ELSEVIER

Progress in Nuclear Magnetic Resonance Spectroscopy 43 (2003) 105–125

PROGRESS IN NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

www.elsevier.com/locate/pnmrs

Automated NMR protein structure calculation

Peter Güntert*

RIKEN Genomic Sciences Center, 1-7-22 Suehiro, Tsurumi, Yokohama 230-0045, Japan

Accepted 23 June 2003

Contents

* Tel.: +81-45-503-9345; fax: +81-45-503-9343.
E-mail address: guentert@gsc.riken.go.jp (P. Güntert).

doi:10.1016/S0079-6565(03)00021-9

Keywords: Protein structure; Chemical shift assignment; Conformational constraints; Automated structure determination; Automated assignment

1. Introduction

The NMR method for protein structure determination in solution is now firmly established besides X-ray crystallography as a second generally applicable technique that can give a detailed picture of the three-dimensional structure of biological macromolecules at atomic resolution. By April 2003, more than 3150 (15%) of the entries deposited in the Protein Data Bank [1] originated from macromolecular structures that had been solved by NMR methods. NMR plays also an important role in the current efforts of structural genomics that are driven by the vision to supplement the knowledge on the sequence of proteins by structural information on a genome-wide scale, determined either experimentally or by theoretical homology modeling [2]. Structural genomics wants to help us understand the molecular 'book of life', the genome, by translating its concise but cryptic DNA or amino acid sequence idiom into the more readily comprehensible language of three-dimensional structures. A massive structure determination effort will be needed to achieve the aim of structural genomics, since of the order of 10^5 new protein structures need to be determined experimentally [3] in order to allow coverage of the rest of sequence space with structures from theoretical methods because at present homology modeling is reliable only for proteins that share high (more than 30%) sequence identity with a protein of known three-dimensional structure.

Until recently NMR protein structure determination has remained a laborious undertaking that occupied a trained spectroscopist over several months for each new protein structure. It has been recognized that many of the time-consuming interactive steps carried out by an expert during the process of spectral analysis could be accomplished by automated, computational approaches [4]. Today automated methods for NMR structure determination are playing a more and more prominent role and will most likely supersede the conventional manual approaches to solving three-dimensional protein structures in solution.

This review gives an introduction to the current state of automated NMR structure calculation. Section 2 gives a general survey of the principles and problems of automated NOESY assignment and structure calculation. Section 3 is devoted to various specific implementations of algorithms for automated NOESY assignment and structure calculation. Aspects of reliability, quality control and troubleshooting in automated NMR structure calculation are discussed in Section 4. Alternative methods for structure calculation without chemical shift assignment are introduced in Section 5. In the three core Sections 3–5 a selection of programs is presented for which either the literature bears testimony of widespread use or that embody concepts of particular interest and future potential.

For consistency and simplicity, the following conventions are used in this review: an interaction between two or more nuclei is manifested by a *signal* in a multidimensional spectrum. A *peak* refers to an entry in a peak list that has been derived from an experimental spectrum by *peak picking*. A peak may or may not represent a signal, and there may be signals that are not represented by a peak. *Chemical shift assignment* is the process and the result of attributing a specific chemical shift value to a nucleus. *Peak assignment* is the process and the result of identifying

in each spectral dimension the nucleus or nuclei that are involved in the signal represented by the peak. *NOESY assignment* is peak assignment in NOESY spectra.

2. General principles of automated NOESY assignment and structure calculation

Many approaches have already been proposed in order to automate parts of the NMR protein structure determination process. So far, all *de novo* NMR protein structure determinations have followed the 'classic' way [5] including the successive steps of sample preparation, NMR experiments, spectrum calculation, peak picking, chemical-shift assignment, NOESY assignment and collection of other conformational constraints, structure calculation, and structure refinement. Alternative approaches that bypass the potentially cumbersome chemical shift and NOESY assignment steps have been proposed, and will be discussed in Section 5 below. The present section introduces basic aspects of automated NOESY assignment that are relevant for any algorithm implementing the standard approach.

2.1. Chemical shift assignment

The assignment of NOESY cross peaks requires as a prerequisite a knowledge of the chemical shifts of the spins from which nuclear Overhauser effects (NOEs) are arising. There have been many attempts to automate this chemical shift assignment step that has to precede the collection of conformational constraints and the structure calculation. These methods have been reviewed recently [4], and will not be discussed in detail here. Some automated approaches [6–21] target the question of assigning the backbone and, possibly, β chemical shifts, usually on the basis of triple resonance experiments that delineate the protein backbone through one- and two-bond scalar couplings, while others [22–33] are concerned with the more demanding problem of complete assignment of the amino acid side-chain chemical shifts. In most cases, these algorithms require peak lists from a specific set of NMR spectra as input, and produce lists of chemical shifts of varying completeness and correctness, depending on the quality and information content of the input data, and on the capabilities of the algorithm.

2.2. The ambiguity of chemical shift-based NOESY assignment

In *de novo* three-dimensional structure determinations of proteins in solution by NMR spectroscopy, the key conformational data are upper distance limits derived from NOEs [34–37]. In order to extract distance constraints from a NOESY spectrum, its cross peaks have to be assigned, i.e. the pairs of interacting hydrogen atoms have to be identified. The NOESY assignment is based on previously determined chemical shift values that result from the chemical shift assignment.

Because of the limited accuracy of chemical shift values and peak positions many NOESY cross peaks cannot be attributed to a single unique spin pair but have an ambiguous NOE assignment comprising multiple spin pairs. A simple mathematical model of the NOESY assignment process by chemical shift matching gives insight into this problem [38]. It assumes a protein with n hydrogen atoms, for which complete and correct chemical shift assignments are available, and N cross peaks picked in a 2D [^{1}H,^{1}H]-NOESY spectrum with an accuracy of the peak position of $\Delta\omega$, i.e. the position of the picked peak differs from the resonance frequency of the underlying signal by no more than $\Delta\omega$ in both spectral dimensions. Under the simplifying assumption of a uniform distribution of the proton chemical shifts over a range $\Delta\Omega$, the chemical shift of a given proton falls within an interval of half-width $\Delta\omega$ about a given peak position with probability $p = 2\Delta\omega/\Delta\Omega$. Peaks with unique chemical shift-based assignment have in both spectral dimensions exactly one out of all n proton shifts inside the tolerance range $\Delta\omega$ from the peak position. Their expected number,

$$N^{(1)} = N(1-p)^{2n-2} \approx N\mathrm{e}^{-2np} = N\mathrm{e}^{-4n\Delta\omega/\Delta\Omega}, \qquad (1)$$

decreases exponentially with increasing size of the protein (n) and increasing chemical shift tolerance range ($\Delta\omega$). For a typical small protein such as the *Williopsis mrakii* killer toxin (WmKT) with

88 amino acid residues, $n = 457$ proton chemical shifts and $N = 1986$ NOESY cross peaks within a range of $\Delta\Omega = 9$ ppm [39], Eq. (1) predicts that less than 2% of the NOEs can be assigned unambiguously based solely on chemical shift information with a accuracy of $\Delta\omega = 0.02$ ppm (Fig. 1), which is an insufficient number to calculate a preliminary three-dimensional structure. For peak lists obtained from ^{13}C- or ^{15}N-resolved 3D [^{1}H,^{1}H]-NOESY spectra, the ambiguity in one of the proton dimensions can usually be resolved by reference to the hetero-spin, so that Eq. (1) is replaced by

$$N^{(1)} \approx Ne^{-np} = Ne^{-2n\Delta\omega/\Delta\Omega}. \quad (2)$$

With regard to assignment ambiguity, ^{13}C- or ^{15}N-resolved 3D [^{1}H,^{1}H]-NOESY spectra are thus equivalent to homonuclear NOESY spectra from a protein of half the size or with twice the accuracy in the determination of the chemical shifts and peak positions.

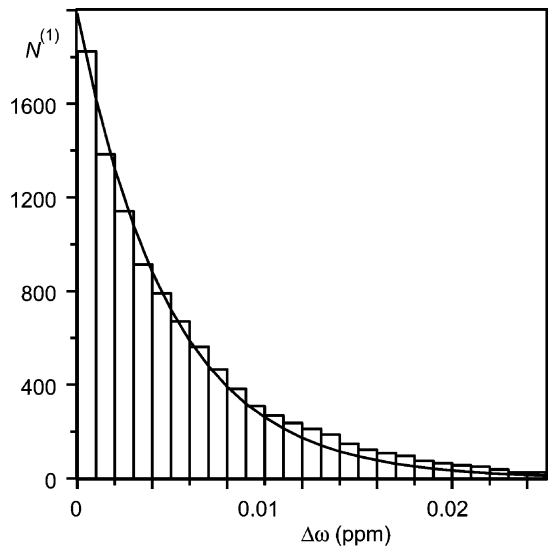

Fig. 1. Number of NOESY cross peaks with a unique chemical shift-based assignment, $N^{(1)}$, plotted as a function of the maximal chemical shift difference, $\Delta\omega$, between peak position and corresponding proton chemical shift [38]. The histogram was obtained using the experimental chemical shift list for the protein WmKT [39] and a homonuclear NOESY peak list that was simulated by postulating $N = 1986$ cross peaks for all pairs of protons that are closer than 4.0 Å in the best NMR conformer [39]. The curved line represents the corresponding values predicted by Eq. (1) for $n = 457$ proton chemical shifts, $N = 1986$ NOESY cross peaks, and $\Delta\Omega = 9.0$ ppm spectral width. No structural information was used to resolve ambiguities.

Once available, a preliminary three-dimensional structure may be used to resolve ambiguous NOE assignments. The ambiguity is resolved if only one out of all chemical shift-based assignment possibilities corresponds to an inter-atomic distance shorter than the maximal NOE-observable distance, d_{max}. Assuming that the hydrogen atoms are evenly distributed within a sphere of radius R that represents the protein, the probability q that two given hydrogen atoms are closer to each other than d_{max} can be estimated by the ratio between the volumes of two spheres with radii d_{max} and R, respectively: $q = (d_{max}/R)^3$. Using $d_{max} = 5$ Å, one obtains $q \approx 4\%$ for WmKT, a nearly spherical protein with a radius of about 15 Å [39]. Thus, only 96% of the peaks with two assignment possibilities can be assigned uniquely by reference to the protein structure. Even by reference to a perfectly refined structure it is therefore impossible, on fundamental grounds, to resolve all assignment ambiguities, since q will always be larger than zero.

