5

Calculation of Structures from NMR Restraints

Peter Guntert

5.1 Introduction

When the NMR method for protein structure determination was introduced in the early 1980s the new approach met with enthusiasm amongst NMR spectroscopists, as well as scepticism and disbelief by structural biologists until the simultaneous but independent determinations of the three-dimensional structure of the protein tendamistat by X-ray crystallography [1] and NMR spectroscopy [2,3] yielded virtually identical results [4]. Since that time, NMR has become a firmly established method for determining the three-dimensional structures of proteins. More than 7800 structures in the Protein Data Bank [5] of March 2009 have been determined by NMR (Figure 5.1). This remarkable achievement would not have been possible without the development of sophisticated computational methods to compute three-dimensional protein structures from NMR-derived conformational restraints, and by increasingly automated approaches for analysing multidimensional NMR spectra.

NMR structure calculations can be performed in several ways that differ essentially by the extent to which the analysis of the spectra is automated. In a basic structure calculation, all spectra are analysed by the spectroscopist who also interprets the data and provides the structure calculation program with geometric restraints in the form of allowed interatomic distance ranges, ranges of allowed torsion angle values, and

Protein NMR Spectroscopy: Practical Techniques and Applications, First Edition.
Edited by Lu-Yun Lian and Gordon Roberts.

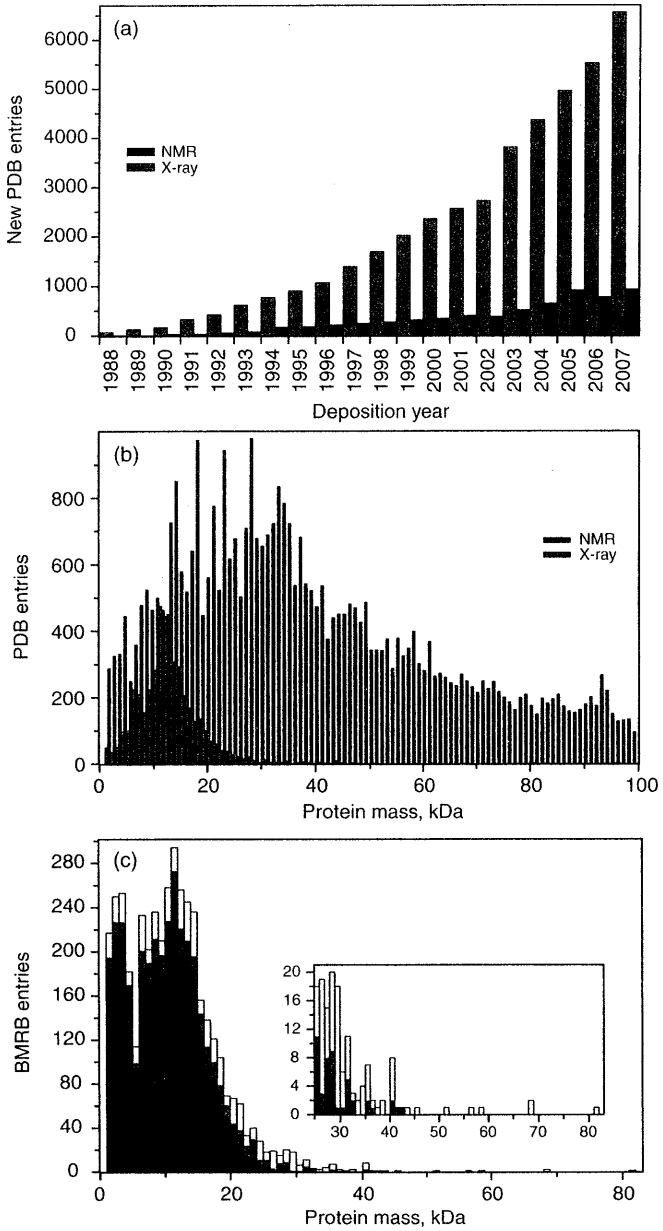

Figure 5.1 *(a) Annual depositions of X-ray and NMR structures 1988–2007. (b) Size distribution of X-ray and NMR structures in the Protein Data Bank of January 2009. (c) Completeness of chemical shift assignments in the Biological Magnetic Resonance Data Bank (BMRB) of January 2009. Completeness of the chemical shift assignments of backbone amide* ^{1}H *and aliphatic* ^{1}H *chemical shifts: 10–30 %, black bars; 30–70 %, dark grey bars; 70–90 %, medium grey bars; more than 90 %, light grey bars*

possibly additional types of restraints. In this case the software deals with the purely geometric problem of finding a three-dimensional arrangement of the atoms that is compatible with the primary structure of the protein, the conformational restraints from NMR, and steric repulsion. Instead of using conformational restraints, NMR structure calculation software can also read assigned NOESY peak lists and convert this information into upper bounds on distances between the corresponding pairs of hydrogen atoms using a given or automatically derived peak volume-to-distance relationship. A first significant degree of automation was reached by approaches that combined the automated assignment of NOESY peaks with the structure calculation. These algorithms start from the given chemical shift assignments and unassigned lists of NOESY peak positions and intensities. Only recently has it become possible to completely automate NMR spectra analysis by a fully automated algorithm that uses as input data a set of uninterpreted, multidimensional NMR spectra. Finally, several lines of unconventional approaches to NMR structure determination have been proposed that do not rely on sequence-specific chemical shift assignments and/or NOESY data.

This chapter gives an overview of the principles, basic algorithms and popular implementations of NMR structure calculation methods, including automated, assignment-free, and chemical shift-based approaches.

For consistency and simplicity, the following conventions will be 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.

5.2 Historical Development

With the first attempts to determine protein structures by NMR it became clear that new computer algorithms for structure calculation would be indispensable for solving three-dimensional protein structures, and that existing techniques from X-ray diffraction data would be as inadequate for the task as manual model building or interactive computer graphics.

The mathematical theory of distance geometry [6] was the first method to be used for protein structure calculation. The basic idea of distance geometry is to formulate the problem not in the Cartesian space of the atom positions but in the high-dimensional space of all interatomic distances where it is straightforward to find configurations that satisfy a network of distance measurements. The crucial step is then the embedding of a solution found in distance space into Cartesian space. For the first time a computer program was used to calculate the solution structure of a nonapeptide on the basis of experimental NOE measurements [7], and later on the NMR solution structure of a 35-residue globular protein [8]. An improved version of the original embedding

algorithm was implemented in DISGEO [9], the first complete program package for NMR protein structure calculation.

Finding molecular conformations that are in agreement with geometrical restraints can be formulated as the minimisation of a suitable 'target function'. The variable target function method in torsion angle space [10] used the method of conjugate gradients [11] for the minimisation of a multidimensional function. Recognising that fluctuations of the covalent bond lengths and bond angles around their equilibrium values are small, fast, and not measurable by NMR, only the torsion angles were retained as degrees of freedom. A fast recursive method made it possible to rapidly calculate the gradient of the target function against torsion angles [12]. However, as a local minimiser that takes exclusively downhill steps, conjugate gradient minimisation of a target function representing the complete network of NMR-derived restraints and the steric repulsion in a protein was virtually always trapped in local minima far from the correct solution. To alleviate this problem, the variable target function method implemented in the programs DISMAN [10] and DIANA [13], went through a series of minimisations of different target functions that gradually included restraints between atoms further and further separated along the polypeptide chain, thereby increasing step-by-step the complexity of the target function. This was a natural idea for helical proteins, but less successful for β-sheet topologies that are characterised by many nonlocal contacts. This convergence problem could later be cured in part by the usage of redundant torsion angle restraints [14]. In this iterative procedure redundant torsion angle restraints were generated on the basis of the torsion angle values found in a previous round of structure calculations.

In parallel with these developments, NMR structure calculation methods based on simulated annealing [15] driven by molecular dynamics simulation were developed. By numerically solving Newton's equations of motion of classical mechanics, trajectories for the atoms of a protein can be obtained. In the context of protein structure calculation the basic advantage of molecular dynamics simulation over minimisation techniques is the presence of kinetic energy that allows the system to escape from local minima. The efficiency of structure calculations using molecular dynamics simulation was enhanced by replacing the full 'physical' force field [16] by a simplified 'geometric' energy function, a modified potential for NOE restraints with asymptotically linear slope for large violations [17–19], and simulated annealing [15]. Three different protocols for simulated annealing by molecular dynamics, each using a different way to produce the starting structure for the molecular dynamics run, were established: 'Hybrid distance geometry-dynamical simulated annealing' [17] used a start conformation obtained from metric matrix distance geometry, the second method started from an extended polypeptide chain [18], and the third from a random array of atoms [19]. These protocols were implemented in the molecular dynamics program X-PLOR [20], that was written especially for biomolecular structure determination by NMR and X-ray diffraction, and its later successors CNS [21] and Xplor-NIH [22].

It became clear that a method working in torsion angle space and using simulated annealing by molecular dynamics would benefit from the advantages of both approaches because the absence of high-frequency bond length and bond angle vibrations in torsion angle space would allow for longer integration time steps and/or higher temperatures during the simulated annealing. Mazur and Abagyan [23,24] derived explicit formulas

for Lagrange's equations of motion of a polymer using internal coordinates as degrees of freedom. Independently, Bae and Haug [25] and Jain *et al.* [26] found improved torsion angle dynamics algorithms whose computational effort scaled linearly with the system size, as in Cartesian space molecular dynamics, such that the advantage of longer integration time steps in torsion angle dynamics could be exploited for systems of any size. Both algorithms were adapted for protein structure calculation on the basis of NMR data, the first [25] in the program X-PLOR [27], the other [26] in the programs DYANA and CYANA [28], and in the NIH version of X-PLOR [29]. Experience with these programs confirmed that torsion angle dynamics was the most efficient way to calculate NMR structures of biological macromolecules, and showed that the computation time with DYANA and CYANA was about one order of magnitude shorter than with other programs [28]. Simulated annealing by torsion angle dynamics became the standard method to calculate NMR protein structures. A recent survey (Table 1 in [30]) revealed that the structure calculation programs that are cited most often in the NMR protein structures deposited to the Protein Data Bank [5] in September 2005–2008 were CYANA [28] (1160 citations), CNS [21] (242 citations), Xplor-NIH [22] (153 citations), ARIA [31,32] (122 citations), DYANA [28] (114 citations), AutoStructure [33] (103 citations), and X-PLOR [20] (75 citations).

When the basic problem of NMR protein structure calculation was solved satisfactorily by these programs, interest turned towards automating the most time-consuming part of NMR spectral analysis, namely the assignment of multidimensional NOESY spectra for the collection of conformational restraints. Because of the extensive degeneracy of the chemical shifts this task is cumbersome and error-prone if done manually. After semiautomatic approaches [34,35], the feasibility of automated NOESY cross-peak assignment was afforded by the NOAH algorithm [36,37] implemented in the program DIANA [13]. Automated NOESY assignment became of practical relevance with the introduction of ambiguous distance restraints [38] that allowed one to make use of NOESY cross-peaks in the structure calculation even if they had multiple possible assignments [39]. Ambiguous distance restraints became the central feature of the ARIA algorithm [31,32,40]. The CANDID [41] algorithm implemented in DYANA and CYANA [42] made use of ambiguous distance restraints, and improved the robustness of structure calculations with automated NOESY assignment by 'network anchoring' and 'constraint combination'. Network anchoring reduced the initial ambiguity of NOESY cross-peak assignments by inducing self-consistency with the network of other assigned NOEs, and constraint combination minimised the impact of erroneous distance restraints on the resulting structure. At present, the combination of the automated assignment of NOESY cross-peaks and the structure calculation with CYANA or ARIA have become the standard approach to protein structure analysis by NMR [30]. Alternative approaches for the automated assignment of NOESY cross-peaks were implemented in the AutoStructure [33], PASD [43], and KNOWNOE [44] algorithms, and in a Bayesian approach [45].

