METHODS IN MOLECULAR BIOLOGY™

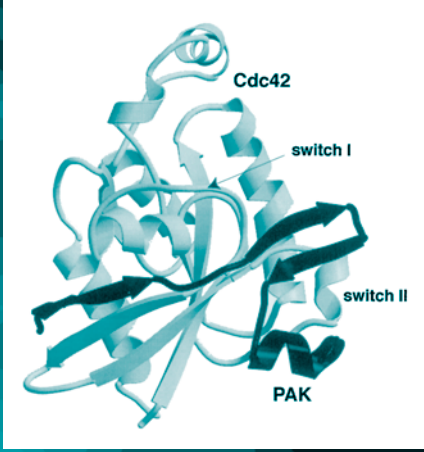
Volume 278

Protein NMR Techniques

SECOND EDITION

Edited by

A. Kristina Downing



HUMANA PRESS

17

Automated NMR Structure Calculation With CYANA

Peter Güntert

Summary

This chapter gives an introduction to automated nuclear magnetic resonance (NMR) structure calculation with the program CYANA. Given a sufficiently complete list of assigned chemical shifts and one or several lists of cross-peak positions and columes from two-, three-, or four-dimensional nuclear Overhauser effect spectroscopy (NOESY) spectra, the assignment of the NOESY cross-peaks and the three-dimensional structure of the protein in solution can be calculated automatically with CYANA.

Key Words: Protein structure; NMR structure determination; conformational constraints; automated structure determination; automated assignment; NOESY assignment; CYANA program; network anchoring; constraint combination; torsion angle dynamics.

1. Introduction

Until recently, nuclear magnetic resonance (NMR) protein structure determination has remained a laborious undertaking that occupied a trained spectroscopist over several months for each new protein structure. It was 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 ***(1)***. Today, automated methods for NMR structure determination are playing a more and more prominent role and are superseding the conventional manual approaches to solving three-dimensional (3D) protein structures in solution.

In *de novo* 3D structure determinations of proteins in solution by NMR spectroscopy, the key conformational data are upper distance limits derived from nuclear Overhauser effects (NOEs) ***(2–4)***. To extract distance constraints from a nuclear Overhauser effect spectroscopy (NOESY) spectrum, its cross-peaks have to be assigned—that is, the pairs of interacting hydrogen atoms have

From: *Methods in Molecular Biology, vol. 278: Protein NMR Techniques*
Edited by: A. K. Downing © Humana Press Inc., Totowa, NJ

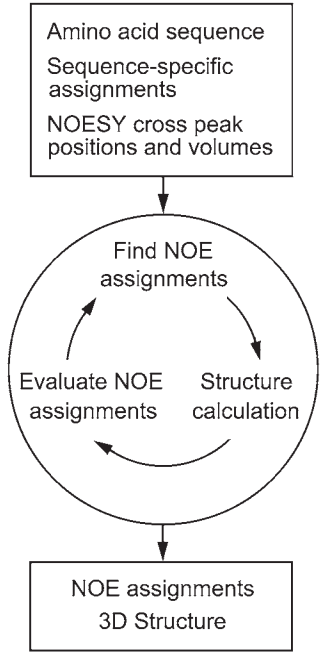


Fig. 1. General scheme of automated combined NOESY assignment and structure calculation.

to be identified (*see* **Note 1** for a summary of conventions used in this chapter). The NOESY assignment is based on previously determined chemical shift values that result from chemical shift assignment (*see* **Note 2**). 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 composed of multiple spin pairs (*see* **Note 3**). Obtaining a comprehensive set of distance constraints from a NOESY spectrum is by no means straightforward under these conditions 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 the 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 or conformational exchange. These inevitable shortcomings of NMR data collection are the main reason why until recently laborious interactive procedures have dominated 3D protein structure determinations. The automated NOESY assignment method CANDID ***(5)*** implemented in CYANA follows the same general scheme but does not require manual intervention during the assignment/structure calculation cycles (**Fig. 1**).

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 shifts is, as just pointed out, 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 substitute the intuitive decisions of an experienced spectroscopist in dealing with the imperfections of experimental NMR data by automated devices that can cope with the erroneously picked or inaccurately positioned peaks and with the incompleteness of the chemical shift assignment of typical experimental data sets. If used sensibly, automated NOESY assignment with CYANA has no disadvantage compared to the conventional, interactive approach but is a lot faster and more objective. With CYANA, the evaluation of NOESY spectra is no longer the time-limiting step in protein structure determination by NMR.

2. Materials

1. Computer with a UNIX -based operating system (e.g., Linux, Silicon Graphics IRIX, Compaq Alpha OSF1, IBM AIX, or MacOS X). The time-consuming structure calculations are most efficiently performed on a cluster of Linux computers using the message passing interface (MPI) for interprocess communication ***(6)***, or on a shared-memory multiprocessor system. A minimum of 256 megabytes of memory per processor is recommended.
2. CYANA software package for structure calculation using torsion angle dynamics, automated NOESY assignment, and structure analysis.
3. MOLMOL software package ***(7)*** for molecular graphics and structure analysis.
4. Input file with the amino acid sequence.
5. One or several input files with lists of cross-peaks from 2D [^{1}H,^{1}H]–NOESY, 3D or 4D ^{13}C- or ^{15}N-resolved [^{1}H,^{1}H] NOESY spectra. The input NOESY peak lists can be prepared either using interactive spectrum analysis programs such as XEASY ***(8)***, NMRView ***(9)***, ANSIG ***(10,11)***, or by automated peak-picking methods such as AUTOPSY ***(12)*** or ATNOS ***(13)***, which permit starting the NOE assignment and structure calculation process directly from the NOESY spectra. The peak lists must give the positions and volumes of the NOESY cross-peaks, but initial assignments are not required for the NOESY cross-peaks.
6. Input file(s) with ^{1}H and, if available, ^{13}C and ^{15}N chemical shifts in the format of the program XEASY ***(8)*** or of the BioMagResBank ***(14)***.
7. Optional: Previously assigned NOE upper distance constraints or other previously assigned conformational constraints. These will not be touched during automated NOE assignment but will be used for the CYANA structure calculation.

3. Methods

In this section, automated structure calculation with CYANA is described. Automated NOESY assignment is described in **Subheading 3.1.**, and structure

calculation by torsion angle dynamics-driven simulated annealing in **Subheading 3.2.** The effect of incomplete chemical shift assignment is discussed in **Subheading 3.4.**, quality control of structure calculations using automated NOESY assignment in **Subheading 3.5.**, and strategies to overcome problems with insufficient input data in **Subheading 3.5.** The approach presented here has been used successfully in many NMR structure determinations of proteins with hitherto unknown structure (*see* **Note 4**).