Obtaining a comprehensive set of distance constraints from a NOESY spectrum is thus by no means straightforward but becomes an iterative process in which preliminary structures, calculated from limited numbers of distance constraints, serve to reduce the ambiguity of cross peak assignments. In addition to this problem of resonance and peak overlap, considerable difficulties may arise from spectral artifacts and noise, and from the absence of expected signals because of fast relaxation. These inevitable shortcomings of NMR data collection are the main reason that until recently laborious interactive procedures have dominated 3D protein structure determinations.

2.3. Automated versus manual NOESY assignment

Automated procedures follow the same general scheme but do not require manual intervention during the assignment/structure calculation cycles (Fig. 2). Two main obstacles have to be overcome by an automated approach starting without any prior knowledge of the structure. First, the number of cross peaks with unique assignment based on chemical

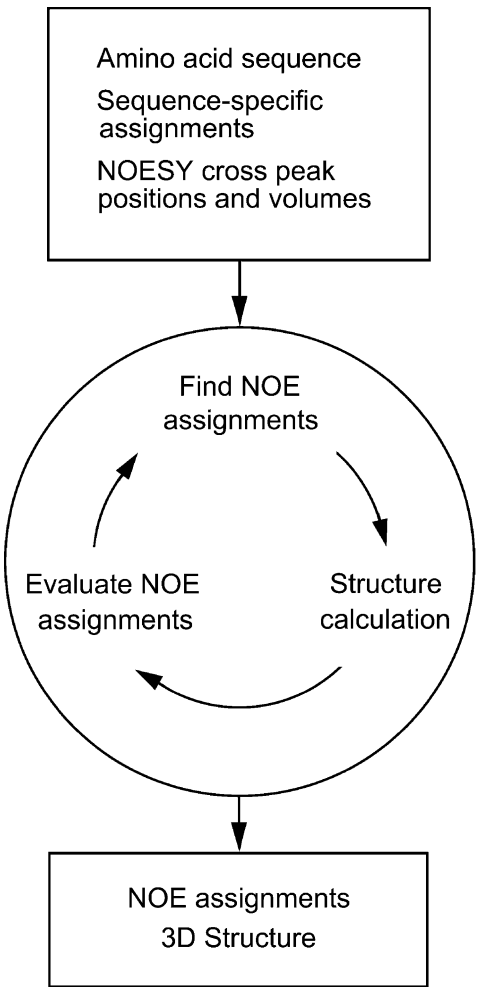

Fig. 2. General scheme of automated combined NOESY assignment and structure calculation.

shifts is, as pointed out before, in general not sufficient to define the fold of the protein. Therefore, the automated method must have the ability to make use also of NOESY cross peaks that cannot yet be assigned unambiguously. Second, the automated program must be able to cope with the erroneously picked or inaccurately positioned peaks and with the incompleteness of the chemical shift assignment of typical experimental data sets. An automated procedure needs devices to substitute the intuitive decisions made by an experienced spectroscopist in dealing with the imperfections of experimental NMR data.

3. Algorithms for automated NOESY assignment

3.1. Semi-automatic methods

Semi-automatic NOESY assignment methods relieve the spectroscopist from the burden of checking the two straightforward criteria for NOESY assignments, i.e. the agreement of chemical shifts and the compatibility with a preliminary structure, while entrusting the assignment decisions to the spectroscopist who may have additional relevant information available. Such approaches (e.g. [40–42]) use the chemical shifts and a model or preliminary structure to provide the user with a list of possible assignments for each cross peak. The user decides interactively about the assignment and/or temporary removal of individual NOESY cross peaks, possibly taking into account supplementary information such as line shapes or secondary structure data, and performs a structure calculation with the resulting, usually incomplete input. In practice, several cycles of NOESY assignment and structure calculation are required to obtain a high-quality structure.

3.1.1. The ASNO method

A prototype of this semi-automatic approach is the program ASNO [40]. The input for ASNO consists of a list of the proton chemical shifts, a peak list containing the chemical shift coordinates of the cross peaks in the NOESY spectrum, and a bundle of conformers calculated using a previous, in general preliminary set of input of NOE distance constraints. Alternatively, the structural input can consist of the crystal structure of the protein under investigation or originate from a homologous protein. However, in such applications care must be exercised to rule out possible bias by the imported reference data. In addition, the user specifies the maximally allowed chemical shift differences between corresponding cross peak coordinates and proton chemical shift values to be used for chemical shift-based assignments, the maximal proton–proton distance $d_{\max}$ in the structure that may give rise to an observable NOE, and the minimal number of conformers for which a given proton–proton distance must be shorter than $d_{\max}$ for an acceptable NOE assignment. For each NOESY cross peak ASNO first determines the set of all possible chemical shift-based assignments.

These are then checked against the corresponding $^1H-^1H$ distances in the available group of preliminary conformers and retained only if the distance between the two protons is shorter than d_{max} in at least the required number conformers. After several rounds of structure calculation, NOE assignment with ASNO, and interactive checking and refinement of the assignments, a final, high-quality structure is obtained.

3.1.2. The SANE method

The program Structure Assisted NOE Evaluation (SANE) [42] is an alternative protocol in which ambiguous distance constraints (see Section 3.3.1 below) are generated for cross peaks with multiple possible assignments. The user is directly involved in violation analysis after each round of structure calculation. Throughout the structure determination the user provides input that can help to circumvent erroneous local structures and reduce the number of iterations required to reach acceptable structures. Like ASNO, the SANE program includes a distance filter that is based on an initial search model structure, which may be an X-ray structure, an ensemble of solution structures, or even a homology-modeled structure. To minimize the problem of multiple possible assignments SANE makes use of a suite of filters that take into account existing partial assignments, the average distance between protons in one or more structures, relative NOE contributions calculated from the structures, and the expected secondary structure in order to iterate to an accurately assigned NOE cross peak list, including both unambiguous and ambiguous NOEs for the structure calculation.

3.2. The NOAH method

In a first approach and proof of feasibility of automated NOESY assignment, the programs DIANA [43] and DYANA [44] were supplemented with the automated NOESY assignment routine NOAH [38,45]. In NOAH, the multiple assignment problem is treated by temporarily ignoring cross peaks with too many (typically, more than two) assignment possibilities and instead generating independent distance constraints for each of the assignment possibilities of the remaining, low-ambiguity cross peaks, where one has to accept that part of these distance constraints may be incorrect. In order to reduce the impact of these incorrect constraints on the structure, an error-tolerant target function is used [38,45]. NOAH requires a high accuracy of the input chemical shifts and peak positions. It makes use of the fact that only a set of correct assignments can form a self-consistent network, and convergence towards the correct structure has been achieved for several proteins [38,46–48].

As an illustration, experimental 2D and 3D NOESY cross peak lists were analyzed for six proteins for which almost complete sequence-specific ^1H assignments were available for the polypeptide backbone and the amino acid side chains. The automated NOAH method assigned 70–90% of all NOESY cross peaks, which is on average 10% less than with the interactive approach, and only between 0.8 and 2.4% of the automatically assigned peaks had a different assignment than in the corresponding manually assigned peak lists. The structures obtained with NOAH/DIANA were in close agreement with those from manually assigned peak lists, and with both approaches the small remaining constraint violations indicate high-quality NMR structure determinations. Systematic comparisons of the automatically and interactively determined structures documented the absence of significant bias in either approach, indicating that an important step had been made towards automation of structure determination from NMR spectra.

In the initial assignment cycle with NOAH all peaks with one or two assignment possibilities are included into the structure calculation. In view of the large number of erroneous conformational constraints that are likely to be included at this stage, it seems non-trivial that the NOAH/DIANA approach may ultimately converge to the correct structure. The explanation is related to the fact that the structure calculation algorithm attempts to satisfy a maximum number of conformational constraints simultaneously. The correctly assigned constraints form a large subset of self-consistent constraints, whereas, in contrast, the erroneously assigned constraints are randomly distributed in space, generally contradicting each other. As a consequence, erroneously assigned constraints may distort the structure but will not lead to

a distinctly different protein fold. One must keep in mind that the elimination of erroneously assigned constraints through contradiction with correct constraints will in general be less efficient in regions of low NOE density, such as chain ends, surface loops or the periphery of long side chains, than in the well defined protein core. Another peculiarity of the randomly distributed erroneously assigned constraints is that they are more likely to be long-range than short-range or intra-residual. This contrasts with the overall constraint distribution of a correctly assigned NOESY spectrum, where more than 50% of all cross peaks are from short-range NOEs [5].

3.3. The ARIA method

The widely used automated NOESY assignment procedure ARIA [49–52] has been interfaced initially with the structure calculation program XPLOR [53] and later with the program CNS [54]. ARIA introduced many new concepts, most importantly the use of ambiguous distance constraints [55,56] for handling ambiguities in the initial, chemical shift-based NOESY cross peak assignments. Prior to the introduction of ambiguous distance constraints, in general only unambiguously assigned NOEs could be used as distance constraints in a structure calculation. Since the majority of NOEs cannot be assigned unambiguously from chemical shift information alone, this lack of a general way to directly include ambiguous data into the structure calculation considerably hampered the performance of automatic NOESY assignment algorithms.

3.3.1. Ambiguous distance constraints

When using ambiguous distance constraints, each NOESY cross peak is treated as the superposition of the signals from each of its multiple assignments, using relative weights proportional to the inverse sixth power of the corresponding inter-atomic distance. A NOESY cross peak with a unique assignment possibility gives rise to an upper bound b on the distance between two hydrogen atoms, α and β. A NOESY cross peak with $n > 1$ assignment possibilities can be seen as the superposition of n degenerate signals and interpreted as an ambiguous distance constraint, $\bar{d} \leq b$, with

$$\bar{d} = \left(\sum_{k=1}^{n} d_k^{-6} \right)^{-1/6}. \quad (3)$$

Each of the distances $d_k = d(\alpha_k, \beta_k)$ in the sum of Eq. (3) corresponds to one assignment possibility to a pair of hydrogen atoms, α_k and β_k. Because the 'r^{-6}-summed distance' $\bar{d}$ is always shorter than any of the individual distances d_k, an ambiguous distance constraint is never falsified by including incorrect assignment possibilities, as long as the correct assignment is present.