The complete automation of protein structure determination is one of the challenges of biomolecular NMR spectroscopy that has, despite early optimism [46], proved difficult to achieve. The unavoidable imperfections of experimental NMR spectra and the intrinsic ambiguity of peak assignments that results from the limited accuracy of frequency measurements turn the tractable problem of finding the chemical shift assignments from

Peak picking	→	Signal frequencies
Shift assignments	→	Chemical shifts
NOESY assignment	→	Structural restraints
Structure calculation	→	3D structure
Refinement, validation	→	Final structure

Figure 5.2 *Steps of an NMR protein structure determination and their resulting data*

ideal spectra into a formidably difficult one under realistic conditions. Many attempts have been made to automate further parts of the structure determination process, including peak identification [46–60], and the sequence-specific assignment of the chemical shifts [61–107]. However, fully automated NMR structure determination was more demanding than automating individual parts of NMR structure analysis because the cumulative effect of imperfections at successive steps could easily render the overall process unsuccessful (Figure 5.2). Systems designed to handle the whole process therefore generally required certain human interventions [55,62]. Only recently was the purely computational FLYA algorithm [108], that is capable of determining the 3D structure of proteins on the basis of uninterpreted spectra, developed.

Nowadays, most NMR protein structure determinations make use of sophisticated computational methods but nevertheless follow in essence the original approach (Figure 5.2) that was introduced in the early 1980s [109]. Alternative methods that circumvented the chemical shift assignment step [110–116], or replaced the NOESY information by residual dipolar couplings [117–121] or chemical shift data [122,123], have been developed. *De novo* protein structure determination by these approaches have not been reported yet and it remains to be seen whether they will provide the reliability and the structural quality of the conventional method.

5.3 Structure Calculation Algorithms

This section presents the core algorithms for NMR protein structure calculation by simulated annealing in torsion angle space, as implemented in the widely applied programs CYANA [28] and X-PLOR/CNS [20,21].

5.3.1 Molecular Dynamics Simulation versus NMR Structure Calculation

There is a fundamental difference between molecular dynamics simulation that has the aim of simulating the trajectory of a molecular system as realistically as possible in order to extract molecular quantities of interest and NMR structure calculation that is driven by experimental restraints. Classical molecular dynamics simulations rely on a full 'physical' 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 physical 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 optimise 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 is close to the global minimum of the target function. Therefore, the efficiency of NMR structure calculation can be enhanced by simplifying the force field and/or the algorithm without significantly altering the location of the global minimum (the correctly folded structure) but shortening, in terms of the computation time needed, the path 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.

5.3.2 Potential Energy – Target Function

For simulated annealing a simplified potential energy function, the 'target function', is used that includes a simple repulsive potential instead of the Lennard-Jones and electrostatic nonbonded interactions, as well as terms for distance and torsion angle restraints. In Cartesian space the target function also comprises terms to maintain the covalent geometry of the structure by means of harmonic bond length and bond angle potentials, torsion angle potentials, and terms to enforce the proper chiralities and planarities. These terms are not needed in torsion angle space. For instance, in the program X-PLOR [20],

$$E_{\text{pot}} = \sum_{\text{bonds}} k_b(r-r_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} k_\phi(1+\cos(n\phi+\delta)) + \sum_{\text{impropers}} k_\psi(\psi-\psi_0)^2 + \sum_{\text{nonbonded pairs}} k_{\text{repel}}\left[\max\left(0,(sR_{\min})^2\right)-R^2\right]^2 + \sum_{\text{distance restraints}} k_d\Delta_d^2 + \sum_{\text{angle restraints}} k_a\Delta_a^2.$$

k_b, k_θ, k_ϕ, k_ψ, k_{repel}, k_d and k_a denote the various force constants, r the actual and r_0 the correct bond length, respectively, θ the actual and θ_0 the correct bond angle, ϕ the actual torsion angle, ψ the improper angle and ψ_0 the correct improper angle, n the number of minima of the torsion angle potential, δ an offset of the torsion angle and improper potentials, $R_{\min}$ the distance where the van der Waals potential has its minimum, R the actual distance between a nonbonded atom pair, s a scaling factor, and Δ_d and Δ_a the size of the distance or torsion angle restraint violation. As an alternative to the square-well potential, distance restraints are often represented by a potential with linear asymptote for

large violations [20,124], which limits the maximal force exerted by a violated distance constraint. In this case the violation Δ_d of a single distance restraint is computed as

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

Here, d denotes the actual distance, l and u are the lower and upper distance bounds, γ is the slope of the asymptotic potential, and a is the violation at which the potential switches from harmonic to asymptotic behaviour.

In the program CYANA the target function [13,28] is defined such that it is zero if and only if all experimental distance restraints and torsion angle restraints are fulfilled and all nonbonded atom pairs satisfy a check for the absence of steric overlap. A conformation that satisfies the restraints more closely than another one will lead to a lower target function value. The CYANA target function for distance restraints and torsion angle restraints is defined by

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

Upper and lower bounds, $b_{\alpha\beta}$, on distances $d_{\alpha\beta}$ between two atoms α and β, and restraints 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 violated 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 overall weighting factors for the different types of restraints, and $w_{\alpha\beta}^c$ and w_i are relative weighting factors for individual restraints. $\Gamma_i = \pi-(\theta_i^{\max}-\theta_i^{\min})/2$ denotes the half-width of the forbidden range of torsion angle values, and Δ_i is the size of the torsion angle restraint violation. The target function may include additional terms for restraints on vicinal scalar coupling constants, residual dipolar couplings, and pseudocontact shifts, as well as identity and symmetry restraints for symmetric multimers. Alternatives to the simple square potential for violated distance restraints have also been implemented.

5.3.3 Torsion Angle Dynamics

Torsion angle dynamics, i.e. molecular dynamics simulation using torsion angles instead of Cartesian coordinates as degrees of freedom [23–26], provides at present the most efficient way to calculate NMR structures of biological macromolecules. The only degrees of freedom are the torsion angles, i.e. rotations about single bonds, such that the conformation of the molecule is uniquely specified by the values of all torsion angles. The efficiency of the torsion angle dynamics algorithm [26] implemented in the program CYANA, and, previously, in DYANA [28], is due to the fact that it requires a computational effort that increases only linearly with the system size. In contrast, the computation

time for 'naïve' approaches to torsion angle dynamics rises with the third power of the system size [24], 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 that much longer integration time steps can be used, are effective for molecules of all sizes.

5.3.3.1 *Tree Structure*

A key idea of the fast torsion angle dynamics algorithm in CYANA [26,28] 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 or 'clusters' that are connected by n rotatable bonds. 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 side-chains and at the C-terminus. The only degrees of freedom are rotations about single bonds, and parameters that define the position and orientation of the molecule in space. The clusters are numbered from 0 to n. The base cluster has the number $k = 0$. Each of the other clusters, with numbers $k \ge 1$, has a single nearest neighbour in the direction toward the base, which has a number $p(k) < k$. The torsion angle between the rigid bodies $p(k)$ and k is denoted by θ_k. The conformation of the molecule is uniquely specified by the values of all torsion angles, $(\theta_1, \ldots, \theta_n)$.

The following quantities are defined for each cluster k (Figure 5.3): the 'reference point', r_k, which is the position vector of the end point of the bond between the clusters $p(k)$ and k; $v_k = \dot{r}_k$, the velocity of the reference point; ω_k, the angular velocity of the cluster; Y_k, the vector from the reference point to the centre of mass of the cluster; m_k, the mass of the cluster k; I_k, the inertia tensor of the cluster k with respect to the reference point, given by

$$I_k = \sum\nolimits_\alpha m_\alpha I(y_\alpha),$$

where the sum runs over all atoms in the cluster k, m_α is the mass of the atom α, y_α is the vector from the reference point of cluster k to the atom α, and $I(y_\alpha)$ is the symmetric 3×3 matrix defined by the relation $I(\boldsymbol{y})\boldsymbol{x} = \boldsymbol{y} \wedge (\boldsymbol{x} \wedge \boldsymbol{y})$ for all three-dimensional vectors $\boldsymbol{x}$. The symbol '$\wedge$' denotes the vector product. All position vectors are in an inertial frame of reference that is fixed in space.

5.3.3.2 *Kinetic Energy*

The angular velocity vector $\boldsymbol{\omega}_k$ and the linear velocity $\boldsymbol{v}_k$ of the reference point of the rigid body k are calculated recursively from the corresponding quantities of the preceding rigid body $p(k)$:

$$\begin{aligned} \boldsymbol{\omega}_k &= \boldsymbol{\omega}_{p(k)} + \boldsymbol{e}_k\dot{\theta}_k, \\ \boldsymbol{v}_k &= \boldsymbol{v}_{p(k)} - (\boldsymbol{r}_k - \boldsymbol{r}_{p(k)}) \wedge \boldsymbol{\omega}_{p(k)}. \end{aligned}$$

The kinetic energy can then be computed as a sum over all rigid bodies:

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

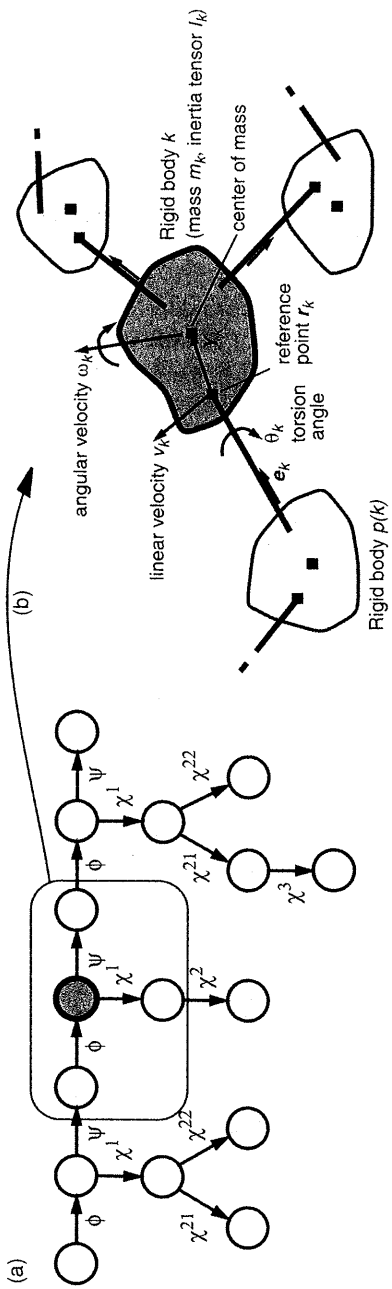

Figure 5.3 *(a) Tree structure of torsion angles for the tripeptide Val-Ser-Ile. Circles represent rigid units. Rotatable bonds are indicated by arrows that point towards the part of the structure that is rotated if the corresponding torsion angle is changed. (b) Excerpt from the tree structure formed by the torsion angles of a molecule, and definition of quantities required by the CYANA torsion angle dynamics algorithm*

5.3.3.3 *Forces = Torques =− Gradient of the Target Function*

The torques about the rotatable bonds, i.e. the negative gradients of the potential energy or target function with respect to torsion angles, $-\nabla V(\theta)$, are calculated by a fast recursive algorithm [12]. The gradient of the target function can be calculated efficiently because the target function is a sum of functions of individual interatomic distances and torsion angles. The partial derivative of the target function V with respect to a torsion angle θ_k is given by

$$\frac{\partial V}{\partial \theta_k} = -\boldsymbol{e}_k \cdot \boldsymbol{f}_k - (\boldsymbol{e}_k \wedge \boldsymbol{r}_k) \cdot \boldsymbol{g}_k + 2w_a \sum_{i \in I_a} w_i \left[1 - \left(\frac{\Delta_i}{\Gamma_i}\right)^2\right] \Delta_i \delta_{ik}$$

with

$$\boldsymbol{f}_k = \sum_{c=u.l.v} w_c \sum_{(\alpha.\beta) \in I_c} \sum_{\alpha \in M_k} w_{\alpha\beta}^c 2 \frac{d_{\alpha\beta} - b_{\alpha\beta}}{d_{\alpha\beta}} (\boldsymbol{r}_\alpha \wedge \boldsymbol{r}_\beta),$$

$$\boldsymbol{g}_k = \sum_{c=u.l.v} w_c \sum_{(\alpha.\beta) \in I_c} \sum_{\alpha \in M_k} w_{\alpha\beta}^c 2 \frac{d_{\alpha\beta} - b_{\alpha\beta}}{d_{\alpha\beta}} (\boldsymbol{r}_\alpha - \boldsymbol{r}_\beta).$$

r_α and r_β are the position vectors of the atoms α und β, and M_k denotes the set of all atoms whose position is affected by a change of the torsion angle θ_k if the base cluster is held fixed. Because M_k is a subset of $M_{p(k)}$, the quantities f_k and g_k can be calculated recursively for $k = n, n-1, \ldots, 1$ starting from the leaves of the tree structure by evaluating the interaction for each atom pair (α, β) only once.

