR-NIH [291], and INSIGHT II-CHARMM [292].

In restrained molecular dynamics (mostly combined with a simulated annealing protocol), the total potential energy V_{total} of the molecule is minimized. V_{total} usually comprises the following terms:

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{dihedr}} + V_{\text{vdW}} + V_{\text{coulomb}} + V_{\text{exp}} \tag{2.22}$$

The first five terms are empirical energy terms that describe the physical interactions of the atoms such as the strength of the covalent bond (V_{bond}), the bond angle (V_{angle}), the dihedral angle (V_{dihed}), the van-der-Waals interaction (V_{vdW}), and the electrostatic interaction (V_{coulomb}). In contrast, V_{exp} contains the experimental NMR information such as distance and dihedral angle restraints. It should be noted that V_{exp} does not correspond to any real physical force. Thus the form of the potential is not predefined and is often assumed to be a simple harmonic potential. For distance restraints from NOE data V_{NOE} can thus be defined as

$$V_{\text{NOE}} = \sum V_{\text{NOE}}(ij) = k_{\text{NOE}} \sum (r_{ij} - r_{ij,0})^2 \tag{2.23}$$

with k_{NOE} an arbitrary constant, r_{ij} and $r_{ij,0}$ the actual and the expected distances between the atoms i and j. For computational stability potentials should not exceed some threshold values with the condition $V_{\text{NOE}} = T_{\text{NOE}}$ if it would be larger than the threshold. The estimated error of r_{ij}^0 is usually expressed in the form of a lower and upper limit $r_{ij,\text{low}}$ and $r_{ij,\text{up}}$ of the distance.

To most users it is not clear that these values are only defined if a confidence level of, e.g. 0.995 is predefined for the hypothesis that the distance is expected to be located in these error limits. Even worse, the error is usually just set to an arbitrary value, e.g. a certain percentage of r_{ij}^0 without considering if the data would require completely different values. However, it is possible to obtain reasonable error estimates for individual cross peaks directly from the data (Trenner et al., in preparation).

For including the presumed error equation (2.23) is often modified to

$$V_{\text{NOE}}(ij) = 0, \qquad \text{for} \qquad r_{ij,\text{low}} \le r_{ij} \le r_{ij,\text{up}} \tag{2.24a}$$

$$V_{\text{NOE}}(ij) = k_{\text{NOE}}(r_{ij} - r_{ij,\text{low}})^2, \qquad \text{for} \qquad r_{ij,\text{low}} \ge r_{ij} \tag{2.24b}$$

$$V_{\text{NOE}}(ij) = k_{\text{NOE}}(r_{ij} - r_{ij,\text{up}})^2, \qquad \text{for} \qquad r_{ij,\text{up}} \le r_{ij} \tag{2.24c}$$

and

$$V_{\text{NOE}} = \min(V_{\text{NOE}}(ij), T_{\text{NOE}}) \tag{2.24d}$$

There is a similar definition of the NOE potential for ambiguous NOEs (Eq. (2.4)). However, since the potential should reflect the error distribution p of the measured quantity one can derive more reasonable potentials. According to Sippl [293] the appropriate potential V is related to the probability p and the state integral Z by

$$V = -k_B T \ln(pZ) \tag{2.25a}$$

with

$$Z = \int \cdots \int \exp\left[-\frac{V(x)}{kT}\right] \mathrm{d}x \tag{2.25b}$$

where $V(x)$ is the potential of a particular state x. This means for a normal distributed quantity (as, for example, to a first approximation the peak volume), a harmonic potential is in fact an adequate description. For the distances derived from NOEs one obtains (Kalbitzer and Gronwald, to be published)

$$V_{\text{NOE}}(ij) = k_B T \left(\ln\sqrt{2\pi}\sigma + \frac{1}{2\sigma^2}\left(\frac{1}{r_{ij}^{12}} - \frac{2}{r_{ij}^6 r_{ij,0}^6} + \frac{1}{r_{ij,0}^{12}} \right) - \ln Z \right) \tag{2.26}$$

Here, T is the absolute temperature and σ the standard deviation of the measured values of the cross-peak volume.

In molecular dynamics, structures are calculated by solving Newton's equation of motion (for a review, see Ref. [294])

$$\mathbf{F}_i = m_i \mathbf{a}_i \tag{2.27}$$

The force $\mathbf{F}_i$ on atom i can be obtained from the derivative of V_{total} with respect to the coordinates r_i as

$$\mathbf{F}_i = -\frac{\mathrm{d}V}{\mathrm{d}\mathbf{r}_i} = m_i \frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t_i^2} \tag{2.28}$$

Since the atom masses m_i and V_{total} are known this equation can be solved in order to obtain future atom positions in time t_i. If the temperature T of the simulation is given, the atomic velocities are related to the temperature by

$$\frac{3N}{2} k_B T = \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2 \tag{2.29}$$

and should be described by a Maxwell-distribution. Thus temperature can be simulated by assigning to the individual proteins velocities from a Maxwell distribution. Usually rotation around and translation of the center of mass are removed from the system. During the simulations the temperature T can be kept constant by scaling the velocities v_i after each step Δt, with a scaling factor λ this is often referred to as coupling to a bath of constant temperature T_0. The scaling factor λ is given by

$$\lambda = [1 + (T_0/T - 1)\Delta t/\tau_T]^{1/2} \tag{2.30}$$

where τ_T is the time constant for the temperature coupling.

To overcome local minima a restrained molecular dynamics and simulated annealing protocol can be used. First the system is allowed to increase its kinetic energy (temperature) and then the kinetic energy is slowly decreased. Compared to distance geometry methods the required amount of computational time is drastically increased especially in cases that start from random structures. However, with the constantly increasing computer power molecular dynamics calculations can nowadays also be started directly from random conformations. Key advantages include that local minima can be overcome, an even sampling of the conformational space, that the deviations from ideal geometry of, e.g. planar rings are usually small, and that usually good non-bonded interactions can be obtained.

Therefore, both methods are often combined where an initial structure is calculated using distance geometry which is then refined using simulated annealing. Also a considerable amount of time can be saved by simulated annealing techniques operating not in Cartesian space but in dihedral angle space as it is implemented in DYANA [287] and CYANA [247] or by using programs that were especially designed for parallelizing the molecular dynamics calculations, e.g. DL_POLY [286] and NAMD2 [289].

Distance geometry algorithms are used by the programs DISGEO and DISMAN [2]. The DISMAN program first calculates local conformations of the polypeptide chain from short-range restraints and then gradually adds long-range restraints to obtain the global structure. In contrast DISGEO tries to obtain the global structure first and then improves the local geometry. The next step was the distance geometry program DIANA [285] that, like DISMAN, has a variable target function that is minimized during structure calculations in torsion-angle space.

The successor of DIANA called DYANA [287] performs torsion-angle dynamics to obtain three-dimensional structures from NMR derived distance and torsion-angle constraints. In comparison to calculations in Cartesian space the required amount of computational time is drastically reduced due to the reduced number of degrees of freedom and the absence of high frequency bond and angle vibrations. DYANA was further developed into the program CYANA [247] that contains as a main new feature the combination with the automated assignment module CANDID.

INSIGHT II [292] contains the CHARMM module that enables molecular dynamics calculations and energy minimizations to be done in Cartesian space of proteins, nucleic acids and other biological macromolecules. Experimentally derived distance and dihedral angle restraints can be incorporated. X-PLOR, which was initially derived from CHARMM is one of the most widely used programs. It allows simulated annealing structure calculations in torsion angle and/or Cartesian space. A new development based upon X-PLOR is CNS [288]. Data input for CNS includes NOE-derived distances, NOE intensities, torsion-angle restraints, coupling constants, homo- and heteronuclear chemical shift information, residual dipolar couplings, heteronuclear T_1/T_2 ratios, and a full relaxation matrix approach can be used for the direct refinement against NOE intensities.

The original X-PLOR and CNS programs are no longer in active development. New additional NMR specific features are now incorporated in X-PLOR-NIH [291] and this contains all the functionality of X-PLOR 3.851. In addition to the features mentioned for CNS XPLOR-NIH includes the automated NOE assignment module ARIA, additional potentials for NMR observables, e.g. $^1J_{\mathrm{C}\alpha-\mathrm{H}\alpha}$ coupling constant restraints related to ϕ and ψ angles, and three bond amide deuterium isotope effects on $^{13}\mathrm{C}'$ shifts related to ψ angles. An important addition is the inclusion of knowledge-based potentials of mean force generated from known high resolution 3D structures. These potentials should be especially important in regions of the molecule where only a limited number of experimental restraints are available. Multidimensional torsion angle database potentials of mean force for proteins and nucleic acids are also included. Joint NMR/X-ray refinement calculations are also possible.

The name AMBER [290] usually refers to a set of molecular mechanical force fields for molecular dynamics simulations of biomolecules, and to several programs to perform these simulations. The AMBER module SANDER is the general molecular dynamics and energy minimization program working in Cartesian space. It allows simulated annealing structure calculation based on NMR derived restraints. Experimental input includes NOE-derived

distances, NOE intensities, torsion-angle restraints, scalar coupling constants, proton chemical shifts and residual dipolar couplings. For the refinement against NOE intensities a full relaxation matrix approach is included.

The GROMOS [272] program also allows the NMR based three-dimensional structure calculation of biological macromolecules in Cartesian space using molecular dynamics. Restraints obtained from experimental data can be defined for NOE-derived distances, dihedral angles and *J*-coupling values.

In contrast to most other known programs, Gippert et al. [295] have developed an approach (DTAGS/NEWMOL) for systematic grid searches in torsion angle conformational space to obtain all structures in agreement with experimental restraints. The main idea to increase efficiency in this method is to use the allowed conformational space of smaller fragments to restrict trial conformations of larger fragments. It is feasible to apply the method to smaller peptides and to overlapping fragments of medium sized proteins. For larger macromolecules such as proteins the main applications will probably include systematic searches for allowed loop conformations and the validation of experimental restraints and stereospecific assignments.

One of the most versatile programs for biomolecular structure determination from NMR data is X-PLOR in its current version since it allows for a very broad range of input from experimental data. Also it is capable of performing distance geometry calculations as well as molecular dynamics simulations in Cartesian and dihedral angle space. A computationally very efficient program that allows calculating a set of structures in a comparably short amount of time is DYANA/CYANA. This might provide an important advantage in cases where automated peak assignments are combined with structure calculations in an iterative manner.

Structure determination using sparse NMR data. One avenue to speed up the structure determination process is to reduce the required number of restraints and/or to use only restraints that are relatively easily available, e.g. backbone dihedral angles, chemical shifts, residual dipolar couplings, hydrogen bonds, or HN–HN NOEs. These methods should be applicable in particular to cases where one is more interested in the global fold of the molecule than in a highly detailed structure. In the easiest application one can check if the trial protein adopts a previously known fold [296]. Several of the methods published so far rely on the combination of modeling and NMR techniques. Bowers et al. describe an approach that combines the ROSETTA ab initio protein structure prediction method with sparse NMR data [297]. The ROSETTA method assembles protein structures from fragments of known structures with sequences similar to the target protein [298,299]. Here in addition chemical shift and NOE data are used in the fragment selection process.

More specifically chemical shift data are employed to generate backbone dihedral angles, using the TALOS algorithm described above [278], which are then used together with NOE data in the fragment selection process. From the selected fragments models are built by minimizing an energy function that emphasizes terms for hydrophobic burial, β-strand pairing, and NOE restraint satisfaction.

In the MFR approach of Delagio et al. [300] a starting structure is generated searching a database of three-dimensional protein structure fragments generated from the PDB database. Fragments are selected whose predicted residual dipolar couplings best fit the set of measured values, and also the fit between measured and predicted chemical shifts is considered to a less degree in fragment selection. Average values for the backbone dihedral angles of the selected fragments are calculated, and are used to generate initial models. Models are further refined by adjusting the backbone dihedral angles to optimize the fit between predicted and measured values of the residual dipolar couplings and chemical shifts, respectively.

Since it is possible to back-calculate residual dipolar couplings from known structures experimentally measured residual dipolar couplings can be used to search a database of structural fragments to find the best-fitting fragments [301]. Using overlapping fragments a structural model of the trial proteins backbone is generated.

In a different approach it is demonstrated that fold prediction by protein threading can be shown to be improved by including experimental distance constraints obtained from, e.g. mass spectroscopy or NOE measurements [302].

Other methods rely solely on the use of NMR data [296, 303–306]. The approach by Bonvin et al. uses hydrogen bond restraints obtained from experimentally measured cross-hydrogen bond $^{3hb}J_{NC'}$ coupling constants and backbone dihedral angle restraints obtained from an analysis of secondary chemical shifts. Using this limited set of restraints calculations with CNS resulted in a set of structures which did not converge into one single fold. Therefore, the structures were further analyzed by grouping similar structures into clusters. It was shown that the correct fold could be obtained by calculating the average structure from the cluster containing the most similar structures.

In another investigation it was shown that three different residual dipolar coupling measurements together with sparse long-range HN–HN NOE contacts are sufficient to define the global fold of Ubiquitin [304]. Using only residual dipolar couplings it is possible to determine if a protein adopts a previously known fold [296]. The method is based on the comparison between measured residual dipolar couplings and the corresponding values computed from known structures. In this process an explicit structure determination of the target protein is not necessary.

In fully deuterated proteins it is possible to obtain long range NOEs between amide protons corresponding to distances up to 0.8 nm. Together with secondary structure restraints obtained from a chemical shift analysis this

information is sufficient to calculate structures of medium precision [305].

Another method is based on the use of deuterated proteins with selectively protonated side-chain methyl groups [306]. Here the programs AUTOASSIGN and AUTOSTRUCTURE (described above) are combined with the STAC algorithm that allows classification of spin systems based on the presence of methyl groups. Structures are calculated using relatively few NOEs, hydrogen bonds and residual dipolar couplings. This limited set of constraints is sufficient for fold determination.

2.2.5. Structure validation

One of the most important points in any automated or manual structure determination process is the assessment of the quality of the resulting structures. As a final aim one wants to know if the solved structure really reflects the true structure present in the natural environment of the molecule of interest. However, an intermediate but still worthwhile aim is to show that the obtained structures do optimally explain all the experimental evidence available. In fact, the question about the true structure is ill posed as long as it is not defined what is meant by 'structure'. Since the ensemble of all structural states of a protein is infinite when characterized by continuous variables such as Cartesian coordinates or torsion angles, it can only be characterized by a limited subset of all structures. Here, the 'lowest-energy' structure that is the structure with lowest Gibbs free energy at given external conditions is a rather well-defined concept although more than one structure could possibly have the same energy. The use of a 'mean' structure makes sense only when a subset of closely related structures are selected since averaging of very different structures (e.g. a second structural state or random-coil structures) does not give information of practical use in structural biology.

The overall precision of an NMR structure is usually expressed either as an average pairwise root-mean-square deviation (rmsd) of the coordinates of the selected ensemble of structures or as an rmsd of the structures relative to the mean coordinates of the ensemble. However, rmsd values are a measure of the precision of the structures in the ensemble but not necessarily for their accuracy.

Another measure for the quality of an NMR structure is how well the obtained structures agree with the experimental data. Therefore, the number and sizes of violated restraints, such as distance, dihedral angle, hydrogen bond, and residual dipolar coupling restraints can be analyzed. An NMR *R*-factor provides a direct measure of how well the resulting structures fit the corresponding experimental NOESY spectra. Often the overall quality of the experimental data itself is judged by the number of restraints per residue. A measure that is independent of the experimental data is the quality of the geometrical properties of the molecule, e.g. the comparison of bond lengths, bond angles, dihedral angles, etc. with standard values obtained for example from a set of high resolution structures. To judge the overall quality of a protein structure it will, in most cases, not be sufficient to rely solely on one of the indicators mentioned above but to consider most of them simultaneously.

A general overview of the validation of experimentally derived X-ray and NMR structures is given by Laskowski et al. [307]. However, we will summarize in the following some of the methods applicable to NMR spectroscopy.

NMR R-factors. A comparison between experimental and back-calculated NOESY spectra leads to an error function similar to the *R*-factor (residual factor) used in crystallography [308]. An NMR *R*-factor gives a direct value for the quality of the NMR structure obtained; more precisely it gives a measure for the agreement between experimental data and the estimated structure. As in the case of X-ray crystallography, the *R*-factor can only be interpreted in the context of the quality of the data because poor data can result in low *R*-factors and hence falsely indicate a high quality of the structures. This is a well known fact in X-ray crystallography where *R*-factors are only interpreted in conjunction with the resolution of the crystals. In NMR a corresponding measure does not exist yet, although the completeness of the NOESY-spectra defined by Doreleijers et al. [309] provides a reasonable but not ideal measure for the quality of the data.

In the literature different NMR *R*-factor definitions can be found. Generally, they are based on the comparison of experimental NOEs with the corresponding back-calculated NOEs obtained from a relaxation matrix analysis. However, differences exist which parameters, such as motional models, are taken into account in the back-calculations. One major application of *R*-factors is quality determination [254,310–312]. In the paper by Gonzalez et al. various linear and quadratic *R*-factor definitions are compared with each other. However, no major differences were found on the sensitivity of these *R*-factors to structural changes. Also the conversion of volumes into distance-like quantities by using the sixth-root of the volume was tested to define an *R*-factor that is more closely related to the NOE energy used in structure calculations. Similar *R*-factor definitions were successfully tested by Thomas et al. and Xu et al. [313,314]. An *R*-factor taking into account integration errors due to noise and spectral overlap was defined by Nilges et al. [315].

In other applications *R*-factors are used for structural refinement. In these generally iterative approaches the difference between simulated and experimental intensities is minimized during structure calculations [253,310, 316–320]. In the MARDIGAS approach by Borgias and James an iterative relaxation matrix approach is used to obtain accurate distance information from NOE spectra. *R*-factors are used here to provide a stop criterion for the iteration process. A hybrid relaxation matrix approach is used for the refinement of a nucleotide structure by Nikonowicz et al. [321]. In the approach by Mertz et al. [322] the refinement is performed in dihedral angle space employing the distance geometry program DIANA.

To our knowledge in all of these approaches manually assigned NOESY peaks are compared to their corresponding back-calculated counterparts. Therefore, these *R*-factor calculations are only possible after the time consuming manual assignment of the NOESY spectra has been performed. However, it often would be useful to estimate the agreement of a structural model with the experimental data independent of the completion of the NOESY assignment. Typical examples could include the selection of a proper starting model for automated or manual structure based NOE assignments or for answering the question of whether a structure solved by X-ray crystallography also applies in solution.

For that we have developed the computer program RFAC [323] which allows the automated estimation of *R*-factors from protein NMR-structures and gives a reliable measure for the quality of the structures. The *R*-factor calculation is based on the comparison of experimental and simulated ^{1}H NOESY NMR spectra. The approach comprises an automatic peak picking and a Bayesian analysis of the data, followed by an automated structure based assignment of the NOESY spectra and the calculation of the *R*-factor.

The major difference to previously published *R*-factor definitions is that RFAC takes the non-assigned experimental peaks into account as well. The number and the intensities of the non-assigned signals are an important measure for the quality of a NMR structure. It turns out that optimally adapted *R*-factors should be used for different problems. RFAC (which is implemented in AUREMOL) allows the computation of a global *R*-factor, different *R*-factors for the intra-residual NOEs, the inter-residual NOEs, sequential NOEs, medium-range NOEs and long-range NOEs. *R*-factors can be calculated for various user-defined parts of the molecule or it is possible to obtain a residue-by-residue *R*-factor. Another possibility is to sort the *R*-factors according to their corresponding distances. The summary of all these different *R*-factors should allow the user to judge the structure in detail. A comparison with a previously published *R*-factor definition shows that the approach of RFAC is more sensitive to errors in the calculated structure.

Automated R-factor determination. The automated *R*-factor analysis consists in principle of two separate parts: (1) the comparison of the experimental NOESY spectrum with the NOESY spectrum back-calculated from a given structure, and (2) the calculation of the *R*-factor(s) from the data. In the first part the NOESY spectrum has to be calculated from the trial structure using the sequential assignments; that is, for a meaningful *R*-factor the spin systems must have been assigned completely or almost completely. The back-calculation of the NOESY-spectra should be as perfect as possible, that is the application of a full relaxation matrix approach should be used, which corresponds to the state of the art, although initial slope approaches could also be used.

A schematic representation of the steps essential for automated *R*-factor determination are given in Fig. 6. Since practical NMR spectra have a limited quality and contain a large number of noise and artifact peaks, information about the validity of a given cross-peak must enter in the *R*-factor calculation. This can be done by using the probabilities p_i of the peaks i to be true NMR signals and not noise or artifact peaks. They can be calculated according to Bayes theorem

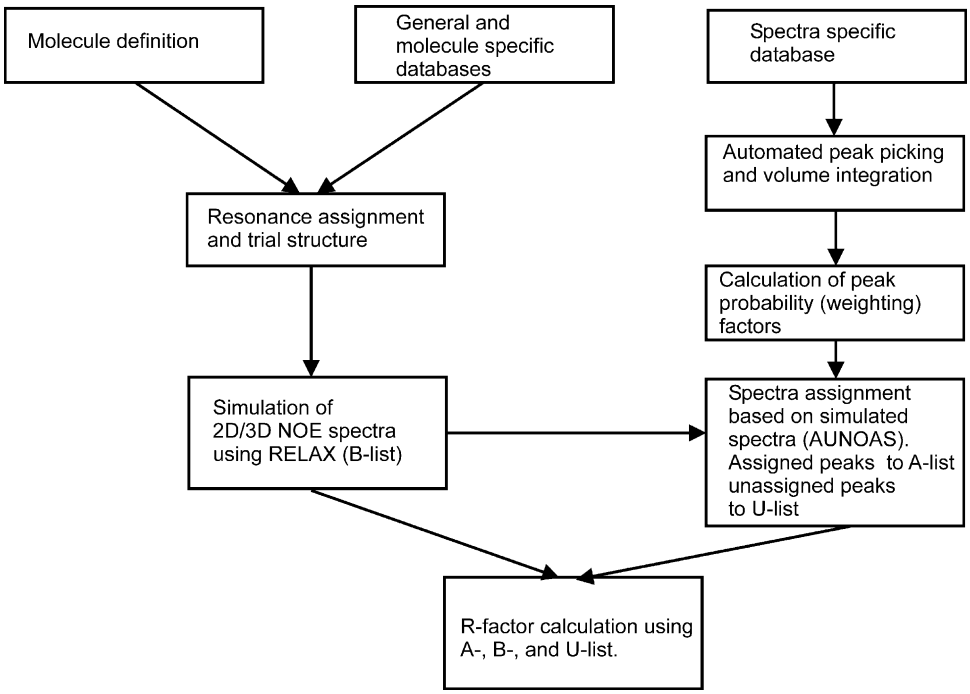

Fig. 6. Schematic representation of the computation steps required for automated *R*-factor determination with the program RFAC.

[182,183], for a detailed description, see Section 2.2.4.2, and can be used as weighting factors during the calculation of the *R*-factors. In a next step of the *R*-factor calculation the experimental cross peaks must be automatically assigned to entries of a basic assignment table. Since the general chemical shift assignments are usually the result of many different experiments they do not fit exactly to the spectrum under consideration. This means that one has first to optimally adapt the chemical shift values obtained from the general sequential resonance assignment to the actual experimental data by a global comparison of the back-calculated spectrum with the experimental spectrum. The next step would then be the actual assignment of cross peaks, which is usually not unique, since a given cross peak often has more than one explanation.

The experimental signals for which no corresponding simulated peaks were found and which therefore remain unassigned will be called the set *U* in the following. The *U*-list can be further reduced by applying a lattice algorithm which can be used if one assumes that the sequential assignment is true and almost complete. In this algorithm only non-assigned peaks are taken into account where at least one back-calculated peak in each dimension can be found within user defined search radii, e.g. 0.01 ppm for 2D spectra. In this context it is important to note that for each atom at least the structure independent diagonal peak is back calculated. In case that more than one back-calculated peak is assigned to a single experimental peak, the mean volume of the corresponding back-calculated peaks is estimated before the comparison is done while the volume of the experimental peak is divided by the number of corresponding back-calculated peaks.

In general, the *R*-factor should measure the agreement between the experimental data set and the data back-calculated from the structure. In its simplest form it is defined by:

$$R_1 = \frac{\sum_{i\in A} |I_{\mathrm{exp},i} - \mathrm{sf}\cdot I_{\mathrm{calc},i}|}{\sum_{i\in A} |I_{\mathrm{exp},i}|} \tag{2.31}$$

The summation is performed over the data points i with intensities I_i in a given set A. With this definition R is 0 if the agreement is perfect and >0 for all other cases. In NMR spectroscopy and X-ray crystallography one has to normalize the experimental data (or the calculated data) since the experimental values are scaled by a constant factor depending on not exactly known instrumental and experimental parameters. The optimal scale factor sf is found when the likelihood function $L(\mathrm{sf})$ adopts its maximum value

$$L(\mathrm{sf}) = \prod_{i\in A} p\cdot(\mathrm{sf}\cdot I_{\mathrm{calc},i}, I_{\mathrm{exp},i}) \tag{2.32}$$

Here, p is the probability that for a calculated value $\mathrm{sf}\cdot I_{\mathrm{calc},i}$ the value $I_{\mathrm{exp},i}$ is measured. From Eq. (2.32) the scale factor sf is given as:

$$\mathrm{sf} = \frac{\sum_{i\in A} I_{\mathrm{exp},i}\cdot I_{\mathrm{calc},i}}{\sum_{i\in A} I^2_{\mathrm{calc},i}} \tag{2.33}$$

The above definition of the *R*-factor is well suited for X-ray crystallography: the exact positions of the X-ray reflections are determined by the crystal lattice and are exactly known. Therefore, the assignments of the reflection spots are usually unambiguous and only the intensities of these spots determine the *R*-factor. The data set *A* corresponding to a given resolution can easily be assigned and used for the calculation of the *R*-factor (which is always dependent on *A*).

