147 Fernandez, C.; Adeishvili, K.; Wüthrich, K. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 2358–2363.

148 Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C.A.; Motoshima, H.; Fox, B.A.; Le Trong, I.; Teller, D.C.; Okada, T.; Stenkamp, R.E.; Yamamoto, M.; Miyano, M. *Science* **2000**, *289*, 739–745.

149 Spirin, A.S.; Baranov, V.I.; Ryabova, L.A.; Ovodov, S.Y.; Alakhov, Y.B. *Science* **1988**, *242*, 1162–1164.

150 Kim, D.M.; Kigawa, T.; Choi, C.Y.; Yokoyama, S. *Eur. J. Biochem.* **1996**, *239*, 881–886.

151 Kigawa, T.; Yabuki, T.; Yoshida, Y.; Tsutsui, M.; Ito, Y.; Shibata, T.; Yokoyama, S. *FEBS Lett.* **1999**, *442*, 15–19.

152 Madin, K.; Sawasaki, T.; Ogasawara, T.; Endo, Y. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 559–564.

153 Kim, D.M.; Swartz, J.R. *Biotechnol. Prog.* **2000**, *16*, 385–390.

154 Kim, D.M.; Swartz, J.R. *Biotechnol. Bioeng.* **2001**, *74*, 309–316.

155 Kigawa, T.; Muto, Y.; Yokoyama, S. *J. Biomol. NMR* **1995**, *6*, 129–134.

156 Yabuki, T.; Kigawa, T.; Dohmae, N.; Takio, K.; Terada, T.; Ito, Y.; Laue, E. D.; Cooper, J.A.; Kainosho, M.; Yokoyama, S. *J. Biomol. NMR* **1998**, *11*, 295–306.

157 See www.proteinexpression.com for more information.

158 Fernholz, E.; Besir, H.; Kühlewein, A.; Mayr, D.; Schmitt, R.; Schwaiger, M. unpublished results **2002**.

159 Spirin, A.S. *Cell-Free Translation Systems*; Springer Verlag: Berlin, **2002**.

160 Fernholz, E.; Zaiss, K.; Besir, H.; Mutter, W. In *Cell-Free Translation Systems*; Spirin, A. S., Ed.; Springer Verlag: Berlin, **2002**; pp 175–179.

161 Tarui, H.; Imanishi, S.; Hara, T. *J. Biosci. Bioeng.* **2000**, *90*, 508–514.

162 LeMaster, D.M.; Kushlan, D.M. *J. Am. Chem. Soc.* **1996**, *118*, 9255–9264.

163 Ishima, R.; Louis, J.M.; Torchia, D.A. *J. Am. Chem. Soc.* **1999**, *121*, 11589–11590.

164 Toyoshima, C.; Nakasako, M.; Nomura, H.; Ogawa, H. *Nature* **2000**, *405*, 647–655.

165 Hunte, C.; Koepke, J.; Lange, C.; Rossmanith, T.; Michel, H. *Struct. Fold. Des.* **2000**, *8*, 669–684.

166 Yoshikawa, S.; Shinzawa-Itoh, K.; Nakashima, R.; Yaono, R.; Yamashita, E.; Inoue, N.; Yao, M.; Fei, M. J.; Libeu, C. P.; Mizushima, T.; Yamaguchi, H.; Tomizaki, T.; Tsukihara, T. *Science* **1998**, *280*, 1723–1729.

167 Stock, D.; Leslie, A. G.; Walker, J. E. *Science* **1999**, *286*, 1700–1705.

168 Lancaster, C.R.; Kroger, A.; Auer, M.; Michel, H. *Nature* **1999**, *402*, 377–385.

169 Unger, V.M.; Kumar, N.M.; Gilula, N. B.; Yeager, M. *Science* **1999**, *283*, 1176–1180.

170 Fu, D.; Libson, A.; Miercke, L.J.; Weitzman, C.; Nollert, P.; Krucinski, J.; Stroud, R. M. *Science* **2000**, *290*, 481–486.

171 Kolbe, M.; Besir, H.; Essen, L.O.; Oesterhelt, D. *Science* **2000**, *288*, 1390–1396.

172 Auer, M.; Scarborough, G.A.; Kühlbrandt, W. *Nature* **1998**, *392*, 840–843.

173 Chang, G.; Spencer, R.H.; Lee, A.T.; Barclay, M.T.; Rees, D. C. *Science* **1998**, *282*, 2220–2226.

174 Williams, K.A. *Nature* **2000**, *403*, 112–115.

175 Miyazawa, A.; Fujiyoshi, Y.; Stowell, M.; Unwin, N. *J. Mol. Biol.* **1999**, *288*, 765–786.

176 Lancaster, C.R.; Michel, H. *J. Mol. Biol.* **1999**, *286*, 883–898.

177 Zouni, A.; Witt, H.T.; Kern, J.; Fromme, P.; Krauss, N.; Saenger, W.; Orth, P. *Nature* **2001**, *409*, 739–743.

178 Meyer, J.E.; Hofnung, M.; Schulz, G.E. *J. Mol. Biol.* **1997**, *66*, 761–765.

179 Doyle, D.A.; Morais Cabral, J.; Pfuetzner, R.A.; Kuo, A.; Gulbis, J.M.; Cohen, S.L.; Chait, B. T.; MacKinnon, R. *Science* **1998**, *280*, 69–77.

2
Structure Calculation Using Automated Techniques

Peter Güntert

2.1
Introduction

Understanding the relationship between structure and function of biological macromolecules is one of the key elements of rational drug design. In this context, the three-dimensional structure has a pivotal role, since its knowledge is essential to understand the physical, chemical, and biological properties of a protein [1, 2]. Until 1984 structural information at atomic resolution could only be determined by X-ray diffraction techniques with protein single crystals [3]. The introduction of NMR spectroscopy [4] as a technique for protein structure determination [5] has made it possible to obtain structures with comparable accuracy also in a solution environment that is much closer to the natural situation in a living being than the single crystals required for protein crystallography.

It has been recognized that many of the time-consuming interactive tasks carried out by an expert during the process of spectral analysis could be done more efficiently by automated computational systems [6]. Over the past few years, this potential has been realized to some degree. Today automated methods for NMR structure determination are playing a more and more prominent role and can be expected to largely supersede the conventional manual approaches to solving three-dimensional protein structures in solution.

The structure of this chapter is as follows: Section 2.2 introduces the various types of conformational constraints used in NMR structure calculations. Section 2.3 is devoted to modern structure calculation algorithms. Section 2.4 gives an account of the general principles and the practice of automated NOESY assignment.

2.2
Conformational Constraints for NMR Structure Calculations

For use in a structure calculation, geometric conformational constraints have to be derived from suitable conformation-dependent NMR parameters. These geometric constraints should, on the one hand, convey to the structure calculation as much as possible of the structural information inherent in the NMR data, and, on the other hand, be simple enough to be used efficiently by the structure calculation algorithms. NMR parameters with a clearly understood physical relation to a corresponding geometric parameter

generally yield more trustworthy conformational constraints than NMR data for which the conformation dependence was deduced merely from statistical analyses of known structures.

2.2.1 Constraints from Covalent Structure

NMR data alone would not be sufficient to determine the positions of all atoms in a biological macromolecule. It has to be supplemented by information about the covalent structure of the protein – the amino acid sequence, bond lengths, bond angles, chiralities, and planar groups – and the steric repulsion between nonbonded atom pairs. Depending on the degrees of freedom used in the structure calculation, the covalent parameters are maintained by different methods: in Cartesian space, where in principle each atom moves independently, the covalent structure has to be enforced by potentials in the force field, whereas in torsion angle space the covalent geometry is fixed at the ideal values and there are no degrees of freedom that affect covalent structure parameters.

Depending on the structure calculation program used, special covalent bonds such as disulfide bridges or cyclic peptide bonds have to be enforced by distance constraints. Disulfide bridges may be fixed by restraining the distance between the two sulfur atoms to 2.0–2.1 Å and the two distances between the C^b and the sulfur atoms of different residues to 3.0–3.1 Å [7].

2.2.2 Steric Repulsion

Usually a simple geometric force field is used for the structure calculation that retains only the most dominant part of the nonbonded interaction, namely the steric repulsion in the form of lower bounds for all interatomic distances between pairs of atoms separated by three or more covalent bonds from each other. Steric lower bounds are generated internally by the structure calculation programs by assigning a repulsive core radius to each atom type and imposing lower distance bounds given by the sum of the two corresponding repulsive core radii. For instance, the following repulsive core radii are used in the program Dyana [8]: 0.95 Å (1 Å=0.1 nm) for amide hydrogen, 1.0 Å for other hydrogen, 1.35 Å for aromatic carbon, 1.4 Å for other carbon, 1.3 Å for nitrogen, 1.2 Å for oxygen, and 1.6 Å for sulfur and phosphorus atoms [9]. To allow the formation of hydrogen bonds, potential hydrogen bond contacts are treated with lower bounds that are smaller than the sum of the corresponding repulsive core radii.

2.2.3 Distance Constraints from Nuclear Overhauser Effects

The principle source of experimental conformational data in an NMR structure determination is constraints on short interatomic distances between hydrogen atoms obtained from NMR measurements of the nuclear Overhauser effect (NOE). NOEs result from cross-relaxation mediated by the dipole-dipole interaction between spatially proximate nuclear spins in a molecule undergoing Brownian motion [10] and are manifested by cross peaks in NOESY spectra [11, 12]. NOEs connect pairs of hydrogen atoms separated by less than about 5 Å in amino acid residues that may be far away along the protein sequence but close together in space.

The intensity of an NOE, given by the volume V of the corresponding cross peak in a NOESY spectrum [11, 13, 14] is related to the distance r between the two interacting spins by

$$V = \langle r^{-6} \rangle f(\tau_c) . \quad (1)$$

The averaging indicates that in molecules with inherent flexibility the distance r may vary and thus has to be averaged appropriately. The remaining dependence of the magnetization transfer on the motion enters through the function $f(\tau_c)$, which includes effects of global and internal motions of the molecule. Since, with the exceptions of the protein surface and disordered segments of the polypeptide chain, globular proteins are relatively rigid, the structure calculation is usually based on the assumption that there exists a single rigid conformation that is compatible with all NOE data simultaneously, provided that the NOE data are interpreted in a conservative, semi-quantitative manner [5]. More sophisticated treatments that take into account the fact that the result of a NOESY experiment represents an average over time and space are, if used at all, usually deferred until the structure refinement stage [15].