3.3.2. Overview of the ARIA algorithm

ARIA starts from lists of peaks and chemical shifts in the format of the common spectral analysis programs ANSIG [57,58], NMRView [59], PIPP [60] or XEASY [61] and proceeds in cycles of NOE assignment and structure calculation. Constraints on dihedral angles, J-couplings, residual dipolar couplings, disulfide bridges and hydrogen bonds can be used in addition, if available. In each cycle, ARIA calibrates and assigns the NOESY spectra, merges the constraint lists from different spectra, and calculates a bundle of (typically 20) conformers with the program CNS [54]. Normally, an internally generated extended start structure is used in the initial cycle 0. In all later cycles, NOE assignment, calibration and violation analysis are based on the average distances $\langle d \rangle$ calculated from the (typically 7 out of 20) lowest energy conformers from the previous cycle.

3.3.3. Calibration of distance constraints

The target distances d_{NOE} can be obtained by a simple calibration function, $d_{\mathrm{NOE}} = (CV)^{-1/6}$. The calibration constant is given by $C = \sum_{\mathrm{NOEs}} \langle d \rangle^{-6} / V$, where the sum runs over all NOEs with a corresponding average distance $\langle d \rangle$ smaller than a cutoff (typically 6 Å). An upper bound $u = d_{\mathrm{NOE}} + \varepsilon d_{\mathrm{NOE}}^2$ and a lower bound $l = d_{\mathrm{NOE}} - \varepsilon d_{\mathrm{NOE}}^2$ (typically $\varepsilon = 0.125$ Å^{-1}) are derived from each target distance d_{NOE} [51]. Alternatively, spin diffusion effects [62] can be taken into account by a relaxation matrix approach based on the simulation of the NOE spectrum rather than the direct use of the individual distances $\langle d \rangle$ [52]. A fast matrix squaring scheme performs the potentially time-consuming relaxation matrix analysis efficiently, and the deviation of the calculated NOE from the value resulting from

the isolated spin pair approximation is used to derive a correction factor for the target distance. In this way, severe cases of spin diffusion can be detected and corrected within the framework of the automated algorithm.

3.3.4. Partial NOE assignment

Despite the property of ambiguous distance constraints that additional, even wrong assignment possibilities added to an ambiguous distance constraint that contains one or several correct assignments do not render the constraint incompatible with the correct structure, it is important to reduce the ambiguity of NOE assignments as much as possible in order to obtain a well-defined structure because additional assignment possibilities 'dilute' the information contained in an ambiguous distance constraint and make it more difficult for the structure calculation algorithm to converge to the correct structure.

To this end, the relative contribution C_k of each assignment possibility to the total peak intensity is estimated from the three-dimensional structure of the previous cycle by

$$C_k = \left(\frac{\overline{\langle d \rangle}}{\langle d_k \rangle} \right)^6, \tag{4}$$

or, in the case of the relaxation matrix treatment, by the back-calculated NOE intensity [52], normalized such that the sum over all contributions to a given peak equals 1. A partial assignment is then achieved by ordering the contributions by decreasing size, and discarding the smallest contributions such that

$$\sum_{k=1}^{N_p} C_k > p, \tag{5}$$

where p is the 'assignment cutoff' and N_p the number of contributions to the peak necessary to account for a fraction of the peak volume larger than p. The parameter p is decreased from cycle to cycle and typically takes the values 1.0, 0.9999, 0.999, 0.99, 0.98, 0.96, 0.93, 0.9, 0.8 in cycles 0–8, respectively [51]. To give an intuitive meaning to the assignment cutoff p, a cross peak with two assignments may be considered [50]: If the shorter of the two distances is 2.5 Å, a value $p = 0.999$ will exclude a second distance of 7.9 Å, a value $p = 0.95$ a second distance of 4.1 Å, and a value $p = 0.8$ a second distance of 3.3 Å. If the shorter distance is 4 Å, the corresponding minimal excluded distances are 12.6, 6.6 and 5.2 Å, respectively.

3.3.5. Removal of erroneous constraints by violation analysis

Experimental peak lists can in practice not be assumed to be completely free of errors, especially in the early stages of a structure determination or if they originate from automatic peak picking. In addition, if the chemical shift assignment is incomplete, even the most carefully prepared peak list will contain peaks that cannot be assigned correctly, namely those involving unassigned spins, because the ARIA algorithm does not attempt to extend or modify chemical shift assignments provided by the user. When building a three-dimensional structure from NOE data, most erroneous distance constraints will be inconsistent with each other and with the correct ones. The erroneous constraints can therefore, in principle, be detected by analyzing the violations of constraints with respect to the bundle of three-dimensional structures from the previous cycle of calculation. The problem is to distinguish violations arising from incorrect constraints from those of correct constraints that appear as a result of insufficient convergence of the structure calculation algorithm, or as an indirect effect of structural distortions caused by other erroneous constraints. Violations due to incorrect constraints can be expected to occur in the majority of conformers rather than sporadically. Therefore, a violation analysis is performed by counting the conformers in which a given constraint is violated by more than a cutoff that is decreased gradually from 1.0 Å in the second to 0.1 Å in the final cycle of ARIA. If this is the case in more than, typically, 50% of all conformers, three options are possible [51]: The peak is either reported as a problem but still used without change, or the upper distance bound may be increased to 6 Å, or the constraint may be removed from the input for the structure calculation in the current cycle. Obviously, this kind of violation analysis can be applied only *after* a first preliminary structure has been obtained.

3.3.6. Target function with linear asymptote

In order to reduce distortions in the structures that are caused by the presence of erroneous constraints that passed undetected through this violation analysis, ARIA uses in the structure calculation with CNS a target function with a linear asymptote for large violations which limits the maximal force exerted by a violated distance constraint. The target function for a single distance constraint is [50]:

$$f(\bar{d}) = \begin{cases} (\bar{d} - l)^2 & \text{if } \bar{d} < l; \\ 0 & \text{if } l \le \bar{d} \le u; \\ (\bar{d} - u)^2 & \text{if } u < \bar{d} < u + a; \\ a(3a - 2\gamma) + \dfrac{a^2(\gamma - 2a)}{\bar{d} - u} & \\ \qquad + \gamma(\bar{d} - u) & \text{if } \bar{d} \ge u + a. \end{cases} \tag{6}$$

Here, $\bar{d}$ denotes the r^{-6}-summed distance of Eq. (3), l and u are the lower and upper distance bounds, γ is the slope of the asymptotic potential, and a is the violation at which the potential switches from harmonic to asymptotic behavior.

3.3.7. Refinement in explicit solvent

Strongly simplified, 'soft' force fields are generally used for the *de novo* calculation of NMR structures. There are two reasons for this: computational efficiency and, the need to allow for a reasonably smooth folding pathway of the polypeptide chain from a random initial structure to the native conformation that is not obstructed by high energy barriers which occur if steep, divergent potentials such as the Lennard–Jones potential of standard classical molecular dynamics force fields are used. The stiffness incurred by potentials that impede the interpenetration of parts of the molecule during the initial stages of the simulated annealing procedure would result in most conformers being trapped in local minima at unfavorable energies and far from the native structure.

However, since the physical reality of the non-bonded attractive and repulsive interactions is only crudely approximated in this way, the resulting structures have often appeared to be of low quality when submitted to common structure validation programs that put much emphasis on such features as the appearance of the Ramachandran plot, staggered rotamers of side-chain torsion angles, covalent and hydrogen bond geometry, and electrostatic interactions. To remedy this situation, a short molecular dynamics trajectory in explicit solvent may be used to refine the final structure in ARIA [63]. It has been shown that a thin layer of solvent molecules around the protein is sufficient to obtain a significant improvement in validation parameters over unrefined structures, while maintaining reasonable computational efficiency [63,64].

3.3.8. Use of ARIA in practice

The ARIA algorithm is particularly efficient for improving and completing the NOESY assignment once a correct preliminary polypeptide fold is available. On the other hand, obtaining a correct initial fold at the outset of a *de novo* structure determination can be challenging because the powerful structure-based filters used for the elimination of erroneous cross peak assignments are not yet operational at that stage. It is of great help for the initial phase of the algorithm if the user can supply a limited number of already assigned long-range distance constraints. ARIA has been used in the NMR structure determinations of more than 50 proteins [51]. A similar algorithm that also relies on ambiguous distance constraints and the program XPLOR for the structure calculation has been implemented [65,66].

3.4. The AutoStructure method

An approach that uses rules for assignments similar to those used by an expert to generate an initial protein fold has been implemented in the program AutoStructure, and applied to protein structure determination [4,67]. AutoStructure is aimed at identifying iteratively self-consistent NOE contact patterns, without using any 3D structure model, and delineating secondary structures, including alignments between β-strands, based upon a combined pattern analysis of secondary structure-specific NOE contacts, chemical shifts, scalar coupling constants, and slow amide proton exchange data. The software automatically generates conformational

constraints, e.g. distance, dihedral angle and hydrogen bond constraints, and submits parallel structure calculations with the program DYANA [44]. The resulting structure is then refined automatically by iterative cycles of self-consistent assignment of NOESY cross peaks and regeneration of the protein structure with the program DYANA.

