Journal of Biomolecular NMR, **23:** 271–287, 2002.
KLUWER/ESCOM
© 2002 *Kluwer Academic Publishers. Printed in the Netherlands.*

Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE)

Wolfram Gronwald[a], Sherif Moussa[a], Ralph Elsner[a], Astrid Jung[a], Bernhard Ganslmeier[a], Jochen Trenner[a], Werner Kremer[a], Klaus-Peter Neidig[b] & Hans Robert Kalbitzer[a,*]

[a]*Department of Biophysics and Physical Biochemistry, University of Regensburg, Postfach, D-93040 Regensburg, Federal Republic of Germany and* [b]*Bruker Analytik GmbH, Software Department, Rudolf Plank-Str. 23, D-76275 Ettlingen, Federal Republic of Germany*

Received 18 April 2002; Accepted 1 July 2002

Abstract

Automated assignment of NOESY spectra is a prerequisite for automated structure determination of biological macromolecules. With the program KNOWNOE we present a novel, knowledge based approach to this problem. KNOWNOE is devised to work directly with the experimental spectra without interference of an expert. Besides making use of routines already implemented in AUREMOL, it contains as a central part a knowledge driven Bayesian algorithm for solving ambiguities in the NOE assignments. These ambiguities mainly arise from chemical shift degeneration which allows multiple assignments of cross peaks. Using a set of 326 protein NMR structures, statistical tables in the form of atom-pairwise volume probability distributions (VPDs) were derived. VPDs for all assignment possibilities relevant to the assignments of interproton NOEs were calculated. With these data for a given cross peak with N possible assignments A_i (i = 1,. . . ,N) the conditional probabilities $P(A_i, a|V_0)$ can be calculated that the assignment A_i determines essentially all (a-times) of the cross peak volume V_0. An assignment A_k with a probability $P(A_k, a|V_0)$ higher than 0.8 is transiently considered as unambiguously assigned. With a list of unambiguously assigned peaks a set of structures is calculated. These structures are used as input for a next cycle of iteration where a distance threshold D_{max} is dynamically reduced. The program KNOWNOE was tested on NOESY spectra of a medium size protein, the cold shock protein (TmCsp) from *Thermotoga maritima*. The results show that a high quality structure of this protein can be obtained by automated assignment of NOESY spectra which is at least as good as the structure obtained from manual data evaluation.

Abbreviations: TmCsp – cold shock protein from *Thermotoga maritima*; VPD – volume probability distribution.

Introduction

The use of two-dimensional homonuclear (Wagner and Wüthrich, 1982) and/or three- and four-dimensional heteronuclear edited NOESY spectra (Fesik and Zuiderweg, 1988; Clore and Gronenborn 1991) in conjunction with experiments using the J-coupling for delineating spin systems and their sequential assignments is still the standard method for the structure determination of proteins in solution. However, the necessary assignment of these spectra is quite time consuming and usually the time limiting step in the structure determination process.

Ideally, the NMR structure determination or parts of it should be completely automated or at least should not require an expert for this task. Usually, the sequential resonance assignment is done before the detailed analysis of the structural relevant data. Even if the complete identification of spin systems (Nagayama and Wüthrich, 1981) and their sequential assignments (Wüthrich et al., 1982) has been performed successfully, one has to extract structural restraints from the NMR data. For small and medium-size proteins the dominant source of information is the nuclear Over-

*To whom correspondence should be addressed. E-mail: hans-robert.kalbitzer@biologie.uni-regensburg.de

hauser effect, although additional information is obtained from J-couplings, residual dipolar couplings, and chemical shifts. The assignment of the NOE-cross peaks in two or three dimensional spectra is a tedious and error-prone process simply because the number of cross peaks is very large. Consequently, it appears reasonable to automate at least this part of the spectra evaluation.