5.3.3.4 *Equations of Motion*

The calculation of the torsional accelerations, i.e. 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 generalised coordinates are the Lagrange equations

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

with the Lagrange function $L = E_{\text{kin}} - E_{\text{pot}}$. They lead to equations of motion of the form

$$M(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) = 0.$$

In the case of n torsion angles as degrees of freedom, the $n \times n$ mass matrix $M(\theta)$ and the n-dimensional vector $C(\theta, \dot{\theta})$ can be calculated explicitly [23,24]. To generate a trajectory

this linear set of n equations would have to be solved in each time step for the torsional accelerations $\ddot{\theta}$, formally by

$$\ddot{\theta} = M(\theta)^{-1} C(\theta, \dot{\theta}).$$

This requires a computational effort proportional to n^3, which is prohibitively expensive for larger systems. Therefore, in CYANA the fast recursive algorithm of [26] is implemented to compute the torsional accelerations, which makes explicit use of the tree structure of the molecule in order to obtain $\ddot{\theta}$ with a computational effort that is only proportional to n. The mathematical details and a proof of correctness of the CYANA torsion angle dynamics algorithm are given in [26].

5.3.3.5 *Torsional Accelerations*

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 algorithm [26] computes a factorisation of the inverse of the mass matrix, $M(\theta)^{-1}$, into a product of highly sparse matrices with nonzero elements only in 6×6 blocks on or near the diagonal. As a result, the torsional accelerations can be obtained by executing a series of three linear loops over all rigid bodies similar to the single loop that is needed to compute the kinetic energy, E_{kin}.

The computation of the torsional accelerations $\ddot{\theta}_k$ is initialised by calculating for all rigid bodies, $k = 1, \ldots, n$, the six-dimensional vectors a_k, e_k and z_k:

$$a_k = \begin{bmatrix} (\boldsymbol{\omega}_k \wedge e_k)\,\dot{\theta}_k \\ \boldsymbol{\omega}_{p(k)} \wedge (\boldsymbol{v}_k - \boldsymbol{v}_{p(k)}) \end{bmatrix},$$

$$e_k = \begin{bmatrix} \boldsymbol{e}_k \\ \boldsymbol{0} \end{bmatrix},$$

$$z_k = \begin{bmatrix} \boldsymbol{\omega}_k \wedge \boldsymbol{I}_k \boldsymbol{\omega}_k \\ (\boldsymbol{\omega}_k \cdot m_k \boldsymbol{Y}_k)\boldsymbol{\omega}_k - |\boldsymbol{\omega}_k|^2 m_k \boldsymbol{Y}_k \end{bmatrix},$$

and the 6×6 matrices P_k and ϕ_k:

$$P_k = \begin{bmatrix} I_k & m_k A(\boldsymbol{Y}_k) \\ -m_k A(I_k) & m_k 1_3 \end{bmatrix},$$

$$\phi_k = \begin{bmatrix} 1_3 & A(r_k - r_{p(k)}) \\ 0_3 & 1_3 \end{bmatrix}.$$

The three-dimensional zero vector is denoted by $\boldsymbol{0}$, 0_3 is the 3×3 zero matrix, 1_3 is the 3×3 unit matrix, and $A(\boldsymbol{y})$ denotes the antisymmetric 3×3 matrix associated with the cross product, defined by the relation $A(\boldsymbol{y})\boldsymbol{x} = \boldsymbol{y} \wedge \boldsymbol{x}$ for all vectors $\boldsymbol{x}$.

Next, several auxiliary quantities are calculated by executing a recursive loop over all rigid bodies in the backward direction, $k = n, n-1, \ldots, 1$:

$$\begin{aligned} D_k &= e_k \cdot P_k e_k \\ G_k &= P_k e_k / D_k \\ \varepsilon_k &= -e_k \cdot (z_k + P_k a_k) - \frac{\partial V}{\partial \theta_k} \end{aligned}$$

$$\begin{aligned} P_{p(k)} &\leftarrow P_{p(k)} + \phi_k (P_k - G_k e_k^T P_k)\, \phi_k^T \\ z_{p(k)} &\leftarrow z_{p(k)} + \phi_k (z_k + P_k a_k - G_k \varepsilon_k). \end{aligned}$$

D_k and ε_k are scalars, G_k is a six-dimensional vector, and '$\leftarrow$' means: 'assign the result of the expression on the right-hand side to the variable on the left-hand side.' Finally, the torsional accelerations are obtained by executing another recursive loop over all rigid bodies in the forward direction, $k = 1, \ldots, n$:

$$\begin{aligned} \alpha_k &= \phi_k^T \alpha_{p(k)} \\ \ddot{\theta}_k &= \varepsilon_k / D_k - G_k \cdot \alpha_k \\ \alpha_k &\leftarrow \alpha_k + e_k \ddot{\theta}_k + a_k. \end{aligned}$$

The auxiliary quantities α_k are six-dimensional vectors, with α_0 being equal to the zero vector.

5.3.3.6 *Time Step*

The integration scheme for the equations of motion in torsion angle dynamics is a variant of the 'leap-frog' algorithm [125] used in Cartesian space molecular dynamics. To obtain a trajectory, the equations of motion are numerically integrated by advancing the $i = 1, \ldots, n$ (generalised) coordinates q_i and velocities $\dot{q}_i$ that describe the system by a small but finite time step Δt:

$$\begin{aligned} \dot{q}_i(t + \Delta t/2) &= \dot{q}_i(t - \Delta t/2) + \Delta t \ddot{q}_i(t) + O(\Delta t^3) \\ q_i(t + \Delta t/2) &= q_i(t) + \Delta t \dot{q}_i(t + \Delta t/2) + O(\Delta t^3) \end{aligned}$$

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 bon 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 [28]. The temperature is controlled by weak coupling to an external bath [126] and the length of the time step is adapted automatically based on the accuracy of energy conservation [28]. 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 [28], whereas time steps in Cartesian space molecular dynamics simulation generally have to be in the range of 2 ns. The concomitant fast exploration of conformation space provides the basis for the efficient CYANA structure calculation protocol.

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.

Furthermore, since 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 speed-up, i.e. an overall computation time almost inversely proportional to the number of processors, can be achieved with CYANA [28].

5.3.4 Simulated Annealing

The potential energy landscape of a protein is complex and studded with many local minima, even in the presence of experimental restraints and when using a simplified target function. Because the temperature, i.e. the 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 [17,20]. In addition to the temperature, other parameters such as force constants and repulsive core radii are varied in these schedules that may involve several stages of heating and cooling. However, the fast exploration of conformation space with torsion angle dynamics allows for simpler schedules.

Protocol for Simulated Annealing

The standard simulated annealing protocol in the program CYANA includes N torsion angle dynamics time steps. It starts from a conformation with all torsion angles treated as independent, uniformly distributed random variables and consists of five stages:

1. *Initial minimisation.* A short conjugate gradient minimisation is applied to reduce high energy interactions that could otherwise disturb the torsion angle dynamics algorithm: 100 conjugate gradient minimisation steps are performed, including only distance restraints between atoms up to 3 residues apart along the sequence, followed by a further 100 minimisation steps including all restraints. For efficiency, all hydrogen atoms are excluded from the check for steric overlap, the repulsive core radii of heavy atoms without covalently bound hydrogen atoms are decreased by 0.2 Å with respect to their standard values, and the radii of heavy atoms with covalently bound hydrogens are decreased by 0.05 Å. The weighting factors in the target function are set to 1 for user-defined upper and lower distance bounds, and to 0.5 for steric lower distance bounds.
2. *First simulated annealing stage with reduced heavy atom radii.* A torsion angle dynamics trajectory with $(N - 200)/3$ time steps is generated. Typically, one-fifth of these torsion angle dynamics steps are performed at a constant high reference temperature T_{high} of, typically, 10 000 K, followed by slow cooling according to a fourth-power law to an intermediate reference temperature $T_{med} = T_{high}/20$. The time step is initialised to 2 fs. The list of van der Waals lower distance bounds is updated every 50 steps using a cutoff equal to twice the largest van der Waals radius plus 1 Å (=4.2 Å for proteins) for the van der Waals pair list generation throughout all torsion angle dynamics phases.
3. *Second simulated annealing stage with normal heavy atom radii and, later, normal hydrogen atom radii.* The repulsive core radii of all heavy atoms are reset to their standard values, 50 conjugate gradient minimisation steps are performed, and the torsion angle dynamics trajectory is continued for $2(N - 200)/3$ time step starting with an initial time step that is half as long as the last preceding time step. The reference temperature is decreased according to a fourth-power law from the intermediate temperature T_{med} to zero reference temperature. After two-thirds of these time steps, the hydrogen atoms are included, with their standard radii, in the steric overlap check, and 50 conjugate gradient minimisation steps are performed before continuing the trajectory, starting with a time step that is half as long as the last preceding time step.
4. *Low temperature phase with increased weight for steric repulsion.* The weighting factor for steric restraints is increased to 2, and 50 conjugate gradient minimisation steps are performed, followed by 200 torsion angle dynamics steps at zero reference temperature, starting with a time step that is half as long as the last preceding time step.
5. *Final minimisation.* A final minimisation with a maximum of, typically, 1000 conjugate gradient steps is applied.

5.4 Automated NOE Assignment

Obtaining a comprehensive set of distance restraints from a NOESY spectrum is in practice by no means straightforward. Resonance and peak overlap turn NOE assignment into an iterative process in which preliminary structures, calculated from limited numbers of distance restraints, serve to reduce the ambiguity of the cross-peak assignments. Additional difficulties may arise from spectral artifacts and noise, and from the absence of expected signals because of fast relaxation. These inevitable shortcomings of NMR data collection are the main reason why laborious interactive procedures have dominated this central step of NMR protein structure determination for a long time. Automated procedures follow the same general scheme as the interactive approach but do not require manual intervention during the assignment/structure calculation cycles. Two main obstacles have to be overcome by an automated method starting without any prior knowledge of the structure. First, the number of cross-peaks with unique assignments based on chemical shift alignment alone is in general not sufficient to define the fold of the protein [127]. An automated method must therefore also have the capability to use NOESY cross-peaks that cannot (yet) be assigned unambiguously. Secondly, the automated program must be able to cope with the erroneously picked or inaccurately positioned peaks and with the incompleteness of the chemical shift assignment of typical experimental data sets. An automated procedure needs devices to substitute for the intuitive decisions made by an experienced spectroscopist in dealing with the imperfections of experimental NMR data.

