doi:10.1016/S0022-2836(02)00241-3 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 319, 209–227

JMB

# Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA

**Torsten Herrmann, Peter Güntert* and Kurt Wüthrich**

*Institut für Molekularbiologie und Biophysik Eidgenössische Technische Hochschule-Hönggerberg CH-8093 Zürich, Switzerland*

Combined automated NOE assignment and structure determination module (CANDID) is a new software for efficient NMR structure determination of proteins by automated assignment of the NOESY spectra. CANDID uses an iterative approach with multiple cycles of NOE cross-peak assignment and protein structure calculation using the fast DYANA torsion angle dynamics algorithm, so that the result from each CANDID cycle consists of exhaustive, possibly ambiguous NOE cross-peak assignments in all available spectra and a three-dimensional protein structure represented by a bundle of conformers. The input for the first CANDID cycle consists of the amino acid sequence, the chemical shift list from the sequence-specific resonance assignment, and listings of the cross-peak positions and volumes in one or several two, three or four-dimensional NOESY spectra. The input for the second and subsequent CANDID cycles contains the three-dimensional protein structure from the previous cycle, in addition to the complete input used for the first cycle. CANDID includes two new elements that make it robust with respect to the presence of artifacts in the input data, i.e. network-anchoring and constraint-combination, which have a key role in *de novo* protein structure determinations for the successful generation of the correct polypeptide fold by the first CANDID cycle. Network-anchoring makes use of the fact that any network of correct NOE cross-peak assignments forms a self-consistent set; the initial, chemical shift-based assignments for each individual NOE cross-peak are therefore weighted by the extent to which they can be embedded into the network formed by all other NOE cross-peak assignments. Constraint-combination reduces the deleterious impact of artifact NOE upper distance constraints in the input for a protein structure calculation by combining the assignments for two or several peaks into a single upper limit distance constraint, which lowers the probability that the presence of an artifact peak will influence the outcome of the structure calculation. CANDID test calculations were performed with NMR data sets of four proteins for which high-quality structures had previously been solved by interactive protocols, and they yielded comparable results to these reference structure determinations with regard to both the residual constraint violations, and the precision and accuracy of the atomic coordinates. The CANDID approach has further been validated by *de novo* NMR structure determinations of four additional proteins. The experience gained in these calculations shows that once nearly complete sequence-specific resonance assignments are available, the automated CANDID approach results in greatly enhanced efficiency of the NOESY spectral analysis. The fact that the correct fold is obtained in

Present address: P. Güntert, RIKEN Genomic Sciences Center, 1-7-22 Suehiro, Tsurumi, Yokohama 230-0045, Japan.

Abbreviations used: NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser enhancement spectroscopy; CANDID, combined automated NOE assignment and structure determination module; DYANA, dynamics algorithm for NMR applications; 2D, 3D, 4D, two, three, four-dimensional.

E-mail address of the corresponding author: guentert@gsc.riken.go.jp

cycle 1 of a *de novo* structure calculation is the single most important advance achieved with CANDID, when compared with previously proposed automated NOESY assignment methods that do not use network-anchoring and constraint-combination.



*Keywords:* automated NOE assignment; network-anchoring; constraint combination; CANDID; DYANA

*Corresponding author

## Introduction

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.[1] NOEs result from cross-relaxation due to the dipole–dipole interactions between nearby pairs of nuclear spins in a molecule undergoing Brownian motion,[2] and in two-dimensional (2D) or higher-dimensional heteronuclear-resolved [$^1$H,$^1$H]-NOESY spectra they are manifested by cross-peaks.[3,4] In order to extract informative distance constraints from a NOESY spectrum, its cross-peaks have to be assigned, i.e. the pairs of hydrogen atoms that give rise to the observed cross-peaks need to be identified. These NOESY assignments are based on $^1$H chemical shift values that result from previous sequence-specific resonance assignments.[1] However, because of the limited accuracy with which NOESY cross-peak positions and chemical shift values can be measured, it is in general not possible to unambiguously assign all NOESY cross-peaks on the basis of the known chemical shift values alone, not even for small proteins. The number of NOESY cross-peaks that can be unambiguously assigned from knowledge of the $^1$H chemical shifts decreases rapidly with increasing uncertainty of the information on chemical shifts and NOE peak positions,[5] and may drop below 10% of the total number of NOE cross-peaks.[5] Obtaining a comprehensive set of distance constraints from a NOESY spectrum of a protein has therefore conventionally been an iterative process in which preliminary protein three-dimensional (3D) structures calculated from a fraction of the total number of distance constraints are used to reduce the ambiguity of additional cross-peak assignments.[6] Additional difficulties may arise from spectral artifacts and noise, and from the absence of expected signals because of fast relaxation. These inevitable shortcomings of current NMR data collection are the main reason that laborious interactive procedures are still prominent in 3D protein structure determinations.

Interactive computer-supported NOESY assignment methods have been introduced that use chemical shift fits for initial assignments, and a molecular model from a preliminary structure determination to validate individual ones out of the list of possible assignments for each cross-peak.[6–8] The user decides interactively about the assignment and/or temporary removal of individual NOESY cross-peaks, possible 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.[9] Automated NOESY assignment/structure calculation methods attempt to work along the same general scheme without interactive interventions.[10] Because in the initial phase of a structure determination the number of cross-peaks with unique assignments based on chemical shift fitting may not be sufficient to define the protein fold, automated methods should be able to initially extract information also from NOESY cross-peaks that cannot yet be assigned unambiguously. Furthermore, an automated procedure must be able to substitute for the intuitive decisions made by an experienced spectroscopist in dealing with spectral noise, other artifacts, and possibly inaccurately positioned real NOE cross-peaks.

The programs DIANA[9,11] and DYANA[12] have previously been supplemented with the automated NOESY assignment routine NOAH.[5,13] 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 all assignment possibilities of the remaining cross-peaks, where one takes into account that part of these distance constraints may be incorrect. NOAH requires high accuracy of the chemical shifts and peak positions in the input. 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. Another automated NOESY assignment procedure, ARIA, has been interfaced with the programs XPLOR and CNS,[14–17] and a similar approach has been implemented by Savarin *et al*.[18] ARIA introduced the concept of ambiguous distance constraints for handling of ambiguities in the initial, chemical shift-based NOESY cross-peak assignments. When using ambiguous distance constraints each individual NOESY cross-peak is treated as the superposition of the signals from each of its multiple initial assignments, using relative weights proportional to the inverse sixth power of the corresponding interatomic distance in a

preliminary model of the molecular structure.[19,20] In this way, information from cross-peaks with an arbitrary number of initial assignment possibilities can be used for the structure calculation, and although inclusion of erroneous assignments for a given cross-peak results in a loss of information, it will not lead to inconsistencies as long as one or several correct assignments are among the initial assignments. Both of these automated methods are quite efficient for improving and completing the NOESY assignment once a correct preliminary polypeptide fold is available, for example, based on a limited set of interactively assigned NOEs. On the other hand, obtaining a correct initial fold at the outset of a *de novo* structure determination often proves to be difficult, because the structure-based filters used in both of these procedures for the elimination of erroneous cross-peak assignments are then not operational. A third approach that uses rules for assignments similar to the ones used by an expert to generate an initial protein fold has been implemented in the program AUTOSTRUCTURE, and applied to protein structure determination.[10,21]

The CANDID procedure described here 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.

## Algorithms

This section starts with a brief overview of the process of automated protein structure determination with CANDID, and then presents a technical description of the individual steps of the procedure in the order in which they appear in Figure 1. The flow diagram (Figure 1) emphasizes the new elements implemented in the CANDID algorithm with thick-framed boxes. The key new features are "network-anchoring" of the initial, chemical shift-based NOE cross-peak assignments, and "constraint-combination", which represents an extension of the concept of ambiguous NOE assignments.[19,20] These two elements are of critical importance for the generation of the correct polypeptide fold during the first cycle of a *de novo* protein structure determination.

The automated CANDID method proceeds in iterative *cycles*, each consisting of exhaustive, in part ambiguous NOE assignment followed by a structure calculation with the DYANA torsion angle dynamics algorithm. Between subsequent cycles, information is transferred exclusively through the intermediary 3D structures, in that the protein molecular structure obtained in a given cycle is used to guide further NOE assignments in the following cycle. Otherwise, the same input data are used for all cycles, i.e. 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 four-dimensional (4D) NOESY spectra. The input may further include previously assigned NOE upper distance constraints or other previously assigned conformational constraints, which will then not be changed by CANDID, but will be used for the structure calculation.

A CANDID cycle starts by generating for each NOESY cross-peak an initial assignment list, i.e. hydrogen atom pairs are identified that could, from the fit of chemical shifts within the user-defined tolerance range, contribute to the peak. Subsequently, for each cross-peak these initial assignments are weighted with respect to several criteria (listed in Figure 1), and initial assignments with low overall score are then discarded. 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.[19,20] All 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" (Figure 1). The distance constraints are then included in the input for the structure calculation with the DYANA torsion angle dynamics algorithm.

The structure calculations described here comprise seven CANDID cycles. The second and subsequent CANDID 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 3D protein structure in the form of a bundle of conformers.

In the final CANDID cycle (cycle 7), 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 CANDID NOE assignments in subsequent refinement and analysis programs that do not handle ambiguous distance constraints,