One major problem in manual and automated evaluation of NOE data of proteins is the chemical shift degeneracy in the NMR spectra. Therefore, for a high percentage of the NOESY signals no unambiguous assignments can be obtained based on chemical shifts alone. The problem of calculating three dimensional structures with ambiguous restraints was approached by a number of groups. ARIA developed by Nilges (1995) and NOAH developed by Mumenthaler and Braun (1995) are both using an iterative combination of resonance assignments and structure calculations. Both programs list for each peak all possible assignments that are compatible with the resonance assignment using a fixed chemical shift tolerance value. However, differences exist how ambiguous assignments are treated. In ARIA the restraints from these assignments were included as a r^{-6} weighted sum in the NOE target function. In the NOAH approach an unambiguous restraint is calculated for each assignment possibility (Mumenthaler and Braun, 1995; Xu et al., 1999). In both approaches ambiguous assignment possibilities will be judged after the following structure calculations, based on the fit of the corresponding restraints to the obtained structures.

AutoStructure (Y. Huang, R. Tejero, and G.T. Montelione, unpublished data) is an expert system for the iterative NOE assignment using rules that are similar to those used by a human expert in the structure determination process. GARANT developed by Bartels et al. (1997) compares predicted peaks with experimental ones to obtain assignments. Predicted peaks are hereby generated using a set of rules about magnetization transfer pathways and, if available, structural information. Sane (Duggan et al., 2001) is a semi-automated iterative approach for NOE assignment where the user is directly involved in violation analysis.

The ultimate goal in the evaluation of NOE information is to obtain a protein structure from 2D or 3D NOESY data without interference of the expert but with equal or higher quality. This is the aim of KNOWNOE, the program presented in the following. For the automation it can make use of already existing routines which are part of the program AURELIA/AUREMOL (Neidig et al., 1995; Ganslmeier, to be published): Two- and three- dimensional NOESY spectra can be back calculated from a given structure on the basis of the complete relaxation matrix formalism (Görler and Kalbitzer, 1997; Görler et al., 1999), the spectra themselves can be analyzed with the automated routines for peak picking (Neidig et al., 1990), peak integration (Geyer et al., 1995), and signal and artifact recognition based on a Bayesian method (Antz et al., 1995; Schulte et al., 1997).

Material and methods

NMR-samples

The spectra used were recorded from a sample of 1.5 mM cold shock protein (Csp) from *Thermotoga maritima* in 92% H_2O/8% D_2O (v/v), pH 6.5.

NMR spectroscopy

The NMR spectra were measured on a Bruker DMX-800 spectrometer operating at a proton frequency of 800 MHz. The NOESY spectrum (Jeener et al., 1979) was recorded with a mixing time of 160 ms at 303 K. Phase-sensitive detection in the t_1-direction was obtained using time-proportional phase increments (TPPI) (Marion and Wüthrich, 1983). The time domain data set consisted of 512 real data points in the t_1-direction and 2048 complex data points in the t_2-direction. Data were multiplied by a Gaussian (Ferrige and Lindon, 1978) or a shifted square sine-bell filter and baseline-corrected with the routines contained in the program XWINNMR (Bruker). The final size of the real part of the spectrum was 1024 × 2048 data points corresponding to a digital resolution of 9.4 × 4.7 Hz/point. The sequential assignments were taken from Kremer et al. (2001).

Software

The NMR-data were processed with the program XWINNMR (Bruker). Peak picking, integration, Bayesian analysis, back calculation of NOESY spectra and data inspection were performed with the program AUREMOL. Structures for $TmCsp$ were obtained using the CNS 1.0 package (Brünger et al., 1998) and the standard dynamical annealing protocol delivered with CNS. However, the slope of the asymptote of the NOE energy function was decreased from 6.3 to 1.89 MJ $\text{mol}^{-1}\text{nm}^{-1}$ to allow proper folding of the molecules even in the presence of some erroneous NOE assignments. The program for automated NOE assignment (KNOWNOE) is written in standard ANSI C and is implemented in the graphical environment of AUREMOL, a new program package of AURELIA aimed at the automated structure determination of biological macromolecules. A beta-version of AUREMOL is available free of charge from the following web address: http://www.uniregensburg.de/Biophysik/Kalbitzer/software/index.html.