Besides semi-automatic approaches [34,35,128], several algorithms have been developed for the automated analysis of NOESY spectra given the chemical shift assignments of the backbone and side chain resonances, namely NOAH [36,37], ARIA [31,32,40,129], AUTOSTRUCTURE [33], KNOWNOE [44], CANDID [41] and a similar algorithm implemented in CYANA [130], PASD [43], and a Bayesian approach [45]. Automated NOE assignment algorithms generally require a high degree of completeness of the backbone and side-chain chemical shift assignments [131].

5.4.1 Ambiguity of Chemical Shift Based NOESY Assignment

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

Because of the limited accuracy of chemical shift values and peak positions many NOESY cross-peaks cannot be attributed to a single unique spin pair but have an ambiguous NOE assignment comprising multiple spin pairs. A simple mathematical model of the NOESY assignment process by chemical shift matching gives insight into this problem [37]. 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 NOESY spectrum with an accuracy of the peak position of $\Delta\omega$, i.e. the position of the picked peak differs from the resonance frequency of the underlying signal by no more than $\Delta\omega$ in both spectral dimensions. Under the simplifying assumption of a uniform distribution of the proton chemical shifts over a range $\Delta\Omega$, the chemical shift of a given proton falls within an interval of half-width $\Delta\omega$ about a given peak position with probability $p = 2\Delta\omega/\Delta\Omega$. Peaks with unique chemical shift-based assignment have in both spectral dimensions exactly 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 with 100 amino acid residues, $n = 500$ proton chemical shifts, and $N = 2000$ NOESY cross-peaks within a range of $\Delta\Omega = 10$ ppm, one expects that only about 2 % of the NOEs can be assigned unambiguously based solely on chemical shift information with an accuracy of $\Delta\omega = 0.02$ ppm, which is an insufficient number to calculate a preliminary three-dimensional structure. For peak lists obtained from 3D ^{13}C- or ^{15}N-resolved NOESY spectra, the ambiguity in one of the proton dimensions can usually be resolved by reference to the hetero-spin, so that the expected number of unambiguously assignable NOEs becomes

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

With regard to 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.

5.4.2 Ambiguous Distance Restraints

Ambiguous distance restraints [39] provide a powerful concept for handling ambiguities in the initial, chemical shift-based NOESY cross-peak assignments. Prior to the introduction of ambiguous distance restraints in the ARIA algorithm [40], in general only unambiguously assigned NOEs could be used as distance restraints in the structure calculation. Since the majority of NOEs cannot be assigned unambiguously from chemical shift information alone, this lack of a general way to include ambiguous data into the structure calculation considerably hampered the performance of early automatic NOESY assignment algorithms. When using ambiguous distance restraints, every NOESY cross-peak is treated as the superposition of the signals from each of its possible assignments by applying relative weights proportional to the inverse sixth power of the corresponding interatomic distances. 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 interpreted as the superposition of n degenerate signals and interpreted as an ambiguous distance restraint, $d_{\text{eff}} \leq b$, with the 'effective' or 'r^{-6}-summed' distance

$$d_{\text{eff}} = \left(\sum_{k=1}^{n} d_k^{-6}\right)^{-1/6}.$$

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. The effective distance d_{eff} is always shorter than any of the individual distances d_k. Thus, an ambiguous distance restraint will be fulfilled by the correct structure provided that the correct assignment is included amongst its assignment possibilities, regardless of the possible presence of other, incorrect assignment possibilities. Ambiguous distance restraints make it possible to interpret NOESY cross-peaks as correct conformational restraints also if a unique assignment cannot be determined at the outset of a structure determination. Including multiple assignment possibilities, some but not all of which may later turn out to be incorrect, does not result in a distorted structure but only in a decrease of the information content of the ambiguous distance restraints.

5.4.3 Combined Automated NOE Assignment and Structure Calculation with CYANA

A widely used algorithm for the automated interpretation of NOESY spectra is implemented in the NMR structure calculation program CYANA [28,130]. This algorithm is a re-implementation of the former CANDID algorithm [41] on the basis of a probabilistic treatment of the NOE assignment, combined in an iterative process that comprises seven cycles of automated NOE assignment and structure calculation, followed by a final structure calculation using only unambiguously assigned distance restraints. Between successive

cycles, information is transferred exclusively through the intermediary 3D structures. The molecular structure obtained in a given cycle is used to guide the NOE assignments in the following cycle. Otherwise, the same input data are used for all cycles, that is, the amino acid sequence of the protein, one or several chemical shift lists from the sequence-specific resonance assignment, and one or several lists containing the positions and volumes of cross-peaks in 2D, 3D or 4D NOESY spectra. The input may further include previously assigned NOE upper distance bounds or other previously assigned conformational restraints for the structure calculation.

In each cycle, first all assignment possibilities of a peak are generated on the basis of the chemical shift values that match the peak position within given tolerance values, and the quality of the fit is expressed by a Gaussian probability, P_{shifts}. Secondly, in all but the first cycle the probability $P_{\text{structure}}$ for agreement with the preliminary structure from the preceding cycle, represented by a bundle of conformers, is computed as the fraction of the conformers in which the corresponding distance is shorter than the upper distance bound plus the acceptable distance restraint violation cutoff. Thirdly, each assignment possibility is evaluated for its network anchoring (see below), which is quantified by the probability P_{network}. Only assignment possibilities for which the product of the three probabilities is above a threshold,

$$P_{\text{tot}} = P_{\text{shifts}} \cdot P_{\text{structure}} \cdot P_{\text{network}} \geq P_{\text{min}},$$

are accepted (Figure 5.4). Cross-peaks with a single accepted assignment yield a conventional unambiguous distance restraint. Otherwise, an ambiguous distance restraint is generated that embodies multiple accepted assignments.

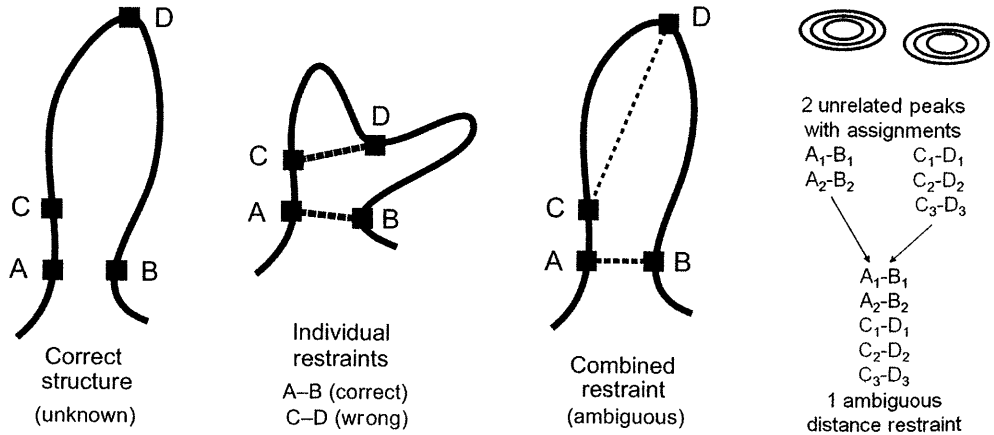

Figure 5.4 *Schematic illustration of the effect of constraint combination in the case of two distance restraints, a correct one connecting atoms A and B, and a wrong one between atoms C and D. A structure calculation that uses these two restraints as individual restraints that have to be satisfied simultaneously will, instead of finding the correct structure (shown, schematically, in the first panel), result in a distorted conformation (second panel), whereas a combined restraint, which will be fulfilled already if one of the two distances is sufficiently short, leads to an almost undistorted solution (third panel). The formation of a combined restraint from the assignments of two peaks is shown in the right panel*

5.4.4 Network-Anchoring

Each assignment possibility is evaluated for its network anchoring, i.e. its embedding in the network formed by the assignment possibilities of all the other peaks and the covalently restricted short-range distances. The network anchoring probability P_{network} that the distance corresponding to an assignment possibility is shorter than the upper distance bound plus the acceptable violation is computed given the assignments of the other peaks but independent of knowledge of the three-dimensional structure. Contributions to the network anchoring probability for a given, 'current', possible assignment result from other peaks with the same assignment, from pairs of peaks that connect indirectly the two atoms of the current possible assignment via a third atom, and from peaks that connect an atom in the vicinity of the first atom of the current assignment with an atom in the vicinity of the second atom of the current assignment. For network anchoring, short-range distances that are constrained by the covalent geometry take the same role as an unambiguously assigned NOE. Individual contributions to the network anchoring of the current assignment possibility are expressed as probabilities, P_1, P_2, ..., that the distance corresponding to the current assignment possibility satisfies the upper distance bound. The network anchoring probability is obtained from the individual probabilities as $P_{\text{network}} = 1 = (1 - P_1) \cdot (1 - P_2) \ldots$, which is never smaller than the highest probability of an individual network anchoring contribution.

5.4.5 Constraint Combination

In practice, spurious distance restraints may arise from the misinterpretation of noise and spectral artifacts, in particular at the outset of a structure determination, before 3D structure-based filtering of the restraint assignments can be applied. The key technique used in CYANA to reduce structural distortions from erroneous distance restraints is 'constraint combination' [41]. Ambiguous distance restraints are generated with combined assignments from different, in general unrelated, cross-peaks (Figure 5.5). The basic property of ambiguous distance restraints that the restraint will be fulfilled by the correct structure whenever at least one of its assignments is correct, regardless of the presence of additional, erroneous assignments, then implies that such combined restraints have a lower probability of being erroneous than the corresponding original restraints, provided that the fraction of erroneous original restraints is smaller than 50 %. Constraint combination aims at minimising the impact of such imperfections on the resulting structure at the expense of a temporary loss of information. It is applied to medium- and long-range distance restraints in the first two cycles of combined automated NOE assignment and structure calculation with CYANA.

5.4.6 Structure Calculation Cycles

The distance restraints are then included in the input for the structure calculation with simulated annealing by the fast CYANA torsion angle dynamics algorithm [28]. The structure calculations typically comprise seven cycles. The second and subsequent cycles differ from the first cycle by the use of additional selection criteria for cross-peaks and NOE assignments that are based on assessments relative to the protein 3D structure from the preceding cycle. The precision of the structure determination normally improves with each

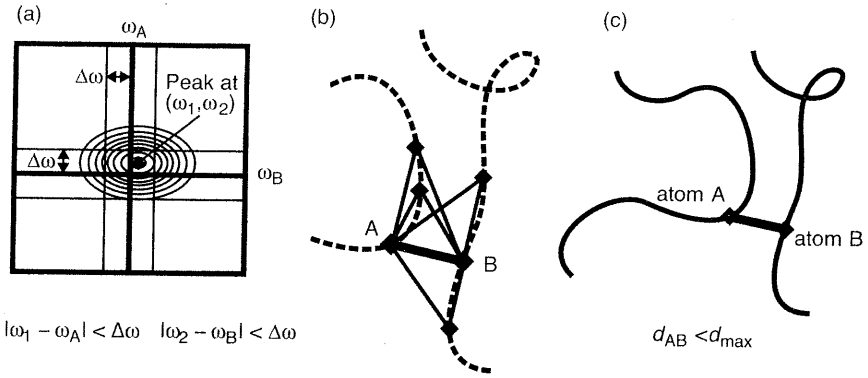

Figure 5.5 *Three conditions that must be fulfilled by a valid assignment of a NOESY cross-peak to two protons A and B in the CYANA automated NOESY assignment algorithm: (a) Agreement between the proton chemical shifts ω_A and ω_B and the peak position (ω_1,ω_2) within a tolerance of $\Delta\omega$. (b) Spatial proximity in a (preliminary) structure. (c) Network-anchoring. The NOE between protons A and B must be part of a network of other NOEs or covalently restricted distances that connect the protons A and B indirectly through other protons*

subsequent cycle. Accordingly, the cutoff for acceptable distance restraint violations in the calculation of $P_{\text{structure}}$ is tightened from cycle to cycle. 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 facilitates the subsequent use of refinement and analysis programs that cannot handle ambiguous distance restraints.