In principle, all hydrogen atoms of a protein form a single network of spins, coupled by the dipole-dipole interaction. Magnetization can be transferred from one spin to another not only directly but also by "spin diffusion", that is, indirectly via other spins in the vicinity [11, 16]. The approximation of isolated spin pairs is valid only for very short mixing times in the NOESY experiment. However, in order to detect an observable NOE the mixing time cannot be made arbitrarily short. In practice, a compromise has to be made between the suppression of spin diffusion and sufficient cross-peak intensities, usually with mixing times in the range of 40–100 ms for high-quality structures. Spin diffusion effects can be included in the structure calculation by complete relaxation matrix refinement [17–19]. Because also parameters about internal and overall motions that are difficult to measure experimentally enter into the relaxation matrix refinement, care has to be taken not to bias the structure determination by overinterpretation of the data. Relaxation matrix refinement has been used mostly in situations where the conservative and robust interpretation of NOEs as upper distance limits would not be sufficient to define the three-dimensional structure, especially in the case of nucleic acids [20–22].

The quantification of an NOE amounts to determining the volume of the corresponding cross peak in the NOESY spectrum. Since the linewidths can vary appreciably for different resonances, cross-peak volumes should in principle be determined by integration over the peak area rather than by measuring peak heights. However, one should also keep in mind that, according to Eq. (1), the relative error of the distance estimate is only one sixth of the relative error of the volume determination. Furthermore, Eq. (1) involves factors that have their origin in the complex internal dynamics of the macromolecule and are beyond practical reach such that even a very accurate measurement of peak volumes will not yield equally accurate conformational constraints.

On the basis of Eq. (1), NOEs are usually treated as upper bounds on interatomic distances rather than as precise distance constraints, because the presence of internal motions and, possibly, chemical exchange may diminish the strength of an NOE [23]. In fact, much of the robustness of the NMR structure determination method is due to the use of upper distance bounds instead of exact distance constraints in conjunction with the observation that internal motions and exchange effects usually reduce rather than increase the NOEs [5]. For the same reason, the absence of an NOE is in general not interpreted as a lower bound on the distance between the two interacting spins.

Upper bounds b on the distance between two hydrogen atoms are derived from the corresponding NOESY cross peak volumes V according to "calibration curves", $V=f(b)$. Assuming a rigid molecule, the calibration curve is

$$V = \frac{k}{b^6}, \tag{2}$$

with a constant k that depends on the arbitrary scaling of the NOESY spectrum. The value b obtained from the equation may either be used directly as an upper distance bound, or NOEs may be calibrated into different classes according to their volume, using the same upper bound b for all NOEs in a given class. In this case, it is customary to set the upper bound to 2.7 Å for "strong" NOEs, 3.3 Å for "medium" NOEs, and 5.0 Å for "weak" NOEs [7]. The constant k in Eq. (2) can be determined on the basis of known distances, for example the sequential distances in regular secondary structure elements or by reference to a preliminary structure [24]. In an automatic NOESY assignment procedure it is convenient to get an estimate of the calibration constants k independently of knowledge of certain distances or preliminary structures. This can be obtained by automated structure-independent calibration or by automated structure-based calibration. Automated structure-independent calibration [25] defines the calibration constant such that the average of the upper distance bounds for all peaks involving a given combination of atom types attains a predetermined value that has been found to vary little among different structures. Structure-based automated calibration [26] sets the calibration constant such that an available preliminary structure does not violate more than a predetermined (small) percentage of the upper distance bounds.

NOEs that involve groups of protons with degenerate chemical shifts, in particular methyl groups, may be referred to pseudoatoms located in the center of the protons that they represent, and the upper bound is increased by a pseudoatom correction equal to the proton-pseudoatom distance [27, 28]. Another method that usually incurs a smaller loss of information [29] is to treat NOEs for groups of protons with degenerate chemical shifts as ambiguous distance constraints (see Eq. (13) below).

A related but not identical problem occurs because the standard method for obtaining resonance assignments in proteins [5] cannot provide stereospecific assignments, i.e. individual assignments for the two diastereotopic substituents of a prochiral center, for example in methylene groups and in the isopropyl groups of valine and leucine. In the absence of stereospecific assignments, restraints involving diastereotopic substituents can also be referred to pseudoatoms [27] or otherwise treated such that they are invariant under exchange of the two diastereotopic substituents, which inevitably results in a loss of

information and less well-defined structures [30]. To minimize such loss of information, the absence of stereospecific assignments for diastereotopic groups can be treated by periodic optimal swapping of the pairs of diastereotopic atoms for minimal target function value during the simulated annealing [31]. It is also possible to determine stereospecific assignments by various methods, including biosynthetical fractional ^{13}C-labeling of valine and leucine isopropyl groups [32, 33], systematic analysis of the local conformation through grid searches [30], or comparison with preliminary three-dimensional structures [24, 28].

2.2.4
Hydrogen Bond Distance Constraints

Slow hydrogen exchange indicates that an amide proton is involved in a hydrogen bond [34]. However, hydrogen exchange measurements cannot identify the acceptor oxygen or nitrogen atom. Recently, NMR experiments have been developed that can unambiguously identify hydrogen bonds by experimental observation of scalar couplings over hydrogen bonds [35]. If the acceptor oxygen or nitrogen atom cannot be identified experimentally, one has to rely on NOEs in the vicinity of the postulated hydrogen bond or on assumptions about regular secondary structure to define the acceptor. In this way, the standard backbone-backbone hydrogen bonds in regular secondary structure can be identified with higher reliability than hydrogen bonds with side-chains. Therefore, unless based on cross-hydrogen bond scalar couplings, hydrogen bond constraints are either largely redundant with the NOE network or involve structural assumptions and should be used with care or not at all. They can, however, be useful during preliminary structure calculations of larger proteins when not enough NOE data are available yet. Hydrogen bonds constraints are introduced into the structure calculation as distance constraints, typically by confining the acceptor-hydrogen distance to the range 1.8–2.0 Å and the distance between the acceptor and the atom to which the hydrogen atom is covalently bound to 2.7–3.0 Å. The second distance constraint restricts the angle of the hydrogen bond. Being tight medium- or long-range distance constraints, their impact on the resulting structure is considerable. In regular secondary structure elements they significantly enhance their regularity. In fact, α-helices and, to a lesser extent, β-sheets become well defined by hydrogen bond constraints alone without the use of NOE constraints [36]. On the other hand, hydrogen bond constraints may lead, if assigned mechanically without clear-cut evidence, to overly regular structures in which subtle features such as a 3_{10}-helix-like final turn of an α-helix may be missed.

2.2.5
Torsion Angle Constraints from Chemical Shifts

Chemical shifts are very sensitive probes of the molecular environment of a spin. However, in many cases their dependence on the structure is complicated and either not fully understood or too intricate to allow the derivation of reliable conformational constraints [37, 38]. An exception in this respect are the deviations of $^{13}C^{\alpha}$ (and, to some extent, $^{13}C^{\beta}$) chemical shifts from their random coil values that are correlated with the local

backbone conformation [39, 40]: $^{13}C^{\alpha}$ chemical shifts larger than the random coil values occur for amino acid residues in α-helical conformation, whereas deviations towards smaller values are observed for residues in β-sheet conformation. Such information can be included in a structure calculation by restricting the local conformation of a residue to the α-helical or β-sheet region of the Ramachandran plot, using torsion angle constraints in the form of allowed intervals for the backbone torsion angles ϕ and ψ [41]. Some care should be applied because the correlation between chemical shift deviation and structure is not perfect. Similarly to hydrogen bond constraints, conformational constraints based on $^{13}C^{\alpha}$ chemical shifts are therefore in general only used as auxiliary data in special situations, in particular at the beginning of a structure calculation when the NOE network is still sparse. There have also been attempts to use ^{1}H chemical shifts as direct constraints in structure refinement [42, 43]. This is more difficult than with $^{13}C^{\alpha}$ shifts because the secondary structure is not the dominant determinant of ^{1}H chemical shifts. Therefore, ^{1}H chemical shifts are more often used indirectly to delineate secondary structure elements by the "chemical shift index" [44].

2.2.6
Torsion Angle Constraints from Scalar Coupling Constants

Vicinal scalar coupling constants, ^{3}J, between atoms separated by three covalent bonds from each other are related to the enclosed torsion angle, θ, by Karplus relations [45].

$$^{3}J(\theta) = A\cos^2\theta + B\cos\theta + C. \qquad (3)$$

The parameters A, B and C have been determined for various types of couplings by a best fit of the measured J values to the corresponding values calculated with Eq. (3) for known protein structures [36]. When interpreting scalar coupling constants using Eq. (3) one has to take into account not only the measurement error but also that there may be averaging due to internal mobility and that both the functional form and the parameters of the Karplus curves are approximations. In contrast to NOEs, scalar coupling constants give information only on the local conformation of a polypeptide chain. They can nevertheless be useful to accurately define the local conformation, to obtain stereospecific assignments for diastereotopic protons (usually for the β-protons) [30], and to detect torsion angles (usually χ^1) that occur in multiple rotamer states.

Torsion angle constraints in the form of an allowed interval are used to incorporate scalar coupling information into the structure calculation. Using Eq. (3), an allowed range for a scalar coupling constant value in general leads to several (up to four) allowed intervals for the enclosed torsion angle [36]. Restraining the torsion angle to a single interval that encloses all torsion angle values compatible with the scalar coupling constant then often results in a loss of structural information because the torsion angle constraint may encompass large regions that are forbidden by the measured coupling constant. It is therefore often advantageous to combine local data – for example all distance constraints and scalar coupling constants within the molecular fragment defined by the torsion angles ϕ, ψ, and χ^1 – in a systematic analysis of the local conformation and to derive torsion angle constraints from the results of this grid search rather than from the individual NMR parameters [30].