3.1. Automated NOESY Assignment

In the program, CYANA automated NOESY assignment is performed by the CANDID algorithm ***(5)*** that combines features from NOAH ***(15,16)*** and ARIA ***(17–20)***, such as the use of 3D structure-based filters and ambiguous distance constraints, with the new concepts of network anchoring and constraint combination that enable an efficient and reliable search for the correct fold already in the initial cycle of *de novo* NMR structure determinations.

3.1.1. Overview of the CANDID Algorithm for Automated NOE Assignment

In CYANA, the automated CANDID method ***(5)*** proceeds in iterative cycles of ambiguous NOE assignment followed by structure calculation using torsion angle dynamics (**Fig. 1**):

1. *Read experimental input data.* Amino acid sequence, chemical shift list from sequence-specific resonance assignment, list of NOESY cross-peak positions and volumes, and, optionally, conformational constraints from other sources for use in addition to the input from automated NOE assignment.
2. *Create initial assignment list.* For each NOESY cross-peak, one or several initial assignments are determined based on chemical shift agreement within a user-defined tolerance range.
3. *Rank initial assignments.* For each individual NOESY cross-peak the initial assignments are weighted with respect to several criteria, and initial assignments with low overall score are discarded. The 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* "network anchoring" in **Subheading 3.1.3.**), and, if available (i.e., in cycles 2, 3, . . .), the compatibility with the 3D structure from the preceding cycle (**Fig. 2**). The assessment of self-consistency also includes a check for the presence of symmetry-related cross-peaks.
4. *Calibrate distance constraints.* From the NOESY peak volumes or intensities upper distance bounds are derived for the corresponding, ambiguous or unambiguous distance constraints.
5. *Eliminate spurious NOESY cross-peaks*. Only those cross-peaks are retained that have at least one assignment with a network-anchoring score above a user-defined threshold and that are compatible with the intermediate 3D protein structure generated in the preceding cycle (cycles 2, 3, . . .).

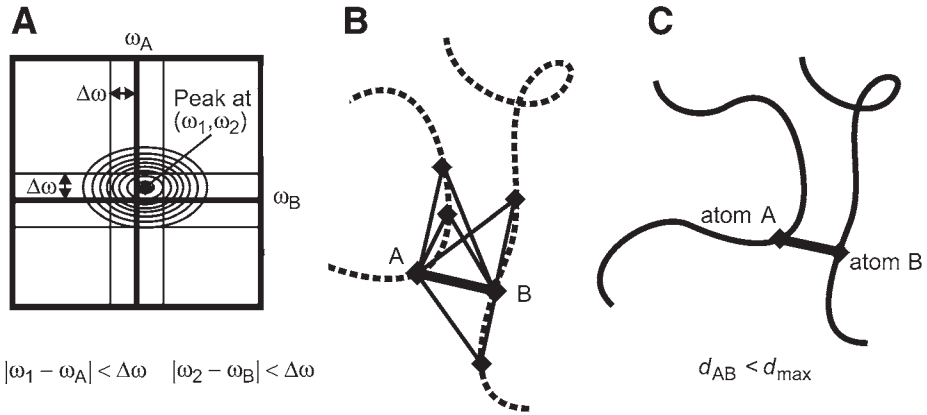


Fig. 2. Three conditions that must be fulfilled by a valid assignment of a NOESY cross-peak to two protons A and B in the CANDID-automated NOESY assignment algorithm *(5)*: **(A)** agreement between chemical shifts and the peak position, **(B)** network anchoring, and **(C)** spatial proximity in a (preliminary) structure.

6. *Constraint combination.* In cycles 1 and 2 groups of 2 or 4, *a priori* unrelated long-range distance constraints are combined into new virtual distance constraints that each carry the assignments from two of the original constraints (*see* **Subheading 3.2.2.**).
7. *Structure calculation.* Using torsion angle dynamics (*see* **Subheading 3.1.4.**) a 3D structure of the protein is calculated that is added to the input for the following cycle. Distance constraints from NOEs with multiple assignments and those resulting from constraint combination are introduced as ambiguous distance constraints into the structure calculation. Return to **step 1**.

Between subsequent cycles, information is transferred exclusively through the intermediary 3D structures, in that the molecular structure obtained in a given cycle is used to guide the NOE assignment in the following cycle (**Fig. 1**). Otherwise, the same input data is used for all cycles. 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. Otherwise an ambiguous distance constraint is generated that embodies several assignments (*see* **Subheading 3.1.2.**). Cross-peaks with a poor score are temporarily discarded. 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* "constraint combination" in **Subheading 3.1.4.**). The distance constraints are then included in the input for the structure calculation with the CYANA torsion angle dynamics algorithm. An automated structure calculation typically

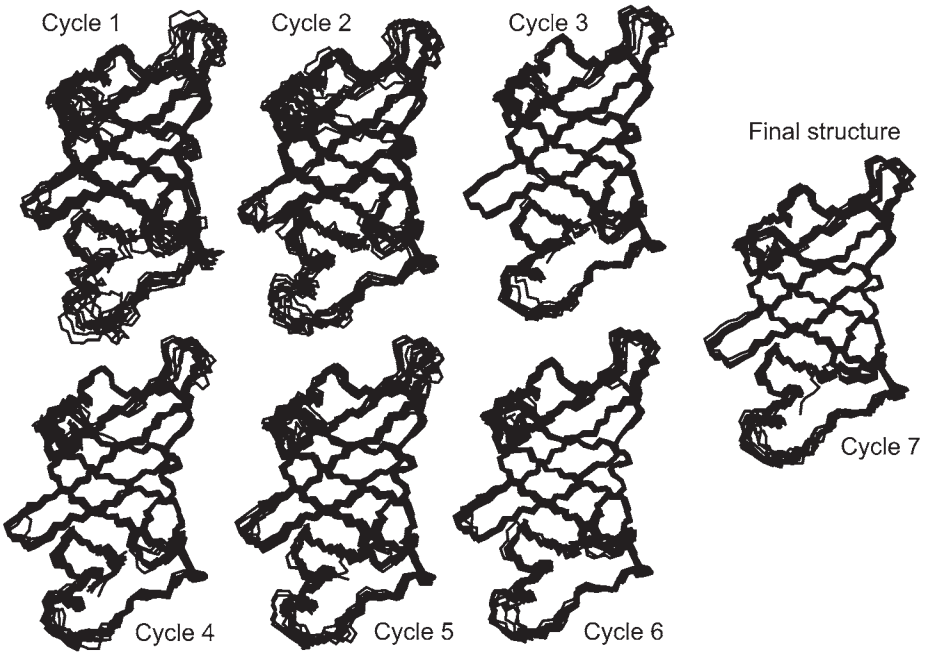


Fig. 3. Structures of the heme chaperone CcmE *(37)* obtained with the program CYANA in seven consecutive cycles of combined automated NOESY assignment with CANDID *(5)* and structure calculation with torsion angle dynamics. The backbones of the 10 conformers with lowest target function value in each cycle were drawn with the program MOLMOL *(7)*.