In NMR-spectroscopy, however, many assignments of experimental peaks are ambiguous and many experimental peaks are artifacts. Therefore, in the literature only the set *A* of manually assigned peaks is used for the calculation of the *R*-factor. By application of Eq. (2.31) to NOESY spectra one can define a measure for the error (Eq. (2.34)) that corresponds to a normalized mean deviation [312]. Note that in the following for all *R*-factor calculations the intensities will be replaced by their corresponding volumes *V*

$$R_2 = \sqrt{\frac{\sum_{i\in A} (V_{\mathrm{exp},i} - \mathrm{sf}\cdot V_{\mathrm{calc},i})^2}{\sum_{i\in A} V^2_{\mathrm{exp},i}}} \tag{2.34}$$

Since unlike in X-ray crystallography the set of peaks is incomplete and dominated by the (structurally less important) strong short range NOEs, *R* is dominated also by the volumes of these 'trivial' peaks. Therefore, usually *V* is replaced by a more meaningful function $f(V)$ which emphasizes the more important long range NOEs. The most common form of $f(V)$ is

$$f(V) = V^{\alpha} \tag{2.35}$$

Thus a more general form of R_2 is then given by

$$R_2(\alpha) = \sqrt{\frac{\sum_{i\in A} (V^{\alpha}_{\mathrm{exp},i} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{calc},i})^2}{\sum_{i\in A} V^{2\alpha}_{\mathrm{exp},i}}} \tag{2.36}$$

If $f(V)$ as described in Eq. (2.35) is used, the calculation of the scale factor must be changed accordingly

$$\mathrm{sf}_{\alpha} = \frac{\sum_{i\in A} (V_{\mathrm{exp},i}\cdot V_{\mathrm{calc},i})^{\alpha}}{\sum_{i\in A} V^{2\alpha}_{\mathrm{calc},i}} \tag{2.37}$$

As Eq. (2.33) the expression fulfils the important condition that it gives the correct value of sf in the error free case where all experimental and back-calculated peak volumes differ only by a proportionality factor. With $\alpha = -1/6$ $f(V)$ is in first-order proportional to the internuclear distance and one

obtains the distance related *R*-factor already defined by Gonzales et al. [312].

In the automated *R*-factor calculation there is no user intervention in deciding (1) which peak in the NOESY spectrum is a true resonance and (2) if an assignment of a cross peak is correct. In principle only probabilities $p_{\mathrm{exp},i}$ and $p_{\mathrm{calc},i}$ exist for case (1) and (2), respectively. A method for estimating $p_{\mathrm{exp},i}$ has already been developed (see above). An algorithm does not yet exist for estimating $p_{\mathrm{calc},i}$ which would contain information about the validity of the model used for the simulation, including not completely adequate motional models, and the local validity of the structural model itself. Consequently, in the following we will explicitly make use only of the probabilities $p_{\mathrm{exp},i}$. With these probabilities, Eqs. (2.34) and (2.36) can be rewritten as

$$R_3(\alpha) = \sqrt{\frac{\sum_{i\in A}(V^{\alpha}_{\mathrm{exp},i} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{calc},i})^2\cdot p^2_{\mathrm{exp},i}}{\sum_{i\in A} V^{2\alpha}_{\mathrm{exp},i}\cdot p^2_{\mathrm{exp},i}}} \tag{2.38}$$

with suitable coordinates as $V_{\mathrm{calc},i}$. However, since for $\alpha = -1/6$ very small volumes in R_4 dominate the *R*-value more stable results can be expected in this case if a lower limit is set for $V_{\mathrm{calc},i}$. It is computationally efficient to set $V_{\mathrm{calc},i}$ to a value which just cannot be detected safely in the experimental spectrum that is to the intensity V_{noise} of a standard noise peak. In the present implementation, it is possible to calculate the noise intensity automatically or a user specified noise intensity could be employed. If the noise volume is calculated by the program the weakest back-calculated intensity where the corresponding distance is not greater than the detection limit is selected.

In the automatic routine a detection limit of 0.5 nm is assumed. Since in R_4 ($\alpha = -1/6$) the large distances (small volumes) dominate the expression the above normalization of the *R*-factor leads to a strong dependence on the exact value of the V_{noise} term. This influence can be diminished by inclusion of V_{noise} in the denominator:

$$R_5(\alpha) = \sqrt{\frac{\sum_{i\in A}(V^{\alpha}_{\mathrm{exp},i} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{calc},i})^2\cdot p^2_{\mathrm{exp},i} + \sum_{i\in U}(V^{\alpha}_{\mathrm{exp},i} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{noise}})^2\cdot p^2_{\mathrm{exp},i}}{\sum_{i\in A} V^{2\alpha}_{\mathrm{exp},i}\cdot p^2_{\mathrm{exp},i} + \sum_{i\in U}(V^{\alpha}_{\mathrm{exp},i} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{noise}})^2\cdot p^2_{\mathrm{exp},i}}} \tag{2.40}$$

The above *R*-factor only estimates how well the assigned peaks are explained by the structural model but they do not provide information on how well all experimental peaks are explained. For doing this, the *R*-factor should decrease if more peaks are assigned correctly and explained by the structural model. A practical expansion of Eq. (2.38) including the non-assigned peaks can be defined by

In the case of $\alpha = 1$ the standard noise intensity V_{noise} for the *R*-factors R_4 and R_5 can be set to 0, since strong unassigned signals will lead to increasing *R*-factors in this equation. With this definition the two *R*-factors in Eqs. (2.39) and (2.40) become equal.

The *R*-factors $R_{4,5}$ indicate how well the experimental signals are explained by back-calculated peaks. However, one can also define an *R*-factor to check how well the back-

$$R_4(\alpha) = \sqrt{\frac{\sum_{i\in A}(V^{\alpha}_{\mathrm{exp},i} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{calc},i})^2\cdot p^2_{\mathrm{exp},i} + \sum_{i\in U}(V^{\alpha}_{\mathrm{exp},i} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{noise}})^2\cdot p^2_{\mathrm{exp},i}}{\sum_{i\in A} V^{2\alpha}_{\mathrm{exp},i}\cdot p^2_{\mathrm{exp},i} + \sum_{i\in U} V^{2\alpha}_{\mathrm{exp},i}\cdot p^2_{\mathrm{exp},i}}} \tag{2.39}$$

The first summation is performed over all assigned experimental peaks (set *A*) and the second summation is performed over the list of unassigned peaks (set *U*). $V_{\mathrm{calc},i}$ are the corresponding calculated intensities

calculated signals are explained by experimental data. A definition analogous to R_5 uses the non-assigned back-calculated signals instead of the non-assigned experimental peaks:

$$R_6(\alpha) = \sqrt{\frac{\sum_{i\in A}(V^{\alpha}_{\mathrm{exp},i} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{calc},i})^2\cdot p^2_{\mathrm{exp},i} + \sum_{i\in U'}(\mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{noise}} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{calc},i})^2}{\sum_{i\in A} V^{2\alpha}_{\mathrm{exp},i}\cdot p^2_{\mathrm{exp},i} + \sum_{i\in U'}(\mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{noise}} - \mathrm{sf}_{\alpha}\cdot V^{\alpha}_{\mathrm{calc},i})^2}} \tag{2.41}$$

(volumes). For set *U* the logical extension of R_3 would assign the strongest back-calculated cross peak

The summation of the unassigned calculated peaks has now to be performed over a different set U' which

contains all back-calculated non-assigned peaks with volumes $V_{\mathrm{calc},i} \geq V_{\mathrm{noise}}$.

Instead of the standard noise intensity used in R_4 and R_5 one can assign specific volumes to all experimental peaks which could not be assigned unambiguously by the standard assignment routine. In this way, a new R-factor (R_7) can be defined by using Eq. (2.38) but performing the summation over all experimental peaks. The majority of peaks which are contained in set U are originally not assigned since the back-calculation from the test structure has not produced a corresponding peak with sufficient intensity. Appropriate assignments for the yet unassigned peaks could be first obtained on chemical shifts alone and the corresponding simulated volumes could be obtained with, for example, the initial slope approximation from the distances. For a more detailed description please see the original RFAC paper [323].

The above-defined R-factors are devised primarily for judging global properties. It is further possible to calculate the R-factor for previously specified regions of the molecule of interest. This allows judging how well for example a given α-helix or β-strand is defined. In this case R_3 seems to be appropriate where only the subset of the assigned peaks A is taken into account.

Another possibility using R_3 is to calculate a separate R-factor for each residue. This can be a useful tool for finding miss-assigned signals. A different way to look at R-factors is to sort them by distance. In this case R_3 is used again and signals are sorted by the corresponding distances of the calculated intensities. Fig. 7 shows an example of the application of RFAC on the HPr-protein. The separation of the R-factor in distance classes allows checking if, for example, NOEs corresponding to short distances are over proportionally violated. And this in turn could give a hint if the upper and lower bounds in the structure calculation procedure have been correctly defined.

In principle the R-factor calculation could be improved further by including the true peak shape which in addition contains J-coupling information (dihedral angles) and transverse relaxation (motional and distance information).

An R-factor can be defined analogous to the NOE based NMR R-factors which is based on the comparison of

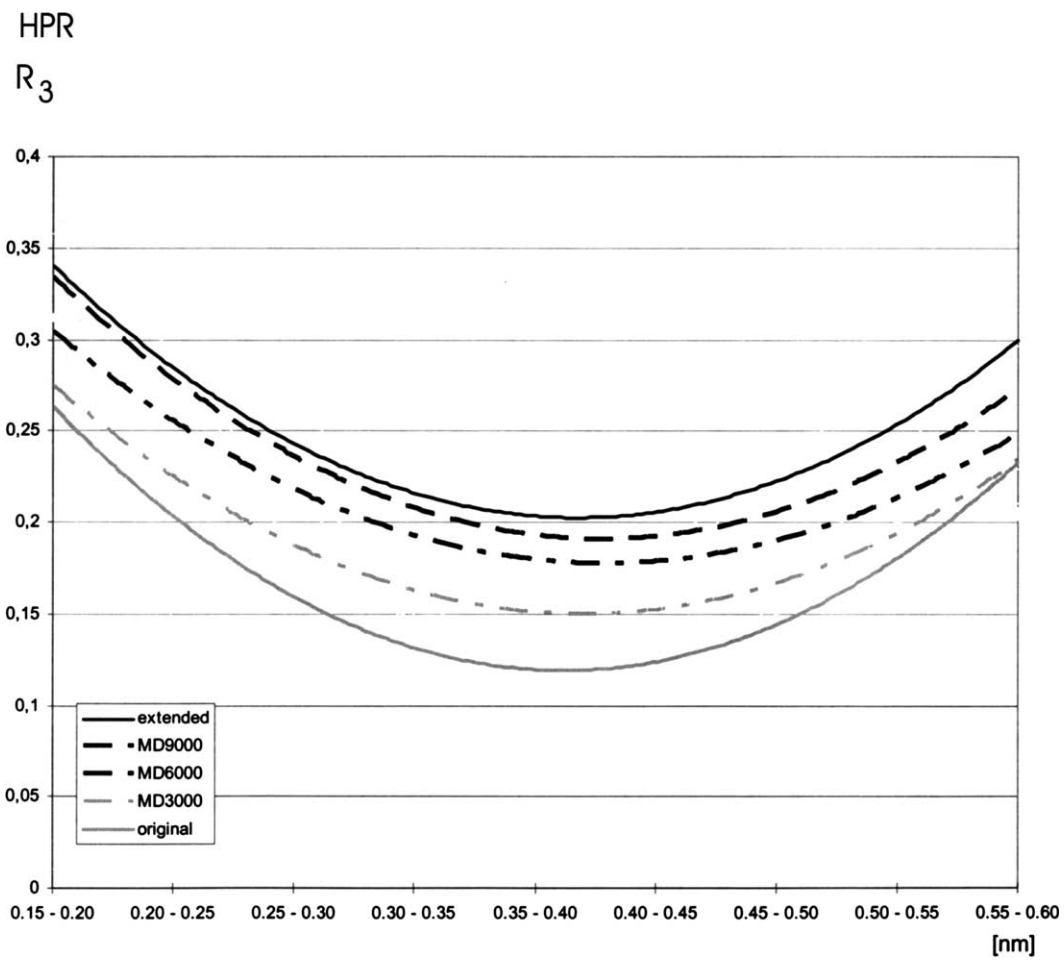

Fig. 7. Distance dependent R-factors for HPr. A polynomial smoothing of second order was applied to the data to enhance the readability of the figure. R-factors are calculated according to R_3 ($\alpha = -1/6$) using only the inter-residual NOEs. For the tests shown one of the final NMR structure was subjected to 3000, 6000 and 9000 steps of 0.005 ps of unrestrained molecular dynamics simulation to obtain increasingly disordered structures (MD3000, MD6000 and MD9000 structure). Figure adapted from Ref. [323].

experimental residual dipolar couplings with residual dipolar couplings predicted from a structure [324]. Also a cross-validated free *R*-factor can be calculated based on residual dipolar couplings, where this *R*-factor is calculated for a subset of dipolar couplings that were not included in the refinement.

Other methods for the validation of structures. In the following some other computational methods for the analysis of protein NMR structures are summarized. One of the major problems in NMR structure determination is the identification of bad restraints. The program AQUA [325] automatically calculates restraint violations between the ensemble of structures and experimental data. Violations are classified, for example, as violations per model, per residue, and per restraint. Another approach for restraint analysis was pursued with the genetic FINGAR algorithm that was developed originally for NMR-based structural refinement [326]. Here, it is extended to allow the automated identification of problem restraints [327]. It is applicable to distance and dihedral angle restraints.

Complete cross validation is a statistical method to determine the quality of a structure [317]. In this method randomly selected subsets of the experimental data, e.g. NOEs are excluded, while a model is fitted against the remaining data. The resulting models are then used to predict values for the excluded sets and the agreement between predicted and experimental data can be evaluated as a measure for the quality of a structure.

The completeness of observed NOEs can be used as a measure for the quality of an NMR structure [309]. Completeness is measured as the ratio of the number of experimentally observed NOEs, NOE_{obs} and the number of NOEs expected for this structure, NOE_{exp}.

$$\text{Completeness} = 100\% \frac{NOE_{obs}}{NOE_{exp}} \tag{2.42}$$

It is shown that the completeness is independent of the residue type, contrary to the number of restraints per residue, which makes it easier to detect problematic regions. It is advisable to use various distance cutoffs when calculating the expected NOEs, to adjust for the maximum observable distance in a spectrum. The measure of completeness is in some regards related to the non-assigned NOEs used in the RFAC program.

Other methods for the quality determination of protein structures are independent of the experimental data. Instead general structural parameters such as bond lengths, bond angles, packing quality, etc. are analyzed [328]. The program PROCHECK-NMR [325] offers various tools to investigate the 'stereochemical quality' of an ensemble of structures. For example, quantities like bond-angles, bond-lengths, and backbone and side-chain dihedral angles are taken into account. In addition probability values are given for certain conformations. The program WHAT_CHECK [329] is part of the larger WHAT_IF package. In addition to standard quantities like bond-angles, bond-lengths, etc. other properties such as the packing quality of the trial molecule and hydrogen bonding networks can also be analyzed.

Using a set of previously solved three-dimensional structures one can construct a force field consisting of potentials of mean force. In this way the energy potentials for the atomic interactions between the various amino acid pairs are derived as a function of the distance between the involved atoms. Employing such a force field one can compute energy graphs for a given structure to identify problematic regions as is done within PROSA II [330]. High energies correspond to stressed or strained sections of the chain. The similarity to the approach used for the automated NOE assignment with KNOWNOE [233] can be noted.

2.2.6. Data deposition

A constantly increasing amount of spectroscopic and structural information on proteins is published. To access all the data, several databases have been published. For structural data the PDB database [331] and for spectroscopic data, e.g. chemical shifts, coupling constants, and relaxation data the BioMagRes [216] database are the most important ones. Currently, developments are underway to allow an automatic deposition in these data bases see, e.g. the CCPN project [http://www.bio.cam.ac.uk/nmr/ccp/]. Also it has become clear from the previous sections that a wide variety of programs exist that perform various tasks of the structure determination process. To allow an easy data exchange between the various programs a common data model is required, that is also currently being developed in the CCPN project.

Several programs exist to automatically access the different databases, e.g. SHIFTY [220] to search for chemical shifts of homologous proteins in the BioMagResBank and the BBReader [332] to search the BioMagResBank databank in an inverse manner. That means for given chemical shift values the program searches the databank for matching possible assignments.

Also local database programs exist to organize NMR related data in the laboratory. SPINS [333] is a local relational database to manage raw NMR data measured in one institute. It also creates header file information required for data deposition in the BioMagResbank.

3. Classical bottom-up computer-aided structure determination

In the second part of the review we will concentrate on integrated approaches for NMR structure determination and describe a new approach for macromolecular structure determination in solution (Section 2.2). The automated assignment that is a key step of any automated structure determination process would be trivial if exactly one cross peak in a given spectrum corresponded to one *n*-tuple of

possible chemical shifts, and vice versa. This is not the case, since practical spectra are *incomplete* (peaks are missing), contain *artifacts* (peaks not arising from the protein under consideration), and show *overlap* (degeneracy of chemical shifts). In addition for the automated assignment of spectra containing structurally relevant information, e.g. NOESY spectra, the resonance frequencies of all relevant spins are usually not known (incomplete assignment), may be wrongly determined (false assignments) or may not fit exactly to the spectrum used (chemical shift variations). Also proteins can occur in different conformational states, a difficulty that makes the assignment process even more complicated. An optimal automated strategy must tolerate the above difficulties at least to some degree and lead to a stable solution in the presence of incomplete data. In fact, a complete assignment is not feasible in practice but the assignments must be sufficiently complete to obtain an accurate structure of the protein with enough redundancy to cross check the results. That is, typical bottom-up strategies make use of experiments with some degree of redundancy where ambiguous solutions can be checked for the overall consistency of the solution.

An advantage of bottom-up strategies is that many sub problems of the automated structure determination are already solved, and automated assignment procedures tackling specific problems have already been reported. In the following, the partial steps of the automated structure determination in solution which were analyzed in detail in the previous section will be carefully examined under the aspect of a complete automated data evaluation and structure determination (Fig. 8). We omit in this section the steps which are not NMR-specific, but nevertheless are required in the automated structure determination, such as target selection and protein expression (Table 2). We will concentrate rather on those aspects which are essential in software development.

3.1. General strategies

In the following we will discuss in some detail the different assignment strategies published under the aspect of a full automation of the assignment procedure. They have to be of course supplemented by the automated routines for data processing and peak recognition, for the routines for

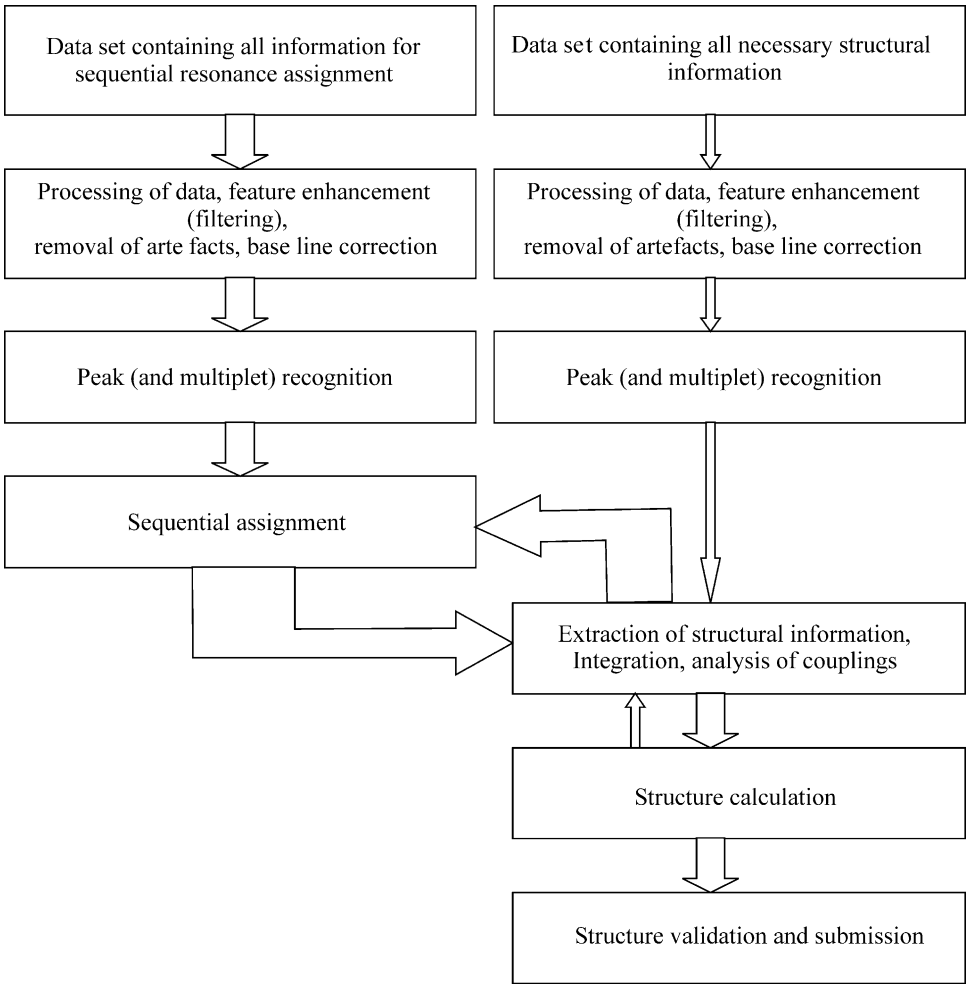

Fig. 8. Scheme for classical automated structure determination. The main emphasis is here on obtaining an almost complete resonance assignment as a basis for structure determination.

automated extraction of structurally relevant information as contained, for example, in ARIA, NOAH or KNOWNOE, and by the structure calculation itself (Fig. 8). Implicitly, the routines published also select for a specific set of NMR-experiments required. In automation studies one should state explicitly what experiments are optimal with respect to the convergence of the assignment procedure and the spectrometer time required.

(1) *Grouping of resonances from one or more spectra to spin systems.* The resonance lines which are part of specific spin systems must be identified at some point in the assignment procedure. Procedures used for this assignment process can be modeled on strategies applied in manual evaluations. The reliability of the assignment procedure is increased greatly when several different, complementary NMR spectra are evaluated together. In general two different approaches are used for spin system generation. In the first resonances are grouped via common 'root' resonances found in all or most of the spectra, e.g. 3D and 4D heteronuclear triple resonance spectra [334–345]. Often N–H pairs are used for this purpose that can easily be identified in 2D $^{1}H-^{15}N$ HSQC spectra. SPI [344] uses Bayesian statistics for compiling spin systems from 2D and 3D experiments and ambiguities are solved by using sets of diverse 2D and 3D experiments. In the program by Lukin et al. [339], Bayesian statistics are used to assemble signals originating from different experiments to partial spin systems. The so-called 'root resonance' approach has been shown to be sufficiently robust for the analysis of large proteins [345].

The second method, and the one applied in the earliest attempts at automatic evaluations, is to search the peak (multiplet) list for cross peaks expected for the primary building blocks of the biopolymer (e.g. amino acids in the case of proteins) in a particular NMR experiment. These methods are usually referred to as bond pattern methods, where for each residue a pattern is generated [210, 346–352]. As supplementary information, the expected chemical shift ranges can also be taken into account. With this information, which reduces the size of the data set to be searched and, hence, reduces the number of possible solutions, spectra of peptides and small to medium sized proteins can be analyzed automatically. Alternatively, the known probability distributions of the individual chemical shifts can be used [216,218,221,353] to assess the probability associated with an identified pattern [210,354]. It should be noted that bond pattern methods are sensitive to missing signals and spectral overlap, therefore they are best suited for the analysis of peptides and small proteins.

The method described by Oschkinat et al. [346] aims at the analysis of homonuclear 3D spectra. Graphs representing spin systems are generated from 3D TOCSY–TOCSY spectra. In the program from Croft et al. [350] the Fourier transformed 3D triple resonance spectra are convoluted with a mask function to emphasize real peaks and to de-emphasize features like noise and artifacts before the following analysis. In the exhaustive pattern search, multiple spectra are simultaneously searched against a list of predefined patterns, and matches exceeding a certain threshold value are stored for further analysis. A set of rules is applied to evaluate the results of the previous step. The results from this approach can then be used in a subsequent manual or automated sequential assignment procedure. The program GARANT [347,348] represents spin systems as graphs where peaks are connected by common resonances.

(2) *Association of spin systems with amino acid types.* One of the next steps in the assignment procedure is usually to assign amino acid types to the previously identified spin systems. It can be performed before or after spin systems are linked to longer fragments. Again different general approaches exist for this task. In the first, spin systems are matched to bond pattern templates that are typical for the various residue types. Since this method is in general very sensitive to missing data it is usually only applicable for small proteins for which almost complete, well-resolved spectra can be obtained. Improved results can be obtained when this method is supplemented with chemical shift information [349–351,355–357].