Alternatively, scalar coupling constants can also be introduced into the structure calculation as direct constraints by adding a term of the type

$$V_J = k_J \sum_i ({}^{3}J_i^{\exp} - {}^{3}J_i^{\mathrm{calc}})^2 \qquad (4)$$

to the target function of the structure calculation program [46, 47]. The sum in Eq. (4) extends over all measured couplings, k_J is a weighting factor, and $^{3}J_i^{\exp}$ and $^{3}J_i^{\mathrm{calc}}$ denote the experimental and calculated value of the coupling constant, respectively. The latter is obtained from the value of the corresponding torsion angle using Eq. (3).

2.2.7
Orientation Constraints

Orientation constraints originate from residual dipolar couplings in partially aligned molecules and provide information on angles between covalent bonds and globally defined axes in the molecule, namely those of the magnetic susceptibility tensor [48, 49]. In contrast to vicinal scalar couplings or ^{13}C secondary chemical shifts that probe exclusively local features of the conformation, residual dipolar couplings can provide information on long-range order that is not directly accessible from other commonly used NMR parameters.

Residual dipolar couplings arise because the strong internuclear dipolar couplings are no longer completely averaged out – as is the case in a solution of isotropically oriented molecules – if there is a small degree of molecular alignment with the static magnetic field due to an anisotropy of the magnetic susceptibility. The degree of alignment depends on the strength of the external magnetic field and results in residual dipolar couplings that are proportional to the square of the magnetic field strength [50]. They are manifested in small, field-dependent changes of the splitting normally caused by one-bond scalar couplings between directly bound nuclei and can thus be obtained from accurate measurements of ^{1}J couplings at different magnetic field strengths [48, 49]. The magnetic susceptibility anisotropy is relatively large in paramagnetic proteins but in general very small for diamagnetic globular proteins. It can, however, be enhanced strongly if the protein is brought into a liquid-crystalline environment [51, 52]. One obtains structural information on the angle θ between the covalent bond connecting the two scalar coupled atoms and the main axis of the magnetic susceptibility tensor, which can be incorporated into the structure calculation by adding orientation constraints that measure the deviation between the experimental residual dipolar coupling value and the corresponding value calculated from the structure. It has been shown [53] that such orientation constraints can be used in conjunction with conventional distance and angle constraints during the structure calculation, and that they can improve the quality of the resulting structure.

2.3 Structure Calculation Algorithms

The calculation of the three-dimensional structure forms a cornerstone of the NMR method for protein structure determination. Because of the complexity of the problem – a protein typically consists of more than a thousand atoms which are restrained by a similar number of experimentally determined constraints in conjunction with stereochemical and steric conditions – it is in general neither feasible to do an exhaustive search of allowed conformations nor to find solutions by interactive model building. In practice, the calculation of the three-dimensional structure is therefore usually formulated as a minimization problem for a target function that measures the agreement between a conformation and the given set of constraints. At present, the most widely used algorithms are based on the idea of simulated annealing [54]. These will be discussed in detail here. Earlier methods have been reviewed extensively already [55–58], and most of them are rarely used today. Special emphasis is thus given to the currently most efficient way of calculating NMR structures of biological macromolecules by torsion angle dynamics.

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

2.3.1 Simulated Annealing by Molecular Dynamics Simulation in Cartesian Space

One major method for NMR structure calculation is based on numerically solving Newton's equation of motion in order to obtain a trajectory for the molecular system [59]. The degrees of freedom are the Cartesian coordinates of the atoms. In contrast to "standard"

molecular dynamics simulations [60–62] that try to simulate the behavior of a real physical system as closely as possible (and do not include constraints derived from NMR), the purpose of a molecular dynamics calculation in an NMR structure determination is simply to search the conformation space of the protein for structures that fulfill the constraints, i.e. that minimize a target function which is taken as the potential energy of the system. Therefore, simulated annealing [54, 56, 63] is performed at high temperature using a simplified force field that treats the atoms as soft spheres without attractive or long-range (i.e. electrostatic) nonbonded interactions and does not include explicit consideration of the solvent. The distinctive feature of molecular dynamics simulation when compared to the straightforward minimization of a target function is the presence of kinetic energy that allows barriers of the potential surface to be crossed, thereby greatly reducing the problem of becoming trapped in local minima. Since molecular dynamics simulation cannot generate conformations from scratch, a start structure is needed, and this can be generated either by metric matrix distance geometry [63] or by the variable target function method [9, 28], but – at the expense of increased computation time – it is also possible to start from an extended structure [64] or even from a set of atoms randomly distributed in space [65]. Any general molecular dynamics program, such as Charmm [66], Amber [67], or Gromos [62], can be used for the simulated annealing of NMR structures, provided that pseudoenergy terms for distance and torsion angle constraints have been incorporated. In practice, the programs best adapted and most widely used for this purpose are Xplor [68] and its successor, CNS [69].

The classical dynamics of a system of n particles with masses m_i and positions $\mathbf{r}_i$ is governed by Newton's equation of motion,

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i \,, \tag{5}$$

where the forces $\mathbf{F}_i$ are given by the negative gradient of a potential energy function E_{pot} with respect to the Cartesian coordinates: $\mathbf{F}_i = -\nabla E_{\text{pot}}$. For simulated annealing, a simplified potential energy function is used that includes terms to maintain the covalent geometry of the structure by means of harmonic bond length and bond angle potentials, torsion angle potentials, terms to enforce the proper chiralities and planarities, a simple repulsive potential instead of the Lennard-Jones and electrostatic nonbonded interactions, as well as terms for distance and torsion angle constraints. For example, in the program Xplor [68],

$$\begin{aligned} E_{\text{pot}} = & \sum_{\text{bonds}} k_b(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi(1 + \cos(n\phi + \delta)) \\ & + \sum_{\text{dihedrals}} k_\phi(\phi - \delta)^2 + \sum_{\text{nonbonded pairs}} k_{\text{repel}}(\max(0, (sR_{\min})^2 - R^2)) \\ & + \sum_{\text{distance constraints}} k_d\Delta_d^2 + \sum_{\text{angle constraints}} k_a\Delta_a^2 \end{aligned} \tag{6}$$

k_b, k_θ, k_ϕ, 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 tor-

sion angle or improper angle value, 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 constraint violation. As an alternative to the square-well potential of Eq. (6), distance constraints are often represented by a potential with linear asymptote for large violations [68]. To obtain a trajectory, the equations of motion are numerically integrated by advancing the coordinates $\mathbf{r}_i$ and velocities $\mathbf{v}_i$ of the particles by a small but finite time step Δt, for example according to the "leap-frog" integration scheme [59, 70]:

$$\begin{aligned} \mathbf{v}_i(t + \Delta t/2) &= \mathbf{v}_i(t - \Delta t/2) + \Delta t F_i(t)/m_i + O(\Delta t^3) \\ \mathbf{r}_i(t + \Delta t) &= \mathbf{r}_i(t) + \Delta t \mathbf{v}_i(t + \Delta t/2) + O(\Delta t^3)\,. \end{aligned} \tag{7}$$

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, i.e. of the order of 10^{-15} s. In general the highest frequency motions are bond length oscillations. Therefore, the time step can be increased if the bond lengths are constrained to their correct values by the Shake method [71]. To control the temperature the system is loosely coupled to a heat bath [73]. For the simulated annealing of a (possibly distorted) start structure, certain measures have to be taken in order to achieve sampling of the conformation space within reasonable time [63]. In a typical simulated annealing protocol [68], the simulated annealing is performed for a few picoseconds at high temperature, say 2000 K, starting with a very small weight for the steric repulsion that allows atoms to penetrate each other, and gradually increasing the strength of the steric repulsion during the calculation. Subsequently, the system is cooled down slowly for another few picoseconds and finally energy-minimized. This process is repeated for each of the start conformers. The alternative of selecting conformers that represent the solution structure at regular intervals from a single trajectory is used rarely because it is difficult to judge whether the spacing between the "snapshots" is sufficient for good sampling of conformation space. Simulated annealing by molecular dynamics requires substantially more computation time per conformer [68] than pure minimization methods such as the variable target function approach [9, 28, 72], but this potential disadvantage is in general more than compensated by a higher success rate of 40–100% of the start conformers ending up in a conformation in the vicinity of the global minimum. This effect is due to the ability of the simulated annealing algorithm to escape from local minima.

2.3.2 Torsion Angle Dynamics

Torsion angle dynamics, i.e. molecular dynamics simulation using torsion angles instead of Cartesian coordinates as degrees of freedom [8, 74–82], provides at present the most efficient way to calculate NMR structures of biomacromolecules. This is in stark contrast to a widespread but incorrect belief that dynamics in generalized coordinates is hopelessly complicated and cannot be done efficiently. In this section the torsion angle dynamics algorithm implemented in the program Dyana [8] is described in some detail. Dyana employs

the fast torsion angle dynamics algorithm of Jain et al. [78] that requires a computational effort proportional the system size, as is also the case for molecular dynamics simulation in Cartesian space. "Naïve" approaches to torsion angle dynamics require a computational effort proportional to the third power of the system size (e.g. Ref. [77]), and are therefore not suitable for macromolecules. With the fast torsion angle dynamics algorithm in Dyana, the advantages of torsion angle dynamics, especially the much longer integration time steps that can be used, are effective for molecules of all sizes. There is a close analogy between molecular dynamics simulation in Cartesian and torsion angle space [36].