A CYANA structure calculation with automated NOE assignment can be completed in less than one hour for a 10–15 kDa protein, provided that the structure calculations can be performed in parallel, for instance on a Linux cluster system.

5.5 Nonclassical Approaches

Nonclassical approaches that do not rely on sequence-specific resonance assignments and methods using residual dipolar couplings or chemical shifts in conjunction with molecular modelling to determine the backbone structure without the need for side-chain assignments have also been proposed.

5.5.1 Assignment-Free Methods

Much of the NMR measurement time and the spectral analysis effort is devoted to finding sequence-specific resonance assignments. However, the chemical shift assignment by itself has no biological relevance. It is required only as an intermediate step in the interpretation of the NMR spectra. Consequently, strategies for NMR protein structure determination were

sought that circumvented the chemical shift assignment step. Assignment-free NMR structure calculation methods exploit the fact that NOESY spectra provide distance information even in the absence of chemical shift assignments. This proton-proton distance information is used to calculate a spatial proton distribution. Since there is no association with the covalent structure at this point, the protons of the protein are treated as a cloud of unconnected particles. Provided that the emerging proton distribution is sufficiently clear, a model can then be built into the proton density in a manner analogous to X-ray crystallography where a structural model is placed into the electron density.

This general idea was first tested with simulated data [110–114]. The most recent approach to NMR structure determination without chemical shift assignment is the CLOUDS protocol [115,116] which has demonstrated the feasibility of assignment-free structure determination using experimental rather than simulated data. A 'gas' of unassigned, unconnected hydrogen atoms is condensed into a structured proton distribution (cloud) via a molecular dynamics simulated annealing scheme in which the internuclear distances and van der Waals repulsive terms are the only active restraints. Proton densities are generated by combining a large number of such clouds, each computed from a different trajectory. The primary structure is threaded through the unassigned proton density by a Bayesian approach, for which the probabilities of sequential connectivity hypotheses are inferred from likelihoods of H^{N+}-H^N, H^N-H^α, and H^α-H^α interatomic distances as well as ^{1}H NMR chemical shifts, both derived from public databases. Side chains are placed by a similar procedure. As for all NMR spectrum analysis, resonance overlap presents a major difficulty also in applying assignment-free strategies. At present, a *de novo* protein structure determination by the assignment-free approach has not yet been reported.

5.5.2 Methods Based on Residual Dipolar Couplings

Methods using residual dipolar couplings to determine the backbone structure without the need for side-chain assignments have been developed [117]. In a first approach [118] the Protein Data Bank was searched for fragments of seven contiguous amino acid residues that fitted the measured residual dipolar couplings. From consensus values of the torsion angles for the nonterminal residues of these fragments, an initial structure was built from overlapping fragments by 'molecular fragment replacement' (MFR). Errors in the MFR-derived backbone torsion angles accumulate when building the initial model because the long-range information contained in the residual dipolar couplings is not yet used. However, this global orientational information could be reintroduced when using these rough models as starting structures in a subsequent refinement procedure based on a simple iterative gradient approach that adjusted the values of the backbone torsion angles ϕ and ψ to minimise the difference between measured and best-fitted dipolar couplings, and between measured chemical shifts and those predicted by the model. It was demonstrated that the 3D structure of large protein backbone segments, and in favourable cases an entire small protein, could be calculated exclusively from dipolar couplings and chemical shifts [118]. This and similar approaches [119] require assignments of the backbone chemical shifts as input.

In a further step, automated algorithms were developed that simultaneously perform the assignment and the determination of low resolution backbone structures on the basis of

unassigned chemical shifts and residual dipolar couplings [120,121]. The latter method relied on the *de novo* protein structure prediction algorithm ROSETTA [132] and a Monte Carlo search for chemical shift assignments that produced the best fit of the experimental NMR data to a candidate 3D structure.

5.5.3 Chemical Shift-Based Structure Determination

The chemical shift is the NMR parameter than can be measured most easily and accurately. Because the chemical shifts are highly sensitive to their local environment they are widely used to monitor conformational changes or ligand binding, and they can yield information about specific features of protein conformations, notably dihedral angles [133] and secondary structure [134]. However, the complex relationship between chemical shifts and 3D structure has impeded their direct use for tertiary structure determination. Recently, however, two approaches to 3D protein structure determination have been developed that use exclusively chemical shifts as experimental input data [122,123]. The methods do not rely on the quantum mechanical calculation of chemical shifts from first principles but exploit the availability of an ever-growing database of 3D protein structures [5] and corresponding chemical shifts [135] to extract from known protein structures molecular fragment conformations that match the experimentally determined secondary chemical shifts of the protein under study. A secondary chemical shift is the deviation of a chemical shift from the residue type dependent random coil chemical shift value of the corresponding atom. This separates the conformation dependence of the chemical shift from its residue-type dependence, which is a prerequisite for the sequence independent identification of molecular fragments with similar conformation. The molecular fragment conformations are found by extending the database search method of the program TALOS [133] to contiguous segments of several residues [122,136]. The fragment conformations are then assembled into a 3D structure of the entire protein using molecular modelling approaches.

The CHESHIRE algorithm was the first program to generate near-atomic resolution structures from chemical shifts [122]. It first uses the $^{1}H^{\alpha}$, ^{15}N, $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ secondary chemical shifts to predict the secondary structure of the protein and the backbone torsion angles, followed by the identification of three- and nine-residue segments on the basis of the secondary chemical shifts, the predicted secondary structure and the predicted backbone dihedral angles. Low resolution structures in which the side-chains are represented by a single C^{β} atom are calculated by a Monte Carlo algorithm using the CHARMM force field [16] complemented with terms for secondary structure packing and cooperative hydrogen bonding. The previously determined three- and nine-residue fragments guide Monte Carlo moves. All atom conformers are generated. Finally, the 500 best scoring all atom conformers are refined by an Monte Carlo protocol during which an additional energy term is active that describes the correlation between experimental and predicted chemical shifts. The CHESHIRE algorithm yielded the structures of 11 proteins of 46–123 residues with an accuracy of 2 Å or better for the backbone RMSD.

The CS-ROSETTA method is based on the same concept [123]. It combines the ROSETTA structure prediction program [137] with a recently enhanced empirical relation between structure and chemical shifts [136], which allows the selection of database fragments that better match the structure of the unknown protein. Generating new protein structures by CS-ROSETTA involves two separate stages. First, polypeptide fragments are

selected from a protein structural database, based on the combined use of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $13C'$, ^{15}N, $^{1}H^{\alpha}$ and $^{1}H^{N}$ chemical shifts and the amino acid sequence pattern. In the second stage, these fragments are used for *de novo* structure generation, using the standard ROSETTA Monte Carlo assembly and relaxation methods. The method was calibrated using 16 proteins of known structure, and then successfully tested for nine proteins with 65–147 residues under study in a structural genomics project. For these, the CS-ROSETTA algorithm yielded full-atom models with 0.6–2.1 Å RMSD for the backbone atoms relative to the independently determined NMR structures.

Both methods require as experimental input the chemical shift assignments for the backbone and $^{13}C^{\beta}$ atoms. These shifts are generally available at an early stage of the traditional NMR structure determination process, before the collection and analysis of structural restraints. Side chain chemical shift assignments beyond C^{β}, which are considerably harder to obtain than those for the backbone, are not necessary.

In contrast to the NOE-based conventional approach, for which a well-established theory exists relating each piece of NMR data (the NOESY peak volume) to a corresponding conformational restraint, chemical shift-based structure determination is an empirical approach in which it is assumed that the entire sequence of the protein can be covered by overlapping fragments with a similar conformation in already existing structures. There are no experimentally derived long-range conformational restraints. This implies that the correct tertiary structure has to be found – or may be missed – by the underlying molecular modelling algorithm. In practice, convergence decreases with increasing protein size, and is adversely affected by the presence of long, disordered loops [123]. The CS-ROSETTA approach works for proteins up to about 130 residues.

5.6 Fully Automated Structure Analysis

Fully automated NMR structure determination is more demanding than automating individual parts of NMR structure analysis because the cumulative effect of imperfections at successive steps can easily render the overall process unsuccessful. For example, it has been demonstrated that reliable automated NOE assignment and structure calculation requires around 90 % completeness of the chemical shift assignment [41,131], which is not straightforward to achieve by unattended automated peak picking and automated resonance assignments. Present systems designed to handle the whole process therefore generally require certain human interventions [55,62]. The interactive validation of peaks and assignments, however, still constitutes a time-consuming obstacle for high-throughput NMR protein structure determination. The crucial indicator for a fully automated NMR structure determination method is the accuracy of the resulting 3D structures when real experimental input data is used and any human interventions at intermediate steps are avoided. Even 'small' manual corrections, or the use of idealised input data, can lead to substantially altered conclusions, and prejudice the assessment of different methods.

Fully automated structure determination of proteins in solution (FLYA) yields, without human intervention, 3D protein structures starting from a set of multidimensional NMR spectra [108]. As in the classical manual approach, structures are determined by a set of experimental NOE distance restraints without reference to already existing structures or empirical molecular modelling information. In addition to the 3D structure of the protein,

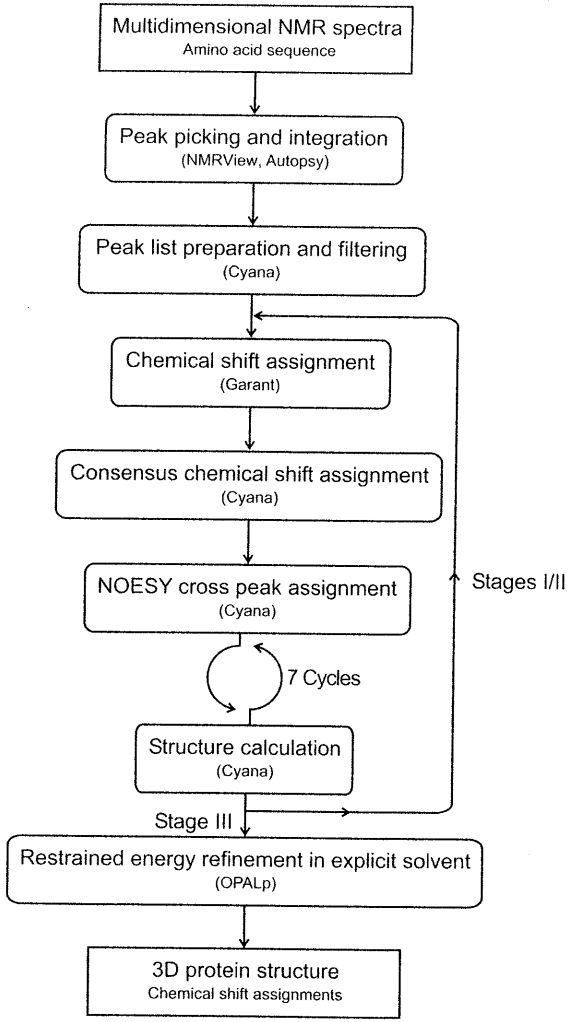

Figure 5.6 *Flowchart of the fully automated structure determination algorithm FLYA*

FLYA yields backbone and side-chain chemical shift assignments, and cross-peak assignments for all spectra.

