

Structure calculation of biological macromolecules from NMR data

PETER GÜNTERT

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

1. INTRODUCTION	146
2. HISTORICAL OVERVIEW	148
3. PROTEIN STRUCTURES SOLVED BY NMR	156
4. NMR DATA FOR PROTEIN STRUCTURE CALCULATION	160
4.1 Nuclear Overhauser effects	161
4.2 Scalar coupling constants	166
4.3 Hydrogen bonds	170
4.4 Chemical shifts	170
4.5 Residual dipolar couplings	172
4.6 Other sources of conformational restraints	173
5. PRELIMINARIES OF A STRUCTURE CALCULATION	173
5.1 Systematic analysis of local conformation	173
5.2 Stereospecific assignments	175
5.3 Treatment of distance restraints to diastereotopic protons	177
5.4 Removal of irrelevant restraints	178
6. STRUCTURE CALCULATION ALGORITHMS	180
6.1 Metric matrix distance geometry	180
6.2 Variable target function method	182
6.3 Molecular dynamics in Cartesian space	184
6.4 Torsion angle dynamics	186
6.4.1 Tree structure of the molecule	187
6.4.2 Potential energy	188
6.4.3 Kinetic energy	190
6.4.4 Torsional accelerations	191
6.4.5 Integration of the equations of motion	192
6.4.6 Energy conservation and time step length	193
6.4.7 Simulated annealing schedule	194
6.4.8 Computation times	196
6.4.9 Application to biological macromolecules	197
6.5 Other algorithms	200

7.	STRUCTURE ANALYSIS	200
7.1	<i>Restraint violations</i>	200
7.2	<i>Atomic root-mean-square deviations</i>	201
7.3	<i>Torsion angle distributions</i>	205
7.4	<i>Hydrogen bonds</i>	205
7.5	<i>Molecular graphics</i>	206
7.6	<i>Check programs</i>	206
7.7	<i>A single, representative conformer</i>	208
8.	GENERAL ASPECTS OF NMR STRUCTURE CALCULATION	209
8.1	<i>Ensemble size</i>	209
8.2	<i>Different NOE calibrations</i>	211
8.3	<i>Completeness of the data set</i>	212
8.4	<i>Wrong restraints and their elimination</i>	214
9.	AUTOMATED ANALYSIS OF NOESY SPECTRA	216
9.1	<i>Chemical shift tolerance range</i>	216
9.2	<i>Semiautomatic methods</i>	219
9.3	<i>Ambiguous distance restraints</i>	219
9.4	<i>Iterative combination of NOE assignment and structure calculation</i>	220
10.	STRUCTURE REFINEMENT	223
10.1	<i>Restrained energy minimization</i>	223
10.2	<i>Molecular dynamics simulation</i>	224
10.3	<i>Time- or ensemble averaged restraints</i>	224
10.4	<i>Relaxation matrix refinement</i>	224
11.	ACKNOWLEDGEMENTS	225
12.	REFERENCES	225

I. INTRODUCTION

The relationship between amino acid sequence, three-dimensional structure and biological function of proteins is one of the most intensely pursued areas of molecular biology and biochemistry. In this context, the three-dimensional structure has a pivotal role, its knowledge being essential to understand the physical, chemical and biological properties of a protein (Branden & Tooze, 1991; Creighton, 1993). Until 1984 structural information at atomic resolution could only be determined by X-ray diffraction techniques with protein single crystals (Drenth, 1994). The introduction of nuclear magnetic resonance (NMR) spectroscopy (Aragam, 1961) as a technique for protein structure determination (Wüthrich, 1986) 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.

The NMR method for the study of molecular structures depends on the sensitive variation of the resonance frequency of a nuclear spin in an external magnetic field with the chemical structure, the conformation of the molecule, and the solvent environment (Ernst *et al.* 1987). The dispersion of these chemical

shifts ensures the necessary spectral resolution, although it usually does not provide direct structural information. Different chemical shifts arise because nuclei are shielded from the externally applied magnetic field to differing extent depending on their local environment. Three of the four most abundant elements in biological materials, hydrogen, carbon and nitrogen, have naturally occurring isotopes with nuclear spin $\frac{1}{2}$, and are therefore suitable for high-resolution NMR experiments in solution. The proton (^1H) has the highest natural abundance (99.98 %) and the highest sensitivity (due to its large gyromagnetic ratio) among these isotopes, and hence plays a central role in NMR experiments with biopolymers. Because of the low natural abundance and low relative sensitivity of ^{13}C and ^{15}N (1.11 % and 0.37 %, respectively) NMR spectroscopy with these nuclei normally requires isotope enrichment. This is routinely achieved by overexpression of proteins in isotope-labelled media. Structures of small proteins with molecular weight up to 10 kDa can be determined by homonuclear ^1H NMR. Heteronuclear NMR experiments with ^1H , ^{13}C and ^{15}N (Cavanagh *et al.* 1996) are indispensable for the structure determination of larger systems (e.g. Clore & Gronenborn, 1990; Edison *et al.* 1994).

Today many, if not most, NMR measurements with proteins are performed with the ultimate aim of determining their three-dimensional structure. However, NMR is not a 'microscope with atomic resolution' that would directly produce an image of a protein. Rather, it is able to yield a wealth of indirect structural information from which the three-dimensional structure can only be uncovered by extensive calculations. The pioneering first structure determinations of peptides and proteins in solution (e.g. Arseniev *et al.* 1984; Braun *et al.* 1981; Clore *et al.* 1986*b*; Williamson *et al.* 1985; Zuiderweg *et al.* 1984) were year-long struggles, both fascinating and tedious because of the lack of established NMR techniques and numerical methods for structure calculation, and hampered by limitations of the spectrometers and computers of the time. Recent experimental, theoretical and technological advances – and the dissemination of the methodological knowledge – have changed this situation decisively: Given a sufficient amount of a purified, water-soluble, monomeric protein with less than about 200 amino acid residues, its three-dimensional structure in solution can be determined routinely by the NMR method, following the procedure described in the classical textbook of Wüthrich (1986) and outlined in Fig. 1.

There is a close mutual interdependence, indicated by circular arrows in Fig. 1, between the collection of conformational restraints and the structure calculation, which forms the subject of this work. In its framework, structure calculation is the *de novo* computation of three-dimensional molecular structures on the basis of conformational restraints derived from NMR. Structure calculation is distinguished from structure refinement by the fact that no well-defined start conformation is used, whereas structure refinement aims at improving a given, well-defined structure with respect to certain features, for example its conformational energy.

After a historical outline of the development of NMR structure calculation methods in Section 2, and an overview of NMR structures deposited in the Protein

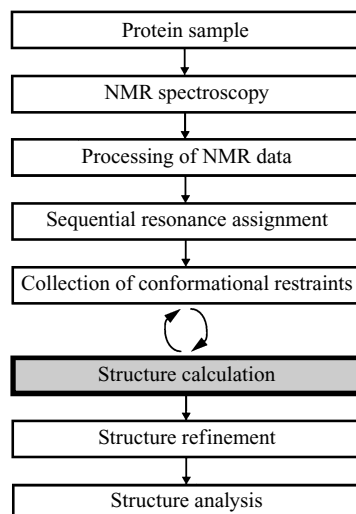


Fig. 1. Outline of the procedure for protein structure determination by NMR.

Data Bank in Section 3, the core part of the presentation starts in Section 4 with a discussion of various types of structurally relevant NMR data and their conversion into conformational restraints. Section 5 explains preliminary steps that precede a structure calculation. The central Section 6 is devoted to algorithms used for structure calculation. Special emphasis is given to molecular dynamics in torsion angle space, the currently most efficient method for biomolecular structure calculation. Measures to analyse the outcome of a structure calculation are introduced in Section 7. The relation between the conformational restraints in the input of a structure calculation and the quality of the resulting structure is discussed in Section 8. The combination of NOE assignment and structure calculation in automated procedures is introduced in Section 9. The text concludes with a glance at various structure refinement methods in Section 10.

2. HISTORICAL OVERVIEW

The aim of this section is to give a brief overview of the history of NMR structure calculation in the period from its beginning in the early 1980s until now. No attempt is made to cover the history of NMR spectroscopy in general, or of other aspects of the NMR method for biomolecular structure determination besides structure calculation, since a lavish account of this exciting story has been published recently in the opening volume of the *Encyclopedia of NMR* (Grant & Harris, 1996), together with an entertaining collection of personal reminiscences from the pioneers in the field. The new method was confronted initially with much scepticism and also utter disbelief, partly because the early solution structure determinations were done for systems for which the three-dimensional structure had already been known, or could be inferred from that of a homologous protein.

Suspensions could be allayed only when simultaneous but completely independent determinations of the three-dimensional structure of the protein tendamistat, for which no structural information was available before the project was started, by X-ray crystallography (Pflugrath *et al.* 1986) and by NMR (Kline *et al.* 1986, 1988) yielded virtually identical results (Billeter *et al.* 1989).

In the early development of the NMR method for protein structure determination it became clear that computer algorithms for structure calculation would be an indispensable prerequisite for solving the three-dimensional structures of objects as complex as a protein. It emerged that the key data measured by NMR would consist of a network of distance restraints between spatially proximate hydrogen atoms (Dubs *et al.* 1979; Kumar *et al.* 1980), for which existing techniques for structure determination from X-ray diffraction data would be inappropriate. Manual model building or interactive computer graphics could not provide solutions either because the intricacies of the distance restraint network precluded a manual analysis at atomic level, virtually restricting manual approaches to strongly simplified, cartoon-like representations of a protein (Zuiderweg *et al.* 1984). Hence new ways had to be developed.

The mathematical theory of distance geometry (Blumenthal, 1953) was the first method to be used for protein structure calculation. (Since distance geometry was first, NMR structure calculations were and are often termed 'distance geometry calculations', regardless of the principles underlying the algorithm used. Here, this practice is not followed, and the term is used only for algorithms based on distance space and the metric matrix.) The basic idea of distance geometry is to formulate the problem not in the Cartesian space of the atom positions but in the much higher 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. Algorithms for this purpose had been devised (Crippen, 1977; Crippen & Havel, 1978; Havel *et al.* 1979; Kuntz *et al.* 1979) already before their use in NMR protein structure determination could be envisioned, but the advent of NMR as a – however imprecise – microscopic ruler with which a large number of interatomic distances could be measured in a biological macromolecule spurred vigorous research in the field of distance geometry. For the first time a computer program was used to calculate the solution structure of a biological macromolecule on the basis of NOE measurements (Braun *et al.* 1981). The program, based on metric matrix distance geometry, was applied to a nonapeptide of 109 atoms. 23 distance restraints had been determined by NMR. Later, the same program was used for the first calculation of the NMR solution structure of a globular protein, a scorpion insectotoxin of 35 amino acid residues comprising both α -helical and β -sheet secondary structure (Arseniev *et al.* 1984). Presumably because of memory limitations, not all atoms of the protein could be treated explicitly. Instead, a simplified representation with two pseudoatoms per residue was used. Havel, Kuntz & Crippen (1983) provided an improved version of the original embedding algorithm, which was implemented in DISGEO (Havel & Wüthrich, 1984), the first complete program package for NMR

protein structure calculation. Calculations with simulated NMR data sets (Havel & Wüthrich, 1985), and a structure calculation of a protein on the basis of experimental NMR data (Williamson *et al.* 1985), both performed with DISGEO, made it clear that even very imprecise measurements of distances that are short compared with the size of a protein were sufficient to define the three-dimensional structure of a protein, provided that a sufficient number of such distance restraints was available. At the time this finding convincingly refuted a widespread argument against NMR protein structure determination, namely that short distance restraints could never consistently determine the relative orientation of parts of a molecule that are much further apart than the longest upper distance bound.

For a molecular system with N atoms, metric matrix distance geometry calls for storage of a matrix with N^2 elements, and the computation time is proportional to N^3 . Both requirements posed formidable challenges to the computer hardware in these early years of protein structure calculation. Even for a small protein like the basic pancreatic trypsin inhibitor (BPTI), with 58 amino acid residues and about 900 atoms, special devices had to be introduced to cut down the number of atoms in the embedding step such as performing the embedding on only a substructure. Nevertheless, the computation time for a single BPTI conformer was of the order of 10 hours on a DEC 10, then a state-of-the-art computer (Havel & Wüthrich, 1984). The DISGEO program was in use for several years, and it could have been expected that such practical problems would be alleviated by the steady advancement of computer technology. However, other, more fundamental problems were looming.

In the meantime algorithms based on very different ideas came into being. The problem of finding molecular conformations that are in agreement with certain geometrical restraints can always be formulated as one of minimization of a suitable objective or target function. The global minimum of the target function, or a close enough approximation of it, is sought, whereas local minima are to be avoided. The target function can be defined on different spaces. Metric matrix distance geometry took refuge from the local minimum problem in a very high-dimensional space, from which it could be difficult at times to come back to our three-dimensional world, not least because the notions of chirality or mirror images are unknown in distance space. Another approach went the opposite way by reducing the dimensionality of conformation space as far as possible. Recognizing that fluctuations of the covalent bond lengths and bond angles around their equilibrium values are small and fast, and cannot be measured by NMR, Braun & Go (1985) retained only the essential degrees of freedom of a macromolecule, namely the torsion angles. In this way, the number of degrees of freedom was reduced by about an order of magnitude compared with Cartesian coordinate space. Their variable target function method in torsion angle space (Braun & Go, 1985) used the method of conjugate gradients (Powell, 1977), a standard algorithm for the minimization of a multidimensional function. In the times of severely limited computer memories this algorithm had the advantage that no large matrices had to be stored. However, two problems had to be

overcome to enable its use in protein structure calculation. For efficient minimization it is essential to know not only the value of the target function but also its gradient, that is the partial derivatives with respect to the coordinates, the torsion angles in this case. At first the calculation of the gradient appeared to be very computation intensive. However, Abe *et al.* (1984) had removed this obstacle with their discovery of a fast recursive method to accomplish this task. The other, more daunting difficulty was the local minimum problem. Being a minimizer that takes exclusively downhill steps, the conjugate gradient algorithm is effective in locating a local minimum in the vicinity of the current conformation, but not as a method to search conformation space for the global minimum of the target function. Therefore, straightforward conjugate gradient minimization of a target function representing the complete network of NMR-derived restraints and the steric repulsion among all pairs of atoms in a protein was found to get stuck virtually always in local minima very far from the correct solution. The variable target function method, devised by Braun & Go (1985), and implemented in their program DISMAN, offered a partial answer to this question by going through a series of minimizations 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, where first, under the influence of short- and medium-range distance restraints, the helical segments are formed and subsequently, when the long-range restraints gradually come into play, positioned relative to each other. Not surprisingly, the variable target function method performed well for helical peptides but much less so for β -sheet proteins like tendamistat, where the fraction of acceptable conformers dropped to 9% (Kline *et al.* 1988); a situation that was calling for enhancements of the original variable target function idea. Reassuring, on the other hand, was the result obtained in the course of the solution structure determination of BPTI, where both algorithms, DISGEO and DISMAN, yielded essentially equivalent structure bundles, both in close agreement with the X-ray structures (Wagner *et al.* 1987).

In parallel with these developments, another powerful computing technique was recruited for protein structure calculation: molecular dynamics simulation. The method is based on classical mechanics and proceeds by numerically solving Newton's equation of motion in order to obtain a trajectory for the molecular system. The Cartesian coordinates of the atoms are the degrees of freedom. In the context of protein structure calculation the basic advantage of molecular dynamics simulation over minimization techniques is the presence of kinetic energy. It allows the system to escape from local minima that would be traps for minimizers bound to take exclusively downhill steps. By 1985, molecular dynamics simulation had existed already for more than two decades. Initially it had been used to simulate simple gases (Alder & Wainwright, 1959; Rahman, 1964; Verlet, 1967), but calculations with proteins had become feasible as well, starting with the first simulation of BPTI by McCammon *et al.* (1977). The first calculation of protein tertiary structure on the basis of NMR distance measurements by molecular dynamics simulation was performed by Kaptein *et al.* (1985) for the *lac* repressor

headpiece, using the program that was to become GROMOS (van Gunsteren & Berendsen, 1982). This was, however, not really a *de novo* structure calculation by molecular dynamics simulation because first ‘a molecular model was built using the three helices as building blocks [...] which, after measurement of the atomic co-ordinates, was subjected to refinement’ (Kaptein *et al.* 1985). Clore *et al.* (1985) used the molecular dynamics program CHARMM (Brooks *et al.* 1983) to compute the solution structure of a single helix of 17 amino acid residues, starting from three different initial conformations, an α -helix, a β -strand, and a 3_{10} -helix. The viability of restrained molecular dynamics simulation as a method for *de novo* structure calculation of complete globular proteins was demonstrated by Brünger *et al.* (1986), using simulated data for crambin, a small protein of 46 amino acid residues. Shortly thereafter, a method that has been in use for NMR structure calculation ever since was introduced and employed to calculate the globular structure of a protein with 45 amino acids (Clore *et al.* 1986*b*): the combination of metric matrix distance geometry to obtain a rough but correctly folded structure followed by restrained energy minimization and molecular dynamics refinement.

So far, these molecular dynamics approaches had relied on a full empirical force field (Brooks *et al.* 1983) to ensure proper stereochemistry, and were generally run at a constant temperature, close to room temperature. Substantial amounts of computation time were required because the empirical energy function included long-range pair interactions that were time-consuming to evaluate, and because conformation space was explored slowly at room temperature. Both features had been inherited from molecular dynamics programs created with the aim of simulating the time evolution of a molecular system as realistically as possible in order to extract from the complete trajectories molecular quantities of interest. When these algorithms are used for structure calculations, however, the objective is quite different. Here, they simply provide a means to efficiently optimize a target function that takes the role of the potential energy. The course of the trajectory is unimportant, as long as its end point comes close to the global minimum of the target function. Therefore, the efficiency of a structure calculation by molecular dynamics 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 the number of integration steps needed) the trajectory by which it can be reached from the start conformation. Based on this observation new ingredients to the method made the folding process much more efficient (Nilges *et al.* 1988*a*, 1991): a simplified ‘geometric’ energy function, a modified potential for NOE restraints with asymptotically linear slope for large violations, and simulated annealing. The geometric force field retained only the most important part of the non-bonded interaction by a simple repulsive potential that replaced the Lennard-Jones and electrostatic interactions in the full empirical energy function. This short-range repulsive function could be calculated much faster, and it significantly facilitated the large-scale conformational changes required during the folding process by lowering energy barriers induced by the overlap of atoms. A similar effect could be expected from replacing the originally

quadratic distance restraining potential by a function that was dominated less by the most strongly violated restraints. The most decisive new concept was, however, the amalgamation of molecular dynamics with simulated annealing, an optimization method derived from concepts in statistical mechanics (Kirkpatrick *et al.* 1983). Simulated annealing mimics on the computer the annealing process by which a solid attains its minimum energy configuration through slow cooling after having been heated up to high temperature at the outset. Simulated annealing uses a target function, the 'energy', and requires a mechanism to generate Boltzmann ensembles at each temperature $T_1 > T_2 > \dots > T_n$ of the annealing schedule. In the case of protein structure calculation, molecular dynamics is the method of choice to generate the Boltzmann ensemble because it restricts conformational changes to physically reasonable pathways, while the inertia of the system enables transitions over barriers up to a height that is controlled by the temperature. Monte Carlo (Metropolis *et al.* 1953), the other familiar technique to create a Boltzmann distribution, relies on random conformational changes that are accepted or rejected randomly with a probability that depends on the energy change incurred by the move. Monte Carlo has never become popular in the field of protein structure calculation because it is extremely difficult to devise schemes for choosing 'random' moves that are not either physically unreasonable (i.e. leading to a large increase of the energy and, hence, almost certain rejection) or too small for efficient exploration of conformation space. Three different protocols for simulated annealing by molecular dynamics, each using a different way to produce the start structure for the molecular dynamics run, were established: 'Hybrid distance geometry-dynamical simulated annealing' (Nilges *et al.* 1988*a*) used a start conformation obtained from metric matrix distance geometry, the second method started from an extended polypeptide chain (Nilges *et al.* 1988*c*), and the third from a random array of atoms (Nilges *et al.* 1988*b*). Obviously, from the first to the third method the simulated annealing protocols had to cope with progressively less realistic start conformations. From a theoretical point of view it was an impressive demonstration of the power of simulated annealing by molecular dynamics that a correctly folded protein could result starting from a cloud of randomly placed atoms. In practice, however, the combination of substructure embedding by distance geometry and simulated annealing by molecular dynamics became most popular because its – still considerable – demand on computation time was much lower than for the other protocols. Together with these protocols, a new molecular dynamics program entered the stage. XPLOR (Brünger, 1992) drew on the molecular dynamics simulation package CHARMM (Brooks *et al.* 1983), but was written especially with the aim of structure calculation and refinement in mind. Being a versatile tool for biomolecular structure determination by NMR and X-ray diffraction, it soon gained, and maintained ever since, high popularity. The original protocols by Nilges *et al.* (1988*a-c*) were improved (Brünger, 1992), and a metric matrix distance geometry module was incorporated into XPLOR (Kuszewski *et al.* 1992).

The success of the hybrid distance geometry-simulated annealing technique

brought about a gradual change in the way metric matrix distance geometry was used. Rather than being employed as a self-contained method for complete structure calculation, it became more and more a device to efficiently build a crude, but globally correctly folded start conformation for subsequent simulated annealing. Times were troubled for metric matrix distance geometry temporarily by a problem that had been noticed already in the first comparison with another structure calculation method (Wagner *et al.* 1987): the possibility of insufficient sampling of conformation space. Since the beginning of the NMR method for biomolecular structure determination, the precision with which the experimental data defined the structure had been estimated by the spread among a group of conformers calculated from the same input data by the same computational protocol but starting from different, randomly chosen initial conditions. The NMR solution structure of a protein was hence represented by a bundle of equivalent conformers, each of which proffering an equally good fit to the data, rather than by a single set of coordinates. This approach was in line with the fact that the experimental measurements were not interpreted as yielding a single best value for, say, an interatomic distance but an allowed range within that no particular value should be favoured *a priori* over another. Obviously, this method would picture faithfully the real situation only if the algorithm used performed a uniform sampling of the conformation space that is accessible to the molecule subjected to a set of experimental restraints, yielding at least a coarse approximation to a Boltzmann ensemble. There had been early indications that this was not the case for certain implementations of metric matrix distance geometry (Wagner *et al.* 1987), especially in regions not or hardly restrained by experimental data, where structures tended to be clustered and artificially expanded as if a mysterious force was to drive them away from the centre of the molecule. This problem was most clearly exposed by calculations made without any experimental distance restraints (Metzler *et al.* 1989; Havel, 1990), and vigorous and ultimately successful research set in to discover the cause of the problem and to offer possible solutions to it. Distance geometry algorithms compute the metric matrix, with elements $G_{ij} = \mathbf{r}_i \cdot \mathbf{r}_j$, from the complete distance matrix, with elements $D_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$. But only a tiny fraction of these distances are given by the covalent structure or restrained by experimental data. Distances for which the exact value is not known, have to be selected 'randomly' between their lower and upper bound. It was discovered soon (Havel, 1990) that details of how the missing distances were chosen had paramount influence on the sampling properties, and that the commonly used, straightforward strategy for selecting distances (Havel & Wüthrich, 1984) was responsible for the artificial expansion and spurious clustering of structures because it tended to produce distance values that were too long. A 'metrization' procedure had been proposed (Havel & Wüthrich, 1984) to cull them in accord, such that the triangle inequality $D_{ik} \leq D_{ij} + D_{jk}$ was fulfilled for all triples (i, j, k) of atoms, albeit at the cost of considerably increased computation time. However, the implementation in the DISGEO program still induced a bias, and a solution to the sampling problem came with improved 'random metrization' procedures (Havel, 1990) that were

implemented in a new program package, DG-II (Havel, 1991). The inconvenience of long computation times could be alleviated by the partial metrization algorithm of Kuszewski *et al.* (1992) without deteriorating the sampling properties.

In contrast to the early implementations of metric matrix distance geometry, the variable target function method, into which randomness entered through completely randomized start conformations in torsion angle space, was not beset by sampling problems but had the drawback that for all but the most simple molecular topologies only a small percentage of the calculations converged to solutions with small restraint violations. A new implementation of the variable target function method in the program DIANA (Güntert *et al.*, 1991*a*) initially offered a symptomatic therapy to the problem by dramatically reducing the computation time needed to carry out the variable target function minimization for a conformer, but later also a cure of the causes of the problem by the usage of redundant torsion angle restraints (Güntert & Wüthrich, 1991). In this iterative procedure redundant restraints were generated on the basis of the torsion angle values found in a previous round of structure calculations.

It had been obvious for a long time that a method working in torsion angle space and using simulated annealing by molecular dynamics could benefit from the advantages of both approaches but it seemed very difficult to implement an algorithm for molecular dynamics with torsion angles as degrees of freedom. Provided that an efficient implementation could have been found, such a 'torsion angle dynamics' algorithm would have been more efficient than conventional molecular dynamics in Cartesian coordinate space, simply because the absence of high-frequency bond length and bond angle vibrations would have allowed for much longer integration time steps or higher temperatures during the simulated annealing schedule. An algorithm for Langevin-type dynamics (neglecting inertial terms) of biopolymers in torsion angle space had been presented already by Katz *et al.* (1979), and the authors stated laconically without further elaborating on the point that for the full equations of motion including inertial terms 'all constituents [...] and its derivatives are calculated when the matrix elements of the Hessian [i.e. the second derivatives of the potential energy with respect to torsion angles] are evaluated. Thus it is a trivial matter to assemble these.' More than a decade later, Mazur & Abagyan (1989) derived explicit formulas for Lagrange's equations of motion of a polymer, using internal coordinates as degrees of freedom. Calculations for a poly-alanine peptide of nine residues using the CHARMM force field demonstrated that time steps of 13 fs – an order of magnitude longer than in standard molecular dynamics simulation based on Newton's equations of motion in Cartesian space – were feasible when torsion angles were the only degrees of freedom (Mazur *et al.* 1991). Nevertheless, in practical applications with larger proteins the algorithm would have been much slower than a standard molecular dynamics simulation in Cartesian space because in every integration time step a system of linear equations had to be solved with a computational effort proportional to the third power of the number of torsion angles. Solutions to this problem were found in other branches of science where questions of simulating the dynamics of complex multibody systems such as robots, spacecraft and

vehicles were pondered. Independently, Bae & Haug (1987) and Jain *et al.* (1993) found torsion angle dynamics algorithms whose computational effort scaled linearly with the system size, as in Cartesian space molecular dynamics. The advantage of longer integration time steps in torsion angle dynamics could be exploited for systems of any size with these algorithms. Both algorithms have been used for protein structure calculation on the basis of NMR data, one (Bae & Haug, 1987) in the program XPLOR (Stein *et al.* 1997), the other (Jain *et al.* 1993) in the program DYANA (Güntert *et al.* 1997). Experience with both programs indeed confirmed expectations that torsion angle dynamics constituted the most efficient way to calculate NMR structures of biomacromolecules, but showed as well that the computation time with DYANA is about one order of magnitude shorter than with XPLOR (Güntert *et al.* 1997).

With this, the history of NMR structure calculation has reached the present but certainly not its end. Inevitably, writing a succinct account of this story solicited many subjective decisions, to skip important contributions, and not to follow numerous original side lines. The impulse to solve ever larger biomolecular structures, the strive for automation of NMR structure determination, and the advent, for the first time since the method was introduced, of a new class of generally applicable conformational restraints (Tjandra & Bax, 1997), will confront structure calculation with new challenges and offer renewed chances for success.

3. PROTEIN STRUCTURES SOLVED BY NMR

The increasing importance of NMR as a method for structure determination of biological macromolecules is manifested in the steadily rising number of NMR structures that are deposited in the Protein Data Bank (PDB; Bernstein *et al.* 1977). In December 1997, there were a total of 1059 (or 1008, if duplicate entries for the same protein are excluded) files available from the PDB with Cartesian coordinates of proteins, nucleic acids and macromolecular complexes that have been obtained by NMR techniques.

The development of NMR structure determination since 1988, when the first two NMR structures entered the PDB, is summarized in Fig. 2. The number of NMR structures in the PDB has increased at a faster rate than the total number of coordinate files in the PDB, resulting in a continuous increase of the percentage of NMR structures among all PDB structures. In December 1997 NMR structures comprised 16% of all coordinate files in the PDB. The average size of unique NMR structures in the PDB has also increased, albeit at a slow rate, reaching 8.65 kDa in December 1997. The size distribution of unique NMR structures in the PDB, shown in Fig. 3, indicates that structures of small proteins with a molecular weight below 15 kDa are solved routinely, whereas structure determinations for systems above 20 kDa are still rare.

Since 1990, it was possible also to submit files to the PDB containing experimental data that was used in the structure calculation. Typically, these files include the distance and torsion angle restraints used in the final round of

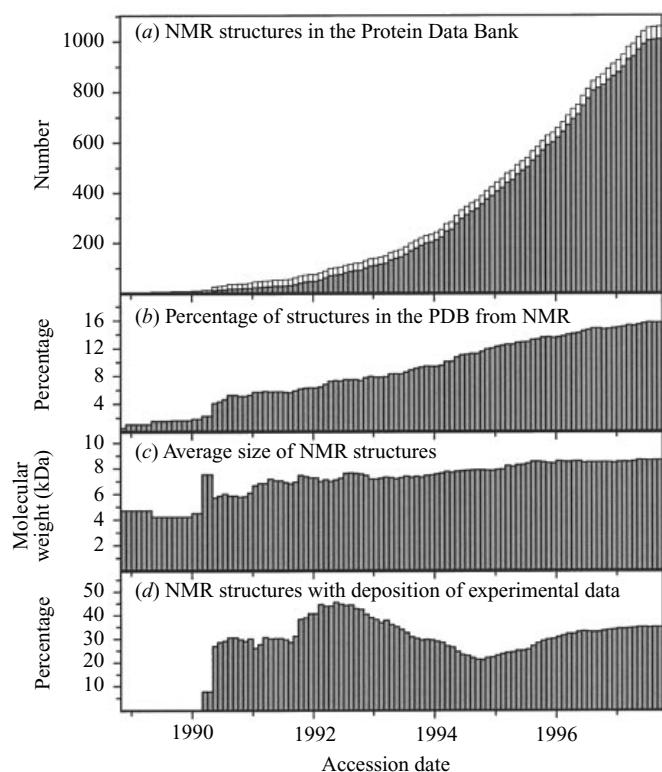


Fig. 2. NMR structures in the Protein Data Bank (PDB; Bernstein *et al.* 1977) until December 1997. (a) Number of coordinate entries in the PDB that were derived from NMR data plotted versus the accession date. White bars show all NMR structures, and shaded bars indicate all unique NMR structures that have been deposited with the PDB until a given date. (b) Percentage of all coordinate files in the PDB that represent NMR structures until a given date. (c) Average molecular weight of all unique NMR structures that have been deposited until a given date. (d) Percentage of NMR structures for which experimental NMR data have been submitted until a given date. Labels on the horizontal axis indicate the beginning of a year. Definitions: *NMR structure*: Coordinate file with the word 'NMR' in the EXPDTA record. *Accession date*: The date given in the HEADER record. *Unique NMR structure*: If there are several NMR structures with PDB codes that differ only in the first digit, only one of them is retained. (This happens, for example, if a bundle of conformers and a minimized mean structure were submitted for the same protein.) *Molecular weight*: Sum of atomic masses of all atoms listed in ATOM records and for which coordinates are available.

structure calculations. Although these data can be essential to judge the quality of a structure determination by NMR, only a minority of the PDB coordinate files derived from NMR measurements are accompanied by a file with experimental data. There was no clear trend towards a higher percentage of NMR structures with corresponding experimental data files during the period 1990–97. In December 1997 experimental data were available for only 35% of the NMR structures in the PDB, a lower percentage than in 1992.

The large majority (82%) of NMR structures for which data have been

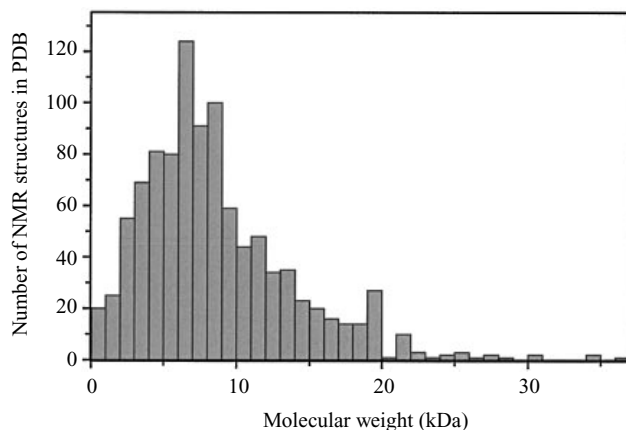


Fig. 3. Size distribution of the 1008 unique NMR structures in the Protein Data Bank in December 1997. The molecular weight is the sum of atomic masses of all atoms in the protein or nucleic acid for which coordinates are available.