The key idea of the fast torsion angle dynamics algorithm in Dyana is to exploit the fact that a chain molecule such as a protein or nucleic acid can be represented in a natural way as a tree structure consisting of n+1 rigid bodies that are connected by n rotatable bonds (Fig. 2.1) [74, 83]. Each rigid body is made up of one or several mass points (atoms) with invariable relative positions. The tree structure starts from a base, typically

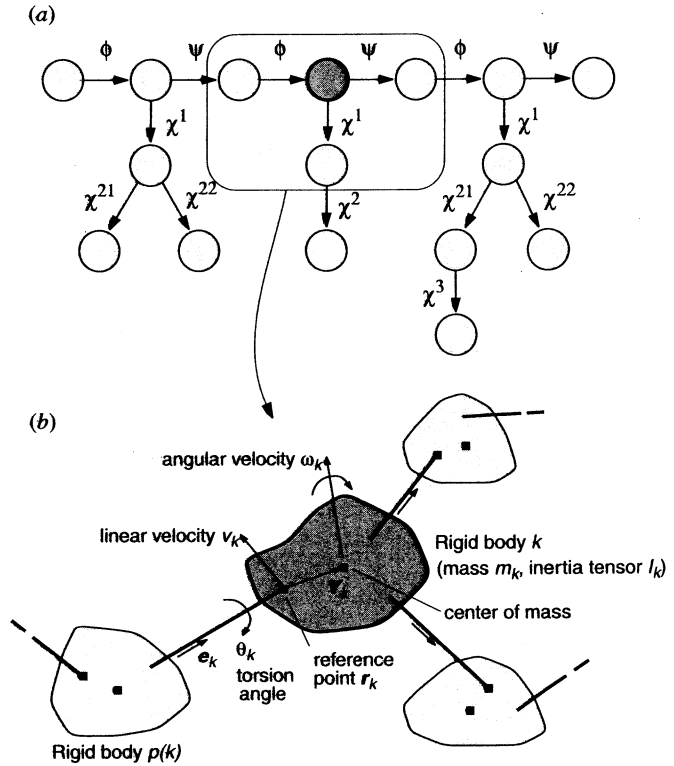

Fig. 2.1 **a** Tree structure of torsion angles for the tripeptide Val-Ser-Ile. Circles represent rigid units. Rotatable bonds are indicated by arrows that point toward the part of the structure that is rotated if the corresponding dihedral angle is changed. **b** Excerpt from the tree structure formed by the torsion angles of a molecule, and definition of quantities required by the Dyana fast torsion angle dynamics algorithm.

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 degrees of freedom are n torsion angles, i.e. rotations about single bonds. The conformation of the molecule is thus uniquely specified by the values of all torsion angles. Covalent bonds that are incompatible with a tree structure because they would introduce closed flexible rings, for example disulfide bridges, are treated, as in Cartesian space dynamics, by distance constraints.

The role of the potential energy is taken by the Dyana target function [8, 28] that is defined such that it is zero if and only if all experimental distance constraints and torsion angle constraints are fulfilled and all nonbonded atom pairs satisfy a check for the absence of steric overlap. A conformation that satisfies the constraints more closely than another one will lead to a lower target function value. The exact definition of the Dyana target function is:

$$V = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta)\in I_c} (d_{\alpha\beta} - b_{\alpha\beta})^2 + w_a \sum_{i\in I_a} \left[1 - \frac{1}{2}\left(\frac{\Delta_i}{\Gamma_i}\right)^2\right] \Delta_i^2 \qquad (8)$$

Upper and lower bounds, $b_{\alpha\beta}$, on distances $d_{\alpha\beta}$ between two atoms α and b, and constraints on individual torsion angles θ_i in the form of allowed intervals $[\theta_i^{\min}, \theta_i^{\max}]$ are considered. I_u, I_l and I_v are the sets of atom pairs (α, β) with upper, lower or van der Waals distance bounds, respectively, and I_a is the set of restrained torsion angles. w_u, w_l, w_v and w_a are weighting factors for the different types of constraints. $\Gamma_i = \pi - (\theta_i^{\max} - \theta_i^{\min})/2$ denotes the half-width of the forbidden range of torsion angle values, and Δ_i is the size of the torsion angle constraint violation. The torques about the rotatable bonds, i.e. the negative gradients of the potential energy with respect to torsion angles, are calculated by the fast recursive algorithm of Abe et al. [83].

The angular velocity vector $\boldsymbol{\omega}_k$ and the linear velocity $\boldsymbol{v}_k$ of the reference point of the rigid body k (Fig. 2.1b) are calculated recursively from the corresponding quantities of the preceding rigid body $p(k)$:

$$\begin{aligned} \omega_k &= \omega_{p(k)} + \boldsymbol{e}_k \dot{\theta}_k \,, \\ \boldsymbol{v}_k &= \boldsymbol{v}_{p(k)} - (\boldsymbol{r}_k - \boldsymbol{r}_{p(k)}) \wedge \omega_{p(k)} \,. \end{aligned} \qquad (9)$$

Denoting the vector from the reference point to the center of mass of the rigid body k by $\boldsymbol{Y}_k$, its mass by m_k, and its inertia tensor by I_k (Fig. 2.1b), the kinetic energy can be computed in a linear loop over all rigid bodies

$$E_{\text{kin}} = \frac{1}{2} \sum_{k=0}^{n} [m_k \boldsymbol{v}_k^2 + \omega_k \cdot I_k \omega_k + 2\boldsymbol{v}_k \cdot (\omega_k \wedge m_k \boldsymbol{Y}_k)] \,. \qquad (10)$$

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 generalized coordinates are the Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_k}\right) - \frac{\partial L}{\partial \theta_k} = 0 \qquad (k = 1, K, n) \qquad (11)$$

with the Lagrange function $L = E_{\text{kin}} - E_{\text{pot}}$ [84]. They lead to equations of motion of the form

$$M(\ddot{\theta}) + C(\theta, \dot{\theta}) = 0 \,. \qquad (12)$$

In the case of torsion angles as degrees of freedom, the mass matrix $M(\theta)$ and the n-dimensional vector $C(\theta, \dot{\theta})$ can be calculated explicitly [76, 77]. However, to integrate the equations of motion, Eq. (12) would have to be solved in each time step for the torsional accelerations $\ddot{\theta}$). This requires the solution of a system of n linear equations and hence entails a computational effort proportional to n^3 that would become prohibitively expensive for larger systems. Therefore, in Dyana the fast recursive algorithm of Jain et al. [78] 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 Dyana algorithm is too involved to be explained in detail here. Suffice it to say that the torsional accelerations can be obtained by executing a series of three linear loops over all rigid bodies similar to the one in Eq. (10) used to compute the kinetic energy.

The integration scheme for the equations of motion in torsion angle dynamics is a variant of the leap-frog algorithm used in Cartesian dynamics. In addition to the basic scheme of Eq. (7), the temperature is controlled by weak coupling to an external bath [73], and the length of the time step is adapted automatically based on the accuracy of energy conservation [8]. 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 (10000 K) temperatures, respectively, can be used in torsion angle dynamics calculations with Dyana [8], 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 Dyana structure calculation protocol.

The potential energy landscape of a protein is complex and studded with many local minima, even in the presence of experimental constraints in a simplified target function of the type of Eq. (8). Because the temperature, i.e. kinetic energy, determines the maximal height of energy barriers that can be overcome in a molecular dynamics simulation, the temperature schedule is important for the success and efficiency of a simulated annealing calculation. Consequently, elaborated protocols have been devised for structure calculations using molecular dynamics in Cartesian space [63, 68]. In addition to the temperature, other parameters such as force constants and repulsive core radii are varied in these schedules, which may involve several stages of heating and cooling. The faster exploration of conformation space with torsion angle dynamics allows for much simpler schedules. The standard simulated annealing protocol used by the program Dyana [8] will serve as an example here.

The structure calculation is started from a conformation with all torsion angles treated as independent uniformly distributed random variables and consists of five stages:

Stage 1. Short minimization to reduce high energy interactions that could otherwise disturb the torsion angle dynamics algorithm: 100 conjugate gradient minimization steps are performed, including only distance constraints between atoms up to 3 residues apart along the sequence, followed by a further 100 minimization steps including all constraints. For efficiency, until step 4 below, all hydrogen atoms are excluded from the check for steric overlap, and the repulsive core radii of heavy atoms with covalently bound hydrogens are increased by 0.15 Å with respect to their standard values. The weights in the target function of Eq. (8) are set to 1 for user-defined upper and lower distance bounds, to 0.5 for steric lower distance bounds, and to 5 $Å^2$ for torsion angle constraints.

Stage 2. Torsion angle dynamics calculation at constant high temperature: One fifth of all N torsion angle dynamics steps are performed at a constant high reference temperature of, typically, 10,000 K. The time step is initialized to 2 fs.

Stage 3. Torsion angle dynamics calculation with slow cooling close to zero temperature: The remaining $4N/5$ torsion angle dynamics steps are performed during which the reference value for the temperature approaches zero according to a fourth-power law.

Stage 4. Incorporation of all hydrogen atoms into the check for steric overlap: After resetting the repulsive core radii to their standard values and increasing the weighting factor for steric constraints to 2.0, 100 conjugate gradient minimization steps are performed, followed by 200 torsion angle dynamics steps at zero reference temperature.

Stage 5. A final minimization consisting of 1000 conjugate gradient steps.

Throughout the torsion angle dynamics calculation the list of van der Waals lower distance bounds is updated every 50 steps using a cutoff of 4.2 Å for the interatomic distance.

With the Dyana 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 Dyana can execute several thousand torsion angle dynamics steps within minutes of computation time. Computation times are of the order of one minute for NMR-size proteins on generally available computers. Furthermore, since an NMR structure calculation always involves the computation of a group of conformers, it is highly efficient to run calculations of multiple conformers in parallel. Nearly ideal speedup, i.e. a reduction of the computation time by a factor close to the number of processors used, can be achieved with Dyana [8].

2.4 Automated NOESY Assignment

2.4.1 The NOESY Assignment Problem

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) [11, 14]. In order to extract distance constraints from a NOESY spectrum, its cross peaks have to be assigned, i.e. the pairs of hydrogen atoms that give rise to cross peaks have to be identified. The basis for the NOESY assignment

are previously determined ^1H chemical shift values that result from sequence-specific resonance assignment. However, because the accuracy with which NOESY cross peak positions and chemical shift values can be measured experimentally is limited, it is in general not possible to unambiguously assign all NOESY cross peaks on the basis of the known chemical shift values alone. It can be shown [25] that the number of NOESY cross peaks that can be assigned unambiguously from knowledge of the ^1H chemical shifts decreases exponentially with increasing uncertainty of the chemical shift or peak position information and drops below 10% of the total number of cross peaks for typical protein data sets. Obtaining a comprehensive set of distance constraints from a NOESY spectrum is thus by no means straightforward but becomes an iterative process in which preliminary structures, calculated from limited numbers of distance constraints, serve to reduce the ambiguity of cross peak assignments. In addition to this problem of resonance and peak overlap, considerable difficulties may arise from spectral artifacts and noise, and from the absence of expected signals because of fast relaxation. These inevitable shortcomings of NMR data collection are the main reason that until recently laborious interactive procedures have dominated three-dimensional protein structure determinations.