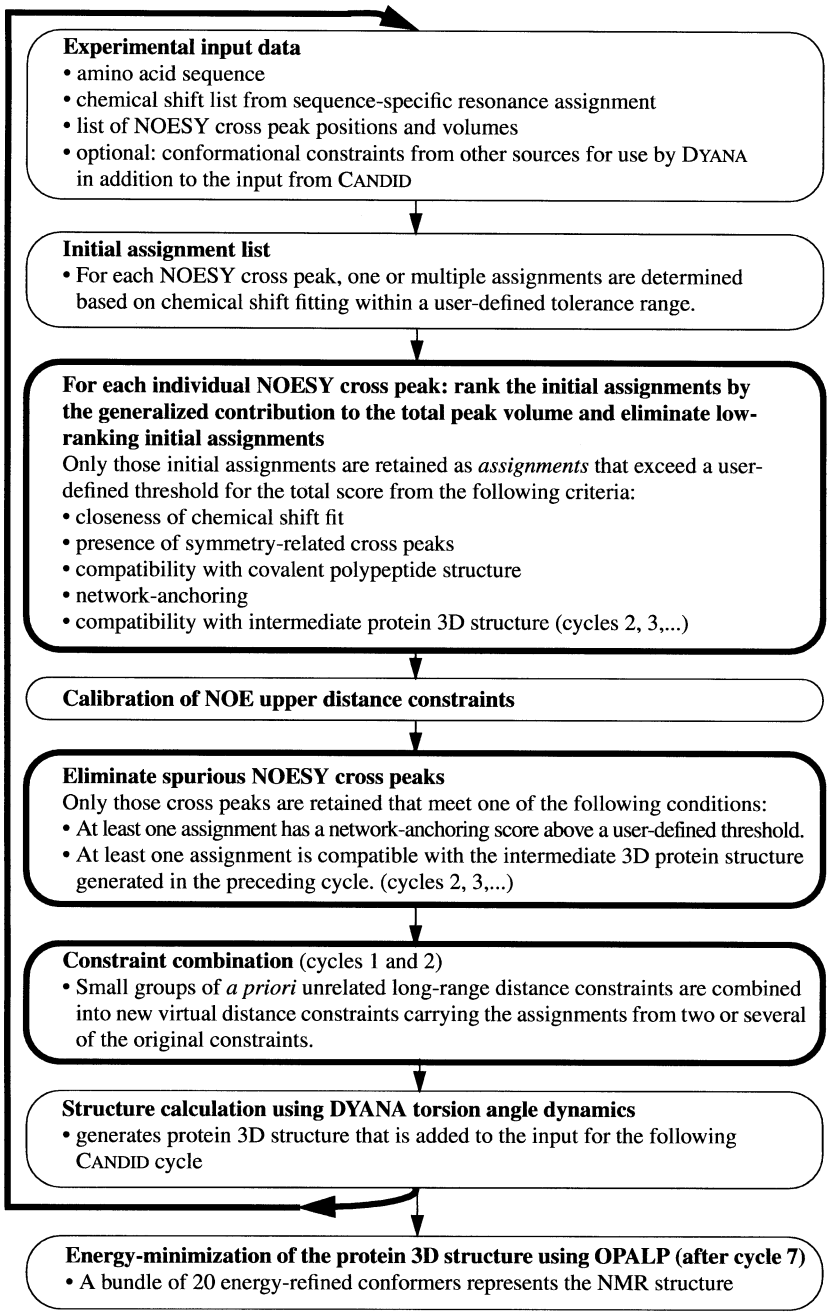


**Figure 1.** Flow chart of NMR structure determination using the CANDID method for automated NOE cross-peak assignment.

and enables here direct comparisons of the CANDID results with the corresponding data obtained by conventional, interactive procedures.

## Experimental input data

The formats of the input files containing the amino acid sequence, chemical shifts, NOE cross-peak positions and volumes, and possible previously assigned cross-peaks or distance constraints for use in the structure calculation (Figure 1) are compatible with the programs XEASY,[22] which supports interactive sequential assignment protocols, and DYANA.[12]

## Initial assignment list

The chemical shift value of an atom, α, in the input chemical shift list for a NOESY spectrum, $S$, is denoted $\Omega_\alpha^S \pm \Delta\Omega_\alpha^S$, where $\Delta\Omega_\alpha^S$ describes a tolerance range that allows for the limited precision of experimental chemical shift determinations. The absence of a chemical shift value for the atom α is indicated by setting $\Omega_\alpha^S = \infty$. A different chemical shift list may be used for each NOESY spectrum from which peaks are extracted. However, if experimental conditions are carefully matched when recording multiple NOESY spectra, it is advantageous to generate a single list of chemical shifts for all the spectra.

A peak, $p$, in an input peak list for the 2D, 3D or 4D NOESY spectrum $S$ is characterized by $D$ chemical shift coordinates, $\omega_i^p (i = 1, 2, \ldots, D)$, by corresponding tolerance ranges, $\Delta\omega_i^p$, and by its volume, $I^p$. The dimensions of the spectrum are chosen such that $\omega_1^p$ and $\omega_2^p$ denote the two $^1$H chemical shifts. Each hydrogen atom, $\alpha$, is covalently bound to a "heavy" atom, $h(\alpha)$. If applicable, $\omega_3^p$ and $\omega_4^p$ refer to the chemical shifts of the $^{13}$C or $^{15}$N atoms that are covalently bound to the $^1$H atoms represented by the dimensions 1 and 2, respectively. CANDID can work simultaneously with several NOESY peak lists of different dimensionality.

A $^1$H atom, $\alpha$, is assigned to dimension $i$ ($i = 1, 2$) of a peak, $p$, from the NOESY spectrum $S$ if the chemical shift value $\Omega_\alpha^S$ agrees with the peak position $\omega_i^p$ within a given tolerance range, i.e. if:

$$\left|\omega_i^p - \Omega_\alpha^S\right| \le \max\left(\Delta\omega_i^p, \Delta\Omega_\alpha^S\right) \tag{1}$$

and in the case of 3D or 4D NOESY spectrum:

$$\left|\omega_{i+2}^p - \Omega_{h(\alpha)}^S\right| \le \max\left(\Delta\omega_{i+2}^p, \Delta\Omega_{h(\alpha)}^S\right) \tag{2}$$

If $A_i^p (i = 1, 2)$ is the set of all hydrogen atoms $\alpha$ that satisfy the conditions of equations (1) and (2), then the chemical-shift based initial assignments for the peak $p$ are given by the direct product of the sets $A_1^p$ and $A_2^p$, i.e. by the ordered pairs of hydrogen atoms, ($\alpha$,$\beta$), with $\alpha \in A_1^p$ and $\beta \in A_2^p$. All peaks with at least one initial assignment on the diagonal, i.e. $\alpha = \beta$, are eliminated from further consideration as cross-peaks.

## Ranking of the initial assignments

A NOESY cross-peak with a single initial assignment gives rise to a conventional upper distance constraint, whereas a NOESY cross-peak with $n \ge 2$ initial assignments gives rise to an ambiguous distance constraint[19,20] that will not distort the protein structure by the inadvertent inclusion of incorrect initial assignments as long as the correct assignment is also present among the initial assignments. However, since an ambiguous distance constraint has a reduced information content, it may nonetheless be difficult for the structure calculation to converge to the correct structure. It is therefore important, in as far as possible, to eliminate incorrect initial assignments before the start of the structure calculation. For this filtering process, the initial assignments are ranked by their generalized relative contributions, and only the assignments with sufficiently high contributions are retained as conventional or ambiguous distance constraints for the structure calculation.

This generalized relative contribution, $V_k$, of an initial assignment, $k$, to a cross-peak volume is given by the normalized total score from four structure-independent and one structure-based term (see Figure 1), and is defined by:

$$V_k = \frac{C_k \min(T_k O_k N_k, S_{\max}) D_k}{\sum_{i=1}^{n} C_i \min(T_i O_i N_i, S_{\max}) D_i} \tag{3}$$

such that $\sum_{k=1}^{n} V_k = 1$. $C_k$ is the weight for the closeness of the chemical shift fit. The other weighting factors, $T_k$, $O_k$, $N_k$ and $D_k$ are related to the presence of symmetry-related cross-peaks in the NOESY spectra, the compatibility with the covalent polypeptide structure, the convergence of network-anchoring, and the compatibility with the protein 3D structure. These factors will be defined in the following sections. The product of the three weighting factors, $T_k$, $O_k$ and $N_k$ is capped at a user-defined maximal value, $S_{\max}$.

## Ranking of the initial assignments by the closeness of the chemical shift fit

For the purpose of discriminating between multiple initial assignments, the agreement between peak coordinates, $\omega_i^p (i = 1, 2, \ldots, D)$, and the chemical shifts of the corresponding initial assignment, $\Omega_{\alpha_i}^S (\alpha_1 = \alpha,\ \alpha_2 = \beta,\ \alpha_3 = h(\alpha),\ \alpha_4 = h(\beta))$, is quantified by a Gaussian weighting factor:

$$C_k = C_{\alpha,\beta}^p = \exp\left(-\frac{1}{2}\sum_{i=1}^{D}\left(\frac{\omega_i^p - \Omega_{\alpha_i}^S}{\Gamma \max\left(\Delta\omega_i^p, \Delta\Omega_{\alpha_i}^S\right)}\right)^2\right) \tag{4}$$

which has the value 1.0 for a perfect fit. $\Gamma$ is a user-defined parameter that determines the weight of close chemical shift alignment relative to the other ranking criteria.

## Ranking of the initial assignments by the presence of symmetry-related cross-peaks in 3D and 4D heteronuclear-resolved NOESY spectra

Assume that a peak, $p$, in a 3D NOESY spectrum has an initial assignment ($\alpha$,$\beta$,$h(\alpha)$). Then, peak $p^*$, at the position $(\omega_1^{p^*}, \omega_2^{p^*}, \omega_3^{p^*})$ is in the transposed position with respect to the initial assignment ($\alpha$,$\beta$,$h(\alpha)$) if:

$$\left|\omega_i^{p^*} - \Omega_{\alpha_i}^S\right| \le \max\left(\Delta\omega_i^{p^*}, \Delta\Omega_{\alpha_i}^S\right) \quad (i = 1, 2, 3) \tag{5}$$

where $\alpha_1 = \beta$, $\alpha_2 = \alpha$ and $\alpha_3 = h(\beta)$. 4D NOESY spectra can be treated in an analogous manner. If a transposed peak is found by the criterion of equation (5), a weight $T_k = T \gg 1$ is attached to the corresponding initial assignment, $k$, where $T$ is a user-defined constant, and otherwise $T_k$ is set to unity. To prevent arbitrary discrimination among the initial assignments for a given peak, values of $T_k \ne 1$ are used only if the heavy atoms $h(\beta)$ have been assigned for all initial assignments of this peak.

### Ranking of the initial assignments by the compatibility with the covalent polypeptide structure

The fixed bond lengths, bond angles and chiralities of the covalent structure impose NOE-observable upper limits on certain intraresidual and sequential distances.[1,9,23,24] In CANDID these conformation-independent upper limits, $u^{(cc)}_{\alpha\beta}$, are computed analytically for atom pairs (α,β) that are separated by one or two torsion angles. CANDID gives a weight $O_k = O \gg 1$ to an initial assignment $k$ if the corresponding distance cannot exceed a user-defined maximal value, $d_{max}$, where $O$ is a user-defined constant, and otherwise $O_k$ is set to unity. This discrimination in favor of assignments that are expected to yield observable NOEs in all possible conformations of the protein corresponds to the common treatment of short-range $^1H-^1H$ connectivites by experienced spectroscopists in the course of interactive NOE assignments.[1]

### Ranking of the initial assignments by the compatibility with the intermediate 3D protein structure

The structure-based volume contribution $D_k$, of an initial assignment, $k$ ($k = 1, \ldots, n$), is calculated as an average over all $M$ conformers in the preliminary structure bundle:

$$D_k = \frac{1}{M}\sum_{j=1}^{M}\left(d^{(j)}_{\alpha_k\beta_k}/\bar{d}^{(j)}\right)^{-\eta} \tag{6}$$

where $d^{(j)}_{\alpha_k,\beta_k}$ denotes the distance between the two atoms $\alpha_k$ and $\beta_k$ in conformer $j$, and:

$$\bar{d} = \left(\sum_{k=1}^{n} d^{-6}_{\alpha_k\beta_k}\right)^{-1/6} \tag{7}$$

For the isolated spin pair approximation, $\eta = 6$, but smaller values may be used to reduce the sensitivity of $D_k$ to structural variations in the situation where only an imprecise preliminary structure is available as a reference. In the absence of a structure during the first CANDID cycle, uniform weights, $D_k = 1/n$, are applied for all initial assignments.

### Ranking of the initial assignments of network-anchoring

Network-anchoring exploits the observation that the correctly assigned constraints form a self-consistent 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. 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 constrains might not lead to systematic constraint violations during the structure calculation, and could therefore not be eliminated by 3D structure-based peak filters.