involves seven cycles (**Fig. 3**). In the first cycle, the structure-independent NOE self-consistency check (*see* "network anchoring" in **Subheading 3.1.3.**) has a dominant impact because structure-based criteria cannot be applied yet. The second and subsequent cycles differ from the first cycle by the use of additional selection criteria for cross-peaks and NOE assignments that exploit the protein 3D structure from the preceding cycle. Because the precision of the structure determination normally improves with each subsequent cycle (**Fig. 3**), the criteria for accepting assignments and distance constraints are tightened in more advanced cycles of the structure 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 structure in the form of a bundle of conformers. In the final 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 distance constraints in subsequent refinement and analysis programs that do not handle ambiguous distance constraints.

3.1.2. Ambiguous Distance Constraints

Ambiguous distance constraints ***(21,22)*** provide a very important concept 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 the structure calculation. Because the majority of NOEs cannot be assigned unambiguously from chemical shift information alone, this lack of a general way to incorporate ambiguous data into the structure calculation considerably hampered the performance of early automatic NOESY assignment algorithms ***(15)***.

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 interatomic distance. A NOESY cross-peak with a unique assignment possibility gives rise to an upper-bound b on the distance $d(\alpha,\beta)$ 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, $\bar{d} \le b$, with

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

Each of the distances $d_k = d(\alpha_k, \beta_k)$ in the sum 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.

3.1.3. Network Anchoring

Network anchoring ***(5)*** 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 it compensates for the absence of 3D structural information at the outset of a *de novo* structure determination (**Fig. 2**). 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 artifact constraints would not lead to

systematic constraint violations during the structure calculation and, therefore, can 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 ***(5)*** 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 involves 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.

3.1.4. Constraint Combination

Spurious distance constraints may arise in NMR protein structure determinations from misinterpretation of noise and spectral artifacts. This situation is particularly critical at the outset of a structure determination, before a preliminary structure is available for 3D structure-based filtering of constraint assignments. Constraint combination ***(5)*** 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 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, namely 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, implies that such combined constraints have a lower probability of being erroneous than the corresponding original constraints (provided that less than half of the original constraints are erroneous; *see* **Note 5**). Two modes of constraint combination are provided in CYANA ***(5)***: "$2 \rightarrow 1$" 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

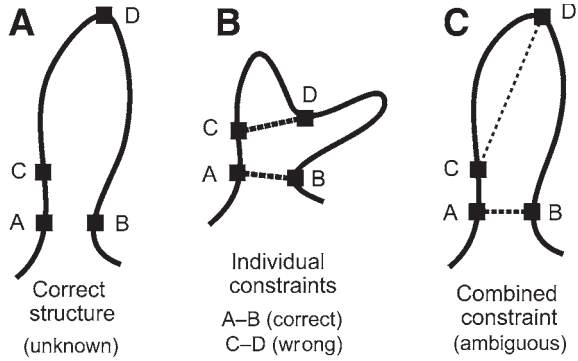


Fig. 4. Schematic illustration of the effect of constraint combination ***(6)*** in the case of two distance constraints, a correct one connecting atoms A and B, and an incorrect one between atoms C and D. A structure calculation that uses these two constraints as individual constraints that have to 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 set $A \cup B$, the union of sets A and B (**Fig. 4**). The $4 \rightarrow 4$ 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, that is, the peaks with all assignments to pairs of atoms separated by five or more 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 the $2 \rightarrow 1$ combination but stays constant on $4 \rightarrow 4$ pairwise combination. Therefore, the latter approach preserves more of the original structural information. Furthermore, it can take into account that certain peaks and their assignments are more reliable than others because the peaks with assignment sets A, B, C, D are used 3, 2, 2, and 1 times, respectively, to form combined constraints. To this end, the long-range peaks are sorted according to their total residue-wise network-anchoring score ***(5)***, and the $4 \rightarrow 4$ combination is performed by selecting the assignments A, B, C, D from the first, second, third, and fourth quarter of the sorted list, respectively. 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, $b = \left(b_1^{-6} + b_2^{-6}\right)^{-1/6}$, or as the maximum, $b = \max\left(b_1, b_2\right)$. 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.

3.2. Structure Calculation

The calculation of the 3D structure forms a cornerstone of the NMR method for protein structure determination. Because of the complexity of the problem—a protein typically consists of more than 1000 atoms, which are restrained by a similar number of experimentally determined constraints in conjunction with stereochemical and steric conditions—it is in neither feasible to do an exhaustive search of allowed conformations nor to find solutions by interactive model building. Therefore, the calculation of the 3D structure is formulated in CYANA as a minimization problem for a target function that measures the agreement between a structure and the given set of constraints.

3.2.1. Target Function

The CYANA target function ***(23,24)*** is defined such that it is zero if and only if all experimental distance constraints and torsion angle constraints are fulfilled and all nonbonded atom pairs satisfy a check for the absence of steric overlap. A conformation that satisfies the constraints more closely than another one will lead to a lower target function value. The exact definition of the CYANA target function is:

$$V = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta)\in I_c} \left(d_{\alpha\beta} - b_{\alpha\beta}\right)^2 + w_a \sum_{i\in I_a} \left[1 - \frac{1}{2}\left(\frac{\Delta_i}{\Gamma_i}\right)^2\right]\Delta_i^2$$

Upper and lower bounds, $b_{\alpha\beta}$, on distances $d_{\alpha\beta}$ between two atoms α and β, and constraints on individual torsion angles θ_i in the form of allowed intervals $[\theta_i^{\min}, \theta_i^{\max}]$ are considered. I_u, I_l, and I_v are the sets of atom pairs (α,β) with upper, lower, or van der Waals distance bounds, respectively, and I_a is the set of restrained torsion angles. w_u, w_l, w_v, and w_a are weighting factors for the different types of constraints. $\Gamma_i = \pi - \left(\theta_i^{\max} - \theta_i^{\min}\right)/2$ denotes the half width of the forbidden range of torsion angle values, and Δ_i is the size of the torsion angle constraint violation.