2.4.2 Semi-Automatic Methods

Semi-automated approaches to NOESY assignment [85–87] use the chemical shifts and a model or preliminary structure to provide the user with the list of possible assignments for each cross peak. The user decides interactively about the assignment and/or temporary removal of individual NOESY cross peaks, possibly taking into account supplementary information such as line shapes or secondary structure data, and performs a structure calculation with the resulting (usually incomplete) input. In practice, several cycles of NOESY assignment and structure calculation are required to obtain a high-quality structure.

2.4.3 General Principles of Automatic NOESY Assignment

Automated procedures follow the same general scheme but do not require manual intervention during the assignment/structure calculation cycles (Fig. 2.2). Two main obstacles have to be overcome by an automated approach starting without any prior knowledge of the structure: First, because the number of cross peaks with unique assignment based on chemical shifts is, as pointed out before, in general not sufficient to define the fold of the protein, the automated method must make use also of those NOESY cross peaks that cannot yet be assigned unambiguously. Second, the automated program must be able to deal with the amount of erroneously picked or inaccurately positioned peaks and with the incompleteness of the chemical shift assignment that is present in typical experimental data sets. An automated procedure needs devices to substitute the intuitive decisions made by an experienced spectroscopist in dealing with the imperfections of experimental NMR data.

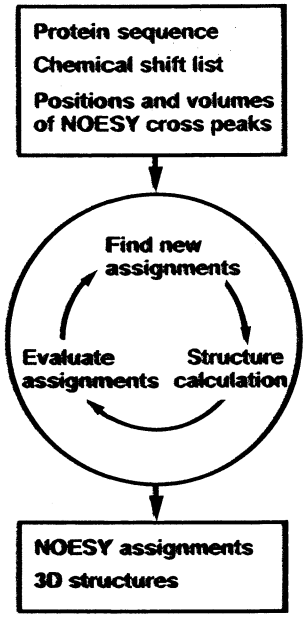

Fig. 2.2 General scheme of automated combined NOESY assignment and structure calculation.

2.4.4 Requirements on Input Data

Current automated NOESY assignment procedures do not attempt to correct or extend the sequence-specific resonance assignments and cannot normally make up for lack of chemical shift assignments. It is therefore important that the input chemical shift list includes nearly complete sequence-specific resonance assignments. Two requirements that the input data should meet in order to be a sufficient basis for a safe and successful automated *de novo* NMR structure determination of a globular protein emerged from test calculations and experience gained in *de novo* structure determinations with the automated NOESY assignment method Candid (see Sect. 2.4.6 below). Other automated NOESY assignment algorithms with fewer built-in safeguards against erroneous input data might call for more stringent requirements:

Requirement 1. Completeness of assignment: The input chemical shift list must contain more than 90% of the nonlabile and backbone amide ^{1}H chemical shifts. If three-dimensional or four-dimensional heteronuclear-resolved [^{1}H,^{1}H]-NOESY spectra are used, more than 90% of the ^{15}N and/or ^{13}C chemical shifts must also be available.

Requirement 2. Self-consistency: The peak lists must be faithful representations of the NOESY spectra, and the chemical shift positions of the NOESY cross peaks must be correctly calibrated to fit the chemical shift lists within the chemical shift tolerances. The range of allowed chemical shift variations ("tolerances") for ^{1}H should not exceed 0.02 ppm when working with homonuclear [^{1}H,^{1}H]-NOESY spectra, or 0.03 ppm when working with heteronuclear-resolved three-dimensional or four-dimensional NOESY spectra, and the tolerances for the ^{15}N and/or ^{13}C shifts should not exceed 0.6 ppm.

The requirement on the completeness of the chemical shift list is very important. A missing (or wrong) entry in the chemical shift list will make it impossible for the algorithm to correctly assign any of the NOEs of the corresponding atom. Therefore, the more NOESY cross peaks are expected for a certain atom, the more important it is to know its chemical shift. Special care should be taken to assign as extensively as possible the chemical shifts of the backbone and the hydrophobic core side-chains, whereas leniency is more tolerable for chemical shifts of flexible hydrophilic side-chains. The second requirement ensures that assignments already present in the input NOESY peak list can be reproduced by Candid. Typically, this includes many intra-residual and sequential NOEs that have been assigned in the preceding sequence-specific assignment and have been used to generate the chemical shift list(s) that are adapted to the NOESY spectra used for structure determination. Chemical shift tolerances should be chosen as small as possible but such that the second requirement is always fulfilled.

2.4.5 Overview of Algorithms

In a first approach to automated NOESY assignment, the programs Diana [28] and Dyana [8] were supplemented with the automated NOESY assignment routine Noah [25, 88]. In Noah, the multiple assignment problem is treated by temporarily ignoring cross peaks with too many (typically, more than two) assignment possibilities and instead generating independent distance constraints for all assignment possibilities of the remaining cross peaks, where one takes into account that part of these distance constraints may be incorrect. Noah requires high accuracy of the chemical shifts and peak positions in the input. It makes use of the fact that only a set of correct assignments can form a self-consistent network, and convergence towards the correct structure has been achieved for several proteins [25].

Another automated NOESY assignment procedure, Aria [89, 90], has been interfaced with the programs Xplor [68] and CNS [69], and a similar approach has been implemented by Savarin et al. [91]. Aria introduced the important concept of ambiguous distance constraints [92] for handling of ambiguities in the initial, chemical shift-based NOESY cross-peak assignments. When ambiguous distance constraints are used, each individual NOESY cross peak is treated as the superposition of the signals from each of its multiple assignments, using relative weights proportional to the inverse sixth power of the corresponding interatomic distance in a preliminary model of the molecular structure. A NOESY cross peak with a unique assignment possibility gives rise to an upper bound b on the distance $d_{\alpha\beta}$ between two hydrogen atoms, α and β. A NOESY cross peak with $n>1$ assignment possibilities can be seen as the superposition of n degenerate signals and interpreted as an ambiguous distance constraint, $\bar{d} \leq b$, with

$$\bar{d} = \left(\sum_{k=1}^{n} d_{a_k \beta_k}^{-6} \right)^{-1/6} . \qquad (13)$$

Each of the distances $d_{\alpha_k\beta_k}$ in the sum corresponds to one assignment possibility, (α_k, β_k). Because the "r^{-6}-summed distance" $\bar{d}$ is always shorter than any of the individual distances $d_{\alpha_k\beta_k}$, an ambiguous distance constraint is never falsified by including incorrect assignment possibilities, as long as the correct assignment is present. However, having more assignment possibilities decreases the information content of an ambiguous distance constraint and makes it more difficult for the structure calculation algorithm to converge to the correct structure.

It is therefore important to eliminate as far as possible incorrect assignment possibilities before the start of the structure calculation. To this end, the assignment possibilities are weighted by their generalized volume contributions, and only those with a sufficiently high contribution enter the ambiguous distance constraints used for the structure calculation. If the three-dimensional structure is known, the normalized relative contribution of the kth individual assignment possibility to the total volume of the cross peak can be estimated by $(d_{\alpha_k\beta_k}/\bar{d})^{-6}$ [89]. In this way, information from cross peaks with an arbitrary number of assignment possibilities can be used for the structure calculation, and although inclusion of erroneous assignments for a given cross peak results in a loss of information, it will not lead to inconsistencies as long as one or several correct assignments are among the initial assignments.

Both of these automated methods are quite efficient for improving and completing the NOESY assignment once a correct preliminary polypeptide fold is available, for example, based on a limited set of interactively assigned NOEs. On the other hand, unless a fair number of long-range assignments is provided by the user, obtaining a correct initial fold at the outset of a *de novo* structure determination often proves to be difficult because the structure-based filters used in both of these procedures for the elimination of erroneous cross peak assignments are then not operational. Aria has been used in the NMR structure determinations of various proteins [90].

A third approach that uses rules for assignments similar to the ones used by an expert to generate an initial protein fold has been implemented in the program AutoStructure, and applied to protein structure determination [6, 93].

The latest approach to automated NOESY assignment is the Candid algorithm [26], which will be explained in detail in the following sections.

2.4.6 The Candid Algorithm

Candid [26] combines features from Noah and Aria, such as the use of three-dimensional structure-based filters and ambiguous distance constraints, with the new concepts of network-anchoring and constraint combination that further enable an efficient and reliable search for the correct fold in the initial cycle of *de novo* NMR structure determinations. A flowchart of the Candid algorithm is given in Fig. 2.3.