Table 1. Journals that have published NMR structures available from the Protein Data Bank^a

Journal	Structures
<i>Biochemistry</i>	191
<i>Journal of Molecular Biology</i>	121
<i>Nature Structural Biology</i>	62
<i>Structure</i>	42
<i>Science</i>	33
<i>Protein Science</i>	23
<i>European Journal of Biochemistry</i>	21
<i>Nature</i>	20
<i>Other journals</i>	109

^a The information was taken from the JRNAL REF records of all unique coordinate files with NMR structures that were available from the Protein Data Bank in December 1997. About one third of these PDB coordinate files could not be considered because no precise reference is given (e.g. 'to be published').

deposited in the PDB, and for which a precise reference is given in the PDB coordinate file, have been published in only eight journals with 20 or more structures in each of them (Table 1). *Biochemistry* and the *Journal of Molecular Biology* with 191 and 121 structures, respectively, are the most popular places for the publication of NMR structures that are available from the PDB. This statistics may of course also reflect different coordinate deposition policies, and the extent to which these are enforced. Anyway, since not the text or figures of a paper but the Cartesian coordinates of the atoms constitute the main result of a structure determination, the value of structures that are not freely available to the scientific community is limited.

The wide-spread dissemination of the methodology of macromolecular

Table 2. Structure calculation programs

Program ^a	Structures ^b	Reference
Metric matrix distance geometry		
DG-II	86	Havel (1991)
DISGEO	22	Havel & Wüthrich (1984)
DSPACE	27	Biosym, Inc.
EMBOSS	19	Nakai <i>et al.</i> (1993)
TINKER	3	Hodsdon <i>et al.</i> (1996)
Variable target function method		
DIANA	124	Güntert <i>et al.</i> (1991a)
DISMAN	17	Braun & Go (1985)
Cartesian space molecular dynamics		
AMBER	135	Pearlman <i>et al.</i> (1991)
CHARMM	88	Brooks <i>et al.</i> (1983)
DISCOVER	99	Molecular Simulations, Inc.
GROMOS	22	van Gunsteren <i>et al.</i> (1996)
SYBYL	6	Tripos, Inc.
XPLOR	570	Brünger (1992)
Torsion angle dynamics		
DYANA	5	Güntert <i>et al.</i> (1997)

^a Programs that are specified in the 'PROGRAM' entry of more than one unique NMR structure coordinate file available from the Protein Data Bank in December 1997, excluding those used exclusively for relaxation matrix refinement or energy refinement of structures that have been calculated with another program. Also excluded are programs that have been used only for peptides of less than 20 amino acids. Each program is listed only once; if a program offers different structure calculation algorithms (e.g. DYANA or XPLOR), it is listed under the method for which it is most commonly used. Some of the programs, e.g. DISGEO, DSPACE and DISMAN, are virtually out of use today.

^b Number of unique NMR structure coordinate files available from the Protein Data Bank in December 1997 that mention the name of the program anywhere in their text. According to this simple criterion many structures are counted for which the molecular dynamics simulation programs AMBER, CHARMM, DISCOVER, GROMOS and SYBYL have been used not for the actual structure calculation but only for refinement purposes, or that contain merely a reference to the corresponding force field. Note also that for many structure determinations hybrid methods employing more than one program have been used.

structure determination by NMR within about a dozen years is probably best illustrated by the fact that by December 1997 a total of 1850 different persons have become co-authors of an NMR structure in the PDB, 40 of which have contributed to ten or more unique NMR structures in the PDB. The field is thus no longer exclusively 'in the hands' of the limited number of specialists who have developed the technique.

An attempt to classify the NMR structures in the PDB according to the program used in the structure calculation has been made in Table 2, although this statistics is beset with many uncertainties because the Protein Data Bank does not use a consistent format for information about the structure calculation. In

particular, it is in general not possible to determine in an automatic search whether a program has been used for the actual structure calculation or only for a subsequent energy refinement. With very few exceptions, structure calculation programs can be assigned to just four different types of algorithms (some programs offer several of these simultaneously): metric matrix distance geometry, variable target function method in torsion angle space, molecular dynamics simulation in Cartesian space, and molecular dynamics simulation in torsion angle space.

4. NMR DATA FOR PROTEIN STRUCTURE CALCULATION

For use in a structure calculation, geometric conformational restraints have to be derived from suitable, conformation-dependent NMR parameters. These geometric restraints 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 restraints than NMR data for which the conformation dependence was deduced merely from statistical analyses of known structures. Advances in the theoretical treatment of biological systems can lead to better physical understanding and predictability of an NMR parameter such as the chemical shift that allows to put its structural interpretation – formerly deduced from empirical statistics (Spera & Bax, 1991) – on a firmer physical basis (de Dios *et al.* 1993).

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 – as well as by the steric repulsion between non-bonded 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 because there are no degrees of freedom that affect covalent structure parameters. Usually a simple geometric force field is used for the structure calculation that retains only the most dominant part of the non-bonded 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* (Güntert *et al.* 1997): 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, 1.6 Å for sulphur and phosphorus atoms (Braun & Go, 1985). 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. Depending on the structure calculation program used, special covalent bonds such as disulphide bridges or cyclic peptide bonds have to be enforced by distance restraints. Disulphide bridges may be fixed by restraining the distance between the two sulphur atoms to 2.0–2.1 Å and the two distances between the C^β and the sulphur atoms of different residues to 3.0–3.1 Å (Williamson *et al.* 1985).

4.1 Nuclear Overhauser effects

The NMR method for protein structure determination relies on a dense network of distance restraints derived from nuclear Overhauser effects (NOEs) between nearby hydrogen atoms in the protein (Wüthrich, 1986). NOEs are the essential NMR data to define the secondary and tertiary structure of a protein because they 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 NOE reflects the transfer of magnetization between spins coupled by the dipole–dipole interaction in a molecule that undergoes Brownian motion in a liquid (Solomon, 1955; Macura & Ernst, 1980; Neuhaus & Williamson, 1989). The intensity of a NOE, i.e. the volume V of the corresponding cross peak in a NOESY spectrum (Jeener *et al.* 1979; Kumar *et al.* 1980; Macura & Ernst, 1980), 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 $f(\tau_c)$ that 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, it is generally assumed 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 (Wüthrich, 1986). More sophisticated treatments that take into account that the result of a NOESY experiment represents an average over time and space are usually deferred to the structure refinement stage (Torda *et al.* 1989, 1990).

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’, i.e. indirectly *via* other spins in the vicinity (Kalk & Berendsen, 1976; Macura & Ernst, 1980). The approximation of isolated spin pairs is valid only for very short mixing times in the NOESY experiment. However, the mixing time cannot be made arbitrarily short because (in the limit of short mixing times) the intensity of a NOE is proportional to the mixing time (Kumar *et al.* 1980). In practice, a compromise has to be made

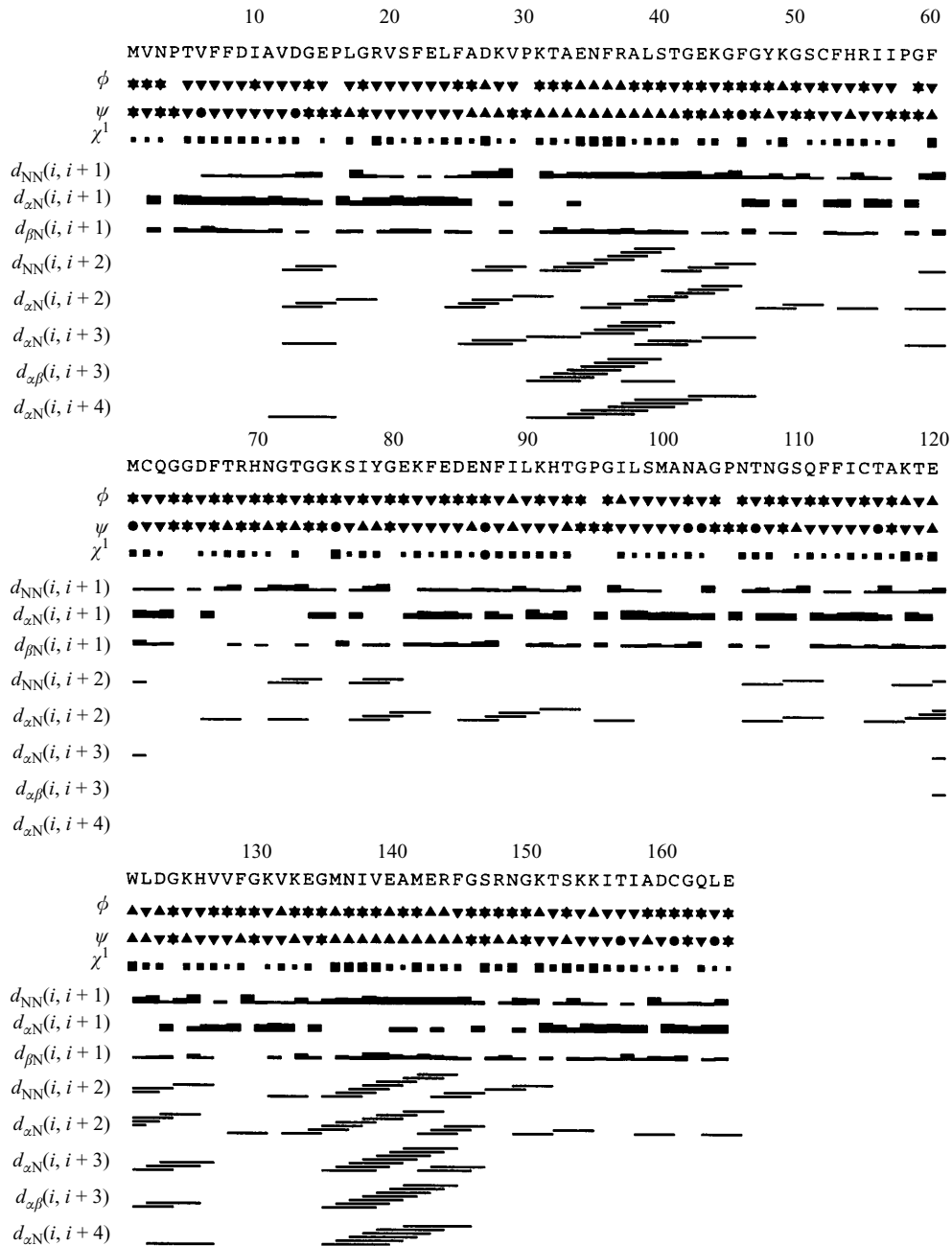


Fig. 4. Short- and medium-range restraints in the experimental NMR data set for the protein cyclophilin A (Ottiger *et al.* 1997). The first three lines below the amino acid sequence represent torsion angle restraints for the backbone torsion angles ϕ and ψ , and for the side-chain torsion angle χ^1 . For ϕ and ψ a triangle pointing upwards indicates a restraint that allows the torsion angle to take the values observed in an ideal α -helix ($\phi = -57^\circ$, $\psi = -47^\circ$) or 3_{10} -helix ($\phi = -60^\circ$, $\psi = -30^\circ$); a triangle pointing downwards indicates compatibility with an ideal parallel or antiparallel β -strand ($\phi = -119^\circ$, $\psi = -113^\circ$, or $\phi = -139^\circ$, $\psi = -135^\circ$, respectively; Schultz & Schirmer, 1979); a restraint represented by a

between the suppression of spin diffusion and sufficient cross peak intensities, usually with mixing times in the range of 40–80 ms for high-quality structures. Spin diffusion effects can be included in the structure calculation by complete relaxation matrix refinement (Keepers & James, 1984; Yip & Case, 1989; Mertz *et al.* 1991). 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 (Wijmenga *et al.* 1993; Pardi, 1995; Varani *et al.* 1996).

The quantification of an NOE amounts to determining the volume of the corresponding cross peak in the NOESY spectrum (Ernst *et al.* 1987). Since the line-widths can vary appreciably for different resonances, cross peak volumes should be determined by integration over the peak area rather than by measuring peak heights, for example by counting contour lines. Integration is straightforward for isolated cross peaks. For clusters of overlapping cross peaks deconvolution methods have been proposed to distribute the total volume among the individual signals (e.g. Denk *et al.* 1986; Koradi *et al.* 1998). While the reliable quantification of NOEs is important to obtain a high-quality protein structure, one should also keep in mind that, according to equation (1), the relative error of the distance estimate is only one sixth of the relative error of the volume determination.

On the basis of equation (1), NOEs are usually treated as upper bounds on interatomic distances rather than as precise distance restraints because the presence of internal motions and, possibly, chemical exchange may diminish the strength of an NOE (Ernst *et al.* 1987). 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 restraints in conjunction with the observation that internal motions and exchange effects usually reduce rather than increase the NOEs (Wüthrich, 1986). For the same reason, the absence of a NOE is in general not interpreted as a lower bound on the distance between the two interacting spins. Certain NOEs, however, may also be enhanced by internal motions or chemical exchange and then be incompatible with the assumption of a rigid structure that fulfils all NMR data simultaneously (Torda *et al.* 1990; Brüschweiler *et al.* 1991).

star encloses conformations of both α and β secondary structure types; and a filled circle marks a restraint that excludes the torsion angle values of these regular secondary structure elements. Torsion angle restraints for χ^1 are depicted by filled squares of three different decreasing sizes, depending on whether they allow for none, one, two or all three of the staggered rotamer positions $\chi^1 = -60, 60, 180^\circ$. Torsion angle restraints for χ^1 that exclude all three staggered rotamer positions are shown as filled circles. Upper distance limits for sequential and medium-range distances are shown by horizontal lines connecting the positions of the two residues involved. The thickness of the lines for the sequential distances $d_{\text{NN}}(i, i+1)$, $d_{\text{zN}}(i, i+1)$ and $d_{\text{pN}}(i, i+1)$ is inversely proportional to the squared upper distance bound. The plot was produced with the program DYANA (Güntert *et al.* 1997).

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

$$V = \frac{k}{u^6} \quad (2)$$

with a constant k that depends on the arbitrary scaling of the NOESY spectrum. The value u obtained from equation (2) 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 u 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 (Williamson *et al.* 1985; Clore *et al.* 1986*b*).

The constant k in equation 2 can be determined on the basis of known distances, for example the sequential distances $d(\text{H}_i^{\alpha}, \text{H}_{i+1}^{\text{N}})$ and $d(\text{H}_i^{\text{N}}, \text{H}_{i+1}^{\text{N}})$ in regular secondary structure elements (Billeter *et al.* 1982), or by reference to a preliminary structure (Güntert *et al.* 1991*b*). Sometimes, especially in the course of an automatic NOESY assignment procedure, it is convenient to get an estimate of k independent from the knowledge of certain distances or preliminary structures. This can be obtained, based on the observation that the average value \bar{u} of the upper distance bounds u for NOEs among the backbone and β protons is similar in all globular proteins, by setting k such that the average upper distance bound becomes $\bar{u} = 3.4$ Å (Mumenthaler *et al.* 1997).

In practice, it has been observed that more conservative calibration curves, for example of the type $V = k/u^n$, with $n = 4$ or 5 , may be advantageous for NOEs with peripheral side-chain protons (Güntert *et al.* 1991*b*). The uniform average model (Braun *et al.* 1981) provides another, very conservative, calibration curve by making the assumption that, due to internal motions, the interatomic distance, r , assumes all values between the steric lower limit, l , and an upper limit, u , with equal probability:

$$V = \frac{k}{u-l} \int_l^u \frac{dr}{r^6} = \frac{k'}{u-l} \left(\frac{1}{l^5} - \frac{1}{u^5} \right). \quad (3)$$

NOEs that involve groups of protons with degenerate chemical shifts, in particular methyl groups, are commonly referred to pseudoatoms located in the centre of the protons that they represent, and the upper bound is increased by a pseudoatom correction equal to the proton-pseudoatom distance (Wüthrich *et al.* 1983).

Sometimes, especially in the case of nucleic acid structure determination, where the standard, conservative interpretation of NOEs might not be sufficient to obtain a well-defined structure, also lower distance limits have been attributed to NOEs, either based on the intensity of the NOE or to reflect the absence of a corresponding cross peak in the NOESY spectrum (Pardi, 1995; Varani *et al.* 1996). Such practices should be exercised with care because they may carry the danger of overinterpretation of the experimental data and potentially jeopardize the robustness of the NMR approach to structure determination.

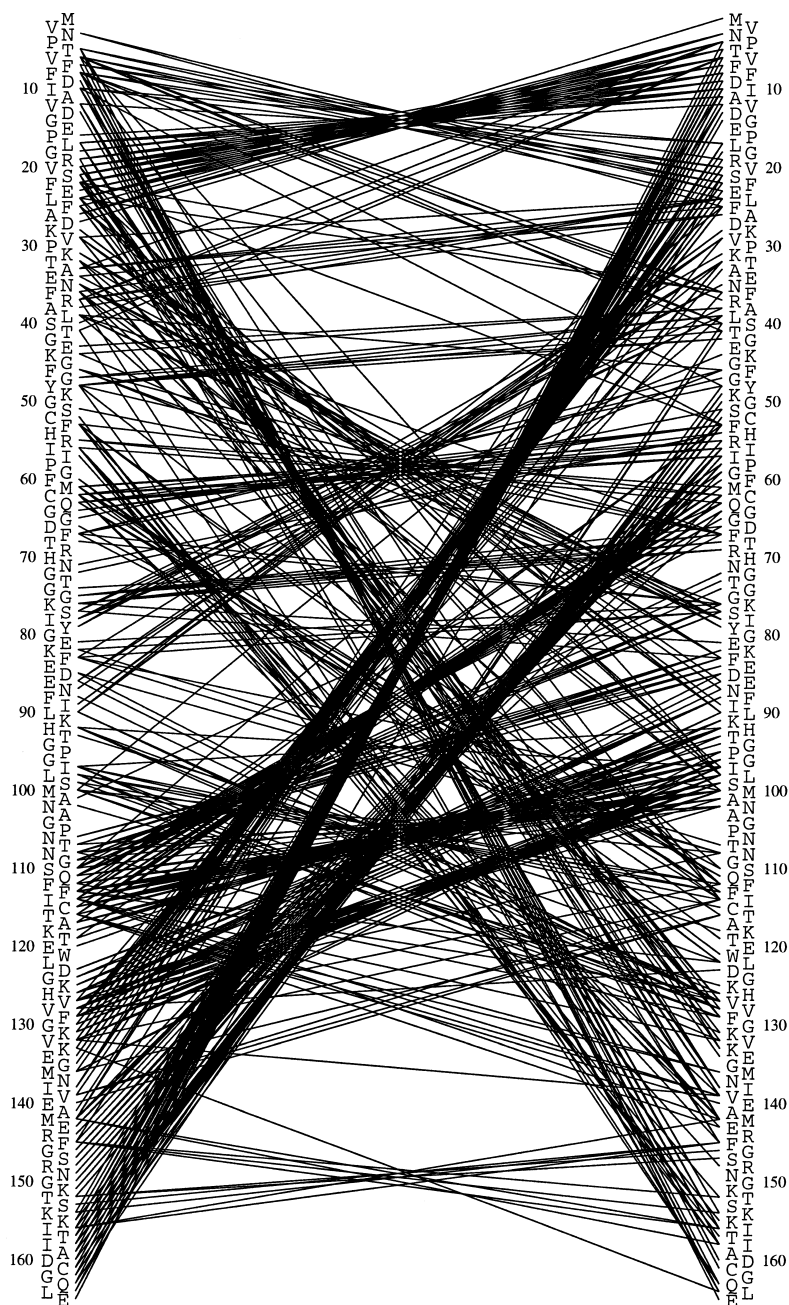


Fig. 5. Long-range distance restraints in the experimental NMR data set for the protein cyclophilin A (Ottiger *et al.* 1997). Restraints between atoms five or more residues apart in the sequence are represented by lines going from upper left to lower right (restraints between side-chain atoms), or from lower left to upper right (restraints involving backbone atoms). On the left and right hand sides the amino acid sequence of cyclophilin A is given.

Distance restraints can be visualized in a number of different ways. Short- and medium-range restraints are best plotted against the sequence, as in Fig. 4 which was produced automatically by the program DYANA on the basis of the experimental

Table 3. Karplus relations, ${}^3\mathcal{J}(\theta) = A \cos^2 \theta + B \cos \theta + C$, for proteins between a vicinal scalar coupling constant ${}^3\mathcal{J}$ and the corresponding torsion angle θ defined by the three covalent bonds between the two scalar coupled atoms

Angle	Coupling	A (Hz)	B (Hz)	C (Hz)	Offset ^a (degrees)	Reference
ϕ	H ^N -H ^{α}	6.98	-1.38	1.72	-60	Wang & Bax (1996)
	H ^N -C'	4.32	0.84	0.00	180	Wang & Bax (1996)
	H ^N -C ^{β}	3.39	-0.94	0.07	60	Wang & Bax (1996)
	C' _{<i>i-1</i>} -H ^{α}	3.75	2.19	1.28	120	Wang & Bax (1996)
	C' _{<i>i-1</i>} -C ^{β}	1.59	-0.67	0.27	-120	Hu & Bax (1997)
ψ	H ^{α} -N _{<i>i+1</i>}	-0.88	-0.61	-0.27	-120	Wang & Bax (1995)
χ^1	H ^{α} -H ^{β}	9.50	-1.60	1.80	-120/0	de Marco <i>et al.</i> (1978a)
	N-H ^{β}	-4.40	1.20	0.10	120/-120	de Marco <i>et al.</i> (1978b)
	C'-H ^{β}	7.20	-2.04	0.60	0/120	Fischman <i>et al.</i> (1980)

^a Difference between θ and the standard torsion angle ϕ , ψ or χ^1 . In the case of β -methylene protons the first number is for H ^{$\beta 2$} , the second for H ^{$\beta 3$} .

NMR data set for cyclophilin A (Ottiger *et al.* 1997). It is possible to identify secondary structure elements, especially helices, from characteristic patterns in such plots (Wüthrich, 1986). The distribution of long-range distance restraints can be shown schematically as in Fig. 5, or overlaid over the structure as in Fig. 21d below.

4.2 Scalar coupling constants

Vicinal scalar coupling constants, ${}^3\mathcal{J}$, between atoms separated by three covalent bonds from each other are related to the enclosed torsion angle, θ , by Karplus relations (Karplus, 1963):

$${}^3\mathcal{J}(\theta) = A \cos^2 \theta + B \cos \theta + C. \quad (4)$$

The parameters A , B and C have been determined for various types of couplings by a best fit of the measured \mathcal{J} values to the corresponding values calculated with equation (4) for known protein structures. The most commonly used Karplus relations in proteins are given in Table 3 and illustrated in Fig. 6.

In contrast to NOEs, scalar coupling constants give information only on the local conformation of a polypeptide chain. They are nevertheless important to accurately define the local conformation, to obtain stereospecific assignments for diastereotopic protons (usually for the β protons), and to detect torsion angles (usually χ^1) that occur in multiple states.

Scalar couplings are manifested in the cross peak fine structures of most NMR spectra (Ernst *et al.* 1987). Many NMR experiments have been proposed for the measurement of scalar coupling constants (Kessler *et al.* 1988; Biamonti *et al.* 1994; Case *et al.* 1994; Vuister *et al.* 1994; Cavanagh *et al.* 1996). Scalar coupling constants are conventionally measured from the separation of fine-structure components in anti-phase spectra. One has to be aware, however, of the

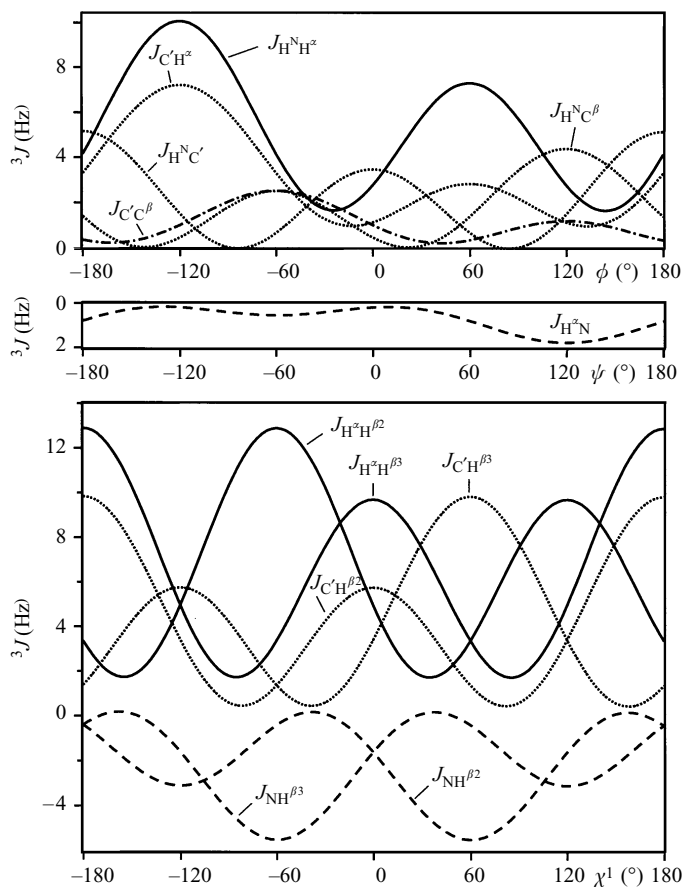


Fig. 6. Karplus relations between vicinal scalar coupling constants and the torsion angles ϕ , ψ and χ^1 in proteins. Karplus curves are drawn as solid lines for couplings between two hydrogen atoms, as dotted lines for couplings between a carbon and a hydrogen atom, as dot-dashed lines for couplings between two carbon atoms, and as dashed lines for couplings between a nitrogen and a hydrogen atom. See also Table 3.

cancellation effects between positive and negative fine-structure elements that lead both to an overestimation of the coupling constant and to a decrease of the overall cross peak intensity (Neuhaus *et al.* 1985; Cavanagh *et al.* 1996, pp. 315–320). These effects inhibit the determination of coupling constant values that are much smaller than the line-width from anti-phase cross-peaks. The cancellation effects can be reduced in E. COSY type spectra (Griesinger *et al.* 1985) where the cross peak fine-structure is simplified by suppression of certain components of the fine-structure. Other methods to determine coupling constants rely on the measurement of cross peak intensity ratios (Vuister *et al.* 1994), on a series of spectra with cross peak volumes modulated by the coupling constant (Neri *et al.* 1990), or on in-phase spectra (Szyperki *et al.* 1992*a*). When interpreting scalar coupling constants using equation (4) 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.

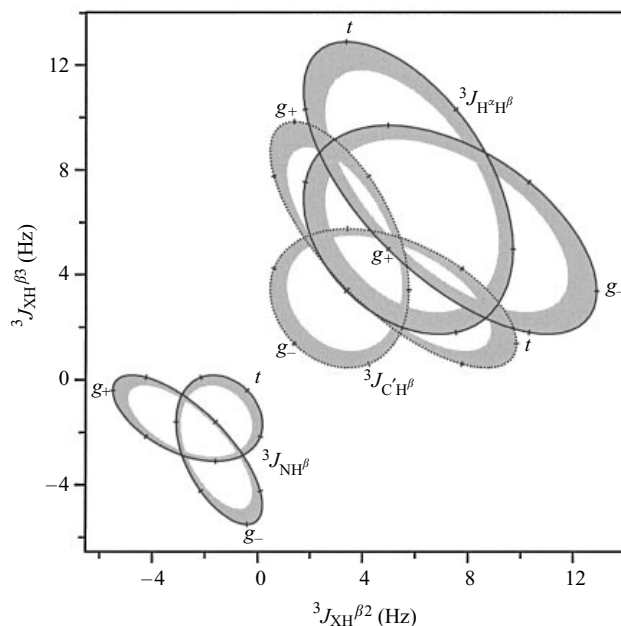


Fig. 7. Relations between vicinal scalar coupling constants for β -methylene protons in proteins. Solid and dotted lines correspond to the Karplus curves given in Table 3. Values in the shaded areas result from additional mobility leading to χ^1 torsion angles that are uniformly distributed within up to $\pm 30^\circ$ around a given value. Ticks on the curves are set in intervals of 30° , and the rotamer positions *gauche+* ($\chi^1 = 60^\circ$), *gauche-* ($\chi^1 = -60^\circ$) and *trans* ($\chi^1 = 180^\circ$) are labelled.

Motions that lead to fluctuations of θ about an average value reduce the amplitude of the Karplus curve but do not strongly change the functional form. For instance, if the torsion angle values are distributed uniformly in an interval $[\theta - \Delta\theta/2, \theta + \Delta\theta/2]$ of width $\Delta\theta$ centred at θ , the resulting Karplus curve maintains the shape of equation (4) but with parameters A , B , C replaced by

$$A' = A \frac{\sin \Delta\theta}{\Delta\theta}, \quad B' = B \frac{\sin \Delta\theta}{\Delta\theta}, \quad C' = C + \frac{A}{2} \left(1 - \frac{\sin \Delta\theta}{\Delta\theta} \right). \quad (5)$$

In the case of uniform fluctuations of, say, the ϕ torsion angle of $\pm 30^\circ$ around a central value, this amounts to a maximal deviation of the Karplus curve given by equations (5) from the corresponding one for a rigid molecule of 1.45 Hz for ${}^3\mathcal{J}_{\text{H}^{\text{NH}^\beta}}$ (Table 3).

A different situation arises if the torsion angle fluctuates between distinctively different conformations, for example if several rotamers of the χ^1 torsion angle are populated. Under such conditions a direct interpretation of the measured \mathcal{J} value with equations (4) and (5) becomes meaningless. However, if several scalar coupling constant values can be measured for a given torsion angle, the set of values provides a method to detect whether the torsion angle is significantly disordered because only certain combinations of the \mathcal{J} values for different atom pairs are compatible with a rigid structure (Fig. 7).

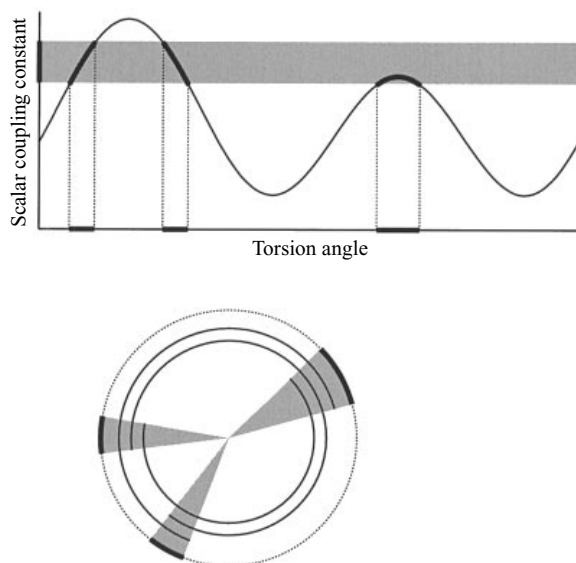


Fig. 8. Conversion of scalar coupling constants to torsion angle restraints. Top: The shaded range for the measured value of a scalar coupling constant leads, according to a Karplus curve, to three separate allowed intervals of the corresponding torsion angle. Bottom: The same allowed angle ranges, shown as sectors of a circle, and three torsion angle restraints, represented by the inner concentric circles. Each restraint restricts the torsion angle to one allowed interval. Applied simultaneously, the three restraints confine the torsion angle to values that are in agreement with the measured coupling constant.

Torsion angle restraints in the form of an allowed interval are used to incorporate scalar coupling information into the structure calculation. Using equation (4), an allowed range for a scalar coupling constant value in general leads to several (up to four) allowed intervals for the enclosed torsion angle (Fig. 8). 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 restraint 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 restraints 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 restraints from the results of this grid search rather than from the individual NMR parameters (see Section 5.1 below; Güntert *et al.* 1989).

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

$$V_J = k_J \sum_i (J_i^{\text{exp}} - J_i^{\text{calc}})^2 \quad (6)$$

to the target function of the structure calculation program (Kim & Prestegard, 1990; Torda *et al.* 1993). The sum in equation (6) extends over all measured

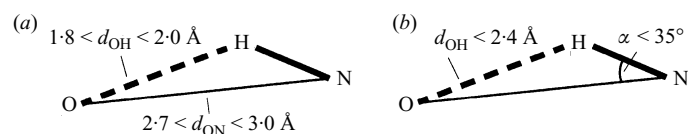


Fig. 9. (a) Hydrogen bond restraints used during a structure calculation (Williamson *et al.* 1985). (b) Criterion used to detect hydrogen bonds when analyzing a structure (Billeter *et al.* 1990; Koradi *et al.* 1996).

couplings, k_j is a weighting factor, and $^3J_i^{\text{exp}}$ and $^3J_i^{\text{calc}}$ denote the experimental and calculated value of the coupling constant, respectively. The latter is obtained from the value of the corresponding torsion angle by virtue of equation (4).