Theoretical considerations and algorithms

General problems

For smaller proteins the distance information contained in NOESY spectra is still the main source of information for the structure calculation. When the spin systems are sequentially assigned, the possible positions of the NMR cross peaks in the NOESY spectra of a protein consisting of N residues are given by a combination of all occurring chemical shifts $\delta_{\text{i,j}}$ (i=1,...,N, j=1,...,J(k), k=1,...,20) with i the position of the residue in the sequence, j(k) the atom j of the residue i consisting of J(k) protons and k the amino acid type of residue i. With ultimate sensitivity a two-dimensional homonuclear NOESY spectrum would thus consist of the set $\{(\delta_{\text{i,j}}, \delta_{\text{i}',\text{j}'})$, i,i′ =1,...,N, j,j′ =1,...,J(k), k=1,...,20; NMR-visible protons$\}$. Three dimensional spectra would accordingly consist of triples of chemical shifts with the additional constraint that only those cross peaks are visible which contain at least one proton coupled to the heteronucleus. In this paper we will only deal with two-dimensional spectra and therefore will limit the discussion to this case.

The automated assignment would be trivial if exactly one cross peak in a given spectrum would correspond to one n-tuple of possible chemical shifts and vice versa. This is not the case, since practical spectra are *incomplete* (peaks are missing), contain *artifacts* (peaks not arising from the protein under consideration), and show *overlap* (degeneracy of chemical shifts). In addition, the resonance frequencies of all relevant spins are usually not known (*incomplete assignment*), may be wrongly determined (*false assignments*) or may not fit exactly to the NOESY spectrum used (*chemical shift variations*). In principle, proteins can occur in different conformational states, a difficulty we will not discuss in the following, but can be treated analogously if the resonances of the other states are also completely assigned.

An optimal automated strategy must tolerate the above difficulties at least to some degree, where the main aim is not a completely correct assignment but a three-dimensional structure which represents a solution consistent with all experimental data. The general assessment of the accuracy and reliability of the obtained structure(s) is important but is not directly part of an automated structure calculation from NOE data. Of course, the obtained structures should be at least as accurate as those obtained from manual data evaluation.

General strategy for automated structure determination from NOESY data

From the difficulties described above the strategy presented in Figure 1 can be derived: (1) Optimal processing of the NOESY spectra including the best possible baseline correction since the quality of the base line determines the number of artifacts at a given intensity level and the number of artifacts the accuracy of the assignments. Here, many different methods were published in the past (e.g., Glaser and Kalbitzer 1986; Mitschang et al., 1990; Saffrich et al., 1993; Koradi et al., 1998), and some of them are implemented in AURELIA (Neidig et al., 1995); (2) the resulting spectrum is automatically peak picked; (3) separation of artifacts and noise from true signals is very important since a too large number of additional false cross peaks leads to an instability in the assignment procedure. We use a Bayesian analysis of the data implemented in AUREMOL (Antz et al., 1995; Schulte et al., 1997). It calculates the conditional probabilities $P(i)$ of the peaks i to be true NMR signals and not noise or artifact peaks. The probabilities $P_{\text{signal}}(i)$ provide a measure of how reliable the peaks i are. Peaks with $P_{\text{signal}}(i)$ values below a user defined threshold $P_{\text{signal}-\text{min}-0}$ can be automatically removed. A reasonable limit would be 0.5 meaning that one cannot decide whether the peak is a signal or not. Since all peaks are still present, optionally higher threshold values $P_{\text{signal}-\text{min}}$ can be used first and then be decreased in later steps of the iterative assignment. Alternative methods were published by Koradi et al. (1998); (4) for the calculation of the structure the cross peaks must be integrated. Since the structure determination should be automated only fully automated procedures as the integration by iterative segmentation (Geyer et al, 1995) implemented in AUREMOL should be used; (5) our approach uses a trial structure that is iteratively refined e.g., an extended strand starting structure. For optimal success of the method the sequential assignments should fit the experimental spectrum as good as possible. Therefore, in AUREMOL routines are implemented that in case of 2D NOESY spectra can adapt the general sequential assignment to the spectrum in use; the steps (6) and (7) concerning the automated assignment of NOE cross peaks and structure calculations will be described in more detail below; (8) for the assessment of the quality of the structure the automated *R*-factor determination implemented in AUREMOL is used, but it has to be supplemented with other more general tools for structure validation as they are e.g., implemented in PROCHECK (Laskowski et al., 1996).