The automated Candid method proceeds in iterative *cycles*, each consisting of exhaustive, in part ambiguous, NOE assignment followed by a structure calculation with the Dyana torsion angle dynamics algorithm. Between subsequent cycles, information is transferred exclusively through the intermediary three-dimensional structures, in that the protein molecular structure obtained in a given cycle is used to guide further NOE as-

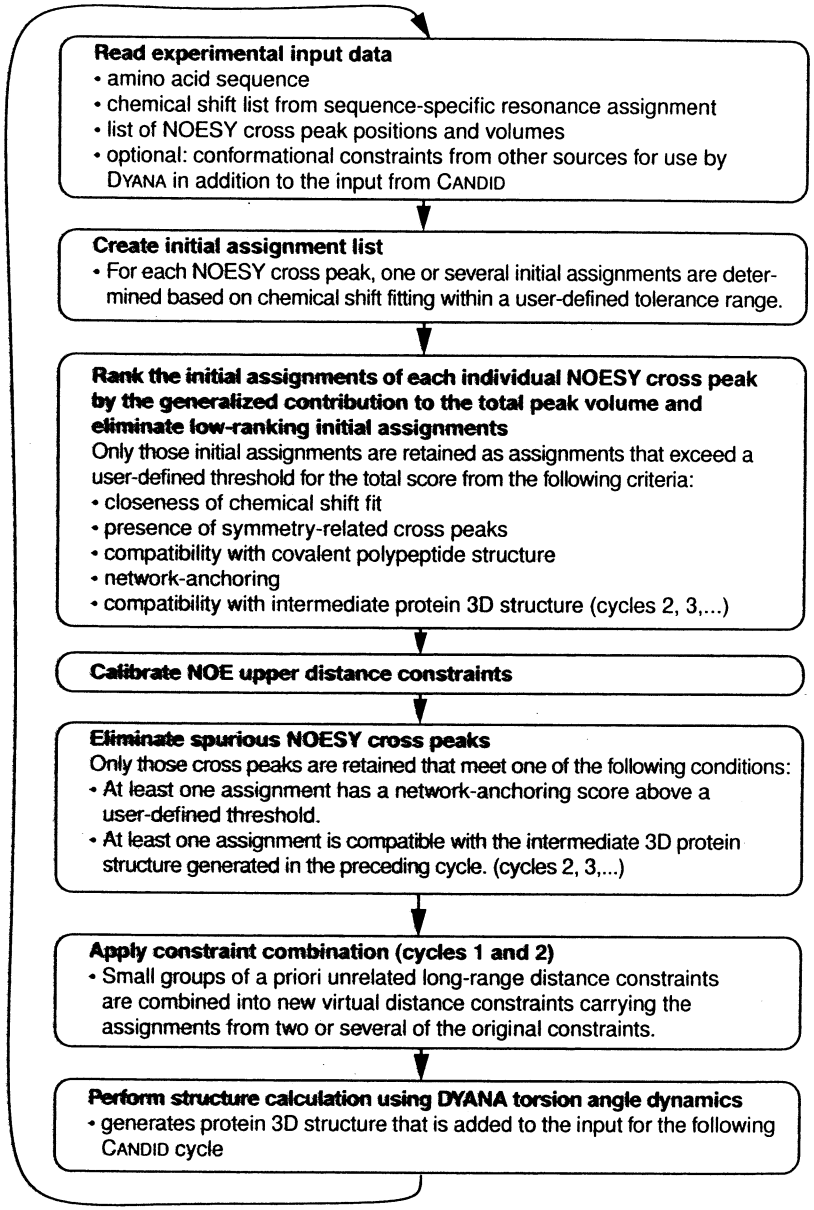

Fig. 2.3 Flowchart of NMR structure determination using the Candid method for automated NOE cross peak assignment.

signments 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 constraints or other previously assigned conformational constraints. These will then not be changed by Candid, but used for the Dyana structure calculation.

A Candid cycle starts by generating for each NOESY cross peak an initial assignment list, i.e., hydrogen atom pairs are identified that could, from the fit of chemical shifts within the user-defined tolerance range, contribute to the peak. Subsequently, for each cross peak these initial assignments are weighted with respect to several criteria (listed in Fig. 2.3), and initial assignments with low overall scores are then discarded. In the first cycle, network anchoring has a dominant impact, since structure-based criteria cannot be applied yet. For each cross peak, the retained assignments are interpreted in the form of an upper distance limit derived from the cross peak volume. Thereby, a conventional distance constraint is obtained for cross peaks with a single retained assignment, and otherwise an ambiguous distance constraint is generated that embodies several assignments. All cross peaks with a poor score are temporarily discarded. In order to reduce deleterious effects on the resulting structure from erroneous distance constraints that may pass this filtering step, long-range distance constraints are incorporated into "combined distance constraints" (Fig. 2.3). The distance constraints are then included in the input for the structure calculation with the Dyana torsion angle dynamics algorithm.

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

In the final Candid cycle, an additional filtering step ensures that all NOEs have either unique assignments to a single pair of hydrogen atoms or are eliminated from the input for the structure calculation. This allows for the direct use of the Candid NOE assignments in subsequent refinement and analysis programs that do not handle ambiguous distance constraints, and in this paper enables direct comparisons of the Candid results with the corresponding data obtained by conventional interactive procedures.

The core of the current version of Candid is implemented in standard Fortran-77 and has been built upon the data structures and into the framework of the user interface of the program Dyana. The standard schedule and parameters for a complete automated structure determination with Candid and Dyana are specified in a script written in the interpreted command language Inclan that gives the user high flexibility in the way automated structure determination is performed without the need to modify the compiled core part of Candid [26].

2.4.7
Network-Anchoring of NOE Assignments

Network-anchoring exploits the observation that the correctly assigned constraints form a self-consistent subset in any network of distance constraints that is sufficiently dense for the determination of a protein three-dimensional structure. Network-anchoring thus evaluates the self-consistency of NOE assignments independently of knowledge of the three-dimensional protein structure, and in this way compensates for the absence of three-dimensional structural information at the outset of a *de novo* structure determination (Fig. 2.4). The requirement that each NOE assignment must be embedded in the network of all other assignments makes network-anchoring a sensitive approach for detecting erroneous, "lonely" constraints that might artificially constrain unstructured parts of the protein. Such constraints would not otherwise lead to systematic constraint violations during the structure calculation, and could therefore not be eliminated by three-dimensional structure-based peak filters.

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

The calculation of the network-anchoring score is recursive in the sense that its calculation for a given peak requires the knowledge of the generalized volume contributions

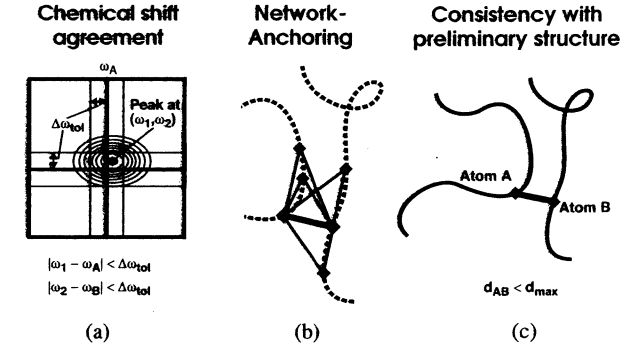

Fig. 2.4 Three conditions that must be fulfilled by valid NOESY cross peak assignments in Candid: **a** Agreement between chemical shifts and the peak position, **b** network-anchoring, and **c** spatial proximity in a (preliminary) structure.

from other peaks, which in turn involve the corresponding network-anchored assignment contributions. Therefore, the calculation of these quantities is iterated three times, or until convergence. Note that the peaks from all peak lists contribute simultaneously to network-anchored assignment.

2.4.8
Constraint-Combination

In the practice of NMR structure determination with biological macromolecules, spurious distance constraints in the input may arise from misinterpretation of stochastic noise, and similar. This situation is particularly critical at the outset of a structure determination, before the availability of a preliminary structure for three-dimensional structure-based screening of constraint assignments. Constraint-combination aims at minimizing the impact of such imperfections on the resulting structure at the expense of a temporary loss of information. Constraint combination is applied in the first two Candid cycles. It consists of generating distance constraints with combined assignments from different, in general unrelated, cross peaks (Fig. 2.5). The basic property of ambiguous distance constraints that the constraint will be fulfilled by the correct structure whenever at least one of its assignments is correct, regardless of the presence of additional, erroneous assignments, then implies that such combined constraints have a lower probability of being erroneous than the corresponding original constraints, provided that the fraction of erroneous original constraints is smaller than 50%.

Candid provides two modes of constraint combination (further combination modes can be envisaged readily): "$2 \rightarrow 1$" combination of all assignments of two long-range peaks each into a single constraint, and "$4 \rightarrow 4$" pair wise combination of the assignments of

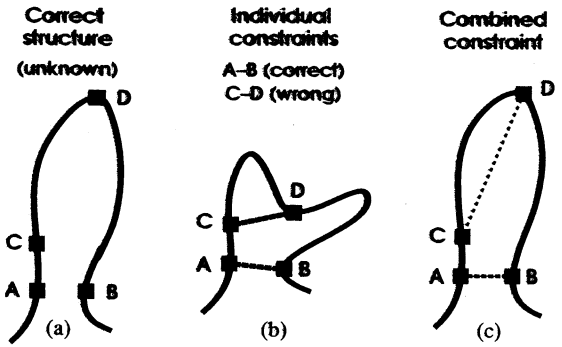

Fig. 2.5 Schematic illustration of the effect of constraint combination in the case of two distance constraints, a correct one connecting atoms A and B, and a wrong one between atoms C and D. A structure calculation that uses these two constraints as individual constraints that have to be satisfied simultaneously will, instead of finding the correct structure (**a**), result in a distorted conformation (**b**), whereas a combined constraint, that will be fulfilled already if one of the two distances is sufficiently short, leads to an almost undistorted solution (**c**).

four long-range peaks into four constraints. Let A, B, C, D denote the sets of assignments of four peaks. Then, $2 \rightarrow 1$ combination replaces two constraints with assignment sets A and B, respectively, by a single ambiguous constraint with assignment set $A \cup B$ (the union of sets A and B). $4 \rightarrow 4$ pair-wise combination replaces four constraints with assignments A, B, C, D by four combined ambiguous constraints with assignment sets $A \cup B$, $A \cup C$, $A \cup D$, and $B \cup C$, respectively. In both cases constraint combination is applied only to the long-range peaks, i.e. the peaks with all assignments to pairs of atoms separated by at least 5 residues in the sequence, because in case of error their effect on the global fold of a protein is much stronger than that of erroneous short- and medium-range constraints. The number of long-range constraints is halved by $2 \rightarrow 1$ combination but stays constant upon $4 \rightarrow 4$ pair-wise combination. The latter approach therefore preserves more of the original structural information, and can furthermore take into account the fact that certain peaks and their assignments are more reliable than others, because the peaks with assignment sets A, B, C, D are used 3, 2, 2, 1 times, respectively, to form combined constraints. To this end, the long-range peaks are sorted according to their total residue-wise network-anchoring and $4 \rightarrow 4$ combination is performed by selecting the assignments A, B, C, D from the first, second, third, and fourth quarter of the sorted list.