The weighting factor for network-anchoring for an initial assignments of a NOESY cross-peak, $N_k = N_{\alpha\beta}$, is calculated as follows. All atoms $\gamma \neq \alpha,\beta$ are searched that are connected simultaneously to both atoms α and β, either by initial assignments of other NOE cross-peaks, or because the covalent polypeptide structure implies that the distance must be sufficiently short to produce a NOE. In addition, the atoms γ are required to be in the same residue as either α or β, or in one of the neighboring residues. $N_{\alpha\beta}$ is then defined as the sum over all indirect pathways that connect the atoms α and β *via* a third atoms, γ:

$$N_k = N_{\alpha\beta} = \sum_{\gamma}\sqrt{\nu_{\alpha\gamma}\nu_{\beta\gamma}} \tag{8}$$

$N_{\alpha\beta}$ thus represents the number of indirect connections between the atoms α and β through a third atom γ and their impact on the network-anchoring, with the sum of equation (8) running over all atoms γ as defined above. The weights for the connections α to γ, $\nu_{\alpha\gamma}$, are defined by:

$$\nu_{\alpha\gamma} = \tilde{\nu}_{\alpha\gamma}\theta(\tilde{\nu}_{\alpha\gamma} - \nu_{\min}), \text{ with } \tilde{\nu}_{\alpha\gamma} = \max\left(\sum_{p} V^{(p)}_{\alpha\gamma}, V^{(cc)}_{\alpha\gamma}\right) \tag{9}$$

where θ is the Heaviside function, and $\nu_{min}$ is a threshold for the minimal contribution that will be considered. $\tilde{\nu}_{\alpha\gamma}$ is the sum of all generalized volume contributions (equation (3)) taken over all the peaks with an assignment (α,γ), where $V^{(cc)}_{\alpha\gamma}$ ensures that there is a minimal contribution for pairs of hydrogen atoms with intraresidual and sequential relative positioning:

$$V^{(cc)}_{\alpha\gamma} = \begin{cases} \nu_{max} & \text{if } u^{(cc)}_{\alpha\gamma} \leq d_{max}; \\ \nu_{min} & \text{all other intraresidual or sequential combinations;} \\ 0 & \text{all longer-range connectivities.} \end{cases} \tag{10}$$

The calculation of the network-anchoring contribution by equations (8)–(10) is recursive in the sense that its evaluation for a given peak requires the knowledge of the generalized volume contributions (equation (3)) from other peaks, which in turn is a function of some of these same network-anchoring contributions. Therefore, the calculation of these quantities is iterated alternatingly three times. If separate peak lists from different NOESY

spectra are used, the peaks from all peak lists contribute simultaneously to network-anchoring.

Finally, a measure of residue-wise network-anchoring between residues $A$ and $B$, defined by:

$$\bar{N}_{AB} = \sum_{\alpha \in A} \sum_{\beta \in B} N_{\alpha\beta} \tag{11}$$

is computed by CANDID.

### Elimination of low-ranking initial assignments

The list of initial assignments is screened for high values of the generalized contributions, $V_k$, (equation (3)) to the total peak volume, and only those assignments are retained for which $V_k \geq V_{\min}$.[15] These assignments will then be used to generate distance constraints for the structure calculation.

In the last CANDID cycle, the remaining ambiguous NOE assignments are either replaced with unambiguous assignments to unique pairs of protons or the corresponding NOESY cross-peaks are discarded as a source of information for the structure calculation. Technically, this filtering is achieved by increasing the acceptable minimal generalized volume contribution for an initial assignment to be retained in the input for the structure calculation, $V_{\min}$, to a value larger than 50%.

### Calibration of NOE upper distance constraints

A NOESY cross-peak with a single assignment, (α,β), gives rise to an upper bound, $b$, on the distance between the two hydrogen atoms, α and β, $d_{\alpha\beta}$, in the molecular structure: $d_{\alpha\beta} \leq b$. A NOESY cross-peak with $n \geq 2$ assignments can be interpreted as the superposition of $n$ signals giving rise to an ambiguous distance constraint:[19,20]

$$\bar{d} = \left( \sum_{k=1}^{n} d_{\alpha_k \beta_k}^{-6} \right)^{-1/6} \leq b \tag{12}$$

Each of the distances $d_{\alpha_k\beta_k}$ corresponds to one assignment, $(\alpha_k,\beta_k)$. In CANDID the upper distance bound $b$ from a peak $p$ in the NOESY spectrum $S$ with volume $I^p$ and $n$ assignments $(\alpha_k, \beta_k)$, $k = 1, \ldots, n$, is computed as:

$$b = \left( \sum_{k=1}^{n} \frac{I^p V_k}{\sqrt{Q_{\alpha_k}^S Q_{\beta_k}^S}} \right)^{-1/6} \tag{13}$$

$Q_\alpha^S$ and $Q_\beta^S$ are atomic calibration constants for the atoms α and β in the spectrum $S$. In common practice, equal calibration constants are used for given types of atoms, such as for all backbone atoms, all side-chain atoms, or all methyl groups.[9] CANDID provides a choice of three ways to set the atomic calibration constants, i.e. fixed user-defined calibration, automated structure-independent calibration, or automated structure-based calibration. Fixed calibration uses $Q_\alpha^S$ and $Q_\beta^S$ values chosen by the user, which will be held constant throughout all cycles of a CANDID calculation. The two automated methods do not require such explicit input from the user. Automated structure-independent calibration defines the calibration constant in such a manner that the average of the upper distance bounds for all peaks involving a given combination of atom types attains a predetermined value.[5] Structure-based automated calibration sets the calibration constant such that the available preliminary structure does not violate more than a predetermined percentage of the upper distance bounds.

### Elimination of spurious NOE cross-peaks

Identification and elimination of potentially erroneous NOE cross-peaks is an essential step in finding the correct protein 3D structure, since the experimental input data typically contain hardly avoidable imperfections. Usually, a limited number of peaks can therefore not be assigned correctly, e.g. because they correspond to noise artifacts, some peak positions have been determined with a larger error than the chemical shift tolerance, the chemical shift list is incomplete, some peak integrals have been severely overestimated, and similar.

To minimize deleterious effects on the resulting structure, CANDID applies four different peak filters, so that a distance constraint is derived from a given peak only if the following conditions are met: (i) at least one of the generalized contributions, $V_k$ (equation (3)), exceeds the threshold value $V_{\min}$; (ii) the number of assignments is below a user-defined maximal value $n_{\max}$; (iii) the corresponding distance constraint is not violated by more than a cutoff value, $d_{\text{cut}}$, in more than a user-defined percentage of the number of conformers used to represent a preliminary structure (only in the second and subsequent CANDID cycles, see Figure 1); (iv) the assignments of the peak are well anchored in the network of the assignments of all peaks, as quantified by the following: $\langle f \rangle_p = \sum_{k=1}^{n} V_k f_k$ is the average of a quantity, $f$, over the $n$ assignments of a peak, $p$, weighted by their generalized volume contributions, $V_k$. To be accepted, a peak, $p$, must either satisfy the single condition of having a high average network-anchoring per residue (equation (11)), $\langle \bar{N} \rangle_p \geq \bar{N}_{\text{high}}$, or the combined condition of having a minimal value of average network-anchoring per residue and per atom, $\langle \bar{N} \rangle_p \geq \bar{N}_{\min}$ and $\langle N \rangle_p \geq N_{\min}$, where $\bar{N}_{\text{high}}$, $\bar{N}_{\min}$ and $N_{\min}$ are constant parameters of CANDID, with $\bar{N}_{\text{high}} > \bar{N}_{\min}$.

### Constraint-combination

In the practice of NMR structure determination with biological macromolecules, spurious distance constraints in the input may arise from misinterpretation of stochastic noise, and similar, as

well as from real signals that involve atoms that are not included in the chemical shift list. This situation is particularly critical at the outset of a structure determination, before the availability of a preliminary structure for 3D structure-based screening 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 early CANDID cycles (the first two cycles in the calculations here, unless noted otherwise). It consists of generating virtual distance constraints with combined assignments from different, in general unrelated cross-peaks. The basic property of ambiguous distance constraints is that the constraint will be satisfied by the correct protein structure provided that at least one of the assignments in the combined constraint is correct. Overall, combined constraints therefore have a lower probability of being erroneous than the individual constraints.

Two different modes of constraint-combination have been implemented in CANDID (further combination modes can readily be envisaged), i.e. "$2 \rightarrow 1$ combination" of all long-range assignments of two peaks into a single new, virtual constraint, and "$4 \rightarrow 4$ pairwise combination" of the long-range assignments of four peaks into four new, virtual constraints. Constraint-combination is applied only to the long-range peaks, i.e. peaks for which all assignments are to pairs of atoms separated by at least five residues in the sequence,[1] because the effect of an erroneous long-range constraint on the global fold of a protein is much stronger than that of erroneous short and medium-range constraints. For further description of the two modes of constraint-combination we denote the set of all assignments of a given peak with an upper case letter. $2 \rightarrow 1$ combination replaces two constraints with the assignment sets $A$ and $B$ by a single, ambiguous constraints with assignment set $A \cup B$ (the union of sets $A$ and $B$). $4 \rightarrow 4$ pairwise combination replaces four constraints with the assignment sets, $A$, $B$, $C$ and $D$ by four ambiguous constraints with assignment sets $A \cup B$, $A \cup C$, $A \cup D$ and $B \cup C$. To increase the efficiency of $4 \rightarrow 4$ pairwise combination, the long-range peaks are sorted according to their total residue-wise network-anchoring, i.e. by the sum of the quantities defined in equation (11) over all assignments of the peak, and the assignment sets $A$, $B$, $C$ and $D$ are selected from the first, second, third, and fourth quarter of the sorted peak list, respectively. The number of long-range constraints is halved by $2 \rightarrow 1$ combination, whereas $4 \rightarrow 4$ pairwise combination preserves more of the intrinsic structural information, since the number of constraints is unchanged. It furthermore takes into account that certain peaks and their assignments are more reliably documented than others, because in the combined constraints the assignment sets $A$, $B$, $C$ and $D$ are used 3, 2, 2 and 1 times, respectively.

A quantitative impression of the effect of constraint-combination on the input for a structure calculation can be obtained from the following considerations. For an experimental data set containing $N$ long-range peaks, we assume a uniform probability, $p \ll 1$, that any one of these peaks would lead to an erroneous constraint. By $2 \rightarrow 1$ constraint-combination the $N$ experimental constraints are substituted by $N/2$ virtual constraints with a uniform error probability $p^2 \ll p$. In the case of $4 \rightarrow 4$ constraint-combination, we assume that $N$ long-range peaks can be grouped into four classes with error probabilities $\alpha p$, $p$, $p$ and $(2-\alpha)p$, so that the overall probability for an input constraint to be erroneous is again $p$. The parameter $\alpha$, with $0 \leq \alpha \leq 1$, accounts for the stronger experimental evidence for the presence of the peaks in the first class when compared to those in the two middle classes and the fourth class (see above). The $N$ virtual long-range constraints obtained after $4 \rightarrow 4$ combination have an overall error probability of $(\alpha + (1 - \alpha^2)/4)p^2$, which is smaller than $p^2$ for $\alpha < 1$, i.e. whenever the ranking of the experimental constraints for variable reliability was successful. For example, $4 \rightarrow 4$ constraint-combination with $\alpha = 0.5$ will transform an input data set of 900 correct and 100 erroneous long-range cross-peaks (i.e. $N = 1000$, $p = 0.1$) into a new set of approximately 993 correct and seven erroneous virtual combined constraints. For the same system, $2 \rightarrow 1$ constraint-combination will yield approximately 495 correct and five erroneous virtual combined constraints. For all calculations here, $4 \rightarrow 4$ constraint-combination was used in the first two CANDID cycles, unless noted otherwise.