3.5. The KNOWNOE method

The program KNOWNOE [68] presents a 'knowledge-based' approach to the problem of automated assignment of NOESY spectra that is, in principle, devised to work directly with the experimental spectra without interference of an expert. Its central part is a 'knowledge-driven Bayesian algorithm' for resolving ambiguities in the NOE assignments. NOE cross peak volume probability distributions were derived for various classes of proton–proton contacts by a statistical analysis of the corresponding inter-atomic distances in 326 protein NMR structures. For a given cross peak with n possible assignments $A_1, \ldots, A_n$, the conditional probabilities $P(A_k, a|V)$ that an assignment A_k is responsible for at least a fraction a of the cross peak volume V can then be calculated from the volume probability distributions using Bayes' theorem. Peaks with one assignment A_k with a probability $P(A_k, a|V_0)$ higher than a cutoff, typically in the range 0.8–0.9, are transiently considered as unambiguously assigned. Note that a preliminary structure is not needed to achieve this discrimination, which therefore yields a higher number of unambiguous assignments than would be possible based on chemical shifts alone (see Section 2.2). With this list of unambiguously assigned peaks a set of structures is calculated. These structures are used as input for a next cycle in which only those assignments are accepted that correspond to distances shorter than a threshold $d_{\max}$, which is decreased from cycle to cycle until 5 Å, the assumed detection limit for NOEs. Since this algorithm essentially relies on the unambiguously assigned NOEs in order to calculate the intermediate structures (only for the final structure calculation are some ambiguous distance constraint used), it requires, like NOAH (see Section 3.2), a high accuracy of the chemical shifts of typically 0.01 ppm. The program KNOWNOE was tested successfully on 2D NOESY spectra of the 66 amino acid cold shock protein from *Thermotoga maritima* for which automated assignment of NOESY spectra yielded a structure of comparable quality to the one obtained from manual data evaluation [68].

3.6. The CANDID method

The CANDID algorithm [69] in the program CYANA [70] combines features from NOAH and ARIA, such as the use of three-dimensional structure-based filters and ambiguous distance constraints, with the new concepts of network-anchoring and constraint combination that further enable an efficient and reliable search for the correct fold in the initial cycle of *de novo* NMR structure determinations.

3.6.1. Overview of the CANDID algorithm

The automated CANDID method proceeds in iterative cycles of ambiguous NOE assignment followed by structure calculation with the CYANA torsion angle dynamics algorithm. Between subsequent cycles, information is transferred exclusively through the intermediary three-dimensional structures, in that the molecular structure obtained in a given cycle is used to guide the NOE assignments in the following cycle. Otherwise, the same input data are used for all cycles, that is, the amino acid sequence of the protein, one or several chemical shift lists from the sequence-specific resonance assignment, and one or several lists containing the positions and volumes of cross peaks in 2D, 3D or 4D NOESY spectra. The NOESY peak lists can be prepared either using interactive spectrum analysis programs such as XEASY [61], NMRView [59], ANSIG [57,58], or automated peak picking methods such as AUTOPSY [71] or ATNOS [72] that allow to start the NOE assignment and structure calculation process directly from the NOESY spectra. The input may further include previously assigned NOE upper distance constraints or other previously assigned conformational constraints. These will not be touched during NOE assignment with CANDID, but used for the CYANA structure calculation.

A CANDID cycle starts by generating for each NOESY cross peak an initial assignment list containing the hydrogen atom pairs that could, from the fit of chemical shifts within a user-defined tolerance range, contribute to the peak. Subsequently, for each cross peak these initial assignments are weighted with

respect to several criteria, and initial assignments with low overall score are discarded. These filtering criteria include the agreement between the values of the chemical shift list and the peak position, self-consistency within the entire NOE network (see Section 3.6.2 below), and, if available, the compatibility with the three-dimensional structure from the preceding cycle (Fig. 3). In the first cycle, network-anchoring has a dominant impact, since structure-based criteria cannot be applied yet. For each cross peak, the retained assignments are interpreted in the form of an upper distance limit derived from the cross peak volume. Thereby, a conventional distance constraint is obtained for cross peaks with a single retained assignment, and otherwise an ambiguous distance constraint is generated that embodies several assignments. Cross peaks with a poor score are temporarily discarded. In order to reduce deleterious effects on the resulting structure from erroneous distance constraints that may pass this filtering step, long-range distance constraints are incorporated into 'combined distance constraints' (see Section 3.6.3 below). The distance constraints are then included in the input for the structure calculation with the CYANA torsion angle dynamics algorithm.

The structure calculations typically comprise seven cycles. The second and subsequent cycles differ from the first cycle by the use of additional selection criteria for cross peaks and NOE assignments that are based on assessments relative to the protein 3D structure from the preceding cycle. Since the precision of the structure determination normally improves with each subsequent cycle, the criteria for accepting assignments and distance constraints are tightened in more advanced cycles of the CANDID calculation. The output from a CANDID cycle includes a listing of NOESY cross peak assignments, a list of comments about individual assignment decisions that can help to recognize potential artifacts in the input data, and a three-dimensional protein structure in the form of a bundle of conformers.

In the final CANDID cycle, an additional filtering step ensures that all NOEs have either unique assignments to a single pair of hydrogen atoms, or are eliminated from the input for the structure calculation. This allows for the direct use of the NOE assignments in subsequent refinement and analysis programs that do not handle ambiguous distance constraints.

The core of the CANDID algorithm has been implemented in the program CYANA [70]. The standard schedule and parameters for a complete automated structure determination with CYANA are specified in a script written in the interpreted command language INCLAN [44] that gives the user high flexibility in the way automated structure determination is performed without the need to modify the compiled core part of the algorithm.

3.6.2. Network-anchoring

Network-anchoring exploits the observation that the correctly assigned constraints form a self-consistent

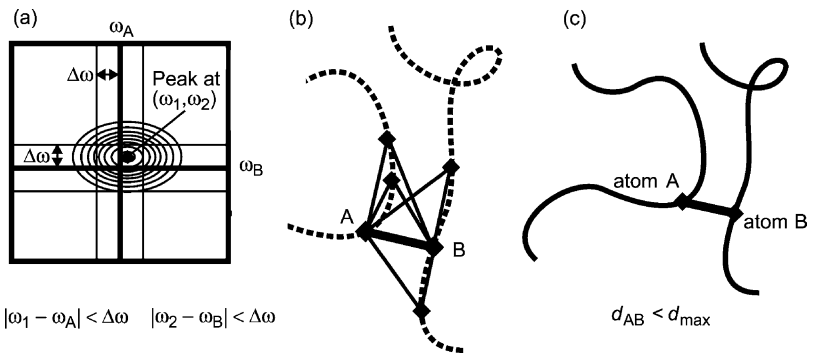

Fig. 3. Three conditions that must be fulfilled by a valid assignment of a NOESY cross peak to two protons A and B in the CANDID automated NOESY assignment algorithm [69]: (a) Agreement between chemical shifts and the peak position, (b) network-anchoring, and (c) spatial proximity in a (preliminary) structure.

subset in any network of distance constraints that is sufficiently dense for the determination of a protein 3D structure. Network-anchoring thus evaluates the self-consistency of NOE assignments independent of knowledge on the 3D protein structure, and in this way compensates for the absence of 3D structural information at the outset of a *de novo* structure determination (Fig. 3). The requirement that each NOE assignment must be embedded in the network of all other assignments makes network-anchoring a sensitive approach for detecting erroneous, 'lonely' constraints that might artificially constrain unstructured parts of the protein. Such constraints would not otherwise lead to systematic constraint violations during the structure calculation, and could therefore not be eliminated by 3D structure-based peak filters.

The network-anchoring score $N_{\alpha\beta}$ for a given initial assignment of a NOESY cross peak to an atom pair (α, β) is calculated by searching all atoms γ in the same or in the neighboring residues of either α or β that are connected simultaneously to both atoms α and β. The connection may either be an initial assignment of another peak (in the same or in another peak list) or the fact that the covalent structure implies that the corresponding distance must be short enough to give rise to an observable NOE. Each such indirect path contributes to the total network-anchoring score for the assignment (α, β) an amount given by the product of the generalized volume contributions of its two parts, $\alpha \rightarrow \gamma$ and $\gamma \rightarrow \beta$. $N_{\alpha\beta}$ has an intuitive meaning as the number of indirect connections between the atoms α and β through a third atom γ, weighted by their respective generalized volume contributions.

The calculation of the network-anchoring score is recursive in the sense that its calculation for a given peak requires the knowledge of the generalized volume contributions from other peaks, which in turn involve the corresponding network-anchored assignment contributions. Therefore, the calculation of these quantities is iterated three times, or until convergence. Note that the peaks from all peak lists contribute simultaneously to the network-anchored assignment.

3.6.3. Constraint combination

In the practice of NMR structure determination with biological macromolecules, spurious distance constraints may arise from misinterpretation of noise and spectral artifacts. This situation is particularly critical at the outset of a structure determination, before the availability of a preliminary structure for 3D structure-based filtering of constraint assignments. Constraint combination aims at minimizing the impact of such imperfections on the resulting structure at the expense of a temporary loss of information. Constraint combination is applied in the first two CANDID cycles. It consists of generating distance constraints with combined assignments from different, in general unrelated, cross peaks (Fig. 4). The basic property of ambiguous distance constraints that the constraint will be fulfilled by the correct structure whenever at least one of its assignments is correct, regardless of the presence of additional, erroneous assignments, then implies that such combined constraints have a lower probability of being erroneous than the corresponding original constraints, provided that the fraction of erroneous original constraints is smaller than 50%.