3.2.2. Torsion Angle Dynamics

The minimization algorithm in CYANA is based on the idea of simulated annealing ***(25)*** by molecular dynamics simulation (*see* **Note 6**) in torsion angle space. The distinctive feature of molecular dynamics simulation when compared to the straightforward minimization of a target function is the presence of kinetic energy that allows overcoming barriers of the potential surface, which reduces greatly the problem of becoming trapped in local minima. Torsion

angle dynamics, that is, molecular dynamics simulation using torsion angles instead of Cartesian coordinates as degrees of freedom ***(24,26–34)***, provides at present the most efficient way to calculate NMR structures of biological macromolecules. The only degrees of freedom are the torsion angles, that is, rotations about single bonds, such that the conformation of the molecule is uniquely specified by the values of all torsion angles. Covalent bonds that are incompatible with a tree structure because they would introduce closed flexible rings, for example, disulphide bridges, are treated, as in Cartesian space dynamics, by distance constraints. The efficiency of the torsion angle dynamics algorithm ***(30)*** implemented in the program CYANA, and previously in DYANA ***(24)***, is high because of the fact that it requires a computational effort that increases only linearly with the system size (*see* **Note 7**). In contrast, the computation time for "naïve" approaches to torsion angle dynamics rises with the third power of the system size (e.g., **ref. *29***), which renders these algorithms unsuitable for use with macromolecules. With the fast torsion angle dynamics algorithm in CYANA, the advantages of torsion angle dynamics, especially the much longer integration time steps that can be used, are effective for molecules of all sizes.

3.2.3. Simulated Annealing

The potential energy landscape of a protein is complex and studded with many local minima, even in the presence of experimental constraints and when using the simplified target function of **Subheading 3.2.1.** Because the temperature, that is, kinetic energy, determines the maximal height of energy barriers that can be overcome in a molecular dynamics trajectory, the temperature schedule is important for the success and efficiency of a simulated annealing calculation. Elaborated protocols have been devised for structure calculations using molecular dynamics in Cartesian space ***(35,36)***. In addition to the temperature, other parameters, such as force constants and repulsive core radii, are varied in these schedules, which may involve several stages of heating and cooling. The fast exploration of conformation space with torsion angle dynamics allows for much simpler schedules. The standard simulated annealing protocol in the program CYANA ***(24)*** starts from a conformation with all torsion angles treated as independent, uniformly distributed random variables and consists of five stages:

1. *Initial minimization.* A short minimization to reduce high-energy interactions that could otherwise disturb the torsion angle dynamics algorithm: 100 conjugate gradient minimization steps are performed, including only distance constraints between atoms up to 3 residues apart along the sequence, followed by a further 100 minimization steps including all constraints. For efficiency, until **step 4**, all hydrogen atoms are excluded from the check for steric overlap, and the repulsive core radii of heavy atoms with covalently bound hydrogens are increased by

0.15Å with respect to their standard values. The weights in the target function of **Subheading 3.2.1.** are set to 1 for user-defined upper and lower distance bounds, to 0.5 for steric lower distance bounds, and to 5Å^2 for torsion angle constraints.

2. *High-temperature phase.* A torsion angle dynamics calculation at constant high temperature: One-fifth of all N torsion angle dynamics steps are performed at a constant high reference temperature of, typically, 10,000 K. The time step is initialized to 2 fs (= 2×10^{-15} s). The list of van der Waals lower distance bounds is updated every 50 steps using a cutoff of 4.2Å for the interatomic distance throughout all torsion angle dynamics phases.
3. *Slow cooling.* Torsion angle dynamics calculation with slow cooling close to zero temperature: The remaining 4N/5 torsion angle dynamics steps are performed during which the reference value for the temperature approaches zero according to a fourth-power law.
4. *Low-temperature phase with individual hydrogen atoms.* Incorporation of all hydrogen atoms into the check for steric overlap: After resetting the repulsive core radii to their standard values, and increasing the weighting factor for steric constraints to two, 100 conjugate gradient minimization steps are performed, followed by 200 torsion angle dynamics steps at zero reference temperature.
5. *Final minimization.* A final minimization consisting of 1000 conjugate gradient steps.

With the CYANA torsion angle dynamics algorithm it is possible to efficiently calculate protein structures on the basis of NMR data. Even for a system as complex as a protein, the program CYANA can execute thousands of torsion angle dynamics steps within minutes of computation time. For instance, the computation time for the calculation of one conformer of the 136-residue heme chaperone protein CcmE on the basis of 2453 NOE upper distance bounds and 56 torsion angle constraints ***(37)*** using 10,000 torsion angle dynamics steps on a single processor is below 1 min on up-to-date hardware:

Linux PC, Pentium IV, 3.06 GHz:	29 s
Linux PC, Pentium IV, 1.8 GHz:	42 s
Compaq Alpha server GS 320:	23 s
Silicon Graphics, R16000, 700 MHz:	39 s
Silicon Graphics, R12000, 400 MHz:	59 s

Furthermore, because an NMR structure calculation always involves the computation of a group of conformers, it is highly efficient and straightforward with CYANA to run calculations of multiple conformers in parallel. Nearly ideal speedup, that is, an overall computation time almost inversely proportional to the number of processors, can be achieved with CYANA ***(24)***.

3.3. Effect of Incomplete Chemical Shift Assignments

A limiting factor for the application of the automated NOE assignment algorithm CANDID is that it relies 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 semiautomated and interactive methods, in spite of promising attempts toward automation ***(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, or if chemical shift assignments could be completed during the structure calculation (*see* **Note 8**), provided that the reliability and robustness of the NMR method for protein structure determination is not compromised.

It has been shown ***(38)*** that for reliable automated NOESY, assignment with the CANDID algorithm around 90% completeness of the chemical shift assignment is necessary (*see* **Note 9**). In certain cases, the lack of a small number of "essential" chemical shifts can lead to a significant deviation of the structure. On the other hand, in practice the algorithm might be expected to tolerate a slightly higher degree of incompleteness in the chemical shift assignments provided that most missing assignments are of "unimportant" atoms that are involved in only 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 few 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. Network anchoring and constraint combination are two methods that have been designed and shown to be effective in minimizing the impact of incomplete and/or erroneous pieces of input data (*see* **Subheadings 3.1.3.** and **3.1.4.**). Chemical shift assignment-based automated NOE assignment without the safeguards of network anchoring and constraint combination is expected to be more susceptible to deleterious effects from missing chemical shift assignments and artifacts in the input data. In contrast to missing or incorrect entries in the chemical shift list, the algorithm is remarkably tolerant regarding incompleteness of the NOESY peak list (*see* **Note 10**). This suggests that it is better to strive for correctness than for ultimate completeness of the input NOESY peak lists.