The upper distance bound $b$ for a virtual combined constraint is derived from the two upper distance bounds $b_1$ and $b_2$ of the two parent experimental constraints either as the $r^{-6}$-sum, $b = (b_1^{-6} + b_2^{-6})^{-1/6}$ (which was used for the calculations here), or as the maximum, $b = \max(b_1, b_2)$. The first choice minimizes the loss of information in situations where two correct constraints are combined, whereas the second choice avoids introducing a spurious shorter upper bound if a correct and an erroneous constraint are combined.

### Structure calculation using DYANA torsion angle dynamics

The program DYANA was adapted for automated structure determination in conjunction with CANDID, so that it accepts ambiguous distance constraints in the input and generally interacts efficiently with CANDID. Each CANDID cycle is completed by a structure calculation using the fast DYANA torsion angle dynamics algorithm with the standard simulated annealing schedule, whereby the input comprises the list of distance constraints from CANDID, and possibly additional conformational constraints from other sources (Figure 1). To minimize loss of information,

**Table 1.** Experimental chemical shift assignments and NOESY peak lists used for the final structure calculations based on interactive analysis of the NOSEY spectra of four proteins that have been used here to validate the CANDID procedure

| Protein[a] | Size (residues) | Assigned chemical shifts (%)[b] | NOESY spectra[c] | Peaks picked[d] |
|---|---|---|---|---|
| CopZ | 68 | 94.9 | 2D, $H_2O$ | 1175 |
| | | | 3D ($^{15}$N), $H_2O$ | 1063 |
| WmKT | 88 | 97.0 | 2D, $H_2O$ | 1998 |
| bPrP(121–230) | 110 | 98.1 | 3D ($^{15}$N), $H_2O$ | 1893 |
| | | | 3D ($^{13}$C), $H_2O$ | 3859 |
| P14a | 135 | 99.4 | 3D ($^{15}$N), $H_2O$ | 1457 |
| | | | 3D ($^{13}$C), $H_2O$ | 3055 |
| | | | 2D, $H_2O$ | 1925 |
| | | | 2D, $^2H_2O$ | 2001 |

[a] CopZ: apo-from of the copper chaperone Z;[38] WmKT: killer toxin from the yeast *Williopsis mrakii*;[32] bPrP(121–230): globular domain of the bovine prion protein comprising residues 121–230;[39] P14a: pathogenesis-related protein from tomato leaves.[36]
[b] Percent of the total number of non-labile hydrogen atoms and backbone amide protons for which the chemical shifts were assigned. Pairs of diastereotopic protons or methyl groups are considered to be assigned when at least one of the $^1$H chemical shifts is known.
[c] Notation used: 2D, 2D [$^1$H,$^1$H]-NOESY; $H_2O$, solvent of 95% $H_2O$/5% $^2H_2O$; $^2H_2O$, solvent of 100% $^2H_2O$; 3D ($^{15}$N), 3D $^{15}$N-resolved [$^1$H,$^1$H]-NOESY; 3D ($^{13}$C), 3D $^{13}$C-resolved [$^1$H,$^1$H]-NOESY.
[d] Number of interactively picked NOESY cross-peaks. Not all of these peaks were assigned by the spectroscopist in the final stage of the structure determination (see Table 2 below).

constraints relating to degenerate groups of protons are expanded into ambiguous distance constraints involving all hydrogen atoms of the corresponding degenerate group(s),[25] and the absence of stereospecific assignments for diastereotopic groups is treated by periodic optimal swapping of the pairs of diastereotopic atoms for minimal target function value during the simulated annealing.[26]

**Table 2.** Experimental input for the final structure calculations of the four test proteins of Table 1 and statistics of the structure determinations using either interactive or automated NOESY assignment

| Quantity | CopZ | WmKT | bPrP(121–230) | P14a |
|---|---|---|---|---|
| A. *CANDID* | | | | |
| NOE cross-peaks assigned[a] | 1025 | 1865 | 1670 | 1340 |
| | 887 | | 3538 | 2916 |
| | | | | 1674 |
| | | | | 1715 |
| NOE upper distance limits[b] | 937 | 1223 | 2091 | 1885 |
| Residual DYANA target function (Å$^2$)[c] | 1.56 | 1.74 | 2.19 | 4.16 |
| RMSD (Å)[d] | 0.53 | 0.76 | 0.57 | 0.82 |
| B. *Interactive structure determination* | | | | |
| NOE cross-peaks assigned[a] | 1024 | 1421 | 1493 | 1248 |
| | 947 | | 3310 | 2636 |
| | | | | 244 |
| | | | | 210 |
| NOE upper distance limits[b] | 993 | 1053 | 1797 | 1701 |
| Residual DYANA target function (Å$^2$)[c] | 1.67 | 3.51 | 0.79 | 3.13 |
| RMSD (Å$^2$)[d] | 0.42 | 0.68 | 0.58 | 0.88 |
| C. *Comparisons between CANDID and interactive assignment* | | | | |
| Peaks with identical assignment (%)[e] | 92.3 | 92.3 | 94.1 | 92.6 |
| Peaks with different assignments (%)[f] | 4.7 | 4.0 | 3.2 | 2.4 |
| Peaks assigned interactively but not by CANDID (%) | 3.0 | 3.7 | 2.7 | 5.0 |
| Average rank of the interactive assignment after cycle 1[g] | 1.14 | 1.08 | 1.11 | 1.08 |
| Average rank of the interactive assignment after cycle 6[g] | 1.06 | 1.05 | 1.04 | 1.03 |
| RMSD between mean structures (Å) | 0.67 | 0.80 | 1.11 | 1.09 |

[a] On the basis of the experimental input data of Table 1. The number of assigned NOE cross-peaks given for each protein from top to bottom corresponds to the listing of the NOESY spectra in Table 1.
[b] Number of NOE upper distance limits that represent conformational restraints on the polypeptide fold.
[c] The residual DYANA target function value is the average for the bundles of conformers representing the NMR structure. The target function values before energy minimization are given.
[d] The RMSD is the average of the RMSD values between the individual conformers in the bundle and their mean coordinates for the backbone atoms N, $C^\alpha$, $C'$ of residues 2–67 for CopZ, 4–39 and 47–87 for WmKT, 128–166 and 172–223 for bPrP(121–230), and 2–134 for P14a. The RMSD values after energy minimization are given.
[e] Peaks with identical assignments are those for which the interactive assignment is the same as the assignment made by CANDID.
[f] Peaks with different assignment have been assigned by both approaches, but the interactive assignment is to a different pair of protons than by CANDID.
[g] The initial assignment of each NOE cross-peak are sorted by decreasing generalized volume contribution (equation (3)). The average rank given for an assignment from the interactive approach is relative to all retained CANDID assignments.

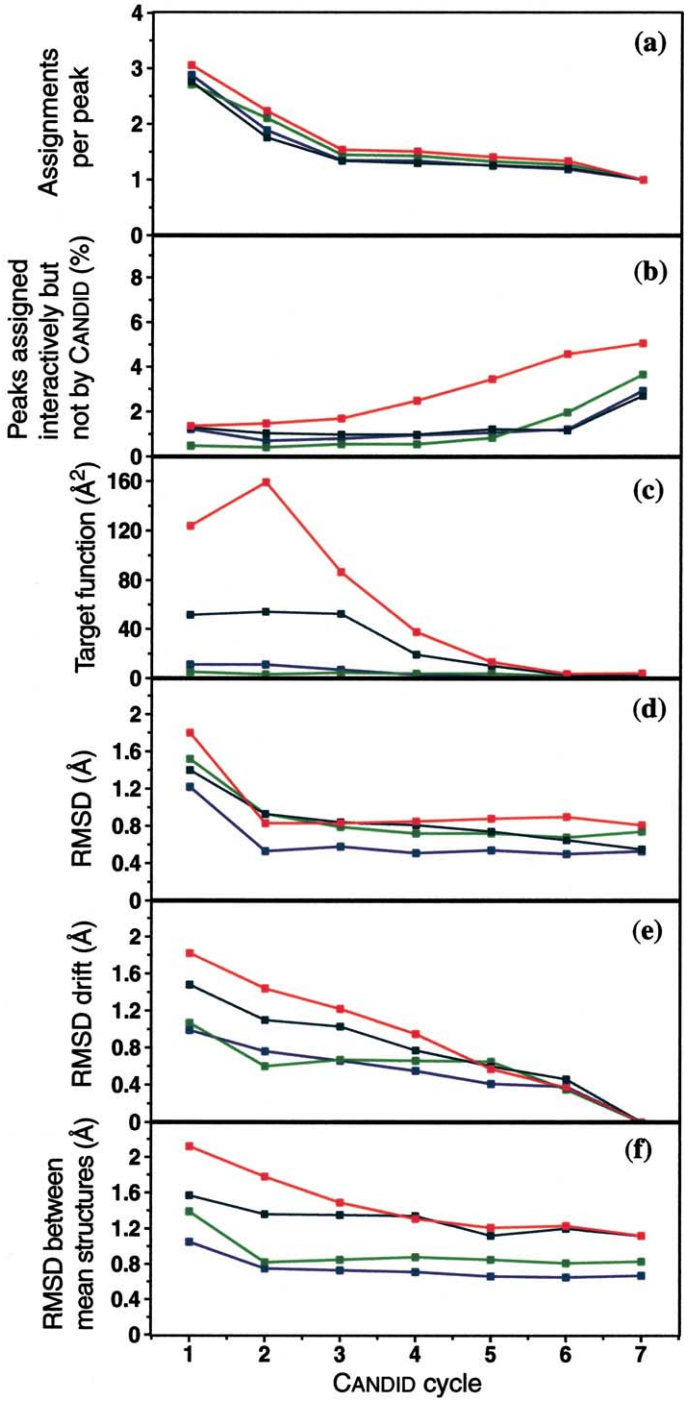


**Figure 2**. Evolution of characteristic parameters for NMR structures in the course of the seven cycles of CANDID structure calculation for the four proteins CopZ (blue), WmKT (green), bPrP (121–130) (black) and P14a (red). (a) Average number of assignments per NOE cross-peak retained in the input for the structure calculation. In cycle 7, an additional filtering step enforces the value 1.0 (see text). (b) Percentage of interactively assigned peaks that were discarded by CANDID. (c) Average final target function value for the bundle of conformers representing the result of the DYANA structure calculation. (d) RMSD calculated as the average of the RMSD values between the individual conformers in the bundles and their mean coordinates. (e) RMSD drift, calculated as the RMSD between the mean coordinates of the bundles of conformers obtained after the *k*th and the seventh CANDID cycles. (f) RMSD between mean structures, calculated as the RMSD between the mean coordinates of the bundles of conformers obtained after the *k*th CANDID cycle and the final result of the interactive reference structure determination. All RMSD values are calculated for the backbone atoms N, $C^{\alpha}$ and $C'$ of the well-defined segments of the polypeptide chains given in Table 2.