Assignment of ambiguous NOEs

A main difficulty for the automated assignment of NOESY spectra is the ambiguous assignment of NOESY cross peaks. There are three different reasons for the ambiguity: (a) The ambiguity caused by the superposition of more than one cross peak, (b) the ambiguity in the assignment of resolved cross peaks caused by experimental limitations, and (c) the ambiguity caused by an insufficient assignment. In case (a) the chemical shifts ($\delta_{i,j}$, $\delta_{i',j'}$) of cross peak A and ($\delta_{m,l}$, $\delta_{m',l'}$) of cross peak B are identical or almost identical. Such a situation occurs when the chemical shifts are exactly identical or when they cannot be separated under the experimental conditions used. The genuine limitation is the inhomogeneous line width $\Delta\nu_{1/2}$ after filtering of the data in the two directions. For a small protein (molecular mass < 10 kDa) at room temperature $\Delta\nu_{1/2}$ is of the order of 8 to 30 Hz that is at 800 MHz of the order of 0.01 ppm to 0.03 ppm. Two Gaussians separated by the frequency $\Delta\nu$ give just one peak if $\Delta\nu \leq \sqrt{1/2 \ln 2} * \Delta\nu_{1/2}$. A second point is the digital resolution of the frequency domain data DR_1 and DR_2 in frequency dimension 1 and 2. In two-dimensional data the resolution in the indirect dimension influences the total measurement time and is usually of the order of 0.01 ppm. In the direct dimension it can be chosen smaller by a factor of 0.5 to 0.25 without negative effects on the duration of the experiment or the signal-to-noise ratio.

For resolved peaks the main limitation is the digital resolution, since the peak center can be determined

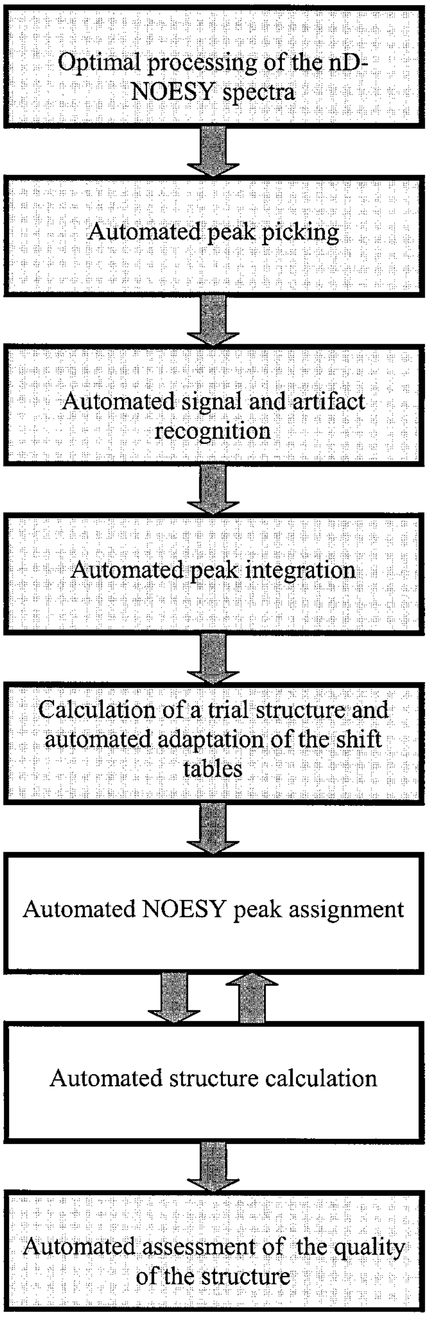
Optimal processing of the nD-NOESY spectra
Automated peak picking
Automated signal and artifact recognition
Automated peak integration
Calculation of a trial structure and automated adaptation of the shift tables
Automated NOESY peak assignment
Automated structure calculation
Automated assessment of the quality of the structure

Figure 1. General scheme for automated structure calculation from NOESY-data.

with arbitrary accuracy under favorable conditions. However, an insufficient signal-to noise ratio reduces the accuracy of the estimated peak position and can lead to an additional shift of the peak position (case b). When peaks are partly superimposed, in two-dimensional spectroscopy of proteins the resolution is limited by the inhomogeneous line widths and is of the order of 0.01 ppm to 0.03 ppm for small proteins. Case (c) is often encountered for methylene protons which are usually not assigned stereo-specifically. In addition, as a rule the assignment table is not complete and some peaks are missing. Computationally, a solution for non-stereospecifically assigned resonances is the introduction of pseudo atoms. Another reason for ambiguities of type (c) is the insufficient precision of the assignment table. It can be largely avoided by carefully adapting the chemical shifts to the NOESY spectrum in use.