3.4. Quality Control

In this section, simple criteria based on the output of CYANA are given that allow assessing the reliability of the resulting structure without cumbersome recourse to independent interactive verification of the NOESY assignments.

Final structures from an automatic algorithm that have a low root-mean-square deviation (RMSD) within the bundle of conformers but differ significantly from the "correct" reference structure are problematic because, without knowledge of an independently determined "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 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. Test calculations 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 ***(38)***. 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 ***(5)***. These comprised six criteria that should be met simultaneously:

1. Average CYANA target function value of cycle 1 below 250 $Å^2$.
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.
6. RMSD between the mean structures of the first and last cycle below 3 Å.

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. Criteria 3 and 4 impose a limit on 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 CYANA runs was confirmed by the fact that all the structure calculations in a systematic study ***(38)*** with an RMSD bias ***(39)*** to the reference structure higher than 2 Å violated one or several of the six criteria. On the other hand, the same test calculations 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 just two conditions for successful structure calculation with CYANA:

1. Less than 25% of the long-range NOEs must have been discarded by the automated NOESY assignment algorithm for the final structure calculation (*see also* **Note 11**).
2. The backbone RMSD to the mean coordinates for the structure bundle of the *first* cycle must not exceed 3 Å.

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 an important factor that strongly influences the accuracy of the final structure. This can be understood by considering the iterative nature of automated NOESY assignment, 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 toward a precise but inaccurate final structure. In practice, it is safe to apply both criteria, even though in test calculations ***(38)*** the percentage of discarded long-range NOEs alone would have been sufficient to detect all runs that resulted in a structure with more than 2 Å RMSD bias. In these test calculations a large dispersion in the accuracy of the final structure was reflected reliably by the percentage of discarded long-range NOEs and the RMSD in cycle 1, but it could not readily be discerned from the values of the target function after cycle 1 or 7, the RMSD at cycle 7, or, in a few cases, the percentage of unassigned NOEs.

3.5. Troubleshooting

If the output of a CYANA structure calculation based on automated NOESY assignment with CANDID does not fulfill the guidelines of **Subheading 3.4.**, then the structure will in many cases still be essentially correct but should not be accepted without further validation. Within the framework of CYANA, the recommended approach is to improve the quality of the input chemical shift and peak lists and to perform a new complete CYANA run with seven cycles, until the criteria are met. Usually, this can be achieved efficiently because the output from an unsuccessful CYANA run, even though the structure should not be trusted *per se,* clearly points out problems in the input—for example, peaks that cannot be assigned and might therefore be artifacts or indications of erroneous or missing sequence-specific assignments. To facilitate this task, the program gives for each peak informational output that includes 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 **Subheading 3.4.** are already met, 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 seven 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 to prepare an improved, more complete NOESY peak list for a second CYANA calculation. This can be done more efficiently than it 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.

4. Notes

1. *Definitions.* For consistency and simplicity, the following conventions are used: An interaction between two or more atoms 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 an atom. *Peak assignment* is the process and the result of identifying in each spectral dimension the atom(s) that are involved in the signal represented by the peak. *NOESY assignment* is peak assignment in NOESY spectra.
2. *Automated chemical shift assignment algorithms.* There have been many attempts to automate the chemical shift assignment that has to precede the collection of conformational constraints and the structure calculation. These methods have been reviewed recently ***(1)***. Some automated approaches to chemical shift assignment 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, whereas others 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.
3. *Ambiguity of chemical shift based NOE assignment.* A simple mathematical model of the NOESY assignment process by chemical shift matching gives insight into this problem ***(16)***. 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 $\Delta\omega$, that is, 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 spectral width $\Delta\Omega$, 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 1 out of all n proton shifts inside the tolerance range $\Delta\omega$?from the peak position. Their expected number, $N^{(1)} = N(1-p)^{2n-2} \approx Ne^{-2np} = Ne^{-4n\Delta\omega/\Delta\Omega}$,

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 ***(40)***, this model predicts that only about 4% of the NOEs can be assigned unambiguously based solely on chemical shift information with an accuracy of $\Delta\omega$ = 0.02 ppm—an insufficient number to calculate a preliminary 3D 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 heterospin, so that $N^{(1)} \approx Ne^{-np} = Ne^{-2n\Delta\omega/\Delta\Omega}$. Regarding assignment ambiguity, 3D 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. Once available, a preliminary 3D 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 interatomic 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: $q = \left(d_{\max}/R\right)^3$. Using $d_{\max}$ = 5 Å, one obtains for WmKT, a nearly spherical protein with a radius of about 15 Å, $q \approx 4\%$. Hence, not more than 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 because q will always be larger than 0.

4. *Structure determinations with automated NOE assignment by CANDID.* The automated structure calculation method described in this chapter has been evaluated in test calculations ***(5,13,38)*** and used for various *de novo* structure determinations, including four variants of the human prion protein ***(41,42)***, the calreticulin P-domain ***(43)***, two distinct forms of the pheromone-binding protein from *Bombyx mori* ***(44,45)***, the class I human ubiquitin-conjugating enzyme 2b ***(46)***, the heme chaperone CcmE ***(37)*** (**Fig. 3**), and the nucleotide-binding domain of Na,K-ATPase ***(47)***. The NOESY assignments and the corresponding distance constraints for these *de novo* structure determinations were made automatically by the program, confining interactive work to the stage of the preparation of the input chemical shift and peak lists. These structure determinations have confirmed the viability of CYANA for automated NOESY assignment and structure calculation without prior knowledge about NOESY assignments or the 3D structure.
5. *Effect of constraint combination.* The effect of constraint combination on the expected number of erroneous distance constraints can be estimated quantitatively in the case of 2 → 1 combination by assuming an original data set containing N long-range peaks and a uniform probability $p << 1$ that a long-range peak would

lead to an erroneous constraint ***(5)***. 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 may be assumed that the same N long-range peaks can be classified into four equally large classes with probabilities to be erroneous of αp, p, p, $(2 - \alpha)p$, respectively. The overall probability for an input constraint to be erroneous is again p. The parameter α, $0 \le \alpha \le 1$, 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 $\left(\alpha + \left(1 - \alpha^2\right)/4\right)p^2$, 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 (10%) 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 approx 993 correct and 7 (0.7%) erroneous combined constraints. Alternatively, $2 \rightarrow 1$ combination will yield under these conditions approx 495 correct and 5 (1%) 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.