4.3 Hydrogen bonds

Slow hydrogen exchange indicates that an amide proton is involved in a hydrogen bond (Wagner & Wüthrich, 1982). Unfortunately, the acceptor oxygen or nitrogen atom cannot be identified directly by NMR, and 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. The standard backbone-backbone hydrogen bonds in regular secondary structure can be identified with much higher certainty than hydrogen bonds with side-chains. Hydrogen bond restraints are thus 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 bond restraints are introduced into the structure calculation as distance restraints, 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 (Fig. 9a). The second distance restraint restricts the angle of the hydrogen bond. Being tight medium- or long-range distance restraints, their impact on the resulting structure is considerable. Restraints for architectural hydrogen bonds in secondary structures enhance the regularity of the secondary structure elements. In fact, helices and, to a lesser extent, β -sheets can be defined by hydrogen bond restraints alone without the use of NOE restraints (Fig. 10). On the other hand, hydrogen bond restraints 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.

4.4 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 restraints (Oldfield, 1995; Williamson & Asakura, 1997). An

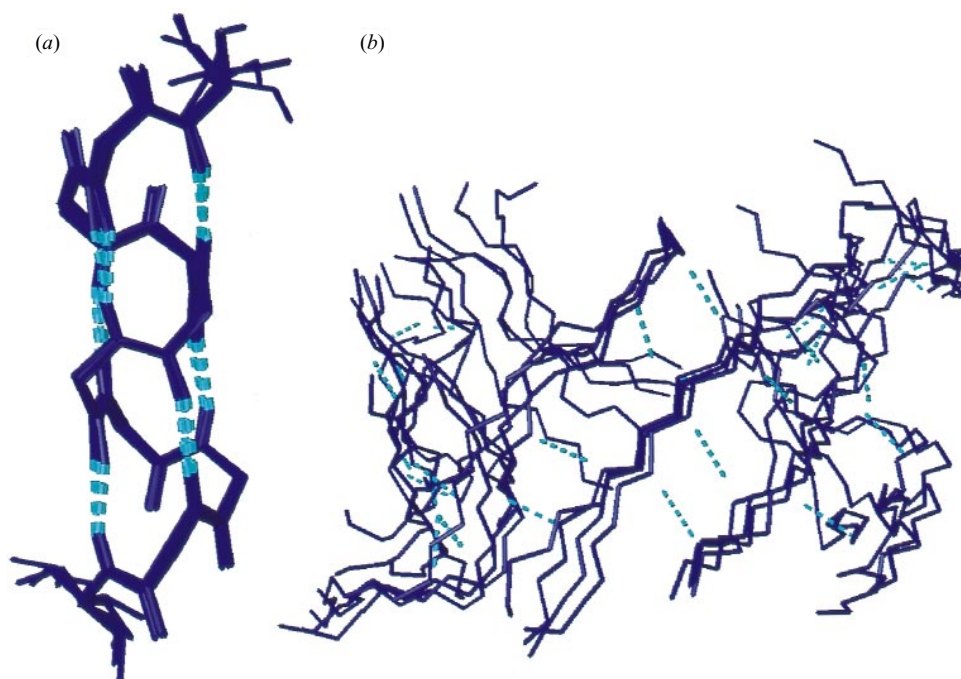


Fig. 10. Potential of hydrogen bond restraints to define three-dimensional protein structures. (a) Group of ten conformers calculated with the program DYANA for a 10-residue polyalanine fragment. Only standard α -helical hydrogen bond restraints, steric lower limits, and loose restraints that restrict the torsion angle ϕ to the range $-180 \leq \phi \leq 0^\circ$ (in order to exclude mirror images) were used in the structure calculation. (b) Group of four conformers for the β -barrel structure in the protein cyclophilin A, calculated with the program DYANA. The same input was used as in (a) except that hydrogen bond restraints were imposed on all HN–O pairs that are connected by a hydrogen bond in more than half of the 20 energy-refined conformers of the solution structure of cyclophilin A (Ottiger *et al.* 1997).

exception in this respect are the deviations of $^{13}\text{C}^\alpha$ (and, to some extent, $^{13}\text{C}^\beta$) chemical shifts from their random coil values that are correlated with the local backbone conformation (Spera & Bax, 1991; de Dios *et al.* 1993): $^{13}\text{C}^\alpha$ chemical shifts larger than the random coil values tend to 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, either through torsion angle restraints (Luginbühl *et al.* 1995) or by a special potential (Kuszewski *et al.* 1995*b*) although care should be applied because the correlation between chemical shift deviation and structure is not perfect. Similar to hydrogen bond restraints, conformational restraints based on $^{13}\text{C}^\alpha$ 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 ^1H chemical shifts as direct restraints in structure refinement

(Ösapay *et al.* 1994; Kuszewski *et al.* 1995a). More often, however, they are used to delineate secondary structure elements by virtue of the ‘chemical shift index’ (Wishart *et al.* 1992) or to assess the quality of a structure (Williamson *et al.* 1995).

4.5 Residual dipolar couplings

Recently, a new class of conformational restraints has been introduced that originates from residual dipolar couplings in partially aligned molecules and gives information on angles between covalent bonds and globally defined axes in the molecule, namely those of the magnetic susceptibility tensor (Tolman *et al.* 1995; Tjandra *et al.* 1996). 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 which 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 it 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 (Gayathri *et al.* 1982). 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 $^1\mathcal{J}$ couplings at different magnetic field strengths (Tolman *et al.* 1995; Tjandra *et al.* 1996). The magnetic susceptibility anisotropy is relatively large in paramagnetic proteins (Tolman *et al.* 1995) 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 (Losonczi & Prestegard, 1998; Tjandra & Bax, 1997).

Assuming an axially symmetric magnetic susceptibility tensor and neglecting the very small contribution from ‘dynamic frequency shifts’, the difference $\Delta\mathcal{J}^{\text{obs}}$ between the apparent \mathcal{J} -values at two different magnetic field strengths, $B_0^{(1)}$ and $B_0^{(2)}$, is given by (Tjandra *et al.* 1997)

$$\Delta\mathcal{J}^{\text{obs}} = \frac{h\gamma_\alpha\gamma_\beta\chi_a S}{60\pi^2 d_{\alpha\beta}^3 k_B T} (B_0^{(2)} - B_0^{(1)})^2 (3 \cos^2 \theta - 1), \quad (7)$$

where h is Planck’s constant, γ_α and γ_β are the gyromagnetic ratios of the two spins α and β , $d_{\alpha\beta}$ the distance between them, k_B the Boltzmann constant, T the temperature, χ_a the axial component of the magnetic susceptibility tensor, and S the order parameter for internal motions (Lipari & Szabo, 1982). The structural information is contained in the angle θ between the covalent bond connecting the two scalar coupled atoms α and β and the main axis of the magnetic susceptibility tensor; given all other constants, equation (7) yields an experimental value for $\cos^2\theta$. It is then straightforward to add an orientation restraint term to the target

function of a structure calculation program that measures the deviation between the value of $\cos^2\theta^{\text{obs}}$ obtained from equation (7) and the corresponding value $\cos^2\theta^{\text{calc}}$ calculated from the structure. Provided that the structure calculation program allows free global reorientation of the molecule in space, the angle θ^{calc} can be measured simply with respect to an arbitrary fixed axis, for instance the z -axis of the global coordinate system. The molecule will then rotate during the structure calculation such as to align, under the simultaneous influence of all orientation restraints, the main axis of the magnetic susceptibility tensor with the z -axis. Tjandra *et al.* (1997) have shown that such orientation restraints can be used in conjunction with conventional distance and angle restraints during the structure calculation, and that they can have a beneficial effect on the quality of the resulting structure. Being in an early stage of their application, the potential of orientation restraints in biomolecular structure calculation remains to be assessed by further research. A particularly interesting, as yet unanswered question is whether they will open an avenue towards non-NOE-based NMR structure determination.

4.6 Other sources of conformational restraints

Additional types of conformational restraints have been used occasionally in NMR structure determination. These included, for example, *ad hoc* restraints to enforce certain conformations that were believed to be present but could not be found unambiguously on the basis of the 'normal' conformational restraints, conformational database potentials to confine the structure to those regions of the Ramachandran plot or side-chain conformation space that are populated in high-resolution X-ray structures (Kuszewski *et al.* 1996), and restraints that are available only for special systems. The latter include, for example, restraints derived from pseudocontact shifts in paramagnetic proteins (Banci *et al.* 1996).

5. PRELIMINARIES OF A STRUCTURE CALCULATION

5.1 Systematic analysis of local conformation

Due to the size and complexity of the conformation space of a biological macromolecule, it is not possible to perform an exhaustive search that would yield a complete description of the accessible regions in conformation space. However, for limited fragments of the macromolecule exhaustive searches are feasible if conformation space is discretized in the form of a multidimensional grid in which each dimension corresponds to a torsion angle or to a group of dependent torsion angles. The basic advantage of such 'grid searches' over statistical sampling methods is that no conformations (on the grid) will be missed and that therefore definite answers to questions like: 'Are the local experimental data consistent with a single rigid structure?' can be given.

The role of grid searches in NMR structure calculation is three-fold: First, inconsistencies among the experimental data can be detected. In general, a

significant fraction of the experimental restraints are short-range in nature and can therefore be checked by considering limited fragments of the complete molecule. Second, stereospecific assignments of diastereotopic protons or isopropyl groups can be obtained by performing separate grid searches for both possible stereospecific assignments (Güntert *et al.* 1989; Nilges *et al.* 1990; Polshakov *et al.* 1995). Third, experimental data such as scalar coupling constants and NOEs can be converted into direct restraints on torsion angles that can be used by structure calculation programs in order to improve both the precision of the resulting structure and the success rate of the calculation.

Grid searches have been used most commonly to analyze the local conformation of protein fragments involving the three torsion angles ϕ , ψ and χ^1 of an amino acid residue, for example employing the programs HABAS (Güntert *et al.* 1989) or STEREOSEARCH (Nilges *et al.* 1990), with the aims of determining the stereospecific assignment of the β -protons and obtaining restraints for the torsion angles ϕ , ψ and χ^1 on the basis of intramolecular and sequential NOEs, and scalar coupling constants within the fragment. Torsion angle restraints can be generated regardless of whether or not an unambiguous stereospecific assignment for the β -protons is obtained, simply by merging the results of the two grid searches for both possible stereospecific assignments. Compared to manual, qualitative conversions of scalar coupling constants into angle restraints (Wüthrich, 1986) and manually derived stereospecific assignments based on the assumption that the χ^1 angle adopts only the three staggered rotamer positions, the automatic grid search approach is more convenient and more objective. Grid search methods are expected to be especially useful also in DNA/RNA structure determination, because long-range NOEs are relatively scarce in nucleic acids and hence short-range restraints become more important (Wijmenga *et al.* 1993; Pardi, 1995; Varani *et al.* 1996).

A new grid search algorithm, implemented as a module of the DYANA structure calculation program (Güntert *et al.* 1997), extends the previous HABAS approach in various directions (P. Güntert, M. Billeter, O. Ohlenschläger, unpublished). It can be applied in a straightforward manner as an automated step of a structure calculation and is described here as an example of a versatile grid search algorithm. In contrast to HABAS that was specific for ϕ , ψ and χ^1 fragments in proteins, arbitrary fragments, defined as connected subsets of torsion angles in the molecule, can be investigated. Covalent geometry parameters (bond lengths, bond angles etc.) and Karplus curve relations (Table 3) are stored in the standard DYANA residue library. This allows the treatment of any type of molecule, in particular proteins and nucleic acids. The fragment size is limited in practice only by the computing power available. All upper and lower limit distance restraints, scalar couplings and angle restraints available within the fragment are used. Criteria to accept or reject conformations can be based either on the maximal restraint violation in the fragment or on the local DYANA target function value (see below), calculated for all restraints within the fragment. Stereospecific assignments can be determined also for other groups than βCH_2 and even if the fragment contains more than one pair of diastereotopic substituents. Sets of dependent torsion angles

are treated as a single degree of freedom. For example, the dependencies of the ribose ring torsion angles $\nu_0, \nu_1, \nu_2, \nu_3, \nu_4$ on the pseudorotation angle P are given by (Saenger, 1984):

$$\nu_k = \nu_{\max} \cos\left(P + \frac{4\pi}{5}(k-2)\right) \quad (P \in [0, 2\pi]; \nu_{\max} = 40^\circ). \quad (8)$$

The parameter P is used as a single degree of freedom in the grid search. This ensures that the relations among the dependent torsion angles are always fulfilled and significantly increases the efficiency of the algorithm. Output torsion angle restraints may comprise more than one allowed interval for one torsion angle (Fig. 8). Results from different grid searches for overlapping fragments can be combined, both in order to reduce the computational effort required for higher-dimensional grid searches and in order to generate the narrowest possible output torsion angle restraints.

The molecular fragment to be analysed in the grid search is defined by selecting a connected subset of torsion angles (Fig. 11*a*). The program then extracts from the data for the complete molecule the subset of conformational restraints within the chosen fragment and evaluates them in a multidimensional grid search. The grid search is implemented by n nested loops, each of which corresponds to a degree of freedom. In routine applications the results of the grid search comprise the number of allowed conformations, N , and the sets of allowed values for each torsion angle in the fragment (Fig. 11*b*), i.e. information about correlations between torsion angles is discarded. It is, however, possible to save all allowed conformations for further analysis. The number of allowed conformations yields the important information whether the set of restraints for a fragment is compatible, within the given limits on the target function value or the sizes of violations, with a rigid conformation ($N > 0$) or not ($N = 0$). Inconsistencies among the local restraints can be detected in this way at the beginning of an NMR structure determination. If the fragment under investigation contains M pairs of diastereotopic substituents for which the stereospecific assignment is not known, 2^M grid searches will be performed, one for each combination of possible stereospecific assignments, and their results will be combined. Alternatively, the algorithm can be used also to find stereospecific assignments on the basis of local restraints by checking whether the restraints are compatible with only one of the two possible stereospecific assignment of a diastereotopic pair. In practice, a complete local conformation analysis of a macromolecule includes many grid searches for (possibly overlapping) fragments, and finally restraints are generated for all torsion angles that formed part of at least one of the fragments.

5.2 Stereospecific assignments

The standard method for obtaining resonance assignments in proteins (Wider *et al.* 1982; Wüthrich *et al.* 1982) cannot provide stereospecific assignments, i.e. individual assignments for the two diastereotopic substituents of a prochiral centre, for example in methylene groups and in the isopropyl groups of valine and

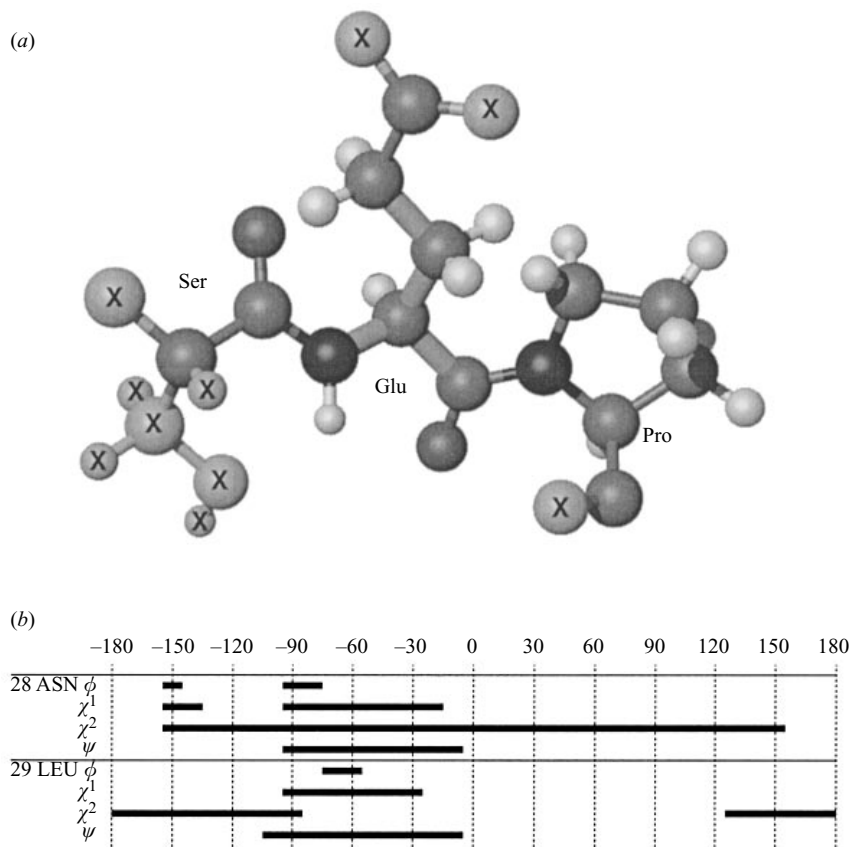


Fig. 11. (a) Ball-and-stick representation of a molecular fragment defined by the four torsion angles ϕ , ψ , χ^1 and χ^2 (indicated by thick bonds) of the central Glu amino acid residue. Atoms not included in the fragment are marked with 'X'. (b) Allowed torsion angle ranges (horizontal bars) obtained with the grid search module of the program DYANA for two ϕ - ψ - χ^1 - χ^2 -fragments of the protein P14a, using the experimental NMR data set of Fernández *et al.* (1997).

leucine. In the absence of stereospecific assignments restraints involving diastereotopic substituents have to be referred to pseudoatoms (Wüthrich *et al.* 1983), or otherwise treated such that they are invariant under exchange of the two diastereotopic substituents (see next section), which inevitably results in a loss of information and less well defined structures (Güntert *et al.* 1989; Table 8 below). It is therefore important for obtaining a high-quality structure that as many stereospecific assignments as possible are determined. Stereospecific assignments of valine and leucine isopropyl groups can be determined experimentally by biosynthetic fractional ^{13}C -labelling (Senn *et al.* 1989; Neri *et al.* 1989). Stereospecific assignments for methylene protons have to be determined in the course of the structure calculation, either manually (Hyberts *et al.* 1987), by systematic analysis of the local conformation around a methylene group, or by reference to preliminary three-dimensional structures.

The local method, introduced by Güntert *et al.* (1989), consists of two separate grid searches, one for each of the two assignment possibilities. An unambiguous stereospecific assignment results if allowed conformations occur only for one of the two possible assignments. It relies exclusively on scalar coupling constants and local distance restraints. Assuming realistic error ranges for experimental data it will not be possible to obtain unambiguous stereospecific assignments by the local method in all cases. Using complete simulated sets of local distance restraints and homonuclear coupling constants with an accuracy of ± 2 Hz, it was estimated that the program HABAS can yield unambiguous stereospecific assignments for about 50% of the β -methylene protons (Güntert *et al.* 1989).

In contrast to the local method, global methods aim at the determination of stereospecific assignments either during the calculation of a three-dimensional structure or by reference to preliminary three-dimensional structures. They have the potential advantage over the local method that all conformational restraints, not only local ones, can be exploited, but, on the other hand, a systematic search of allowed conformations is no longer feasible, and the stereospecific assignments have to be based on a statistical analysis of a limited number of conformers. In conjunction with structure calculation programs working in Cartesian coordinate space, the so-called method of 'floating stereospecific assignments' (Weber *et al.* 1988) can be used: At the beginning of a structure calculation a strong reduction of the corresponding potential energy terms allows the two diastereotopic substituents to interchange freely under the influence of the restraints before they later become fixed when the potential energy terms are slowly restored to their normal values (which inhibit an interchange of the diastereotopic substituents). A stereospecific assignment is considered to be unambiguous if it is found consistently in all conformers that were calculated. Another simple method for obtaining stereospecific assignments is implemented in the program GLOMSA of the DIANA package (Güntert *et al.* 1991a) and consists of an analysis of preliminary three-dimensional structures: If there are two NOEs of significantly different strength from a given proton to both diastereotopic substituents of a prochiral centre and if the distances from the given proton to the two diastereotopic substituents differ consistently in the structures, the stronger NOE can be identified with the diastereotopic substituent that is closer to the given proton.

5.3 Treatment of distance restraints to diastereotopic protons

Distance restraints involving diastereotopic substituents that could not be assigned stereospecifically have to be modified such that they are invariant under exchange of the two diastereotopic substituents. Traditionally, this is achieved by referring the restraints to a pseudoatom located centrally with respect to the two diastereotopic substituents and a concomitant increase of the upper distance bound, b_Q , by a pseudoatom correction, c_Q , equal to the distance from the pseudoatom to the individual protons, i.e. $b_Q = \min(b_1, b_2) + c_Q$, where b_1 and b_2 are two individual upper bounds (Wüthrich *et al.* 1983). This approach, however, completely discards the weaker of the two possible NOEs from a given proton to

the two diastereotopic substituents. Another straightforward method symmetrizes the two restraints by imposing the less restrictive of the two upper bounds, $\max(b_1, b_2)$, simultaneously on both distances to the individual diastereotopic substituents. The pseudoatom and symmetrization treatments are not equivalent and a combination of the two treatments, implemented in the programs DIANA (Güntert *et al.* 1991a) and DYANA (Güntert *et al.* 1997), can give improved results. In addition, it makes use of the information from both upper bounds, b_1 and b_2 , also by assigning a more restrictive upper limit, b_Q , to the restraint to the pseudoatom,

$$b_Q = \frac{b_1^2 + b_2^2}{2} - c_Q^2. \quad (9)$$

Yet another approach, used for example in the program XPLOR (Brünger, 1992), does not introduce a pseudoatom explicitly but assumes that the NOE originates from both diastereotopic substituents simultaneously and imposes a distance restraint on a 'sum-averaged' distance,

$$\langle d \rangle = (d_1^{-6} + d_2^{-6})^{-\frac{1}{6}}, \quad (10)$$

rather than on the individual distances to the two diastereotopic substituents, d_1 and d_2 . Such sum-averaged distance restraints can be used for all groups of non-stereassigned or degenerate protons without the need for multiplicity or pseudoatom corrections (Fletcher *et al.* 1996).

Fig. 12 illustrates and compares the various methods in the two cases of significantly different or similar upper bounds on the distances from the two diastereotopic substituents to a third proton. Obviously, none of the methods can completely make up for the loss of information that results from the absence of a stereospecific assignment but there are significant differences between the various procedures. In general, the loss of information is largest with the original pseudoatom concept and smallest with the DIANA/DYANA treatment, which leads to more precise structures (Güntert *et al.* 1991a).

5.4 Removal of irrelevant restraints

The number of experimental distance restraints used in a structure calculation is an important parameter that determines the precision of the resulting structure. To allow for meaningful comparisons it is therefore important to report the number of relevant distance restraints, i.e. of those that actually restrict the allowed conformation space, rather than the total number of NOESY cross peaks that have been assigned. In addition, the removal of irrelevant distance restraints increases slightly the efficiency of the structure calculation by obviating unnecessary computations. In practice, often more than half of the intraresidual and many sequential restraints are irrelevant. Those include restraints for fixed distances, for example between geminal protons among the protons attached to an aromatic ring, and distance bounds that cannot be reached by any conformation, for example an upper bound of 3.5 Å for the intraresidual distance between the

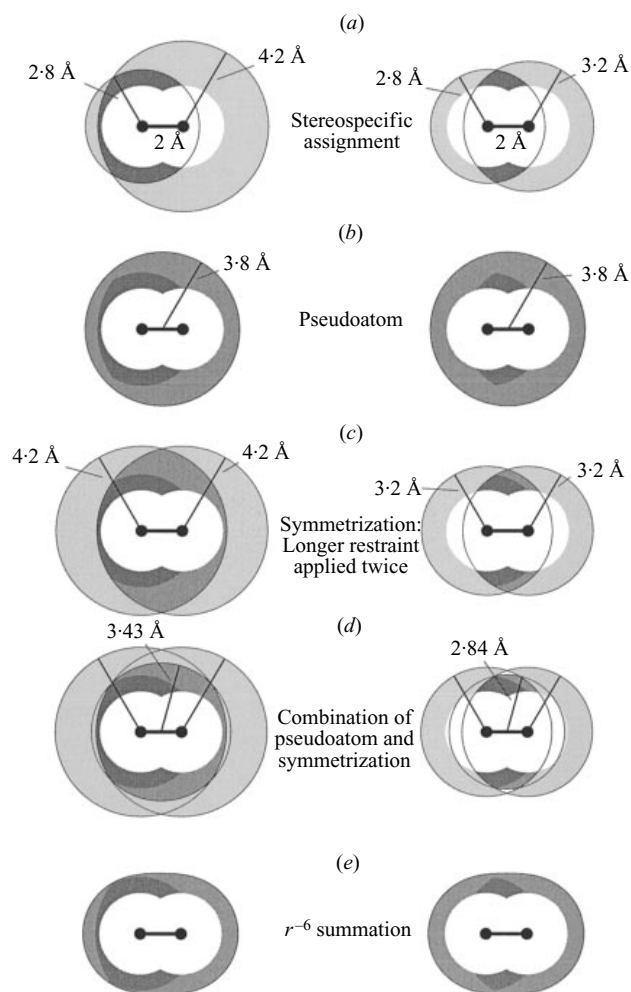


Fig. 12. Treatment of distance restraints with diastereotopic protons. The situation of two distance restraints of 2.8 Å and 4.2 Å (left column), or 2.8 Å and 3.2 Å (right column) from two geminal methylene protons with a separation of 2 Å to a common third atom is shown. The methylene protons are shown as black dots, surrounded by white spheres with a radius of 2 Å that represent the sterically forbidden area around them. Dark shading indicates the area of allowed positions for the third atom if the methylene protons are stereospecifically assigned. Additionally allowed areas in the absence of a stereospecific assignment are shaded in medium grey, and the individual distance restraints are shown as lightly shaded thin circles. (a) Stereospecific assignment is available. (b) Conventional pseudoatom treatment (Wüthrich *et al.* 1983). (c) The larger of the two upper distance bounds is applied to both methylene protons. (d) Combination of a pseudoatom treatment with minimal pseudoatom correction and the method of two identical upper bounds as implemented in the programs DIANA and DYANA. (e) The smaller of the two distance bounds is applied to the r^{-6} sum of the distances to the two methylene protons.

backbone amide- and the α -proton. Assuming rigid bond lengths and bond angles, the latter condition can be checked readily for distances that depend on one or two torsion angles (Güntert *et al.* 1991a).

6. STRUCTURE CALCULATION ALGORITHMS

The calculation of the three-dimensional structure forms a cornerstone of the NMR method for protein structure determination. Due to 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 restraints 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 which measures the agreement between a conformation and the given set of restraints. In the following, the four most widely used types of algorithms (Table 2) are discussed. Because the earlier methods have been reviewed extensively already (Braun, 1987; Brünger & Nilges, 1993; Sutcliffe, 1993; James, 1994; Nilges, 1996), special emphasis is given to the new structure calculation method based on torsion angle dynamics which is currently the most efficient way to calculate NMR structures of biological macromolecules.

6.1 Metric matrix distance geometry

Distance geometry based on the metric matrix was the first approach used for the structure calculation of proteins on the basis of NMR data (Braun *et al.* 1981; Havel & Wüthrich, 1984). It relies on the fact that the NOE data and most of the stereochemical data can be represented as distance restraints. Metric matrix distance geometry is based on the theorem (Blumenthal, 1953; Crippen, 1977; Crippen & Havel, 1988) that, given exact values for all distances among a set of points in three-dimensional Euclidean space, it is possible to determine Cartesian coordinates for these points uniquely except for a global inversion, translation and rotation.

To see this, assume that all $n \times n$ distances $D_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ are known among n points in three-dimensional Euclidean space with (unknown) coordinates $\mathbf{r}_1, \dots, \mathbf{r}_n$ that can be assumed, without loss of generality, to fulfill the relation $\sum_i \mathbf{r}_i = \mathbf{0}$. Then, the $n \times n$ metric matrix G , with elements

$$G_{ij} = \mathbf{r}_i \cdot \mathbf{r}_j = \begin{cases} \frac{1}{n} \sum_{k=1}^n D_{ik}^2 - \frac{1}{2n^2} \sum_{k,l=1}^n D_{kl}^2, & i = j \\ \frac{D_{ij}^2 - G_{ii} - G_{jj}}{2}, & i \neq j \end{cases} \quad (11)$$

can be calculated. G has at most three positive eigenvalues λ^α with corresponding n -dimensional eigenvectors e^α that are related to the Cartesian coordinates $\mathbf{r}_1, \dots, \mathbf{r}_n$ of the n points by

$$r_i^\alpha = \sqrt{\lambda^\alpha} e_i^\alpha \quad (\alpha = 1, 2, 3). \quad (12)$$

Equations (11) and (12) provide a straightforward way to embed a distance matrix in three-dimensional space, i.e. to obtain Cartesian coordinates for a set of points if all distances are known exactly. To make use of this theorem in a structure calculation one has to account for the fact that in practice neither complete nor exact distance information is available. Only for a small subset of all distances d_{ij} , restraints in the form of lower and upper bounds, $l_{ij} < d_{ij} < u_{ij}$, can be determined. Upper bounds result from NOEs, lower bounds from the steric repulsion, and there are some exact distance constraints from known bond lengths and bond angles of the covalent structure. To apply equations (11) and (12), unknown upper bounds are first initialized to a large value, and unknown lower bounds to zero. Subsequently they are reduced by ‘bounds smoothing’ (Crippen, 1977), i.e. repeated application of the triangle inequality until all lower and upper bounds are consistent with the triangle inequality. Then a complete set of distances is produced by ‘randomly’ selecting for each distance a value between the corresponding lower and upper bounds, and the embedding procedure equations (11) and (12) is used to obtain Cartesian coordinates. Because the assumptions of the embedding theorem are not met exactly, the resulting structure will in general have the correct three-dimensional fold (or its mirror image) but will be severely distorted. It needs to be regularized extensively, for example by conjugate gradient minimization of an appropriate target function in Cartesian coordinate space (Havel & Wüthrich, 1984). Nowadays a crudely regularized structure is usually passed as start structure to simulated annealing by molecular dynamics (Nilges *et al.* 1988a; Brünger, 1992). Starting from the smoothed distance bound matrix, the calculation is repeated with different ‘random’ selections of distances, in order to obtain a group of conformers whose spread should give an indication of the allowed conformation space.

Despite the elegance of embedding method given by equations (11) and (12) there are a number of problems that have to be dealt with. Since all conformational data has to be encoded into the distance matrix, it is not possible to introduce any handedness or chirality. A structure and its mirror image are always equivalent for metric matrix distance geometry. The correct handedness is only enforced during regularization. For the same reason, torsion angle restraints cannot be used directly in the embedding; they have to be represented by distance bounds, thereby losing part of their information.

The sampling of conformation space by a group of conformers resulting from metric matrix distance geometry is decisively determined by the ‘random’ selection of distance values between corresponding lower and upper bounds. The most straightforward method, namely selecting the distances as independent, uniformly distributed random variables between the two limits, leads, because meaningful upper bounds exist only for a small subset of all distances, on the

average to an overestimation of the true distances with the consequence of artificially expanded structures (Metzler *et al.* 1989; Havel, 1990). This effect is most pronounced in regions of the polypeptide chain for which only few restraints are available. For example, chain ends that are unstructured in solution tend to be forced into an extended conformation. A method to reduce – at the expense of considerably increased computation time – such biased sampling of the allowed conformation space is metrization (Havel & Wüthrich, 1984): instead of selecting the individual distances independently from each other, the bounds smoothing is repeated each time after a distance value has been chosen, thereby resulting in a more consistent set of distances for the embedding. This introduces, however, a strong dependence of the sampling properties on the order in which the distances are chosen (Havel, 1990). Good sampling can be achieved if the distances are chosen in random order (Havel, 1990, 1991). The computational efficiency of metrization can be enhanced by partial metrization, i.e. by repeating the bounds smoothing only after the selection of the first few percent of the randomly chosen distances (Kuszewski *et al.* 1992).