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

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

2.4.9
Has it worked?

On the basis of experience gained in *de novo* structure determinations with Candid so far, a set of four criteria for the evaluation of proper performance of combined automated NOESY assignment and structure calculation independent of the availability of an interactively determined reference structure was proposed [26]. These guidelines apply to Candid calculations using input data that fulfills the requirements for the input data presented in Sect. 2.4.4 above and are designed to ensure that the resulting structure has the correct fold if all four criteria are met simultaneously. They do not, however, automatically guarantee a high-quality structure. The four output-based criteria for "safe" Candid runs are:

Criterion 1. Target function: The average final target function value of the structures from the first Candid cycle should be below 250 $Å^2$, and the final target function value of the structure from the last Candid cycle should be below 10 $Å^2$.

Criterion 2. RMSD radius: The average backbone RMSD to the mean coordinates (excluding unstructured parts of the polypeptide chain) should be below 3 Å for the structure from Candid cycle 1.

Criterion 3. RMSD drift: The backbone RMSD between the mean structures of the first and last Candid cycles (excluding unstructured parts of the polypeptide chain) should be smaller than 3 Å and should not exceed by more than 25% the average RMSD to the mean coordinates of cycle 1.

Criterion 4. Eliminated peaks: More than 80% of the NOESY peaks should have been assigned by Candid, and less than 20% of the NOESY cross peaks with exclusively long-range assignments (spanning 5 or more residues) should have been eliminated by the peak filters of Candid.

These criteria again emphasize the crucial importance of getting good results from the first Candid cycle. For reliable automated NMR structure determination, the bundle of conformers obtained after cycle 1 should be reasonably compatible with the input data (criterion 1) and show a defined fold of the protein (criterion 2). Structural changes between the first and subsequent Candid cycles should occur essentially within the conformation space determined by the bundle of conformers obtained after cycle 1, with the implicit assumption that this conformation space contains the correct fold of the protein (criterion 3). The output criteria for target function and RMSD values might need to be slightly relaxed for proteins with more than 150 amino acid residues and tightened for small proteins of less than 80 residues.

If the output of a structure calculation based on automated NOESY assignment with Candid does not fulfill these guidelines, the structure might in many cases still be essentially correct, but it should not be accepted without further validation. Within the framework of Candid, the correct approach is to improve the quality of the input chemical shift and peak lists and to perform another Candid run until the criteria are met. Usually, this can be achieved efficiently because the output from an unsuccessful Candid run, even though the structure should not be trusted *per se*, clearly points out problems in the input, e.g., peaks that cannot be assigned and might therefore be artifacts or indications of erroneous or missing sequence-specific assignments. Candid provides for each peak

informational output that greatly facilitates this task: the list of its chemical shift-based assignment possibilities, the assignment(s) finally chosen, and the reasons why an assignment is chosen or not, or why a peak is not used at all. Of course, even when the criteria are already met, a still higher precision and local accuracy of the structure might be achieved by further improving the input.

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

The Candid method has been evaluated in test calculations [26] and in various *de novo* structure determinations, including, for instance, three mutants of the human prion protein [94], the calreticulin P-domain [95] the pheromone-binding protein from *Bombyx mori* [96] (Fig. 2.6), and the class I human ubiquitin-conjugating enzyme 2b [97]. These structure determinations have confirmed that the new methods of network-anchored assignment and constraint combination enable reliable, truly automated NOESY assignment and structure calculation without prior knowledge about NOESY assignments or the three-dimensional structure. All NOESY assignments and the corresponding distance constraints for these *de novo* structure determinations were made by Candid, confining interactive work to the stage of the preparation of the input chemical shift and peak lists.

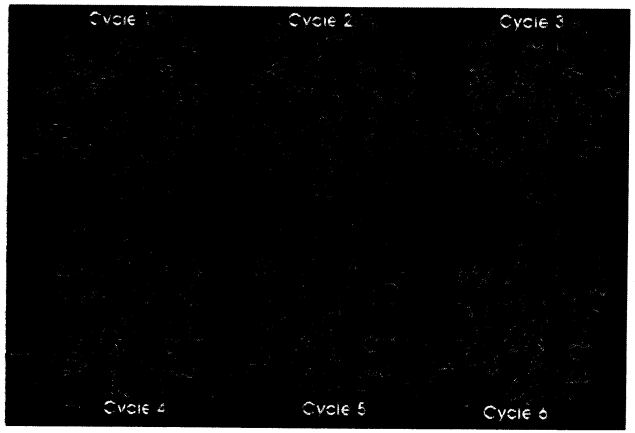

Fig. 2.6 Structures of the pheromone binding protein (form A) from the silk moth *Bombyx mori* [96] obtained in six iterative cycles of combined automated NOESY assignment with Candid and structure calculation with Dyana.

If used sensibly, automated NOESY assignment with Candid has no disadvantage over the conventional interactive approach but is a lot faster and more objective. Network-anchored assignment and constraint combination render the automated Candid method stable also in the presence of the imperfections typical of experimental NMR data sets. Simple criteria basis on the output of Candid allow one to assess the reliability of the resulting structure without cumbersome recourse to independent interactive verification of the NOESY assignments. Candid is a generally applicable, reliable method for automated NOESY assignment. With Candid, the evaluation of NOESY spectra is no longer the time-limiting step in protein structure determination by NMR.

2.5 References

1 Branden, C., Tooze, J. (1991), *Introduction to protein structure*. New York & London: Garland Publishing.

2 Creighton, T. (1993), *Proteins. Structures and molecular properties*. 2nd ed. New York: Freeman.

3 Drenth, J. (1994), *Principles of protein X-ray crystallography*. New York: Springer.

4 Abragam, A. (1961), *Principles of Nuclear Magnetism*. Oxford: Clarendon Press.

5 Wüthrich, K. (1986), *NMR of Proteins and Nucleic Acids*. New York: Wiley.

6 Moseley, H.N.B., Montelione, G.T. (1999), *Curr. Op. Struct. Biol.* **9**, 635–642.

7 Williamson, M.P., Havel, T.F., Wüthrich, K. (1985), *J. Mol. Biol.* **182**, 295–315.

8 Güntert, P., Mumenthaler, C., Wüthrich, K. (1997), *J. Mol. Biol.* **273**, 283–298.

9 Braun, W., Gô, N. (1985), *J. Mol. Biol.* **186**, 611–626.

10 Solomon, I. (1955), *Phys. Rev.* **99**, 559–565.

11 Macura, S., Ernst, R.R. (1980), *Mol. Phys.* **41**, 95–117.

12 Neuhaus, D., Williamson, M.P. (1989), *The nuclear Overhauser effect in structural and conformational analysis*. New York: VCH.

13 Jeener, J., Meier, B.H., Bachmann, P., Ernst, R.R. (1979), *J. Chem. Phys.* **71**, 4546–4553.

14 Kumar, A., Ernst, R.R., Wüthrich, K. (1980), *Biochem. Biophys. Res. Comm.* **95**, 1–6.

15 Torda, A.E., Scheek, R.M., van Gunsteren, W.F. (1989), *Chem. Phys. Lett.* **157**, 289–294.

16 Kalk, A., Berendsen, H.J.C. (1976), *J. Magn. Reson.* **24**, 343–366.

17 Keepers, J.W., James, T.L. (1984), *J. Magn. Reson.* **57**, 404–426.

18 Yip, P., Case, D.A. (1989), *J. Magn. Reson.* **83**, 643–648.

19 Mertz, J.E., Güntert, P., Wüthrich, K., Braun, W. (1991), *J. Biomol. NMR.* **1**, 257–269.

20 Pardi, A. (1995), *Meth. Enzymol.* **261**, 350–380.

21 Varani, G., Aboul-ela, F., Allain, F.H.T. (1996), *Prog. NMR Spectrosc.* **29**, 51–127.

22 Wijmenga, S.S., Mooren, M.M.W., Hilbers, C.W. (1993), In *NMR of macromolecules. A practical approach* (ed. G.C.K. Roberts), pp. 217–288, Oxford: Oxford University Press.

23 Ernst, R.R., Bodenhausen, G., Wokaun, A. (1987), *The principles of nuclear magnetic resonance in one and two dimensions*. Oxford: Clarendon Press.

24 Güntert, P., Qian, Y.Q., Otting, G., Müller, M., Gehring, W.J., Wüthrich K. (1991b), *J. Mol. Biol.* **217**, 531–540.

25 Mumenthaler, C., Güntert, P., Braun, W., Wüthrich, K. (1997), *J. Biomol. NMR* **10**, 351–362.

26 Herrmann, T., Güntert, P., Wüthrich, K. (2002), *J. Mol. Biol.*, **319**, 209–227.

27 Wüthrich, K., Billeter, M., Braun, W. (1983), *J. Mol. Biol.* **169**, 949–961.

28 Güntert, P., Braun, W., Wüthrich, K. (1991), *J. Mol. Biol.* **217**, 517–530.

29 Fletcher, C.M., Jones, D.N.M., Diamond, R., Neuhaus, D. (1996), *J. Biomol. NMR* **8**, 292–310.

30 Güntert, P., Braun, W., Billeter, M., Wüthrich, K. (1989), *J. Am. Chem. Soc.* **111**, 3997–4004.

31 Folmer, R.H.A., Hilbers, C.W., Konings, R.N.H., Nilges, M. (1997), *J. Biomol. NMR* **9**, 245–258.

32 Senn, H., Werner, B., Messerle, B.A., Weber, C., Traber, R., Wüthrich, K. (1989), *FEBS Lett.* **249**, 113–118.

33 Neri, D., Szyperski, T., Otting, G., Senn, H., Wüthrich, K. (1989), *Biochemistry*, **28**, 7510–7516.

34 Wagner, G., Wüthrich, K. (1982), *J. Mol. Biol.* **160**, 343–361.

35 Dingley, A.J., Grzesiek, S. (1998), *J. Am. Chem. Soc.* **120**, 8293–8297.

36 Güntert, P. (1998), *Q. Rev. Biophys.* **31**, 145–237.

37 Oldfield, E. (1995), *Protein Sci.* **5**, 217–225.

38 Williamson, M.P., Asakura, T. (1997), In *Protein NMR techniques* (ed. D. G. Reid), pp. 53–69, Totowa, New Jersey: Humana Press.