## Results

### CANDID calculations with experimental data sets

Automated combined structure determination and NOE assignment with CANDID was validated using experimental input data set that had been prepared for previous conventional NMR structure determination of four proteins (Table 1; see also Materials and Methods). The four proteins represent different molecular sizes, different secondary structure types, and different isotope labeling strategies (Table 1). Between 12% and 49% of the peaks that had been picked were left unassigned in the conventional structure determinations of the four proteins (Tables 1 and 2). For the CANDID calculation the unassigned NOESY peak lists were used, and seven cycles of CANDID assignment and DYANA structure calculation were performed (see Materials and Methods).

Table 2 provides an overview of the results. Between 88% and 93% of all NOESY peaks were assigned by CANDID, and for all proteins a low final target function value and a small RMSD were obtained, which is by conventional criteria

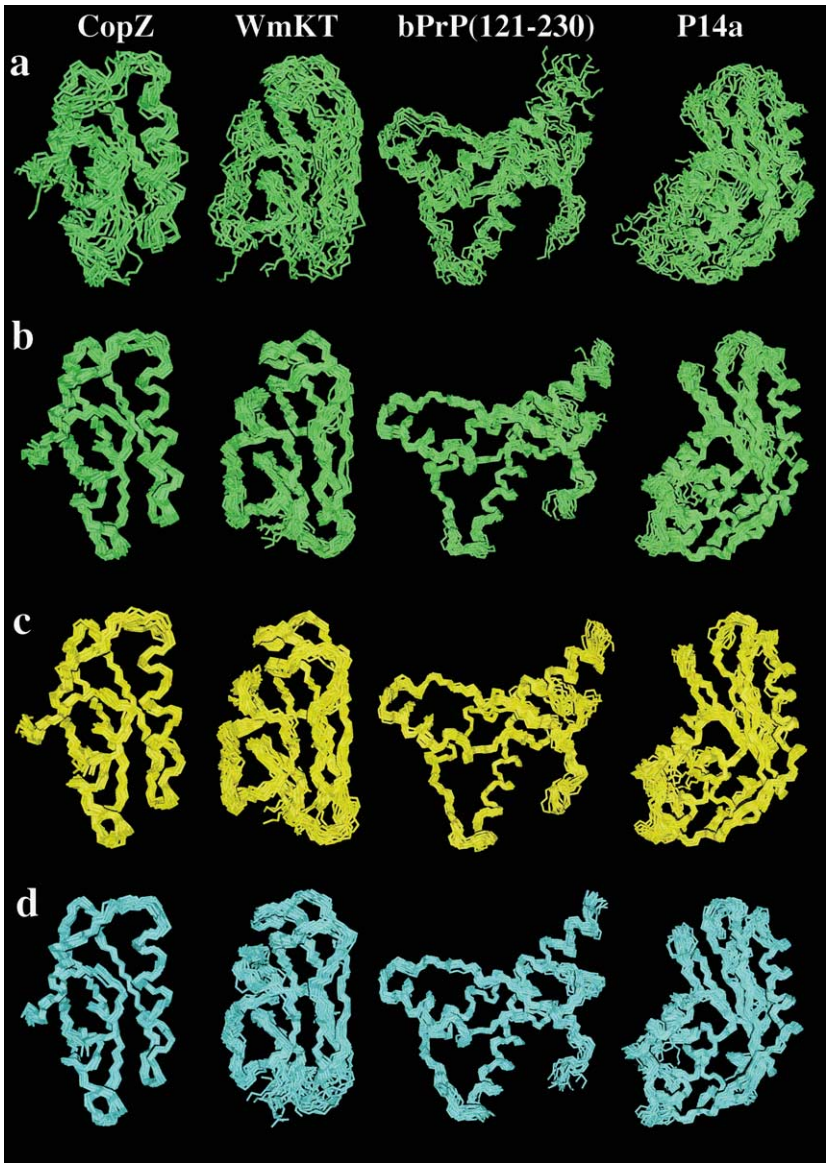


**Figure 3**. Bundles of conformers of the four proteins used for the validation of the CANDID/DYANA procedure. (a) CANDID/DYANA cycle 1 (10 conformers); (b) CANDID/DYANA cycle 6 (20 conformers); (c) CANDID/DYANA cycle 7 (20 conformers after energy-refinement); (d) structure from the previous interactive determination (20 conformers after energy-refinement).

indicative of a high-quality structure determination. The evolution of various quality parameters in the course of the seven CANDID cycles is illustrated in Figure 2. The average number of ambiguous assignments per cross-peak (Figure 2(a)) decreases from about three in cycle 1 to about 1.5 in cycle 6, whereby in cycle 6 most of the ambiguous assignments come from peaks with either two or three possible assignments. In cycle 7 unambiguous assignment is enforced. The fraction of interactively assigned peaks that are left unassigned by CANDID remains below 6% throughout, and varies only slightly in the course of the calculation (Figure 2(b)). A slight increase of this parameter towards the later cycles reflects the increasing stringency of the filtering criteria. The presence of a significant number of artifact cross-peaks in the input peak list, which had been discarded in the conventional structure determinations, manifests itself in relatively high values of the DYANA target function in the early cycles of the calculations for bPrP (121–230) and P14a (Figure 2(c)). Nonetheless, because network-anchoring detected and eliminated many of the artifact peaks, and constraint-combination reduced the impact of the remaining artifacts, the calculations converged to quite well-defined structures already in the first CANDID cycle (Figure 2(d)). Improved precision is achieved during the cycles 2–7 (Figure 2(d)). The system is stable in the sense that the RMSD drift of the mean coordinates is small and decreases monotonously towards the final structure during the entire CANDID calculations (Figure 2(e)). This important result is also apparent from the bundles of conformers obtained after the CANDID cycles 1, 6 and 7 (Figure 3(a)–(c)), which show defined structures for cycle 1 with readily apparent similarity with the final structure (see also Figures 2(f) and 3(c)). Finding a defined and correct fold in the first cycle is the key to reliable automated NOESY assignment and structure calculation, since subsequent cycles are

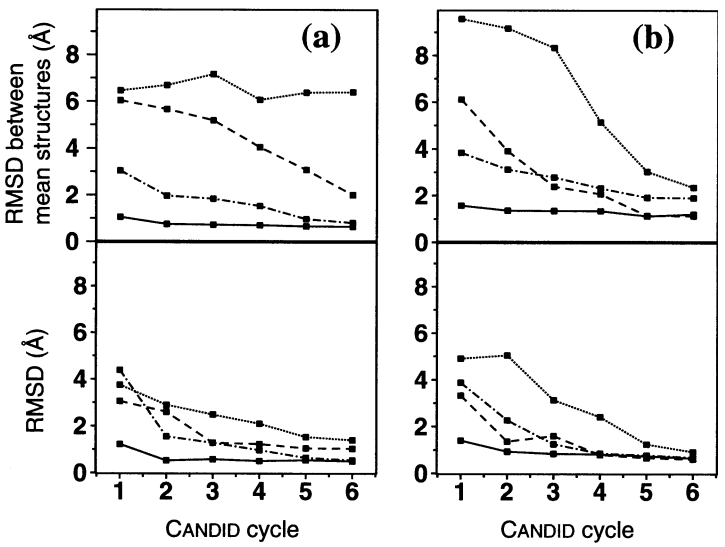


**Figure 4**. Impact of using network-anchoring and constraint-combination in the CANDID/DYANA calculation of the proteins CopZ (a) and bPrP(121–230) (b). The evolution of the average RMSD for the bundle of conformers ("precision"; lower panels) and of the RMSD between the mean structures obtained by the *k*th CANDID cycle and the final structure from the interactive approach ("accuracy"; upper panels) are shown. The same CANDID protocol was used throughout, except that network-anchoring, and constraint-combination were only used as follows: no network-anchoring, no constraint-combination (dotted); network-anchoring, no constraint-combination (broken); no network-anchoring, constraint-combination (dot-dashed); network-anchoring and constraint-combination (continuous).

driven by intermediary 3D structure-based assignment and peak filters (Figure 1).

### Comparison with conventional, interactive structure determination

For these comparisons we used exclusively the results obtained after the CANDID/DYANA cycle 7 and restrained energy-refinement in a water-shell with the program OPALP,[27,28] using the AMBER force field[29] (see Materials and Methods), both because the input for the calculation of the reference structures contained only unambiguous distance constraints, and because these structures had also been energy-refined with OPAL[28] or OPALP.[27]

The input data of NOE upper distance constraints obtained with CANDID are in very good overall agreement with those of the interactive structure determinations that are used as a reference (Tables 1 and 2). For all peak lists except the one for CopZ, a slightly larger number of peaks were assigned by CANDID than had been assigned interactively (Table 2), which indicates that an even more thorough use of the spectral information was possible with CANDID than with the interactive approaches. The overwhelming majority of NOESY cross-peaks, i.e. between 92% and 94% of all peaks assigned previously by interactive approaches for the four proteins, have identical assignments (Table 2). For only between 2.4% and 4.7% of the peaks the interactive approach and CANDID yielded different assignments, and between 2.7% and 5.0% of the peaks that had previously been assigned interactively were not assigned by CANDID. These include some uncertain interactive assignments, but the differences have primarily been caused by a slightly different calibration of the distance constraints. Whereas $r^{-6}$ and $r^{-4}$ relationships between peak volume and upper distance bounds were used in the interactive approach,[9] $r^{-6}$ relationships were used with CANDID (Table 5). Furthermore, the two approaches used different treatments of degenerate diastereotopic groups of protons, i.e. pseudoatoms in the interactive calculation of the reference structures, and $r^{-6}$-summation with swapping of diastereotopic atoms for the structures based on CANDID assignments.[25,26] These different treatments result in slightly different values for some of the upper distance bounds. For example, in cases where CANDID uses a tighter upper bound than the interactive approach, this can lead to a consistent constraint violation and hence to elimination of the corresponding NOESY cross-peak from the input for the structure calculation.