6.2 *Variable target function method*

The basic idea of the variable target function algorithm (Braun & Go, 1985) is to gradually fit an initially randomized starting structure to the conformational restraints collected with the use of NMR experiments, starting with intraresidual restraints only, and increasing the ‘target size’ step-wise up to the length of the complete polypeptide chain. Advantages of the method are its conceptual simplicity and the fact that it works in torsion angle space, strictly preserving the covalent geometry during the entire calculation. The variable target function algorithm was implemented first in the program DISMAN (Braun & Go, 1985) and most commonly used in its implementation in the program DIANA (Güntert *et al.* 1991a), which is discussed here. Today, however, the variable target function method has been superseded largely by the more efficient torsion angle dynamics algorithm. Since both algorithms work in torsion angle space, they have many features in common. These are described in detail in the section about torsion angle dynamics below (see Section 6.4).

The variable target function algorithm is based on the minimization of a target function that includes terms for experimental and steric restraints. To reduce the danger of becoming trapped in a local minimum with a function value much higher than the global minimum, the target function is varied during a structure calculation. At the outset only local restraints with respect to the polypeptide sequence are considered. Subsequently, restraints between atoms further apart with respect to the primary structure are included in a step-wise fashion (Fig. 13). Consequently, in the first stages of a structure calculation the local features of the conformation will be established, and the global fold of the protein will be obtained only towards the end of the calculation. The minimization algorithm used in the program DIANA is the well-known method of conjugate gradients (Powell, 1977) that tries to find the minimum by taking exclusively downhill steps.

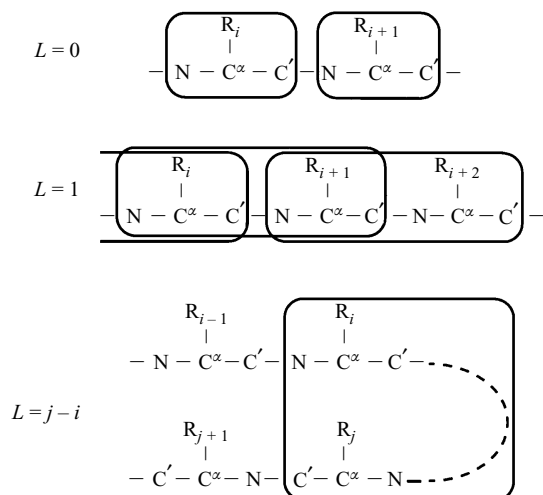


Fig. 13. Active restraints at various minimization levels L of the variable target function algorithm. At a given minimization level L , all distance restraints between residues i and j with $|j-i| \leq L$ are considered.

As an alternative, a Newton–Raphson minimization algorithm that uses the matrix of second derivatives (Abe *et al.* 1984) has been used in the program DADAS90 (Endo *et al.* 1991).

A drawback of the basic implementation of the variable target function algorithm (Braun & Go, 1985) is that for all but the simplest molecular topologies only a small percentage of the calculations converge with small residual restraint violations, which is a typical local minimum problem. Because of the low yield of acceptable conformers, calculations had to be started with a large number of randomized start conformers in order to obtain a group of good solutions, sometimes compromising between the requirements of small restraint violations and the available computing time (Kline *et al.* 1988). The introduction of the optimized program DIANA (Güntert *et al.* 1991a) reduced significantly the computation time needed for the calculation of a single conformer, and a workable situation was achieved for α -helical proteins (Güntert *et al.* 1991b). Nonetheless, the situation for β -proteins with more complex topology remained unsatisfactory and was improved decisively only with the use of redundant dihedral angle restraints (REDAC; Güntert & Wüthrich, 1991).

When using REDAC, the structure calculation is performed in iterative cycles that provide a partial feedback of structural information gathered from the conformers of the preceding cycle. To this end, an amino acid residue in a given conformer is considered to have an acceptable conformation if the target function value due to violations of restraints involving atoms or torsion angles of this residue is below a predefined value, and if the same condition holds for the two sequentially neighbouring residues, too. Redundant torsion angle restraints are then generated and added to the input for the next cycle of DIANA structure calculations for all

residues that were found to be acceptable in a sufficient number of conformers by taking the two extreme torsion angle values in the group of acceptable conformers as upper and lower bounds. This method is able to reduce the computational effort required to obtain a set of converged conformers by a factor of 30 already in the case of a small protein like BPTI. This improvement is achieved without detectable reduction in the sampling of conformation space (Güntert & Wüthrich, 1991). To rationalize the empirically found higher yield of good conformers with the use of REDAC it is important to note that in many regions of a protein structure, in particular in β -strands, the local conformation is determined not only by the local conformational restraints, but also by long-range restraints, such as interstrand distance restraints in β -sheets. The local restraints alone may allow for multiple local conformations at low target levels in a variable target function calculation, of which some may be incompatible with the long-range restraints taken into account later during the calculation. Obviously, incorrect local conformations that satisfy the experimentally available local restraints are potential local minima that could only be ruled out from the beginning if the information contained in the long-range restraints were already available at low levels of the minimization. The use of REDAC achieves this: information contained in the complete data set is translated into (by definition intraresidual) torsion angle restraints. It further makes clear why the yield of good solutions with the original variable target function method (Braun & Go, 1985) was in general higher for α -proteins than for β -proteins, since the conformation of an α -helix is particularly well-determined by sequential and medium-range restraints.

6.3 *Molecular dynamics in Cartesian space*

This third 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 (Allen & Tildesley, 1987). The degrees of freedom are the Cartesian coordinates of the atoms. In contrast to 'standard' molecular dynamics simulations (McCammon & Harvey, 1987; Brooks *et al.* 1988; van Gunsteren & Berendsen, 1990) that try to simulate the behaviour of a real physical system as closely as possible (and do not include restraints 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 fulfil the restraints, i.e. that minimize a target function which is taken as the potential energy of the system. Therefore, simulated annealing (Kirkpatrick *et al.* 1983; Nilges *et al.* 1988a; Scheek *et al.* 1989; Brünger & Nilges, 1993, Brünger *et al.* 1997) 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) non-bonded interactions, and that 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 to cross barriers of the potential surface, thereby reducing greatly the problem of becoming trapped in local minima. Since molecular

dynamics simulation cannot generate conformations from scratch, a start structure is needed, that can be generated either by metric matrix distance geometry (Nilges *et al.* 1988*a*) or by the variable target function method, but – at the expense of increased computation time – it is also possible to start from an extended structure (Nilges *et al.* 1988*c*) or even from a set of atoms randomly distributed in space (Nilges *et al.* 1988*b*). Any general molecular dynamics program, such as CHARMM (Brooks *et al.* 1983), AMBER (Pearlman *et al.* 1991), or GROMOS (van Gunsteren *et al.* 1996), can be used for the simulated annealing of NMR structures, provided that pseudoenergy terms for distance and torsion angle restraints have been incorporated. In practice, the program best adapted and most widely used for this purpose is XPLOR (Brünger, 1992).

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,

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = \mathbf{F}_i \quad (i = 1, \dots, n), \quad (13)$$

where the forces \mathbf{F}_i are given by the negative gradient of the potential energy function E_{pot} with respect to the Cartesian coordinates: $\mathbf{F}_i = -\nabla_i E_{\text{pot}}$. 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 non-bonded interactions, as well as terms for distance and torsion angle restraints. For example, in the program XPLOR (Brünger, 1992),

$$\begin{aligned} E = & \sum_{\text{bonds}} k_b (r - r_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi (1 + \cos(n\phi + \delta)) \\ & + \sum_{\text{impropers}} k_\phi (\phi - \delta)^2 + \sum_{\text{nonbonded pairs}} k_{\text{repel}} (\max(0, (sR_{\text{min}})^2 - R^2))^2 \\ & + \sum_{\text{distance restraints}} k_a \Delta_d^2 + \sum_{\text{angle restraints}} k_a \Delta_a^2 \end{aligned} \quad (14)$$

k_b , k_θ , k_ϕ , k_{repel} , k_a 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 and correct bond angle, ϕ the actual torsion 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 non-bonded 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 of equation (14), distance restraints are often represented by a potential with linear asymptote for large violations (Brünger, 1992). To obtain a trajectory, the equations of motion are numerically integrated by advancing the coordinates \mathbf{r}_i and velocities $\mathbf{v}_i = \dot{\mathbf{r}}_i$ of the particles by a small but finite time step

Δt , for example according to the ‘leap-frog’ integration scheme (Hockney, 1970; Allen & Tildesley, 1987):

$$\mathbf{v}_i(t + \Delta t/2) = \mathbf{v}_i(t - \Delta t/2) + \Delta t \mathbf{F}_i(t)/m_i + O(\Delta t^3), \quad (15)$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \Delta t \mathbf{v}_i(t + \Delta t/2) + O(\Delta t^3). \quad (16)$$

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 (Ryckaert *et al.* 1977). The temperature may be controlled by coupling the system loosely to a heat bath (Berendsen *et al.* 1984). 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 (Nilges *et al.* 1988*a*). In a typical simulated annealing protocol (Brünger, 1992), the simulated annealing is performed for a few picoseconds at high temperature, say $T = 2000$ 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. In general, simulated annealing by molecular dynamics requires substantially more computation time per conformer (Brünger, 1992) than, for example, the variable target function method but this effect may be compensated by a higher success rate of 40–100% of the start conformers which is due to the ability of the algorithm to escape from local minima.

6.4 Torsion angle dynamics

Torsion angle dynamics, i.e. molecular dynamics simulation using torsion angles instead of Cartesian coordinates as degrees of freedom (Bae & Haug, 1987; Güntert *et al.* 1997; Jain *et al.* 1993; Katz *et al.* 1979; Kneller & Hinsen, 1994; Mathiowetz *et al.* 1994; Mazur & Abagyan, 1989; Mazur *et al.* 1991; Rice & Brünger, 1994; Stein *et al.* 1997), provides at present the most efficient way to calculate NMR structures of biomacromolecules. In this section the torsion angle dynamics algorithm implemented in the program DYANA (Güntert *et al.* 1997) is described in some detail. This seems warranted in light of the wide-spread but incorrect belief that dynamics in generalized coordinates is hopelessly complicated and cannot be done efficiently. DYANA employs the torsion angle dynamics algorithm of Jain *et al.* (1993) that requires a computational effort proportional to the system size, as it is the case for molecular dynamics simulation in Cartesian

Table 4. Comparison of molecular dynamics simulation in Cartesian and torsion angle space

Quantity	Cartesian space	Torsion angle space
Degrees of freedom	$3N$ coordinates: $\mathbf{x}_1, \dots, \mathbf{x}_N$	n torsion angles: $\theta_1, \dots, \theta_n$
Equations of motion	Newton's equations: $m_i \ddot{\mathbf{x}}_i = -\frac{\partial E_{\text{pot}}}{\partial \mathbf{x}_i}$	Lagrange equations: $\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{\theta}_k} \right) - \frac{\partial L}{\partial \theta_k} = 0 \quad (L = E_{\text{kin}} - E_{\text{pot}})$
Kinetic energy	$E_{\text{kin}} = \frac{1}{2} \sum_{i=1}^N m_i \dot{\mathbf{x}}_i^2$	$E_{\text{kin}} = \frac{1}{2} \sum_{k,l=1}^n M(\theta)_{kl} \dot{\theta}_k \dot{\theta}_l$
Mass matrix M	Diagonal, elements m_i	$n \times n$, non-diagonal, non-constant
Accelerations	$\ddot{\mathbf{x}}_i = -\frac{1}{m_i} \frac{\partial E_{\text{pot}}}{\partial \mathbf{x}_i}$	$\ddot{\theta} = M(\theta)^{-1} C(\theta, \dot{\theta})$ (n linear equations)
Computational complexity of acceleration calculation	Proportional to N	If solving system of linear equations: proportional to n^3 If exploiting tree structure of molecule: proportional to n

space, too. The advantages of torsion angle dynamics, especially the much longer integration time steps that can be used, are therefore effective for molecules of all sizes, and in particular for large biological macromolecules. A comparison of molecular dynamics simulation in Cartesian and torsion angle space in Table 4 shows the close analogy between the two methods.

6.4.1 Tree structure of the molecule

For torsion angle dynamics calculations with DYANA the molecule is represented as a tree structure consisting of a base rigid body that is fixed in space and n rigid bodies, which are connected by n rotatable bonds (Fig. 14a; Katz *et al.* 1979; Abe *et al.* 1984). The degrees of freedom are exclusively torsion angles, i.e. rotations about single bonds. 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 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 rigid bodies are numbered from 0 to n . The base has the number 0. Each other rigid body, with a number $k \geq 1$, has a single nearest neighbour in the direction toward the base, which has a number $p(k) < k$ (Fig. 14). 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, $\theta = (\theta_1, \dots, \theta_n)$. For each rotatable bond, \mathbf{e}_k denotes a unit vector in the direction of the bond, and \mathbf{r}_k is the position vector of its end point, which is subsequently used as the 'reference point' of the rigid body k . In the following description these and all other three-dimensional vectors are referred to an inertial frame of reference that is fixed in space. Covalent

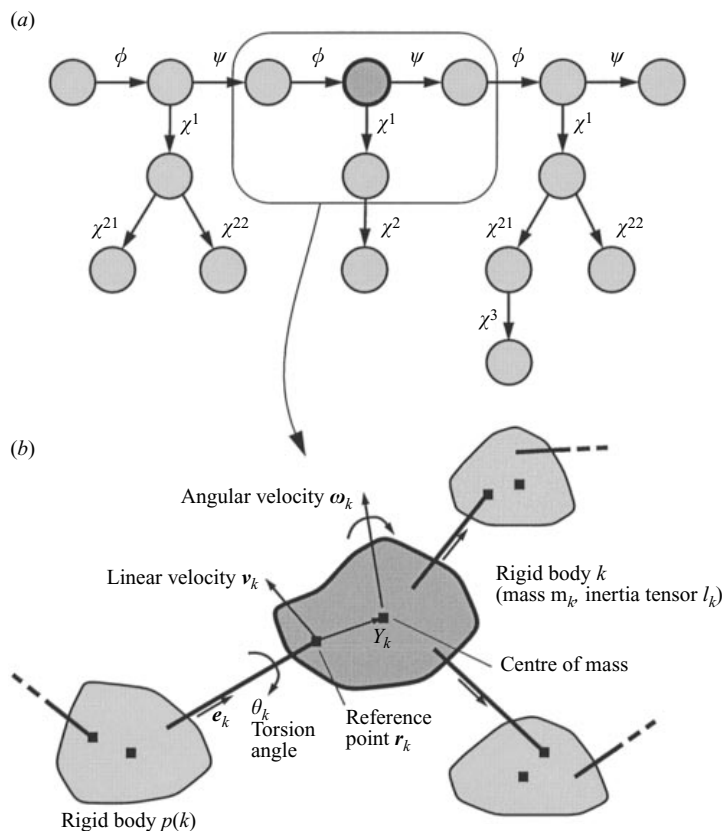


Fig. 14. (a) Tree structure of torsion angles for the tripeptide Val-Ser-Ile. Circles represent rigid units. Rotatable bonds are indicated by arrows that point towards the part of the structure that is rotated if the corresponding dihedral angle is changed. (b) Excerpt from the tree structure formed by the torsion angles of a molecule, and various quantities required by the torsion angle dynamics algorithm of Jain *et al.* (1993).

bonds that are incompatible with a tree structure because they would introduce closed flexible rings, for example disulphide bridges, are treated, as in Cartesian space dynamics, by distance constraints.

6.4.2 Potential energy

The target function V takes the role of the potential energy E_{pot} , i.e. $E_{\text{pot}} = w_0 V$, with an overall weighting factor $w_0 = 10 \text{ kJ mol}^{-1} \text{ \AA}^{-2}$. The target function $V \geq 0$ is defined such that $V = 0$ if and only if all experimental distance restraints and torsion angle restraints are fulfilled and all non-bonded atom pairs satisfy a check for the absence of steric overlap. It measures restraint violations such that $V(\theta) < V(\theta')$ whenever a conformation θ satisfies the restraints more closely than another conformation θ' . The exact definition of the DYANA target function is:

$$V = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta) \in I_c} f_c(d_{\alpha\beta}, b_{\alpha\beta}) + w_a \sum_{i \in I_a} \left(1 - \frac{1}{2} \left(\frac{A_i}{F_i} \right)^2 \right) A_i^2 \quad (17)$$

Upper and lower bounds, $b_{\alpha\beta}$, on distances between two atoms α and β , $d_{\alpha\beta}$, 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 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 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 of equation (17) is continuously differentiable over the entire conformation space, and is chosen such that the contribution of a single small violation δ_c is given by $w_c \delta_c^2$ for all types of restraints. The sets I_u , I_l and I_v of distance restraints that contribute to the target function can include all distance restraints or only those between residues with sequence numbers that differ by not more than a given target level L (Fig. 13).

The function $f_c(d, b)$ that measures the contribution of a violated distance restraint to the target function can be a simple square potential,

$$f_c(d, b) = (d - b)^2, \quad (18)$$

or have the form used in the program DIANA (Güntert *et al.* 1991a),

$$f_c(d, b) = \left(\frac{d^2 - b^2}{2b} \right)^2, \quad (19)$$

or be a function with a linear asymptote for large restraint violations

$$f_c(d, b) = 2\beta^2 b^2 \left[\frac{1}{1 + \left(\frac{d-b}{\beta b} \right)^2} - 1 \right], \quad (20)$$

where β is a dimensionless parameter that weighs large violations relative to small ones. For small restraint violations equations (18)–(20) all yield the same contribution, which is always equal to the square of the restraint violation, but there is a pronounced difference for large violations, where the contributions are proportional to the second, fourth and first power of the restraint violation, respectively (Fig. 15).

The torques about the rotatable bonds, i.e. the negative gradients of the potential energy with respect to torsion angles, $-\nabla E_{\text{pot}}$, are calculated by the fast recursive algorithm of Abe *et al.* (1984). The partial derivative of the function V of equation (17) with respect to a torsion angle θ_k is given by

$$\frac{\partial V}{\partial \theta_k} = -\mathbf{e}_k \cdot \mathbf{g}_k - (\mathbf{e}_k \wedge \mathbf{r}_k) \cdot \mathbf{h}_k + 2w_a \sum_{i \in I_a} \left(1 - \left(\frac{\Delta_i}{\Gamma_i} \right)^2 \right) \Delta_i \delta_{ik}, \quad (21)$$

where

$$\left. \begin{aligned} \mathbf{g}_k &= \sum_{c=u, l, v} w_c \sum_{\substack{(\alpha, \beta) \in I_c \\ \alpha \in M_k}} \frac{\partial f_c(d_{\alpha\beta}, b_{\alpha\beta})}{\partial d_{\alpha\beta}} \frac{\mathbf{r}_\alpha \wedge \mathbf{r}_\beta}{d_{\alpha\beta}}, \\ \mathbf{h}_k &= \sum_{c=u, l, v} w_c \sum_{\substack{(\alpha, \beta) \in I_c \\ \alpha \in M_k}} \frac{\partial f_c(d_{\alpha\beta}, b_{\alpha\beta})}{\partial d_{\alpha\beta}} \frac{\mathbf{r}_\alpha - \mathbf{r}_\beta}{d_{\alpha\beta}}. \end{aligned} \right\} \quad (22)$$

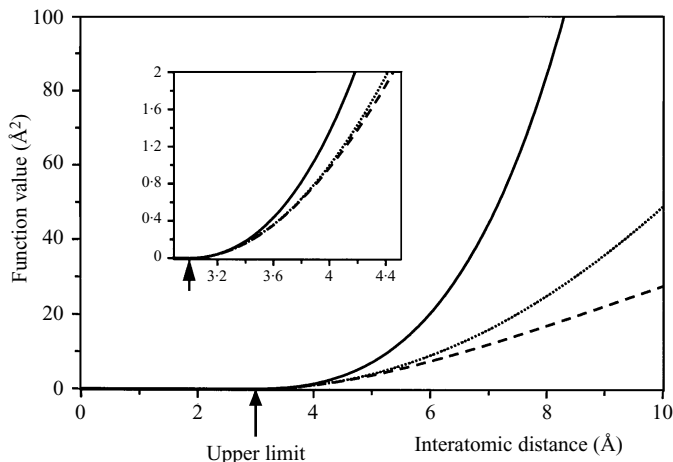


Fig. 15. Contribution of a distance restraint with an upper limit of 3 Å to the target function. The solid line corresponds to the DIANA target function (Güntert *et al.* 1991a), the dotted line to a square potential, and the dashed line to a square potential with linear asymptote for large violations. The inset shows a blow-up of the region of small restraint violations.

\mathbf{r}_α and \mathbf{r}_β denote the position vectors of the atoms α and β , respectively, \mathbf{e}_k denotes the unit vector along the rotatable bond k , \mathbf{r}_k the start point of it (Fig. 14b), and M_k the set of all atoms whose positions are affected by a change of the torsion angle k .

6.4.3 Kinetic energy

For all rigid bodies with $k = 1, \dots, n$ (Fig. 14), the angular velocity vector $\boldsymbol{\omega}_k$ and the linear velocity of the reference point, $\mathbf{v}_k = \dot{\mathbf{r}}_k$, are calculated recursively (Jain *et al.* (1993):

$$\boldsymbol{\omega}_k = \boldsymbol{\omega}_{p(k)} + \mathbf{e}_k \dot{\theta}_k \quad \text{and} \quad \mathbf{v}_k = \mathbf{v}_{p(k)} - (\mathbf{r}_k - \mathbf{r}_{p(k)}) \wedge \boldsymbol{\omega}_{p(k)}. \quad (23)$$

Denoting the vector from the reference point to the centre of mass of the rigid body k by \mathbf{Y}_k , its mass by m_k , and its inertia tensor by I_k (Fig. 14b), the kinetic energy is given by

$$E_{\text{kin}} = \frac{1}{2} \sum_{k=1}^n [m_k v_k^2 + \boldsymbol{\omega}_k \cdot I_k \boldsymbol{\omega}_k + 2\mathbf{v}_k \cdot (\boldsymbol{\omega}_k \wedge m_k \mathbf{Y}_k)]. \quad (24)$$

The inertia tensor I_k is a symmetric 3×3 matrix with elements (Arnold, 1978)

$$(\mathbf{I}_k)_{ij} = \sum_{\alpha} m_{\alpha} (|\mathbf{y}_{\alpha}|^2 \delta_{ij} - y_{\alpha i} y_{\alpha j}). \quad (25)$$

The sum runs over all atoms α with mass m_{α} in the rigid body k . \mathbf{y}_{α} is the vector from the reference point to the atom α , and δ_{ij} is the Kronecker symbol. Since the shape of a rigid body enters the equations of motion only by the inertia tensor and the centre of mass vector, it is not essential to derive these quantities from the masses and relative positions of the individual atoms that constitute the rigid

body, as in equation (25). In fact, the efficiency of the torsion angle dynamics algorithm can be improved by treating the rigid bodies as solid spheres of mass m_k and radius ρ centred at the reference points \mathbf{r}_k :

$$\mathbf{Y}_k = \mathbf{o} \quad \text{and} \quad I_k = \frac{2}{5}m_k\rho^2\mathbf{I}_3, \quad (26)$$

where \mathbf{I}_3 is the 3×3 unit matrix. In DYANA $\rho = 5 \text{ \AA}$ and $m_k = 10\sqrt{n_k}m_0$ are used, where n_k denotes the number of atoms in the rigid body k (not counting pseudoatoms), and $m_0 = 1.66 \times 10^{-27} \text{ kg}$ is the atomic mass unit. In this way, fast motions of light rigid bodies, for example hydroxyl protons, are slowed down, thereby permitting longer integration time steps. Equation (26) does not imply an approximation of the van der Waals interaction: the steric repulsion is still calculated for each individual atom pair.

6.4.4 Torsional accelerations

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

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_k}\right) - \frac{\partial L}{\partial \theta_k} = \mathbf{o} \quad (k = 1, \dots, n), \quad (27)$$

with the Lagrange function $L = E_{\text{kin}} - E_{\text{pot}}$ (Arnold, 1978). They lead to equations of motion of the form

$$M(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) = \mathbf{o}. \quad (28)$$

In the case of 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 (Mazur & Abagyan, 1989; Mazur *et al.* 1991). However, to integrate the equations of motion, equation (28) 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.* (1993) is implemented to compute the torsional accelerations, which makes explicit use of the aforementioned tree structure of the molecule in order to obtain $\ddot{\theta}$ with a computational effort that is only proportional to n .

The algorithm of Jain *et al.* (1993) is initialized by calculating for all rigid bodies, $k = 1, \dots, n$, the six-dimensional vectors

$$a_k = \begin{bmatrix} (\boldsymbol{\omega}_k \wedge \mathbf{e}_k)\dot{\theta}_k \\ \boldsymbol{\omega}_{p(k)} \wedge (\mathbf{v}_k - \mathbf{v}_{p(k)}) \end{bmatrix}, \quad e_k = \begin{bmatrix} \mathbf{e}_k \\ \mathbf{o} \end{bmatrix} \quad \text{and} \quad z_k = \begin{bmatrix} \boldsymbol{\omega}_k \wedge \mathbf{I}_k \boldsymbol{\omega}_k \\ (\boldsymbol{\omega}_k \cdot m_k \mathbf{Y}_k) \boldsymbol{\omega}_k - \boldsymbol{\omega}_k^2 m_k \mathbf{Y}_k \end{bmatrix}, \quad (29)$$

and the 6×6 matrices

$$P_k = \begin{bmatrix} I_k & m_k \mathbf{A}(\mathbf{Y}_k) \\ -m_k \mathbf{A}(\mathbf{Y}_k) & m_k \mathbf{I}_3 \end{bmatrix} \quad \text{and} \quad \phi_k = \begin{bmatrix} \mathbf{I}_3 & \mathbf{A}(\mathbf{r}_k - \mathbf{r}_{p(k)}) \\ \mathbf{o}_3 & \mathbf{I}_3 \end{bmatrix}. \quad (30)$$

\mathbf{o}_3 is the 3×3 zero matrix, and $\mathbf{A}(\mathbf{x})$ denotes the antisymmetric 3×3 matrix associated with the cross product, i.e. $\mathbf{A}(\mathbf{x})\mathbf{y} = \mathbf{x} \wedge \mathbf{y}$ for all vectors \mathbf{y} .

Next, a number of auxiliary quantities is calculated by executing a recursive loop over all rigid bodies in the backward direction, $k = n, n-1, \dots, 1$:

$$\left. \begin{aligned} D_k &= e_k \cdot P_k e_k \\ G_k &= P_k e_k / D_k \\ \epsilon_k &= e_k \cdot (\mathcal{z}_k + P_k a_k) - \frac{\partial V}{\partial \theta_k} \\ P_{p(k)} &\leftarrow P_{p(k)} + \phi_k (P_k - G_k e_k^T P_k) \phi_k^T \\ \mathcal{z}_{p(k)} &\leftarrow \mathcal{z}_{p(k)} + \phi_k (\mathcal{z}_k + P_k a_k + G_k \epsilon_k) \end{aligned} \right\} \quad (31)$$

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, $k = 1, \dots, n$:

$$\left. \begin{aligned} \alpha_k &= \phi_k^T \alpha_{p(k)} \\ \ddot{\theta}_k &= \epsilon_k / D_k - G_k \cdot \alpha_k \\ \alpha_k &\leftarrow \alpha_k + e_k \ddot{\theta}_k + a_k \end{aligned} \right\} \quad (32)$$

The auxiliary quantities α_k are six-dimensional vectors, with α_0 being equal to the zero vector. A proof of the correctness of this algorithm can be found in Jain *et al.* (1993). Equations (29)–(32) also show why the computation of the torsional accelerations requires an effort that is directly proportional to the number of torsion angles: the algorithm consists of a sequence of three linear loops over the rigid bodies (i.e. torsion angles); all three loops involve for each torsion angle only the calculation of quantities that are independent of the system size (e.g. scalars, six-dimensional vectors, and 6×6 matrices).

6.4.5 *Integration of the equations of motion*

The integration scheme for the equations of motion in torsion angle dynamics (Mathiowetz *et al.* 1994) is a variant of the leap-frog algorithm used in Cartesian dynamics. In addition to the basic scheme of equations (15) and (16) the temperature is controlled by weak coupling to an external bath (Berendsen *et al.* 1984) and the time step is adapted based on the accuracy of energy conservation. A slight complication arises because, unlike the situation in Cartesian space dynamics where the accelerations are a function of the positions only, the torsional accelerations also depend on the velocities. These, however, are known in the leap-frog scheme only at half time steps, whereas the positions and accelerations are required at full time steps. The algorithm below therefore employs linear extrapolation from the two former values at half time step to obtain an estimate of the velocity after the full time step, $\dot{\theta}_e(t)$, which is used in the next integration step to calculate the torsional accelerations. It can be shown (Güntert *et al.* 1997) that the intrinsic accuracy of the velocity step remains of order $O(\Delta t^3)$, as in equation

(15). A time step $t \rightarrow t + \Delta t$ that follows a preceding time step $t - \Delta t' \rightarrow t$ is executed as follows:

1. On the basis of the torsional positions $\theta(t)$, calculate the Cartesian coordinates of all atoms (Katz *et al.* 1979; Güntert, 1993), the potential energy $E_{\text{pot}}(t) = E_{\text{pot}}(\theta(t))$, and the torques $-\nabla E_{\text{pot}}(t)$.

2. Adapt the torsional velocities $\dot{\theta}$ (both $\dot{\theta}(t - \Delta t'/2)$ and $\dot{\theta}_e(t)$) to maintain the temperature T^{ref} (Berendsen *et al.* 1984) and adjust the time step to attain a desired relative accuracy of energy conservation ϵ^{ref} :

$$\dot{\theta} = \dot{\theta}' \frac{1 + \frac{T^{\text{ref}} - T(t)}{\tau T(t)}}{1 + \frac{\epsilon^{\text{ref}} - \epsilon(t)}{\tau \epsilon(t)}}, \quad \text{and} \quad \Delta t = \Delta t' \frac{1 + \frac{\epsilon^{\text{ref}} - \epsilon(t)}{\tau \epsilon(t)}}{1 + \frac{T^{\text{ref}} - T(t)}{\tau T(t)}}, \quad (33)$$

where

$$T(t) = \frac{2E_{\text{kin}}(t)}{nk_B} \quad \text{and} \quad \epsilon(t) = \left| \frac{E(t) - E(t - \Delta t')}{E(t)} \right|, \quad (34)$$

respectively, are the instantaneous temperature and the relative change of the total energy, $E = E_{\text{kin}} + E_{\text{pot}}$, in the preceding time step. The time constant, $\tau \gg 1$, is a user-defined parameter, measured in units of the time step, with a typical value of $\tau = 20$; n denotes the number of torsion angles and $k_B = 1.38 \times 10^{-23} \text{ J K}^{-1}$ is the Boltzmann constant. Temperature and time step control can be turned off by setting $\tau = \infty$. To calculate $\epsilon(t)$ in equation (34), $E(t)$ is evaluated before velocity scaling is applied, whereas for $E(t - \Delta t')$ the value after velocity scaling in the preceding time step is used. Thus, the measurement of the accuracy of energy conservation is not affected by the scaling of velocities. An exact algorithm would yield $E(t) = E(t - \Delta t')$ and consequently $\epsilon(t) = 0$.

3. Calculate the torsional accelerations, $\ddot{\theta}(t) = \ddot{\theta}(\theta(t), \dot{\theta}_e(t))$, using equations (29)–(32).

4. Using the leap-frog scheme of equations (15) and (16) (with \mathbf{r} replaced by θ), calculate the new velocities at half time step, $\dot{\theta}(t + \Delta t/2)$, and the new torsional positions $\theta(t + \Delta t)$.

The algorithm is initialized by setting $t = 0$, $\Delta t' = \Delta t$, and the initial torsional velocities are chosen randomly corresponding to a given initial temperature.

Since for optimal efficiency in structure calculations with torsion angle dynamics the time steps are made as long as possible a safeguard against occasional strong violations of energy conservation by more than 10% in a single time step replaces such time steps by two time steps of half length.