39 Spera, S., Bax, A. (1991), *J. Am. Chem. Soc.*, **113**, 5490–5492.

40 de Dios, A.C., Pearson, J.G., Oldfield, E. (1993), *Science* **260**, 1491–1496.

41 Luginbühl, P., Szyperski, T., Wüthrich, K. (1995), *J. Magn. Reson.* **B109**, 229–233.

42 Ösapay, K., Theriault, Y., Wright, P.E., Case, D.A. (1994), *J. Mol. Biol.* **244**, 183–197.

43 Kuszewski, J., Gronenborn, A. M., Clore, G. M. (1995a), *J. Magn. Reson.* **B107**, 293–297.

44 Wishart, D.S., Sykes, B.D., Richards, F.M. (1992), *Biochemistry* **31**, 1647–1651.

45 Karplus, M. (1963), *J. Am. Chem. Soc.* **85**, 2870–2871.

46 Kim, Y., Prestegard, J.H. (1990), *Proteins* **8**, 377–385.

47 Torda, A.E., Brunne, R.M., Huber, T., Kessler, H., van Gunsteren, W.F. (1993), *J. Biomol. NMR.* **3**, 55–66.

48 Tolman, J.R., Flanagan, J.M., Kennedy, M.A., Prestegard, J.H. (1995), *Proc. Natl. Acad. Sci. USA* **92**, 9279–9283.

49 Tjandra, N., Omichinski, J.G., Gronenborn, A.M., Clore, G.M., Bax, A. (1997), *Nature Struct. Biol.* **4**, 732–738.

50 Gayathri, C., Bothner-By, A.A., van Zijl, P.C., MacLean, C. (1982), *Chem. Phys. Lett.* **87**, 192–196.

51 Tjandra, N., Bax, A. (1997), *Science* **278**, 1111–1114.

52 Losonczi, J.A., Prestegard, J.H. (1998), *Biochemistry* **37**, 706–716.

53 Tjandra, N., Grzesiek, S., Bax, A. (1997), *J. Am. Chem. Soc.* **118**, 6264–6272.

54 Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P. (1983), *Science* **220**, 671–680.

55 Braun W. (1987), *Q. Rev. Biophys.* **19**, 115–157.

56 Brünger, A.T., Nilges, M. (1993), *Q. Rev. Biophys.* **26**, 49–125.

57 James, T.L. (1994), *Curr. Opin. Struct. Biol.* **4**, 275–284.

58 Nilges, M. (1996), *Curr. Opin. Struct. Biol.* **6**, 617–623.

59 Allen, M.P., Tildesley, D.J. (1987), *Computer Simulation of Liquids*. Oxford: Clarendon Press.

60 McCammon, J.A., Harvey, S.C. (1987), *Dynamics of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.

61 Brooks III, C.L., Karplus, M., Pettitt, B.M. (1988), *Proteins. A theoretical perspective of dynamics, structure, and thermodynamics*. New York: Wiley.

62 van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Kröger, P., Mark, A.E., Scott, W.R.P., Tironi, I.G. (1996), *Biomolecular Simulation: The Gromos 96 Manual and User Guide*. Zürich: vdf Hochschulverlag.

63 Nilges, M., Clore, G.M., Gronenborn, A.M. (1988a), *FEBS Lett.* **229**, 317–324.

64 Nilges, M., Gronenborn, A.M., Brünger, A.T., Clore, G.M. (1988), *Protein Eng.* **2**, 27–38.

65 Nilges, M., Clore, G.M., Gronenborn, A. M. (1988), *FEBS Lett.* **239**, 129–136.

66 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. (1983), *J. Comp. Chem.* **4**, 187–217.

67 Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz Jr., K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., Kollman, P.A. (1995), *J. Am. Chem. Soc.* **117**, 5179–5197.

68 Brünger, A.T. (1992), *X-PLOR, version 3.1. A system for X-ray crystallography and NMR*. New Haven: Yale University Press.

69 Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges. M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., Warren, G.L. (1998), *Acta Crystallogr.* **D 54**, 905–921.

70 Hockney, R.W. (1970), *Methods Comput. Phys.* **9**, 136–211.

71 Ryckaert, J.-P., Ciccotti, G., Berendsen, H.J.C. (1977), *J. Comput. Phys.* **23**, 327–341.

72 Güntert, P., Wüthrich, K. (1991), *J. Biomol. NMR* **1**, 446–456.

73 Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., Haak, J.R. (1984), *J. Chem. Phys.* **81**, 3684–3690.

74 Katz, H., Walter, R., Somorjay, R.L. (1979), *Computers, Chemistry* **3**, 25–32.

75 Bae, D.S., Haug, E.J. (1987), *Mech. Struct. Mech.* **15**, 359–382.

76 Mazur, A.K., Abagyan, R.A. (1989), *J. Biomol. Struct. Dynam.* **4**, 815–832.

77 Mazur, A.K., Dorofeev, V.E., Abagyan, R.A. (1991), *J. Comp. Phys.* **92**, 261–272.

78 Jain, A., Vaidehi, N., Rodriguez, G. (1993), *J. Comp. Phys.* **106**, 258–268.

79 Kneller, G.R., Hinsen, K. (1994), *Phys. Rev. E* **50**, 1559–1564.

80 Mathiowetz, A.M., Jain, A., Karasawa, N., Goddard III, W.A. (1994), *Proteins* **20**, 227–247.

81 Rice, L.M., Brünger, A.T. (1994), *Proteins* **19**, 277–290.

82 Stein, E.G., Rice, L.M., Brünger, A.T. (1997), *J. Magn. Reson.* **124**, 154–164.

83 Abe, H., Braun, W., Noguti, T., Gô, N. (1984), *Computers, Chemistry* **8**, 239–247.

84 Arnold, V.I. (1978), *Mathematical methods of classical mechanics*. New York: Springer.

85 Güntert, P., Berndt, K.D., Wüthrich, K. (1993), *J. Biomol. NMR* **3**, 601–606.

86 Meadows, R.P., Olejniczak, E.T., Fesik, S.W. (1994), *J. Biomol. NMR* **4**, 79–96.

87 Duggan, B.M., Legge, G.B., Dyson, H.J., Wright, P.E. (2001), *J. Biomol. NMR* **19**, 321–329.

88 Mumenthaler, C., Braun, W. (1995), *J. Mol. Biol.* **254**, 465–480.

89 Nilges, M., Macias, M., O'Donoghue, S.I., Oschkinat, H. (1997), *J. Mol. Biol.* **269**, 408–422.

90 Linge, J.P., O'Donoghue, S.I., Nilges, M. (2001), *Methods Enzymol.* 339, 71–90.

91 Savarin, P., Zinn-Justin, S., Gilquin, B. (2001), *J. Biomol. NMR* **19**, 49–62.

92 Nilges, M. (1993), *Proteins* **17**, 297–309.

93 Greenfield, N.J., Huang, Y.J., Palm, T., Swapna, G.V.T., Monleon, D., Montelione, G.T., Hitchcock-DeGregori, S.E. (2001), *J. Mol. Biol.* **312**, 833–847.

94 Calzolai, L., Lysek, D.A., Güntert, P., von Schroetter, C., Riek, R., Zahn, R., Wüthrich, K. (2000), *Proc. Natl. Acad. Sci. USA* **97**, 8340–8345.

95 Ellgaard, L., Riek, R., Herrmann, T., Güntert, P., Braun, D., Helenius, A., Wüthrich, K. (2001), *Proc. Natl. Acad. Sci. USA* **98**, 3133–3138.

96 Horst, R., Damberger, F., Luginbühl, P., Güntert, P., Peng, G., Nikonova, L., Leal, W. S., Wüthrich, K. (2001), *Proc. Natl. Acad. Sci. USA* **98**, 14374–14379.

97 Miura, T., Klaus, W., Ross, A., Güntert, P., Senn, H. (2002), *J. Biomol. NMR*, **22**, 89–92.

3

Achieving Better Sensitivity, Less Noise and Fewer Artifacts in NMR Spectra

Detlef Moskau and Oliver Zerbe

3.1 Introduction

Both structure determination of large biomolecules by multidimensional NMR and screening techniques require high sensitivity and artifact-free spectra in order to be performed reliably at the lowest possible concentrations. Moreover, automated data analysis software still has not solved the problem of distinguishing artifacts from genuine signals satisfactorily. Hence, the best way to circumvent associated problems is to have as few artifacts or noise as possible. Substantial progress has been made in the last decade in the development of spectrometer hardware and probehead design. Improvements of the hardware have aimed at increasing the signal-to-noise ratio (S/N) and resolution while reducing the amount of spurious, unwanted signals.

In principle, unwanted signals may be due to artifacts or to noise. The main source of "real" noise is Brownian motion of electrons in the receiver coil. In "white noise", frequency components are statistically distributed. In contrast, artifacts are "wrong" signals occurring at well-defined frequencies. Instabilities in hardware may lead to noise and/or artifacts. Improvements in S/N will help to distinguish genuine peaks from noise [1]. Although it is sometimes possible to remove noise peaks by symmetrization routines [2, 3] or when the peak can be recognized as false because either the F1 or F2 frequency is impossible [e.g. in the INADEQUATE experiment, where $\nu(F1) = \nu_A(F2) + \nu_B(F2)$], datasets with insufficient S/N are mostly useless.

The term noise in one-dimensional NMR spectra is related to any noise source in the receiver path which determines the sensitivity or signal-to-noise ratio in the spectrum. In multidimensional NMR spectra, a random modulation of intensities of signals will lead to a continuous band of frequencies translating into t_1 noise bands in the indirect dimension. Although, strictly speaking, t_1 noise is not real noise but rather an artifact, it behaves similarly (no well-defined frequencies of signals), and we will hence use the common term t_1 noise. Another property of (white) noise it that is occurs statistically distributed in frequency and phase and therefore will not add up coherently. Instabilities in any of the system components (either electronics, magnet system or probehead) will lead to elevated noise levels or additional artifacts. Hence, any measure that results in less disturbance of the system will improve the overall performance by reducing noise and/or artifacts. The total noise in the receiver path is a sum of thermal noise, mainly stemming