CANDID provides two modes of constraint combination (further combination modes can be envisaged readily) [69]: '$2 \rightarrow 1$' combination of all assignments of two long-range peaks each into a single constraint, and '$4 \rightarrow 4$' pairwise combination of the assignments of four long-range peaks into four constraints. Let A, B, C, D denote the sets of assignments of four peaks. Then, $2 \rightarrow 1$ combination replaces two constraints with assignment sets A and B, respectively, by a single ambiguous constraint with assignment set $A \cup B$, the union of sets A and B. $4 \rightarrow 4$ pairwise combination replaces four constraints with assignments A, B, C and D by four combined ambiguous constraints with assignment sets $A \cup B$, $A \cup C$, $A \cup D$ and $B \cup C$, respectively. In both cases constraint combination is applied only to the long-range peaks, i.e. the peaks with all assignments to pairs of atoms separated by at least five residues in the sequence, because in case of error their effect on the global fold of a protein is more pronounced than that of erroneous short-and medium-range constraints. The number of long-range constraints is halved by $2 \rightarrow 1$ combination but stays constant upon $4 \rightarrow 4$ pairwise combination. The latter approach therefore preserves more of the original structural information, and can furthermore take into account that certain peaks and their assignments are more reliable than others, because the peaks with

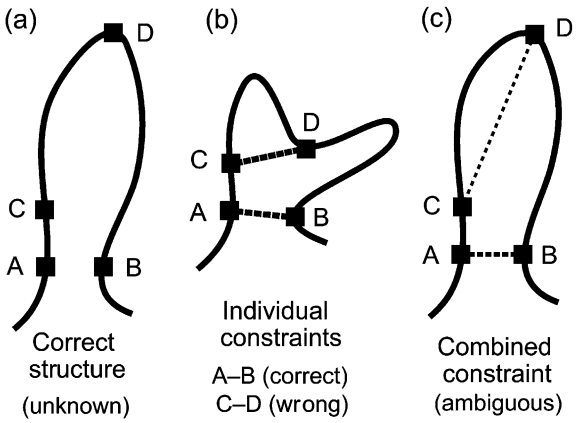

Fig. 4. Schematic illustration of the effect of constraint combination [69] in the case of two distance constraints, a correct one connecting atoms A and B, and a wrong one between atoms C and D. A structure calculation that uses these two constraints as individual constraints that have to be satisfied simultaneously will, instead of finding the correct structure (a), result in a distorted conformation (b), whereas a combined constraint that will be fulfilled already if one of the two distances is sufficiently short leads to an almost undistorted solution (c).

assignment sets A, B, C, D are used 3, 2, 2, 1 times, respectively, to form combined constraints. To this end, the long-range peaks are sorted according to their total residue-wise network-anchoring and $4 \rightarrow 4$ combination is performed by selecting the assignments A, B, C, D from the first, second, third, and fourth quarter of the sorted list.

The effect of constraint combination on the expected number of erroneous distance constraints in the case of $2 \rightarrow 1$ combination may be estimated quantitatively by assuming an original data set containing N long-range peaks, and a uniform probability $p \ll 1$ that a long-range peak would lead to an erroneous constraint. By $2 \rightarrow 1$ constraint combination, these are replaced by $N/2$ constraints that are erroneous with probability p^2. In the case of $4 \rightarrow 4$ combination, it is assumed that the same N long-range peaks can be classified according to the 'safety' of their assignments into four equally large classes with probabilities αp, p, p, $(2 - \alpha)p$, respectively, that they would lead to erroneous constraints. The overall probability for an input constraint to be erroneous is again p. The parameter α, $0 \leq \alpha \leq 1$, expresses how much 'safer' the peaks in the first class are compared to those in the two middle classes, and in the fourth, 'unsafe' class. After $4 \rightarrow 4$ combination, there are still N long-range constraints but with an overall error probability of $(\alpha + (1 - \alpha^2)/4)p^2$, which is smaller than the probability p^2 obtained by simple $2 \rightarrow 1$ combination provided that the classification into more and less safe classes was successful ($\alpha < 1$). For instance, $4 \rightarrow 4$ combination will transform an input data set of 900 correct and 100 erroneous long-range cross peaks (i.e. $N = 1000$, $p = 0.1$) that can be split into four classes with $\alpha = 0.5$ into a new set of approximately 993 correct and 7 erroneous combined constraints. Alternatively, $2 \rightarrow 1$ combination will yield under these conditions approximately 495 correct and 5 erroneous combined constraints. Unless the number of erroneous constraints is high, $4 \rightarrow 4$ combination is thus preferable over $2 \rightarrow 1$ combination in the first two CANDID cycles.

The upper distance bound b for a combined constraint is formed from the two upper distance bounds b_1 and b_2 of the original constraints either as the r^{-6}-sum, $b = (b_1^{-6} + b_2^{-6})^{-1/6}$, or as the maximum, $b = \max(b_1, b_2)$. The first choice minimizes the loss of information if two already correct constraints are combined, whereas the second choice avoids the introduction of too small an upper bound if a correct and an erroneous constraint are combined.

3.6.4. Use of CANDID in practice

If used sensibly, automated NOESY assignment with CANDID has no disadvantage compared to the conventional, interactive approach but is a lot faster, and more objective. Network-anchored assignment and constraint combination render the automated CANDID method stable also in the presence of the imperfections typical for experimental NMR data sets. With CANDID, the evaluation of NOESY spectra is no longer the time-limiting step in protein structure determination by NMR. Furthermore, simple criteria based on the output of CANDID that will be given in Section 4.3 allows the reliability of the resulting structure to be assessed without cumbersome recourse to independent interactive verification of

the NOESY assignments. The CANDID method has been evaluated in test calculations [69] and used in various *de novo* structure determinations, including, for instance, four variants of the human prion protein [73,74], the pheromone binding protein from *Bombyx mori* [75], the calreticulin P-domain [76], the class I human ubiquitin-conjugating enzyme 2b [77], the heme chaperone CcmE [78] (Fig. 5), and the nucleotide-binding domain of Na, K-ATPase [79]. These structure determinations have confirmed that network-anchored assignment and constraint combination enable reliable, truly automated NOESY assignment and structure calculation without prior knowledge about NOESY assignments or the three-dimensional structure. All NOESY assignments and the corresponding distance constraints for these *de novo* structure determinations were made with CANDID, confining interactive work to the stage of the preparation of the input chemical shift and peak lists.

4. Robustness and quality control of automated NMR structure calculation

4.1. Effect of incomplete chemical shift assignments

A limiting factor for the application of all automated NOE assignment methods described in Section 3 is that they rely on the availability of an essentially complete list of chemical shifts from the preceding sequence-specific resonance assignment. At present, chemical shift assignment remains largely the domain of interactive or semi-automated methods, despite promising attempts towards automation (Section 2.1). Experience shows that in general the majority of the chemical shifts can be assigned readily whereas others pose difficulties that may require a disproportionate amount of the spectroscopist's time. Hence, NMR structure determination would be speeded up significantly if NOE assignment and structure calculation could be based on incomplete lists of assigned chemical shifts, provided that

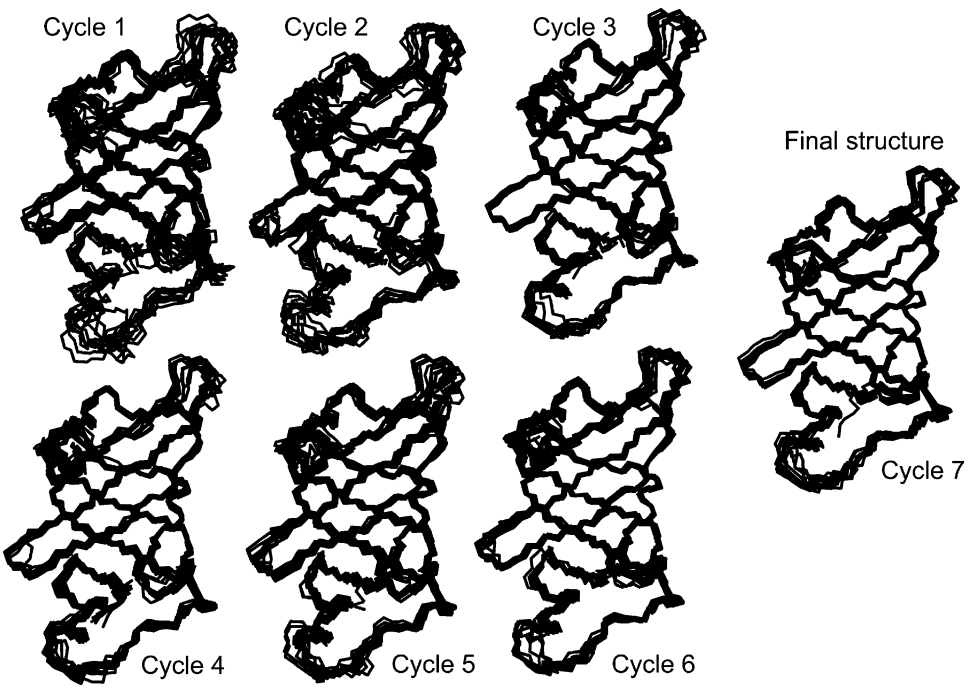

Fig. 5. Structures of the heme chaperone CcmE [78] obtained with the program CYANA [70] in seven consecutive cycles of combined automated NOESY assignment with CANDID [69] and structure calculation with torsion angle dynamics. The backbones of the 10 conformers with lowest target function value in each cycle were drawn with the program MOLMOL [94].

the reliability and robustness of the NMR method for protein structure determination is not compromised.

Methods to find additional chemical shift assignments simultaneously with automated NOESY assignment and the structure calculation have been proposed and applied with some success in the case when a preliminary structure was available [80]. For example, starting from nearly complete chemical shift assignments for the backbone and for 348 side-chain protons of the 28 kDa single-chain T cell receptor protein, the chemical shifts of 40 additional side-chain protons were found by a combination of chemical shift prediction with the program SHIFTS [81,82] and NOE assignment with ARIA [80].

The influence of incomplete chemical shift assignments on the reliability of NMR structures obtained by automated NOESY cross peak assignment has been investigated in detail [83] using the program CYANA for combined automated NOESY assignment with the CANDID algorithm and torsion angle dynamics-based structure calculations at various degrees of completeness of the chemical shift assignment. The effect of missing chemical shift assignments was simulated by randomly omitting entries from the experimental ^{1}H chemical shift lists that had been used for the earlier, conventional structure determinations of two proteins, the *Bombyx mori* pheromone binding protein form A (BmPBP^A) [75] and the *Williopsis mrakii* killer toxin (WmKT) [39]. Sets of structure calculations were performed with different numbers and selections of randomly omitted chemical shifts and the results compared to those obtained when using the complete experimental chemical shift list. The deviation of the structures obtained with incomplete chemical shift assignments from the reference structure was monitored by the 'RMSD bias', the RMSD between the mean coordinates of the two structure bundles [84].