6.4.6 Energy conservation and time step length

Energy conservation is a key feature of proper functioning of any molecular dynamics algorithm (Allen & Tildesley, 1987). The accuracy of energy conservation can be monitored by the standard deviation $\sigma_E = \sqrt{\langle (E - \langle E \rangle)^2 \rangle}$, or by the RMS change of the total energy between successive integration steps, $\delta_E = \sqrt{\langle \Delta E^2 \rangle}$, where $\langle \dots \rangle$ denotes the average over all time steps of a molecular dynamics run (Beeman, 1976; van Gunsteren & Berendsen, 1977). The parameter δ_E is closely related to $\epsilon(t)$ in equation (34) and probes the local error in one time

step of the integration algorithm, whereas the standard deviation δ_E is sensitive both to local errors and to slow drifts of the total energy over many time steps and is thus dependent on the length of molecular dynamics run. The dependence of σ_E and δ_E on the length of the integration time step Δt is plotted in Fig. 16 for a series of torsion angle dynamics runs performed with the experimental NMR data set for cyclophilin A (Ottiger *et al.* 1997) at temperatures of 10600, 390 and 1.2 K under conditions where the total energy should be conserved, i.e. with $\tau = \infty$ in equation (33) (Güntert *et al.* 1997). Fig. 16 shows that long time steps of up to 100 fs are tolerated by the algorithm, that σ_E is proportional to Δt^2 , as expected for Verlet-type integration algorithms (Verlet, 1967; Allen & Tildesley, 1987), and that δ_E is proportional Δt^3 . The most relevant result for practical applications is that long time steps – about 100, 30 and 7 fs at low, medium and high temperatures, respectively – can be used in torsion angle dynamics calculations with DYANA. The concomitant fast exploration of conformation space provides the basis for efficient structure calculation protocols. These time steps should, however, not be compared directly with those of 1–2 fs used in conventional Cartesian space molecular dynamics (Allen & Tildesley, 1987) because DYANA uses more uniform masses (equation (26)) and the much simpler potential energy function of equations (17)–(20) than standard molecular dynamics programs (Brooks *et al.* 1983; Cornell *et al.* 1995; van Gunsteren *et al.* 1996).

6.4.7 *Simulated annealing schedule*

The potential energy landscape of a protein is complex and studded with many local minima, even in the presence of experimental restraints in a simplified target function of the type of equation (17). 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, quite elaborated protocols have been devised for structure calculations using molecular dynamics in Cartesian space (Nilges *et al.* 1988*a*; Brünger, 1992). 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. The faster exploration of conformation space with torsion angle dynamics allows for simpler schedules. The standard simulated annealing protocol used by the program DYANA (Güntert *et al.* 1997) 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 four parts:

1. A short minimization to reduce high energy interactions that could otherwise disturb the torsion angle dynamics algorithm: 100 conjugate gradient minimization steps are performed at target level 3, i.e. including only distance restraints between atoms up to 3 residues apart along the sequence, followed by a further 100 minimization steps including all restraints. 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

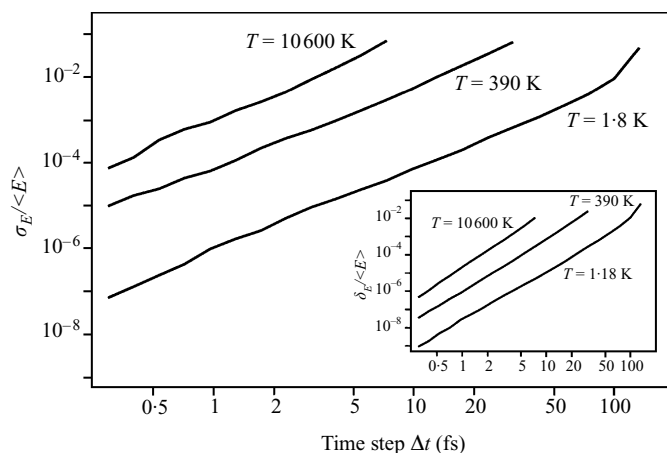


Fig. 16. Dependence of the standard deviation of the total energy, $\sigma_E = \sqrt{\langle (E - \langle E \rangle)^2 \rangle}$, on the length of the integration time step Δt in torsion angle dynamics calculations with the program DYANA using the experimental NMR data set for cyclophilin A (Güntert *et al.* 1997). Each run had a duration of 0.9 ps, and the initial temperatures were 10600, 390, and 1.18 K, respectively. The inset shows the RMS change of the total energy between successive integration steps, $\delta_E = \sqrt{\langle \Delta E^2 \rangle}$, for the same trajectories.

by 0.15 Å with respect to their standard values. The weights in the target function of equation (17) are set to 1 for user-defined upper and lower distance bounds, to 0.5 for steric lower distance bounds, and to 5 Å² for torsion angle restraints.

2. A torsion angle dynamics calculation at constant high temperature: One fifth of all N torsion angle dynamics steps are performed at a constant high reference temperature T_{high} , typically $T_{\text{high}} \approx 10000$ K. The time step is initialized to $\Delta t = 2$ fs, and the reference value for the relative accuracy of energy conservation to $\epsilon_0^{\text{ref}} = 0.005$.

3. A torsion angle dynamics calculation with slow cooling close to zero temperature: The remaining $4N/5$ torsion angle dynamics steps are performed with reference values for the temperature and the relative accuracy of energy conservation of

$$T^{\text{ref}}(s) = (1-s)^4 T_{\text{high}} \quad \text{and} \quad \epsilon^{\text{ref}}(s) = \epsilon_0^{\text{ref}} \times 0.02^s. \quad (35)$$

The parameter s varies linearly from 0 in the first to 1 in the last time-step.

4. The 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 restraints to 2, 100 conjugate gradient minimization steps are performed, followed by 200 torsion angle dynamics steps at zero reference temperature.

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. The temperature schedule, the effect of automatic time step adaptation according to equation (33), and the average structural change per time step of the simulated annealing protocol are illustrated in Fig. 17.

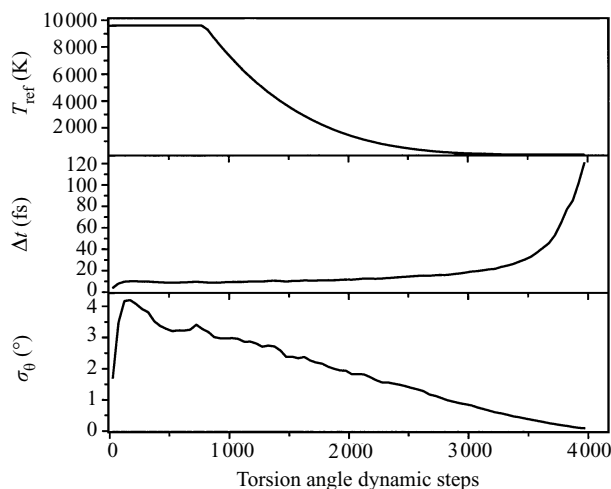


Fig. 17. Plots *versus* the number of torsion angle dynamics steps for a structure calculation using the standard DYANA simulated annealing protocol (see Section 6.4.7) with the experimental NMR data set for cyclophilin A. (a) Temperature of the heat bath to which the system is weakly coupled. (b) Integration time-step Δt which is adapted according to equation (33) in order to achieve a certain accuracy of energy conservation. Values of Δt are averaged over 50 time steps and 80 independently calculated conformers. (c) RMS torsion angle change between successive torsion angle dynamics steps, $\delta_\theta = \sqrt{\langle \Delta\theta^2 \rangle}$, averaged over 50 time steps and all rotatable torsion angles of 80 conformers.

Table 5. *Computation times (seconds) for DYANA structure calculations of the proteins BPTI and cyclophilin A on different computers^a*

Computer	BPTI	Cyclophilin A
NEC SX-4	13	36
DEC Alpha 8400 5/300	20	86
SGI Indigo2 R10000 (175 MHz)	23	127
IBM RS/6000-590	35	141
Cray J-90	44	141
Convex Exemplar	44	177
Hewlett-Packard 735	47	209

^a CPU times are for the calculation of one conformer, using the experimental NMR data sets (Berndt *et al.* 1992; Ottiger *et al.* 1997) and the standard DYANA simulated annealing protocol with 4000 torsion angle dynamics steps.

6.4.8 Computation times

With the torsion angle dynamics algorithm of equations (17)–(35) 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 below 1 min for a small protein like BPTI and less than 3.5 min for cyclophilin A on a wide array of generally available computers (Table 5). These

figures are much lower than those for the variable target function method using redundant torsion angle restraints, or for the torsion angle dynamics and Cartesian space molecular dynamics protocols implemented in the program XPLOR (Stein *et al.* 1997), and show that an improvement of the efficiency of structure calculation by more than an order of magnitude can be achieved.

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 (Güntert *et al.* 1997).

6.4.9 Application to biological macromolecules

Table 6 summarizes structure calculations performed with the torsion angle dynamics algorithm implemented in the program DYANA for six different proteins and a RNA molecule (Güntert *et al.* 1997). For five proteins the experimental NMR data sets were used. These proteins, with sizes ranging from 58 to 165 amino acid residues, represent different topologies and different qualities of input data sets that can be obtained by presently used homo- and heteronuclear NMR measurements. They are complemented by two calculations based on data sets that were simulated from the three-dimensional structures of a large protein of 394 residues and of an RNA structure with 32 nucleotides in order to demonstrate the potential of torsion angle dynamics for structure calculations of larger molecules, for which experimental NMR data sets may become available in the future, and for nucleic acids. NOE distance restraints with an upper bound 0.5 Å larger than the actual distance were generated for all proton-proton pairs (excluding OH and SH) closer than 4.5 Å. Restraints with methyl groups were referred to pseudo atoms, and stereospecific assignments were assumed only for the methyl groups of Val and Leu. Torsion angle restraints of $\pm 60^\circ$ about the value in the structure were simulated for ϕ , ψ and χ^1 in PGK, and for all angles in APK, where the tolerance for torsion angles in sugar rings was reduced to $\pm 5^\circ$ (Güntert *et al.* 1997). Flexibility of the sugar rings was achieved by 'cutting' the bond between C4' and O4' and imposing distance constraints to fix the length of the C4'-O4' bond and the corresponding bond angles.

Structure calculations were performed using the standard simulated annealing schedule of the program DYANA (see Section 6.4.7) with $N = 4000$ torsion angle dynamics steps. For each system 40 acceptable conformers were computed. The results in Table 6 show that this calculation protocol was appropriate for all seven systems, yielding efficiently structures with small restraint violations.

Two factors determine the overall efficiency of an NMR structure calculation: the time needed to calculate one conformer (see Section 6.4.8), and, equally important, the 'success rate', i.e. the percentage of conformers that reach small restraint violations as manifested by low target function values. The success rates for the structures in Table 6 are between 45 and 88%, and the relevant computation times per accepted conformer range from 23 s for the smallest to less than 9 min for the largest system on a commonly available computer.

The structure calculation with a simulated NMR data set for the 394-residue

Table 6. Structure calculations using the torsion angle dynamics algorithm of the program DYANA^a

Quantity ^b	BPTI	Antp	ADB	PrP	Cyp	PGK	APK
Size and experimental input data							
Residues	58	68	81	113	165	394	32
Torsion angles	241	351	379	521	722	1778	318
Upper distance bounds	651	894	795	1598	4093	14161	929
Torsion angle restraints	115	171	168	256	371	1117	284
Restraints/torsion angle	3.2	3.0	2.5	3.4	6.2	8.6	3.8
Structure calculation							
Accepted conformers (%)	88	75	88	47	58	63	45
Computation time (s) ^c	23	36	37	101	159	521	72
Target function (Å ²)	0.33	0.94	0.60	2.19	2.19	3.99	0.32
Backbone RMSD (Å) ^d	0.37	0.38	0.92	1.11	0.68	1.62	3.08
Average or sum ^e of restraint violations							
Upper bounds (Å)	0.004	0.005	0.003	0.006	0.003	0.001	0.001
Steric lower bounds (Å)	1.2	2.5	1.9	5.2	5.1	11.0	0.8
Torsion angle restraints (°)	0.04	0.11	0.04	0.10	0.03	0.01	0.03
Maximal restraint violations							
Upper bounds (Å)	0.19	0.28	0.25	0.41	0.44	0.47	0.22
Steric lower bounds (Å)	0.09	0.17	0.20	0.25	0.23	0.39	0.13
Torsion angle restraints (°)	1.6	4.5	2.7	5.3	3.7	6.9	1.4

^a Structure calculations were performed using the standard simulated annealing protocol of the program DYANA (Güntert *et al.* 1997) with 4000 torsion angle dynamics steps and the experimental NMR data sets for the following proteins: BPTI: basic pancreatic trypsin inhibitor (Berndt *et al.* 1992); Antp: *Antp*(C39S) homodomain (Güntert *et al.* 1991*b*); ADB: activation domain of porcine procarboxypeptidase B (Vendrell *et al.* 1991); PrP: mouse prion protein domain PrP(121–231) (Riek *et al.* 1996); Cyp: human cyclophilin A (Ottiger *et al.* 1997). Additional structure calculations were performed with simulated NMR data sets for the protein 3-phosphoglycerate kinase (PGK, Davies *et al.* 1994) and for the RNA pseudoknot APK (Kang *et al.* 1996). For each protein 40 acceptable conformers were calculated. Conformers were accepted if their target function value was below a cutoff of 1.25 Å² for BPTI, 1.93 Å² for Antp, 1.77 Å² for ADB, 3.02 Å² for PrP, 4.02 Å² for Cyp, 9.03 Å² for PGK, and 1.30 Å² for APK (Güntert *et al.* 1997).

^b Averaged over all accepted conformers, when applicable.

^c CPU times per *accepted* conformer, measured on DEC Alpha 8400 5/300 computers.

^d For the backbone atoms of the following residues: BPTI, 3–55; Antp, 8–60; ADB, 11–76; PrP, 124–168 and 173–227; Cyp, PGK and APK, all residues.

^e Average violation is given for experimental restraints, sum for steric restraints. The average violation equals the sum of violations divided by the number of restraints.

protein PGK was included in Table 6 as an example of a larger molecule, for which experimental NMR data sets may become available in the future. The results show that the same simulated annealing schedule as for smaller proteins could be used, and that the yield of acceptable conformers did not differ significantly from that of the smaller molecules (Table 6). A comparison of the PGK structure calculated using DYANA with the X-ray structure from which the simulated NMR data set was derived (Davies *et al.* 1994) shows that the two structures coincide closely, with an RMSD bias of the DYANA structure bundle from the X-ray structure of 1.05 Å for all backbone atoms N, C^α and C^β, which is significantly smaller than the RMSD radius of 1.62 Å for the bundle of accepted DYANA conformers. In light of these results, the structure calculation is not expected to become a bottleneck for future NMR structure determinations of proteins with up to 400 residues.

NMR structure determination of DNA or RNA molecules is notoriously difficult because the network of NOE distance restraints in nucleic acids is intrinsically less dense than in proteins (Wüthrich, 1986; Wijmenga *et al.* 1993; Pardi, 1995; Varani *et al.* 1996). For instance, the number of simulated distance restraints per degree of freedom is 2.7 times lower for the pseudoknot RNA APK than for the protein PGK in the data sets of Table 6, even though the same criteria were applied to derive NOE distance restraints from the structures. For this reason, structure calculation programs based on Cartesian space molecular dynamics and metric-matrix distance geometry may have difficulties to find nucleic acid conformers that satisfy the experimental data (Stein *et al.* 1997) unless *ad hoc* assumptions about the three-dimensional structure are made, usually by starting the calculations from standard A- or B-type duplex conformations. Stein *et al.* (1997) showed for a 12-base pair DNA duplex that torsion angle dynamics was successful in finding structures that satisfy the experimental restraints in a situation where metric matrix distance geometry and Cartesian space molecular dynamics did not lead to acceptable results.

A structure calculation for the 32-nucleotide pseudoknot RNA (Kang *et al.* 1996) was included in Table 6 to demonstrate that torsion angle dynamics as implemented in DYANA is able to calculate RNA structures starting from conformers with random torsion angle values. The results of the structure calculations show that the standard DYANA simulated annealing protocol for proteins is also adequate for nucleic acid structure calculations, indicating that there is no fundamental difference between the two classes of molecules from the point of view of structure calculation using torsion angle dynamics. In particular, the DYANA calculations do not need to start from well-defined structures, which could introduce a bias into the final result of the structure calculation.

These calculations underline that torsion angle dynamics in its implementation in the program DYANA is a powerful method for the calculation of protein and nucleic acid structures from NMR data that represents a significant advance over the other commonly used methods (see Sections 6.1–6.3), and – as judged by comparison with the available literature data (Table 6 of Stein *et al.* 1997) – other presently available implementations of torsion angle dynamics algorithms for

structure calculation of biological macromolecules from NMR data. Considering its high efficiency for exploring the conformation space of biological macromolecules, future applications of DYANA might well also include problems in molecular modeling, tertiary structure prediction and protein folding.

6.5 Other algorithms

Biomolecular structure calculation has been approached by a number of other methods that have, however, not gained wide-spread use, be it because they employ an algorithm that turned out to be unsuitable to solve the problem, because they were designed for special situations, or because their further development has been stopped for some reason. Examples include a 'heuristic method' (Altman & Jardetzky, 1986, 1989), the ellipsoid algorithm (Billeter *et al.* 1987), Monte Carlo methods (Levy *et al.* 1989; Ulyanov *et al.* 1993; Abagyan & Totrov, 1994), a multiconformational search algorithm for peptides with inherent flexibility (Brüschweiler *et al.* 1991), an algorithm based on 'optimal filtering' (Koehl *et al.* 1992), and the combination of metric matrix distance geometry with a genetic algorithm (van Kampen *et al.* 1996).

7. STRUCTURE ANALYSIS

7.1 Restraint violations

At the end of a structure calculation, the immediate question arises whether the structure calculation was successful, i.e. whether the algorithm was able to find structures that fulfil the given restraints, and, if not, which are the restraints that could not be satisfied. Therefore an analysis of the residual restraint violations seen in the final conformers is performed, which is usually summarized in a table (Table 7). In addition, a list of residual restraint violations that indicates, for each violation separately, the individual conformers where the violation occurs can reveal consistent violations, and distinguish them from insignificant violations resulting from the occurrence of different local minima in different conformers. Consistent violations most likely point to an inconsistency of the input data rather than to a convergence problem of the structure calculation algorithm.

As an example, Fig. 18 shows the overview output file of the program DYANA from a structure calculation with the experimental NMR data set of cyclophilin A. Restraint violations are analysed in the 20 conformers with lowest final target function value that were chosen to represent the solution structure of the protein. The small size and small number of residual restraint violations show that the input data represents a self-consistent set, and that the restraints are well satisfied in all 20 conformers. The table of violated restraints in the lower part of Fig. 18 is typical for the situation of a self-consistent input data set. There are mostly isolated violations that occur in one or very few conformers; only three distance restraints are violated in more than a third of the conformers. This situation indicates the absence of any detectable serious problems in the input data set. If

Table 7. *Experimental data and structural statistics for cyclophilin A*

Quantity ^a	Value	Comments
Resonance assignments		
Sequence-specific ^b	88 %	¹ H 88 %, ¹³ C 91 %, ¹⁵ N, 78 %
Stereospecific and individual NH ₂	73	46 CH ₂ , 16 C(CH ₃) ₂ , 11 NH ₂
Experimental restraints		
Upper distance bounds	4093	
Torsion angle restraints	371	135 ϕ , 135 ψ , 101 χ ¹
Structure calculation		
Target function	1.28 Å ²	Range 0.94–1.66 Å ²
RMSD radius	0.53 Å	For N, C ^α , C' of all residues
Maximal restraint violations		
Upper bounds	0.32 Å	Average violation 0.0021 Å
Steric lower bounds	0.19 Å	Sum of violations 3.5 Å
Torsion angle restraints	2.6°	Average violation 0.016°

^a Averaged over all accepted conformers, when applicable.

^b Includes all protons and methyl groups with the exception of hydroxyl protons, and all ¹³C and ¹⁵N atoms with a directly bound proton.

restraints had been found that were violated in all or almost all conformers, this finding would be the start point for checking their assignment and volume integration.

7.2 Atomic root-mean-square deviations

The standard measure used to quantify differences between three-dimensional structures is the root-mean-square deviation (RMSD) for a given set of corresponding atoms (McLachlan, 1979).

For two sets of n atoms each, $\mathbf{r}_1, \dots, \mathbf{r}_n$ and $\mathbf{q}_1, \dots, \mathbf{q}_n$, with $\sum_i \mathbf{r}_i = \sum_i \mathbf{q}_i = \mathbf{0}$, the RMSD is defined as the root mean square distance between the positions of corresponding atoms after optimal superposition of the two structures:

$$\text{RMSD} = \min_R \sqrt{\frac{1}{n} \sum_{i=1}^n |\mathbf{r}_i - R\mathbf{q}_i|^2}, \quad (36)$$

where R denotes a rotation matrix, and the minimum over all possible rotation matrices is taken. To find the optimal rotation matrix R the 3×3 matrix B with elements

$$B_{kl} = \frac{1}{n} \sum_{i=1}^n q_{ik} r_{il}, \quad (37)$$

where q_{ik} and r_{il} denote the components k and l of the position vectors \mathbf{q}_i and \mathbf{r}_i , respectively, is computed and decomposed by singular value decomposition (Press *et al.* 1986) into a product $B = UWV^T$ of a matrix U with orthonormal column vectors, a diagonal matrix $W = \text{diag}(\omega_1, \omega_2, \omega_3)$ with $\omega_1 \geq \omega_2 \geq \omega_3 \geq 0$, and a

Overview:

```

Number of structures      :      20
Cutoff for upper limits  :      0.20 Å
      lower limits       :      0.20 Å
      van der Waals      :      0.20 Å
      angle constraints   :      5.00 deg

```

struct	target function	upper limits			lower limits			van der Waals			torsion angles		
		#	sum	max	#	sum	max	#	sum	max	#	sum	max
1	0.94	1	6.6	0.35	0	0.0	0.00	0	2.8	0.15	0	2.2	1.7
2	0.98	2	7.3	0.25	0	0.0	0.00	0	2.6	0.16	0	3.8	2.4
3	1.03	4	8.6	0.27	0	0.0	0.00	0	3.2	0.12	0	5.1	1.5
4	1.08	3	8.1	0.34	0	0.0	0.00	0	3.4	0.18	0	4.6	1.5
5	1.13	2	9.4	0.24	0	0.0	0.00	0	3.5	0.12	0	5.5	2.2
6	1.17	1	6.2	0.22	0	0.0	0.00	2	3.3	0.34	0	6.9	4.2
7	1.21	2	7.1	0.26	0	0.0	0.00	1	3.2	0.28	1	9.1	7.3
8	1.22	3	9.1	0.35	0	0.0	0.00	0	3.2	0.12	0	5.6	2.7
9	1.25	3	9.4	0.35	0	0.0	0.00	0	3.2	0.17	0	2.9	1.4
10	1.27	5	9.1	0.36	0	0.0	0.00	0	3.1	0.15	0	7.0	2.0
11	1.30	2	9.9	0.22	0	0.0	0.00	0	4.5	0.15	0	5.8	3.3
12	1.30	2	8.5	0.39	0	0.0	0.00	2	3.2	0.25	0	4.5	2.9
13	1.36	5	9.8	0.35	0	0.0	0.00	0	3.5	0.14	0	2.2	0.9
14	1.40	6	9.0	0.42	0	0.0	0.00	0	3.1	0.18	0	4.7	2.3
15	1.43	4	10.5	0.25	0	0.0	0.00	1	4.2	0.24	0	4.9	1.5
16	1.43	2	9.5	0.38	0	0.0	0.00	0	4.4	0.16	0	9.9	2.2
17	1.44	2	9.3	0.42	0	0.0	0.00	1	3.8	0.24	0	7.1	3.6
18	1.53	4	10.1	0.39	0	0.0	0.00	0	4.0	0.12	0	9.8	2.1
19	1.56	3	8.4	0.29	0	0.0	0.00	2	4.2	0.38	1	10.5	6.1
20	1.66	7	10.6	0.42	0	0.0	0.00	0	3.8	0.16	0	4.0	1.0
Average	1.28	3	8.8	0.32	0	0.0	0.00	0	3.5	0.19	0	5.8	2.6
+/-	0.19	2	1.2	0.07	0	0.0	0.00	1	0.5	0.07	0	2.4	1.6
Minimum	0.94	1	6.2	0.22	0	0.0	0.00	0	2.6	0.12	0	2.2	0.9
Maximum	1.66	7	10.6	0.42	0	0.0	0.00	2	4.5	0.38	1	10.5	7.3

Constraint violation overview:

						max	1	5	10	15	20	
Upper QG	MET	1	-	QB	ALA	26	0.23				*	
Upper HN	ASN	3	-	HB2	ASN	3	0.22	*				
Upper HN	ASN	3	-	HB3	ASN	3	0.42	++		+	*	
Upper HA	ASN	3	-	HD2	PRO	4	0.28	*				
Upper HN	THR	5	-	HN	GLU-	165	0.22		*			
Upper HN	LEU	24	-	HB3	LEU	24	0.22				*	
Upper HN	GLU-	43	-	HB2	GLU-	43	0.20			*		
Upper HA	THR	73	-	QG2	THR	73	0.42	+	+++	+++	+++	+
Upper HA	ILE	78	-	HN	GLY	80	0.25	*	+			
Upper QA	GLY	80	-	HN	LYS+	82	0.22	*				
Upper QB	GLU-	81	-	HN	LYS+	82	0.21			*		
Upper HN	LYS+	82	-	HB2	LYS+	82	0.21	*				
Upper HN	LYS+	82	-	HB3	LYS+	82	0.27			*		
Upper HN	LYS+	82	-	QB	LYS+	82	0.25		+		*	
Upper HN	LEU	90	-	HN	LYS+	91	0.26		+	++	++	
Upper QD2	LEU	98	-	HN	PHE	129	0.25		+	++	*	
Upper HN	SER	99	-	HN	PHE	129	0.27			*		
Upper HA1	GLY	104	-	HD2	PRO	105	0.23		*	+		
Upper HN	ASN	106	-	HN	THR	107	0.23	+				*
Upper HN	ASN	106	-	HN	ASN	108	0.29		+	+	+	++
Upper HA	CYS	115	-	HN	ALA	117	0.27	++++	+	+	+	+++
Upper HB2	LYS+	133	-	HN	GLU-	134	0.24			*		
Upper QG2	ILE	158	-	HG13	ILE	158	0.27		+	*		
Angle PSI	GLU-	81					7.35	*				
Angle CHI1	ASN	106					6.07				*	

Fig. 18. Overview output file of the program DYANA produced at the end of a structure calculation using the experimental NMR data set for the protein cyclophilin A (Ottiger *et al.* 1997). The 20 out of 50 conformers with lowest final target function values were analysed. The upper table shows, with one row for each conformer, the target function value, followed by three restraint violation measures – the number of violations that exceeded the corresponding cutoff given at the top, the sum of violations, and the maximal violation – for each of the four types of restraints: upper distance bounds, lower distance bounds (not used in this calculation), steric lower distance bounds, and torsion angle restraints. Average values, standard deviations, minima and maxima of these quantities are given below. The second part of the file lists all restraint violations larger than the cutoffs and identifies the conformer(s) in which they occur ('+' and '*' signs in the columns on the right).

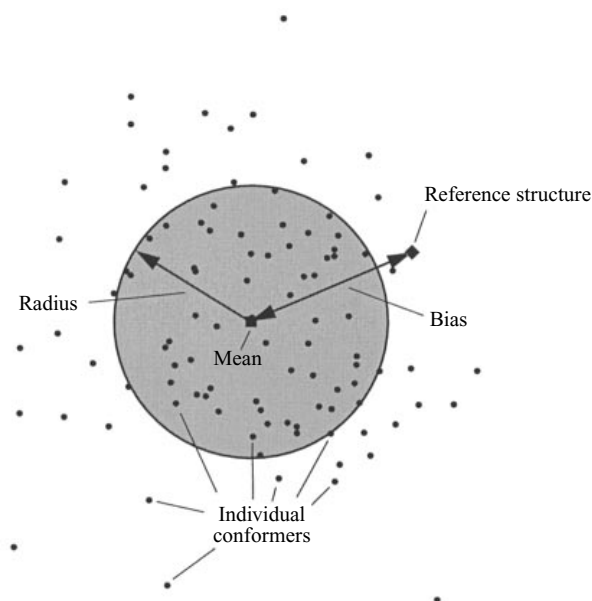


Fig. 19. RMSD radius of a bundle of conformers, and RMSD bias of a bundle of conformers with respect to a reference structure. The RMSD radius is the average of the RMSD values between each individual conformer of the bundle and its mean coordinates. The RMSD bias is the RMSD value between the mean coordinates of the bundle and a reference structure. The mean coordinates of a bundle of n conformers are obtained by superimposing for minimal RMSD the conformers 2, ..., n onto the first conformer and then averaging the Cartesian coordinates.

transposed orthogonal matrix V . The rotation matrix R that minimizes the expression in equation (36) is then given by

$$R = U \text{diag}(1, 1, s) V^T, \quad (38)$$

where $s = \pm 1$ denotes the sign of the determinant of the matrix B and ensures that R is a pure rotation with determinant $+1$ (McLachlan, 1979). Without this precaution incorrect (too small) RMSD values result if the mirror image of a structure yields a better superposition than the structure itself.

The optimal superposition according to equation (36) is also used for the simultaneous display of several conformers on a molecular graphics system (Koradi *et al.* 1996). In practice, RMSD values are usually calculated for the backbone atoms N, C $^\alpha$ and C', or for all heavy (i.e. non-hydrogen) atoms of the residues with well-defined conformation, excluding, for instance, chain termini and loops that are unstructured in solution.

When RMSD values are used to measure the spread among the m conformers in a structure bundle, the quantity with the most intuitive meaning is the 'RMSD radius' (Fig. 19), defined as the average of the m pairwise RMSD values between the individual conformers and their mean structure. To compute the mean

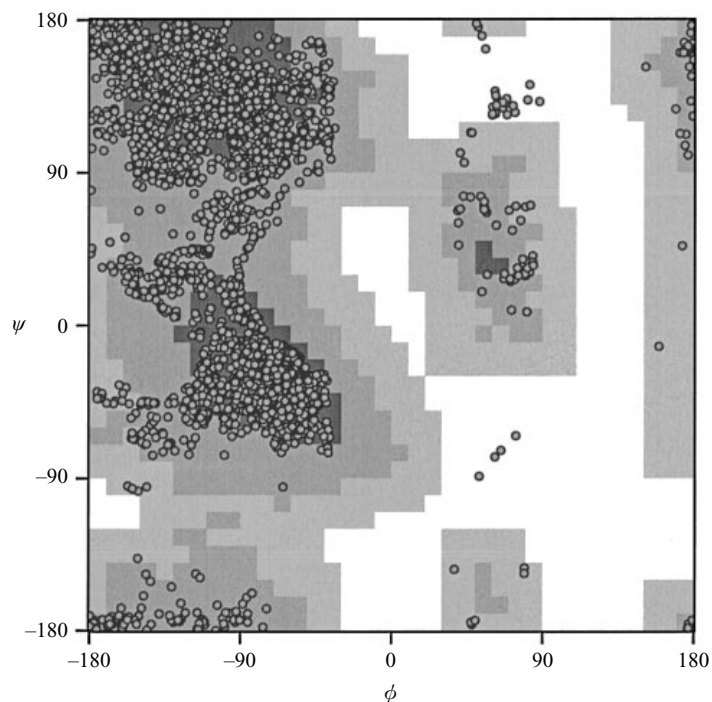


Fig. 20. Ramachandran plot for a bundle of 20 conformers of the protein cyclophilin A. The structures were calculated on the basis of the experimental NMR data set for cyclophilin A (Ottiger *et al.* 1997) with the torsion angle dynamics algorithm of the program DYANA (Güntert *et al.* 1997), which was also used to create the Ramachandran plot. Each circle corresponds to the ϕ/ψ values of a non-glycine residue in one of the 20 conformers. The most favourable, additionally allowed, generously allowed, and disallowed regions according to the program PROCHECK (Laskowski *et al.* 1996) are represented by dark, medium, light and no shading, respectively.

structure, the conformers are superimposed for minimal RMSD onto the first conformer, and the arithmetic mean of the corresponding Cartesian coordinates is taken. As an alternative to the RMSD radius, the average of the $m(m-1)/2$ pairwise RMSD values among the individual conformers can be reported. This quantity is in general about 1.4 times larger than the corresponding RMSD radius. The deviation of a structure bundle from a given 'external' reference structure is most easily described by the RMSD value between the mean coordinates of the bundle and the reference structure, the 'RMSD bias' (Fig. 19). RMSD radius and RMSD bias have the intuitive geometric meaning of the radius of a tube containing the structure bundle and the distance between the centre of the tube and the reference structure, respectively (Fig. 19).

It cannot be overemphasized that a small RMSD value is not by itself indicative of a high-quality structure. It neither conveys any information about the consistency with the experimental data nor does it necessarily correspond to the conformation space that is really allowed by the conformational restraints because

the sampling of conformation space by the structure calculation algorithm may be biased, i.e. there would exist structures that agree with the data but differ significantly from those resulting from the structure calculation. This may be due to the limited statistics – typically only about 20 conformers are analysed – or to an inherent deficiency of the structure calculation algorithm (Metzler *et al.* 1989).