6. *Molecular dynamics simulation vs NMR structure calculation.* There is a fundamental difference between molecular simulation that has the aim of simulating the trajectory of a molecular system as realistically as possible to extract molecular quantities of interest and NMR structure calculation that is driven by experimental constraints. Classical molecular dynamics simulations ***(48)*** rely on a full empirical force field to ensure proper stereochemistry and are generally run at a constant temperature, close to room temperature. Substantial amounts of computation time are required because the empirical energy function includes long-range pair interactions that are time-consuming to evaluate and because conformation space is explored slowly at room temperature. When molecular dynamics algorithms are used for NMR structure calculations, however, the objective is quite different. Here, such algorithms simply provide a means to efficiently optimize a target function that takes the role of the potential energy. Details of the calculation, such as the course of a trajectory, are unimportant, as long as its end point comes close to the global minimum of the target function. Therefore, the efficiency of NMR structure calculation can be enhanced by simplification or modification of the force field and/or the algorithm that does not significantly alter the location of the global minimum (the correctly folded structure) but shortens (in terms of computation time needed) the way by which it can be reached from the start conformation. A typical "geometric" force field used in NMR structure calculation therefore retains only the most important part of the nonbonded interaction by a simple repulsive potential that replaces the Lennard–Jones and electrostatic interactions of the full empirical energy function. This short-range repulsive function can be calculated much faster and significantly facilitates large-scale conformational changes that are required during the folding process by lowering energy barriers induced by the overlap of atoms.

7. *Fast algorithm for torsion angle dynamics.* The key idea of the fast torsion angle dynamics algorithm in CYANA is to exploit the fact that a chain molecule such as a protein or nucleic acid can be represented in a natural way as a tree structure consisting of $n + 1$ rigid bodies that are connected by n rotatable bonds (**Fig. 5A**) ***(28,49)***. Each rigid body is made up of one or several mass points (atoms) with fixed relative positions. The tree structure starts from a base, typically at the N-terminus of the polypeptide chain, and terminates with "leaves" at the ends of the sidechains and at the C-terminus. The angular velocity vector ω_k and the linear velocity v_k of the reference point of the rigid body k (**Fig. 5B**) are calculated recursively from the corresponding quantities of the preceding rigid body

$$\omega_k = \omega_{p(k)} + e_k\dot{\theta}_k,$$
$$\boldsymbol{v}_k = \boldsymbol{v}_{\mathrm{p(k)}} - \left(\mathbf{r}_k - \mathbf{r}_{p(k)}\right) \wedge \omega_{p(k)}.$$

Denoting the vector from the reference point to the center of mass of the rigid body k by $\boldsymbol{Y}_k$, its mass by m_k, and its inertia tensor by I_k (**Fig. 5B**), the kinetic energy can be computed in a linear loop over all rigid bodies:

$$E_{\mathrm{kin}} = \frac{1}{2}\sum_{k=0}^{n}\left[m_k\boldsymbol{v}_k^2 + \omega_k \cdot I_k\omega_k + 2\boldsymbol{v}_k \cdot \left(\omega_k \wedge m_k\boldsymbol{Y}_k\right)\right].$$

The calculation of the torsional accelerations, that is, the second-time derivatives of the torsion angles, is the crucial point of a torsion angle dynamics algorithm. The equations of motion for a classical mechanical system with generalized coordinates are the Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{\theta}_k}\right) - \frac{\partial L}{\partial\theta_k} = 0 \qquad (k = 1,\ldots,n)$$

with the Lagrange function $L = E_{\mathrm{kin}} - E_{\mathrm{pot}}$. They lead to equations of motion of the form $M(\theta)\ddot{\theta} + C(\theta,\dot{\theta}) = 0$. In the case of torsion angles as degrees of freedom, the mass matrix $M(\theta)$ and the n-dimensional vector $C(\theta,\dot{\theta})$ can be calculated explicitly ***(28,29)***. To generate a trajectory, this linear set of n equations would have to be solved in each time step for the torsional accelerations $\ddot{\theta}$, which requires a computational effort proportional to n^3, which is prohibitively expensive for larger systems. Therefore, in CYANA the fast recursive algorithm of Jain et al. ***(30)*** is implemented to compute the torsional accelerations, which makes explicit use of the tree structure of the molecule to obtain $\ddot{\theta}$ with a computational effort that is only proportional to n. The mathematical details of the CYANA torsion angle dynamics algorithm are given in ***(24,30)***. It suffices to note here that the torsional accelerations can be obtained by executing a series of three linear loops over all rigid bodies similar to the single one that is needed to compute the kinetic energy, E_{kin}. The integration scheme for the equations of motion in torsion angle dynamics is a variant of the "leap-frog" algorithm ***(48,50)*** used in Cartesian space molecular dynamics. To obtain a trajectory, the equations of motion are numerically integrated

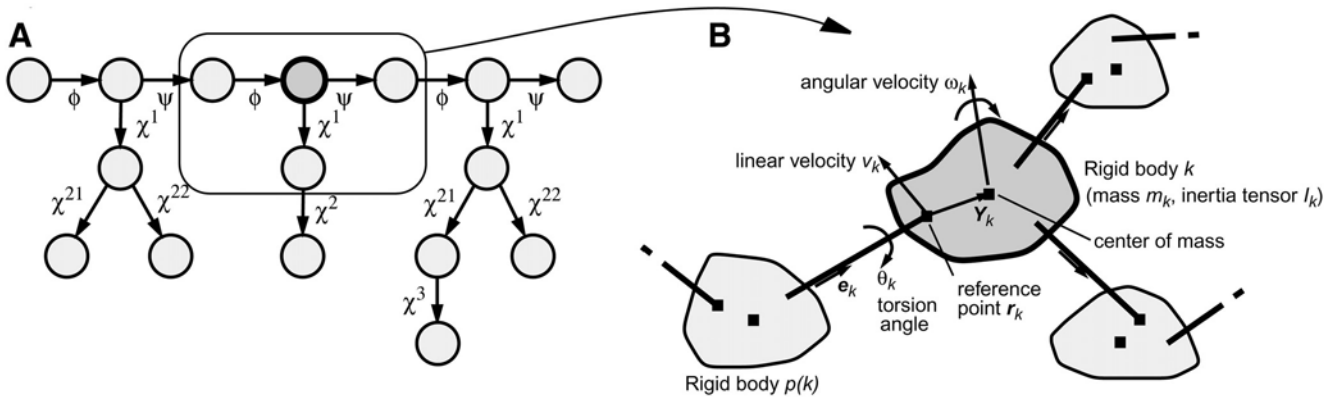


Fig. 5. **(A)** Tree structure of torsion angles for the tripeptide Val-Ser-Ile. *Circles* represent rigid units. Rotatable bonds are indicated by *arrows* that point toward the part of the structure that is rotated if the corresponding torsion angle is changed. **(B)** Excerpt from the tree structure formed by the torsion angles of a molecule, and definition of quantities required by the CYANA torsion angle dynamics algorithm.