The close coincidence of the NOE assignments obtained with CANDID and in the reference structure determinations leads to structures that are closely similar and of comparable quality, as shown by the target function and RMSD values of Table 2. With both approaches, the final target function values are in the range 1–4 Å$^2$ (1Å = 0.1 nm), and the RMSD values vary between 0.4 and 0.9 Å. The RMSD values between the mean structures obtained by the interactive and the automatic approach are in the range 0.7–1.1 Å, which is slightly less than the sum of the average RMSD values for the two bundles of conformers. This good agreement is evident also by inspection of the structures in Figure 3(c) and (d), which show visible deviations almost exclusively for surface loop regions. Evaluation of the stereochemical quality of the energy-minimized protein structures determined by the automated CANDID approach with the program PROCHECK[30] resulted in very similar statistics of the Ramachandran plots as for the reference structures. The CANDID structures have between 93% and 97% of the residues in the

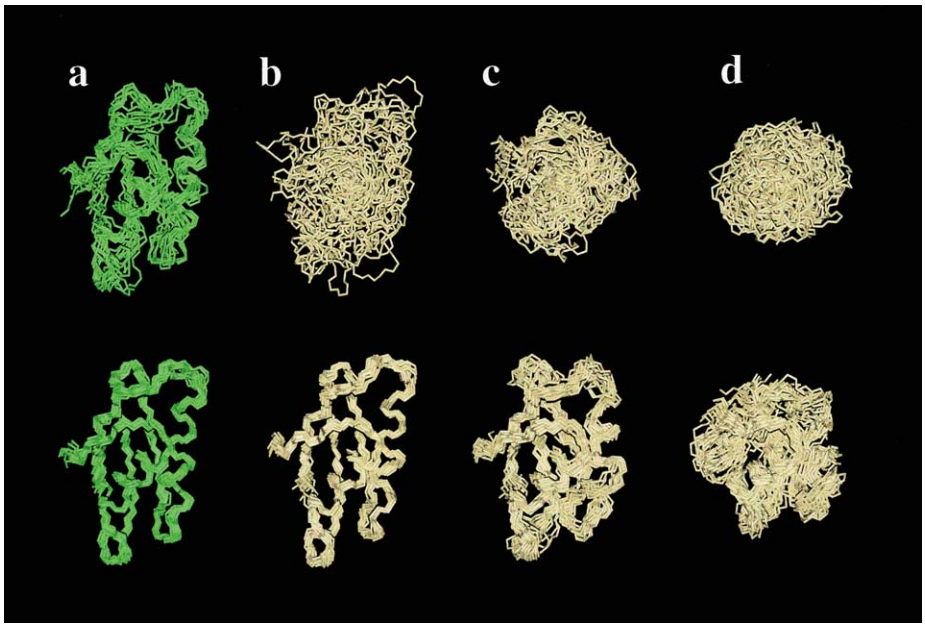


**Figure 5**. Bundles of the conformers with the lowest residual target function values for the protein CopZ after the CANDID cycles 1 (top) and 6 (bottom). The structures were obtained with the four CANDID calculations of Figure 4: (a) using network-anchoring and constraint-combination; (b) no network-anchoring, constraint-combination; (c) network-anchoring, no constraint-combination; (d) no network-anchoring, no constraint-combination.

"most favored" and "additional allowed" regions, as defined by PROCHECK,[30] whereas the corresponding values for the interactive structure determinations are in the range from 95% to 99%.

## Effect of network-anchoring and constraint-combination

To assess the impact of the presently introduced techniques of network-anchoring and constraint-combination on automated NOESY assignment, the CANDID calculations with the data sets of Table 1 were repeated using the same protocol except that either network-anchoring, or constraint-combination, or both were inactivated. The results for CopZ and bPrP(121–230) are shown in Figures 4 and 5 and Table 3 (similar results were obtained for the other two proteins). The standard CANDID schedule using both network-anchoring and constraint-combination yielded in all cases the closest structure to the reference. If one or both of the new techniques were inactivated, the structures after cycle 1 were much more distorted, as manifested by RMSD values between the mean structures obtained by CANDID and the interactive approach of more than 3.0 Å (Figure 4). For CopZ, the computation without network-anchoring and constraint-combination yielded a relatively precisely defined but severely erroneous structure (Figure 4(a)). In the other computations of Figure 4, the convergence towards the correct structure was always fastest when using the standard schedule, and later cycles in the variant protocols were in most instances able to largely correct the severe distortions present in the structures from the first cycle. The fact that the RMSD values

**Table 3.** Consistency with the conventionally determined reference structures of CopZ and bPrP(121–230) of the distance constraints from cycle 1 of CANDID obtained with and without network-anchoring and/or constraint-combination

| Network-anchoring | Constraint-combination | CopZ | | bPrP(121–230) | |
|---|---|---|---|---|---|
| | | Target function[a] (Å$^2$) | Constraint violations[b] > 5 Å | Target function[a] (Å$^2$) | Constraint violations[b] > 5 Å |
| Yes | Yes | 53 | 0 | 385 | 2 |
| No | Yes | 221 | 2 | 605 | 5 |
| Yes | No | 3399 | 25 | 8136 | 43 |
| No | No | 6714 | 56 | 14842 | 98 |

[a] Value of the DYANA target function calculated for the NOE distance constraints of the first CANDID cycle relative to the reference structure. The mean value for the bundle of 20 conformers representing the reference structure (Tables 1 and 2) is listed.
[b] Number of constraint violations larger than 5.0 Å counted for the input of NOE distance constraints of the first CANDID cycle relative to the reference structure.

within the bundles are smaller than the RMSD values between the mean structures reveals that the apparent precision may give a misleading indication of high accuracy of the structure in the absence of network-anchoring and/or constraint-combination. In the case of CopZ, constraint-combination was particularly important, whereas the protocol without network-anchoring yielded a good structure (Figure 4(a)). The observations for CopZ are visualized in Figure 5 with the results from the first and sixth CANDID cycles of calculations with and without network-anchoring and/or constraint-combination. For bPrP(121–230), the use of either network-anchoring or constraint-combination alone was not sufficient to attain comparable convergence and structure quality after the early cycles as when using the standard protocol with both of these two techniques activated (Figure 4(b)). Overall, the results of Table 3 and Figures 4 and 5 indicate that automated NOESY assignment without network-anchoring and constraint-combination is hardly reliable unless additional input data, such as a certain number of previously assigned long-range NOEs or a preliminary polypeptide fold, is available with the input for the first cycle of calculations.

The principal intended role of network-anchoring is to replace the initially unavailable checks for compatibility with the 3D structure within the automated NOESY assignment (Figure 1), whereas constraint-combination reduces the impact of unidentified artifact constraints in the input for the first structure calculation. The plots of Figure 2 illustrate that the results from cycle 1 are quite in line with those of the subsequent "structure-based" cycles, indicating that the envisaged goal had been largely attained. In particular, the low number of only about three possible assignments per peak already for the first cycle (Figure 2(a)) is a direct result of network-anchoring, since solely on the basis of chemical shift information, there would be 5.7, 9.2, 6.9 and 15.5 initial ambiguous assignments per peak for CopZ, WmKT, bPrP(121–230) and P14a, respectively. Network-anchoring is thus very effective in identifying correct assignments among all chemical-shift based initial assignments. This is also reflected by the fact that the interactively determined final reference assignment is usually at top rank if the initial assignments are sorted by decreasing generalized relative contributions (equation (3)). The average rank of the reference assignment therefore changes only from 1.08–1.14 in cycle 1 to 1.04–1.06 in cycle 6 for the different proteins studied (Table 2), so that for at least 86% of the peaks assigned by both methods the reference assignment is also the most highly weighted assignment by CANDID in cycle 1.

The effectiveness of network-anchoring and constraint-combination has also been assessed by evaluating the consistency of the distance constraints produced in the first CANDID cycles with the reference structure. To this end the DYANA target function was calculated (not minimized) and the number of severe constraint violations above a threshold value of 5.0 Å counted for the reference structure when using the NOE distance constraints of the first CANDID cycle as input (Table 3). The results clearly show that the CANDID standard protocol with network-anchoring and constraint-combination yields by far the best set of distance constraints and that constraint-combination is highly effective in reducing the impact of unidentified artifact peaks on the structure.

### *De novo* structure determination of WmKT using an automatically picked peak list

To assess the potential of using CANDID in connection with automation of further parts of the structure determination process, it was applied to a peak list for the 2D [$^1$H,$^1$H]-NOESY spectrum of WmKT that was created automatically by the program AUTOPSY.[31] The resulting list of 3789 peaks picked and integrated by AUTOPSY was used as input for CANDID, and automated NOESY assignment and structure determination were performed with the same protocol and using the same chemical shift list as with the interactively prepared WmKT peak list of Table 1 (since AUTOPSY picks peaks on both sides of the diagonal in a 2D [$^1$H,$^1$H]-NOESY spectrum, the number of peaks in the AUTOPSY peak list is about twice that of the interactively prepared peak list; see Materials and Methods).

Overall, the results obtained on the basis of the AUTOPSY peak list are similar to those from the interactively prepared WmKT peak list (Table 2, Figures 2 and 3). In the seventh CANDID cycle, 88.3% of the peaks were assigned, the final average target function value was 5.34 Å$^2$, the average RMSD value of the energy-refined bundle of conformers relative to the mean coordinates of the backbone atoms N, $C^\alpha$ and $C'$ of residues 4–39 and 47–87 was 0.56 Å, and the RMSD of the mean coordinates to those of the conventionally determined structure was 0.96 Å. Using the AUTOPSY peak list, a slightly lower percentage of the total number of NOESY peaks was thus assigned, the final target function values were somewhat higher and the final structure deviates slightly more from the reference structure. This data manifest the higher quality of the interactively prepared peak list, which had been optimized in multiple rounds of structure refinement and NOE assignment,[32] but the calculation on the basis of the AUTOPSY peak list clearly resulted also in a high-quality structure. In terms of precision as measured by the average RMSD value for the bundle of 20 energy-refined conformers, the automated CANDID method using AUTOPSY actually resulted in a somewhat better-defined structure than either the interactive approach[32] or CANDID using an interactively prepared peak list (Table 2).

## Discussion and Conclusions

This work documents that the CANDID approach for automated NOE assignment and protein structure calculation yields comparable NOE assignments and protein 3D structures to those obtained by conventional, interactive structure determination. The new concepts of network-anchoring and constraint-combination ensured in all applications so far that the correct fold of the protein was obtained already in the first cycle, which is the single most important advance achieved with CANDID when compared with previously proposed automated NOE assignment methods.[5,13,15] The potential of the CANDID procedure to become a generally applicable method for automated NOE assignment is supported by successful applications for structure determinations of three variant human prion proteins,[33] the calreticulin P-domain,[34] the pheromone-binding protein from *Bombyx mori*,[35] the human *Doppel* protein (to be published), and the chicken prion protein (to be published). These *de novo* structure determinations showed that automated NOE assignment with CANDID is faster and more objective than the conventional, interactive approach, so that the analysis of the NOESY spectra is no longer a time-limiting step in *de novo* protein structure determinations by NMR.

For the validation of the CANDID procedure for automated NOE assignment, the current version of CANDID has been interfaced with the program DYANA.[12] The CANDID program could in future applications be used also in conjunction with other structure calculation programs that can handle ambiguous distance constraints. For the evaluation of proper performance of CANDID, independent of the availability of an interactively determined reference structure, we propose a set of general guidelines (see (a)–(e) below). Most important, the CANDID input must include nearly complete sequence-specific resonance assignments, and CANDID cannot normally make up for lack of chemical shift assignments.