The FLYA algorithm (Figure 5.6) uses as input data only the protein sequence and multidimensional NMR spectra. Any combination of commonly used hetero- and homonuclear two-, three- and four-dimensional NMR spectra can be used as input for the FLYA algorithm, provided that it affords sufficient information for the assignment of the backbone and side-chain chemical shifts and for the collection of conformational restraints. Peaks are identified in the multidimensional NMR spectra using the automated peak picking algorithm of NMRView [48], or AUTOPSY [47]. Peak integrals for NOESY cross-peaks are determined simultaneously. Since no manual corrections are applied, the resulting raw peak

lists may contain, in addition to the entries representing true signals, a significant number of artifacts (see Figures 5.2 and 5.4 of [108]). The following steps of the fully automated structure determination algorithm can tolerate the presence of such artifacts, as long as the majority of the true peaks have been identified.

Based on the peak positions and, in the case of NOESY spectra, peak volumes, peak lists are prepared by CYANA [28,127]. Depending on the spectra, the preparation may include unfolding aliased signals, systematic correction of chemical shift referencing, and removal of peaks near the diagonal or water lines. The peak lists resulting from this step remain invariable throughout the rest of the procedure. An ensemble of initial chemical shift assignments is obtained by multiple runs of a modified version of the GARANT algorithm [100,101] with different seed values for the random number generator [138]. The original GARANT algorithms was modified for new spectrum types and for the treatment of NOESY spectra when 3D structures are available. In analogy to NMR structure calculation in which not a single structure but an ensemble of conformers is calculated using identical input data but different randomised start conformers, the initial chemical shift assignment produces an ensemble rather than a single chemical shift value for each ^{1}H, ^{13}C and ^{15}N nucleus. The peak position tolerance is typically set to 0.03 ppm for the ^{1}H dimensions and to 0.4 ppm for the ^{13}C and ^{15}N dimensions. These initial chemical shift assignments are consolidated by CYANA into a single consensus chemical shift list. The most highly populated chemical shift value in the ensemble is computed for each ^{1}H, ^{13}C and ^{15}N spin and selected as the consensus chemical shift value that will be used for the subsequent automated assignment of NOESY peaks. The consensus chemical shift for a given nucleus is the value ω that maximises the function

$$\mu(\omega) = \sum\nolimits_j \exp\left(-(\omega-\omega_j)^2/2\Delta\omega^2\right).$$

where the sum runs over all chemical shift values ω_j for the given nucleus in the ensemble of initial chemical shift assignments, and $\Delta\omega$ denotes the aforementioned chemical shift tolerance. NOESY cross-peaks are assigned automatically [41] on the basis of the consensus chemical shift assignments and the same peak lists and chemical shift tolerance values used already for the chemical shift assignment. The automated NOE assignment algorithm of the program CYANA is used. The overall probability for the correctness of possible NOE assignments is calculated as the product of three probabilities that reflect the agreement between the chemical shift values and the peak position, the consistency with a preliminary 3D structure [34], and network-anchoring [41], i.e. the extent of embedding in the network formed by other NOEs. Restraints with multiple possible assignments are represented by ambiguous distance restraints [39]. Seven cycles of combined automated NOE assignment and structure calculation by simulated annealing in torsion angle space and a final structure calculation using only unambiguously assigned distance restraints are performed. Constraint combination [41] is applied in the first two cycles to all NOE distance restraints spanning at least three residues in order to minimise distortions of the structures by erroneous distance restraints that may result from spurious entries in the peak lists and/or incorrect chemical shift assignments.

A complete FLYA calculation comprises three stages. In the first stage, the chemical shifts and protein structures are generated *de novo* (stage I). In the next stages (stages II

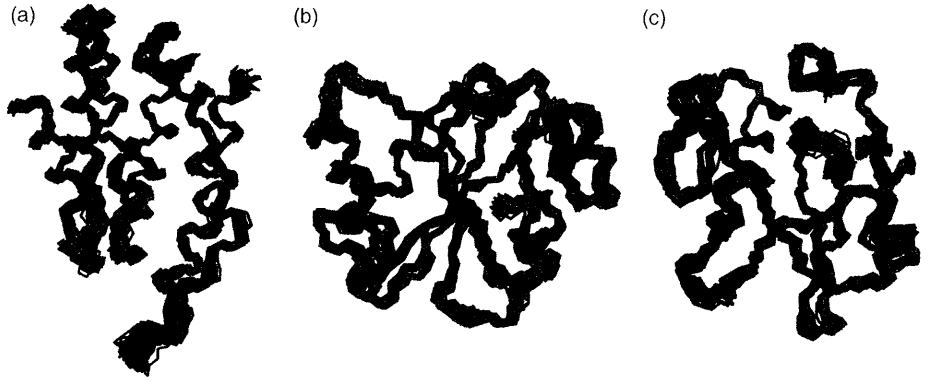

Figure 5.7 *Structures obtained by fully automated structure determination with the FLYA algorithm (blue) superimposed on the corresponding NMR structures determined by conventional methods (dark red). (a) ENTH domain At3g16270(9–135) from Arabidopsis thaliana [147]. (b) Rhodanese homology domain At4g01050(175–295) from Arabidopsis thaliana [148]. (c) Src homology domain 2 (SH2) from the human feline sarcoma oncogene Fes [143]. Please refer to the colour plate section*

and III), the structures generated by the preceding stage are used as additional input for the determination of chemical shift assignments. Stages II and III are particularly important for aromatic residues and other resonances whose assignments rely on through-space NOESY information. At the end of the third stage, the 20 final CYANA conformers with the lowest target function values are subjected to restrained energy minimisation in explicit solvent against the AMBER force field [139] using the program OPALp [140,141]. The complete procedure is driven by the NMR structure calculation program CYANA, which is also used for parallelization of all time-consuming steps. The performance of the FLYA algorithm can be monitored at different steps of the procedure by quality measures that can be computed without referring to external reference assignments or structures [108].

Structure calculations with the FLYA algorithm yielded 3D structures of three 12–16 kDa proteins that coincided closely with the conventionally determined structures (Figure 5.7). Deviations were below 0.95 Å for the backbone atom positions, excluding the flexible chain termini, and 96–97 % of all backbone and side-chain chemical shifts in the structured regions were assigned to the correct residues. The purely computational FLYA method is thus suitable to substitute all manual spectra analysis and overcomes a major efficiency limitation of the NMR method for protein structure determination.

The number of input spectra can be reduced for well-behaved proteins. This is of particular interest because a considerable amount of NMR measurement time was necessary to record the 13–14 input 3D spectra that were used as input for the aforementioned FLYA structure determinations. The influence of reduced sets of experimental spectra on the quality of NMR structures obtained with FLYA was investigated for the 12 kDa Src homology domain 2 from the human feline sarcoma oncogene Fes (Fes SH2) [142]. FLYA calculations were performed for five reduced data sets selected from the complete set of 13 3D spectra of the earlier conventional structure determination [143]. The reduced data sets utilised only CBCA(CO)NH and CBCANH for the backbone assignments and either all,

some or none of the five original side-chain assignment spectra. In four of the five cases tested, the 3D structures deviated by less than 1.3 Å backbone RMSD from the conventionally determined Fes SH2 reference structure. The FLYA algorithm can thus also be used with reduced sets of input spectra.

A further improvement resulted in conjunction with stereo-array isotope labelling (SAIL) [144,145]. SAIL provides a complete stereo- and regiospecific pattern of stable isotopes, which yields much sharper resonance lines and reduced signal overlap without loss of information (see Chapter 2). Automated signal identification can be achieved with higher reliability for the fewer, sharper and more intense peaks of SAIL proteins. The danger of making erroneous assignments decreases with the number of nuclei and peaks to assign, and less spin diffusion allows NOEs to be interpreted more quantitatively. As a result of the superior quality of the SAIL NMR spectra, reliable fully automated analysis of the NMR spectra and structure calculation are possible using fewer input spectra than with conventional uniformly $^{13}C/^{15}N$- labelled proteins. FLYA calculations with SAIL ubiquitin using a single 'through-bond' 3D spectrum in addition to the ^{13}C-edited and ^{15}N-edited NOESY spectra for the restraint collection yielded structures with an accuracy of 0.82–1.15 Å for the backbone RMSD to the conventionally determined solution structure [146], showing the feasibility of fully automated NMR structure analysis from a minimal set of spectra.

References

1. Pflugrath, J.W., Wiegand, G., Huber, R. and Vértesy, L. (1986) Crystal structure determination, refinement and the molecular model of the α-amylase inhibitor Hoe-467a. *J. Mol. Biol.*, **189**, 383–386.
2. Kline, A.D., Braun, W. and Wüthrich, K. (1986) Studies by ^1H nuclear magnetic resonance and distance geometry of the solution conformation of the α-amylase inhibitor tendamistat. *J. Mol. Biol.*, **189**, 377–382.
3. Kline, A.D., Braun, W. and Wüthrich, K. (1988) Determination of the complete three-dimensional structure of the α-amylase inhibitor tendamistat in aqueous solution by nuclear magnetic resonance and distance geometry. *J. Mol. Biol.*, **204**, 675–724.
4. Billeter, M., Kline, A.D., Braun, W. *et al.* (1989) Comparison of the high-resolution structures of the α-amylase inhibitor tendamistat determined by nuclear magnetic resonance in solution and by X-ray diffraction in single crystals. *J. Mol. Biol.*, **206**, 677–687.
5. Berman, H.M., Westbrook, J., Feng, Z. *et al.* (2000) The protein data bank. *Nucleic. Acids Res.*, **28**, 235–242.
6. Blumenthal, L.M. (1953) *Theory and Applications of Distance Geometry*, Cambridge University Press, Cambridge, UK.
7. Braun, W., Bösch, C., Brown, L.R. *et al.* (1981) Combined use of proton-proton overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. Application to micelle-bound glucagon. *Biochim. Biophys. Acta*, **667**, 377–396.
8. Arseniev, A.S., Kondakov, V.I., Maiorov, V.N. and Bystrov, V.F. (1984) NMR solution spatial structure of 'short' scorpion insectotoxin I_5A. *FEBS Lett.*, **165**, 57–62.
9. Havel, T. and Wüthrich, K. (1984) A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular $^1H - ^1H$ proximities in solution. *Bull. Math. Biol.*, **46**, 673–698.
10. Braun, W. and Go, N. (1985) Calculation of protein conformations by proton proton distance constraints - a new efficient algorithm. *J. Mol. Biol.*, **186**, 611–626.
11. Powell, M.J.D. (1977) Restart procedures for the conjugate gradient method. *Math. Program.*, **12**, 241–254.