Within CAMRA [351] improved results can be obtained by the inclusion of predicted chemical shifts. ^{1}H, ^{15}N, and ^{13}C chemical shifts are predicted for the query protein using the module ORB [217]. Predictions are based on a chemical shift database of previously assigned homologous proteins supplemented by a statistically derived chemical shift database in which the shifts are categorized according to their residue, atom and secondary structure type. In the previous step spin systems were generated with CAPTURE [352]. This module generates a list of valid peaks from NMR spectra by filtering out noise peaks and other artifacts and then separates the derived peak list into distinct spin systems. PROCESS generates a set of predicted peaks using the predictions from ORB and the experiment type. The predicted signals are then matched to the spin systems obtained from CAPTURE.

Another possibility for assigning residue types is to use neural networks. Neural networks can be trained to recognize the amino acid type from 2D and/or 3D TOCSY data. In the approach of Hare and Prestegard [358] sets of cross peaks belonging to a single spin system were presented to the neural network one set at a time. The method was refined by additionally including ^{15}N chemical shift data and $^{3}J_{HNH\alpha}$ coupling constants [359]. This allows a reliable prediction of the secondary structure, and since the Hα proton chemical shifts are highly sensitive to the secondary structure, a higher success rate in spin system identification was achieved.

Similarly in the RESCUE approach [360] only signals originating from 2D TOCSY spectra are used. In the approach by Choy et al. [361], neural networks are used to predict the secondary structure of homologous proteins to obtain secondary chemical shift information. This information is then used to improve the chemical shift

and bond-pattern based amino acid recognition in automated resonance assignment procedures. The next common method that is mostly used when the assignment is based upon heteronuclear triple resonance experiments makes use of a statistical chemical shift analysis. Cα and Cβ chemical shifts especially are used in combination for amino acid identification [338,341,342,362].

The program from Meadows et al. [363] uses proton and carbon chemical shifts for amino acid identification. In the programs by Lukin et al. [339] and Zimmermann et al. [340] amino acid type information is given in a probabilistic fashion using Bayesian statistics. The program by Andrec and Levy [364] uses only Cα shifts since it is based on a single HNCA experiment. PLATON [365] is an algorithm for predicting amino acid types and secondary structure types of individual amino acids from chemical shifts. Residue fragments are manually generated using standard triple resonance experiments. The chemical shifts (e.g. Cα, Cβ and C′) of these fragments are then automatically compared to a chemical shift database generated from 51 proteins for which chemical shift and three-dimensional structure information is available. MONTE [345] uses C′, Cα, Cβ, Cγ, Hα, Hβ, and N chemical shift distributions obtained from the BioMagResBank for amino acid type identification. If secondary structural information is available the expected chemical shift values can be adjusted accordingly in case of deuterated proteins carbon chemical shifts are adjusted to take deuterium isotope effects into account.

IBIS [335] also uses backbone and side-chain chemical shift information obtained from the BioMagResBank for amino acid type identification given in a probabilistic fashion. In addition it takes into account whether or not the number of observed side-chain chemical shifts matches the number of corresponding expected chemical shifts.

In PACES [336] Cα, Cβ and C′ chemical shifts are compared to chemical shift distributions derived from the BioMagResBank for amino acid type identification. Within the given chemical shift ranges all of the possible amino acid types for each spin system are recorded.

(2) *Linking of spin systems to shorter or longer fragments*. The next step usually performed by most programs is to link sequential neighboring spin systems to shorter or longer fragments. The linking is either based on NOEs, or on through bond information, or on a combination of both. Using this information basically two algorithms are used for performing the linking step. Deterministic methods require a full comparison between each spin system. For larger systems the required computational time can increase drastically [30,334–336,338–340,342,357,363,364,366,367]. The size of the problem can be reduced by establishing the most reliable links first (best first methods) and by keeping these fixed in the later stages, for example, by creation of small peptides in the first steps that are extended later.

In the other set of methods a pseudo energy is minimized by, for example, a simulated annealing procedure [334,341,345,368–370]. The pseudo energy herein describes the quality of the match between neighboring spin systems. These energy optimization algorithms are applicable to larger systems, however, care must been taken that they do not become trapped in local minima.

If only NOE information is available, as for example in the homonuclear case, intuition and statistical analysis of protein structures tell us that the protons of direct linked neighboring residues in a sequence have a much higher probability of being close to each other than protons of residues which are greatly separated in the biopolymer sequence. Although this information is not sufficient to define a unique path from spin system to spin system, the correct path can be selected by comparing the NMR results with the known amino acid sequence [176,368,371]. The main problem with this strategy is that the NOEs only contain distance information and provide no information concerning covalent bonds. Therefore, because of protein folding, some of the ‘sequential’ NOEs arise from residues which are only close in space but not close to each other in the chain sequence.

The number of ambiguous solutions is reduced by increasing redundancy in the NOE assignment procedure. This can be achieved by considering a whole network of possible intra-residue and inter-residue NOEs. This leads to the already mentioned MCD strategy that tries in its pure form to order the spin systems in a sequence without resorting to the inclusion of information about the amino acid type [212–214]. Strategies which also incorporate readily available information about the type of spin systems expected for a given macromolecule are more robust [212–214,372,373].

In the approach developed by Oschkinat et al., homonuclear 3D TOCSY–NOESY spectra are used to connect graphs that represent spin systems obtained from 3D TOCSY–TOCSY spectra [346]. A deterministic approach is used in which first di-peptides are generated that are then extended to longer fragments. The FIRE [343] procedure uses as input only ^{15}N edited 3D NOESY–HSQC spectra. The program basically creates strips for observed pairs of ^{15}N and HN frequencies. These strips representing spin systems are then compared for signals at common positions to obtain the spatial proximity between spin systems. Spin systems showing the smallest spatial distance are thought to be sequentially linked.

Most of the approaches that make use of inter-residue through bond information are based on the use of heteronuclear triple resonance 3D and 4D experiments. They map intra- to inter-residual Cα, Cβ and CO chemical shifts for linking. The key advantage compared to NOEs is that only sequential information is present. The method is independent of secondary and tertiary structure effects and therefore, much more reliable than pure NOE based methods. Another advantage is that the ^{1}J and ^{2}J scalar couplings used are usually rather large and as a consequence give fast coherence transfer that can compete with relaxation losses

even for larger molecules. Improved results (signal-to-noise) can be obtained by employing relaxation optimized techniques, such as TROSY and CRINEPT or by the use of deuterated molecules. Almost all approaches that are applicable to larger proteins, e.g. MONTE [345], PASTA [341], AUTOASSIGN [340], etc. make use of inter-residue through bond information.

(4) *Mapping of fragments to the primary sequence.* The last step in the sequential assignment procedure is usually to map single or stretches of spin systems to the primary sequence. In addition to the sequential NOE and/or through bond information that was used for linking, residue type information obtained in a previous step is also included for mapping. Basically, the same algorithms that were used for linking are also used for mapping. Fully exhaustive searches, e.g. PACES [336] may become prohibitive for larger proteins therefore, in deterministic methods so-called best first approaches are often used to reduce the size of the problem [334,335,338,340,342,349,351,362–364,366,367,374].

Often sets of rules are included in these methods to improve performance, e.g. redundant mappings are excluded where one fragment is a subset of another fragment or mappings must exceed a certain threshold value to be accepted. However, it should be noted if the data become more ambiguous and/or data are missing best first approaches will probably fail since too many possibilities have to be considered. Alternatives are given by optimization procedures where a pseudo-energy is minimized, e.g. simulated annealing like algorithms [334,337,339,341,345,348,367–370].

One problem encountered in energy optimization procedures are not very smooth energy surfaces caused by discrete changes in fragment mapping. In the approach by Buchler et al. [369], the energy surface is smoothened by considering a large ensemble of assignments. Also genetic algorithms as used within GARANT [347] usually work well with rough energy surfaces. Here, sets of experimental signals are mapped to the corresponding expected peaks. The assignment problem can be simplified when larger fragments are first linked in a deterministic way and the simulated annealing is used for fragment mapping see, e.g. Lukin et al. [339].

Since most of the programs perform more than one of the above-described steps a short overview is given about the capabilities of the various programs. Which program is applicable to a certain problem strongly depends on the available experimental input, the size of the molecule, available homologous chemical shift and/or 3D structural information and the required task.

A deterministic best first method using six 3D and 4D triple resonance experiments to obtain backbone and partial side-chain assignments has been described by Friedrichs et al. [338]. In this approach, which is integrated within FELIX the backbone and Cβ resonances are grouped into partial spin systems. These partial spin systems are then linked using a best first approach based on matches between inter- and intra-residual Hα, Cα and Cβ resonances. Using an exhaustive search procedure the linked fragments are mapped to the primary sequence using residue dependent chemical shift profiles for Cα and Cβ resonances.

A method that relies on the use of homonuclear 3D TOCSY–TOCSY and 3D TOCSY–NOESY spectra is applicable for small- and medium-sized proteins in cases where no ^{15}N and/or ^{13}C labeled protein is available [346,375]. Graphs representing spin systems are generated from the 3D TOCSY–TOCSY spectra. 3D TOCSY–NOESY spectra are used to first link the spin systems to dipeptides which are then combined to longer fragments. Several rules are applied to evaluate the fragments. The mapping of the fragments to the primary sequence is performed manually. It is also shown that the amplitudes in 3D TOCSY–NOESY spectra are characteristic for the various secondary structure elements [376].

A combinatorial oriented automated assignment strategy using a deterministic tree search algorithm has been reported which is based upon only a single triple resonance HNCA experiment [364]. It is aimed at the automated backbone assignment of smaller proteins (<80 residues).

TATAPRO applies an approach where sets of chemical shifts derived from heteronuclear triple resonance assignments are grouped into eight categories according to their chemical shifts. This information is then used together with connectivity information obtained from the triple resonance experiments to search with a deterministic approach for neighboring partners in the primary sequence [342]. As a result backbone and partial (Cβ) side-chain assignments are obtained.

A suite of three programs called CPA, FPRA, and TSA is presented for the automated sequential assignment using 2D proton NMR spectra [366]. CPA automatically generates spin-coupling networks from 2D COSY and TOCSY spectra. The obtained spin systems are then mapped by FPRA using a fuzzy graph pattern recognition algorithm to the various residue types. In this algorithm spin coupling patterns and chemical shifts are considered. A tree search algorithm then assigns the spin systems to the primary sequence using NOE connectivities. The following factors are taken into account (1) maximum number of NOE correlations between neighboring residues, (2) matching of an experimental supergraph to the corresponding theoretical one that was obtained from the primary sequence. A supergraph contains the spin coupling networks of the participating residues plus the corresponding inter-residual NOE connectivities. (3) Using the supergraph that has the maximum number of frequencies assigned. The described algorithm was later refined and extended to the use of non-standard amino acids [377]. It should be noted that the algorithm works fully automated only for relatively small peptides of up to 20 amino acids.

Li and Sanctuary [349,357] present a set of algorithms (DBPA/ASPA/CPA) that first create from a set of 3D

heteronuclear triple resonance experiments a list of dipeptides using a best first approach. Next dipeptides are linked to form polypeptides. Then side-chain information is automatically added to the dipeptides (PBSMA). Amino acid types are identified with a pattern recognition algorithm (AAPR) [355,356] which makes use of spin coupling patterns and chemical shifts. In the last sequential assignment step the polypeptides generated in the previous steps are mapped to the primary sequence (PMA).

The CONTRAST software package [367] for the automated assignment of backbone resonances assembles fragments from a set of 3D heteronuclear triple resonance and TOCSY–HMQC experiments. In contrast to other programs a deterministic best first approach or a simulated annealing like method can be used for fragment mapping. As an example the method is schematically presented in more detail in Fig. 9.

In its published version AutoAssign [340,378,379] can use input from up to eight 3D heteronuclear triple resonance experiments for the automated sequential backbone assignment. From these spectra generic spin systems are generated and residue type probability scores are calculated based on Cα and Cβ shifts. Using a best first approach the spin systems are linked and using the residue type probability scores they are mapped to the primary sequence.

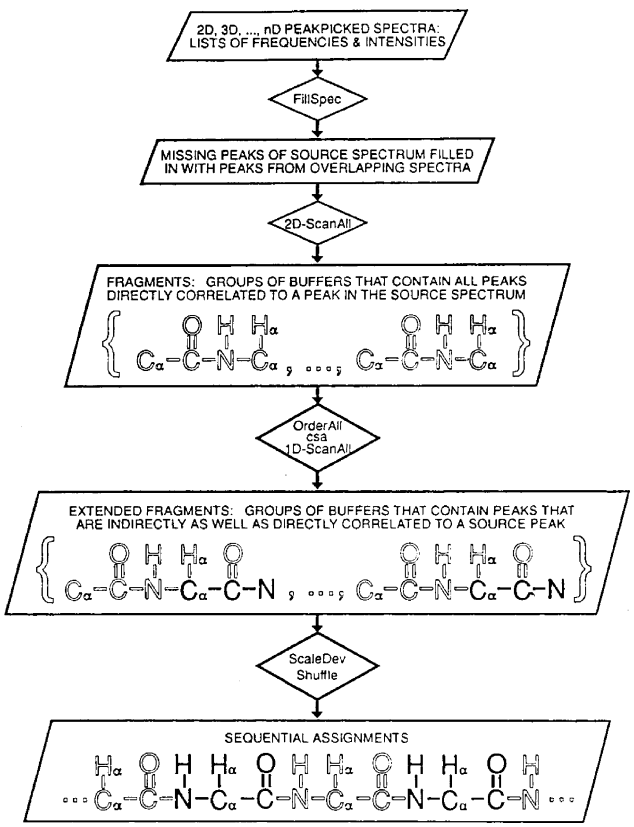

Fig. 9. Flowchart of the SBA (sequential backbone assignment) algorithm as implemented with the CONTRAST software package. The algorithm uses a HNCO source spectrum and HNCA, HN(CO)CA, HCACO, HCA(CO)N and TOCSY–HMQC spectral data sets to generate and order the fragments. Figure adapted from Ref. [367].

In IBIS [335] sequential assignments are obtained using a set of triple resonance spectra supplemented with a (H)CCONH TOCSY for providing side-chain information. The program uses a deterministic approach for linking of spin-systems to fragments and for mapping the fragments to the primary sequence that resembles a manual analysis process. The quality of a certain assignment is evaluated in a probabilistic fashion. The program is interfaced with XEASY for graphical display of the results.

PACES [336] obtains resonance line assignments based on the sequential connectivity and residue type information derived from triple resonance spectra. Especially, Cα, Cβ, C′ and Hα resonances are used by the program. Amino acid type information is generated using amino acid specific chemical shift distributions derived from the BioMagResBank. Exhaustive search procedures are used for linking of residues to fragments and for mapping of fragments to the primary sequence.

In the next approach sequential assignments are obtained using 10 different 3D heteronuclear triple resonance experiments [339] that give intra- and inter-residual information. The likelihood that peaks from overlapping experiments originate from the same nuclei is evaluated using Bayes' theorem. Bayes' theorem is also used to link cross-peak chemical shifts to positions in the primary sequence. These two probability factors are combined into an overall score for a tentative peak assignment. A best first approach is used to link spin systems to fragments. Using a simulated annealing protocol the overall score is maximized by rearranging the assignment.

In the MARS approach (Zweckstetter et al., to be published), global energy minimization and deterministic search (best first) for linking of spin systems to fragments are combined to extract reliable assignments. In this process only consistent results are accepted. In the next step theoretical chemical shift values used for placing linked fragments into the sequence are optimized using sequence based secondary structure predictions. Multiple assignment runs (~100) are performed and in each assignment run the, by secondary structure prediction optimized, theoretical shifts are disturbed by addition of Gaussian noise. Again only consistent assignments are used.

PASTA [341] performs automatic backbone and partial side-chain (Cβ) assignments. As input any experiment containing only HN, N, Hα, Cα, Cβ, and C′ resonances can be used. From a peak list obtained from, e.g. a 2D $^{1}H-^{15}N$ HSQC spectrum a set of so-called pseudo-residues is constructed that is filled with the information obtained from the other experiments. For linking and mapping, these pseudo-residues to the primary protein sequence a threshold accepting algorithm is used within PASTA. It is similar to a simulated annealing approach, but has the advantage of converging significantly faster.

A sequential assignment approach from Buchler et al. [369] uses input from six three-dimensional, mostly triple resonance experiments. Spin systems are manually identified from ^{13}C edited 3D TOCSY experiments and are grouped into different spin system categories. Connectivity information between neighboring spin systems is obtained from 3D triple resonance experiments. A mean-field simulated annealing approach was used to simultaneously link the spin systems and to match them to the primary sequence. Resonance assignments are reported in a probabilistic fashion.

The program st2nmr [370] is a semi-automatic approach for the sequential assignment. It is based on the use of expected NOEs predicted from a known 3D structure. The user has to present spin systems obtained from 2D and/or 3D correlation spectra (COSY and TOCSY) with assigned residue types to the program. A Monte Carlo Algorithm is used to map the spin systems to the primary structure, and the match between expected and experimental NOEs is maximized.

An automated Monte Carlo [345] (MONTE) approach has been developed for resonance assignment of large proteins. The program is quite flexible since it is not dependent on a particular set of experiments. Input can contain intra- and inter-residue chemical shifts obtained from *J*-correlated experiments, NOE connectivity information as well as residue type information obtained from residue specific labeling. In a semi-automatic way all the information associated with one N–H pair is compiled into one so-called spin system. A flexible scoring function is provided that includes, for example, the match between intra- and inter-residual chemical shifts of two in the trial assignment adjacent residues. The Monte Carlo approach is then used to improve a randomly generated start sequential assignment by, for example, swapping one or more consecutive spin systems. In this process the simulated annealing schedule used is defined by the user.

A Monte Carlo approach has been used in the program ASSTOOL [337]. In this approach spin-systems generated from a set of triple resonance spectra are presented to the algorithm. The energy function that is minimized contains a term that describes the match between intra- and inter-residual chemical shifts of two adjacent residues in the trial assignment, and a term is included that describes how well the chemical shifts of the spin-systems match those expected for the residue types to which they are assigned.

CAMRA [351] is a suite of programs for computer assisted residue-specific assignments of proteins. CAMRA consists of three units: ORB, CAPTURE and PROCESS. ORB predicts NMR chemical shifts for unassigned proteins using a chemical shift database of previously assigned homologous proteins supplemented by a statistically derived chemical shift database in which the shifts are categorized according to their residue, atom and secondary structure type. CAPTURE [352] generates a list of valid peaks from NMR spectra by filtering out noise peaks and other artifacts and then separating the derived peak list into distinct spin systems. PROCESS combines the chemical shift predictions from ORB with the spin systems from CAPTURE to obtain residue specific assignments.

GARANT [347,348,380] performs automatic complete resonance assignment for proteins. It is based on the mapping of peaks predicted from the amino acid sequence onto observed experimental signals obtained from two- and three-dimensional spectra. For the mapping process genetic algorithms in combination with a local optimization routine are used. If available, the method is supplemented with homologous structures and/or chemical shifts. Recently, GARANT has been combined with the program AUTOPSY (described above) and a new program PICS to facilitate fully automated resonance line assignments [381]. In this approach the peaks originating from different spectra are automatically identified by AUTOPSY. The program PICS performs calibration and filtering of the resulting peak lists and GARANT the actual resonance line assignment.

Hus et al. [362] have developed an automated assignment procedure for proteins of known three-dimensional structure. It makes use of residual dipolar couplings and the corresponding Cα and Cβ chemical shifts. A combinatorial procedure rearranges the assignment to minimize the difference between measured values and expected chemical shifts and residual dipolar couplings predicted from the structure. The expected chemical shifts are taken to be averaged values from the BioMagResbank. No sequential NMR connectivity information is needed.

Meadows et al. [363] have introduced various tools to partially automate the resonance and NOE assignment process. The approach is mainly based on the use of three- and four-dimensional heteronuclear experiments. For the sequential assignments a deterministic best first approach is used. A structure-based filter can be applied to resolve ambiguities in the NOE assignment when a preliminary structure is available.

The program MAPPER [374] maps manually assigned short stretches of sequentially connected residues to the proteins primary sequence using partial knowledge of the amino acid types. Amino acid type information is either automatically determined from $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts or is provided by the user. Fragments are first individually mapped to the proteins primary sequence and then, based on the accepted individual mappings, an exhaustive search for the self-consistent mapping of all fragments is performed.

3.2. Programs and program packages

In the following a short overview is given about some of the programs in use for manual and/or semi-automatic spectra assignment (Table 3).

EASY [176] and XEASY allow the semi-automatic assignment of homo- and heteronuclear 2D and 3D spectra. Various tools for display, spectra manipulation, automated

Table 3
Available programs for computer aided spectral assignment

Program	Properties and functions	Main algorithms	Source
Data processing and visualization			
AZARA	Multidimensional NMR data processing	Fourier transformation	http://www.bio.cam.ac.uk/azara/
DELTA	Multidimensional NMR data processing	Fourier transformation	Peabody MA, Jeol USA Inc.
NMRLAB	Multidimensional NMR data processing	Wavelet transformation	http://www.bpc.uni-frankfurt.de/~nmrlab/index1.html
NMRPipe	Multidimensional NMR data processing	Fourier transformation	http://spin.niddk.nih.gov/bax/software/NMRPipe/
NMR Toolkit	Multidimensional NMR data processing	Fourier transformation	http://www.rowland.org/rnmrtk/toolkit.html
PROSA	Multidimensional NMR data processing	Fourier transformation	http://www.guentert.com/
TRITON	Multidimensional NMR data processing	Fourier transformation	http://www.nmr.chem.uu.nl/software/
VNMR	Multidimensional NMR data processing	Fourier transformation	Palo Alto, CA, Varian Inc.
XWINNMR	Multidimensional NMR data processing	Fourier transformation	Ettlingen, Bruker Biospin GmbH
Peak and multiplet recognition			
ATNOS	NOESY peak picking	Combination with structure calculation	http://www.mol.biol.ethz.ch/wuthrich/software/
AUTOPSY	Peak picking	Local noise level calculation	http://www.mol.biol.ethz.ch/wuthrich/software/autopsy/
Sequential assignment			
Andrec and Levy	Backbone assignment	Tree search	Available from the authors [364]
ASSTOOL	Backbone and partial side-chain assignment	Monte Carlo	Available from the authors [337]
AUTOASSIGN	Backbone and partial side-chain assignment	Best first approach	http://www-nmr.cabm.rutgers.edu/NMRsoftware/nmr_software.html
CAMRA	Complete sequential assignment	Use of predicted homologous shifts	http://www.pence.ualberta.ca/software/camra/latest/camra.html
CONTRAST	Backbone assignment	Best first or simulated annealing approach	http://www.specres.com/contrast.asp
CPA/FPRA/TSA	Complete sequential assignment	Tree search algorithm	Available from the authors [355]
DBPA/PGA/ASPA/NCPA/PBSMA/AAPR/PMA	Complete sequential assignment	Best first approach	Available from the authors [349]
FELIX/MACROS	Backbone and partial side-chain assignment	Exhaustive search	For users of FELIX the macros are available from the authors [338]
FIRE	Sequential assignment and fold information	Use of pattern similarities	Included in GIFA
GARANT	Complete sequential assignment	Genetic algorithm	http://www.mol.biol.ethz.ch/wuthrich/software/garant/
IBIS	Backbone and partial side-chain assignment	Best first approach	http://gwagner.med.Harvard.edu/ibis/
Lukin et al.	Backbone and partial side-chain assignment	Simulated annealing	Available from the authors [339] http://www.cmu.edu/nmr-center/links.html
MAPPER	Fragment mapping on primary sequence	Exhaustive search algorithm	http://www.guentert.com/
MARS	Backbone assignment	Best first approach and Energy minimization	http://www.mpibpc.mpg.de/abteilungen/030/zweckstetter/_links/software.htm
MONTE	Complete sequential assignment	Monte Carlo	http://www.andrew.cmu.edu/~rule/monte/
PACES	Backbone and partial side-chain assignment	Exhaustive search algorithm	Available from the authors [336]
PASTA	Backbone and partial side-chain assignment	Threshold accepting	http://www.org.chemie.tu-muenchen.de/people/jl/shell_pasta02/pasta_doc.html
PLATON	Residue type prediction	Comparison with reference values	http://www.fmp-berlin.de/~labudde/platon.html

(*continued on next page*)

Table 3 (*continued*)