In the representative case of randomly selecting the omitted chemical shifts among all ^{1}H chemical shift assignments, the RMSD bias increased only slowly with increasing omission ratio P up to about $P = 10\%$, from where onwards the RMSD bias rose abruptly, reflecting that severely distorted structures had been obtained. Higher omission ratios did not only result in high mean values of the RMSD bias but also in pronounced variations among the individual runs at a given P value with different random selections of the omitted shifts. The CYANA target function values of the final structures were, regardless of the omission ratio, almost always in the range below 5 Å^2 that is indicative of a structure that essentially fulfills all the input conformational constraints. The percentages of unassigned NOEs increased and the number of distance constraints for the final cycle of structure calculation decreased almost linearly with the omission rate. The algorithm was more tolerant against the presence of incomplete chemical shifts when run with the data from the uniformly ^{13}C- and ^{15}N-labeled protein BmPBP^A than with the homonuclear data for the protein WmKT despite the fact that BmPBP^A (142 residues) is much larger than WmKT (88 residues). This is due to the availability of ^{13}C and ^{15}N chemical shifts that allow many ^{1}H chemical shift degeneracies to be resolved, such that the probability of accidental erroneous NOE assignments is decreased compared to the case of homonuclear data. The omission of aromatic ^{1}H chemical shift assignments in general causes more severe problems than the omission of the same number of chemical shifts chosen randomly among all assigned ^{1}H chemical shifts [83]. In the case of BmPBP^A the omission of all assigned aromatic chemical shifts, corresponding to 6.0% of all assigned protons, led already to 2 Å RMSD bias. In the case of WmKT, with only homonuclear data, significant deviations from the reference structure were in some cases already observed at 20% omission of the aromatic chemical shifts, which corresponds to an overall omission ratio of merely 1.6% of all assigned ^{1}H chemical shifts.

Overall, the test calculations [83] show that for reliable automated NOESY assignment with the CANDID algorithm, and *a fortiori* other NOE assignment algorithms based on the same principles, around 90% completeness of the chemical shift assignment is necessary. In certain cases the lack of a small number of 'essential' chemical shifts can lead to a significant deviation of the structure. However, in practice the algorithm might be expected to tolerate a slightly higher degree of incompleteness in the chemical shift assignments than the simulations [83] suggest provided that most missing assignments are of 'unimportant' chemical shifts that are involved in only a few NOEs. This is usually the case because the chemical shifts of protons that are

involved in many NOEs, and, if absent, prevent the program from correctly assigning any of these NOEs, are intrinsically easier to assign than those exhibiting only a small number of NOEs. This effect is confirmed by the finding that the lack of aromatic chemical shifts is in general more harmful to the outcome of a structure calculation than that of a similar number of other protons because aromatic protons tend to be located in the hydrophobic core of the protein where they give rise to a higher-than-average number of NOEs.

The CANDID algorithm includes network-anchoring and constraint combination, two exclusive features that have been designed and shown to be effective in minimizing the impact of incomplete and/or erroneous pieces of input data (see Sections 3.6.2 and 3.6.3). Chemical shift assignment-based automated NOE assignment without network-anchoring and constraint combination must be expected to be more susceptible to deleterious effects from missing chemical shift assignments and artifacts in the input data.

4.2. Effect of incomplete NOESY peak picking

In contrast to the effects seen under the omission of chemical shift assignments, the random omission of NOESY peaks does not cause severe problems (Fig. 3 of Ref. [83]). Even when 50% of the NOESY peaks were omitted from the experimental input peak lists for $BmPBP^A$, most RMSD bias values remained in the region of 2 Å. An outlier with RMSD bias close to 4 Å shows that for $BmPBP^A$ the algorithm starts to loose its stability at 50% NOE omission ratio. The results with the homonuclear data from WmKT showed similar patterns, albeit with a somewhat stronger dependence on the omission rate and RMSD bias values occasionally exceeding 2 Å in runs with 30% NOESY peak omission ratio. The CYANA structure calculation protocol is thus remarkably tolerant with respect to incomplete NOESY peak picking, and can tolerate the omission of up to 50% of the NOESY cross peaks with only a moderate decrease in the precision and accuracy of the resulting structure. This suggests that it is better to strive for correctness than for ultimate completeness of the input NOESY peak lists.

4.3. Quality control

Final structures from an automatic algorithm that have a low RMSD within the bundle of conformers but differ significantly from the 'correct' reference structure are problematic because, without a knowledge of a reference structure, they may appear at first glance as good, well-defined solutions. In a conventional structure calculation based on manual NOESY assignment, incomplete or inconsistent input data will be manifested by a large RMSD and/or target function values of the final structure bundle, which will prompt the spectroscopist to correct and/or complete the input data for a next round of structure calculation. The test calculations [83] showed that for structure calculation with automated NOE assignment neither the RMSD value of the final structure nor the final target function value are suitable indicators to discriminate between correct and biased results. Other criteria are needed to evaluate the outcome.

On the basis of the initial experience with the CANDID algorithm, guidelines for successful CANDID runs were proposed [69]. These comprise six criteria that should be met simultaneously: (1) average CYANA target function value of cycle 1 below 250 Å; (2) average final CYANA target function value below 10 $Å^2$; (3) less than 20% unassigned NOEs; (4) less than 20% discarded long-range NOEs; (5) RMSD value in cycle 1 below 3 Å; and (6) RMSD between the mean structures of the first and last cycle below 3 Å. The criterion (4) refers to the percentage of NOEs discarded by the CANDID algorithm among all NOEs with assignments exclusively between atoms separated by four or more residues along the polypeptide sequence. The criteria (3) and (4) limit the number of NOEs that are not used to generate distance constraints for the final structure calculation, and thus measure the completeness with which the picked NOE cross peaks can be explained by the resulting structure.

The validity of the original guidelines as sufficient conditions for successful CANDID runs was confirmed by the fact that all the structure calculations in the systematic study [83] with an RMSD bias to the reference structure of more than 2 Å violated one or several of the six criteria. On the other hand, the test calculations [83] revealed a certain redundancy among the six original criteria. Provided that

the input peak lists do not deliberately misinterpret the underlying NOESY spectra (to which the algorithm has no direct access), the aforementioned criteria can be replaced by only two conditions. Thus, for successful structure calculation with automated NOESY assignment by the CANDID algorithm in CYANA, less than 25% of the long-range NOEs must have been discarded by the automated NOESY assignment algorithm for the final structure calculation, and the backbone RMSD to the mean coordinates for the structure bundle of the *first* cycle must not exceed 3 Å.

The percentage of discarded long-range NOEs cannot be calculated readily outside the CYANA program, because it requires knowledge of the possible assignments also for the NOESY cross peaks that were excluded from the generation of conformational constraints. In this case, an overall percentage of unused cross peaks of less than 15% can be used as an alternative criterion that is straightforward to evaluate from the final assigned output peak lists, in which unused cross peaks remain unassigned.

The ability of the program to find a well-defined structure in the initial cycle of NOE assignment and structure calculation, as measured by the RMSD within the structure bundle in cycle 1, is another important factor that strongly influences the accuracy of the final structure, as measured by the RMSD bias. This can be understood by considering the iterative nature of the CANDID algorithm, by which each cycle except cycle 1 is dependent on the structure obtained in the preceding cycle. Using network-anchoring and constraint-combination, the algorithm tries to obtain a well-defined structure already in the first cycle. A low precision of the structure from cycle 1 may hinder convergence to a well-defined final structure, or, more dangerously, opens the possibility of a structural drift in later cycles towards a precise but incorrect final structure.

4.4. Troubleshooting

If the output of a structure calculation based on automated NOESY assignment with CANDID does not fulfill these guidelines, the structure will in many cases still be essentially correct, but should not be accepted without further validation. Within the framework of CANDID, the normal approach is to improve the quality of the input chemical shift and peak lists, and to perform another CANDID run, until the criteria are met. Usually, this can be achieved efficiently because the output from an unsuccessful CANDID run, even though the structure should not be trusted per se, clearly reveals problems in the input, e.g. peaks that cannot be assigned and might therefore be artifacts or indications of erroneous or missing sequence-specific assignments. CANDID provides informational output for each peak that greatly facilitates this task: the list of its chemical shift-based assignment possibilities, the assignment(s) finally chosen, and the reasons why an assignment is chosen or not, or why a peak is not used at all. Even when the criteria of the previous section are met already, a higher precision and local accuracy of the structure might still be achieved by further improving the input data.

In principle, a *de novo* protein structure determination requires one run of CYANA with 7 cycles of automated NOE assignment and structure calculation. This is realistic when almost complete chemical shift assignments and exhaustive high-quality NOESY peak lists are available. In practice, it is often more efficient to start a first CYANA calculation from an initial, slightly incomplete list of 'safely identifiable' NOESY cross peaks. The results of this first CYANA calculation can then be used as additional information to prepare an improved, more complete NOESY peak list for a second CYANA calculation. This can be done more efficiently than would be possible ab initio because only peaks and regions of the protein that gave rise to problems in the first CYANA calculation need to be checked.

5. Structure calculation without chemical shift assignment

It is almost universally assumed that a protein structure determination by NMR requires the sequence-specific resonance assignments [5]. However, the chemical shift assignment by itself has no biological relevance. It is required only as an intermediate step in the interpretation of the NMR spectra. Several attempts have been made to devise a strategy for NMR protein structure determination that

circumvents the tedious chemical shift assignment step. There is an analogy between these approaches and the direct phasing methods in X-ray crystallography [85]. Although until today no *de novo* NMR protein structure determination has been accomplished without prior chemical shift assignment, an introduction into the concept of assignment-free NMR structure calculation appears warranted because recent progress in this field may open the avenue to an alternative strategy of NMR structure determination.

The underlying idea of assignment-free NMR structure calculation methods is to exploit the fact that NOESY spectra provide distance information even in the absence of any chemical shift assignments. This proton–proton distance information can be exploited to calculate a spatial proton distribution. Since there is no association with the covalent structure at this point, the protons of the protein are treated as a gas of unconnected particles. Provided that the emerging proton distribution is sufficiently clear, a model can then be built into the proton density in a manner analogous to X-ray crystallography in which the structural model is constructed into the electron density.