Displacements (Billeter *et al.* 1989) are a generalization of the RMSD values, since the set of atoms used for the superposition of the conformers, M_{sup} , differs from the set of atoms for which the root mean square deviation of the positions is actually calculated, M_{RMSD} . For example, for the evaluation of the backbone displacement $D_{\text{glob}}^{\text{bb}}$ of a given residue i after global superposition, M_{sup} consists of the backbone atoms N, C $^\alpha$, and C' of the residues used for global superposition, and M_{RMSD} of the backbone atoms N, C $^\alpha$ and C' of residue i . To evaluate local backbone displacements $D_{\text{loc}}^{\text{bb}}$ for a residue i , M_{sup} consists of the backbone atoms N, C $^\alpha$ and C' of the residues $i-1$, i and $i+1$, and M_{RMSD} consists of the backbone atoms of residue i .

7.3 Torsion angle distributions

To analyse the conformation of the polypeptide chain on a local level, plots of the distributions of the individual values of the torsion angles ϕ , ψ and χ^1 versus the amino acid sequence of the protein and Ramachandran plots (Fig. 20) are convenient. They allow, for example, the identification of secondary structure elements, the classification of tight turns, and an assessment of the local precision of the structure determination.

To obtain the average value, $\bar{\phi}$, and the standard deviation, σ , from the torsion angle values, ϕ_1, \dots, ϕ_n , of the individual conformers, one has to take into account the periodicity of the torsion angles, for example by using

$$\bar{\phi} = \arg \sum_k e^{i\theta_k} \quad \text{and} \quad \sigma = \sqrt{-2 \log \left| \frac{1}{n} \sum_k e^{i\theta_k} \right|}. \quad (39)$$

The values defined by equation (39) have a clear meaning only when the torsion angle is well-defined. For example, the common situation that there are two groups of conformers, each with a well-defined value of the torsion angle, but with a large difference between the two groups, cannot be distinguished from the situation of a truly disordered torsion angle by means of equation (39).

7.4 Hydrogen bonds

Another important feature of protein structures are hydrogen bonds. They can be identified readily in the structure, for example by the criterion that the hydrogen-acceptor distance must be shorter than 2.4 Å and that the angle between the hydrogen, the atom to which the hydrogen is covalently bound, and the acceptor must be smaller than 35° (Fig. 9b; Billeter *et al.* 1990). The second condition ensures that the hydrogen bond is more or less linear.

In accord with what was said about the analysis of restraint violations,

significance should only be attributed to hydrogen bonds that occur consistently in a sizeable number of conformers in a structure bundle. Since the geometric force fields generally used for the structure calculation do not contain potentials that favour the formation of hydrogen bonds, their abundance, counted according to the criterion of Fig. 9, often increases considerably during a subsequent energy refinement. On the other hand, the use of electrostatic potentials *in vacuo* tends to yield spurious intra-protein hydrogen bonds (and salt bridges) on the surface of the macromolecule (Guenot & Kollman, 1992; Luginbühl *et al.* 1996), and should be avoided.

7.5 Molecular graphics

Molecular graphics programs are an indispensable tool to visualize and analyse NMR structures. Thanks to real-time transformations, stereo displays and ray-tracing techniques, three-dimensional impressions close to those of real, physically built models can be produced routinely nowadays. Many molecular graphics programs are available, for example the academic packages MIDAS (Ferrin *et al.* 1988), GRASP (Nicholls *et al.* 1991), MOLSCRIPT (Kraulis, 1991), RASMOL (Sayle & Milner-White, 1995), MOLMOL (Koradi *et al.* 1996), and the commercial products INSIGHT (Molecular Simulations, Inc.) and SYBYL (TRIPOS, Inc.).

The program MOLMOL (Koradi *et al.* 1996) is unique in the sense that it has been designed especially for work with NMR structures that are represented by bundles of conformers, which are often handled awkwardly by other programs. In addition to high-quality molecular graphics, a wide array of structure analyses can be performed with MOLMOL, including the calculation and display of mean structures, restraint violations, hydrogen bonds, RMSDs, torsion angle distributions, Ramachandran plots, and electrostatic surfaces. Fig. 21 presents four different representations of the solution structure of cyclophilin A that were produced with the program MOLMOL: A schematic view (Fig. 21 *a*) that emphasizes the secondary structure elements: an eight-stranded antiparallel β -barrel that is closed by two amphipathic α -helices, and a short 3_{10} -helix of residues 120–122 (Ottiger *et al.* 1997); a bundle of ten conformers (Fig. 21 *b*); another schematic view that illustrates the local precision of the structure by virtue of a tube with variable diameter (Fig. 21 *c*); and an all-heavy-atom representation of one conformer overlaid with the dense network of distance restraints (Fig. 21 *d*).

7.6 Check programs

Programs have been developed to perform a ‘quality check’ of a structure determined by X-ray crystallography or NMR spectroscopy. The best known examples of such programs are PROCHECK (Laskowski *et al.* 1993), including its NMR-specific extension PROCHECK-NMR (Laskowski *et al.* 1996) and WHATIF (Vriend & Sander, 1993). These software packages try to assess the quality of a structure primarily by checking whether a number of different parameters are in agreement with their values in databases derived from high-resolution X-ray

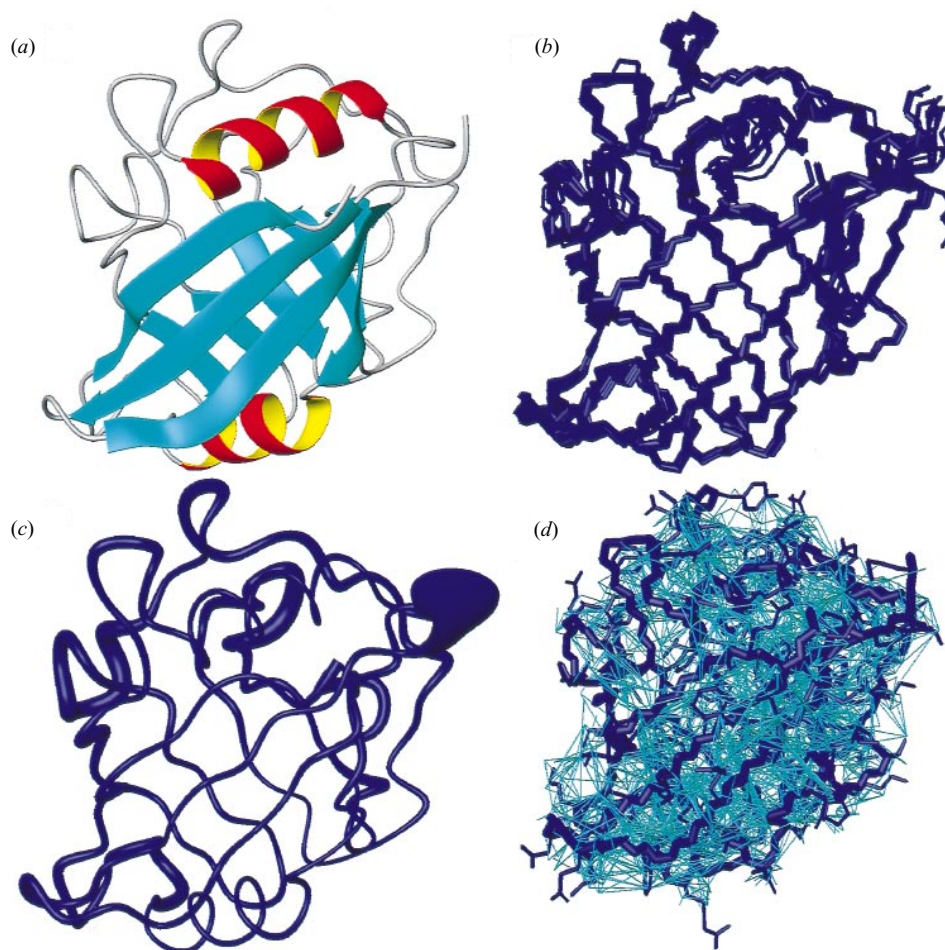


Fig. 21. Solution structure of the protein cyclophilin A, calculated using the torsion angle dynamics algorithm of the program DYANA on the basis of the experimental NMR data set collected by Ottiger *et al.* (1997). Displays produced with the program MOLMOL (Koradi *et al.* 1996): (a) Schematic representation highlighting α -helices and β -strands. (b) Superposition of the ten conformers with lowest target function values. Only bonds between the backbone atoms N, C $^{\alpha}$ and C' are drawn. (c) Another representation that affords an impression of the variable precision of different parts of the polypeptide backbone. The diameter of the hose-shaped object reflects the positional spread in the structure bundle among the corresponding backbone atoms. (d) One of the cyclophilin A conformers and the network of distance restraints used in the structure calculation. The structure is represented by dark cylinders for covalent bonds between heavy atoms; distance restraints are visualized by thin lines. Intraresidual and sequential distance restraints have been omitted for clarity.

structures. Examples of such parameters include: correct values for covalent bond lengths and bond angles (Engh & Huber, 1991), the percentage of residues with ϕ/ψ -values in the most favoured regions of the Ramachandran plot, the clustering of χ^1 -angles at the staggered rotamer positions, the overall quality of packing, the

absence of bad non-bonded contacts, the completeness of the hydrogen bonding network (i.e. a minimal number of atoms with unsatisfied hydrogen bonding capabilities in the core of the molecule), etc. Outliers of these quantities do not necessarily point to errors in the structure – they occur, albeit rarely, also in X-ray structures solved to very high resolution – but should be checked meticulously to rule out a possible misinterpretation of the experimental data. In addition, check programs like PROCHECK-NMR (Laskowski *et al.* 1996) can read experimental restraints in a variety of formats and provide measures for the agreement of the experimental restraints with the structure calculated from them in a way that is independent from the structure calculation program. Programs like WHATIF also look out for straightforward mistakes of the covalent structure, such as wrong chiralities, which seem to occur disquietingly often in protein (Hooft *et al.* 1996) and nucleic acid structures (Schultze & Feigon, 1996). Of course, that a structure fulfils the criteria of a check program does not guarantee it to be correct; most checks probe only local features of the conformation.

7.7 *A single, representative conformer*

The usual representation of an NMR structure as a bundle of conformers, each of which being an equally good fit to the data, provides a wealth of information about the conformational uncertainty, which may be correlated to true flexibility of the molecule. For example, alternative conformations of side-chains and complete loops may be realized in different conformers, a feature that is difficult, if not impossible, to represent in a single structure. Nevertheless, it is often desirable to provide, in addition to the bundle of conformers, a single representative structure that may be used in the same way as an X-ray structure, avoiding the bewildering amount of detail in the bundle, for example in pictures or in comparisons of the structures of different proteins.

Clearly, the Cartesian coordinates averaged over the conformers in the bundle (after suitable superposition) are no good choice: they lie exactly in the centre of the bundle, of course, but the averaging entails unacceptable distortions of the covalent geometry. The average coordinates are thus only used as a reference for the calculation of RMSD values, namely the RMSD radius of Fig. 19. Selecting just one of the conformers in the bundle is another straightforward possibility. In this case, the representative conformer has, by definition, the same quality as the bundle. The selection can be random or based on different criteria, for instance, smallest RMSD to the mean, smallest restraint violations, lowest conformational energy, highest coincidence with the network of consistent hydrogen bonds in the bundle, etc. Since all conformers in the bundle are essentially equivalent, the choice should not be crucial. In general, there will exist structures (not members of the bundle) that fulfil the restraints as well as those in the bundle but that lie closer to its centre than any of its individual members, and hence the representative conformer chosen from among them.

A procedure that can yield such a structure has been introduced by Clore *et al.* (1986a) and is used routinely when structures are determined by simulated

annealing in Cartesian space: From a bundle of conformers the mean structure is computed and subsequently regularized by restrained energy minimization. This results in general in a structure with good stereochemistry and in agreement with the experimental data that is significantly closer to the mean coordinates of the bundle than any of the individual conformers.

8. GENERAL ASPECTS OF NMR STRUCTURE CALCULATION

This chapter discusses a number of general aspects of NMR structure calculation on the basis of the experimental NMR data set for cyclophilin A (Ottiger *et al.* 1997) for which structure calculations by torsion angle dynamics were performed with the program DYANA (Güntert *et al.* 1997). An especially rich set of experimental restraints is available for cyclophilin A (Table 7) which affords a particularly suitable platform for these investigations.

Table 7 and Figs 18, 20 and 21 also show the results of a structure calculation with the complete data set that will serve as a reference for various investigations in this chapter. Fifty random start conformers were subjected to simulated annealing according to the standard schedule of DYANA (see Section 6.4.7), and the 20 conformers with lowest final target function value were chosen to represent the solution structure of the protein.

8.1 Ensemble size

NMR structure calculations are always performed by computing, using the same algorithm, many different conformers, each starting from another random initial conformation. Provided that the input data set is self-consistent (as will be assumed in the following), some of the conformers will be good solutions to the problem, i.e. exhibit small restraint violations, whereas others might be trapped in local minima. For this reason it is customary to compute an ensemble consisting of more conformers than needed, and to select among them the ‘best’ ones that will represent the solution structure of the molecule and be analysed further. Obviously, three choices have to be made in this process: How many conformers should be computed in the first place? How many conformers should be used to represent the solution structure? And how should these be selected from the ensemble of all conformers? The answer to the second question is simple: 20, or any other number that offers a reasonable compromise between sufficient statistics and manageability in graphics and analysis programs. With regard to the third question, it is clear that the selection of acceptable conformers should never rely on a measure of conformational spread, for instance the RMSD value, but be based on how well the experimental and steric restraints are fulfilled and, if the structure calculation program worked in Cartesian space, how close the covalent structure parameters are to their optimal values. Since the target function measures exactly these parameters, the most obvious selection and almost universally applied criterion is therefore to choose the N conformers with lowest

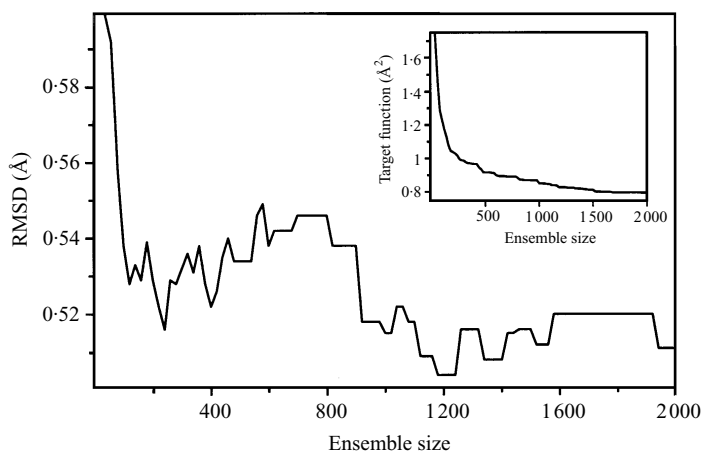


Fig. 22. Dependence of RMSD values on the size of the ensemble from which the 20 conformers with lowest target function values were selected at the end of a DYANA structure calculation. Using the experimental NMR data set for cyclophilin A (Ottiger *et al.* 1997) an ensemble of 2000 conformers was calculated using the standard simulated annealing protocol of the program DYANA with 8000 torsion angle dynamics steps. The inset shows the average final target function values of the 20 conformers with lowest target function values as a function of the ensemble size.

target function value, usually referred to as the ‘ N best conformers’. Alternative criteria, especially if related to the RMSD value or the presence of certain desirable features of the conformation, will inevitably produce a biased selection that neglects certain conformations that are in agreement with the data. All N conformers chosen should be acceptable in the sense that restraint violations are in a (subjectively defined) tolerance range, and it is desirable that the target function values do not vary strongly among them. In the absence of contradicting restraints this can be achieved by generating a large enough ensemble of conformers from which the best ones are taken. Depending on the protein, the data set, and on the structure calculation algorithm used, the distinction between acceptable and unacceptable conformers might be clear-cut, or gradual.

This brings us back to the first question: How many conformers should be computed? Obviously, this depends on the success rate of the algorithm used, and the requirements that are imposed on acceptable conformers. Under the conditions used for the structure calculations in Table 6 it would have been necessary to calculate between 1.14 and 2.2 times more conformers than were used to represent the solution structure of the molecule. However, the success rate depends on the protein and on the restraint data set and is unknown at the outset of the calculation. A common method is to calculate a fixed number of conformers, typically 2.5 times more than used later on. The question arises whether the final results of a structure determination depend crucially on such seemingly arbitrary decisions. Sometimes there is the belief that by selecting the best (as defined above) few conformers from a very large ensemble it would be possible to achieve

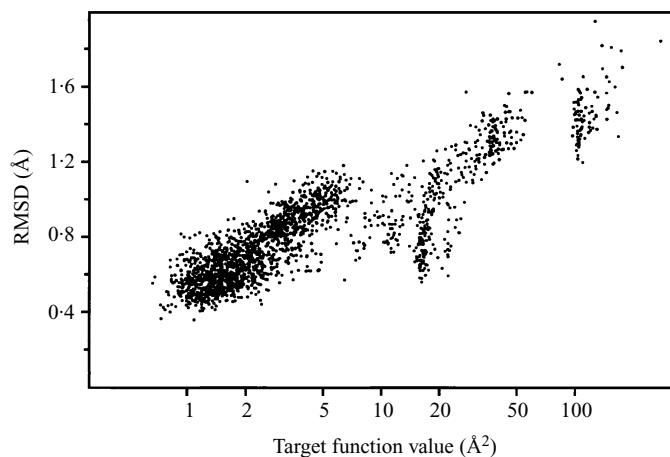


Fig. 23. Correlation between RMSD and final target function values in an ensemble of 2000 cyclophilin A conformers calculated using the standard simulated annealing protocol of the program *DYANA* with 8000 torsion angle dynamics steps. RMSD values are calculated for all backbone atoms of a given conformer relative to the average coordinates of the 20 conformers with lowest target function values in the ensemble.

arbitrarily low RMSD values. To address these questions, an ensemble of 2000 cyclophilin A conformers was produced with the program *DYANA* and the RMSD radius of the bundle of 20 best conformers selected out of the first M conformers (taken in the order in which they were computed) computed. The results, plotted in Fig. 22, show that after an initial drop of the RMSD value with increasing ensemble size, it exhibits only small fluctuations with no clear trend around a non-vanishing value. This behaviour of the RMSD radius roughly parallels that of the average target function value for the 20 best conformers (inset to Fig. 22) and indicates a correlation between target function and RMSD values within an ensemble of conformers, all calculated in the same way and from the same data. Fig. 23 depicts the RMSD (relative to the mean of the 20 best conformers) and final target function values of all 2000 conformers in the ensemble. There is a correlation between the two quantities if a wide range of target function values is considered, which, however, becomes weaker for the best conformers with target function value around 1 \AA^2 . As a side effect, clusters of points at high target function values in Fig. 23 indicate often occurring local minima.

8.2 Different NOE calibrations

The relationship between NOE intensity and upper distance bounds is usually defined by methods with more than a touch of heuristics (see Section 4.1). Nonetheless, the choice of calibration function(s) has a strong influence on the outcome of a structure calculation. To illustrate this, a series of structure calculations has been performed in which all upper distance limits in the experimental NMR data set of cyclophilin A have been scaled by constant factors

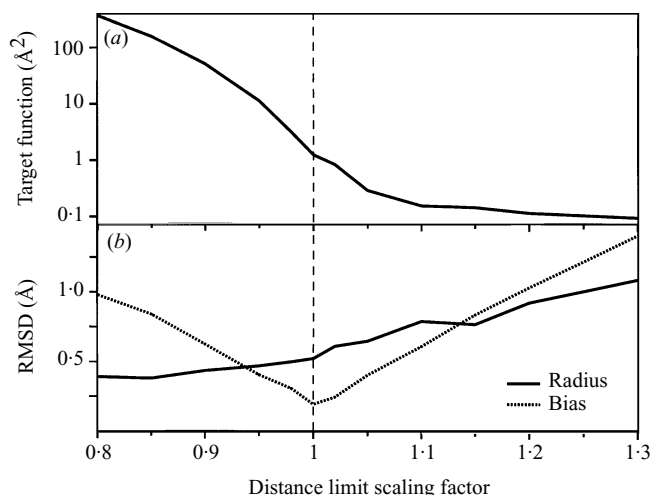


Fig. 24. Influence of scaling of the distance restraints on the outcome of structure calculations with the experimental NMR data set for cyclophilin A (Ottiger *et al.* 1997). All upper distance limits were scaled by the factor given on the horizontal axis. Ensembles of 50 conformers were calculated using the standard simulated annealing protocol of the program DYANA with 8000 torsion angle dynamics steps, and the 20 conformers with lowest target function values were analysed. (a) Average final target function values. (b) RMSD radius (solid), i.e. the average backbone RMSD values of the 20 conformers relative to their mean coordinates, and RMSD bias (dotted), i.e. the backbone RMSD value between the mean coordinates of the bundles obtained with a given scaling factor and without scaling.

in the range of 0.8 to 1.3 in order to mimic equivalent changes of the calibration constants k in equations (2) and (3). A scaling factor of one corresponds to the original experimental data set. The results, plotted in Fig. 24, show a strong increase of the target function values with decreasing distance bounds (note the logarithmic scale in Fig. 24a), and a less pronounced but clear increase of the RMSD radius with increasing scaling factor (Fig. 24b). The RMSD bias of the structure bundle obtained from scaled distance bounds relative to the mean coordinates of the original bundle monotonically increases from a minimum at scaling factor one in both directions (Fig. 24b). These findings indicate that target function and RMSD values have no absolute meaning but depend strongly on the NOE calibration used.

8.3 Completeness of the data set

The collection of an extensive set of NOE distance restraints constitutes a major part of the work involved in solving an NMR structure of a protein, and progressively more effort is required (and increasingly difficult decisions have to be taken) to assign additional NOEs, the more complete the data set becomes. In the case of cyclophilin A, the NOESY spectra were analysed as exhaustively as possible, resulting in a data set of about 25 relevant distance restraints per residue (Ottiger *et al.* 1997). Sometimes, however, such an effort might not be warranted,

Table 8. *DYANA* structure calculations for the protein Cyclophilin A using the complete experimental NMR data set and different subsets thereof^a

Data set ^b	Distance restraints	Success rate (%) ^c	Target function (Å ²)	Backbone RMSD (Å) ^d	
				Radius	Bias
All experimental restraints	4093	82	1.28	0.53	0.17
No stereospecific assignments	4394 ^e	80	1.29	0.61	0.29
No angle restraints	4093	66	1.55	0.53	0.24
% of all NOEs ^f					
75 %	3054	63	1.02	0.64	0.43
50 %	2055	65	0.72	0.79	0.70
25 %	1038	55	0.76	1.07	1.20
10 %	404	52	0.76	1.91	2.46
5 %	213	64	0.83	4.41	4.43
Only backbone and H ^β NOEs	1656	26	3.86	1.35	1.25
Only H ^N -H ^N NOEs	254	38	2.89	9.46	10.09

^a For each data set 50 conformers were calculated using the standard simulated annealing protocol of the program *DYANA* with 8000 torsion angle dynamics steps and a target function with linear asymptote for large violations. The 20 conformers with the lowest final target function values were analysed.

^b The different data sets were derived from the complete experimental NMR data set for Cyclophilin A (Ottiger *et al.* 1997) that comprises 4093 meaningful upper distance limits obtained from NOE measurements and 371 restraints for the torsion angles ϕ , ψ and χ^1 . The same torsion angle restraints were included in all data sets except the one without any torsion angle restraints.

^c Percentage of conformers with final target function values below $f_{\min} + 3.3 \text{ \AA}^2$, where f_{\min} is the lowest target function within the bundle (Güntert *et al.* 1997).

^d Radius: Average of the 20 RMSD values between each individual conformer and the mean coordinates of the bundle. Bias: RMSD value between the mean coordinates of the bundle and the mean coordinates of the bundle obtained with the complete experimental data set. The bias for the complete experimental data set was obtained by performing two structure calculations with different initial structures.

^e This number exceeds that for the complete experimental data set because in the absence of stereospecific assignments pairs of distance restraints to a diastereotopic pair might be replaced by three restraints, two identical ones to the diastereotopic atoms and one to the centrally located pseudoatom (see Section 5.3; Güntert *et al.* 1991a).

^f Each individual distance restraint is retained with the given probability. Results are averages over five different random selections.

and one might ask what quality of structure could be attained on the basis of a less complete but more readily collected data set. To address this question, a number of structure calculations were performed with subsets of the complete experimental data set for cyclophilin A (Table 8). The subsets were created alternatively by retaining randomly only a certain percentage of all distance

restraints, by neglecting stereospecific assignments or torsion angle restraints, or by restricting the data set to only backbone and H^β NOEs or to only H^N-H^N NOEs. As expected, the precision of the structure decreases with decreasing information content of the data set (Table 8). In parallel it becomes more difficult for the structure calculation algorithm to find good solutions, i.e. the success rates sink. However, the absence of stereospecific assignments, torsion angle restraints, or up to 50% of the NOE distance restraints has only a moderate effect and does not preclude the determination of a well-defined structure. Low resolution structures can still be calculated from as little as 10% of the distance restraints, or without any experimental restraints for the side-chain conformation beyond C^β . Not astonishingly, however, the success rate of the structure calculation was only 26% in the absence of side-chain restraints, an unusually low value for the torsion angle dynamics algorithm that presumably resulted from the difficulty to pack the side-chains in the protein core. With only 5% of the NOEs it is no longer possible to unambiguously determine the three-dimensional structure, and even less so if only restraints among backbone amide protons are considered (Table 8). The latter result is in line with the findings of Venters *et al.* (1995) and Smith *et al.* (1996) who have investigated the possibility of global fold determination using deuterated protein samples and found that it would be necessary to measure H^N-H^N distances up to 7 Å to enable an unambiguous global fold determination.

8.4 Wrong restraints and their elimination

In the course of a protein structure determination by NMR it is always possible that NOEs with incorrect assignments enter the data set. The normal way to detect and correct such mistakes is a careful analysis of restraint violations in the structure calculated from the experimental data. Consistent violations, i.e. those that occur in all or in a large majority of the conformers, are most likely not due to imperfections of the structure calculation program but the result of restraints that contradict each other. An ideal structure calculation method from the point of view of error detection would pinpoint all mistakes by reporting consistent violations for all wrong restraints, but not for any other (correct) restraints. In practice, this is not the case because the structure calculation programs minimize a target function that is a sum of contributions from all restraints, and to which the largest violations contribute most. Hence, there is a tendency to ‘smear out’ the problem caused by a wrong restraint over other restraints in the vicinity to the effect that either additional, correct restraints become consistently violated, or that the problem is no longer recognized because it was distributed over many, only slightly violated restraints. The latter problem is normally less severe in torsion angle space than in Cartesian space, where slight, diffuse distortions of the covalent geometry offer additional possibilities to disperse violations.

The ability of the torsion angle dynamics algorithm of DYANA to detect and automatically eliminate erroneous restraints is illustrated in Table 9 using data sets from Table 8 to which 2% of distance restraints with arbitrary, wrong

Table 9. Structure calculations for Cyclophilin A with data sets to which first 2% distance restraints with wrong assignments were added and from which subsequently consistently violated distance restraints were eliminated^a

Data set	Restrains eliminated (%) ^b		Target function (Å ²)	Backbone RMSD (Å) ^c	
	Wrong	Correct		Radius	Bias
All experimental restraints	90	0.71	1.24	0.60	0.38
No stereospecific assignments	91	0.43	1.37	0.66	0.49
No angle restraints	89	0.66	1.59	0.59	0.44
% of all NOEs					
75%	85	0.46	1.45	0.69	0.83
50%	82	0.51	1.33	0.79	1.05
25%	74	0.84	4.02	1.20	1.56
10%	50	0.49	4.02	2.05	3.37
5%	0	0.18	3.38	4.69	5.57
Only backbone and H ^β NOEs	70	0.96	9.54	1.51	1.62

^a Wrong distance restraints were generated by selecting distance restraints arbitrarily from the complete experimental data set and replacing one of the two atoms by an arbitrarily chosen different atom for which the ¹H chemical shift was available. The second atom and the upper distance limit of the restraint remained unchanged. With these wrong restraints added to each data set bundles of 20 conformers were calculated using the same protocol as in Table 8. Consistently violated distance restraints, i.e. those that are violated by more than 0.2 Å in 15 or more of the 20 conformers, were then eliminated in two steps. In the first round, the 25% consistently violated distance restraints with the largest average violations were deleted, and the structure calculation was repeated. In the second round, all remaining consistently violated distance restraints were eliminated, and the structure calculation was repeated again. The resulting 20 conformers for each data set were analysed.

^b Number of wrong distance restraints eliminated, given as a percentage of the total number of wrong distance restraints that were added to the data set, and number of correct distance restraints eliminated, given as a percentage of the total number of distance restraints in the original data set without wrong restraints.

^c RMSD radius and bias are defined as in Table 8.

assignments have been added. Incorrect restraints were detected and eliminated in three rounds of structure calculations where consistent violations found at the end of the first and second round were removed from the data set that became the input for the following structure calculation. The results (Table 9) show that with good data sets around 90% of the erroneous restraints could be detected by this straightforward automatic method, and that significantly less than 1% of the correct restraints were falsely eliminated. The resulting structures are of similar quality as those obtained from correct restraints only, and there is close agreement between them. In the case of sparse data sets, however, the discriminatory power

of the procedure deteriorates to the point that only 50% of the wrong restraints can be found and removed from the data sets comprising one tenth of all NOEs. The many wrong NOEs remaining in the data set lead to significantly higher target function values than the calculations with exclusively correct restraints.

9. AUTOMATED ANALYSIS OF NOESY SPECTRA

The assignment of cross peaks in NOESY spectra for the collection of NOE upper distance limits on ^1H - ^1H distances is an essential part of the determination of three-dimensional protein structures in solution by NMR. Obtaining NOESY cross peak assignments is usually a laborious endeavour, particularly in spectral regions where chemical shift degeneracies result in excessive cross peak overlap. Were it not for these inevitable chemical shift degeneracies and the usually somewhat imprecise cross peak positional information, all assignments could of course be made in a straightforward manner based on the knowledge of the chemical shifts resulting from the sequence-specific resonance assignments. In practice, however, only a fraction of the NOESY cross peaks can be assigned in this direct way and subsequently used to generate a preliminary, 'low resolution' structure of the protein under investigation. Subsequently, these preliminary conformers may be used to reduce the number of previously ambiguous assignments by eliminating pairs of protons which have the chemical shift coordinates of the cross peak considered but, on the basis of the preliminary solution structure, are further apart than a predetermined maximum distance cutoff for the observation of NOEs. The collection of an extensive set of distance restraints and the calculation of a high-quality structure are thus not separate, subsequent steps of an NMR structure determination but intertwined in an iterative process (Fig. 25a), regardless of whether exclusively manual, semiautomatic (Güntert *et al.* 1993; Meadows *et al.* 1994) or automatic methods (Mumenthaler & Braun, 1995; Mumenthaler *et al.* 1997; Nilges, 1995; Nilges *et al.* 1997) are employed.

9.1 Chemical shift tolerance range

The two fundamental requirements for a valid NOE assignment are agreement between chemical shifts and the peak position, and spatial proximity in a (preliminary) structure (Fig. 25b). Typically only a minority of the NOESY cross peaks can be assigned unambiguously based on chemical shift agreement alone because of inevitable small uncertainties in the determination of chemical shifts and peak positions. Such inaccuracies require the introduction of a non-vanishing chemical shift tolerance Δ_{tol} for the agreement between a ^1H chemical shift and a peak position. The size of Δ_{tol} , that is the accuracy of chemical shift and peak position determination, has a very pronounced influence on the number of possible assignments for an NOE cross peak. This can be rationalized as follows (Mumenthaler *et al.* 1997).

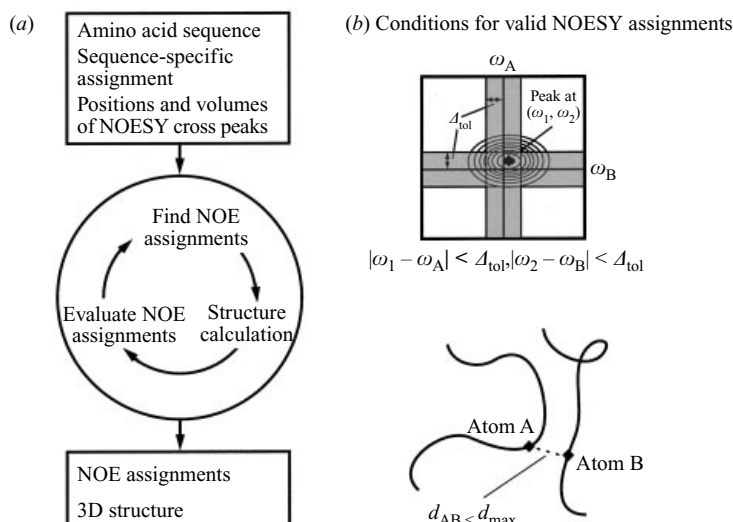


Fig. 25. (a) Flowchart of the iterative process of NOESY cross peak assignment and structure calculation. (b) The two conditions that must be fulfilled by valid NOESY cross peak assignments: Agreement between chemical shifts and the peak position, and spatial proximity in a (preliminary) structure.