Program	Properties and functions	Main algorithms	Source
RESCUE	^{1}H assignments	Neural networks	http://www.infobiosud.cnrs.fr/SERVEUR/RESCUE/welcome.html
SPSCAN	Evaluation of reduced dimensionality experiments	Deterministic approach	http://www.molebio.uni-jena.de/~rwg/spscan/
ST2NMR	Mapping of spin systems	Monte Carlo	http://arg.cmm.ki.si/~primus/st2nmr/
TATAPRO	Backbone and partial side-chain assignment	Deterministic approach	http://tifrc1.tifr.res.in/~hsatreya/interest.html
Global structural information from NMR data			
ARIA	Automated NOE assignment	Use of ambiguous restraints	http://www.pasteur.fr/recherche/unites/Binfs/aria/
AUTOSTRUCTURE	Automated NOE assignment	Rule based expert system	http://www-nmr.cabm.rutgers.edu/NMRsoftware/nmr_software.html
CANDID	Automated NOE assignment	Filtering by violation analysis and use of ambiguous restraints	http://www.guentert.com/
KNOWNOE/AUREMOL	Automated NOE assignment	Knowledge based Bayesian algorithm	http://www.auremol.de/
SANE	Automated NOE assignment	Use of ambiguous restraints	http://garbanzo.scripps.edu/nmrgrp/
Back-calculation of nD-spectra			
BIRDER	Simulation of 2D NOESY spectra	Full relaxation matrix approach	Available from the authors [259]
CORMA	Simulation of 2D NOESY spectra	Full relaxation matrix approach	http://picasso.ucsf.edu/mardihome.html
CORCEMA	Interacting systems	Full relaxation matrix approach	Available from the authors [265]
DINOSAUR	Structure refinement	Full relaxation matrix approach	http://www.nmr.chem.uu.nl/software/
IRMA	Distance determination from NOESY spectra	Full relaxation matrix approach	http://www.nmr.chem.uu.nl/software/
HYDRONMR	Calculation of diffusion tensor	Bead model	http://leonardo.fcu.um.es/macromol/programs/hydronmr/hydronmr.htm
MARDIGRAS	Distance determination from NOESY spectra	Full relaxation matrix approach	http://picasso.ucsf.edu/mardihome.html
MIDGE	Distance determination from NOESY spectra	Full relaxation matrix approach	Available from the authors [255]
MORASS	Distance determination from NOESY spectra	Full relaxation matrix approach	http://www.nmr.utmb.edu/#mrass
NO2DI	Distance determination from NOESY spectra	Full relaxation matrix approach	Available from the authors [256]
RELAX/AUREMOL	Simulation of 2D and 3D NOESY spectra	Full relaxation matrix approach	http://www.auremol.de/
SPIRIT	Simulation of 3D NOESY–HSQC spectra	Full relaxation matrix approach	http://garbanzo.scripps.edu/nmrgrp/
Chemical shift prediction			
ORB	Prediction of ^{1}H, ^{15}N, and ^{13}C shifts	Use of homologous shifts	http://www.pence.ualberta.ca/software/orb/latest/orb.html
PROSHIFT	Prediction of ^{1}H, ^{15}N, and ^{13}C shifts	Neural networks	http://www.jens-meiler.de/proshift.html
SHIFTS	Prediction of ^{15}N, and ^{13}C shifts	Density functional calculations	http://www.scripps.edu/case/casegr-sh-2.3.html
SHIFTY	Prediction of ^{1}H, ^{15}N, and ^{13}C shifts	Use of homologous shifts	http://redpoll.pharmacy.ualberta.ca/shifty/
Various programs by Michael Williamson	Prediction of ^{1}H, and ^{13}C shifts	Semi-empirical	http://www.shef.ac.uk/uni/projects/nmr/resources.html
TANSO	Prediction of ^{13}Cα, and ^{13}Cβ shifts	Empirical database	http://www.tuat.ac.jp/~asakura/reserch/13c/
Local structural information from NMR data			

CSI	Secondary structure prediction	Empirical correlation with known shifts	http://www.pence.ualberta.ca/software/csi/latest/csi.html
FOUND	Stereospecific assignments	Grid search	http://www.guentert.com/
HYPER	Dihedral angles and stereospecific assignments	Multidimensional grid search	http://www-nmr.cabm.rutgers.edu/NMRsoftware/nmr_software.html
MULDER	Dihedral angles	One-dimensional grid search	http://ncbr.chemi.muni.cz/mulder/mulder.html#Intro
Kloiber et al.	Backbone dihedral angels from cross-correlated spin relaxation	One-dimensional grid search	Available from the authors [281]
POP	Prediction of peptide bond conformation	Empirical correlation with known shifts	http://www.fmp-berlin.de/~labudde/pop.html
PSICSI	Secondary structure prediction	Neural networks	http://protinfo.compbio.washington.edu
PSSI	Secondary structure prediction	Empirical correlation with known shifts	http://ccsr3150-p3.stanford.edu/
TALOS	Prediction of backbone dihedral angles	Correlation with known structural elements	Available from the authors http://spin.niddk.nih.gov/bax/
Structure calculation			
AMBER	3D structure calculation	Molecular dynamics	http://www.amber.ucsf.edu/amber/amber.html
CNS	3D structure calculation	Molecular dynamics	http://cns.csb.yale.edu/v1.1/
CYANA	3D structure calculation	Molecular dynamics in torsion angle space	http://www.guentert.com/
DL_POLY	3D structure calculation	Parallel molecular dynamics	http://www.cse.clrc.ac.uk/msi/software/DL_POLY/
DTAGS/NEWMOL	3D structure calculation	Grid search in torsion angle space	http://www.fkem2.lth.se/~garry/programs.html
FANTOM	3D structure calculation	Newton–Raphson torsion angle minimization	http://www.scsb.utmb.edu/fantom/fm_home.html
GROMOS	3D structure calculation	Molecular dynamics	http://www.igc.ethz.ch/gromos/
INSIGHT II-CHARMM	3D structure calculation	Molecular dynamics	San Diego, CA, Acclrys Inc.
NAMD2	3D structure calculation	Parallel molecular dynamics	http://www.ks.uiuc.edu/Research/namd/
XPLOR-NIH	3D structure calculation	Molecular dynamics	http://nmr.cit.nih.gov/xplor-nih/
Structure determination using sparse NMR data			
MFR	3D Structure determination	Use of fragments from known structures	Available from the authors [300]
ROSETTA	3D Modeling	Use of fragments from known structures	http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php and from dabaker@u.washington.edu
Structure validation			
AQUA	Quality of NMR structures	Restraint analysis	http://www.nmr.chem.uu.nl/software/
PROCHECK_NMR	Stereochemical quality of NMR structures	Comparison with standard values	http://www.biochem.ucl.ac.uk/~roman/procheck_nmr/procheck_nmr.html
PROSAII	Energy calculation	Use of potentials of mean force	http://lore.came.sbg.ac.at:8080/CAME/CAME_EXTERN/ProsaII/index_html
RFAC/AUREMOL	NMR *R*-factor	Use of non-assigned signals	http://www.auremol.de/
WHAT_IF/WHAT_CHECK	Stereochemical quality of NMR structures	Comparison with standard values	http://www.cmbi.kun.nl/whatif/
Multipurpose general data evaluation			
ANSIG	Data analysis	Various	http://www-ccmr-nmr.bioc.cam.ac.uk/public/ANSIG/ansig.html
AURELIA	Data analysis	Various	Ettlingen, Bruker Biospin GmbH
AUREMOL	Data analysis	Various	http://www.auremol.de/ and Ettlingen, Bruker Biospin GmbH
FELIX	Processing and data analysis	Various, Fourier transformation	San Diego, CA, Acclrys Inc.

(*continued on next page*)

Table 3 (*continued*)

Program	Properties and functions	Main algorithms	Source
GIFA	Processing and data analysis	Various, Fourier transformation	http://www.cbs.cnrs.fr/GIFA/welcome.html
NMRVIEW	Data analysis	Various	http://www.nmrview.com/
PIPP	Data analysis	Inclusion of line shape information	dgarrett@speck.niddk.nih.gov
PRONTO	Data analysis	Various	http://mail.crc.dk/chem/pronto/welcome.html
SPARKY	Data analysis	Various	http://www.cgl.ucsf.edu/home/sparky/
TRIAD	Processing and data analysis	Various, Fourier transformation	St Louis, MO, Tripos Inc.
XEASY	Data analysis	Various	http://www.mol.biol.ethz.ch/wuthrich/software/
Data storage			
BB-READER	Search of BMRB	Inverse search	http://bmrb.protein.osaka-u.ac.jp/bbreader/BBReader.html
BMRB	NMR data deposition	Relational database	http://www.bmrb.wisc.edu/
CCPN	Data exchange	Various	http://www.bio.cam.ac.uk/nmr/ccp
PDB	Structure deposition	Relational database	http://www.rcsb.org/pdb/
SPINS	NMR data archival	Relational database	http://www-nmr.cabm.rutgers.edu/NMRsoftware/nmr_software.html

and manual peak picking, peak integration, spin system identification from *J*-correlated spectra, sequential assignment based on NOE connectivities, etc. are included.

AURELIA [264] allows the computer-aided analysis of up to four-dimensional NMR spectra. In addition to various display routines AURELIA offers various fully and semi-automated tools for baseline correction, artifact reduction, peak picking, cluster and multiplet analysis, spin system searches, resonance assignments, volume integration, signal and artifact discrimination based on Bayesian statistics, simulation of NOESY spectra employing the full relaxation matrix approach, correlation of structural and NMR data, comparison of spectra via spectra algebra and pattern match techniques. An automated technique for sequential assignments based on 3D triple resonance HNCA spectra is included. To overcome ambiguities in the assignment process, a partial knowledge of the spin system types obtained manually from, e.g. HCCH–TOCSY spectra can be used. Fragments of several residues with partially assigned spin system types are then matched in an exhaustive procedure to the primary sequence and scored according to their fit to the primary sequence.

ANSIG for Windows [382], a new version of the original ANSIG program [383] is a program for the display of up to four-dimensional NMR spectra and it contains tools for semi-automatic sequential assignment. In addition it includes tools for plotting distances from PDB structure files directly in the NMR spectra and it allows the statistical analysis of restraint violations.

Tools for NMR assignment are incorporated in the GIFA NMR processing program [384]. Besides displaying up to three-dimensional spectra several tools are incorporated in GIFA that help the user in the manual assignment process. The main aim is to help in the bookkeeping process during the resonance and NOE assignment.

The program PRONTO [50] is in some respect similar to ANSIG. It also allows the display of up to four-dimensional NMR spectra. Then it includes databases to store information regarding the NMR spectra and their signals and about the protein under investigation. Tools for semi-automatic assignment such as a ‘Spins System Buildup Facility’ and a ‘Sequential Assignment Tool’ are included.

SPARKY [385] is designed for the display, assignment and integration of up to four-dimensional NMR spectra. Assignments can be either manually performed or an automated heteronuclear backbone assignment is possible by using SPARKY as a graphical interface for AUTOASSIGN.

NMRVIEW [386] allows the display and manual assignment of up to four-dimensional NMR spectra in multiple windows. It includes tools for an automated peak picking, spin-system tabulation, creation and analysis of NOE constraints, and structure analysis including constraint violation analysis. Macros can be written in the Tcl command language.

The PIPP [180] package contains several programs PIPP, CAPP, STAPP, PS_CONTOUR, and CONTOUR_SIM, for the manual and semi-automated analysis of up to four-dimensional spectra. CAPP is an automatic peak picker that has been described above. The purpose of STAPP is to perform a semi-automatic NOE-assignment by making use of the sequential assignment and a trial structure. PIPP is then used to visualize the results from the previous two programs. PS_CONTOUR and CONTOUR_SIM are auxiliary programs.

Besides spectral processing FELIX [45] offers various tools for display of up to four-dimensional spectra, capabilities for manual assignment and automated procedures for spin-system identification, sequential assignment and NOE identification.

Another commercial package TRIAD [52] also combines spectra processing and analysis capabilities in one package. It is possible to display and analyze nD spectra. Features include automatic peak picking and the integration of 3D structural data in the NOE assignment process. TRIAD can work in conjunction with MARDIGRAS for obtaining distance information from NOE data and with DYANA for structure determination.

4. Automated top-down NMR-structure determination

Bearing in mind that in many applications of multi-dimensional NMR-spectroscopy, the main aim is not a completely correct spectral assignment but a correct three-dimensional structure then we have to ask for an optimal strategy to obtain this structure with a minimum of experiments in an automated fashion. It is apparent that we have mainly to concentrate on experiments which contain strong structural information since these are indispensable. The experiments typically used for assignment purposes should only be performed when necessary.

For the typical bottom-up strategies discussed above a rather large set of NMR spectra containing assignment information is clearly required. As a consequence structure focused top-down strategies are better suited than bottom-up strategies to fulfill the above requirements. As the most extreme (and with the rapid evolution of structure prediction in bioinformatics not unlikely) case one could reduce the role of NMR spectroscopy to the task of validating a predicted structure without using NMR directly for structure calculation. However, at the present state of the art this seems only to be possible when structures of close homologues are already available.

The validation of the structures has two important aspects: the proof that (1) the obtained structure represents a solution consistent with all experimental data, and (2) that the experimental data are sufficient to define the obtained structure as a unique solution within the limits of a predefined accuracy. For the first condition a number of methods have already been reported, the most important one (but still far from optimal) is probably the calculation of quantities such as *R*-factors. The second problem is still not solved; the calculation of *R*-factors provides only a rather poor solution for this problem (Section 2.2.5). However, the same inherent problem also exists for bottom-up strategies, but is often overlooked. Here, one makes the assumption that a structure is correct when the optimization procedure in the presence of the experimental restraints leads in the majority of trials to similar structures. This is identical to the condition that the majority of all prediction leads to only one set of structures explaining the experimental data.

For practical purposes the required quality of a structure is dependent on the specific problem to be solved. The amount of time and resources needed usually increase rapidly with the demand on quality (resolution). Especially in proteomics one has to optimize the methods with respect to these requirements. One would demand that the structures obtained from automated procedures should be at least as accurate as those obtained from manual data evaluation.

4.1. General strategies

In the conventional approaches described so far, the resonance line assignment is performed followed by the extraction of structurally relevant information as it is contained, for example, in NOESY spectra. In the top-down approaches one starts from a trial structure and uses the structure information contained in the spectra to obtain iteratively improved structures and during this process also resonance assignments (Fig. 10). The trial structure may consist in the two extreme cases either of a random assembly of atoms in space (a cloud of atoms) or of the well-defined structure of a close homologue.

A number of programs is using this strategy where first a three-dimensional structure is calculated from unassigned distance restraints obtained from NOE spectra and then the obtained structure is used for the assignment of the individual spins [387–390]. Oshiro and Kuntz have developed a method that requires as input the identified main-chain protons that are already grouped into unassigned spin systems. Using only inter-residual NOEs and distance geometry trial structures are calculated. However, it was shown that this approach in its present form requires excellent data to succeed.

In the method described by Kraulis [388], a set of 4D ^{13}C and/or ^{15}N edited NOESY spectra is required as input to build a relatively unambiguous network of unassigned NOE restraints. Structures are calculated with a simulated annealing protocol where only one single type of atom which represents the unassigned ^{1}H spin is used. In the next step the yet unassigned ^{1}H spins in the calculated structure are assigned. Here, the probability of a certain assignment for each ^{1}H spin is obtained by considering both the ^{1}H shift and the shift of the covalently bound heteronucleus to predict the amino acid type and atom type from chemical shift probability surfaces derived from published data.

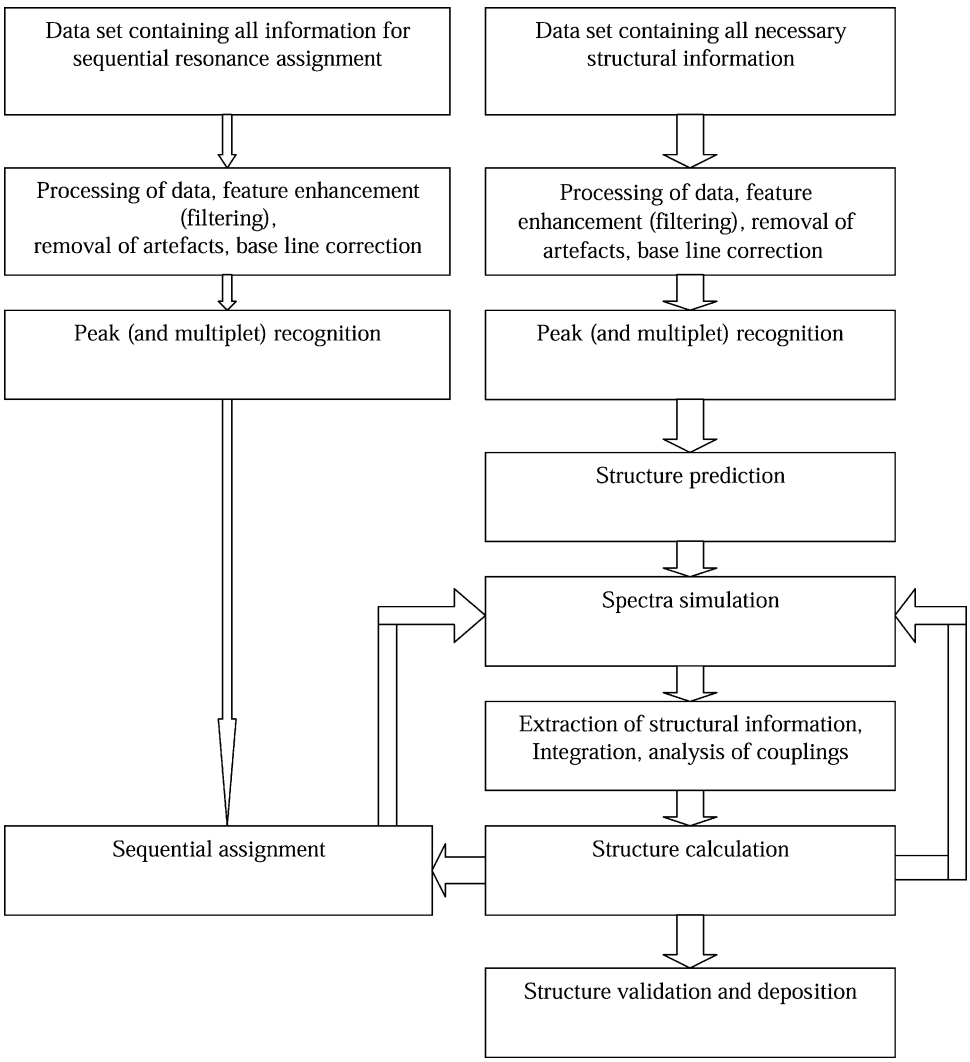

Fig. 10. Schematic representation of top-down approaches for structure determination. The main emphasis is here to obtain a three-dimensional structure that is in agreement with the experimental data.

To obtain the sequence specific assignment it is required that residues adjacent in the protein sequence must be assigned to ^{1}H spin combinations that are spatial neighbors. In this last step the protein chain is traced in the previously calculated 3D structure. The program was tested on semi-synthetic data sets where it performed well. The overall quality of the program is hard to judge since it was not tested on real experimental data.

In CLOUDS [389,390] unassigned NOE signals are extracted from multidimensional NMR spectra. They are converted into interproton distances by relaxation matrix analysis. In the molecular dynamics/simulated annealing procedure which follows a 'collection' of non-connected H-atoms is restrained by these distances to obtain a hydrogen only molecular structure called a cloud. The convergence of the MD calculations was significantly improved by the use of non-NOEs. Selected clouds were superimposed, and from the set of all hydrogen atoms the HN atoms are identified using 2D $^{1}H-^{15}N$ HSQC or exchange experiments. A Bayesian approach is then used to trace the string of backbone atoms within the so-called foc and to identify the side-chains. CLOUDS has been successfully tested on the structure determination of smaller proteins using experimental data.

Most of the previously mentioned approaches are limited to peak lists and are based primarily on the NMR centered evaluation of the spectra without using additional knowledge about the biochemical object. During the last few years various extensive databases containing, for example, chemical shift assignments of biological macromolecules and three-dimensional structure information have been constructed. In our opinion it seems logical to use all this prior knowledge in the structure determination process of a new molecule.

As a consequence we have developed the program AUREMOL, optimized for a molecule-centered approach for the automated structure evaluation in solution. AUREMOL has a general architecture and organization for

the molecule centered data evaluation and is based on and compatible with the program AURELIA [264].

We will discuss this approach in detail in the next section as an example for general-purpose top-down NMR structure determination programs. In this approach, a three-dimensional trial starting structure will be modified in an iterative process until the structure fits the experimental data. Since the starting point of our approach is already a structure it presents a top-down strategy. Based on the trial structure and employing additional knowledge about the sample, NMR parameters or full NMR spectra are predicted and are used as a guideline for the data evaluation. This strategy should lead to faster and better results especially if the automation of the assignment process will be performed using not only peak lists but also the information contained in the NMR spectra like peak shapes, line widths and line splitting. In AUREMOL we are trying to automate the whole structure determination and structure validation process in solution from the point where processed NMR spectra are available.

4.2. Molecule-centered approach (AUREMOL)

4.2.1. Overview

The molecule-centered approach (MCA) as it is currently being implemented in the program AUREMOL is a specific strategy of the general top-down approach. The essentials of this strategy are summarized in Fig. 11. We will describe them here with respect to the existing program AUREMOL, but will also discuss necessary modules which are not yet coded in the program since the aim of the section is to present a general strategy not restricted to a specific program.

(1) A trial three-dimensional structure of the biological macromolecule is the starting point of the data evaluation and a refined three-dimensional structure the final goal. All relevant information about the considered biomolecule should be collected such as primary sequence information, composition of the used buffer and physical parameters, e.g. pH and the temperature of measurement. (2) A general local database provides additional information. It contains data such as the chemical structure of the amino acids, chemical shifts and their distributions, *J*-couplings, Karplus parameters, and temperature dependent viscosities. (3) Structures and sequences of homologous proteins can be loaded from non-local databases. (4) Based on this information a trial assignment and a trial structure are generated. It should be noted that for the basic algorithm one has to allow that these starting values can be far removed from the final results. For example, it must be possible to start with an extended strand as a starting structure. (5) For handling the experimental data an automated processing will be necessary in the longer term. Filtering and Fourier transformation of the data is done outside of AUREMOL by programs such as XWINNMR and the transformed data are accessed by AUREMOL. (6) As discussed above further more involved steps in image (spectrum) analysis are required, most of them are already contained in AUREMOL. They comprise automatic peak picking, calculation of volumes by iterative segmentation [206] and automated removal of noise and artifacts using Bayesian analysis [182,183].

At this point experimental information must be used. The central algorithm selected in AUREMOL is the direct

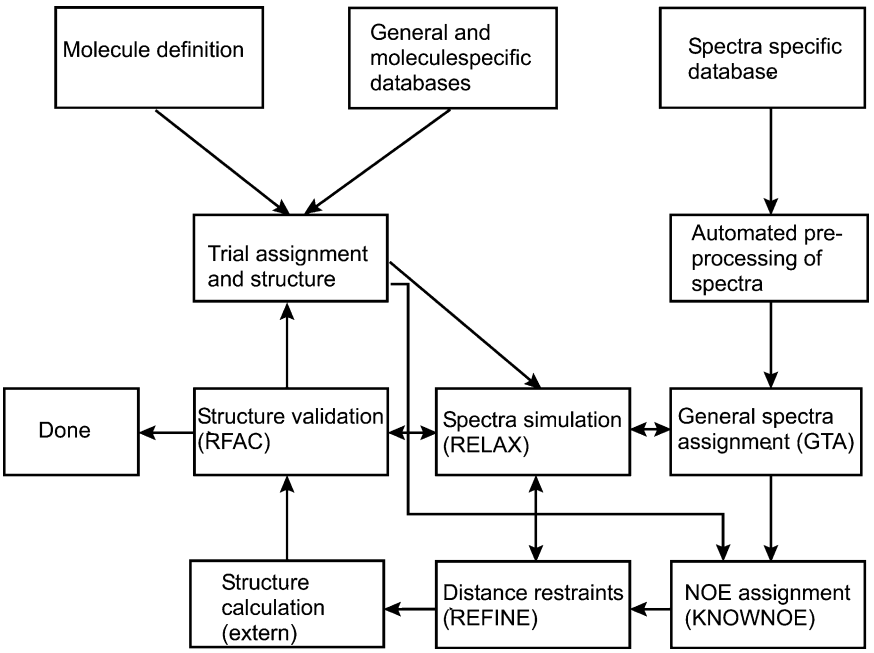

Fig. 11. Representation of the molecule-centered approach used in AUREMOL. The main idea is that a model structure is refined in an iterative process until the resulting structure is in accordance with the experimental data. Details of the algorithm are further explained in the text.

comparison of experimental spectra with spectra simulated on the basis of point (1) to (4). This will lead (7) to an iterative approach where the comparison yields a new spectral assignment which in turn will be used to calculate a new refined three-dimensional structure of the macromolecule. A general advantage of this approach is that characteristics of the (still evolving) pulse sequences are not part of the central algorithm but are contained in the spectrum simulation. Since NOESY spectra are still the dominant source of structurally relevant information we have first concentrated on their automated evaluation. However, simulation of any type of spectra is possible and under development. (8) There are good arguments for interleaving structure calculation (refinement) with the spectral comparison. However, well-developed structure calculation software already exists and is continuously improving. Actually, programs such as X-PLOR or DYANA are connected to the main algorithm by providing parameter files and by reading the resulting structures. The last step (9) would be again the validation, and for this a number of automated routines already exist.

4.2.2. Databases and data structures

The molecule-centered strategy requires well-developed databases which can easily be modified and extended by the user. In addition, the data structures must be flexible and allow one to store and access all necessary information.

The global database contains information that is independent of the molecule under investigation (Table 4). It includes information such as the definition of amino acids in IUPAC format, the chemical structure of each amino acid and possible alias names, definition of various NMR experiments, random coil chemical shift values and chemical shift anisotropy values for proteins, various motion models required for spectra simulation by RELAX, etc.

Table 4
Necessary information contained in the global database

Name	Purpose
as_def.txt	Covalent structure of the amino acids
classes.txt	Standard parameters for motional models required for spectra simulation
cs_table.txt	Average chemical shift values
csa.txt	Standard chemical shift anisotropy values
experiments.txt	Definition of NMR Experiments
IUPAC.txt	Atom names according to IUPAC convention
Modeling_parameters.txt	Parameters for homology modeling in AUREMOL
periodic_table.txt	Weight of common atoms
res_coord.txt	Reference coordinates
Shifts.txt	Secondary structure specific chemical shift values
susc.txt	Parameters for calculation of molecular magnetic susceptibility tensor
topo.jcc	Parameters for calculation of *J*-coupling values

The molecule specific database may include predicted chemical shift values, sequence composition, a homologous three-dimensional structural model, etc., while the spectra specific database contains the experimental data. One of the most important issues for a successful and comprehensible structure determination is the handling of this information in the internal data structure. Since a sample may contain (and usually does contain) more than one molecule the data structure has to admit more than one compound including information about the type and concentration of the components. In addition, since the algorithm relies on the complete simulation of many individual spectra and their simultaneous analysis the data structure must allow the definition of the sample content for all spectra separately.