5.1. Initial approaches

This general idea was first tested in 1992 by Malliavin et al. [86] with 302 NOEs between backbone amide protons of lysozyme that were simulated from the crystal structure, under the assumptions that the NOEs provide distance measurements with an accuracy of $\pm 5\%$, and that the absence of a NOE indicates that the corresponding distance exceeds 4.5 Å. For the distance geometry structure calculations it was further assumed that there is no chemical shift degeneracy, i.e. it is known unambiguously whether any two pairs of NOEs involve the same proton or not. About 100 clouds of backbone hydrogen atoms were calculated using distance geometry. Despite large structural variations reflected by RMSD values of 7–14 Å among these 'structures', some secondary structure elements could be identified. Considering that even in the presence of complete chemical shift assignments the NOEs between backbone amide protons alone are in general not sufficient to determine more than a rough global fold, the results of the simulation are encouraging. Furthermore, a simplistic algorithm could extract from the proton clouds the assignments of the backbone hydrogen atoms with less than 10% error.

The question of direct structure calculation without chemical shift assignments was again investigated in 1993 by Oshiro and Kuntz [87] in simulations with synthetic NOE data for BPTI and combining metric matrix distance geometry with graph theoretical approaches to identify secondary structure elements and, eventually, sequence-specific assignments. It was concluded that 'this approach is only useful with excellent quality stereo-resolved data'.

5.2. The ANSRS method

At that time the most thorough attempt at protein three-dimensional structure determination and sequence-specific assignment of ^{13}C and ^{15}N-separated NOE data using 'a novel real-space ab initio approach' came with Per Kraulis' ANSRS algorithm in 1994 [88]. The input data are a list of NOESY cross peaks including knowledge of the chemical shifts of the ^{13}C or ^{15}N atoms covalently bound to the protons that make the NOE (i.e. a 4D NOESY peak list), and a complete but unassigned list of the chemical shifts of all detectable $^{1}H-^{13}C$ and $^{1}H-^{15}N$ moieties. The ANSRS algorithm then proceeds in three stages. First, 3D structures of unconnected ^{1}H atoms are calculated using dynamical simulated annealing. Second, a list for each residue type of plausible ^{1}H spin combinations with probability scores is generated in a recursive combinatorial search with spatial constraints. Finally, the sequence-specific assignment and a low-resolution 3D structure are obtained by Monte Carlo simulated annealing. The algorithm was tested for two small proteins, a fragment of GAL4 with 32 residues and BPTI with 58 residues using the experimental chemical shifts and synthetic NOE constraints for all distances shorter than 4 Å in the previously known 3D structures. There were 193 $^{1}H-X$ chemical shift pairs and 753 distance constraints for GAL4, and 301 H–X chemical shift pairs and 1173 distance constraints for BPTI. NOEs were interpreted in a conservative manner by using them as upper distance bounds. The resulting average 3D real-space ^{1}H spin structures were within less than 2 Å RMSD from the previously known 3D structure, and

the ANSRS procedure was able to determine the sequence-specific assignments for more than 95% of the spins. These may in turn be used as input for a conventional structure calculation in order to obtain a high-resolution structure. Despite these encouraging figures, the ANSRS program has not become a routine tool for NMR structure determination, presumably because the requirements on the quality of the input data are still formidable from the experimental point of view, and because the algorithm has no facilities to deal with overlap among ^{1}H–X chemical shift pairs.

5.3. Inclusion of information from through-bond spectra

Atkinson and Saudek proposed an interesting algorithm for direct fitting of structure and chemical shift data to NMR spectra [89]. Optimization of four variables per atom, three Cartesian coordinates and the chemical shift value, directly against the NOESY spectrum, rather than peak lists, by simulated annealing was shown to succeed in finding sets of coordinates (i.e. structures) and chemical shifts that match the reference configuration, albeit only in the case of a peptide fragment with six atoms. Subsequently, the same authors realized [90] that the direct determination of protein structures by NMR without chemical shift assignment is not restricted to using only NOESY spectra, but can incorporate, in a natural way, data from the same set of heteronuclear and dipolar coupling experiments as normally used in the conventional approach. NOEs are again interpreted as distances between unassigned and unconnected atoms, while cross peaks in all other spectra are also interpreted as distances instead of being used for assignment purposes. For example, a ^{15}N–^{1}H HSQC peak yields a distance equal to the N–H bond length between the two corresponding atoms, the HNCA spectrum yields, for each N–H pair, four distances to the two adjacent Cα atoms. To validate this principle, synthetic data was produced for the 76 amino acid protein ubiquitin: 1647 exact distances corresponding to the expected peaks from 10 heteronuclear scalar coupling experiments, 2040 4D NOE cross peaks corresponding to the ^{1}H–^{1}H distances shorter than 4 Å in the crystal structure, and 92,570 lower distance bounds of 4 Å for all ^{1}H–^{1}H distances longer than 4 Å in the crystal structure. The structure calculations with the program XPLOR yielded solutions with RMSD values to the crystal structure below 2 Å. These structures were obtained with no prior assignment of any spectral resonance or cross peak, but every hydrogen atom in the structure is labeled by both its own chemical shift and that of the attached heavy atom.

5.4. The CLOUDS method

The most recent approach to NMR structure determination without chemical shift assignment is the CLOUDS protocol of Grishaev and Llinás [91,92]. For the first time, the feasibility of the method has been demonstrated using experimental data rather than simulated data sets. The CLOUDS method relies on precise and abundant inter-proton distance constraints calculated via a relaxation matrix analysis of sets of experimental NOESY cross peaks [93]. A gas of unassigned, unconnected hydrogen atoms is condensed into a structured proton distribution (cloud) via a molecular dynamics simulated annealing scheme in which the inter-nuclear distances and van der Waals repulsive terms are the only active constraints. Proton densities are generated by combining a large number of such clouds, each computed from a different trajectory.

After filtering by reference to the cloud closest to the mean, a minimal dispersion proton density ('family of clouds', foc) is identified that affords a quasi-continuous hydrogen-only probability distribution and conveys immediate information on the shape of the protein.

The NMR-generated foc proton density provides a template to which the molecule has to be fitted to derive the structure. The primary structure is threaded through the unassigned foc by a Bayesian approach, for which the probabilities of sequential connectivity hypotheses are inferred from likelihoods of H^N–H^N, H^N–H^α, and H^α–H^α inter-atomic distances as well as ^{1}H NMR chemical shifts, both derived from public databases. Once the polypeptide sequence is identified, directionality becomes established, and the foc N and C termini are recognized. After a similar procedure, side chain hydrogen atoms are found. The folded structure is then obtained via a molecular dynamics calculation that embeds 3D structures into mirror image-related representations of the foc and selected according to a lowest energy criterion.

The feasibility of the method was tested with experimental NMR data measured for two globular protein domains, the col 2 domain of human matrix metalloproteinase-2 and the kringle 2 domain of human plasminogen, of 60 and 83 amino acid residues, respectively, for which excellent unambiguously identified homonuclear NOESY peak lists were available from the previous, conventional structure determinations. The structures deviate by 1.0–1.4 Å RMSD for the backbone heavy atoms and 1.5–2.1 Å RMSD for all heavy atoms from the previously reported X-ray and NMR structures. These results show that assignment-free NMR structure calculation can successfully generate 3D protein structures from experimental data. Nevertheless, in the course of a *de novo* structure determination it may not be straightforward to produce a NOESY peak list of the completeness and quality used for these test calculations. In particular, it was assumed that the NOEs can be identified unambiguously, i.e. that it is known with certainty whether any two NOESY peaks involve the same proton or not.

As for all NMR spectrum analysis, resonance overlap presents a major difficulty also in applying 'no assignment' strategies. Indeed, if two resonances from nuclei that are far apart in the structure have identical chemical shifts but distinct sets of neighbors they would be represented by a single atom with one set of neighbors, leading to a gross distortion of the calculated structure. In that respect, the use of heteronuclear-edited NOESY spectra drastically reduces the likelihood of overlap. At present, a full *de novo* protein structure determination by the assignment-free approach has not been reported, and it is of great interest to see whether the assignment-free approach will be able to provide the robustness and quality of the structures obtained by the conventional method.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. 28 (2000) 235.
[2] S.E. Brenner, Nat. Rev. 2 (2001) 801.
[3] D. Vitkup, E. Melamud, J. Moult, C. Sander, Nat. Struct. Biol. 8 (2001) 559.
[4] H.N.B. Moseley, G.T. Montelione, Curr. Opin. Struct. Biol. 9 (1999) 635.
[5] K. Wüthrich, NMR of Proteins and Nucleic Acids, Wiley, 1986.
[6] M.S. Friedrichs, L. Mueller, M. Wittekind, J. Biomol. NMR 4 (1994) 703.
[7] J.B. Olson, J.L. Markley, J. Biomol. NMR 4 (1994) 385.
[8] N.E.G. Buchler, E.R.P. Zuiderweg, H. Wang, R.A. Goldstein, J. Magn. Reson. 125 (1996) 34.
[9] J.A. Lukin, A.P. Gove, S.N. Talukdar, C. Ho, J. Biomol. NMR 9 (1997) 151.
[10] 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.
[11] M. Leutner, R.M. Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, H. Kessler, J. Biomol. NMR 11 (1998) 31.
[12] P. Güntert, M. Salzmann, D. Braun, K. Wüthrich, J. Biomol. NMR 18 (2000) 129.
[13] H.S. Atreya, S.C. Sahu, K.V.R. Chary, G. Govil, J. Biomol. NMR 17 (2000) 125.
[14] C. Bailey-Kellog, A. Widge, J.J. Kelley, M.J. Bernardi, J.H. Bushweller, B.R. Donald, J. Comput. Biol. 7 (2000) 537.
[15] N.S. Bhavesh, S.C. Panchal, R.V. Hosur, Biochemistry 40 (2001) 14727.
[16] F. Tian, H. Valafar, J.H. Prestegard, J. Am. Chem. Soc. (2001) 11791.
[17] H.N.B. Moseley, D. Monleon, G.T. Montelione, Curr. Methods Enzymol. 339 (2001) 91.
[18] M. Andrec, R.M. Levy, J. Biomol. NMR 23 (2002) 263.
[19] A. Chatterjee, N.S. Bhavesh, S.C. Panchal, R.V. Hosur, Biochem. Biophys. Res. Commun. 293 (2002) 427.
[20] D. Monleon, K. Colson, H.N.B. Moseley, C. Anklin, R. Oswald, T. Szyperski, G.T. Montelione, J. Struct. Funct. Genom. 2 (2002) 93.
[21] B.E. Coggins, P. Zhou, J. Biomol. NMR 26 (2003) 93.
[22] C. Yu, J.F. Hwang, T.B. Chen, V.W. Soo, J. Chem. Inf. Comput. Sci. 32 (1992) 183.
[23] J. Xu, S.K. Strauss, B.C. Sanctuary, L. Trimble, J. Chem. Inf. Comput. Sci. 33 (1993) 668.
[24] J. Xu, S.K. Strauss, B.C. Sanctuary, L. Trimble, J. Magn. Reson. B103 (1994) 53.
[25] H. Oschkinat, D. Croft, Methods Enzymol. 239 (1994) 308.
[26] C. Bartels, M. Billeter, P. Güntert, K. Wüthrich, J. Biomol. NMR 7 (1996) 207.
[27] C. Bartels, P. Güntert, M. Billeter, K. Wüthrich, J. Comput. Chem. 18 (1997) 139.
[28] W.Y. Choy, B.C. Sanctuary, G. Zhu, J. Chem. Inf. Comput. Sci. 37 (1997) 1086.
[29] W. Gronwald, L. Willard, T. Jellard, R.E. Boyko, K. Rajaranthnam, D.S. Wishart, F.D. Sonnichsen, B.D. Sykes, J. Biomol. NMR 12 (1998) 395.
[30] K.B. Li, B.C. Sanctuary, J. Chem. Inf. Comput. Sci. 37 (1997) 359.
[31] K.B. Li, B.C. Sanctuary, J. Chem. Inf. Comput. Sci. 37 (1997) 467.
[32] P. Pristovšek, H. Rüterjans, R. Jerala, J. Comput. Chem. 23 (2002) 335.
[33] T.K. Hitchens, J.A. Lukin, Y. Zhan, S.A. McCullum, G.S. Rule, J. Biomol. NMR 25 (2003) 1.