by advancing the $i = 1,\ldots,n$ (generalized) coordinates q_i and velocities $\dot{q}_i$ that describe the system by a small but finite time step Δt:

$$\dot{q}\left(t + \Delta t / 2\right) = \dot{q}_i\left(t - \Delta t / 2\right) + \Delta t\ddot{q}_i(t) + O(\Delta t^3)$$
$$q_i\left(t + \Delta t\right) = q_i(t) + \Delta t\dot{q}_i\left(t + \Delta t / 2\right) + O\left(\Delta t^3\right)$$

The degrees of freedom, q_i, are the Cartesian coordinates of the atoms in conventional molecular dynamics simulation, or the torsion angles in CYANA. The $O(\Delta t^3)$ terms indicate that the errors with respect to the exact solution incurred by the use of a finite time step Δt are proportional to Δt^3. The time step Δt must be small enough to sample adequately the fastest motions. Because the fastest motions in conventional molecular dynamics simulation are oscillations of bond lengths and bond angles, which are "frozen" in torsion angle space, longer time steps can be used for torsion angle dynamics than for molecular dynamics in Cartesian space ***(24)***. The temperature is controlled by weak coupling to an external bath ***(51)***, and the length of the time step is adapted automatically based on the accuracy of energy conservation ***(24)***. It could be shown that in practical applications with proteins time steps of about 100, 30, and 7 fs at low (1 K), medium (400 K), and high (10,000 K) temperatures, respectively, can be used in torsion angle dynamics calculations with CYANA ***(24)***, whereas time steps in Cartesian space molecular dynamics simulation generally have to be in the range of 2 fs. The concomitant fast exploration of conformation space provides the basis for the efficient CYANA structure calculation protocol.

8. *Chemical shift assignment during NOE assignment and structure calculation.* 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 ***(52)***: Starting from nearly complete chemical shift assignments for the backbone and for 348 sidechain protons of the 28 kDa single-chain T-cell receptor protein, the chemical shifts of 40 additional sidechain protons could be found by a combination of chemical shift prediction with the program SHIFTS ***(53,54)*** and NOE assignment with ARIA ***(17)***. The same approach can be used with CYANA.
9. *Impact of incomplete chemical shift assignments.* The influence of incomplete chemical shift assignments on the reliability of NMR structures has been investigated using the program CYANA with input data that represents various degrees of completeness of the chemical shift assignment ***(38)***. The effect of missing chemical shift assignments was assessed by randomly omitting entries from the "complete" 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) ***(44)*** and the killer toxin WmKT ***(40)***. 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 ***(39)***. In the representative case of randomly selecting the omitted chemical shifts among all ^{1}H chemical shift assignments of BmPBP^{A}, the RMSD bias increased only slowly with increasing omission ratio p up to about p = 10%, from where onward the RMSD bias rose abruptly, reflecting that severely distorted structures had been obtained. Higher omission ratios not only resulted 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 lack of chemical shifts when run with data from the uniformly ^{13}C- and ^{15}N-labeled protein BmPBP^{A} than with the homonuclear data for the protein WmKT, even though BmPBP^{A} (142 residues) is much larger than WmKT (88 residues). This is because of the availability of ^{13}C and ^{15}N chemical shifts that allow resolution of many ^{1}H chemical shift degeneracies 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 ***(38)***. 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.

10. *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 ***(38)***. 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 lose 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.
11. *Alternative criterion to assess the completeness of the NOESY assignment.* 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 conforma-

tional 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. However, among these two criteria, the percentage of discarded long-range NOEs is a slightly more sensitive indicator of the accuracy of the final structure than the overall percentage of unused cross-peaks because the latter includes also peaks with short-range assignment or with no assignment possibility at all that are expected to have little distorting effect on the resulting structure.

References

1 Moseley, H. N. B. and Montelione, G. T. (1999) Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.* **9,** 635–642.

2 Solomon, I. (1955) Relaxation processes in a system of two spins. *Phys. Rev.* **99,** 559–565.

3. Macura, S. and Ernst, R. R. (1980) Elucidation of cross relaxation in liquids by 2D NMR spectroscopy. *Mol. Phys.* **41,** 95–117.

4. Neuhaus, D. and Williamson, M. P. (1989) *The Nuclear Overhauser Effect in Structural and Conformational Analysis.* VCH, Weinheim, Germany.

5 Herrmann, T., Güntert, P., and Wüthrich, K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319,** 209–227.

6 Gropp, W., Lusk, E., Doss, N., and Skjellum, A. (1996) A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Computing* **22,** 789–828.

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

8. Bartels, C., Xia, T. H., Billeter, M., Güntert, P., and Wüthrich, K. (1995) The program XEASY for computer-supported NMR-spectral analysis of biological macromolecules. *J. Biomol. NMR* **6,** 1–10.

9. Johnson, B. A. and Blevins, R. A. (1994) NMR View—a computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* **4,** 603–614.

10. Kraulis, P. J. (1989) ANSIG—a program for the assignment of protein H-1 2D NMR spectra by interactive computer graphics. *J. Magn. Reson.* **24,** 627–633.

11 Helgstrand, M., Kraulis, P., Allard, P., and Härd, T. (2000) ANSIG for Windows: an interactive computer program for semiautomatic assignment of protein NMR spectra *J. Biomol. NMR* **18,** 329–336.

12 Koradi, R., Billeter, M., Engeli, M., Güntert, P., and Wüthrich, K. (1998) Toward fully automatic peak picking and integration of biomolecular NMR spectra. *J. Magn. Reson.* **135,** 288–297.