As it is shown in Fig. 12 the global database described above and a *sequence file* are used to create a *compound file*. It defines one compound of the NMR sample; this may be a protein or another molecule in the sample. The notation of a *compound file* is similar to NMR-STAR (BioMagResBank) format. The *compound file* is divided into several sections; the first section specifies each atom of the compound in sequential order. The next section describes the chemical structure of the compound. The last section defines dihedral angles and *J*-coupling constants or information for the calculation of *J*-couplings using the Karplus equation. It is important to note that a *compound file* contains no sample specific information such as chemical shift values.

In order to extract all relevant data from these spectra a new peak list format has been developed called *masterlist*. One masterlist contains the information of all picked peaks of one NMR spectrum. It starts with a header which specifies processing data, type of experiment and other parameters. Next all picked peaks and their relevant information, like chemical shifts in each dimension, peak label, comments, volumes, volume errors and a factor which gives you the probability that the peak is a true signal or a noise peak will follow. To get the complete resonance assignment of a protein one has to analyze numerous different NMR spectra, including the simulated ones. Therefore, all masterlists of the experimental and simulated NMR spectra can be collected and used to create a new file, the *MasterMasterfile*. The first section of this file specifies the NMR spectra, respectively, masterlists which are connected to the MasterMasterfile. The following sections contain information of the atoms of the different compounds in the NMR sample which is no longer spectra dependent, that means just chemical shift values in ppm and experimentally determined *J*-coupling values of the atoms will be stored but no peak information such as peak volume or the probability that a peak is a signal or noise.

For each compound of the NMR sample such a section will be created. Next all information needed for the automated data evaluation process will be collected in one

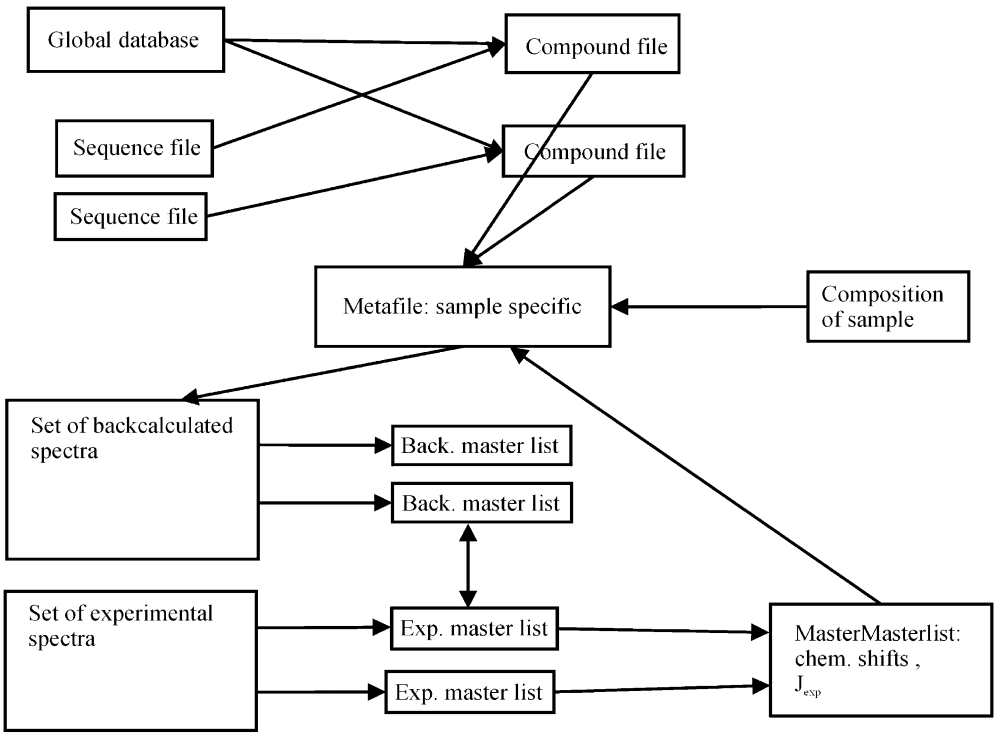

Fig. 12. Data structure of AUREMOL. The central part of the data structure is the meta file required, e.g. by the routines for spectra simulation, spectra assignment, *R*-factor calculation, etc. Details of the data structure are further explained in the text.

file called a *metafile*. The metafile as shown in Fig. 13 is created using the *compound files* defined by the user and the MasterMasterfile. The first section of the metafile defines which compounds are contained in the NMR sample and some specific physical parameters of the NMR sample, e.g. pH or temperature of the measurement. The following sections define the different compounds, namely which motion models are used for the simulation of the NOESY spectra, experimental chemical shift values and *J* coupling values that were extracted from the MasterMasterfile. The Metafile is the basic input file required for the spectra simulation. In the iterative structure determination process in AUREMOL, the masterlists, MasterMasterfile and Metafile will be adapted in each circle.

4.2.3. Preprocessing of experimental NMR spectra

Several tools for the automated preprocessing of NMR spectra are indispensable in automation and are usually implemented in these programs. In AUREMOL already existing routines from AURELIA are used. Signal lists are created using an automated peak-picking algorithm. A practical problem during the evaluation of NMR spectra is the occurrence of noise and artifact peaks. It presents a severe problem in automated assignment routines that may lead to incorrect results or non-converging structures. As a consequence, we have implemented a Bayesian approach coupled to a multivariate linear discriminant analysis of the data to obtain an automated classification of 2D and 3D NMR peaks [182,183]. The method can separate true NMR signals from noise signals, solvent stripes and artifact signals. The analysis relies on the assumption that different signal classes have different distributions of specific properties such as line shapes, line widths and intensities. The classification rule for the signal classes was deduced from Bayes's theorem. Tests have shown that the calculated probabilities for the different class memberships are realistic and reliable, with a high efficiency of discrimination between peaks that are true signals and those that are not.

For the peak volume integration AUREMOL uses an iterative segmentation method combined with a region-growing algorithm. Even for overlapping peaks the volumes can be obtained with sufficient accuracy [206].

4.2.4. Simulation of nD-NMR spectra

As mentioned above the basic idea of the automation process in AUREMOL is the comparison between simulated and experimental NMR spectra. For testing which hypothesis about a given variable (e.g. assignment, volume, *J*-coupling) is most likely the simulation must be as exact as possible since only then can small differences between simulated and experimental spectra be significant. On the other hand, the simulation must be fast enough for allowing a large number of iterations per time unit. This can be obtained by efficient algorithms but usually also involves a compromise in accuracy.

For the simulation of NOESY spectra the complete relaxation matrix formalism is the method of choice (see above). RELAX [261], a program for the back-calculation of NOESY spectra based on complete relaxation matrix formalism, is part of AUREMOL and has been described in

```
section_probedefinition
COMPOUNDS:
_Mol_label
_Compound_file
_Concentration_value
_Concentration_value_units
_Isotopic_labeling
NAME   C:\Programme\Auremol\HPr_Daten\hpr.comp     -     -     -
END_COMPOUNDS
PARAMETER:
_Variable_type
_Variable_value
_Variable_value_unit
pH             7.0    -
Temperature    300.00   K
Pressure        1 Bar
END_PARAMETER
end_section

section_compounddefinition 1
NAME: hpr
CLASSDEF:
 DEFINE_CLASSES

 ANIS

 NAME              METHYL
 NUCLEUS 1H
 OCCUPANCY  1
.
.
END_CLASSDEF
SHIFTS:
_Residue_seq_code
_Atom_num_code
_Atom_alias
_Chem_shift_value
_Chem_shift_value_error
_Chem_shift_ambiguity_code
_Atom_class
_Linewidth
 1   1    -   8.210   0.05   1   3     -
 1   2    -   120.000   0.5   1   8     -
 1   3    -   56.660   0.5   1   7     -
.
.
END_SHIFTS
J_COUPL:
_Coupled_atom_1
_Coupled_atom_2
_Coupling_value
 2,1   2,4    10.1
END_J_COUPL
end_section
```

Fig. 13. Structure of the AUREMOL Metafile. The file consists of two main sections with various subsections. In the first section the sample composition and general sample parameters are defined. For example, if protein complexes are investigated each protein is specified as a separate compound. Each compound is associated with the path of its compound file, the concentration of that specific compound in the sample and if that compound has been isotopically labeled. Next pH, temperature, and pressure of the sample are defined. The next section contains information about each specific compound. In the subsection 'CLASSDEF' the various motional models that are applied during the calculations are defined. These definitions are similar to the definitions described in the original RELAX publication [260]. The 'SHIFTS' subsection associates each atom identified by its residue number and intra-residual number, if applicable with an atom alias identifier, with its chemical shift, with an chemical shift error, with an chemical shift ambiguity code, with a motional model, and if available with an experimentally measured line-width. Please note that only those atoms are listed for which chemical shifts are available. In the subsection 'J_COUPL' experimentally measured *J*-couplings are stored. Each *J*-coupling is identified by the coupled atoms and the coupling constant.

Section 2.2.4.4. RELAX allows the simulation of ^{1}H 2D NOESY spectra and *IS* ($I = {}^{1}$H, $S = {}^{13}$C or ^{15}N) NOESY–HSQC spectra. The *IS* NOESY–HSQC experiment is basically a concatenation of a homonuclear ^{1}H-NOESY and a heteronuclear *IS* HSQC-experiment. Under ideal conditions the net magnetization transfer during the entire experiment can be calculated by multiplying the net polarization transfers of the two basic experiments which can be performed in RELAX. Usually, the NOESY-parts of the 3D pulse sequences differ slightly from the standard homonuclear NOESY pulse sequence, since it is advantageous to decouple the heteronuclei during the evolution period t_1. This is often done by an additional 180°_{S} pulse in the midpoint of t_1. This option is therefore also available. In addition, a number of motional models are available and can be defined for specific spin pairs. Relaxation by chemical shift anisotropy is an option in the simulation.

All experimentally obtained NOESY spectra are acquired with a finite relaxation delay. Therefore, for a more realistic simulation we have included this option in RELAX as well. To include effects such as partial solvent exchange, e.g. for NH groups or to allow the simulation for partial deuterated molecules, separate occupancy levels can be specified for each atom or group of atoms.

The time required for performing the back-calculations for medium sized proteins on a standard PC is usually only a couple of minutes. If even faster calculations are desired or if less accuracy is necessary for a special task, it is possible in the case of 3D NOESY–HSQC spectra to neglect the heteronuclear relaxation during the calculations. By doing so it is possible to reduce the required CPU time by a factor of 2 to 3. In most cases only a small error is introduced in the calculation of the cross-peak intensities if the heteronuclear relaxation is neglected.

Also the simulation of line intensities is sufficient for all cases where only cross-peak volumes are compared. A more elaborate simulation is necessary when line shapes are also used as an additional source of information. It turns out that this is important for a spectrum-based resonance assignment and structure determination. Therefore, individual T_2 times have to be calculated to obtain the line widths in the n dimensions. In addition, J-couplings have to be taken into account to simulate the resolved or non-resolved line splittings as well as possible. An example is shown in Fig. 14.

Although the line shape of a single transition in solution is a Lorentzian we also allowed a Gaussian line shape for the simulation of the frequency domain data. This can serve as a compensation of shape modulation of the experimental data by time domain filtering. However, more exact agreement between calculated and experimental data can be obtained when time domain data are simulated, properly filtered, and Fourier transformed. This should be the method of choice and has to be implemented into the program.

A single trial structure may only explain a limited number of NOEs, and as a possible option, a set of structures

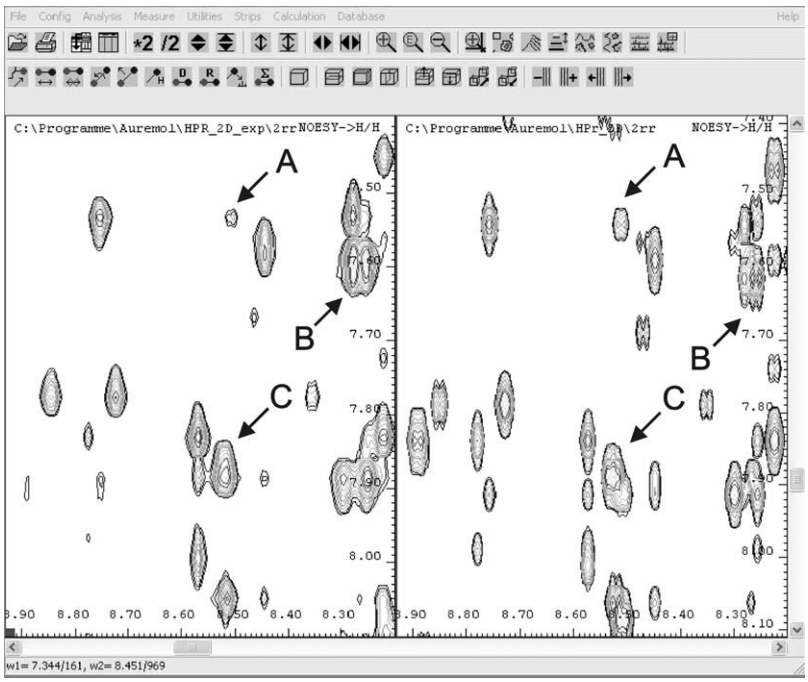

Fig. 14. Comparison of experimental and simulated ^{1}H 800 MHz 2D NOESY spectra of HPr from *S. carnosus*. On the left a section of the NH–NH region of the experimentally determined spectrum of HPr and on the right the corresponding region of the simulated spectrum is shown. By comparing the two spectra it becomes clear that the line-widths and the line splitting due to *J*-coupling are realistically simulated (Ried et al., to be published). E.g. for peaks A and B nicely separated doublets are visible, while for peak C no separate sub-peaks are visible but the line-shape has been considerably altered due to the presence of *J*-coupling. Please note that in the simulated spectrum each peak is labeled with the names of the contributing atoms. However, for reasons of clarity the display of these names has been switched off in Fig. 14.

can be used by the program and an average spectrum can be calculated by averaging the corresponding relaxation matrices. Such an averaging would correspond to a fast exchange between the different conformational states represented by this set of structures. In this case, the chemical shift values given to the program represent the population-averaged shifts.

The output of the program is a list of simulated peaks with their corresponding chemical shift values, labels and intensities and a simulated ^{1}H 2D NOESY spectrum or a ^{15}N- or ^{13}C-edited 3D-NOESY–HSQC spectrum in BRUKER submatrix format which can be used directly by AUREMOL and AURELIA.

4.2.5. Structure based assignment and iterative structure determination

A number of different approaches for a structure-based assignment have been published already and are discussed above. In our opinion, one should use the independent information characterizing the covalent structure of the molecule in any step of the procedure. That is, one should use a proper physical model from the start of the assignment procedure.

In the molecule centered approach of AUREMOL the experimental and simulated spectra are directly compared on the level of individual cross peaks. To do this, the pixels i,j (voxels i,j,k) contributing to a simulated cross peaks S_l are obtained by an iterative segmentation procedure and stored separately. The agreement of their values $I^S(i,j)$ with the intensities $I^E(i,j)$ of the experimental peak E_m can then be measured by means of a target function $M(S_l,E_m)$. Depending on the exact definition of the target function, M should have a global maximum (or minimum) when optimal agreement is obtained. One possibility to define M can be derived by calculating the cosine α between the two vectors $\mathbf{S}$ and $\mathbf{E}$ whose components are $I^S(i,j)$ and $I^E(i,j)$

$$\cos\alpha = \frac{\mathbf{SE}}{|\mathbf{S}||\mathbf{E}|} \tag{4.1}$$

M is then defined by

$$M(\mathbf{S},\mathbf{E}) = \begin{cases} \cos\alpha, & \text{for } \cos\alpha \geq 0, \\ 0, & \text{for } \cos\alpha < 0 \end{cases} \tag{4.2}$$

This definition is optimally suited for comparing shapes independent of the total intensity.

On the basis of this peak comparison, we have tested the convergence of a generalized threshold-accepting algorithm for obtaining the resonance assignment from NOESY spectra only (Ganslmeier et al., to be published). As input a model structure and the partial sequential assignment are required, and arbitrary values are assumed for unknown sequential assignments. NOESY spectra are simulated with these data using the full relaxation matrix approach, and this is combined with the simulation of transversal relaxation times and J-couplings. The general applicability of the algorithm was first tested on synthetic data. Fig. 15 shows the dependence of correctly obtained assignments from a single 2D-NOESY spectrum on the percentage of resonance assignments obtained from other sources. It shows that a large percentage of assignments can be obtained by this method. After these first successful tests we also investigated the performance of the algorithm on experimental data.

These tests on experimental data clearly indicate that in its current stage the algorithm is capable of obtaining the complete assignment of single mutants from a protein with known structure and resonance assignments. It also shows that the general method is most probably working when additional information from typical spectra containing mainly assignment information, such as HNCA spectra, is

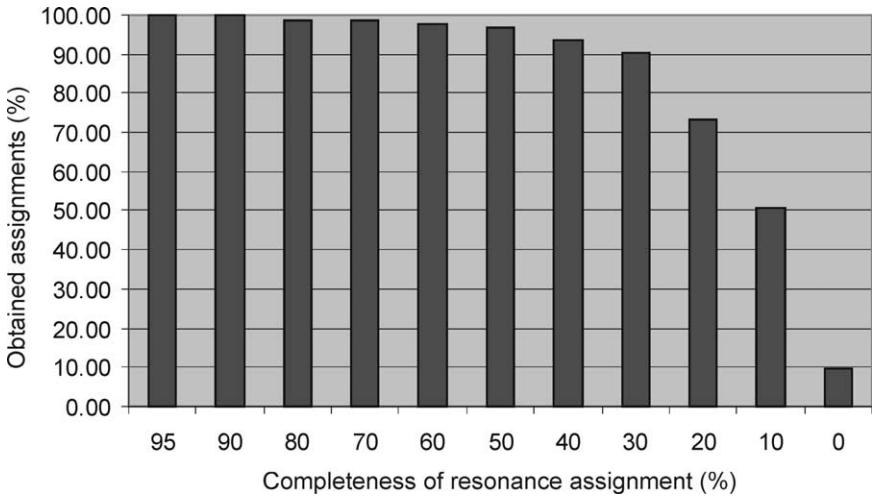

Fig. 15. Resonance assignments obtained from a single NOESY spectrum using the general threshold algorithm implemented in AUREMOL. Dependence on completeness of partial resonance assignment. On the X-axis the percentage of the available partial resonance assignment is shown (input), while on the Y-axis the completeness of the obtained assignment is visualized (output).

used in the algorithm. However, a number of open questions still remain when applying this method, as an example it is not clear how the information from different spectra should be weighted when they are used simultaneously. Another advantage of this method is that are not only complete resonance line assignments obtained, but that the structure relevant information is also extracted from the NOESY spectra at the same time.

4.2.6. Knowledge driven assignment of NOESY spectra

The probabilistic method KNOWNOE for the automated evaluation of NOESY spectra is applicable when the sequential resonance line assignment has been fully or almost completed and no or only partial knowledge about the three-dimensional structure is available. Structural information, when available is helpful in the assignment process especially for larger systems, but it is not required. KNOWNOE contains as a central part a newly developed knowledge driven Bayesian algorithm for solving ambiguities in the NOE assignments. These ambiguities mainly arise from chemical shift degeneracies, which allow multiple assignments of cross peaks.

Statistical tables in the form of atom-pairwise volume probability distributions (VPDs) were derived from a set of 326 protein NMR structures. VPDs for all assignment possibilities relevant to the assignments of interproton NOEs were calculated. With these data for a given cross peak, with N possible assignments $A_i (i = 1, \ldots, N)$, the conditional probabilities $P(A_i, a|V_0)$ that the assignment A_i determines essentially all (a-times) of the cross-peak volume V_0 can be calculated. An assignment A_k with a probability $P(A_k, a|V_0)$ higher than 0.8 is transiently considered as unambiguously assigned. With a list of unambiguously assigned peaks a set of structures is calculated. These structures are used as input for a next cycle of iteration where a distance threshold $D_{\max}$ is dynamically reduced.

Starting with a trial structure (e.g. an extended strand) all assignments of a cross peak possible within the chemical shift tolerances Δ_1 and Δ_2 are considered where the corresponding atoms are separated in the trial structure by a distance $r_{ij} < D_{\max}$. They are stored together with the volume V in the list of unassigned NOEs (U-list). The U-list is automatically created from the corresponding masterlist and it is only visible to the KNOWNOE part of AUREMOL. The required resonance line assignments are contained in the corresponding Metafile. If there is only one assignment possible for a cross peak, this assignment is transferred to the list of unambiguously assigned NOEs (A-list) that is also only visible to KNOWNOE. Based on the observation that cross-peak volume and correct peak assignment are not independent of each other, the U-list is then searched for cross peaks which can be assigned to more than one pair of atoms from their chemical shifts but where there is a large conditional probability $P(A_i, a|V_0)$ that most of the volume V_0 of a cross peak originates from just one assignment A_i. More exactly, if $A_i (i = 1, \ldots, N_{\text{ab}})$ is a possible assignment of a cross peak, it is transferred to the A-list if

$$P(A_i, a|V_0) \geq P_{\min} \tag{4.3}$$

with $V_{\min} = aV_0$ the lower limit of the volume which is explained by the assignment i.

With the NOEs of the A-lists and optionally additional restraints like J-coupling restraints, a set of N_s structures is calculated. In the subset of $bN_s (0 < b \leq 1)$ structures with the lowest total energies, it is checked whether some NOE restraints are systematically violated. These are removed from the A-list if the difference between the distance r_{calc} determined for the calculated structure and the distance $(r_{\text{exp}} + \Delta)$ determined from the experimental data is larger than the tolerance Δ_{viol} in at least N_x cases. That is,

$$r_{\text{calc}} - r_{\text{exp}} - \Delta^+ > \Delta_{\text{viol}} \tag{4.4}$$

with Δ^+ defining the maximum error of r_{exp} allowed in the structure calculation. In the current implementation N_x and Δ_{viol} (typically set to 0.02 nm) have to be specified by the user.

With these restraints, a new set of structures is calculated, and the maximum distance $D_{\max}$ allowed for assignments is reduced and a new A-list is created as described above and by using as a trial structure the structure with the lowest total energy of the previous run. Here, $D_{\max}$ describes the maximum distance in the current trial structure that is allowed between two atoms or groups of atoms contributing to an NOE. This procedure is iterated until $D_{\max}$ reaches its lower limit. The lower limit of $D_{\max}$ is usually set to 0.5 nm, in general the maximum detection range of a NOESY spectrum. Note that in each iteration, the original U-list is used and all previous assignments are discarded. This is done to ensure that the structure determination process does not get trapped in preliminary conformations.

After a last iteration with $D_{\max} = 0.5$ nm there is still a large number of cross peaks which can (or must) be explained by more than one assignment. At this point, a new list of restraints is created out of the A-list and the U-list. The multiply assigned cross peaks from the U-list with volumes V_0 are all taken as possible solutions but the expectation value of the interatomic distance r_{0i} of the assignment A_i is scaled by

$$r_{0i} = \left(\left\langle \frac{\sum_{j=1}^{j=N} V_j}{V_0 V_i} \right\rangle \right)^{1/6} \tag{4.5}$$

with V_j the volumes corresponding to the distance of the atoms in assignment j. With this complete list of assignments (and the list of restraints other than NOE), a new set of structures is calculated. This procedure is similar to the scaling used [391] in the ARIA approach.

Calculation of the assignment ambiguity. In the above strategy the probability $P(A_i, a|V_0)$ that the assignment A_i explains at least the part a of the experimental cross-peak volume V_0 has to be calculated. Starting with Bayes's theorem [392], the probability that more than a-times ($0 \leq a \leq 1$) of the volume V_0 is explained by an assignment A_i can be calculated from

$$P(A_i, a|V_0) = \frac{P(A_i, a)P(V_0|A_i, a)}{\sum_{i=1}^{N_{ab}} P(A_i, a)P(V_0|A_i, a)} \tag{4.6}$$

The simplest case occurs if only one assignment A_1 is possible from the chemical shifts. Here, the a priori probability for the assignment $P(A_1, a) = 1$ and $P(A_i, a, i > 1) = 0$ leads to

$$P(A_i, a|V_0) = 1 \tag{4.7}$$

When based on chemical shifts only two assignment possibilities A_1 and A_2 exist for a given cross peak, the probabilities $P(A_i, a)$ and $P(V_0|A_i, a)$ must be calculated prior to the calculation of $P(A_i, a|V_0)$. Since no other assignments are assumed possible, the a priori probability for $i > 2$ is given by

$$P(Ai, a) = 0, \qquad (i > 2) \tag{4.8}$$

For the non-trivial cases $i = 1$ and 2, the value of $P(A_i, a)$ can be approximated by

$$P(A1, a) = P(A2, a) = 0.5c_s, \qquad \text{with } 0 \leq c_s \leq 1 \tag{4.9}$$

if the expected volumes of the two classes show the same probability distribution. The constant c_s is a normalization constant depending on the shape of the probability distribution which is cancelled during the calculation of $P(A_i, a|V_0)$. A more general expression that is used within KNOWNOE can be derived as

$$P(A_1, a) = \int_{V_0=0}^{\infty} \int_{V_1=aV_0}^{V_0} p_1(V_1)p_2(V_0 - V_1)\mathrm{d}V_1\mathrm{d}V_0 \tag{4.10}$$

and

$$P(A_2, a) = \int_{V_0=0}^{\infty} \int_{V_2=aV_0}^{V_0} p_1(V_0 - V_2)p_2(V_2)\mathrm{d}V_2\mathrm{d}V_0 \tag{4.11}$$

with $p_1(V)$ and $p_2(V)$ the normalized probability densities for finding a volume V for pairs of atoms with the assignments A_1 and A_2, respectively. The probabilities defined by the two equations above are properly normalized when the distributions $p_1(V)$ and $p_2(V)$ are normalized.