12. Abe, H., Braun, W., Noguti, T. and Go, N. (1984) Rapid calculation of 1st and 2nd derivatives of conformational energy with respect to dihedral angles for proteins - General recurrent equations. *Comput. Chem.*, **8**, 239–247.
13. 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.
14. Güntert, P. and 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**, 447–456.
15. Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.
16. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D. *et al.* (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.
17. 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.
18. Nilges, M., Gronenborn, A.M., Brünger, A.T. and Clore, G.M. (1988) Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng.*, **2**, 27–38.
19. Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms - circumventing problems associated with folding. *FEBS Lett.*, **239**, 129–136.
20. Brünger, A.T. (1992) *X-PLOR, Version 3.1. A System for X-ray Crystallography and NMR*, Yale University Press, New Haven, CT.
21. Brünger, A.T., Adams, P.D., Clore, G.M. *et al.* (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D*, **54**, 905–921.
22. Schwieters, C.D., Kuszewski, J.J., Tjandra, N. and Clore, G.M. (2003) The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.*, **160**, 65–73.
23. Mazur, A.K. and Abagyan, R.A. (1989) New methodology for computer-aided modeling of biomolecular structure and dynamics 1. Non-cyclic structures. *J. Biomol. Struct. Dyn.*, **6**, 815–832.
24. 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. Comput. Phys.*, **92**, 261–272.
25. Bae, D.S. and Haug, E.J. (1987) A Recursive formulation for constrained mechanical system dynamics. 1. Open loop-systems. *Mech. Struct. Mach.*, **15**, 359–382.
26. Jain, A., Vaidehi, N. and Rodriguez, G. (1993) A fast recursive algorithm for molecular dynamics simulation. *J. Comput. Phys.*, **106**, 258–268.
27. 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.
28. 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.
29. Schwieters, C.D. and Clore, G.M. (2001) Internal coordinates for molecular dynamics and minimization in structure determination and refinement. *J. Magn. Reson.*, **152**, 288–302.
30. Williamson, M.P. and Craven, C.J. (2009) Automated protein structure calculation from NMR data. *J. Biomol. NMR*, **43**, 131–143.
31. Linge, J.P., Habeck, M., Rieping, W. and Nilges, M. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics*, **19**, 315–316.
32. Rieping, W., Habeck, M., Bardiaux, B. *et al.* (2007) ARIA2: Automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, **23**, 381–382.

33. Huang, Y.J., Tejero, R., Powers, R. and Montelione, G.T. (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins*, **62**, 587–603.
34. Güntert, P., Berndt, K.D. and 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.
35. Duggan, B.M., Legge, G.B., Dyson, H.J. and Wright, P.E. (2001) SANE (Structure assisted NOE evaluation): An automated model-based approach for NOE assignment. *J. Biomol. NMR*, **19**, 321–329.
36. 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.
37. Mumenthaler, C., Güntert, P., Braun, W. and Wüthrich, K. (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J. Biomol. NMR*, **10**, 351–362.
38. Nilges, M. (1993) A calculation strategy for the structure determination of symmetrical dimers by ^{1}H-NMR. *Proteins*, **17**, 297–309.
39. Nilges, M. (1995) Calculation of protein structures with ambiguous distance restraints - Automated assignment of ambiguous NOE crosspeaks and disulfide connectivities. *J. Mol. Biol.*, **245**, 645–660.
40. Nilges, M., Macias, M.J., O'Donoghue, S.I. and Oschkinat, H. (1997) Automated NOESY interpretation with ambiguous distance restraints: The refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. *J. Mol. Biol.*, **269**, 408–422.
41. 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.
42. Güntert, P. (2009) Automated structure determination from NMR spectra. *Eur. Biophys. J.*, **38**, 129–143.
43. Kuszewski, J., Schwieters, C.D., Garrett, D.S. *et al.* (2004) Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. *J. Am. Chem. Soc.*, **126**, 6258–6273.
44. Gronwald, W., Moussa, S., Elsner, R. *et al.* (2002) Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE). *J. Biomol. NMR*, **23**, 271–287.
45. Hung, L.H. and Samudrala, R. (2006) An automated assignment-free Bayesian approach for accurately identifying proton contacts from NOESY data. *J. Biomol. NMR*, **36**, 189–198.
46. Pfändler, P., Bodenhausen, G., Meier, B.U. and Ernst, R.R. (1985) Toward automated assignment of nuclear magnetic resonance spectra - pattern recognition in two-dimensional correlation spectra. *Anal. Chem.*, **57**, 2510–2516.
47. Koradi, R., Billeter, M., Engeli, M. *et al.* (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J. Magn. Reson.*, **135**, 288–297.
48. Johnson, B.A. (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules. *Meth. Mol. Biol.*, **278**, 313–352.
49. Garrett, D.S., Powers, R., Gronenborn, A.M. and Clore, G.M. (1991) A common sense approach to peak picking two-, three- and four-dimensional spectra using automatic computer analysis of contour diagrams. *J. Magn. Reson.*, **95**, 214–220.
50. 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.
51. Antz, C., Neidig, K.P. and Kalbitzer, H.R. (1995) A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. *J. Biomol. NMR*, **5**, 287–296.
52. Kleywegt, G.J., Boelens, R. and Kaptein, R. (1990) A versatile approach toward the partially automatic recognition of cross peaks in 2D ^{1}H NMR spectra. *J. Magn. Reson.*, **88**, 601–608.
53. Rouh, A., Louisjoseph, A. and Lallemand, J.Y. (1994) Bayesian signal extraction from noisy FT NMR spectra. *J. Biomol. NMR*, **4**, 505–518.

54. Dancea, F. and Günther, U. (2005) Automated protein NMR structure determination using wavelet de-noised NOESY spectra. *J. Biomol. NMR*, **33**, 139–152.
55. Huang, Y.P.J., Moseley, H.N.B., Baran, M.C. *et al.* (2005) An integrated platform for automated analysis of protein NMR structures. *Methods Enzymol.*, **394**, 111–141.
56. Moseley, H.N.B., Riaz, N., Aramini, J.M. *et al.* (2004) A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra. *J. Magn. Reson.*, **170**, 263–277.
57. Corne, S.A., Johnson, A.P. and Fisher, J. (1992) An artificial neural network for classifying cross peaks in two-dimensional NMR spectra. *J. Magn. Reson.*, **100**, 256–266.
58. Carrara, E.A., Pagliari, F. and Nicolini, C. (1993) Neural networks for the peak picking of nuclear magnetic resonance spectra. *Neural Networks*, **6**, 1023–1032.
59. Neidig, K.P., Saffrich, R., Lorenz, M. and Kalbitzer, H.R. (1990) Cluster analysis and multiplet pattern recognition in two-dimensional NMR spectra. *J. Magn. Reson.*, **89**, 543–552.
60. Meier, B.U., Bodenhausen, G. and Ernst, R.R. (1984) Pattern recognition in two-dimensional NMR spectra. *J. Magn. Reson.*, **60**, 161–163.
61. 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.
62. Gronwald, W. and Kalbitzer, H.R. (2004) Automated structure determination of proteins by NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.*, **44**, 33–96.
63. Baran, M.C., Huang, Y.J., Moseley, H.N.B. and Montelione, G.T. (2004) Automated analysis of protein NMR assignments and structures. *Chem. Rev.*, **104**, 3541–3555.
64. Altieri, A.S. and Byrd, R.A. (2004) Automation of NMR structure determination of proteins. *Curr. Opin. Struct. Biol.*, **14**, 547–553.
65. Volk, J., Herrmann, T. and Wüthrich, K. (2008) Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *J. Biomol. NMR*, **41**, 127–138.
66. Kamisetty, H., Bailey-Kellogg, C. and Pandurangan, G. (2006) An efficient randomized algorithm for contact-based NMR backbone resonance assignment. *Bioinformatics*, **22**, 172–180.
67. Atreya, H.S., Chary, K.V.R. and Govil, G. (2002) Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. *Curr. Sci.*, **83**, 1372–1376.
68. Friedrichs, M.S., Mueller, L. and Wittekind, M. (1994) An automated procedure for the assignment of protein ^{1}HN, ^{15}N, $^{13}C^{\alpha}$, $^{1}H^{\alpha}$, $^{13}C^{\beta}$ and $^{1}H^{\beta}$ resonances. *J. Biomol. NMR*, **4**, 703–726.
69. Hare, B.J. and Prestegard, J.H. (1994) Application of neural networks to automated assignment of NMR spectra of proteins. *J. Biomol. NMR*, **4**, 35–46.
70. Olson, J.B. and Markley, J.L. (1994) Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances: A demonstration of the connectivity tracing assignment tools (CONTRAST) software package. *J. Biomol. NMR*, **4**, 385–410.
71. Buchler, N.E.G., Zuiderweg, E.R.P., Wang, H. and Goldstein, R.A. (1997) Protein heteronuclear NMR assignments using mean-field simulated annealing. *J. Magn. Reson.*, **125**, 34–42.
72. Li, K.B. and Sanctuary, B.C. (1997) Automated resonance assignment of proteins using heteronuclear 3D NMR. 1. Backbone spin systems extraction and creation of polypeptides. *J. Chem. Inf. Comput. Sci.*, **37**, 359–366.
73. Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) Automated probabilistic method for assigning backbone resonances of (C-13,N-15)-labeled proteins. *J. Biomol. NMR*, **9**, 151–166.
74. Zimmerman, D.E., Kulikowski, C.A., Huang, Y.P. *et al.* (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. *J. Mol. Biol.*, **269**, 592–610.
75. Leutner, M., Gschwind, R.M., Liermann, J. *et al.* (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *J. Biomol. NMR*, **11**, 31–43.
76. Atreya, H.S., Sahu, S.C., Chary, K.V.R. and Govil, G. (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). *J. Biomol. NMR*, **17**, 125–136.
77. Bailey-Kellogg, C., Widge, A., Kelley, J.J. *et al.* (2000) The NOESY JIGSAW: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comput. Biol.*, **7**, 537–558.