In a two-dimensional NOESY spectrum with N cross peaks for a protein containing n hydrogen atoms with chemical shifts distributed evenly over a region of width $\Delta\omega$, the probability of finding a ^1H shift in an interval $[\omega - \Delta_{\text{tol}}, \omega + \Delta_{\text{tol}}]$ about any selected position ω is

$$p = \frac{2\Delta_{\text{tol}}}{\Delta\omega}. \quad (40)$$

In the absence of structural information, the number of peaks that can be assigned unambiguously based on the agreement of chemical shifts within the tolerance is expected to be

$$N^{(1)} = N(1-p)^{2n-2} \approx Ne^{-2np}. \quad (41)$$

Equation (41) predicts that the percentage of peaks that can be assigned unambiguously without knowledge of a preliminary structure decreases exponentially with increasing size of the protein and increasing tolerance range. The number of peaks with exactly two assignment possibilities is expected to be

$$N^{(2)} = N2p(n-1)(1-p)^{2n-3} \approx 2npN^{(1)}. \quad (42)$$

$N^{(2)}$ vanishes for very small Δ_{tol} values, but increases linearly as a function of $N^{(1)}$ with a coefficient that is proportional to the protein size and the Δ_{tol} value. At $\Delta_{\text{tol}} = 0.01$ ppm, $N^{(2)}$ is usually 2–3 times larger than $N^{(1)}$. Fig. 26 shows that the simple model of equations (40)–(42) provides a remarkably good description of the situation in a real protein.

For peak lists obtained from ^{13}C - or ^{15}N -resolved 3D [^1H , ^1H]-NOESY spectra,

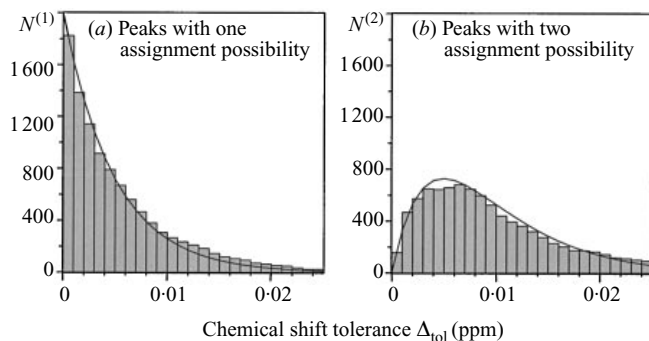


Fig. 26. Numbers of cross peaks with exactly one ($N^{(1)}$, shown in (a)) or exactly two ($N^{(2)}$, shown in (b)) possible assignments on the basis of agreement between ^1H chemical shifts and the peak positions within a tolerance Δ_{tol} in a two-dimensional NOESY spectrum of the protein WmKT (Antuch *et al.* 1996). No structural information has been used to resolve ambiguities. The NOESY peak list was simulated on the basis of the experimental chemical shift list by postulating a cross peak between any pair of protons separated by less than 4 \AA in the best NMR conformer (Antuch *et al.* 1996). In both (a) and (b) the curved lines represent the corresponding values predicted by equations (41) and (42) for $N = 1986$ peaks, $n = 457$ protons, and a spectral width of $\Delta\omega = 9$ ppm.

the ambiguity in the proton dimension correlated to the hetero-spin is normally resolved. Equation (41) then adopts the form

$$N^{(1)} \approx Ne^{-np}. \quad (43)$$

The expected percentage of unambiguously assigned peaks is thus the same as in a two-dimensional NOESY spectrum for a protein of half the size, or for half the chemical shift tolerance.

In order to assign the majority of the NOESY cross peaks, the ambiguity of assignments based exclusively on chemical shifts must be resolved by reference to a preliminary structure. The ambiguity is resolved completely if all but one of the potential assignments correspond to pairs of hydrogen atoms separated by more than a maximal distance d_{max} for which a NOE may be observed. Assuming that the hydrogen atoms are evenly distributed within a sphere of radius R that represents the protein, the probability q to find two randomly selected hydrogen atoms closer to each other than d_{max} is given approximately by the ratio between the volumes of two spheres with radii d_{max} and R , respectively:

$$q = \left(\frac{d_{\text{max}}}{R}\right)^3. \quad (44)$$

For a nearly spherical protein with radius $R = 15 \text{ \AA}$ and $d_{\text{max}} = 5 \text{ \AA}$ this probability becomes approximately 4%, indicating that only 96% of the peaks with two assignment possibilities can be assigned uniquely by reference to the protein structure. The total number of uniquely assigned peaks, N_{unique} , can be increased optimally to

$$N_{\text{unique}} = N^{(1)} + (1-q)N^{(2)} + (1-q)^2N^{(3)} + \dots \quad (45)$$

Even by reference to a perfectly refined structure it is therefore impossible, on fundamental grounds, to resolve all assignment ambiguities because q will never vanish and hence $N_{\text{unique}} < N$.

9.2 Semiautomatic methods

Semiautomatic NOE assignment methods provide for each NOESY cross peak a list of the assignment possibilities according to the criteria of Fig. 25*b*. These are analysed by the spectroscopist who may be able to further reduce their number by visual inspection of the corresponding cross peaks and line shapes in the NOESY spectrum. Peaks that can be assigned unambiguously by this method are added to the input data set for the next round of structure calculation. The program ASNO (Güntert *et al.* 1993) that is normally used in conjunction with the interactive spectrum analysis program XEASY (Bartels *et al.* 1995) uses this principle for automated removal of ambiguities arising from chemical shift degeneracies and thus supports the collection of an extensive set of NOE distance restraints in several rounds of NOESY cross peak assignments and structure calculations.

9.3 Ambiguous distance restraints

An elegant approach to the NOE assignment problem was introduced by Nilges (1993, 1995) who accounted for the ambiguity in the purely chemical-shift-based NOE assignments by ‘ambiguous distance restraints’, i.e. by interpreting the peak volumes as r^{-6} -weighted sums of contributions from all possible peak assignments in the NOE target function. Ambiguous distance restraints are thus a generalization of the r^{-6} -summation method of equation (10) that can be applied to restraints with diastereotopic protons in the absence of stereospecific assignments. An optimization procedure based on simulated annealing by molecular dynamics was described that is capable of using highly ambiguous input data for *ab initio* structure calculations, where it is possible to specify the restraint list directly in terms of the proton chemical shift assignment and the NOESY cross peak positions (Nilges, 1995). This procedure was applied in structure calculations of the basic pancreatic trypsin inhibitor (BPTI) from simulated NOESY spectra (Nilges, 1995), and has been used also for the calculation of symmetric oligomeric structures from NMR data, where all peaks are a superposition of at least two NOE signals (Nilges, 1993; Donoghue *et al.* 1996). In contrast to the normal manual approach, in which unambiguous assignments are sought and peaks that cannot be assigned unambiguously are not used in the structure calculation, the notion of ambiguous distance restraints allows one to exploit the information carried by all NOESY cross peaks, regardless of whether a peak has a unique assignment possibility or not. There are always some NOESY cross peaks that reflect contributions from more than one spatially proximate proton pair; these can be treated more realistically by ambiguous distance restraints than by uniquely assigned NOEs.

Recently, Nilges *et al.* (1997) have proposed a novel structure calculation

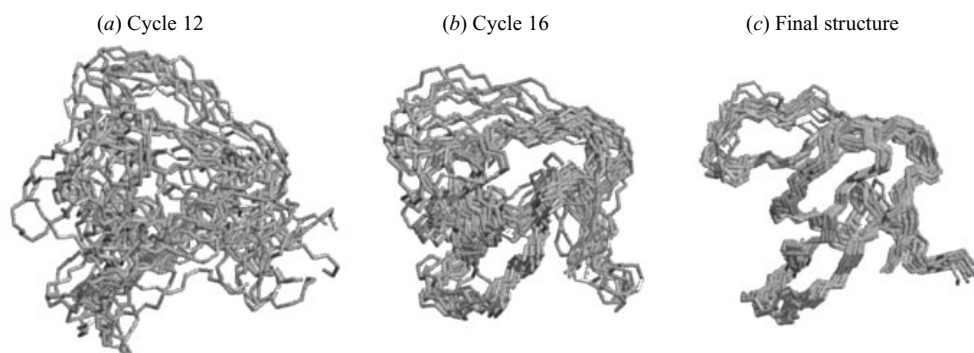


Fig. 27. Structures of the SH₃ domain of human p56 Lck tyrosine kinase (Hiroaki *et al.* 1996; M. Salzmann, unpublished) at various stages of the automated combined NOESY assignment and structure calculation method of Mumenthaler *et al.* (1997). The calculation is based on the experimental NMR data set. Shown are backbone superpositions of ten conformers at the end of the intermediate cycles 12 (a) and 16 (b), as well as the final structure (c).

method that combines ambiguous distance restraints with an iterative assignment strategy (see next section), whereby unambiguous assignments can be derived for many NOE cross peaks that were entered into the calculation initially as ambiguous restraints.

9.4 Iterative combination of NOE assignment and structure calculation

An alternative approach for automatic NOESY assignment was proposed (Mumenthaler & Braun, 1995; Mumenthaler *et al.* 1997) that uses as input only the chemical shift lists obtained from the sequence-specific resonance assignment and a list of NOESY cross peak positions. Ambiguous peak assignments are treated as separate distance restraints in the structure calculations, and erroneous assignments are eliminated in iterative cycles. An error-tolerant target function reduces the impact of erroneous restraints on the calculated structures. In contrast to the approach of Nilges (1995), noise and artifact peaks can be removed automatically during the procedure, and peaks are ultimately assigned to single proton pairs. This allows a critical comparison of the NOE assignments obtained automatically with those from manual procedures not only on the level of the final structures but also on the level of individual NOE assignments.

The method of Mumenthaler *et al.* (1997) performs normally 25 cycles of automatic assignment and structure calculation (Fig. 27), each with the three main steps given in Fig. 25a. For the NOE assignment step in Fig. 25a, a list of the assignment possibilities based on a given chemical shift tolerance Δ_{tol} is prepared. In the first cycle, when no structural information is available, this list is used directly. Otherwise, i.e. from the second cycle onwards, the preliminary structure available from the preceding cycle is used to eliminate assignment possibilities that correspond to proton pairs further apart than a limiting distance d_{max} which

is decreased linearly from 10.8 Å in the first to 5.4 Å in the last cycle. Using the automatic calibration method described in Section 4.1, new distance restraints are then added as 'test assignments' to the input for the structure calculation for all so far unassigned NOESY cross peak with less than M assignment possibilities, where $M = 2$ for the first 15 cycles, $M = 3$ for cycles 16–19, and $M = 4$ for cycles 20–25. It is necessary to use $M > 1$ because otherwise the low number of unambiguous assignments at the outset would preclude convergence to a well-defined structure. In the structure calculation step of Fig. 25*a* an ensemble of conformers is calculated using the standard protocol of the program DYANA (Güntert *et al.* 1997). Because for a given peak up to M different restraints, of which normally only one will turn out to be correct, are included in the input data for the structure calculation it is important to use a functional form of the target function that will not be dominated too much by strongly violated restraints (Mumenthaler & Braun, 1995). In the NOE evaluation step of Fig. 25*a* the ten conformers with lowest target function values are analysed. Each peak for which test assignments have been added to the input is transferred either to the list of unambiguous assignments, or returned to the list of unassigned peaks, which will be analysed again in the next cycle. Peaks that have been classified as unambiguous in previous cycles can be reclassified if the corresponding distance restraint is violated in most conformers.

Applications of this procedure to the experimental data sets for six proteins are summarized in Table 10 (Mumenthaler *et al.* 1997). For all six proteins nearly complete sequence-specific resonance assignments were available, and their solution structure had been calculated on the basis of manually assigned NOE cross peaks. The start point for the automatic procedure were the original peak lists from which all assignments were deleted. Tolerance ranges Δ_{tol} of 0.01, 0.015 and 0.02 ppm were used for protons in two-dimensional NOESY spectra, three-dimensional NOESY spectra of P14a, and three-dimensional NOESY spectra of DnaJ, respectively. The NOE assignments and the structures obtained from the automatic procedure closely resemble those obtained from manually made assignments. On average, the extent of assignments is somewhat lower from the automatic method, and different assignments by the two approaches were obtained for less than 2% of the peaks. The target functions and RMSD radii of the final structures were comparable, and the automatically determined structures show little bias from the original structures (Table 10).

Further calculations conducted by Mumenthaler *et al.* (1997) showed that the automatic assignment method was remarkably robust with respect to imperfect NOE peak lists and could produce acceptable structures from incomplete NOE input. In contrast, the method was quite susceptible to incomplete ^1H chemical shift lists. In spite of the progress made with the automatic method, spectroscopists working interactively with NOESY spectra still have several advantages because they can exclude assignment possibilities by line shape considerations and often intuitively use smaller tolerance ranges between peak positions and chemical shifts. This contributes in general to a more complete NOESY assignment by the interactive method than by the automated approach.

Table 10. Structure calculations using manually and automatically assigned NOESY peak lists^a

Quantity	Er-2	Hirudin	434	WmKT	DnaJ	P14a
Size and experimental input data						
Residues	40	51	63	88	107	135
NOESY spectra ^b	HD	HD	H	H	NC	HDNC
NOESY cross peaks ^c	2207	1273	1282	1998	2841	8438
Percentage of peaks (%) ^d						
Manually assigned	72	99	99	85	96	90
Automatically assigned	74	93	80	83	82	78
Identically assigned	66	91	78	75	80	74
Differently assigned	0.8	1.1	2.4	1.0	1.4	1.0
Inconsistent ^e	5.3	0.5	1.1	3.3	2.0	5.1
Structures obtained from manual assignment						
Target function value (Å ²)	0.4-0.8	0.1-0.2	0.3-0.7	1.9-4.3	0.5-1.3	1.4-4.0
RMSD radius(Å) ^f	0.3	0.4	0.6	0.7	1.0	0.8
Structures obtained from automatic assignment						
Target function value (Å ²)	0.2-0.4	0.1-0.2	0.7-0.8	1.0-1.5	0.1-0.5	5.1-6.8
RMSD radius (Å) ^f	0.4	0.5	0.6	0.6	1.7	1.2
RMSD bias (Å) ^g	0.6	0.8	0.9	0.5	1.0	1.5

^a Experimental NOESY peak list sets for the following proteins were used: Er-2: pheromone Er-2 from *Exphlotes raikovi* (Ottiger *et al.* 1994); Hirudin: hirudin(1-51) (Szyperski *et al.* 1992); 434: 434-repressor(1-63) with mutation R10M (Pervushin *et al.* 1996); WmKT: yeast killer toxin WmKT (Antuch *et al.* 1996); DnaJ: molecular chaperone DnaJ(1-108) (Pellicchia *et al.* 1996); P14a: pathogenesis-related protein P14a from tomato (Fernández *et al.* 1997). Data for structures obtained by manual assignment are from the original publications. Automatic assignment was performed in 25 iterative cycles of NOE assignment and structure calculation (Mumenthaler *et al.* 1997). Structure calculations were performed with the programs DIANA or DYANA using the variable target function method with redundant torsion angle restraints (Güntert & Würthrich, 1991).

^b H: two-dimensional [¹H, ¹H]-NOESY in H₂O; D: two-dimensional [¹H, ¹H]-NOESY in D₂O; N: ¹⁵N-resolved three-dimensional [¹H, ¹H]-NOESY; C: ¹³C-resolved three-dimensional [¹H, ¹H]-NOESY.

^c Total number of cross peaks in all NOESY spectra used.

^d All percentages are relative to the total number of NOESY cross peaks.

^e Percentage of peaks that are inconsistent with the final structure obtained by automatic assignment. For these peaks every possible assignment within the chemical shift tolerance range is violated by more than 1 Å in all conformers.

^f RMSD radii of the 20 conformers used to represent the solution structure, computed for the backbone atoms N, C^α, C' of the following residues: Er-2, 3-37; Hirudin, 3-30 and 37-48; 434, 1-63; WmKT, 3-39 and 47-87; DnaJ, 6-57; P14a, 1-135.

^g RMSD bias to the mean coordinates of the structure bundle obtained from manual assignment.

10. STRUCTURE REFINEMENT

There exist many possibilities for the refinement of NMR structures of proteins. The following brief overview can only mention a few often used refinement methods.

10.1 *Restrained energy minimization*

The structure calculation algorithms for NMR structures usually use a simplified force field that contains only the most dominant parts of the conformational energy. Therefore, the resulting structures may be unfavourable with respect to a full, 'physical' energy function (Momany *et al.* 1975; Brooks *et al.* 1983; Cornell *et al.* 1995; van Gunsteren *et al.* 1996) that includes, in addition to the terms used by the structure calculation algorithms of Section 6, also a Lennard-Jones potential and electrostatic interactions for non-bonded atom pairs, torsion angle potentials, and possibly other terms. The conformational energy of a conformation obtained from a structure calculation program can be reduced significantly by restrained energy minimization, i.e. by locating a local minimum of the conformational energy function in the near vicinity of the input structure.

Restrained energy minimization of a correct structure results in only small changes of the conformation (Billeter *et al.* 1990). Because no large-scale conformational changes are necessary, the restraining potentials for distance and angle restraints may be chosen steeper than in the preceding structure calculation, thereby reducing the maximal restraint violations. Potentials proportional to the sixth (instead of second) power of the distance restraint violation have been used frequently (Billeter *et al.* 1990). Generally the extent and regularity of hydrogen bonds shows a marked improvement upon energy minimization because the geometric force fields used for the structure calculation normally do not contain a driving force for hydrogen bond formation (unless explicit hydrogen bond restraints were used, of course). Since the solvent surrounding the macromolecule is very important for a realistic representation of electrostatic interactions, restrained energy minimizations should be performed in a box or shell of explicit water molecules. Energy minimization in vacuo exaggerates electrostatic interactions and can lead to artifacts such as charged and polar side-chains on the protein surface that bend back to the protein, forming spurious salt-bridges and hydrogen bonds (Luginbühl *et al.* 1996).

10.2 *Molecular dynamics simulation*

An unrestrained or restrained molecular dynamics simulation under physiological conditions using the full physical force field and explicit water to represent the solvent can often give new insights into a protein structure, in particular for the generally disordered protein surface (McCammon & Harvey, 1987; Brooks *et al.* 1988; van Gunsteren & Berendsen, 1990). Such molecular dynamics simulations try to represent the solvated molecular system as faithfully as possible and are

fundamentally different from simulated annealing, where artificial conditions such as high temperature are chosen in order to enhance the sampling of conformation space. A limiting factor in molecular dynamics simulations are the relatively short simulation times of up to a few nanoseconds that are feasible with present computers, because many motions in proteins occur on longer time scales.

10.3 *Time- or ensemble averaged restraints*

The commonly used structure calculation algorithms try to find rigid conformations that fulfill all distance and torsion angle restraints simultaneously, and the effects of internal mobility of the polypeptide chain are taken into account implicitly by interpreting the NOE data as conservative upper distance bounds instead of exact distance restraints (Wüthrich, 1986). In reality, the NOEs and scalar coupling constants measured by NMR constitute an average over time and space. Methods have been devised to include distance and torsion angle restraints as time-averaged rather than instantaneous restraints into a molecular dynamics simulation (Kessler *et al.* 1989; Pearlman & Kollman, 1991; Torda *et al.* 1989, 1990, 1993; van Gunsteren *et al.* 1994). In another approach, a molecular dynamics simulation is performed simultaneously for an ensemble of conformers, such that the restraints are not required to be fulfilled by each individual conformer but only by the ensemble as a whole (Scheek *et al.* 1991; Bonvin & Brünger, 1995, 1996).

10.4 *Relaxation matrix refinement*

Both spin diffusion and internal mobility influence the NOE intensities from which distance restraints are derived for the structure calculation. Complete relaxation matrix refinement (Keepers & James, 1984; Boelens *et al.* 1989; Yip & Case, 1989) can, in principle, take these factors into account and thus may make it possible to make more quantitative use of the NOE data as with the initial rate approximation (Kumar *et al.* 1980) and the semi-quantitative calibration of distance restraints (Wüthrich, 1986). However, there can be a danger of overinterpreting the data because many of the parameters entering the relaxation matrix cannot be measured experimentally. In particular, assumptions are needed about internal and overall motions of the protein (Macura & Ernst, 1980). Two different methods of complete relaxation matrix refinement are in use: Either the relaxation matrix treatment is used to derive a more precise set of distance restraints, which is then used in a conventional structure calculation (Keepers & James, 1984; Boelens *et al.* 1989), or the three-dimensional structure may be refined directly against the observed NOE intensities (Borgias & James, 1988; Yip & Case, 1989; Mertz *et al.* 1991). The second approach is conceptually more attractive but also more time-consuming. In analogy to the practice in X-ray crystallography, it is possible to define *R*-factors that measure the agreement between the NOESY spectrum and the three-dimensional structure (Gonzales *et al.* 1991; Thomas *et al.* 1991; Withka *et al.* 1992).

11. ACKNOWLEDGEMENTS

I thank Prof. Kurt Wüthrich for generous support, Mr Michael Salzmann for providing Fig. 27, and Prof. Martin Billeter for helpful discussions. The use of the computing facilities of the Competence Centre for Computational Chemistry of ETH Zürich is gratefully acknowledged.

12. REFERENCES

- ABAGYAN, R. & TOTROV, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002.
- ABE, H., BRAUN, W., NOGUTI, T. & GO, N. (1984). Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles in proteins. General recurrent equations. *Computers & Chemistry* **8**, 239–247.
- ABRAGAM, A. (1961). *Principles of Nuclear Magnetism*. Oxford: Clarendon Press.
- ALDER, B. J. & WAINWRIGHT, T. E. (1959). Studies of molecular dynamics. I. General method. *J. Chem. Phys.* **31**, 459–466.
- ALLEN, M. P. & TILDESLEY, D. J. (1987). *Computer Simulation of Liquids*. Oxford: Clarendon Press.
- ALTMAN, R. B. & JARDETZKY, O. (1986). New strategies for the determination of macromolecular structure in solution. *J. Biochem.* **100**, 1403–1423.
- ALTMAN, R. B. & JARDETZKY, O. (1989). Heuristic refinement method for determination of solution structure of proteins from nuclear magnetic resonance data. *Meth. Enzymol.* **177**, 218–246.
- ANTUCH, W., GÜNTERT, P. & WÜTHRICH, K. (1996). Ancestral $\beta\gamma$ -crystallin precursor structure in a yeast killer toxin. *Nature Struct. Biol.* **3**, 662–665.
- ARNOLD, V. I. (1978). *Mathematical Methods of Classical Mechanics*. New York: Springer.
- ARSENEV, A. S., KONDAKOV, V. I., MAIOROV, V. N. & BYSTROV, V. F. (1984). NMR solution spatial structure of ‘short’ scorpion insectotoxin I₅A. *FEBS Lett.* **165**, 57–62.
- BAE, D. S. & HAUG, E. J. (1987). A recursive formulation for constrained mechanical system dynamics: Part I. Open loop systems. *Mech. Struct. Mech.* **15**, 359–382.
- BANCI, L., BERTINI, I., BREN, K. L., CREMONINI, M. A., GRAY, H. B., LUCHINAT, C. & TURANO, P. (1996). The use of pseudocontact shifts to refine solution structures of paramagnetic metalloproteins: Met80Ala cyano-cytochrome *c* as an example. *J. Biol. Inorg. Chem.* **1**, 117–126.
- BARTELS, C., XIA, T., BILLETER, M., GÜNTERT, P. & WÜTHRICH, K. (1995). The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **6**, 1–10.
- BEEEMAN, D. (1976). Some multistep methods for use in molecular dynamics calculations. *J. Comput. Phys.* **20**, 130–139.
- BERENDSEN, H. J. C., POSTMA, J. P. M., VAN GUNSTEREN, W. F., DiNOLA, A. & HAAK, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.
- BERNDT, K. D., GÜNTERT, P., ORBONS, L. P. M. & WÜTHRICH, K. (1992). Determination of a high-quality NMR solution structure of the bovine pancreatic trypsin inhibitor (BPTI) and comparison with three crystal structures. *J. Mol. Biol.* **227**, 757–775.

- BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F., JR., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCI, T. & TASUMI, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
- BIAMONTI, C., RIOS, C. B., LYONS, B. A. & MONTELIONE, G. T. (1994). Multi-dimensional NMR experiments and analysis techniques for determining homo- and heteronuclear scalar coupling constants in proteins and nucleic acids. *Adv. Biophys. Chem.* **4**, 51–120.
- BILLETER, M., BRAUN, W. & WÜTHRICH, K. (1982). Sequential resonance assignments in protein ^1H nuclear magnetic resonance spectra. Computation of sterically allowed proton-proton distances and statistical analysis of proton-proton distances in single crystal protein conformations. *J. Mol. Biol.* **155**, 321–346.
- BILLETER, M., HAVEL, M. & WÜTHRICH, K. (1987). The ellipsoid algorithm as a method for the determination of polypeptide conformations from experimental distance constraints and energy minimization. *J. Comp. Chem.* **8**, 132–141.
- BILLETER, M., KLINE, A. D., BRAUN, W., HUBER, R. & WÜTHRICH, K. (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.
- BILLETER, M., SCHAUMANN, T., BRAUN, W. & WÜTHRICH, K. (1990). Restrained energy refinement with two different algorithms and force fields of the structure of the α -amylase inhibitor tendamistat determined by NMR in solution. *Biopolymers* **29**, 695–706.
- BLUMENTHAL, L. M. (1953). *Theory and Applications of Distance Geometry*. Cambridge, UK: Cambridge University Press.
- BOELENS, R., KONING, T. M. G., VAN DER MAREL, G. A., VAN BOOM, J. H. & KAPTEIN, R. (1989). Iterative procedure for structure determination from proton-proton NOEs using a full relaxation matrix approach. Application to a DNA octamer. *J. Magn. Reson.* **82**, 290–308.
- BONVIN A. M. & BRÜNGER, A. T. (1995). Conformational variability of solution nuclear magnetic resonance structures. *J. Mol. Biol.* **250**, 80–93.
- BONVIN A. M. & BRÜNGER, A. T. (1996). Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? *J. Biomol. NMR* **7**, 72–76.
- BORGAS, B. A. & JAMES, T. L. (1988). COMATOSE, a method for constrained refinements of macromolecular structure based on two-dimensional nuclear Overhauser spectra. *J. Magn. Reson.* **79**, 493–512.
- BRANDEN, C. & TOOZE, J. (1991). *Introduction to Protein Structure*. New York & London: Garland Publishing.
- BRAUN W. (1987). Distance geometry and related methods for protein structure determination from NMR data. *Q. Rev. Biophys.* **19**, 115–157.
- BRAUN, W. & GO, N. (1985). Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J. Mol. Biol.* **186**, 611–626.
- BRAUN, W., BÖSCH, C., BROWN, L. R., GO, N. & WÜTHRICH, K. (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.
- BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J., SWAMINATHAN, S. &

- KARPLUS, M. (1983). CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* **4**, 187–217.
- BROOKS III, C. L., KARPLUS, M. & PETTITT, B. M. (1988). *Proteins. A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*. New York: Wiley.
- BRÜNGER, A. T. (1992). *X-PLOR, Version 3.1. A System for X-ray Crystallography and NMR*. New Haven: Yale University Press.
- BRÜNGER, A. T. & NILGES, M. (1993). Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q. Rev. Biophys.* **26**, 49–125.
- BRÜNGER, A. T., CLORE, G. M., GRONENBORN, A. M. & KARPLUS, M. (1986). Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc. Natl. Acad. Sci. USA* **83**, 3801–3805.
- BRÜNGER, A. T., ADAMS, P. D. & RICE, L. M. (1997). New applications of simulated annealing in X-ray crystallography and solution NMR. *Structure* **5**, 325–336.
- BRÜSCHWEILER, R., BLACKLEDGE, M. & ERNST, R. R. (1991). Multi-conformational peptide dynamics derived from NMR data: a new search algorithm and its application to antamanide. *J. Biomol. NMR* **1**, 3–11.
- CASE, D. A., DYSON, H. J. & WRIGHT, P. (1994). Use of chemical shifts and coupling constants in nuclear magnetic resonance structural studies on peptides and proteins. *Meth. Enzymol.* **239**, 392–416.
- CAVANAGH, J., FAIRBROTHER, W. J., PALMER III, A. G. & SKELTON, N. (1996). *Protein NMR spectroscopy. Principles and Practice*. San Diego: Academic Press.
- CLORE, G. M. & GRONENBORN, A. M. (1990). Applications of three- and four-dimensional heteronuclear NMR spectroscopy to protein structure determination. *Prog. NMR Spectrosc.* **23**, 43–92.
- CLORE, G. M., GRONENBORN, A. M., BRÜNGER, A. T. & KARPLUS, M. (1985). Solution conformation of a heptadecapeptide comprising the DNA binding helix F of the cyclic AMP receptor protein of *Escherichia coli*. Combined use of ¹H nuclear magnetic resonance and restrained molecular dynamics. *J. Mol. Biol.* **186**, 435–455.
- CLORE, G. M., BRÜNGER, A. T., KARPLUS, M. & GRONENBORN, A. M. (1986a). Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination: a model study of crambin. *J. Mol. Biol.* **191**, 523–551.
- CLORE, G. M., NILGES, M., SUKUMARAN, D. K., BRÜNGER, A. T., KARPLUS, M. & GRONENBORN, A. M. (1986b). The three-dimensional structure of a1-purothionin in solution: combined use of nuclear magnetic resonance, distance geometry and restrained molecular dynamics. *EMBO J.* **5**, 2729–2735.
- 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). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Amer. Chem. Soc.* **117**, 5179–5197.
- CREIGHTON, T. (1993). *Proteins. Structures and Molecular Properties*. 2nd ed. New York: Freeman.
- CRIPPEN, G. M. (1977). A novel approach to the calculation of conformation: Distance geometry. *J. Comp. Phys.* **26**, 449–452.
- CRIPPEN, G. M. & HAVEL, T. F. (1978). Stable calculation of coordinates from distance information. *Acta Cryst.* **A34**, 282–284.
- CRIPPEN, G. M. & HAVEL, T. F. (1988). *Distance Geometry and Molecular Conformation*. Taunton, UK: Research Studies Press.