For two possible assignments the probabilities $P(V_0|A_i, a)$ can be obtained as

$$P(V_0|A_1, a) = \int_{V_1=aV_0}^{V_0} p_1(V_1)p_2(V_0 - V_1)\mathrm{d}V_1 \tag{4.12}$$

and

$$P(V_0|A_2, a) = \int_{V_2=aV_0}^{V_0} p_1(V_0 - V_2)p_2(V_2)\mathrm{d}V_2 \tag{4.13}$$

When there are three assignment possibilities for a cross peak, the corresponding equations are defined analogously to the case described above.

Scaling of experimental volumes. For performing the calculations mentioned above, the experimental volumes are scaled to adjust to the expected probability distributions of the volumes. In the actual version of KOWNOE a manual approach is used where the user has to specify a reference volume that corresponds to a certain reference distance.

Calculation of the probability distributions. An estimate of the probability distributions p_i of the peak volumes for the assignments A_i is required for the calculation of $P(V_0|A_i, a)$. Although it is possible to formulate a priori assumptions on these distributions, the better way is the extraction of statistical data from known protein structures. For obtaining meaningful distributions, one has to classify the specific assignments A_i of pairs of atoms to obtain a sufficiently high number of class members for the statistical analysis. A powerful way is to extract the information independently of the absolute positions in the sequence S_i and $S_{i'}$. Knowing the absolute position S_i of one amino acid of the pair of atoms considered, the pairwise interaction of any atoms in the protein can be described by the separation in the sequence $\Delta S_i = S_{i'} - S_i$ (without restricting the generality we assume in the following $S_i \leq S_{i'}$), and by the atomic numbers Z_j and $Z_{j'}$. The total sequence information can be coded if in addition the residue types T_i and $T_{i'}$ of amino acids are stored. An assignment $\mathbf{A}_k$ can be stored as a vector $\mathbf{A}_k = (S_i, \Delta S_i, Z_j, Z_{j'}, T_i, T_{i'})$.

If we create sequence independent classes $C_l (l = 1, \ldots, L)$ defined by the sequence independent information $(\Delta S_i, Z_j, Z_{j'}, T_i, T_{i'})$, then $\mathbf{A}_k$ can be written in the reduced form $\mathbf{A}_k = (S_i, C_l)$. In our case the probability distribution $p_k(V)$ of the volume V of a possible assignment A_k can be approximated by the probability distribution $\tilde{p}_k(V)$ of the corresponding class C_l. The same notation can be used for other purposes as well, as has been done for example in the knowledge based structure prediction published by Subramaniam [393].

Calculation of probability tables. The structures of 326 proteins from the PDB databank were taken as data basis for the statistics. Only NMR structures of water-soluble proteins containing no paramagnetic center or larger cofactor were selected. No RNA and DNA structures or complex structures of proteins with RNA or DNA were considered. Using these structures 1577 different assignment class probability tables were calculated containing the corresponding distance distributions. Two examples of distance probability distributions (DPDs) are shown in Fig. 16. Volumes V_{ij} were calculated from the distances r_{ij} between atoms i and j by the relation $V_{ij} = c_V r_{ij}^{-6}$.

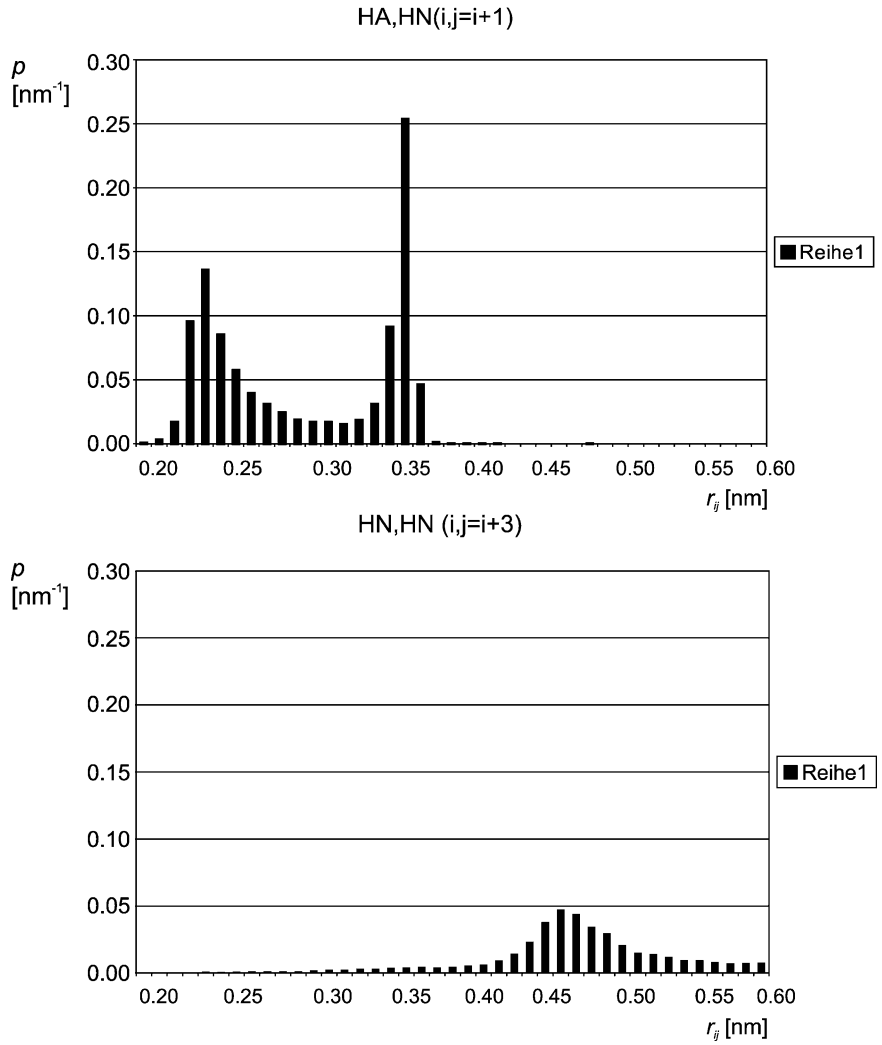

Fig. 16. Examples for probability distributions of distances in 326 selected three-dimensional protein structures. (Upper part) Probability density p of the distances r between Hα of a residue in position i and an HN in position $i + 1$. (Lower part) Probability density p of the distances r between HN of a residue in position i and an HN in position $i + 3$. Figure adapted from Ref. [233].

In principle one can recalculate the VPDs from the DPDs as we did for this study. The result is depicted in Fig. 17 for the same pairs of atoms shown in Fig. 16. The disadvantage of this procedure is that the resolution (the basic widths of the volume classes) is now dependent on the volume.

It is obviously more appropriate to calculate the volume distributions directly from the basic data sets as will be done in the next implementation of KNOWNOE. Nevertheless, it is clear that the two DPDs and VDPs strongly distinguish between the two possible assignments. Probability tables were calculated only for interproton contacts since only these are detected in standard NOESY spectra of proteins.

Although one can argue that one should use a NMR database for NMR data evaluation, this database shows a high degree of heterogeneity of resolution and is biased by the selection of structures. As an alternative one could use a so-called unbiased database of selected X-ray structures, as has been used for structure prediction [394]. We have calculated the corresponding probability distributions which in general do not deviate much from those obtained from the NMR data. It remains to be seen if the new probability tables will provide more reliable results in our applications.

4.2.7. Structure calculation and validation

The three-dimensional structure calculation itself is not part of AUREMOL. The calculations can be performed by some other program such as DYANA or XPLOR. For an overview see the first part of this review.

One of the most important steps in any structure determination project is the validation of the final and/or intermediate structures. Up to now the quality of an NMR structure is mainly judged by factors such as RMSD values or the quality of the Ramachandran plot. However, these methods do not provide a direct measure of how well the structures obtained fit the experimental NMR data. As a consequence we have implemented the program RFAC [323] in AUREMOL, which automatically calculates

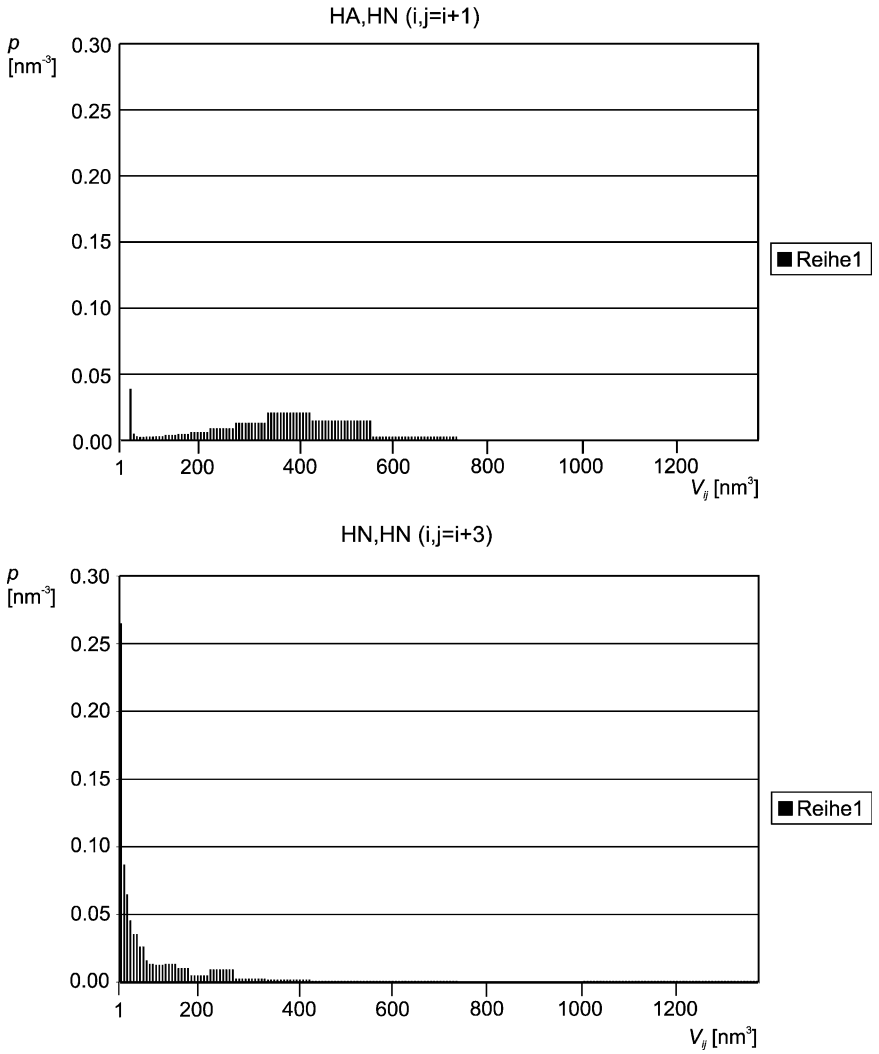

Fig. 17. Examples for probability distributions of volumes in 326 selected three-dimensional protein structures. (Upper part) Probability density p of the expected normalized volumes V of a cross peak between Hα of a residue in position i and an HN in position $i + 1$. (Lower part) Probability density p of the expected normalized volumes V of a cross peak between HN of a residue in position i and an HN in position $i + 3$. Note that in this case the volume distribution was calculated directly from the distance contribution, leading to a volume dependent resolution in the volume space. Figure adapted from Ref. [233].

R-factors for protein NMR structures to provide such a measure (see above for details of the definition of *R*-factors).

The automated *R*-factor analysis envisaged here consists in principle of two separate parts: (1) a comparison of the experimental NOESY spectrum with the NOESY spectrum back-calculated from a given structure, and (2) the calculation of the *R*-factor(s) from the data.

In the first part the NOESY spectrum has to be calculated from the trial structure using the sequential assignments contained in the Metafile; that is the spin systems have to be assigned completely or almost completely. In our implementation we use the full relaxation matrix approach of the program RELAX to obtain accurate simulated peaks defined by their positions and intensities (simulated masterlist).

The corresponding experimental NOESY spectrum is automatically peak picked and integrated in the pre-processing stage of AUREMOL. In addition, the probabilities *pi* of the peaks *i* to be true NMR signals and not noise or artifact peaks are also calculated according to Bayes theorem and are then used as weighting factors during the calculation of the *R*-factors. All this information is contained in the experimental masterlist. For the purpose of *R*-factor calculation the experimental data are automatically assigned with the newly developed AUREMOL routine AUNOAS.

Basically, the program tries first to optimally adapt the chemical shift values obtained from the general sequential resonance assignment to the actual experimental data by a global comparison of the back-calculated spectrum with the experimental spectrum. The peak assignment itself is done on local peak clusters. For each back-calculated peak a search is performed if a corresponding experimental peak exists in a given search radius. If more than one solution exists decision between them is made based on a maximum likelihood criterion.

The experimental and simulated masterlists are fed into the program RFAC. The experimental signals for which no corresponding simulated peaks were found and which therefore remain unassigned will be called the set U in the following while the set A of the experimental masterlist contains all assigned experimental signals. This division of the experimental masterlist into sets U and A is required by RFAC for the calculation of R-factors (Section 2.2.5). The U-list can be further reduced by applying a lattice algorithm which can be used if one assumes that the sequential assignment is true and almost complete. In this algorithm only non-assigned peaks are taken into account where at least one back-calculated peak in each dimension can be found within user defined search radii, e.g. 0.01 ppm for 2D spectra. In this context it is important to note that for each atom at least the structure independent diagonal peak is back calculated. Where more than one back-calculated peak is assigned to a single experimental peak, the mean volume of the corresponding back-calculated peaks is estimated before the comparison is done, while the volume of the experimental peak is divided by the number of corresponding back-calculated peaks.

Application of AUREMOL. AUREMOL has been applied for the automated structure determination of several molecules such as the coldshock protein *Tm*CSP from *T. maritima* and the nucleotide exchange factor of Ral RalGDS. These are two medium size proteins of 66 and 88 residues in length, respectively. In both cases the structure determination began from extended strand structures. As described above the experimental spectra were automatically peak-picked and artifacts and noise were removed.

In these tests of AUREMOL we concentrated on the automated structure determination using KNOWNOE for NOE assignment and RFAC for structure validation. The sequential assignments of *Tm*Csp [395] and RalGDS [396] were taken from the literature and automatically adapted to the corresponding NOESY spectra. In these tests the use of 2D NOESY spectra was sufficient, although ^{15}N or ^{13}C edited NOESY–HSQC spectra could have been used as well. Five and seven iteration cycles were performed for *Tm*CSP and RalGDS, respectively, with the strategy presented in Fig. 18.

The lower error bounds Δ^- were set to the proton van der Waals distance ($\Delta^- = r_{exp} - 0.18$ nm) [211] while the upper error bounds Δ^+ were set to $0.125\, r_{exp}^2$ as proposed by Nilges and O'Donoghue [391]. In addition to the NOE restraints, 28 hydrogen-bonds and 42 ϕ-angle restraints were used for *Tm*CSP and 30 hydrogen-bonds, 82 ϕ-angle and 15 χ_1-angle restraints were used for RalGDS in the following structure calculations.

Using CNS $N_s = 50$ structures were calculated using a standard simulated annealing protocol and the five best structures in terms of total energy were selected for further analysis. The selected structures were then automatically screened for NOE violations. All restraints that were violated by more than $\Delta_{viol} = 0.02$ nm (Eq. (4.4)) in at least two of the selected structures were automatically removed from the restraint file. It should be noted that the violated restraints of the previous step were only removed from the restraint file but not from the corresponding experimental masterlist. Therefore, the corresponding signals could be reassigned in the next iteration. In the course of the iterative structure

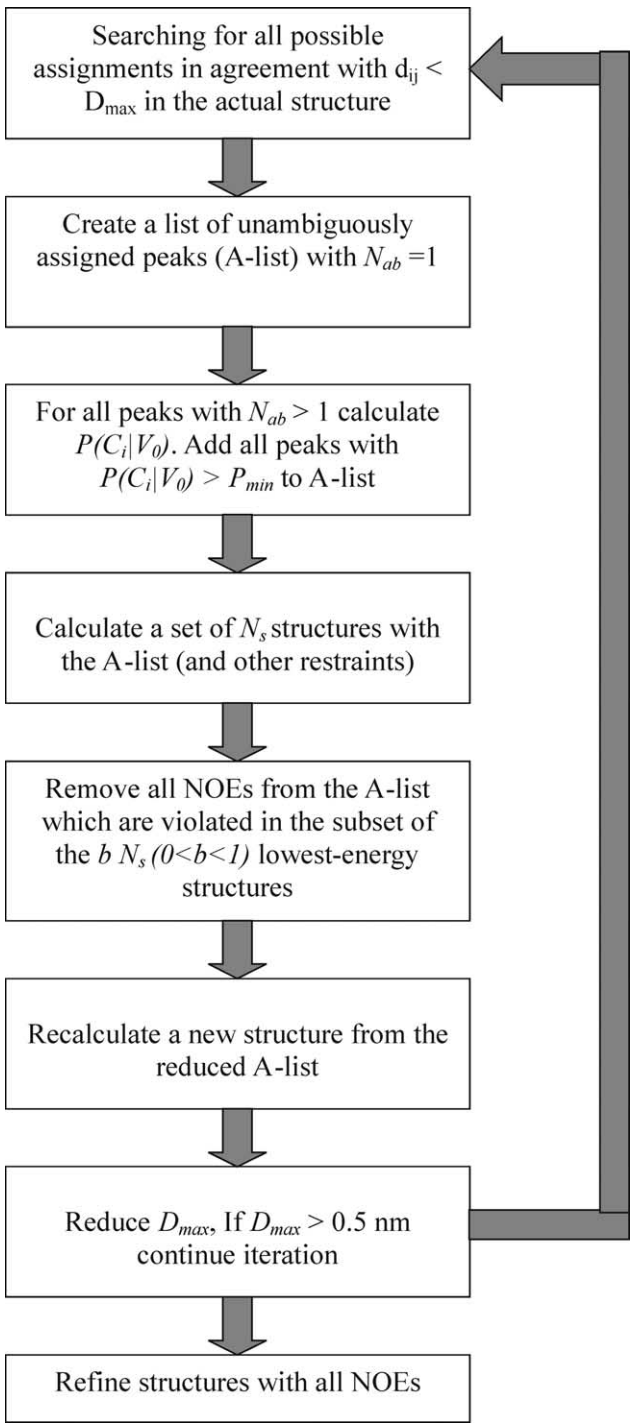

Fig. 18. Schematic representation of the procedure for handling ambiguous NOEs within KNOWNOE. Figure adapted from Ref. [233].

determination the distance cutoff D_{max} used in KNOWNOE was subsequently reduced from ∞ in the first iteration to 0.50 nm in the last iteration. KNOWNOE can only calculate assignment probabilities when not more than three assignments are possible for a given signal based on the sequential assignment. Signals for which no assignment probabilities could be calculated remain unassigned. Therefore, in the first iteration a substantial number of signals remain unassigned. By tightening the distance cutoff D_{max} the amount of ambiguity is reduced which in turn leads to a higher number of assigned signals. In the last iteration 645 and 2096 signals were automatically assigned for *Tm*CSP and RalGDS, respectively. Please note that for *Tm*CSP only a single 2D NOESY spectrum measured in H_2O was used, while for RalGDS in addition a 2D NOESY spectrum measured in D_2O was employed.

Fig. 19 demonstrates for *Tm*CSP the convergence of the structures in the course of the five iterations of the automated structure determination. For both *Tm*CSP and RalGDS, the superposition of the lowest-energy structures obtained from automated structure determination with the corresponding previously manually determined structures shows that for the regular secondary structure elements virtually the same structures are obtained by the two methods (Fig. 20). Minor differences can be observed for the loop regions.

The results (Table 5) show that for both proteins very similar *R*-factors were obtained using RFAC for

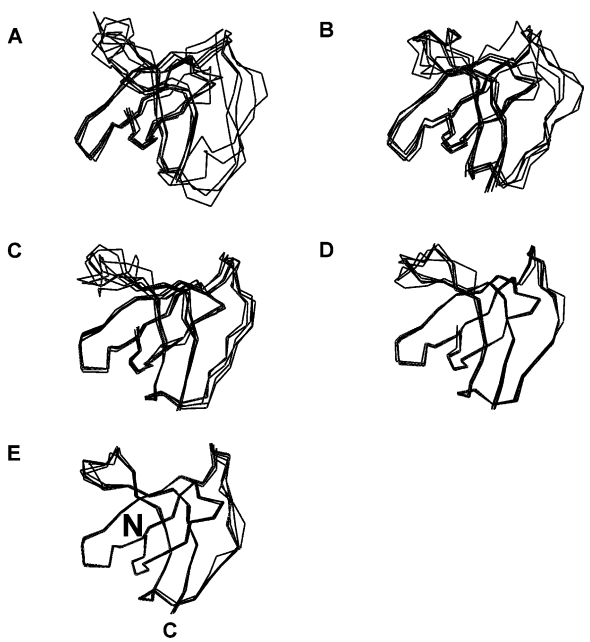

Fig. 19. Example for structural bundles obtained during the iterative structure calculation using KNOWNOE. The five lowest energy structures of *Tm*CSP obtained in different phases of the automated structure calculation are superimposed. Only the Cα-atoms are traced. First iteration (A), second iteration (B), third iteration (C), fourth iteration (D), fifth iteration (D). For each iteration the resulting structures calculated after removal of violated NOEs are shown. Figure adapted from Ref. [233].

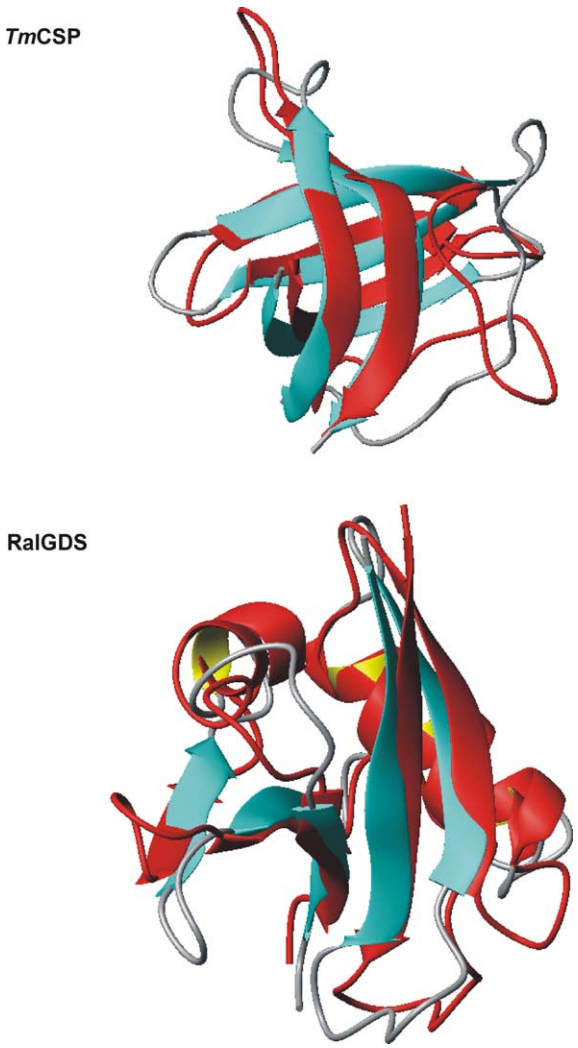

Fig. 20. Comparison of structures obtained by manual and automated NOE assignment. (Top) One of the final solution structures of *Tm*CSP obtained using manual NOE assignments (red) superimposed to one of the final structures obtained using assignments obtained from KNOWNOE. (Bottom) One of the final solution structures of RalGDS obtained using manual NOE assignments (red) superimposed to one of the final structures obtained using assignments obtained from KNOWNOE. Figure adapted from Ref. [233]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Quality of the automatically obtained NMR structures of RalGDS and *Tm*Csp as determined by the calculation of *R*-factors

	RalGDS[a]	*Tm*CSP[a]
Automated[b]	0.38	0.36
Manual[c]	0.41	0.35

[a] Global *R*-factors were calulated using RFAC/AUREMOL. For both molecules the NH-region of the two-dimensional NOESY spectra was used for the *R*-factor calculation.

[b] Structures were determined from the automatically detemined NOE assignment using the program AUREMOL/KNOWNOE.

[c] Structures were determined from the manually derived NOE assignment.

the automatically and manually determined structures indicating that all structures fit the experimental data equally well.

Here, it is important to note that the structures obtained using AUREMOL were solved in a fraction of the time required for the manual structure determination with a minimal set of structurally relevant spectra. In the tests of KNOWNOE described above the solution structure of the molecule was determined using no model structure of, for example, a homologous protein, which is quite a challenging test for the program. In this regard it should be noted that the KNOWNOE part of AUREMOL works best using a complete or almost complete sequential assignment where all main chain and side-chain resonances have been assigned.

In addition, it is preferable for KNOWNOE that the sequential assignment fits the spectra in use as well as possible, enabling the user to use small chemical shift tolerance values.