If a cross peak has the position (δ_a, δ_b) in the NOESY spectrum, possible assignments are all combinations of atoms j, j′ in residues i,i′ whose chemical shifts fulfill the condition that $\delta_{i,j}$ is an element of the closed interval $[\delta_a - \Delta_1, \delta_a + \Delta_1]$ and $\delta_{i',j'}$ is an element of the closed interval $[\delta_b - \Delta_2, \delta_b + \Delta_2]$. The tolerances Δ_1 and Δ_2 in the indirect dimension 1 and the direct dimension 2 must be greater than the digital resolutions DR_1 and DR_2, respectively, since additional factors discussed above increase the uncertainty.

The ambiguity in chemical shift positions leads to the result that on the basis of chemical shifts alone only a small number of the NOESY cross peaks can be uniquely assigned (Table 1). When omitting side chain side chain contacts, which can barely be resolved in homonuclear 2D-NMR spectra, and assuming a resolution of 0.01 ppm only 12% of all cross peaks can be uniquely assigned from the chemical shifts. This number improves somewhat when the stereospecific assignments are known: here about 15% of all cross peaks can be uniquely assigned (Table 1). With a resolution of 0.03 ppm these numbers drop to approximately 1% of peaks which can be assigned unambiguously.

The addition of structural information reduces the number of ambiguities considerably. The simplest strategy uses all possible distance restraints simultaneously in the structure calculation. Together with weighting the ambiguous NOEs according to the number of possible assignments N_{ab} with the highest weight for $N_{ab} = 1$ usually the general fold can be generated (see e.g., Harrieder, 1998). However, large energies and distorted structures are obtained. By addition of structural information the result is greatly improved. This is done in ARIA (Nilges et al., 1995), where the weighting is performed with the assumption that the cross peak volume is distributed according to the partial volumes calculated from a trial structure. The number of ambiguous NOEs can also be reduced by removing the assignments that are violated in a trial structure as it is done (together with a differential weighting) in NOAH (Mumenthaler and Braun, 1995).

For solving this problem we have devised a new algorithm which uses the distance information contained in a general data base and in iteratively improved trial structures (Figure 2). Starting with a trial structure (e.g., an extended strand) all assignments of a cross peak possible within the chemical shift tolerances Δ_1 and Δ_2 are considered where the corresponding atoms are separated in the trial structure by a distance $r_{ij} \leq D_{max}$. They are stored together with the volume V in the list of unassigned NOEs (U-list). If there is only one assignment possible for a cross peak this assignment is transferred to the list of unambiguously assigned NOEs (A-list). Based on the observation that cross peak volume and correct peak assignment are not independent of each other, the U-list is then searched for cross peaks which can be assigned to more than one pair of atoms from their chemical shifts but where the conditional probability $P(A_i,\ a|V_0)$ is large that most of the volume V_0 of a cross peak originates just from one assignment A_i. More exactly, if A_i ($i = 1, \ldots, N_{ab}$) is a possible assignment of a cross peak, it is transferred to the A-list if

$$P(A_i, a|V_0) \geq P_{min} \tag{1}$$

with $V_{min} = aV_0$ the lower limit of the volume which is explained by the assignment i. With the NOEs of the A-lists and optionally additional restraints like J-coupling restraints, a set of N_s structures is calculated. In the subset of bN_s ($0 < b \leq 1$) structures with the lowest total energies, it is checked whether some NOE restraints are systematically violated. These are removed from the A-list if the difference between the distance r_{calc} determined in the calculated structure and the distance ($r_{\text{exp}} + \Delta^+$) determined from the experimental data is in at least N_x cases larger than the tolerance Δ_{viol} that is

$$r_{\text{calc}} - r_{\text{exp}} - \Delta^+ > \Delta_{viol} \tag{2}$$

with Δ^+ defining the maximum error of r_{exp} allowed in the structure calculation. In the current implementation N_x and Δ_{viol} (typically set to 0.02 nm) have to be specified by the user.