- DAVIES, G. J., GAMBLIN, S. J., LITTLECHILD, J. A., DAUTER, Z., WILSON, K. S. & WATSON, H. C. (1994). Structure of the ADP complex of the 3-phosphoglycerate kinase from *Bacillus sterothermophilus* at 1.65 Å. *Acta Cryst.* **D50**, 202–209.
- DE DIOS, A. C., PEARSON, J. G. & OLDFIELD, E. (1993). Secondary and tertiary structural effects on protein NMR chemical shifts: an *ab initio* approach. *Science* **260**, 1491–1496.
- DE MARCO, A., LLINÁS, M. & WÜTHRICH, K. (1978*a*). Analysis of the ¹H-NMR spectra of ferrichrome peptides. I. The non-amide protons. *Biopolymers* **17**, 617–636.
- DE MARCO, A., LLINÁS, M. & WÜTHRICH, K. (1978*b*). ¹H-¹⁵N spin-spin couplings in alumichrome. *Biopolymers* **17**, 2727–2742.
- DENK, W., BAUMANN, R. & WAGNER, G. (1986). Quantitative evaluation of cross-peak intensities by projection of two-dimensional NOE spectra on a linear space spanned by a set of reference resonance lines. *J. Magn. Reson.* **67**, 386–390.
- DONOGHUE, S. I., KING, G. F. & NILGES, M. (1996). Calculation of symmetric multimer structures from NMR data using a priori knowledge of the monomer structure, comonomer restraints, and interface mapping: the case of leucine zippers. *J. Biomol. NMR* **8**, 193–206.
- DRENTH, J. (1994). *Principles of Protein X-ray Crystallography*. New York: Springer.
- DUBS, A., WAGNER, G. & WÜTHRICH, K. (1979). Individual assignments of amide proton resonances in the proton NMR spectrum of the basic pancreatic trypsin inhibitor. *Biochem. Biophys. Acta* **577**, 177–194.
- EDISON, A. S., ABILDGAARD, F., WESTLER, W. M., MOOBERRY, E. S. & MARKLEY, J. L. (1994). Practical introduction to theory and implementation of multinuclear, multidimensional nuclear magnetic resonance experiments. *Meth. Enzymol.* **239**, 3–79.
- ENDO, S., WAKO, H., NAGAYAMA, K. & GO, N. (1991). A new version of DADAS (Distance Analysis in Dihedral Angle Space) and its performance. In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (ed. J. C. Hoch, F. M. Poulsen & C. Redfield), pp. 233–251, New York & London: Plenum Press.
- ENGH, R. A. & HUBER, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr.* **A47**, 392–400.
- ERNST, R. R., BODENHAUSEN, G. & WOKAUN, A. (1987). *The Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford: Clarendon Press.
- FERNÁNDEZ, C., SZYPERSKI, T., BRUYÈRE, T., RAMAGE, P., MÖSINGER, E. & WÜTHRICH, K. (1997). NMR solution structure of the pathogenesis-related protein P14a. *J. Mol. Biol.* **266**, 576–593.
- FERRIN, T. E., HUANG, C. C., JARVIS, L. E. & LANGRIDGE, R. (1988). The MIDAS display system. *J. Mol. Graph.* **6**, 13–27.
- FISCHMAN, A. J., LIVE, D. H., WYSSBROD, H. R., AGOSTA, W. C. & COWBURN, D. (1980). Torsion angles in the cystine bridge of oxytocin in aqueous solution. Measurements of circumjacent vicinal couplings between ¹H, ¹³C, and ¹⁵N. *J. Amer. Chem. Soc.* **102**, 2533–2539.
- FLETCHER, C. M., JONES, D. N. M., DIAMOND, R. & NEUHAUS, D. (1996). Treatment of NOE constraints involving equivalent or nonstereoassigned protons in calculations of biomacromolecular structures. *J. Biomol. NMR* **8**, 292–310.
- GAYATHRI, C., BOTHNER-BY, A. A., VAN ZIJL, P. C. & MACLEAN, C. (1982). Dipolar magnetic field effects in NMR spectra of liquids. *Chem. Phys. Lett.* **87**, 192–196.
- GONZALEZ, C., RULLMANN, J. A. C., BONVIN, A. M. J. J., BOELEN, R. & KAPTEIN, R. (1991). Toward an NMR R factor. *J. Magn. Reson.* **91**, 659–664.

- GRANT, D. M. & HARRIS, R. K. (eds.) (1996). *Encyclopedia of NMR. Vol. 1: Historical Perspectives*. Chichester, UK: Wiley.
- GRIESINGER, C., SØRENSEN, O. W. & ERNST, R. R. (1985). Two-dimensional correlation of connected NMR transitions. *J. Amer. Chem. Soc.* **107**, 6394–6396.
- GUENOT, J. & KOLLMAN, P. (1992). Molecular dynamics studies of a DNA-binding protein: 2. An evaluation of implicit and explicit solvent models for the molecular dynamics simulation of the *Escherichia coli trp* repressor. *Prot. Sci.* **1**, 1185–1205.
- GÜNTERT, P. (1993). *Neue Rechenverfahren für die Proteinstrukturbestimmung mit Hilfe der magnetischen Kernspinresonanz*. Zürich: Ph.D. thesis ETH 10135.
- GÜNTERT, P. & WÜTHRICH, K. (1991). Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints. *J. Biomol. NMR*, **1**, 446–456.
- GÜNTERT, P., BRAUN, W., BILLETTER, M. & WÜTHRICH, K. (1989). Automated stereospecific ¹H NMR assignments and their impact on the precision of protein structure determinations in solution. *J. Amer. Chem. Soc.* **111**, 3997–4004.
- GÜNTERT, P., BRAUN, W. & WÜTHRICH, K. (1991a). 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.
- GÜNTERT, P., QIAN, Y. Q., OTTING, G., MÜLLER, M., GEHRING, W. J. & WÜTHRICH, K. (1991b). Structure determination of the *Antp(C39 → S)* homeodomain from nuclear magnetic resonance data in solution using a novel strategy for the structure calculation with the programs DIANA, CALIBA, HABAS and GLOMSA. *J. Mol. Biol.* **217**, 531–540.
- GÜNTERT, P., BERNDT, K. D. & WÜTHRICH, K. (1993). The program ASNO for computer-supported collection of NOE upper distance constraints as input for protein structure determination. *J. Biomol. NMR* **3**, 601–606.
- GÜNTERT, P., MUMENTHALER, C. & WÜTHRICH, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298.
- HAVEL, T. F. (1990). The sampling properties of some distance geometry algorithms applied to unconstrained polypeptide chains: a study of 1830 independently computed conformations. *Biopolymers* **29**, 1565–1585.
- HAVEL, T. F. (1991). An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Mol. Biol.* **56**, 43–78.
- HAVEL, T. F. & 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.
- HAVEL, T. F. & WÜTHRICH, K. (1985). An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *J. Mol. Biol.* **182**, 281–294.
- HAVEL, T. F., CRIPPEN, G. M. & KUNTZ, I. D. (1979). The effect of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. *Biopolymers* **18**, 73–82.
- HAVEL, T. F., KUNTZ, I. D. & CRIPPEN, G. M. (1983). Theory and practice of distance geometry. *Bull. Math. Biol.* **45**, 665–720.
- HIROAKI, H., KLAUS, W. & SENN, H. (1996). Determination of the SH₃ domain of human p56 Lck tyrosine kinase. *J. Biomol. NMR* **8**, 105–122.

- HOCKNEY, R. W. (1970). The potential calculation and some applications. *Meth. comput. Phys.* **9**, 136–211.
- HODSDON, M. E., PONDER, J. W. & CISTOLA, D. P. (1996). The NMR solution structure of intestinal fatty acid-binding protein complexed with palmitate: application of a novel distance geometry algorithm. *J. Mol. Biol.* **264**, 585–602.
- HOOFT, R. W., VRIEND, G., SANDER, C. & ABOLA, E. E. (1996). Errors in protein structures. *Nature* **381**, 272–273.
- HU, J. S. & BAX, A. (1997). Determination of ϕ and χ^1 angles in proteins from ^{13}C – ^{13}C three-bond J couplings measured by three-dimensional heteronuclear NMR. How planar is the peptide bond? *J. Amer. Chem. Soc.* **119**, 6360–6368.
- HYBERTS, S. G., MÄRKI, W. & WAGNER, G. (1987). Stereospecific assignments of side-chain protons and characterization of torsion angles in eglin c. *Eur. J. Biochem.* **164**, 625–635.
- JAIN, A., VAIDEHI, N. & RODRIGUEZ, G. (1993). A fast recursive algorithm for molecular dynamics simulation. *J. Comp. Phys.* **106**, 258–268.
- JAMES, T. L. (1994). Strategies pertinent to NMR solution structure determination. *Curr. Opin. Struct. Biol.* **4**, 275–284.
- JEENER, J., MEIER, B. H., BACHMANN, P. & ERNST, R. R. (1979). Investigation of exchange processes by two-dimensional NMR spectroscopy. *J. Chem. Phys.* **71**, 4546–4553.
- KALK, A. & BERENDSEN, H. J. C. (1976). Proton magnetic relaxation and spin diffusion in proteins. *J. Magn. Reson.* **24**, 343–366.
- KANG, H., HINES, J. V. & TINOCO JR., I. (1996). Conformation of a non-frameshifting RNA pseudoknot from mouse mammary tumor virus. *J. Mol. Biol.* **259**, 135–147.
- KAPTEIN, R., ZUIDERWEG, E. R. P., SCHEEK, R. M., BOELEN, R. & VAN GUNSTEREN, W. F. (1985). A protein structure from nuclear magnetic resonance data. *lac* repressor headpiece. *J. Mol. Biol.* **182**, 179–182.
- KARPLUS, M. (1963). Vicinal proton coupling in nuclear magnetic resonance. *J. Amer. Chem. Soc.* **85**, 2870–2871.
- KATZ, H., WALTER, R., SOMORJAY, R. L. (1979). Rotational dynamics of large molecules. *Computers & Chemistry* **3**, 25–32.
- KEEPERS, J. W. & JAMES, T. L. (1984). A theoretical study of distance determinations from NMR. Two-dimensional nuclear Overhauser effect spectra. *J. Magn. Reson.* **57**, 404–426.
- KESSLER, H., GEHRKE, M. & GRIESINGER, C. (1988). Two-dimensional NMR spectroscopy: background and overview of the experiments. *Angew. Chem. Int. Ed.* **27**, 490–536.
- KESSLER, H., GRIESINGER, C., LAUTZ, J., MÜLLER, A., VAN GUNSTEREN, W. F. & BERENDSEN, H. J. C. (1989). Conformational dynamics detected by nuclear magnetic resonance NOE values and J coupling constants. *J. Amer. Chem. Soc.* **110**, 3393–3396.
- KIM, Y. & PRESTEGARD, J. H. (1990). Refinement of the NMR structures for acyl carrier protein with scalar coupling data. *Proteins* **8**, 377–385.
- KIRKPATRICK, S., GELATT JR., C. D. & VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.
- KLINE, A. D., BRAUN, W. & 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.

- KLINE, A. D., BRAUN, W. & 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.
- KNELLER, G. R. & HINSEN, K. (1994). Generalized Euler equations for linked rigid bodies. *Phys. Rev. E* **50**, 1559–1564.
- KOEHL, P., LEFÈFRE, J.-F. & JARDETZKY, O. (1992). Computing the geometry of a molecule in dihedral angle space using n.m.r.-derived constraints. A new algorithm based on optimal filtering. *J. Mol. Biol.* **223**, 299–315.
- KORADI, R., BILLETER, M. & WÜTHRICH, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.
- KORADI, R., BILLETER, M., ENGELI, M., GÜNTERT, P. & WÜTHRICH, K. (1998). Towards fully automatic peak picking and integration of biomolecular NMR spectra. *J. Magn. Reson.*, submitted.
- KRAULIS, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946–950.
- KUMAR, A., ERNST, R. R. & WÜTHRICH, K. (1980). A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Comm.* **95**, 1–6.
- KUNTZ, I. D., CRIPPEN, G. M. & KOLLMAN, P. A. (1979). Application of distance geometry to protein tertiary structure calculations. *Biopolymers* **18**, 939–959.
- KUSZEWSKI, J., NILGES, M. & BRÜNGER, A. T. (1992). Sampling and efficiency of metric matrix distance geometry: a novel partial metrization algorithm. *J. Biomol. NMR*, **2**, 33–56.
- KUSZEWSKI, J., GRONENBORN, A. M. & CLORE, G. M. (1995a). The impact of direct refinement against proton chemical shifts on protein structure determination by NMR. *J. Magn. Reson.* **B107**, 293–297.
- KUSZEWSKI, J., QIN, J., GRONENBORN, A. M. & CLORE, G. M. (1995b). The impact of direct refinement against $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ chemical shifts on protein structure determination by NMR. *J. Magn. Reson.* **B106**, 92–96.
- KUSZEWSKI, J., GRONENBORN, A. M. & CLORE, G. M. (1996). Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci.* **5**, 1067–1080.
- LASKOWSKI, R. A., MACARTHUR, M. W., HUTCHINSON, E. G. & THORNTON, J. M. (1993). PROCHECK: a program to check stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291.
- LASKOWSKI, R. A., RULLMANN, J. A. C., MACARTHUR, M. W., KAPTEIN, R. & THORNTON, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477–486.
- LEVY, R. M., BASSOLINO, D. A., KITCHEN, D. B. & PARDI, A. (1989). Solution structures of proteins from NMR data and modeling: alternative folds for neutrophil peptide 5. *Biochemistry* **28**, 9361–9372.
- LIPARI, G. & SZABO, A. (1982). Model-free approach to the interpretation of nuclear magnetic resonance in macromolecules. 1. Theory and range of validity. *J. Amer. Chem. Soc.* **104**, 4546–4559.
- LOSONCZI, J. A. & PRESTEGARD, J. H. (1998). Nuclear magnetic resonance characterization of the myristoylated, N-terminal fragment of ADP-ribosylation factor 1 in a magnetically oriented membrane array. *Biochemistry* **37**, 706–716.

- LUGINBÜHL, P., SZYPERSKI, T. & WÜTHRICH, K. (1995). Statistical basis for the use of $^{13}\text{C}^{\alpha}$ chemical shifts in protein structure determination. *J. Magn. Reson.* **B109**, 229–233.
- LUGINBÜHL, P., GÜNTERT, P., BILLETER, M. & WÜTHRICH, K. (1996). The new program OPAL for molecular dynamics simulation of biological macromolecules. *J. Biomol. NMR* **8**, 136–146.
- MACURA, S. & ERNST, R. R. (1980). Elucidation of cross relaxation in liquids by 2D NMR spectroscopy. *Mol. Phys.* **41**, 95–117.
- MATHIOWETZ, A. M., JAIN, A., KARASAWA, N. & GODDARD III, W. A. (1994). Protein simulations using techniques suitable for large systems: the cell multipole method for nonbond interactions and the Newton–Euler inverse mass operator method for internal coordinate dynamics. *Proteins* **20**, 227–247.
- MAZUR, A. K. & ABAGYAN, R. A. (1989). New methodology for computer-aided modelling of biomolecular structure and dynamics. I. Non-cyclic structures. *J. Biomol. Struct. Dynam.* **4**, 815–832.
- MAZUR, A. K., DOROFEEV, V. E. & ABAGYAN, R. A. (1991). Derivation and testing of explicit equations of motion for polymers described by internal coordinates. *J. Comp. Phys.* **92**, 261–272.
- MCCAMMON, J. A., GELIN, B. R. & KARPLUS, M. (1977). Dynamics of folded proteins. *Nature* **267**, 585–590.
- MCCAMMON, J. A. & HARVEY, S. C. (1987). *Dynamics of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press.
- McLACHLAN, A. D. (1979). Gene duplication in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79.
- MEADOWS, R. P., OLEJNICZAK, E. T. & FESIK, S. W. (1994). A computer-based protocol for semiautomated assignments and 3D structure determination of proteins. *J. Biomol. NMR* **4**, 79–96.
- MERTZ, J. E., GÜNTERT, P., WÜTHRICH, K. & BRAUN, W. (1991). Complete relaxation matrix refinement of NMR structures of proteins using analytically calculated dihedral angle derivatives of NOE intensities. *J. Biomol. NMR.* **1**, 257–269.
- METROPOLIS, N., ROSENBLUTH, M., ROSENBLUTH, A., TELLER, A. & TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. chem. Phys.* **21**, 1087–1092.
- METZLER, W. J., HARE, D. R. & PARDI, A. (1989). Limited sampling of conformational space by the distance geometry algorithm: implications for structures generated from NMR data, *Biochemistry* **28**, 7045–7052.
- MOMANY, F. A., MCGUIRE, R. F., BURGESS, A. W. & SCHERAGA, H. A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. phys. Chem.* **79**, 2361–2381.
- MUMENTHALER, C. & BRAUN, W. (1995). Automated assignment of simulated and experimental Noesy spectra of proteins by feedback filtering and self-correcting distance geometry. *J. Mol. Biol.* **254**, 465–480.
- MUMENTHALER, C., GÜNTERT, P., BRAUN, W. & WÜTHRICH, K. (1997). Automated procedure for combined assignment of Noesy spectra and three-dimensional protein structure determination. *J. Biomol. NMR* **10**, 351–362.
- NAKAI, T., KIDERA, A. & NAKAMURA, H. (1993). Intrinsic nature of the three-dimensional structure of proteins as determined by distance geometry with good sampling properties. *J. Biomol. NMR* **3**, 19–40.

- NERI, D., SZYPERSKI, T., OTTING, G., SENN, H. & WÜTHRICH, K. (1989). Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional ^{13}C labelling. *Biochemistry* **28**, 7510–7516.
- NERI, D., OTTING, G. & WÜTHRICH, K. (1990). New nuclear magnetic resonance experiment for measurements of the vicinal coupling constants $^3J_{\text{HNz}}$ in proteins. *J. Amer. Chem. Soc.* **112**, 3663–3665.
- NEUHAUS, D. & WILLIAMSON, M. P. (1989). *The nuclear Overhauser effect in structural and conformational analysis*. New York: VCH.
- NICHOLLS, A. J., SHARP, K. A. & HONIG, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296.
- NILGES, M. (1993). A calculation strategy for the structure determination of symmetric dimers by ^1H NMR. *Proteins* **17**, 297–309.
- NILGES, M. (1995). Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.* **245**, 645–660.
- NILGES, M. (1996). Structure calculation from NMR data. *Curr. Opin. Struct. Biol.* **6**, 617–623.
- NILGES, M., CLORE, G. M. & GRONENBORN, A. M. (1988a). Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry–dynamical simulated annealing calculations. *FEBS Lett.* **229**, 317–324.
- NILGES, M., CLORE, G. M. & GRONENBORN, A. M. (1988b). Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. *FEBS Lett.* **239**, 129–136.
- NILGES, M., GRONENBORN, A. M., BRÜNGER, A. T. & CLORE, G. M. (1988c). Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine protease inhibitor 2. *Protein Eng.* **2**, 27–38.
- NILGES, M., CLORE, G. M. & GRONENBORN, A. M. (1990). ^1H -NMR stereospecific assignments by conformational data-base searches. *Biopolymers* **29**, 813–822.
- NILGES, M., KUSZEWSKI, J. & BRÜNGER, A. T. (1991). Sampling properties of simulated annealing and distance geometry. In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (ed. J. C. Hoch, F. M. Poulsen & C. Redfield), pp. 451–455, New York & London: Plenum Press.
- NILGES, M., MACIAS, M., O'DONOGHUE, S. I. & OSCHKINAT, H. (1997). Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β -spectrin. *J. Mol. Biol.* **269**, 408–422.
- OLDFIELD, E. (1995). Chemical shifts and three-dimensional protein structures. *Protein Sci.* **5**, 217–225.
- ÖSAPAY, K., THERIAULT, Y., WRIGHT, P. E. & CASE, D. A. (1994). Solution structure of carbonmonoxy myoglobin determined from nuclear magnetic resonance distance and chemical shift constraints. *J. Mol. Biol.* **244**, 183–197.
- OTTIGER, M., SZYPERSKI, T., LUGINBÜHL, P., ORTENZI, C., LUPORINI, P., BRADSHAW, R. A. & WÜTHRICH, K. (1994). The NMR solution structure of the pheromone Er-2 from the ciliated protozoan *Euplotes raikovi*. *Protein Science* **3**, 1515–1526.
- OTTIGER, M., ZERBE, O., GÜNTERT, P. & WÜTHRICH, K. (1997). The NMR solution conformation of unligated human Cyclophilin A. *J. Mol. Biol.* **272**, 64–81.

- PARDI, A. (1995). Multidimensional heteronuclear NMR experiments for structure determination of isotopically labeled RNA. *Meth. Enzymol.* **261**, 350–380.
- PEARLMAN, D. A., CASE, D. A., CALDWELL, J. C., SEIBEL, G. L., CHANDRA SINGH, U., WEINER, P. & KOLLMAN, P. A. (1991). AMBER 4.0, University of California, San Francisco.
- PEARLMAN, D. A. & KOLLMAN, P. A. (1991). Are time-averaged restraints necessary for nuclear magnetic resonance refinement? *J. Mol. Biol.* **220**, 457–479.
- PELLECCHIA, M., SZYPERSKI, T., WALL, D., GEORGOPOULOS, C. & WÜTHRICH, K. (1996). NMR structure of the J-domain and the Gly/Phe-rich region of the *Escherichia coli* DnaJ chaperone. *J. Mol. Biol.*, **260**, 236–250.
- PERVUSHIN, K., BILLETER, M., SIEGAL, G. & WÜTHRICH, K. (1996). Structural role of the buried salt bridge in the 434 repressor DNA-binding domain. *J. Mol. Biol.*, **264**, 1002–1012.
- PFLUGRATH, J., WIEGAND, E., HUBER, R. & VÉRTESY, L. (1986). Crystal structure determination, refinement and the molecular model of the α -amylase inhibitor Hoe-467A. *J. Mol. Biol.* **189**, 383–386.
- POLSHAKOV, V. I., FRENKIEL, T. A., BIRDSALL, B., SOTERIU, A. & FEENEY, J. (1995). Determination of stereospecific assignments, torsion-angle constraints, and rotamer populations in proteins using the program ANGLESEARCH. *J. Magn. Reson.* **B108**, 31–43.
- POWELL, M. J. D. (1977). Restart procedures for the conjugate gradient method. *Math. Programming* **12**, 241–254.
- PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. & VETTERLING, W. T. (1986). *Numerical Recipes. The Art of Scientific Computing*. Cambridge, UK: Cambridge University Press.
- RAHMAN, A. (1964). Correlations in the motion of atoms in liquid argon. *Phys. Rev.* **A136**, 405–411.
- RICE, L. M. & BRÜNGER, A. T. (1994). Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins* **19**, 277–290.
- RIEK, R., HORNEMANN, S., WIDER, G., BILLETER, M., GLOCKSHUBER, R. & WÜTHRICH, K. (1996). NMR structure of the mouse prion protein domain PrP(121–231). *Nature* **382**, 180–182.
- RYCKAERT, J.-P., CICCOTTI, G. & BERENDSEN, H. J. C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **23**, 327–341.
- SAENGER, W. (1984). *Principles of Nucleic Acid Structure*. New York: Springer.
- SAYLE, R. A. & MILNER-WHITE, E. J. (1995). RASMOL: Biomolecular graphics for all. *Trends. Biochem. Sci.* **20**, 374–376.
- SCHEEK, R. M., VAN GUNSTEREN, W. F. & KAPTEIN, R. (1989). Molecular dynamics simulation techniques for determination of molecular structures from nuclear magnetic resonance data. *Methods Enzymol.* **177**, 204–218.
- SCHEEK, R. M., TORDA, A. E., KEMMINK, J. & VAN GUNSTEREN, W. F. (1991). Structure determination by NMR: The modeling of NMR parameters as ensemble averages. In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (ed. J. C. Hoch, F. M. Poulsen & C. Redfield), pp. 209–217, New York & London: Plenum Press.
- SCHULTZE, P. & FEIGON, J. (1997). Chirality errors in nucleic acid structures. *Nature* **387**, 668.

- SCHULZ, G. E. & SCHIRMER, R. H. (1979). *Principles of Protein Structure*. New York: Springer.
- SENN, H., WERNER, B., MESSERLE, B. A., WEBER, C., TRABER, R. & WÜTHRICH, K. (1989). Stereospecific assignment of the methyl ^1H NMR lines of valine and leucine in polypeptides by non-random ^{13}C labelling. *FEBS Lett.* **249**, 113–118.
- SMITH, B. O., ITO, Y., RAINE, A., TEICHMANN, S., BEN-TOVIM, L., NIETLISPACH, D., BROADHURST, R. W., TERADA, T., KELLY, M., OSCHKINAT, H., SHIBATA, T., YOKOYAMA, S. & LAUE, E. D. (1996). An approach to global fold determination using limited NMR data from larger proteins selectively protonated at specific residue types. *J. Biomol. NMR*, **3**, 19–40.
- SOLOMON, I. (1955). Relaxation processes in a system of two spins. *Phys. Rev.* **99**, 559–565.
- SPERA, S. & BAX, A. (1991). Empirical correlation between protein backbone conformation and $\text{C}\alpha$ and $\text{C}\beta$ ^{13}C nuclear magnetic resonance chemical shifts. *J. Amer. Chem. Soc.*, **113**, 5490–5492.
- STEIN, E. G., RICE, L. M. & BRÜNGER, A. T. (1997). Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *J. Magn. Reson.* **124**, 154–164.
- SUTCLIFFE, M. J. (1993). Structure determination from NMR data. II. Computational approaches. In *NMR of Macromolecules. A Practical Approach* (ed. G. C. K. Roberts), pp. 359–390, Oxford: Oxford University Press.
- SZYPERSKI, T., GÜNTERT, P., OTTING, G. & WÜTHRICH, K. (1992a). Determination of scalar coupling constants by inverse Fourier transformation of in-phase multiplets. *J. Magn. Reson.* **99**, 552–560.
- SZYPERSKI, T., GÜNTERT, P., STONE, S. R. & WÜTHRICH, K. (1992b). The NMR solution structure of hirudin(1–51) and comparison with corresponding three-dimensional structures determined using the complete 65-residue hirudin polypeptide chain. *J. Mol. Biol.* **228**, 1193–1205.
- THOMAS, P. D., BASUS, V. J. & JAMES, T. L. (1991). Protein solution structure determination using distances from 2D NOE experiments: effect of approximations on the accuracy of derived structures. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 1237–1241.
- TJANDRA, N. & BAX, A. (1997). Direct measurements of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278**, 1111–1114.
- TJANDRA, N., GRZESIEK, S. & BAX, A. (1996). Magnetic field dependence of nitrogen-proton \mathcal{J} -splittings in ^{15}N -enriched human ubiquitin resulting from relaxation interference and residual dipolar couplings. *J. Amer. Chem. Soc.* **118**, 6264–6272.
- TJANDRA, N., OMICHINSKI, J. G., GRONENBORN, A. M., CLORE, G. M. & BAX, A. (1997). Use of dipolar ^1H - ^{15}N and ^1H - ^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nature Struct. Biol.* **4**, 732–738.
- TOLMAN, J. R., FLANAGAN, J. M., KENNEDY, M. A. & PRESTEGARD, J. H. (1995). Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9279–9283.
- TORDA, A. E., SCHEEK, R. M. & VAN GUNSTEREN, W. F. (1989). Time-dependent distance restraints in molecular dynamics simulations. *Chem. Phys. Lett.* **157**, 289–294.
- TORDA, A. E., SCHEEK, R. M. & VAN GUNSTEREN, W. F. (1990). Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat. *J. Mol. Biol.* **214**, 223–235.
- TORDA, A. E., BRUNNE, R. M., HUBER, T., KESSLER, H. & VAN GUNSTEREN, W. F.

- (1993). Structure refinement using time-averaged \mathcal{J} -coupling constant constraints. *J. Biomol. NMR* **3**, 55–66.
- ULYANOV, N. B., SCHMITZ, U. & JAMES, T. L. (1993). Metropolis Monte Carlo calculations of DNA structure using internal coordinates and NMR distance restraints: an alternative method for generating a high-resolution solution structure. *J. Biomol. NMR* **3**, 547–568.
- VAN GUNSTEREN, W. F. & BERENDSEN, H. J. C. (1977). Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.* **34**, 1311–1327.
- VAN GUNSTEREN, W. F. & BERENDSEN, H. J. C. (1982). Molecular dynamics: Perspective for complex systems. *Biochem. Soc. Trans.* **10**, 301–305.
- VAN GUNSTEREN, W. F. & BERENDSEN, H. J. C. (1990). Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry. *Angew. Chem. Int. Ed.* **29**, 992–1023.
- VAN GUNSTEREN, W. F., BRUNNE, R. M., GROS, P., VAN SCHAIK, R. C., SCHIFFER, C. A. & TORDA, A. E. (1994). Accounting for molecular mobility in structure determination based on nuclear magnetic resonance spectroscopic and X-ray diffraction data. *Meth. Enzymol.* **239**, 619–654.
- VAN GUNSTEREN, W. F., BILLETER, S. R., EISING, A. A., HÜNENBERGER, P. H., KRÜGER, P., MARK, A. E., SCOTT, W. R. P. & TIRONI, I. G. (1996). *Biomolecular Simulation: the GROMOS96 Manual and User Guide*. Zürich: vdf Hochschulverlag.
- VAN KAMPEN, A. H., BUYDENS, L. M., LUCASIU, C. B. & BLOMMERS, M. J. (1996). Optimisation of metric matrix embedding by genetic algorithms. *J. Biomol. NMR* **7**, 214–224.
- VARANI, G., ABOUL-ELA, F. & ALLAIN, F. H. T. (1996). NMR investigation of RNA structure. *Prog. NMR Spectrosc.* **29**, 51–127.
- VENDRELL, J., BILLETER, M., WIDER, G., AVILÉS, F. X. & WÜTHRICH, K. (1991). The NMR structure of the activation domain isolated from porcine procarboxypeptidase B. *EMBO J.* **10**, 11–15.
- VENTERS, R. A., METZLER, W. J., SPICER, L. D., MUELLER, L. & FARMER II, B. T. (1995). Use of $^1\text{H}_\text{N}$ - $^1\text{H}_\text{N}$ NOEs to determine protein global folds in perdeuterated proteins. *J. Amer. Chem. Soc.* **117**, 9592–9593.
- VERLET, L. (1967). Computer ‘experiments’ on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159**, 98–103.
- VRIEND, G. & SANDER, C. (1993). Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystallogr.* **26**, 47–60.
- VUISTER, G. W., GRZESIEK, S., DELAGLIO, F., WANG, A. C., TSCHUDIN, R., ZHU, G. & BAX, A. (1994). Measurement of homo- and heteronuclear \mathcal{J} couplings from quantitative \mathcal{J} correlation. *Meth. Enzymol.* **239**, 79–105.
- WAGNER, G. & WÜTHRICH, K. (1982). Amide proton exchange and surface conformation of the basic pancreatic trypsin inhibitor in solution. *J. Mol. Biol.* **160**, 343–361.
- WAGNER, G., BRAUN, W., HAVEL, T. F., SCHAUMANN, T., GO, N. & WÜTHRICH, K. (1987). Protein structures in solution by nuclear magnetic resonance and distance geometry. The polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN. *J. Mol. Biol.* **196**, 611–639.
- WANG, A. C. & BAX, A. (1995). Reparametrization of the Karplus relation for $^3\mathcal{J}(\text{H}^2\text{-N})$ in peptides from uniformly $^{13}\text{C}/^{15}\text{N}$ enriched human ubiquitin. *J. Amer. Chem. Soc.* **117**, 1810–1813.
- WANG, A. C. & BAX, A. (1996). Determination of the backbone dihedral angles ϕ in

- human ubiquitin from reparametrized empirical Karplus equations. *J. Amer. Chem. Soc.* **118**, 2483–2494.
- WEBER, P. L., MORRISON, R. & HARE, D. (1988). Determining stereo-specific ^1H nuclear magnetic resonance assignments from distance geometry calculations. *J. Mol. Biol.* **204**, 483–487.
- WIDER, G., LEE, K. H. & WÜTHRICH, K. (1982). Sequential resonance assignments in protein ^1H nuclear magnetic resonance spectra. Glucagon bound to perdeuterated dodecylphosphocholine micelles. *J. Mol. Biol.* **155**, 367–388.
- WIJMEGA, S. S., MOOREN, M. M. W. & HILBERS, C. W. (1993). NMR of nucleic acids; from spectrum to structure. In *NMR of Macromolecules. A Practical Approach* (ed. G. C. K. Roberts), pp. 217–288, Oxford: Oxford University Press.
- WILLIAMSON, M. P. & ASAKURA, T. (1997). Protein chemical shifts. In *Protein NMR techniques* (ed. D. G. Reid), pp. 53–69, Totowa, NJ: Humana Press.
- WILLIAMSON, M. P., HAVEL, T. F. & WÜTHRICH, K. (1985). Solution conformation of proteinase inhibitor IIa from bull seminal plasma by ^1H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.* **182**, 295–315.
- WILLIAMSON, M. P., KIKUCHI, J. & ASAKURA, T. (1995). Application of ^1H NMR chemical shifts to measure the quality of protein structures. *J. Magn. Reson.* **B101**, 63–71.
- WISHART, D. S., SYKES, B. D. & RICHARDS, F. M. (1992). The chemical shift index: A fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **31**, 1647–1651.
- WITHKA, J. M., SRINIVASAN, J. & BOLTON, P. H. (1992). Problems with, and alternatives to, the NMR *R* factor. *J. Magn. Reson.* **98**, 611–617.
- WÜTHRICH, K. (1986). *NMR of Proteins and Nucleic Acids*. New York: Wiley.
- WÜTHRICH, K., WIDER, G., WAGNER, G. & BRAUN, W. (1982). Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J. Mol. Biol.* **155**, 311–319.
- WÜTHRICH, K., BILLETER, M. & BRAUN, W. (1983). Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton–proton distance constraints with nuclear magnetic resonance. *J. Mol. Biol.* **169**, 949–961.
- YIP, P. & CASE, D. A. (1989). A new method for refinement of macromolecular structures based on nuclear Overhauser effect spectra. *J. Magn. Reson.* **83**, 643–648.
- ZUIDERWEG, E. R. P., BILLETER, M., BOELEN, R., SCHEEK, R. M., WÜTHRICH, K. & KAPTEIN, R. (1984). Spatial arrangement of the three α helices in the solution structure of *E. coli lac* repressor DNA-binding domain. *FEBS Lett.* **174**, 243–247.