AUREMOL provides routines to optimally adapt the chemical shifts to a specific spectrum. This is especially important for the first iteration when a model structure is not available. In this case large chemical shift tolerance values usually result in many assignment possibilities for a given signal. However, if the number of possible assignments exceeds three, this signal will be excluded by KNOWNOE leading to a relatively small number of NOE restraints. Therefore, in unfortunate cases it may be possible that for large chemical shift tolerance values insufficient NOE restraints might be obtained to allow proper folding of the molecule.

In case where a model structure is available, the number of possible assignments will be limited depending on the used distance cut-off, enabling the user to apply increased chemical shift tolerance values.

5. Conclusions and outlook

Computer automated determination of solution structures will continue to grow in importance. Fully automated procedures for small and medium sized proteins are nowadays feasible. However, for large biopolymers, aid from human experts is still required.

Typically, in macromolecular NMR, the information content of the NMR spectra alone is hardly sufficient for a complete three-dimensional structure determination. Signal-to-noise ratios are necessarily limited because of the limited solubility of the biopolymers. Furthermore, superpositions of resonance lines often lead to interpretational ambiguities. Therefore, it is still of importance that the computer program permits intervention at any step in the automated analysis to allow structural and spectroscopic information to be included from other sources and to supervise various aspects of the evaluation process. Furthermore, as long as new experimental multidimensional NMR techniques are being developed, new strategies for automated analysis, which must be implemented in existing computer programs, are likely to arise.

Acknowledgements

We thank K.-P. Neidig for interesting discussions, help and ideas during the AUREMOL project. This work was supported by a grant from the EU program.

References

[1] J.C. Kendrew, G. Bodo, H.M. Dintzis, R.G. Parrish, H. Wyckoff, D.C. Phillips, Nature 181 (1958) 662.
[2] G. Wagner, W. Braun, T.F. Havel, T. Schaumann, N. Go, K. Wüthrich, J. Mol. Biol. 196 (1987) 611.
[3] S.K. Burley, Nat. Struct. Biol. 7 (2000) 932.
[4] S.E. Brenner, Nat. Struct. Biol. 7 (2000) 967.
[5] A. Savchenko, A. Yee, A. Khachatryan, T. Skarina, E. Evdokimova, M. Pavolva, A. Semesi, J. Northey, S. Beasley, N. Lan, R. Das, M. Gerstein, C.H. Arrowsmith, A.M. Edwards, Proteins 50 (2003) 392.
[6] M.J. Wood, E.A. Komives, J. Biomol. NMR 13 (1999) 149.
[7] D.L. Jarvis, Virology 310 (2003) 1.
[8] D. Nathans, G. Notani, J.H. Schwartz, N.D. Zinder, Proc. Natl Acad. Sci. USA 48 (1962) 1424.
[9] A.M. Edwards, C.H. Arrowsmith, D. Christendat, A. Dharamsi, J.D. Friesen, J.F. Greenblatt, M. Vedadi, Nat. Struct. Biol. 7 (2000) 970.
[10] J. DeVries, G. Zubay, Proc. Natl Acad. Sci. USA 57 (1967) 1010.
[11] G. Zubay, Annu. Rev. Genet. 7 (1973) 267.
[12] J.M. Pratt, in: B.D. Hames, S.J. Higgins (Eds.), Transcription and Translation. A Practical Approach, IRL Press, Oxford, 1984, p. 179.
[13] A.S. Spirin, V.I. Baranov, L.A. Ryabova, S.Y. Ovodov, Y.B. Alakhov, Science 242 (1988) 1162.
[14] V.I. Baranov, I.Yu. Morozov, S.A. Ortlepp, A.S. Spirin, Gene 84 (1989) 463.
[15] D.E. Nevin, J.M. Pratt, FEBS Lett. 291 (1991) 259.
[16] A.G. Ryazanov, B.B. Rudkin, A.S. Spirin, FEBS Lett. 285 (1991) 170.
[17] T. Kigawa, Y. Muto, S. Yokoyama, J. Biomol. NMR 6 (1995) 129.
[18] T. Kigawa, T. Yabuki, Y. Yoshida, M. Tsutsui, Y. Ito, T. Shibata, S. Yokoyama, FEBS Lett. 442 (1999) 15.
[19] D.-M. Kim, C.-Y. Choi, Biotechnol. Prog. 12 (1996) 645.
[20] R. Hofweber, In-vitro translation und ENDOR spektroskopische Analyse von $p21^{Ras}$. Diploma Thesis, University of Regensburg (2003).
[21] C.J. Noren, S.J. Anthony-Cahill, M.C. Griffith, P.G. chultz, Science 244 (1989) 182.
[22] H.D. Ou, H.C. Lai, Z. Serber, V. Dotsch, J. Biomol. NMR 21 (2001) 269.
[23] P.M. Kane, C.T. Yamashiro, D.F. Wolczyk, N. Neff, M. Goebl, T.H. Stevens, Science 250 (1990) 651.
[24] M.W. Southworth, E. Adam, D. Panne, R. Byer, R. Kautz, F.B. Perler, EMBO J. 17 (1998) 918.
[25] F.B. Perler, Cell 92 (1998) 1.
[26] T. Yamazaki, T. Otomo, N. Oda, Y. Kyogoku, K. Uegaki, N. Ito, Y. Ishino, H. Nakamura, J. Am. Chem. Soc. 120 (1998) 5591.
[27] R. Xu, B. Ayers, D. Cowburn, T.W. Muir, Proc. Natl Acad. Sci. USA 96 (1999) 388.
[28] M. Sattler, J. Schleucher, C. Griesinger, Prog. NMR Spectrosc. 34 (1999) 93.
[29] D.S. Wishart, D.A. Case, Meth. Enzymol. 338 (2003) 3.

[30] T. Szyperski, B. Banecki, R.W. Glaser, J. Biomol. NMR 11 (1998) 387.
[31] Y. Xia, C.H. Arrowsmith, T. Szyperski, J. Biomol. NMR 24 (2002) 41.
[32] T. Szyperski, D.C. Yeh, D.K. Sukumaran, H.N. Moseley, G.T. Montelione, Proc. Natl Acad. Sci. USA 99 (2002) 8009.
[33] S. Kim, T. Szyperski, J. Am. Chem. Soc. 125 (2003) 1385.
[34] L. Frydman, T. Scherf, A. Lupulescu, Proc. Natl Acad. Sci. USA 99 (2002) 15858.
[35] M. Schubert, M. Smalla, P. Schmieder, H. Oschkinat, J. Magn. Reson. 141 (1999) 34.
[36] M. Schubert, H. Oschkinat, P. Schmieder, J. Magn. Reson. 148 (2001) 61.
[37] M. Schubert, H. Oschkinat, P. Schmieder, J. Magn. Reson. 153 (2001) 186.
[38] P. Schmieder, M. Leidert, M. Kelly, H. Oschkinat, J. Magn. Reson. 131 (1998) 199.
[39] V. Dötsch, G. Wagner, J. Magn. Reson. B 111 (1996) 310.
[40] V. Dötsch, H. Matsuo, G. Wagner, J. Magn. Reson. B 112 (1996) 95.
[41] C.B. Rios, W. Feng, M. Tashiro, Z. Shang, G. Montelione, J. Biomol. NMR 8 (1996) 345.
[42] W. Feng, C.B. Rios, G. Montelione, J. Biomol. NMR 8 (1996) 98.
[43] AZARA Program, Wayne Boucher, Department of Biochemistry, University of Cambridge (2002).
[44] DELTA Program, Jeol USA Inc., Peabody, MA (2003).
[45] FELIX Program, Accelrys Inc., San Diego, CA (2003).
[46] J.L. Pons, T.E. Malliavin, M.A. Delsuc, J. Biomol. NMR 8 (1996) 445.
[47] U.L. Günther, C. Ludwig, H. Rüterjans, J. Magn. Reson. 145 (2000) 201.
[48] F. Delagio, S. Grzesiek, G.W. Vuister, G. Zhu, J. Pfeifer, A. Bax, J. Biomol. NMR 6 (1995) 277.
[49] NMRZ Program, New Methods Research Inc., Syracuse, NY (2003).
[50] M. Kjaer, K.V. Andersen, F.M. Poulsen, Meth. Enzymol. 239 (1994) 288.
[51] P. Güntert, V. Dötsch, G. Wider, K. Wüthrich, J. Biomol. NMR 2 (1992) 619.
[52] TRIAD Program, Tripos Inc., St Louis, MO (2003).
[53] VNMR Program, Varian, Palo Alto, CA (2003).
[54] XWINNMR Program, Bruker, Biospin GmbH, Ettlingen (2003).
[55] E.J. Delikatny, W.E. Hull, C.E. Mountford, J. Magn. Reson. 94 (1991) 563.
[56] R.R. Ernst, G. Bodenhausen, A. Wokaun, Principles of Nuclear Magnetic Resonance in One and Two Dimensions, Clarendon Press, Oxford, 1987.
[57] A.G. Ferrige, J.C. Lindon, J. Magn. Reson. 31 (1978) 337.
[58] G.A. Pearson, J. Magn. Reson. 74 (1987) 541.
[59] J. Cavanagh, W.J. Fairbrother, A.G. Palmer III, N.J. Skelton, Protein NMR Spectroscopy Principles and Practice, Academic Press, San Diego, 1996.
[60] G. Otting, H. Widmer, G. Wagner, K. Wüthrich, J. Magn. Reson. 66 (1986) 187.
[61] A. Bax, M. Ikura, L.E. Kay, G. Zhu, J. Magn. Reson. 91 (1991) 174.
[62] M.A. Delsuc, J.-Y. Lallemand, J. Magn. Reson. 69 (1986) 504.
[63] G. Wider, J. Magn. Reson. 89 (1990) 406.
[64] A.G. Redfield, S.A. Kunz, J. Magn. Reson. A 108 (1994) 234.
[65] P. Koehl, Prog. NMR Spectrosc. 34 (1999) 257.
[66] P.L. Fortier, M.A. Delsuc, P. Kahn, J.-Y. Lallemand, J. Magn. Reson. 95 (1991) 161.
[67] Z. Zolani, S. Macura, J.L. Markley, J. Magn. Reson. 82 (1989) 496.
[68] I.L. Barsukov, A.S. Arseniev, J. Magn. Reson. 73 (1987) 148.
[69] R. Saffrich, W. Beneicke, K.-P. Neidig, H.R. Kalbitzer, J. Magn. Reson. B 101 (1993) 304.
[70] W. Dietrich, C.H. Rüdel, M. Neumann, J. Magn. Reson. 91 (1991) 1.
[71] S. Golotvin, A. Williams, J. Magn. Reson. 146 (2000) 122.
[72] P. Güntert, K. Wüthrich, J. Magn. Reson. 96 (1992) 403.
[73] R.E. Klevit, J. Magn. Reson. 62 (1985) 551.
[74] H.R. Kalbitzer, K.-P. Neidig, M. Geyer, R. Saffrich, M. Lorenz, in: J.C. Hoch (Ed.), Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy, Plenum Press, New York, 1991, p. 175.
[75] Z. Zolani, S. Macura, J.L. Markley, Comp. Enh. Spectrosc. 3 (1986) 141.
[76] Z. Zolani, S. Macura, J.L. Markley, J. Magn. Reson. 80 (1988) 60.
[77] D. Marion, A. Bax, J. Magn. Reson. 83 (1989) 205.
[78] J.J. Led, F. Abildgaard, H. Gesmar, J. Magn. Reson. 93 (1991) 659.
[79] L. Mitschang, K.-P. Neidig, H.R. Kalbitzer, J. Magn. Reson. 90 (1990) 359.
[80] M. Adler, G. Wagner, J. Magn. Reson. 91 (1991) 450.
[81] M.S. Friedrichs, W.J. Metzler, L. Mueller, J. Magn. Reson. 95 (1991) 178.
[82] D. Marion, M. Ikura, A. Bax, J. Magn. Reson. 84 (1989) 425.
[83] P. Tsang, P.E. Wright, M. Rance, J. Magn. Reson. 88 (1990) 210.
[84] Y. Kuroda, A. Wada, T. Yamazaki, K. Nagayama, J. Magn. Reson. 84 (1989) 604.
[85] L. Mitschang, C. Cieslar, T.A. Holak, H. Oschkinat, J. Magn. Reson. 92 (1991) 208.
[86] J.K. Hardy, P.L. Rinaldi, J. Magn. Reson. 88 (2003) 320.
[87] D.E. Brown, T.W. Campbell, J. Magn. Reson. 89 (1990) 255.
[88] D. Barache, J.-P. Antoine, J.-M. Dereppe, J. Magn. Reson. 128 (1997) 1.
[89] J.-P. Antoine, A. Coron, J.-M. Dereppe, J. Magn. Reson. 144 (2000) 189.
[90] U.L. Günther, C. Ludwig, H. Rüterjans, J. Magn. Reson. 156 (2002) 19.
[91] K. Stadlthanner, A.M. Tome, F.J. Theis, W. Gronwald, H.R. Kalbitzer, E.W. Lang, Proc. ICA2003 (2003) 167.
[92] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[93] P.H. Bolton, J. Magn. Reson. 64 (1985) 352.
[94] K.-P. Neidig, H.R. Kalbitzer, Magn. Reson. Chem. 26 (1988) 848.
[95] K.-P. Neidig, H.R. Kalbitzer, J. Magn. Reson. 91 (1991) 155.
[96] H. Gesmar, J.J. Led, F. Abildgaard, Prog. NMR Spectrosc. 22 (1990) 255.
[97] H. Barkhuijsen, R. De Beer, W.M.M.J. Bovee, D. van Ormondt, J. Magn. Reson. 61 (1985) 465.
[98] J. Tang, J.R. Norris, J. Chem. Phys. 84 (1986) 5210.
[99] P. Barone, L. Guidoni, E. Massaro, V. Vitti, J. Magn. Reson. 73 (1987) 23.
[100] J. Tang, J.R. Norris, J. Magn. Reson. 79 (1988) 190.
[101] J. Tang, J.R. Norris, J. Magn. Reson. 78 (1988) 23.
[102] D.S. Stephenson, Prog. NMR Spectrosc. 20 (1988) 515.
[103] H. Gesmar, J.J. Led, J. Magn. Reson. 83 (1989) 53.
[104] Y. Zeng, J. Tang, C.A. Bush, J.R. Norris, J. Magn. Reson. 83 (1989) 473.
[105] H. Yan, J.C. Gore, J. Magn. Reson. 88 (1990) 354.
[106] G. Zhu, A. Bax, J. Magn. Reson. 98 (1992) 192.
[107] S.M. Kristensen, M.D. Sörensen, H. Gesmar, J.J. Led, J. Magn. Reson. B 112 (1996) 193.
[108] J. Tang, J.R. Norris, J. Magn. Reson. 69 (1986) 180.
[109] A.E. Schussheim, D. Cowburn, J. Magn. Reson. 71 (1987) 371.
[110] C.F. Tirendi, J.F. Martin, J. Magn. Reson. 81 (1989) 577.
[111] C. Cieslar, T.A. Holak, H. Oschkinat, J. Magn. Reson. 89 (1990) 184.
[112] G. Zhu, A. Bax, J. Magn. Reson. 90 (1990) 405.
[113] A. Knijn, R. De Beer, D. van Ormondt, J. Magn. Reson. 97 (1992) 444.
[114] Z. Bi, A.P. Bruner, J. Li, K.N. Scott, Z.-S. Liu, C.B. Stopka, H.-W. Kim, D.C. Wilson, J. Magn. Reson. 140 (1999) 108.
[115] M.A. Delsuc, F. Ni, G.C. Levy, J. Magn. Reson. 73 (1987) 548.
[116] P. Mutzenhardt, J. Brondeau, F. Humbert, D. Canet, J. Magn. Reson. 94 (1991) 543.
[117] W.W.F. Pijnappel, A. van den Boogaart, R. De Beer, D. van Ormondt, J. Magn. Reson. 97 (1992) 122.

[118] R. Kumaresan, C.S. Ramalingam, D. van Ormondt, J. Magn. Reson. 89 (1990) 562.
[119] E.T. Jaynes, Proc. IEEE 70 (1982) 939.
[120] S. Sibisi, Nature 301 (1983) 134.
[121] S. Sibisi, J. Skilling, R.G. Bereton, E.D. Laue, J. Staunton, Nature 311 (1984) 446.
[122] J.C. Hoch, J. Magn. Reson. 64 (1985) 436.
[123] P.J. Hore, J. Magn. Reson. 62 (1985) 561.
[124] E.D. Laue, J. Skilling, J. Staunton, S. Sibisi, R.G. Bereton, J. Magn. Reson. 62 (1985) 437.
[125] E.D. Laue, J. Skilling, J. Staunton, J. Magn. Reson. 63 (1985) 418.
[126] J.F. Martin, J. Magn. Reson. 65 (1985) 291.
[127] E.T. Jaynes, in: C.R. Smith, W.T. Grandy Jr. (Eds.), Maximum-Entropy and Bayesian Methods in Inverse Problems, D. Reidel, Dordrecht, 1985, p. 21.
[128] J. Skilling, S.F. Gull, in: C.R. Smith, W.T. Grandy Jr. (Eds.), Maximum-Entropy and Bayesian Methods in Inverse Problems, D. Reidel, Dordrecht, 1985, p. 83.
[129] E.T. Jaynes, in: C.R. Smith, W.T. Grandy Jr. (Eds.), Maximum-Entropy and Bayesian Methods in Inverse Problems, D. Reidel, Dordrecht, 1985, p. 443.
[130] F. Ni, G.C. Levy, H.A. Sheraga, J. Magn. Reson. 66 (1986) 385.
[131] E.D. Laue, M.R. Mayger, J. Skilling, J. Staunton, J. Magn. Reson. 68 (1986) 14.
[132] K.M. Wright, P.S. Belton, Mol. Phys. 58 (1986) 485.
[133] E.D. Laue, K.O.B. Pollard, J. Skilling, J. Staunton, A.C. Sutkowski, J. Magn. Reson. 72 (1987) 493.
[134] J.C.J. Barna, E.D. Laue, M.R. Mayger, J. Skilling, S.J.P. Worrall, J. Magn. Reson. 73 (1987) 69.
[135] M.L. Waller, P.S. Tofts, Magn. Reson. Med. 4 (1987) 385.
[136] J.C.J. Barna, E.D. Laue, J. Magn. Reson. 75 (1987) 384.
[137] J.C.J. Barna, S.M. Tan, E.D. Laue, J. Magn. Reson. 78 (1988) 327.
[138] S.J. Davies, C. Bauer, P.J. Hore, R. Freeman, J. Magn. Reson. 76 (1988) 476.
[139] R.H. Newman, J. Magn. Reson. 79 (1988) 448.
[140] G.L. Bretthorst, C.-C. Hung, D.A. D'Avignon, J.J.H. Ackerman, J. Magn. Reson. 79 (1988) 369.
[141] M.A. Delsuc, G.C. Levy, J. Magn. Reson. 76 (1988) 306.
[142] F. Ni, H.A. Sheraga, J. Magn. Reson. 82 (1989) 413.
[143] A. Heuer, U. Haeberlen, J. Magn. Reson. 85 (1989) 79.
[144] A.R. Mazzeo, M.A. Delsuc, A. Kumar, G.C. Levy, J. Magn. Reson. 81 (1989) 512.
[145] G.J. Daniell, P.J. Hore, J. Magn. Reson. 84 (1989) 515.
[146] P.J. Hore, D.S. Grainger, S. Wimperis, G.J. Daniell, J. Magn. Reson. 89 (1990) 415.
[147] D.L. Donoho, I.M. Johnstone, A.S. Stern, J.C. Hoch, Proc. Natl Acad. Sci. USA 87 (1990) 5066.
[148] J.C. Hoch, A.S. Stern, D.L. Donoho, I.M. Johnstone, J. Magn. Reson. 86 (1990) 236.
[149] G.L. Bretthorst, J. Magn. Reson. 93 (1991) 369.
[150] J.A. Jones, P.J. Hore, J. Magn. Reson. 92 (1991) 363.
[151] J.A. Jones, P.J. Hore, J. Magn. Reson. 92 (1991) 276.
[152] M. Robin, M.A. Delsuc, E. Guitettet, J.-Y. Lallemand, J. Magn. Reson. 92 (1991) 645.
[153] P. Schmieder, A.S. Stern, G. Wagner, J.C. Hoch, J. Biomol. NMR 3 (1993) 569.
[154] P. Schmieder, A.S. Stern, G. Wagner, J.C. Hoch, J. Biomol. NMR 4 (1994) 483.
[155] P. Schmieder, A.S. Stern, G. Wagner, J.C. Hoch, J. Magn. Reson. 125 (1997) 332.
[156] K.-B. Li, A.S. Stern, J.C. Hoch, J. Magn. Reson. 134 (1998) 161.
[157] N. Shimba, A.S. Stern, C.S. Craik, J.C. Hoch, V. Dötsch, J. Am. Chem. Soc. 125 (2003) 2382.
[158] M. Andrec, J.H. Prestegard, J. Magn. Reson. 130 (1998) 217.
[159] R.E. Hoffman, A. Kumar, K.D. Bishop, P.N. Borer, G.C. Levy, J. Magn. Reson. 83 (1989) 586.
[160] G. Zhu, W.Y. Choy, B.C. Sanctuary, J. Magn. Reson. 135 (1998) 37.
[161] R.A. Chylla, J.L. Markley, J. Biomol. NMR 5 (1995) 245.
[162] F.S. DiGennaro, D. Cowburn, J. Magn. Reson. 96 (1992) 582.
[163] A.S. Stern, K.-B. Li, J.C. Hoch, J. Am. Chem. Soc. 124 (2002) 1982.
[164] V.A. Mandelshtam, J. Magn. Reson. 144 (2000) 343.
[165] H. Hu, A.A. De Angelis, V.A. Mandelshtam, A.J. Shaka, J. Magn. Reson. 144 (2000) 357.
[166] W. Beneicke, H.R. Kalbitzer, J. Magn. Reson. A 107 (1994) 134.
[167] V.Y. Orekhov, I.V. Ibraghimov, M. Billeter, J. Biomol. NMR 20 (2001) 49.
[168] D.H. Live, D.G. Davis, W.C. Agosta, D. Cowburn, J. Am. Chem. Soc. 106 (1984) 1939.
[169] T. Maurer, H.R. Kalbitzer, J. Magn. Reson. B 113 (1996) 177.
[170] J.L. Markley, A. Bax, Y. Arata, C.W. Hilbers, R. Kaptein, B.D. Sykes, P.E. Wright, K. Wüthrich, Pure Appl. Chem. 70 (1998) 117.
[171] K.-P. Neidig, H. Bodenmüller, H.R. Kalbitzer, Biochem. Biophys. Res. Commun. 125 (1984) 1143.
[172] S. Glaser, H.R. Kalbitzer, J. Magn. Reson. 74 (1987) 450.
[173] P. Pfändler, G. Bodenhausen, Magn. Reson. Chem. 26 (1988) 888.
[174] M. Novic, U. Eggenberger, G. Bodenhausen, J. Magn. Reson. 77 (1988) 394.
[175] V. Stoven, A. Mikou, D. Piveteau, E. Guitettet, J.-Y. Lallemand, J. Magn. Reson. 82 (1989) 163.
[176] C. Eccles, P. Guntert, M. Billeter, K. Wuthrich, J. Biomol. NMR 1 (1991) 111.
[177] T. Herrmann, P. Güntert, K. Wüthrich, J. Biomol. NMR 24 (2002) 171.
[178] K.-P. Neidig, H.R. Kalbitzer, J. Magn. Reson. 88 (1990) 155.
[179] G.J. Kleywegt, R. Boelens, R. Kaptein, J. Magn. Reson. 88 (1990) 601.
[180] D.S. Garrett, R. Powers, A.M. Gronenborn, G.M. Clore, J. Magn. Reson. 95 (1991) 214.
[181] R. Koradi, M. Billeter, M. Engeli, P. Guntert, K. Wuthrich, J. Magn. Reson. 135 (1998) 288.
[182] C. Antz, K.-P. Neidig, H.R. Kalbitzer, J. Biomol. NMR 5 (1995) 287.
[183] A.C. Schulte, A. Gorler, C. Antz, K.P. Neidig, H.R. Kalbitzer, J. Magn. Reson. 129 (1997) 165.
[184] A. Rough, M.A. Delsuc, G. Bertrand, J.-Y. Lallemand, J. Magn. Reson. A 102 (1993) 357.
[185] E. Brunner, J. Ogle, M. Wenzler, H.R. Kalbitzer, Biochem. Biophys. Res. Commun. 272 (2000) 694.
[186] A. Bax, G. Kontaxis, N.L. Tjandra, Meth. Enzymol. 339 (2001) 127.
[187] C. Cieslar, G.M. Clore, A.M. Gronenborn, J. Magn. Reson. 80 (1988) 119.
[188] Z. Madi, B.U. Meier, R.R. Ernst, J. Magn. Reson. 72 (1987) 584.
[189] K.-P. Neidig, R. Saffrich, M. Lorenz, H.R. Kalbitzer, J. Magn. Reson. 89 (1990) 543.
[190] H.R. Kalbitzer, in: W.R. Croasmun, R. Carlson (Eds.), Two-Dimensional NMR-Spectroscopy: Applications for Chemists and Biochemists, Verlag Chemie, Weinheim, 1994, p. 581.
[191] M. Kjaer, F.M. Poulsen, J. Magn. Reson. 94 (1991) 659.
[192] P.H. Bolton, J. Magn. Reson. 70 (1986) 344.
[193] B.U. Meier, Z. Madi, R.R. Ernst, J. Magn. Reson. 74 (1987) 565.
[194] J.C. Hoch, S. Hengyi, M. Kjaer, S. Ludvigsen, F.M. Poulsen, Carlsberg Res. Commun. 52 (1987) 111.
[195] B.U. Meier, R.R. Ernst, J. Magn. Reson. 79 (1988) 540.
[196] M. Novic, G. Bodenhausen, Anal. Chem. 60 (1988) 582.
[197] H. Shen, S. Ludvigsen, F.M. Poulsen, J. Magn. Reson. 90 (1990) 346.
[198] H. Shen, F.M. Poulsen, J. Magn. Reson. 97 (1992) 385.
[199] J. Xu, B.C. Sanctuary, B.N. Gray, J. Chem. Inf. Comput. Sci. 33 (1993) 475.
[200] P. Pfändler, G. Bodenhausen, B.U. Meier, R.R. Ernst, Anal. Chem. 57 (1985) 2510.
[201] P. Pfändler, G. Bodenhausen, J. Magn. Reson. 70 (1986) 71.
[202] P. Pfändler, G. Bodenhausen, J. Magn. Reson. 79 (1988) 99.
[203] B.U. Meier, G. Bodenhausen, R.R. Ernst, J. Magn. Reson. 60 (1984) 161.