Table 1. General ambiguity of NOE assignments in Csp[a]

N_{ab}	D/ppm	Number of NOE cross peaks with ambiguous assignments								
		Without structure[b]			Extended[c]			Correctly folded[d]		
		A	B	C	A	B	C	A	B	C
1	0.01	8556	7500	11830	268	242	364	548	462	614
	0.02	2256	1836	2850	80	70	100	170	136	176
	0.03	650	518	720	18	16	22	54	40	50
2	0.01	16926	13758	22670	420	368	560	790	654	1100
	0.02	7776	6348	9864	176	152	244	390	302	456
	0.03	3224	2528	3542	86	76	102	196	148	186
3	0.01	6138	4620	10114	148	108	280	282	202	424
	0.02	3360	2814	6624	122	96	200	190	150	326
	0.03	1664	1496	3584	38	30	94	62	50	152
4	0.01	19908	14484	15100	478	398	388	876	710	682
	0.02	11472	9216	12102	302	266	308	564	484	596
	0.03	6642	5206	6060	180	138	150	336	256	290
5	0.01	4278	3156	3202	94	76	74	176	122	158
	0.02	2784	1650	2106	48	32	56	118	68	84
	0.03	1352	1040	1826	34	28	50	82	64	94
6	0.01	12516	8196	9694	280	222	250	506	382	392
	0.02	8454	6646	12648	220	192	320	418	352	538
	0.03	5320	4458	9558	138	130	250	260	246	456
7	0.01	1302	840	390	32	24	12	66	38	22
	0.02	1536	1116	1380	32	30	30	74	58	78
	0.03	832	568	1078	20	20	30	28	26	54
8	0.01	14628	8898	4120	286	200	136	534	360	196
	0.02	9768	7386	6996	222	172	178	378	294	306
	0.03	7652	5928	5108	202	152	140	406	292	264
>8	0.01	62820	24858	9190	1454	674	248	2472	1046	388
	0.02	99666	49298	31740	2258	1302	876	3948	2132	1416
	0.03	119736	64568	54834	2744	1722	1474	4826	2854	2430
≥ 1[e]	–[e]	147072	86310	86310	3460	2312	2312	6250	3976	3976

[a] Shown are the numbers of signals for which at least one possible assignment was found. The resonance assignments of TmCsp were taken from (Harrieder, 1998; Kremer et al., 2001). The number of ambiguous NOEs was calculated for three different sets A, B, C. N_{ab} specifies hereby the number of assignment possibilities for a given experimental crosspeak. Set A considers all possible NOE crosspeaks and for the assignment process it is assumed that the correct stereospecific assignments of the resonances are not known. In comparison to set A all NOEs corresponding to side chain to side chain contacts are excluded in set B. In set C the same NOE cross peaks as in set B are considered, however in this case it is assumed that the stereospecific assignment is known (as given in the assignment table). In the application of the automated assignment described, D is the maximum separation of the chemical shift δ_{ij} of a candidate assignment from the experimental cross peak, that is two resonances can be separated by 2 D and correspond to one cross peaks. For the calculations two resonances δ_{ij} and δ_{ml} are assumed as not distinguishable if $|\delta_{ij} - \delta_{ml}| \leq 2$ D.

[b] No distance cutoff was assumed.

[c] Distance cutoff of 0.5 nm for possible assignments, an extended strand of TmCsp was used as test structure.

[d] Distance cutoff of 0.5 nm for possible assignments, the final structure of TmCsp was used as test structure.

[e] Total number of cross peaks considered.