78. Bailey-Kellogg, C., Chainraj, S. and Pandurangan, G. (2005) A random graph approach to NMR sequential assignment. *J. Comput. Biol.*, **12**, 569–583.
79. Güntert, P., Salzmann, M., Braun, D. and Wüthrich, K. (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. *J. Biomol. NMR*, **18**, 129–137.
80. Bhavesh, N.S., Panchal, S.C. and Hosur, R.V. (2001) An efficient high-throughput resonance assignment procedure for structural genomics and protein folding research by NMR. *Biochemistry*, **40**, 14727–14735.
81. Moseley, H.N.B., Monleon, D. and Montelione, G.T. (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. *Method Enzymol.*, **339**, 91–108.
82. Andrec, M. and Levy, R.M. (2002) Protein sequential resonance assignments by combinatorial enumeration using $^{13}C\alpha$ chemical shifts and their $(i, i-1)$ sequential connectivities. *J. Biomol. NMR*, **23**, 263–270.
83. Chatterjee, A., Bhavesh, N.S., Panchal, S.C. and Hosur, R.V. (2002) A novel protocol based on HN(C)N for rapid resonance assignment in (^{15}N, ^{13}C) labeled proteins: implications to structural genomics. *Biochem. Biophys. Res. Commun.*, **293**, 427–432.
84. Coggins, B.E. and Zhou, P. (2003) PACES: Protein sequential assignment by computer-assisted exhaustive search. *J. Biomol. NMR*, **26**, 93–111.
85. Bernstein, R., Cieslar, C., Ross, A. *et al.* (1993) Computer-assisted assignment of multidimensional NMR spectra of proteins - application to 3D NOESY-HMQC and TOCSY-HMQC Spectra. *J. Biomol. NMR*, **3**, 245–251.
86. Chen, Z.Z., Lin, G.H., Rizzi, R. *et al.* (2005) More reliable protein NMR peak assignment via improved 2-interval scheduling. *J. Comput. Biol.*, **12**, 129–146.
87. Kjaer, M., Andersen, K.V. and Poulsen, F.M. (1994) Automated and semiautomated analysis of homonuclear and heteronuclear multidimensional nuclear magnetic resonance spectra of proteins - the program PRONTO. *Methods Enzymol.*, **239**, 288–307.
88. Lin, H.N., Wu, K.P., Chang, J.M. *et al.* (2005) GANA - a genetic algorithm for NMR backbone resonance assignment. *Nucleic Acids Res.*, **33**, 4593–4601.
89. Masse, J.E. and Keller, R. (2005) AutoLink: Automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. *J. Magn. Reson.*, **174**, 133–151.
90. Vitek, O., Bailey-Kellogg, C., Craig, B. *et al.* (2005) Reconsidering complete search algorithms for protein backbone NMR assignment. *Bioinformatics*, **21**, 230–236.
91. Vitek, O., Bailey-Kellogg, C., Craig, B. and Vitek, J. (2006) Inferential backbone assignment for sparse data. *J. Biomol. NMR*, **35**, 187–208.
92. Wang, J.Y., Wang, T.Z., Zuiderweg, E.R.P. and Crippen, G.M. (2005) CASA: An efficient automated assignment of protein mainchain NMR data using an ordered tree search algorithm. *J. Biomol. NMR*, **33**, 261–279.
93. Wu, K.P., Chang, J.M., Chen, J.B. *et al.* (2006) RIBRA - An error-tolerant algorithm for the NMR backbone assignment problem. *J. Comput. Biol.*, **13**, 229–244.
94. Xu, Y., Xu, D., Kim, D. *et al.* (2002) Automated assignment of backbone NMR peaks using constrained bipartite matching. *Comput. Sci. Eng.*, **4**, 50–62.
95. Xu, Y.Z., Wang, X.X., Yang, J. *et al.* (2006) PASA - A program for automated protein NMR backbone signal assignment by pattern-filtering approach. *J. Biomol. NMR*, **34**, 41–56.
96. Eghbalnia, H.R., Bahrami, A., Wang, L.Y. *et al.* (2005) Probabilistic identification of spin systems and their assignments including coil-helix inference as output (PISTACHIO). *J. Biomol. NMR*, **32**, 219–233.
97. Masse, J.E., Keller, R. and Pervushin, K. (2006) SideLink: Automated side-chain assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. *J. Magn. Reson.*, **181**, 45–67.
98. Xu, J., Straus, S.K., Sanctuary, B.C. and Trimble, L. (1993) Automation of protein 2D proton NMR assignment by means of fuzzy mathematics and graph theory. *J. Chem. Inf. Comput. Sci.*, **33**, 668–682.

99. Xu, J., Straus, S.K., Sanctuary, B.C. and Trimble, L. (1994) Use of fuzzy mathematics for complete automated assignment of peptide ^{1}H 2D NMR spectra. *J. Magn. Reson. B*, **103**, 53–58.
100. Bartels, C., Billeter, M., Güntert, P. and Wüthrich, K. (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *J. Biomol. NMR*, **7**, 207–213.
101. Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) GARANT - A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J. Comput. Chem.*, **18**, 139–149.
102. Choy, W.Y., Sanctuary, B.C. and Zhu, G. (1997) Using neural network predicted secondary structure information in automatic protein NMR assignment. *J. Chem. Inf. Comput. Sci.*, **37**, 1086–1094.
103. Croft, D., Kemmink, J., Neidig, K.P. and Oschkinat, H. (1997) Tools for the automated assignment of high-resolution three-dimensional protein NMR spectra based on pattern recognition techniques. *J. Biomol. NMR*, **10**, 207–219.
104. Li, K.B. and Sanctuary, B.C. (1997) Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. *J. Chem. Inf. Comput. Sci.*, **37**, 467–477.
105. Gronwald, W., Willard, L., Jellard, T. *et al.* (1998) CAMRA: Chemical shift based computer aided protein NMR assignments. *J. Biomol. NMR*, **12**, 395–405.
106. Pristovšek, P., Rüterjans, H. and Jerala, R. (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure - the program st2nmr. *J. Comput. Chem.*, **23**, 335–340.
107. Hitchens, T.K., Lukin, J.A., Zhan, Y.P. *et al.* (2003) MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *J. Biomol. NMR*, **25**, 1–9.
108. López-Méndez, B. and Güntert, P. (2006) Automated protein structure determination from NMR spectra. *J. Am. Chem. Soc.*, **128**, 13112–13122.
109. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, Inc., New York.
110. Malliavin, T.E., Rouh, A., Delsuc, M.A. and Lallemand, J.Y. (1992) Approche directe de la détermination de structures moléculaires à partir de l'effet Overhauser nucléaire. *C. R. Acad. Sci. II*, **315**, 653–659.
111. Oshiro, C.M. and Kuntz, I.D. (1993) Application of distance geometry to the proton assignment problem. *Biopolymers*, **33**, 107–115.
112. Kraulis, P.J. (1994) Protein three-dimensional structure determination and sequence-specific assignment of ^{13}C-separated and ^{15}N-separated NOE data - a novel real-space *ab-initio* approach. *J. Mol. Biol.*, **243**, 696–718.
113. Atkinson, R.A. and Saudek, V. (1997) Direct fitting of structure and chemical shift to NMR spectra. *J. Chem. Soc. Faraday T*, **93**, 3319–3323.
114. Atkinson, R.A. and Saudek, V. (2002) The direct determination of protein structure by NMR without assignment. *FEBS Lett.*, **510**, 1–4.
115. Grishaev, A. and Llinás, M. (2002) CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proc. Natl. Acad. Sci. USA*, **99**, 6707–6712.
116. Grishaev, A. and Llinás, M. (2002) Protein structure elucidation from NMR proton densities. *Proc. Natl. Acad. Sci. USA*, **99**, 6713–6718.
117. Prestegard, J.H., Mayer, K.L., Valafar, H. and Benison, G.C. (2005) Determination of protein backbone structures from residual dipolar couplings. *Methods Enzymol.*, **394**, 175–209.
118. Delaglio, F., Kontaxis, G. and Bax, A. (2000) Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc.*, **122**, 2142–2143.
119. Rohl, C.A. and Baker, D. (2002) De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.*, **124**, 2723–2729.
120. Jung, Y.S., Sharma, M. and Zweckstetter, M. (2004) Simultaneous assignment and structure determination of protein backbones by using NMR dipolar couplings. *Angew. Chem. Int. Edit.*, **43**, 3479–3481.
121. Meiler, J. and Baker, D. (2003) Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15404–15409.
122. Cavalli, A., Salvatella, X., Dobson, C.M. and Vendruscolo, M. (2007) Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 9615–9620.

123. Shen, Y., Lange, O., Delaglio, F. *et al.* (2008). Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 4685–4690.
124. Nilges, M. and O'Donoghue, S.I. (1998) Ambiguous NOEs and automated NOE assignment. *Prog. Nucl. Magn. Reson. Spectrosc.*, **32**, 107–139.
125. Allen, M.P. and Tildesley, D.J. (1987) *Computer Simulation of Liquids*, Clarendon Press, Oxford.
126. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F. *et al.* (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690.
127. Güntert, P. (2003) Automated NMR protein structure calculation. *Prog. Nucl. Magn. Reson. Spectrosc.*, **43**, 105–125.
128. Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) A computer-based protocol for semiautomated assignments and 3D structure determination of proteins. *J. Biomol. NMR*, **4**, 79–96.
129. Habeck, M., Rieping, W., Linge, J.P. and Nilges, M. (2004) NOE assignment with ARIA 2.0: the nuts and bolts. *Meth. Mol. Biol.*, **278**, 379–402.
130. Güntert, P. (2004) Automated NMR structure calculation with CYANA. *Meth. Mol. Biol.*, **278**, 353–378.
131. Jee, J. and Güntert, P. (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *J. Struct. Funct. Genom.*, **4**, 179–189.
132. Simons, K.T., Kooperberg, C., Huang, E. and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
133. Cornilescu, G., Delaglio, F. and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289–302.
134. Wishart, D.S. and Sykes, B.D. (1994) The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR*, **4**, 171–180.
135. Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, **1**, 217–236.
136. Shen, Y. and Bax, A. (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, **38**, 289–302.
137. Bradley, P., Misura, K.M. and Baker, D. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science*, **309**, 1868–1871.
138. Malmodin, D., Papavoine, C.H.M. and Billeter, M. (2003) Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. *J. Biomol. NMR*, **27**, 69–79.
139. Cornell, W.D., Cieplak, P., Bayly, C.I. *et al.* (1995) A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.
140. Luginbühl, P., Güntert, P., Billeter, M. and Wüthrich, K. (1996) The new program OPAL for molecular dynamics simulations and energy refinements of biological macromolecules. *J. Biomol. NMR*, **8**, 136–146.
141. Koradi, R., Billeter, M. and Güntert, P. (2000) Point-centered domain decomposition for parallel molecular dynamics simulation. *Comput. Phys. Commun.*, **124**, 139–147.
142. Scott, A., López-Méndez, B. and Güntert, P. (2006) Fully automated structure determinations of the Fes SH2 domain using different sets of NMR spectra. *Magn. Reson. Chem.*, **44**, S83–S88.
143. Scott, A., Pantoja-Uceda, D., Koshiba, S. *et al.* (2005) Solution structure of the Src homology 2 domain from the human feline sarcoma oncogene Fes. *J. Biomol. NMR*, **31**, 357–361.
144. Kainosho, M., Torizawa, T., Iwashita, Y. *et al.* (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature*, **440**, 52–57.
145. Takeda, M., Ikeya, T., Güntert, P. and Kainosho, M. (2007) Automated structure determination of proteins with the SAIL-FLYA NMR method. *Nature Protocols*, **2**, 2896–2902.

146. Ikeya, T., Takeda, M., Yoshida, H. *et al.* (2009) Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the SAIL-FLYA system. *J. Biomol. NMR*, **44**, 261–272.
147. López-Méndez, B., Pantoja-Uceda, D., Tomizawa, T. *et al.* (2004) Letter to the Editor: NMR assignment of the hypothetical ENTH-VHS domain At3g16270 from Arabidopsis thaliana. *J. Biomol. NMR*, **29**, 205–206.
148. Pantoja-Uceda, D., López-Méndez, B., Koshiba, S. *et al.* (2005) Solution structure of the rhodanese homology domain At4g01050(175–295) from Arabidopsis thaliana. *Protein Sci.*, **14**, 224–230.

6

Paramagnetic Tools in Protein NMR

Peter H.J. Keizers and Marcellus Ubbink

6.1 Introduction

Unpaired electrons have a strong magnetic moment and consequently affect nearby nuclear spins. The nuclear resonances can be broadened or shifted and these paramagnetic effects are distance and orientation dependent in a well-understood fashion. Stable unpaired electrons are found on metals and protected radicals, but this does not mean that the observation of paramagnetic effects is limited to metal proteins. It is possible to generate such effects also in other proteins by the introduction of paramagnetic centres, for example by attaching a paramagnetic tag or substitution of a diamagnetic metal, like Ca^{2+}, with a paramagnetic one, like a lanthanide. Paramagnetic effects offer distance restraints up to 60 Å, a way to cause partial alignment for the generation of RDCs without the need of external media, the possibility to study protein dynamics and to visualise minor populations. Thus, paramagnetic effects are amazingly powerful and complement more classical restraints, like the NOE.

This chapter aims to provide the uninitiated reader sufficient knowledge of paramagnetic NMR tools to enable him/her to select the method of choice for the particular problem at hand. After discussing the type of restraints, choice of metals and the available paramagnetic tags, practical hints are given in a protocol as well as several examples that illustrate the current possibilities. It is not meant as a theoretical description of paramagnetism, for which the reader is referred to other sources [1–3].

Protein NMR Spectroscopy: Practical Techniques and Applications, First Edition.
Edited by Lu-Yun Lian and Gordon Roberts.