The guidelines (a)–(e) for checks on the successful performance of CANDID for automated NMR structure determination of globular proteins emerged from the test calculations here, and from experience gained in the aforementioned *de novo* structure determinations.[33–35] They consist of two straightforward requirements on the input chemical shift lists and NOESY cross-peak lists, (a) and (b), and of three output criteria to judge the reliability of the resulting structure, (c)–(e). All the presently described CANDID test calculations using network-anchoring and constraint-combination, as well as the applications for *de novo* structure determinations[33–35] fulfilled these five criteria.

(a) The input chemical shift list must contain more than 90% of the non-labile and backbone amide $^{1}$H chemical shifts. If 3D or 4D heteronuclear-resolved [$^{1}$H,$^{1}$H]-NOESY spectra are used, more than 90% of the $^{15}$N and/or $^{13}$C chemical shifts must also be available.

(b) The peak lists must be faithful representations of the NOESY spectra, and the chemical shift positions of the NOESY cross-peaks must be correctly calibrated to fit the chemical shift lists within the chemical shift tolerances. The range of allowed chemical shift variations ("tolerances") for $^{1}$H should not exceed $\pm 0.02$ ppm when working with homonuclear [$^{1}$H,$^{1}$H]-NOESY spectra, or $\pm 0.03$ ppm when working with heteronuclear-resolved 3D or 4D NOESY spectra, and the tolerances for the $^{15}$N and/or $^{13}$C shifts should not exceed $\pm 0.6$ ppm.

(c) The average final DYANA target function value for the bundle of conformers used to represent the structure from the first CANDID cycle should be below 250 Å$^2$, and the corresponding value for the last CANDID cycle should be below 10 Å$^2$, with more than 80% of all picked NOESY cross-peaks assigned and less than 20% of the peaks with exclusively long-range assignments[1] eliminated by the peak filters of CANDID. If CANDID is used with a different structure calculation algorithm, the parameters given for the target function will need to be redefined accordingly.

(d) The average backbone RMSD to the mean coordinates for the structured parts of the polypeptide chain should be below 3.0 Å for the bundle of conformers used to represent the structure from CANDID cycle 1.

(e) The RMSD drift between the mean atom coordinates after the first and the last CANDID cycles calculated for the backbone heavy atoms of the structured part of the polypeptide chain should be smaller than 3.0 Å, and it should not exceed the average RMSD to the mean coordinates after cycle 1 by more than 25%.

The input requirements (a) and (b) are imposed to ensure that in most instances the correct assignment is among the initial assignments of a cross-peak. Incomplete chemical shift lists make it impossible for CANDID to correctly assign any of the NOEs involving the unassigned atoms. Therefore, the assignment criterion (a) is particularly important for atoms with a large number of expected NOEs, such as the hydrogen atoms of the polypeptide backbone and the hydrophobic core side-chains, whereas incomplete chemical shift information for hydrophilic surface side-chains is more tolerable. The criterion (b) is needed because one often uses different NMR spectra for obtaining resonance assignments, and for the collection of the conformational constraints, respectively. Clearly, if the difference between the NOESY cross-peak positions and the chemical shift value of a given atom is larger than the tolerance range, then CANDID cannot make

correct NOE assignments. Such difficulties do not usually arise in structure determinations with homonuclear $^{1}H$ NMR, where the assignments are based on sequential NOEs observed in the same data sets that are also used for the collection of conformational constraints.[1] In projects with heteronuclear NMR, the chemical shift lists may need to be updated by reference to the corresponding NOESY cross-peak positions, using NOESY cross-peaks that have been assigned as part of the sequence-specific resonance assignment procedure, which typically makes use of numerous intra-residual and sequential NOEs, in addition to the data from heteronuclear triple resonance experiments.

The three output criteria (c)–(e) emphasize the crucial importance of getting good results from the first CANDID cycle. For reliable automated NMR structure determination, the bundle of conformers obtained after cycle 1 should be reasonably compatible with the input data (criterion (c)) and show a defined fold of the protein (criterion (d)). Structural changes between the first and subsequent CANDID cycles should occur within the conformation space determined by the bundle of conformers obtained after cycle 1, with the implicit assumption that this conformation space contains the correct fold of the protein (criterion (e)). The output criteria for target function and RMSD values might need to be slightly relaxed for proteins with more than 150 amino acid residues, and tightened for small proteins of less than 80 residues.

In principle, a *de novo* protein structure determination requires one round of 7 CANDID cycles (Figure 1). This is realistic for projects where an essentially complete chemical shift list is available and much effort was made to prepare a complete, high-quality input of NOESY peak lists. In practice, our experience so far has been that it may be more efficient to start a first round of CANDID analysis without excessive work for the preparation of the input peak list, using an incomplete list of "safely identifiable" NOESY cross-peaks, and then use the result of the first round of CANDID assignment and structure determination as additional information from which to prepare an improved, more complete NOESY peak list as input for a second round of 7 CANDID cycles.

## Materials and Methods

### Experimental NMR data sets used for the validation of automated structure determination with CANDID/DYANA

For the evaluation of the performance of CANDID/DYANA the experimental NMR data sets of four proteins were used for which high-quality NMR structures had previously been determined by a conventional, interactive approach (Tables 1 and 2; PDB entries: CopZ, 1CPZ; WmKT, 1WKT; bPrP(121–230), 1DWZ; P14a, 1CFE). For all four proteins nearly complete sequence-specific resonance assignments for the backbone and the side-chains are available. The chemical shifts of the $^{1}H$, $^{15}N$, and $^{13}C$ atoms were used as deposited in the BioMagResBank (accession codes: CopZ, 4344; WmKT, 5255; bPrP(121–130), 4563; P14a, 4301). A single chemical shift list was used for each of the proteins CopZ, WmKT and bPrP(121–130). For P14a, separate chemical shift lists were utilized for each of the four peak lists derived from four different NOESY data sets, as it had been done in the interactive structure determination in order to account for slight deviations between corresponding chemical shifts in the different NOESY spectra.[36]

The positions and volumes of all the peaks in the NOESY peak lists from the original structure determination were used as input for CANDID. These peak lists resulted from interactive peak picking with the program XEASY.[22] They contain peaks that had been unambiguously assigned and the corresponding upper distance constraints used for the structure calculation, as well as unassigned peaks that were not included in the input for the final structure calculation and may therefore be artifacts (Tables 1 and 2).

For all four proteins the experimentally determined $^{3}J$-coupling constants were used as in the previous, interactive structure determination. In each CANDID cycle, these scalar coupling constants were converted in conjunction with the updated list of NOE upper distance constraints into torsion angle constraints by the grid search procedure FOUND.[24] Stereospecific assignments from the interactive structure calculations were not included into the input for the CANDID/DYANA structure determination. Each disulfide bridge was constrained by a set of upper and lower distance constraints.[37]

To explore the potential of CANDID for more fully automated structure determination, an additional calculation was performed using a peak list for WmKT that had been produced automatically with the program AUTOPSY, as described by Koradi *et al*.[31]

### Standard protocol for automated structure determination with CANDID and DYANA

The CANDID/DYANA calculations comprised either six iterative cycles of NOESY assignment and structure calculation with the parameters given in Tables 4 and 5, or seven cycles for all calculations used for comparison with the reference structures of Table 1. Upper distance bounds were derived from NOESY cross-peak intensities according to equation (13). In the CANDID calculations only the two automated procedures for determining the calibration constants were applied, which ensured that no explicit input from the user was required (see Algorithms). The atomic calibration constants of the backbone and non-methyl β-protons were determined in the first CANDID cycle by automated, structure-independent calibration with a target average upper distance bound of 3.8 Å. In the following cycles, automated structure-based calibration was used, requiring that less than 15 and 10% of the constraints violate the input structures of cycles 2–3 and 4–7, respectively. For the remaining methyl and non-methyl protons the calibration constants were set to 3.0 and 1.5 times the value determined for backbone and non-methyl β-protons, respectively.

DYANA structure calculations using the standard simulated annealing schedule with 8000 torsion angle dynamics steps were started from 80 and 60 randomized

**Table 4.** Cycle-independent CANDID parameters used in the structure calculations here

| Symbol | Parameter | Value |
|---|---|---|
| $\Delta\omega_1^p$ | Tolerance range for peak positions in the indirect $^1$H dimension (equation (1)) | 0.02 ppm (bPrP: 0.03 ppm) (P14a: 0.025 ppm) |
| $\Delta\omega_2^p$ | Tolerance range for peak positions in the direct $^1$H dimension (equation (1)) | 0.02 ppm (P14a: 0.025 ppm) |
| $\Delta\omega_3^p$ | Tolerance range for peak positions in the $^{13}$C or $^{15}$N dimension (equation (2)) | 0.4 ppm |
| $\Delta\Omega_\alpha^S$ | Tolerance range of chemical shift (equations (1) and (2)) | 0.0 ppm |
| $\Gamma$ | Scaling factor for chemical shift agreement (equation (4)) | 0.5 |
| $d_{max}$ | Maximal value for covalent structure-constrained $^1$H–$^1$H distances for which a NOE is expected | 5.5 Å |
| $\nu_{min}$ | Minimal network-anchoring contribution for all other intraresidual and sequential distances (equation (10)) | 0.1 |
| $\nu_{max}$ | Maximal network-anchoring contribution for covalent structure-constrained distances (equation (10)) | 1.0 |
| $n_{max}$ | Acceptable maximal number of ambiguous assignments per peak | 20 |
| $M_{vio}$ | Acceptable maximal number of conformers with violation $> d_{cut}$ among all $M$ conformers | $M/2$ |
| $\bar{N}_{high}$ | Lower limit to qualify a cross-peak for having a "high network-anchoring per residue" (equation (11)) | 4.0 |

conformers in cycles 1 and 2, 3–7, respectively.[12] The 10 best conformers from the cycles 1 and 2, and the 20 best conformers from the cycles 3–6 were used as input structure for the next CANDID cycle.

Computations were performed on shared-memory multiprocessor Compaq computers using four Alpha processors in parallel for the structure calculations. The computation time for a complete automated structure determination with CANDID and DYANA ranged from 3.9 h for CopZ to 15.4 h for P14a on a single processor, and was spent predominantly in the DYANA structure calculation of, in total, 460 conformers per structure determination.

### Energy-minimization of the protein 3D structures using the program OPALP

Since the program DYANA does not optimize against a general energy force field that would also account for electrostatic interactions,[12] the bundles of conformers to be used for representation of the NMR structure were energy-minimized in a water-shell with the program OPALP,[27,28] using the AMBER force field.[29] OPALP accepts as input a bundle of conformers and the conformational constraints used in the final DYANA structure calculation, whereby the NOE upper distance constraints must be assigned to single pairs of hydrogen atoms. The energy-refined bundle of 20 conformers is used to represent the 3D NMR structure, which can then be directly compared with the reference structure determinations of Table 1. These have also used exclusively unambiguous NOE assignments in the input for the structure calculation, and were similarly refined with OPAL or OPALP.[32,36,38,39]

### Reference NOESY assignments and 3D NMR structures

The outcome of the automated CANDID structure determinations was evaluated by comparison with the published, interactively determined NOESY assignments and structures of CopZ, WmKT, bPrP(121–230), and P14a.[32,36,38,39] These structures had been calculated on the basis of the experimental data of Table 1 by the program DYANA[12] for CopZ and bPrP(121–230), and by its predecessor DIANA using the REDAC strategy for WmKT and P14a.[9,11] All reference calculations were thus performed in torsion angle space, with fixed values of the bond lengths, bond angles, planar groups and chiralities according to the ECEPP/2 force field,[40] and using the same functional form and weighting factors for the target function as in the presently introduced CANDID/DYANA protocol.