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

14 Doreleijers, J. F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Markley, J. L., et al. (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR* **26,** 139–146.
15 Mumenthaler, C. and 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.
16 Mumenthaler, C., Güntert, P., Braun, W., and Wüthrich, K. (1997) Automated procedure for combined assignment of NOESY spectra and three-dimensional protein structure determination. *J. Biomol. NMR* **10,** 351–362.
17 Nilges, M., Macias, M., O'Donoghue, S. I., and Oschkinat, H. (1997) Automated NOESY interpretation with ambiguous distance constraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. *J. Mol. Biol.* **269,** 408–4228
18 Nilges, M. and O'Donoghue, S. I. (1998) Ambiguous NOEs and automated NOE assignment. *Prog. NMR Spectrosc.* **32,** 107–139.
19 Linge, J. P., O'Donoghue, S. I., and Nilges, M. (2001) Automated assignment of ambiguous nuclear Overhauser effects with ARIA. *Methods Enzymol.* **339,** 71–90.
20 Linge, J. P., Habeck, M., Rieping, W., and Nilges, M. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19,** 315–316.
21 Nilges, M. (1993) A calculation strategy for the structure determination of symmetric dimers by ^{1}H NMR. *Proteins* **17,** 297–309.
22 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.
23. Güntert, P., Braun, W., and 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.
24. Güntert, P., Mumenthaler, C., and Wüthrich, K. (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273,** 283–298.
25. Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P. (1983) Optimization by simulated annealing. *Science* **220,** 671–680.
26 Katz, H., Walter, R., and Somorjay, R. L. (1979) Rotational dynamics of large molecules. *Computers Chemistry* **3,** 25–32.
27. Bae, D. S. and Haug, E. J. (1987) A recursive formulation for constrained mechanical system dynamics, part I: open loop systems. *Mech. Struct. Mech.* **15,** 359–382.
28. Mazur, A. K. and Abagyan, R. A. (1989) New methodology for computer-aided modelling of biomolecular structure and dynamics (I): non-cyclic structures. *J. Biomol. Struct. Dyn.* **4,** 815–832.
29. Mazur, A. K., Dorofeev, V. E., and Abagyan, R. A. (1991) Derivation and testing of explicit equations of motion for polymers described by internal coordinates. *J. Comp. Phys.* **92,** 261–272.

30 Jain, A., Vaidehi, N., and Rodriguez, G. (1993) A fast recursive algorithm for molecular dynamics simulation. *J. Comp. Phys.* **106,** 258–268.
31 Kneller, G. R. and Hinsen, K. (1994) Generalized Euler equations for linked rigid bodies. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **50,** 1559–1564.
32 Mathiowetz, A. M., Jain, A., Karasawa, N., and Goddard, W. A., III. (1994) Protein simulations using techniques suitable for large systems: the cell multipole method for nonbond interactions and the Newton-Euler inverse mass operator method for internal coordinate dynamics. *Proteins* **20,** 227–247.
33 Rice, L. M. and Brünger, A. T. (1994) Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins* **19,** 277–290.
34 Stein, E. G., Rice, L. M., and Brünger, A. T. (1997) Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *J. Magn. Reson.* **124,** 154–164.
35 Nilges, M., Clore, G. M., and Gronenborn, A. M. (1988) Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Lett.* **229,** 317–324.
36. Brünger, A. T. (1992) *X-PLOR version 3.1: a system for X-ray crystallography and NMR.* Yale University Press, New Haven, CT.
37 Enggist, E., Thöny-Meyer, L., Güntert, P., and Pervushin, K. (2002) NMR structure of the heme chaperone CcmE reveals a novel functional motif. *Structure* **10,** 1551–1557.
38 Jee, J. G. and Güntert, P. (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *J. Struct. Funct. Genomics* 4, 179–189.
39. Güntert, P. (1998) Structure calculation of biological macromolecules from NMR data. *Q. Rev. Biophys.* **31,** 145–237.
40 Antuch, W., Güntert, P., and Wüthrich, K. (1996) Ancestral βγ-crystallin precursor structure in a yeast killer toxin, *Nat. Struct. Biol.* **3,** 662–665.
41 Calzolai, L., Lysek, D. A., Güntert, P., von Schroetter, C., Riek, R., Zahn, R., et al. (2000) NMR structures of three single-residue variants of the human prion protein. *Proc. Natl. Acad. Sci. USA* **97,** 8340–8345.
42 Zahn, R., Güntert, P., von Schroetter, C., and Wüthrich, K. (2003) NMR structure of a human prion protein with two disulfide bridges. *J. Mol. Biol.* **326,** 225–234.
43 Ellgaard, L., Riek, R., Herrmann, T., Güntert, P., Braun, D., Helenius, A., et al. (2001) NMR structure of the calreticulin P-domain. *Proc. Natl. Acad. Sci. USA* **98,** 3133– 3138.
44 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,** 14,374–14,379.
45 Lee, D., Damberger, F. D., Peng, G., Horst, R., Güntert, P., Nikonova, L., et al. (2002) NMR structure of the unliganded *Bombyx mori* pheromone-binding protein at physiological pH. *FEBS Lett.* **531,** 314–318.

46 Miura, T., Klaus, W., Ross, A., Güntert, P., and Senn, H. (2002) The NMR structure of the class I human ubiquitin-conjugating enzyme 2b. *J. Biomol. NMR* **22,** 89–92.
47 Hilge, M., Siegal, G., Vuister, G. W., Güntert, P., Gloor, S. M., and Abrahams, J. P. (2003) ATP-induced conformational changes of the nucleotide binding domain of Na,K-ATPase. *Nat. Struct. Biol.* **10,** 10–18.
48. Allen, M. P. and Tildesley, D. J. (1987) *Computer Simulation of Liquids.* Clarendon Press, Oxford, UK.
49 Abe, H., Braun, W., Noguti, T. and Go, N. (1984) Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles in proteins: general recurrent equations. *Computers Chemistry* **8,** 239–247.
50. Hockney, R. W. (1970) The potential calculation and some applications. *Meth. Comput. Phys.* **9,** 136–211.
51 Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684–3690.
52 Hare, B. J. and Wagner, G. (1999) Application of automated NOE assignment to three-dimensional structure refinement of a 28 kDa single-chain T cell receptor. *J. Biomol. NMR* **15,** 103–113.
53. Ösapay, K. and Case, D. A. (1991) A new analysis of proton chemical shifts in proteins. *J. Am. Chem. Soc.* **113,** 9436–9444.
54. Sitkoff, D. and Case, D. A. (1997) Density functional calculations of proton chemical shifts in model peptides. *J. Am. Chem. Soc.* **119,** 12,262–12,273.