[204] Z. Madi, R.R. Ernst, J. Magn. Reson. 79 (1988) 513.
[205] H. Shen, F.M. Poulsen, J. Magn. Reson. 89 (1990) 585.
[206] M. Geyer, K.-P. Neidig, H.R. Kalbitzer, J. Magn. Reson. B 109 (1995) 31.
[207] W. Denk, R. Baumann, G. Wagner, J. Magn. Reson. 67 (1986) 386.
[208] G.H. Weiss, J.E. Kiefer, J.A. Ferretti, J. Magn. Reson. 97 (1992) 227.
[209] G. Wagner, K. Wüthrich, J. Mol. Biol. 155 (1982) 347.
[210] G.J. Kleywegt, R. Boelens, M. Cox, M. Llinas, R. Kaptein, J. Biomol. NMR 1 (1991) 23.
[211] K. Wüthrich, NMR of Proteins and Nucleic Acids, Wiley, New York, 1986.
[212] S.W. Englander, A.J. Wand, Biochemistry 26 (1987) 5953.
[213] S.J. Nelson, D.M. Schneider, A.J. Wand, Biophys. J. 59 (1991) 1113.
[214] A.J. Wand, S.J. Nelson, Biophys. J. 59 (1991) 1101.
[215] B.C.M. Potts, W.J. Chazin, J. Biomol. NMR 11 (1998) 45.
[216] B.R. Seavey, E.A. Farr, W.M. Westler, J.L. Markley, J. Biomol. NMR 1 (1991) 217.
[217] W. Gronwald, R.F. Boyko, F.D. Sönnichsen, D.S. Wishart, B.D. Sykes, J. Biomol. NMR 10 (1997) 165.
[218] D.S. Wishart, B.D. Sykes, F.M. Richards, J. Mol. Biol. 222 (1991) 311.
[219] D.S. Wishart, R.F. Boyko, B.D. Sykes, CABIOS 10 (1994) 687.
[220] D.S. Wishart, M.S. Watson, R.F. Boyko, B.D. Sykes, J. Biomol. NMR 10 (1997) 329.
[221] M. Iwadate, T. Asakura, M.P. Williamson, J. Biomol. NMR 13 (1999) 199.
[222] K. Ösapay, D.A. Case, J. Am. Chem. Soc. 113 (1991) 9436.
[223] M.P. Williamson, T. Asakura, J. Magn. Reson. 94 (1991) 557.
[224] M.P. Williamson, T. Asakura, J. Magn. Reson. B 101 (1993) 63.
[225] K. Ösapay, D.A. Case, J. Biomol. NMR 4 (1994) 215.
[226] A.C. de Dios, J.G. Pearson, E. Oldfield, Science 260 (1993) 1491.
[227] E. Oldfield, J. Biomol. NMR 5 (1995) 217.
[228] A.C. de Dios, Prog. NMR Spectrosc. 29 (1996) 229.
[229] D. Sitkoff, D.A. Case, J. Am. Chem. Soc. 119 (1997) 12262.
[230] I. Ando, T. Kameda, N. Asakawa, S. Kuroki, H. Kurosu, J. Mol. Struct. 441 (1998) 213.
[231] X.P. Xu, D.A. Case, J. Biomol. NMR 21 (2001) 321.
[232] J. Meiler, J. Biomol. NMR 26 (2003) 25.
[233] W. Gronwald, S. Moussa, R. Elsner, A. Jung, B. Ganslmeier, J. Trenner, W. Kremer, K.P. Neidig, H.R. Kalbitzer, J. Biomol. NMR 23 (2002) 271.
[234] M. Nilges, J. Mol. Biol. 245 (1995) 645.
[235] M. Nilges, M.J. Macias, S.I. O'Donoghue, H. Oschkinat, J. Mol. Biol. 269 (1997) 408.
[236] J.P. Linge, S.I. O'Donoghue, M. Nilges, Meth. Enzymol. 339 (2001) 71.
[237] J. Boisbouvier, M. Blackledge, A. Sollier, D. Marion, J. Biomol. NMR 16 (2000) 197.
[238] B.J. Hare, G. Wagner, J. Biomol. NMR 15 (2002) 103.
[239] C. Dominguez, R. Boelens, A.M.J.J. Bonvin, J. Am. Chem. Soc. 125 (2003) 1731.
[240] B.M. Duggan, G.B. Legge, H.J. Dyson, P.E. Wright, J. Biomol. NMR 19 (2001) 321.
[241] P. Savarin, S. Zinn-Justin, B. Gilquin, J. Biomol. NMR 19 (2001) 49.
[242] C. Mumenthaler, W. Braun, J. Mol. Biol. 254 (1995) 465.
[243] Y. Xu, J. Wu, D. Gorenstein, W. Braun, J. Magn. Reson. 136 (1999) 76.
[244] Y. Xu, M.J. Jablonsky, P.L. Jackson, W. Braun, N.R. Krishna, J. Magn. Reson. 148 (2001) 35.
[245] N. Oezguen, L. Adamian, Y. Xu, K. Rajarathnam, W. Braun, J. Biomol. NMR 22 (2002) 249.
[246] C. Mumenthaler, P. Guntert, W. Braun, K. Wuthrich, J. Biomol. NMR 10 (1997) 351.
[247] T. Herrmann, P. Guntert, K. Wuthrich, J. Mol. Biol. 319 (2002) 209.
[248] T. Maurer, R. Döker, A. Görler, W. Hengstenberg, H.R. Kalbitzer, Eur. J. Biochem. 268 (2001) 635.
[249] J.W. Keepers, T.L. James, J. Magn. Reson. 57 (1984) 404.
[250] K.M. Banks, D.R. Hare, B.R. Reid, Biochemistry 28 (1989) 6996.
[251] R. Boelens, M.G. Koning, G.A. van der Marel, J.H. van Boom, R. Kaptein, J. Magn. Reson. 82 (1989) 290.
[252] C.B. Post, R.P. Meadows, D.G. Gorenstein, J. Am. Chem. Soc. 112 (1990) 6796.
[253] B.A. Borgias, T.L. James, J. Magn. Reson. 87 (1990) 475.
[254] A.M.J.J. Bonvin, R. Boelens, R. Kaptein, J. Biomol. NMR 1 (1991) 305.
[255] M. Madrid, E. Llinas, M. Llinas, J. Magn. Reson. 93 (1991) 329.
[256] F.J.M. van de Ven, M.J.J. Blommers, R.E. Schouten, C.W. Hilbers, J. Magn. Reson. 94 (1991) 140.
[257] S.-G. Kim, B.R. Reid, J. Magn. Reson. 100 (1992) 382.
[258] A.T. Brünger, X-PLOR Version 3.1, Yale University Press, New Haven/London, 1992.
[259] L. Zhu, B.R. Reid, J. Magn. Reson. B 106 (1995) 227.
[260] A. Görler, H.R. Kalbitzer, J. Magn. Reson. 124 (1997) 177.
[261] A. Gorler, W. Gronwald, K.P. Neidig, H.R. Kalbitzer, J. Magn. Reson. 137 (1999) 39.
[262] L. Zhu, H.J. Dyson, P.E. Wright, J. Biomol. NMR 11 (1998) 17.
[263] S. Shibata, K. Akasaka, Magn. Reson. Chem. 28 (1990) 129.
[264] K.-P. Neidig, M. Geyer, A. Görler, C. Antz, R. Saffrich, W. Beneicke, H.R. Kalbitzer, J. Biomol. NMR 6 (1995) 255.
[265] H.N.B. Moseley, E.V. Curto, N.R. Krishna, J. Magn. Reson. B 108 (1995) 243.
[266] J. Jeener, B.H. Meier, P. Bachmann, R.R. Ernst, J. Chem. Phys. 71 (1979) 4546.
[267] S. Macura, R.R. Ernst, Mol. Phys. 41 (1980) 95.
[268] J.M. Schurr, H.P. Babcock, B.S. Fujimoto, J. Magn. Reson. B 105 (1994) 211.
[269] M.M. Tirado, J.G. de la Torre, J. Chem. Phys. 73 (1980) 1986.
[270] H. Liu, D.L. Banville, V.J. Basus, T.L. James, J. Magn. Reson. B 107 (1995) 51.
[271] H. Liu, M. Tonelli, T.L. James, J. Magn. Reson. B 111 (1996) 85.
[272] W.F. van Gunsteren, P.H. Hünenberger, A.E. Mark, P.E. Smith, I.G. Tironi, Computer Phys. Commun. 91 (1995) 305.
[273] D.S. Wishart, B.D. Sykes, F.M. Richards, Biochemistry 31 (1992) 1647.
[274] D.S. Wishart, B.D. Sykes, J. Biomol. NMR 4 (1994) 171.
[275] Y. Wang, O. Jardetzky, Protein Sci. 11 (2002) 852.
[276] L.-H. Hung, R. Samudrala, Protein Sci. 12 (2003) 288.
[277] R.D. Beger, P.H. Bolton, J. Biomol. NMR 10 (1997) 129.
[278] G. Cornilescu, F. Delagio, A. Bax, J. Biomol. NMR 13 (1999) 289.
[279] M. Schubert, D. Labudde, H. Oschkinat, P. Schmieder, J. Biomol. NMR 24 (2002) 149.
[280] P. Padrata, V. Sklenar, J. Biomol. NMR 24 (2002) 339.
[281] K. Kloiber, W. Schuler, R. Konrat, J. Biomol. NMR 22 (2002) 349.
[282] V.I. Polshakov, T.A. Frenkiel, B. Birdsall, J. Feeney, J. Magn. Reson. B108 (1995) 31.
[283] P. Güntert, M. Billeter, O. Ohlenschläger, L.R. Brown, K. Wüthrich, J. Biomol. NMR 12 (1998) 543.
[284] R. Tejero, D. Monleon, B. Celda, R. Powers, G.T. Montelione, J. Biomol. NMR 15 (1999) 251.
[285] P. Güntert, W. Braun, K. Wüthrich, J. Mol. Biol. 217 (1991) 517.
[286] W. Smith, T.R. Forester, J. Mol. Graphics 14 (1996) 136.
[287] P. Güntert, C. Mumenthaler, K. Wüthrich, J. Mol. Biol. 273 (1997) 283.
[288] A.T. Brünger, P.D. Adams, G.M. Clore, W.L. DeLano, P. Gros, R.W. Grossekunstleve, J.-S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson, G.L. Warren, Acta Cryst. D54 (1998) 905.
[289] L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, K. Schulten, J. Comput. Phys. 151 (1999) 283.
[290] D.A. Case, D.A. Pearlman, J.W. Caldwell, T.E. Cheatham III, J. Wang, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R.J. Radmer, Y. Duan, J. Pitera, I. Massova, G.L. Seibel, U.C. Singh,

P.K. Weiner, P.A. Kollman, AMBER 7 Program, University of California, San Francisco, CA, 2002.
[291] C.D. Schwieters, J. Kuszewski, N.L. Tjandra, G.M. Clore, J. Magn. Reson. 160 (2003) 65.
[292] InsightII Program, Accelrys Inc., San Diego, CA (2003).
[293] M.J. Sippl, J. Mol. Biol. 213 (1990) 859.
[294] T. Hansson, C. Oostenbrink, W.F. van Gunsteren, Curr. Opin. Struct. Biol. 12 (2002) 190.
[295] G.P. Gippert, P.E. Wright, D.A. Case, J. Biomol. NMR 11 (1998) 241.
[296] A. Annila, H. Aito, E. Thulin, T. Drakenberg, J. Biomol. NMR 14 (1999) 223.
[297] P.M. Bowers, C.E.M. Strauss, D. Baker, J. Biomol. NMR 18 (2000) 311.
[298] K.T. Simons, C. Kooperberg, E. Huang, D. Baker, J. Mol. Biol. 268 (1997) 209.
[299] K.T. Simons, I. Ruczinski, C. Kooperberg, B.A. Fox, C. Bystroff, D. Baker, Proteins 34 (1999) 82.
[300] F. Delagio, G. Kontaxis, A. Bax, J. Am. Chem. Soc. 122 (2000) 2142.
[301] M. Andrec, Y. Harano, M.P. Jacobson, R.A. Friesner, R.M. Levy, J. Struct. Funct. Genomics 2 (2002) 103.
[302] M. Albrecht, D. Hanisch, R. Zimmer, T. Lengauer, Silico Biol. 2 (2002) 1.
[303] A.M.J.J. Bonvin, K. Houben, M. Guenneugues, R. Kaptein, R. Boelens, J. Biomol. NMR 21 (2001) 221.
[304] A.W. Giesen, S.W. Homans, J.M. Brown, J. Biomol. NMR 25 (2003) 63.
[305] L.M. Koharudin, A.M.J.J. Bonvin, R. Kaptein, R. Boelens, J. Magn. Reson. 163 (2003) 228.
[306] D. Zheng, Y. Huang, H.N. Moseley, R. Xiao, J. Aramini, G.V.T. Swapna, G.T. Montelione, Protein Sci. 12 (2003) 1232.
[307] R.A. Laskowski, M.W. MacArthur, J.M. Thornton, Curr. Opin. Struct. Biol. 8 (1998) 631.
[308] A.T. Brünger, R.L. Campbell, G.M. Clore, A.M. Gronenborn, M. Karplus, G.A. Petsko, M.M. Teeter, Science 235 (1987) 1049.
[309] J.F. Doreleijers, M.L. Raves, T. Rullmann, R. Kaptein, J. Biomol. NMR 14 (1999) 123.
[310] G. Gupta, M.H. Sarma, R.H. Sarma, Biochemistry 27 (1988) 7909.
[311] A.N. Lane, Biochim. Biophys. Acta 1049 (1990) 189.
[312] C. Gonzalez, J.A.C. Rullmann, A.M.J.J. Bonvin, R. Boelens, R. Kaptein, J. Magn. Reson. 91 (1991) 659.
[313] P.D. Thomas, V.J. Basus, T.L. James, Proc. Natl Acad. Sci. USA 88 (1991) 1237.
[314] Y. Xu, I.P. Sugar, N.R. Krishna, J. Biomol. NMR 5 (1995) 37.
[315] M. Nilges, J. Habazettl, A.T. Brünger, T.A. Holak, J. Mol. Biol. 219 (1991) 499.
[316] J.-F. Lefevre, A.N. Lane, O. Jardetzky, Biochemistry 26 (1987) 5076.
[317] A.T. Brünger, G.M. Clore, A. Gronenborn, R. Saffrich, M. Nilges, Science 261 (1993) 328.
[318] G.M. Clore, M.A. Robien, A. Gronenborn, J. Mol. Biol. 231 (2002) 82.
[319] J.D. Baleja, J. Moult, B.D. Sykes, J. Magn. Reson. 87 (1990) 375.
[320] B.A. Borgias, M. Gochin, D.J. Kerwood, T.L. James, Prog. NMR Spectrosc. 22 (1990) 83.
[321] E.P. Nikonowicz, R.P. Meadows, D.G. Gorenstein, Biochemistry 29 (1990) 4193.
[322] J.E. Mertz, P. Güntert, K. Wüthrich, W. Braun, J. Biomol. NMR 1 (1991) 257.
[323] W. Gronwald, R. Kirchhofer, A. Gorler, W. Kremer, B. Ganslmeier, K.P. Neidig, H.R. Kalbitzer, J. Biomol. NMR 17 (2000) 137.
[324] G.M. Clore, D.S. Garrett, J. Am. Chem. Soc. 121 (1999) 9008.
[325] R.A. Laskowski, J.A.C. Rullmann, M.W. MacArthur, R. Kaptein, J.M. Thornton, J. Biomol. NMR 8 (1996) 477.
[326] D.A. Pearlman, J. Biomol. NMR 8 (1996) 49.
[327] D.A. Pearlman, J. Biomol. NMR 13 (1999) 325.
[328] A.L. Morris, M.W. MacArthur, E.G. Hutchinson, J.M. Thornton, Proteins 12 (1992) 345.
[329] R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola, Nature 381 (2003) 272.
[330] M.J. Sippl, Proteins 17 (1993) 355.
[331] J. Westbrook, Z. Feng, L. Chen, H. Yang, H.M. Berman, Nucleic Acids Res. 31 (2003) 489.
[332] R. Wimmer, N. Müller, S.B. Petersen, J. Biomol. NMR 9 (1997) 101.
[333] M.C. Baran, H.N.B. Moseley, G. Sahota, G.T. Montelione, J. Biomol. NMR 24 (2002) 113.
[334] N. Morelle, B. Brutscher, J.-P. Simorre, D. Marion, J. Biomol. NMR 5 (1995) 154.
[335] S.G. Hyberts, G. Wagner, J. Biomol. NMR 26 (2003) 335.
[336] B.E. Coggins, P. Zhou, J. Biomol. NMR 26 (2003) 93.
[337] M.A.C. Reed, A.M. Hounslow, K.H. Sze, I.L. Barsukov, L.L.P. Hosszu, A.R. Clarke, C.J. Craven, J.P. Waltho, J. Mol. Biol. 330 (2003) 1189.
[338] M.S. Friedrichs, L. Mueller, M. Wittekind, J. Biomol. NMR 4 (1994) 703.
[339] J.A. Lukin, A.P. Gove, S.N. Talukdar, C. Ho, J. Biomol. NMR 9 (1997) 151.
[340] D.E. Zimmerman, C.A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, G.T. Montelione, J. Mol. Biol. 269 (1997) 592.
[341] M. Leutner, R.M. Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, H. Kessler, J. Biomol. NMR 11 (1998) 31.
[342] H.S. Atreya, S.C. Sahu, K.V. Chary, G. Govil, J. Biomol. NMR 17 (2000) 125.
[343] T.E. Malliavin, P. Barthe, M.A. Delsuc, Theor. Chem. Acc. 106 (2001) 91.
[344] A. Grishaev, M. Llinas, J. Biomol. NMR 24 (2002) 203.
[345] T.K. Hitchens, J.A. Lukin, Y. Zhan, S.A. McCallum, G.S. Rule, J. Biomol. NMR 25 (2003) 1.
[346] H. Oschkinat, T.A. Holak, C. Cieslar, Biopolymers 31 (1991) 699.
[347] C. Bartels, M. Billeter, P. Güntert, K. Wüthrich, J. Biomol. NMR 7 (1996) 207.
[348] C. Bartels, P. Güntert, M. Billeter, K. Wüthrich, J. Comput. Chem. 18 (1997) 139.
[349] K.B. Li, B.C. Sanctuary, J. Chem. Inf. Comput. Sci. 37 (1997) 467.
[350] D. Croft, J. Kemmink, K.-P. Neidig, H. Oschkinat, J. Biomol. NMR 10 (1997) 207.
[351] W. Gronwald, L. Willard, T. Jellard, R.F. Boyko, K. Rajarathnam, D.S. Wishart, F.D. Sönnichsen, B.D. Sykes, J. Biomol. NMR 12 (1998) 395.
[352] C. Bigam, T. Jellard, W. Gronwald, B.D. Sykes, Magn. Moments 9 (1998) 9.
[353] K.-H. Groß, H.R. Kalbitzer, J. Magn. Reson. 76 (1987) 87.
[354] G.J. Kleywegt, R.M.J.N. Lamerichs, R. Boelens, R. Kaptein, J. Magn. Reson. 85 (1989) 186.
[355] J. Xu, B.C. Sanctuary, J. Chem. Inf. Comput. Sci. 33 (1993) 491.
[356] J. Xu, S.K. Straus, B.C. Sanctuary, L. Trimble, J. Magn. Reson. B 103 (1994) 53.
[357] K.B. Li, B.C. Sanctuary, J. Chem. Inf. Comput. Sci. 37 (1997) 359.
[358] B.J. Hare, J.H. Prestegard, J. Biomol. NMR 4 (1994) 35.
[359] K. Huang, M. Andrec, S. Heald, P. Blake, J.H. Prestegard, J. Biomol. NMR 10 (1997) 45.
[360] J.L. Pons, M.A. Delsuc, J. Biomol. NMR 15 (1999) 15.
[361] W.Y. Choy, B.C. Sanctuary, G. Zhu, J. Chem. Inf. Comput. Sci. 37 (1997) 1086.
[362] J.-C. Hus, J.J. Prompers, R. Brüschweiler, J. Magn. Reson. 157 (2002) 119.
[363] R.P. Meadows, E.T. Olejniczak, S.W. Fesik, J. Biomol. NMR 4 (1994) 79.
[364] M. Andrec, R.M. Levy, J. Biomol. NMR 23 (2002) 263.
[365] D. Labudde, D. Leitner, M. Krüger, H. Oschkinat, J. Biomol. NMR 25 (2003) 41.
[366] J. Xu, S.K. Straus, B.C. Sanctuary, L. Trimble, J. Chem. Inf. Comput. Sci. 33 (1993) 668.
[367] J.B. Olson Jr., J.L. Markley, J. Biomol. NMR 4 (1994) 385.

[368] R. Bernstein, C. Cieslar, A. Ross, H. Oschkinat, J. Freund, T.A. Holak, J. Biomol. NMR 3 (1993) 245.
[369] N.E.G. Buchler, E.R.P. Zuiderweg, H. Wang, R.A. Goldstein, J. Magn. Reson. 125 (1997) 34.
[370] P. Pristovsek, H. Ruterjans, R. Jerala, J. Comput. Chem. 23 (2002) 335.
[371] F.J. van de Ven, J. Magn. Reson. 86 (1990) 633.
[372] M. Billeter, V.J. Basus, I.D. Kuntz, J. Magn. Reson. 76 (1988) 400.
[373] C.D. Eads, I.D. Kuntz, J. Magn. Reson. 82 (1989) 467.
[374] P. Güntert, M. Salzmann, D. Braun, K. Wüthrich, J. Biomol. NMR 18 (2000) 129.
[375] A. D'Ursi, H. Oschkinat, C. Cieslar, D. Picone, G. D'Alessio, P. Amodeo, P.A. Temussi, Eur. J. Biochem. 229 (1995) 494.
[376] H. Oschkinat, C. Cieslar, C. Grisinger, J. Magn. Reson. 86 (1990) 453.
[377] J. Xu, P.L. Weber, P.N. Borer, J. Biomol. NMR 5 (1995) 183.
[378] D. Zimmerman, C. Kulikowski, L. Wang, B. Lyons, G.T. Montelione, J. Biomol. NMR 4 (1994) 241.
[379] S. Shimotakahara, C.B. Rios, J.H. Laity, D.E. Zimmerman, H.A. Scheraga, G.T. Montelione, Biochemistry 36 (1997) 6915.
[380] J.F. O'Connell, K.D. Pryor, S.K. Grant, B. Leiting, J. Biomol. NMR 13 (1999) 311.
[381] D. Malmodin, C.H.M. Papavoine, M. Billeter, J. Biomol. NMR 27 (2003) 69.
[382] M. Helgstrand, P. Kraulis, P. Allard, T. Hard, J. Biomol. NMR 18 (2000) 329.
[383] P.J. Kraulis, J. Magn. Reson. 84 (1989) 627.
[384] T.E. Malliavin, J.L. Pons, M.A. Delsuc, Bioinformatics 14 (1998) 624.
[385] T.D. Goddard, D.G. Kneller, SPARKY 3 Program, University of California, San Francisco, CA (2003).
[386] B. Johnson, R.A. Blevins, J. Biomol. NMR 4 (1994) 603.
[387] C.M. Oshiro, I.D. Kuntz, Biopolymers 33 (1993) 107.
[388] P.J. Kraulis, J. Mol. Biol. 243 (1994) 696.
[389] A. Grishaev, M. Llinas, Proc. Natl Acad. Sci. USA 99 (2002) 6707.
[390] A. Grishaev, M. Llinas, Proc. Natl Acad. Sci. USA 99 (2002) 6713.
[391] M. Nilges, S.I. O'Donoghue, Prog. NMR Spectrosc. 32 (1998) 107.
[392] J. Cornfield, Biometrics 25 (1969) 617.
[393] S. Subramaniam, D.K. Tcheng, J.M. Fenton, ISMB 96 (1996) 218.
[394] B. Rost, R. Schneider, C. Sander, J. Mol. Biol. 270 (1997) 471.
[395] W. Kremer, B. Schuler, S. Harrieder, M. Geyer, W. Gronwald, C. Welker, R. Jaenicke, H.R. Kalbitzer, Eur. J. Biochem. 268 (2001) 2527.
[396] M. Geyer, C. Herrmann, S. Wohlgemuth, A. Wittinghofer, H.R. Kalbitzer, Nat. Struct. Biol. 4 (1997) 694.