### Structure analysis and comparison

RMSD values are used for three different types of comparisons: The RMSD of a bundle of $n$ conformers is the average of the $n$ RMSD values between the individual conformers and the mean coordinates for the bundle. The RMSD between mean structures is the RMSD value between the mean coordinates of two

**Table 5.** Cycle-dependent CANDID parameters used in the structure calculations here

| Symbol | Parameter | Value in cycle $i$ ($i = 1, \ldots, 7$) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $V_{min}$ | Acceptable minimal volume contribution per assignment (%) (equation (3)) | 1.0 | 0.1 | 0.5 | 1.0 | 2.5 | 5.0 | 50.1 |
| T | Weight for presence of transposed peak (equation (3)) | 10 | 10 | 10 | 10 | 1 | 1 | 1 |
| O | Weight for covalent structure-constrained distances (equation (3)) | 10 | 10 | 10 | 1 | 1 | 1 | 1 |
| $\eta$ | Exponent for 3D structure-based volume contribution (equation (6)) | 3 | 3 | 6 | 6 | 6 | 6 | 6 |
| $S_{max}$ | Maximal structure-independent generalized volume contribution (equation (3)) | 20 | 20 | 20 | 20 | 10 | 10 | 10 |
| $d_{cut}$ | Upper limit on acceptable distance violation for elimination of spurious NOESY cross-peaks (Å) | – | 1.5 | 0.9 | 0.6 | 0.3 | 0.1 | 0.1 |
| $\bar{N}_{min}$ | Threshold for acceptable lower limit of network-anchoring per residue | 1.0 | 0.75 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| $N_{min}$ | Threshold for acceptable lower limit of network-anchoring per atom | 0.25 | 0.25 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |

bundles of conformers, for example, corresponding bundles obtained by CANDID and by the interactive approach. The RMSD drift for the CANDID cycle $k$ is the RMSD between the mean coordinates of the bundles of conformers obtained in the last cycle and in cycle $k$. The mean coordinates of a structure bundle are calculated by superimposing conformers $2, \ldots, n$ onto the first conformer for minimal RMSD for the backbone atoms N, $C^{\alpha}$ and $C'$ and subsequent calculation of the arithmetic average of the Cartesian coordinates. RMSD values were calculated for well-defined segments of the polypeptide chain, as published with the original structure determinations and given in Table 2. The program MOLMOL was used to visualize the 3D structures and for the calculation of RMSD values.[41]

### Implementation and availability of CANDID

The core of the current version of CANDID is implemented in standard Fortran-77 and has been built upon the data structures and into the framework of the user interface of the program DYANA. The standard schedule and parameters for a complete automated structure determination with CANDID and DYANA are specified in a script written in the interpreted command language INCLAN,[12] that gives the user high flexibility in the way automated structure determination is performed without need to modify the compiled core part of CANDID. The current versions of CANDID and DYANA are available from P.G.

## Acknowledgments

Financial support by the Schweizerischer Nationalfonds (project 31.49047.96), and the use of the high-performance computing facilities of ETH Zürich, EPF Lausanne, and the Centro Svizzero di Calcolo Scientifico are gratefully acknowledged.

## References

1. Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*, Wiley, New York.
2. Solomon, I. (1955). Relaxation processes in a system of two spins. *Phys. Rev.* **99**, 559–565.
3. Anil-Kumar, Ernst, R. R. & Wüthrich, K. (1980). A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton–proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Commun.* **95**, 1–6.
4. Macura, S. & Ernst, R. R. (1980). Elucidation of cross relaxation in liquids by 2D NMR spectroscopy. *Mol. Phys.* **41**, 95–117.
5. Mumenthaler, C., Güntert, P., Braun, W. & Wüthrich, K. (1997). Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J. Biomol. NMR*, **10**, 351–362.
6. Güntert, P., Berndt, K. D. & Wüthrich, K. (1993). The program ASNO for computer-supported collection of NOE upper distance constraints as input for protein structure determination. *J. Biomol. NMR*, **3**, 601–606.
7. Meadows, R. P., Olejniczak, E. T. & Fesik, S. W. (1994). A computer-based protocol for semi-automated assignments and 3D structure determination of proteins. *J. Biomol. NMR*, **4**, 79–96.
8. Duggan, B. M., Legge, G. B., Dyson, H. J. & Wright, P. E. (2001). SANE (structure assisted NOE evaluation): an automated model-based approach for NOE assignment. *J. Biomol. NMR*, **19**, 321–329.
9. Güntert, P., Braun, W. & Wüthrich, K. (1991). Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J. Mol. Biol.* **217**, 517–530.
10. Moseley, H. N. B. & Montelione, G. T. (1991). Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.* **9**, 635–642.
11. Güntert, P. & Wüthrich, K. (1991). Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints. *J. Biomol. NMR*, **1**, 446–456.
12. Güntert, P., Mumenthaler, C. & Wüthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298.
13. Mumenthaler, C. & Braun, W. (1995). Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J. Mol. Biol.* **254**, 465–480.
14. Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.* **245**, 645–660.
15. Nilges, M., Macias, M., O'Donoghue, S. I. & Oschkinat, H. (1997). Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. *J. Mol. Biol.* **269**, 408–422.
16. Brünger, A. T. (1992). *X-PLOR, Version 3.1. A System for X-ray Crystallography and NMR*, Yale University Press, New Haven, CT.
17. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W. *et al.* (1998). Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallog. sect. D*, **54**, 905–921.
18. Savarin, P., Zinn-Justin, S. & Gilquin, B. (2001). Variability in automated assignment of NOESY spectra and three-dimensional structure determination: a test case on three small disulfide-bonded proteins. *J. Biomol. NMR*, **19**, 49–62.
19. Nilges, M. (1993). A calculation strategy for the structure determination of symmetric dimers by $^{1}$H NMR. *Proteins: Struct. Funct. Genet.* **17**, 297–309.
20. Nilges, M. & O'Donoghue, S. I. (1998). Ambiguous NOEs and automated NOE assignment. *Prog. NMR Spectrosc.* **32**, 107–139.
21. Greenfield, N. J., Huang, Y. J., Palm, T., Swapna, G. V. T., Monleon, D., Montelione, G. T. & Hitchcock-DeGregori, S. E. (2001). Solution NMR structure and folding dynamics of the N terminus of a rat non-muscle alpha-tropomyosin in an engineered chimeric protein. *J. Mol. Biol.* **312**, 833–847.
22. Bartels, C., Xia, T., Billeter, M., Güntert, P. & Wüthrich, K. (1995). The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR*, **6**, 1–10.
23. Wüthrich, K., Billeter, M. & Braun, W. (1983). Pseudo-structures for the 20 common amino acids

for use in studies of protein conformation by measurements of intramolecular proton–proton distance constraints with nuclear magnetic resonance. *J. Mol. Biol.* **169**, 949–961.
24. Gütert, P., Billeter, M., Ohlenschläger, O., Brown, L. R. & Wüthrich, K. (1998). Conformational analysis of protein and nucleic acid fragments with the new grid search algorithm FOUND. *J. Biomol. NMR*, **12**, 543–548.
25. Fletcher, C. M., Jones, D. N. M., Diamond, R. & Neuhaus, D. (1996). Treatment of NOE constraints involving equivalent or nonstereoassigned protons in calculations of biomacromolecular structures. *J. Biomol. NMR*, **8**, 292–310.
26. Folmer, R. H. A., Hilbers, C. W., Konings, R. N. H. & Nilges, M. (1997). Floating stereospecific assignment revisited: application to an 18 kDa protein and comparison with *J*-coupling data. *J. Biomol. NMR*, **9**, 245–258.
27. Koradi, R., Billeter, M. & Güntert, P. (2000). Point-centered domain decomposition for parallel molecular dynamics simulation. *Comput. Phys. Commun.* **124**, 139–147.
28. Luginbühl, P., Güntert, P., Billeter, M. & Wüthrich, K. (1996). The new program OPAL for molecular dynamics simulations and energy refinements of biological macromolecules. *J. Biomol. NMR*, **8**, 136–146.
29. Cornell, W. D., Cieplak, P., Bayly, I., Gould, I. R., Merz, K. M., Ferguson, D. M. *et al.* (1996). A second generation force field for the simulation of proteins. Nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **118**, 2309–2309.
30. Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Struct. Funct. Genet.* **12**, 345–364.
31. Koradi, R., Billeter, M., Engeli, M., Güntert, P. & Wüthrich, K. (1998). Towards fully automatic peak picking and integration of biomolecular NMR spectra. *J. Magn. Reson.* **135**, 288–297.
32. Antuch, W., Güntert, P. & Wüthrich, K. (1996). Ancestral βγ-crystallin precursor structure in a yeast killer toxin. *Nature Struct. Biol.* **3**, 662–665.
33. Calzolai, L., Lysek, D. A., Güntert, P., von Schroetter, C., Riek, R., Zahn, R. & Wüthrich, K. (2000). NMR structures of three single-residue variants of the human prion protein. *Proc. Natl Acad. Sci. USA*, **97**, 8340–8345.
34. Ellgaard, L., Riek, R., Herrmann, T., Güntert, P., Braun, D., Helenius, A. & Wüthrich, K. (2001). NMR structure of the calreticulin P-domain. *Proc. Natl Acad. Sci. USA*, **98**, 3133–3138.
35. Horst, R., Damberger, F., Luginbühl, P., Güntert, P., Peng, G., Nikonova, L. *et al.* (2001). NMR structure reveals intramolecular regulation mechanism for pheromone binding and release. *Proc. Natl Acad. Sci. USA*, **98**, 14374–14379.
36. Fernández, C., Szyperski, T., Bruyère, T., Ramage, P., Mössinger, E. & Wüthrich, K. (1997). NMR solution structure of the pathogenesis-related protein P14a. *J. Mol. Biol.* **266**, 576–593.
37. Williamson, M., Havel, R. F. & Wüthrich, K. (1985). Solution conformation of proteinase inhibitor IIA from bull seminal plasma by $^1$H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.* **155**, 311–319.
38. Wimmer, R., Herrmann, T., Solioz, M. & Wüthrich, K. (1999). NMR structure and metal interactions of the CopZ copper chaperone. *J. Biol. Chem.* **274**, 22597–22603.
39. Lopez García, F., Zahn, R., Riek, R. & Wüthrich, K. (2000). NMR structure of the bovine prion protein. *Proc. Natl Acad. Sci. USA*, **97**, 8334–8339.
40. Némethy, G., Pottle, S. M. & Scheraga, H. A. (1983). Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem.* **87**, 1883–1887.
41. Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.

***Edited by M. F. Summers***

*(Received 23 January 2002; received in revised form 13 March 2002; accepted 13 March 2002)*