[34] I. Solomon, Phys. Rev. 99 (1955) 559.
[35] S. Macura, R.R. Ernst, Mol. Phys. 41 (1980) 95.
[36] A. Kumar, R.R. Ernst, K. Wüthrich, Biochem. Biophys. Res. Commun. 95 (1980) 1.
[37] D. Neuhaus, M.P. Williamson, The Nuclear Overhauser Effect in Structural and Conformational Analysis, VCH, 1989.
[38] C. Mumenthaler, P. Güntert, W. Braun, K. Wüthrich, J. Biomol. NMR 10 (1997) 351.
[39] W. Antuch, P. Güntert, K. Wüthrich, Nat. Struct. Biol. 3 (1996) 662.
[40] P. Güntert, K.D. Berndt, K. Wüthrich, J. Biomol. NMR 3 (1993) 601.
[41] R.P. Meadows, E.T. Olejniczak, S.W. Fesik, J. Biomol. NMR 4 (1994) 79.
[42] B.M. Duggan, G.B. Legge, H.J. Dyson, P.E. Wright, J. Biomol. NMR 19 (2001) 321.
[43] P. Güntert, W. Braun, K. Wüthrich, J. Mol. Biol. 217 (1991) 517.
[44] P. Güntert, C. Mumenthaler, K. Wüthrich, J. Mol. Biol. 273 (1997) 283.
[45] C. Mumenthaler, W. Braun, J. Mol. Biol. 254 (1995) 465.
[46] Y. Xu, J. Wu, D. Gorenstein, W. Braun, J. Magn. Reson. 136 (1999) 76.
[47] Y. Xu, M.J. Jablonsky, P.L. Jackson, W. Braun, N.R. Krishna, J. Magn. Reson. 148 (2001) 35.
[48] N. Oezguen, L. Adamian, Y. Xu, K. Rajarathnam, W. Braun, J. Biomol. NMR 22 (2002) 249.
[49] M. Nilges, M.J. Macias, S.I. O'Donoghue, H. Oschkinat, J. Mol. Biol. 269 (1997) 408.
[50] M. Nilges, S.I. O'Donoghue, Prog. NMR Spectrosc. 32 (1998) 107.
[51] J.P. Linge, S.I. O'Donoghue, M. Nilges, Methods Enzymol. 339 (2001) 71.
[52] J.P. Linge, M. Habeck, W. Rieping, M. Nilges, Bioinformatics 19 (2003) 315.
[53] A.T. Brünger, X-PLOR Version 3.1. A system for X-ray crystallography and NMR, Yale University Press, 1993.
[54] A.T. Brünger, P.D. Adams, G.M. Clore, W.L. DeLano, P. Gros, R.W. Grosse-Kunstleve, J.S. Jiang, J. Kuszewski, M. Nilges, N.S. Pannu, R.J. Read, L.M. Rice, T. Simonson, G.L. Warren, Acta Crystallogr. D 54 (1998) 905.
[55] M. Nilges, Proteins 17 (1993) 297.
[56] M. Nilges, J. Mol. Biol. 245 (1995) 645.
[57] P.J. Kraulis, J. Magn. Reson. 24 (1989) 627.
[58] M. Helgstrand, P. Kraulis, P. Allard, T. Härd, J. Biomol. NMR 18 (2000) 329.
[59] B.A. Johnson, R.A. Blevins, J. Biomol. NMR 4 (1994) 603.
[60] D.S. Garrett, R. Powers, A.M. Gronenborn, G.M. Clore, J. Magn. Reson. 95 (1991) 214.
[61] C. Bartels, T. Xia, M. Billeter, P. Güntert, K. Wüthrich, J. Biomol. NMR 6 (1995) 1.
[62] A. Kalk, H.J.C. Berendsen, J. Magn. Reson. 24 (1976) 343.
[63] J.P. Linge, M.A. Williams, C.A.E.M. Spronk, A.M.J.J. Bonvin, M. Nilges, Proteins 50 (2003) 496.
[64] C.A.E.M. Spronk, J.P. Linge, C.W. Hilbers, G.W. Vuister, J. Biomol. NMR 22 (2002) 281.
[65] B. Gilquin, A. Lecoq, F. Desné, M. Guenneugues, S. Zinn-Justin, A. Ménez, Proteins 34 (1999) 520.
[66] P. Savarin, S. Zinn-Justin, B. Gilquin, J. Biomol. NMR 19 (2001) 49.
[67] N.J. Greenfield, Y.J. Huang, T. Palm, G.V.T. Swapna, D. Monleon, G.T. Montelione, S.E. Hitchcock-DeGregori, J. Mol. Biol. 312 (2001) 833.
[68] 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.
[69] T. Herrmann, P. Güntert, K. Wüthrich, J. Mol. Biol. 319 (2002) 209.
[70] CYANA version 1.0, www.guentert.com.
[71] R. Koradi, M. Billeter, M. Engeli, K. Wüthrich, J. Magn. Reson. 135 (1998) 288.
[72] T. Herrmann, P. Güntert, K. Wüthrich, J. Biomol. NMR 24 (2002) 171.
[73] L. Calzolai, D.A. Lysek, P. Güntert, C. von Schroetter, R. Riek, R. Zahn, K. Wüthrich, Proc. Natl Acad. Sci. USA 97 (2000) 8340.
[74] R. Zahn, P. Güntert, C. von Schroetter, K. Wüthrich, J. Mol. Biol. 326 (2003) 225.
[75] R. Horst, F. Damberger, P. Luginbühl, P. Güntert, G. Peng, L. Nikonova, W.S. Leal, K. Wüthrich, Proc. Natl Acad. Sci. USA 98 (2001) 14374.
[76] L. Ellgaard, R. Riek, T. Herrmann, P. Güntert, D. Braun, A. Helenius, K. Wüthrich, Proc. Natl Acad. Sci. USA 98 (2001) 3133.
[77] T. Miura, W. Klaus, A. Ross, P. Güntert, H. Senn, J. Biomol. NMR 22 (2002) 89.
[78] E. Enggist, L. Thöny-Meyer, P. Güntert, K. Pervushin, Structure 10 (2002) 1551.
[79] M. Hilge, G. Siegal, G.W. Vuister, P. Güntert, S.M. Gloor, J.P. Abrahams, Nat. Struct. Biol. 10 (2003) 468.
[80] B.J. Hare, G. Wagner, J. Biomol. NMR 15 (1999) 103.
[81] K. Ösapay, D.A. Case, J. Am. Chem. Soc. 113 (1991) 9436.
[82] D.F. Sitkoff, D.A. Case, J. Am. Chem. Soc. 119 (1997) 12262.
[83] J.G. Jee, P. Güntert, J. Struct. Funct. Genom. (2003) in press.
[84] P. Güntert, Q. Rev. Biophys. 31 (1998) 145.
[85] J. Drenth, Principles of Protein X-ray Crystallography, Springer, 1994.
[86] T.E. Malliavin, A. Rouh, M. Delsuc, J.-Y. Lallemand, C. R. Acad. Sci. Ser. II 315 (1992) 635.
[87] C.M. Oshiro, I.D. Kuntz, Biopolymers 33 (1993) 107.
[88] P.J. Kraulis, J. Mol. Biol. 243 (1994) 696.
[89] R.W. Atkinson, V. Saudek, J. Chem. Soc. Faraday Trans. 93 (1997) 3319.
[90] R.W. Atkinson, V. Saudek, FEBS Lett. 510 (2002) 1.
[91] A. Grishaev, M. Llinás, Proc. Natl Acad. Sci. USA 99 (2002) 6707.
[92] A. Grishaev, M. Llinás, Proc. Natl Acad. Sci. USA 99 (2002) 6713.
[93] M. Madrid, E. Llinás, M. Llinás, J. Magn. Reson. 93 (1991) 329.
[94] R. Koradi, M. Billeter, K. Wüthrich, J. Mol. Graph. 14 (1996) 51.