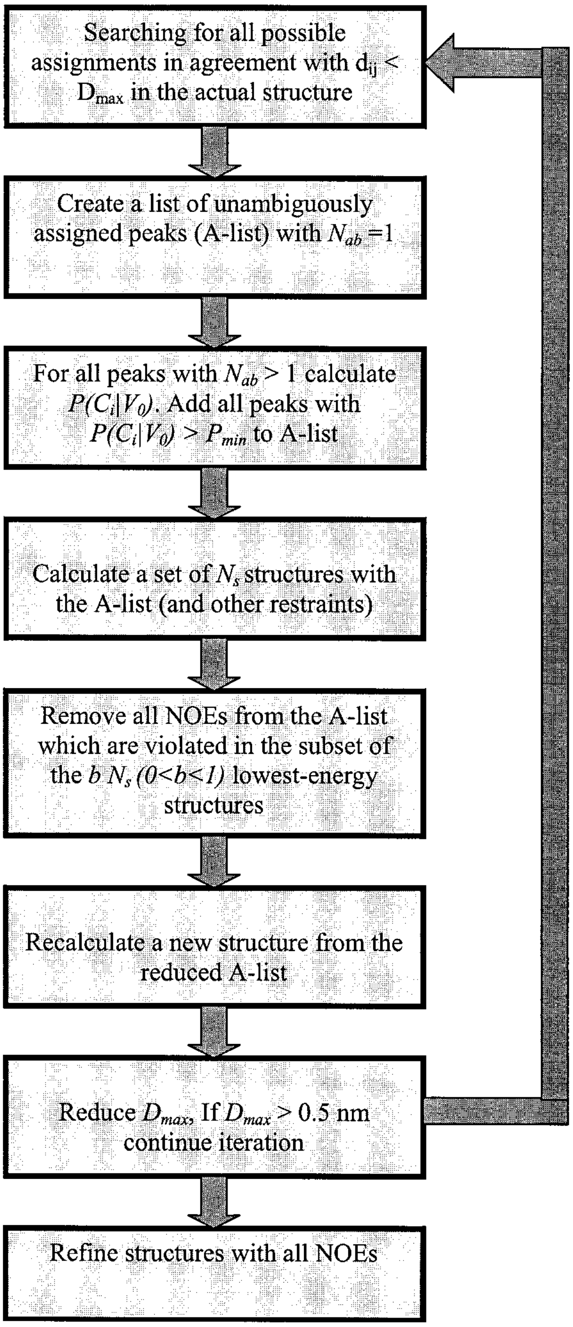
Searching for all possible assignments in agreement with d_{ij} < D_{max} in the actual structure
Create a list of unambiguously assigned peaks (A-list) with N_{ab} =1
For all peaks with N_{ab} > 1 calculate $P(C_i|V_0)$. Add all peaks with $P(C_i|V_0)$ > P_{min} to A-list
Calculate a set of N_s structures with the A-list (and other restraints)
Remove all NOEs from the A-list which are violated in the subset of the b N_s $(0<b<1)$ lowest-energy structures
Recalculate a new structure from the reduced A-list
Reduce D_{max}, If D_{max} > 0.5 nm continue iteration
Refine structures with all NOEs

Figure 2. Schematic representation of the procedure for handling ambiguous NOEs.

With theses restraints a new set of structures is calculated, the maximum distance $D_{\max}$ allowed for assignments is reduced and a new A-list is created as described above. This procedure is iterated until $D_{\max}$ is equal 0.5 nm, in general the maximum detection range of a NOESY spectrum.

After a last iteration with $D_{\max} = 0.5$ nm there is still a large number of cross peaks which can (or must) be explained by more than 1 assignment. At this point, a new list of restraints is created out of the A-list and the U-list. The multiply assigned cross peaks from the U-list with volumes V_0 are all taken as possible solutions but the expectation value of the interatomic distance r_{0i} of the assignment A_i is scaled by

$$r_{0i} = \left(\left\langle \frac{\sum_{j=1}^{N} V_j}{V_0 V_i} \right\rangle \right)^{1/6} \qquad (3)$$

with V_j the volumes corresponding to the distance of the atoms in assignment j. As before the average is performed for the b N_s lowest energy structures. With this complete list of assignments (and the list of restraints other than NOE) a new set of structures is calculated. This procedure is similar to the scaling used by Nilges et al. (1998).

Calculation of the assignment ambiguity

In the above strategy the probability $P(A_i, a|V_0)$ that the assignment A_i explains at least the part a of the experimental cross peak volume V_0 has to be calculated. Starting with Bayes's theorem (Cornfield, 1967, 1969), the probability that more than a-times ($0 \leq a \leq 1$) of the volume V_0 is explained by an assignment A_i can be calculated

$$P(A_i, a|V_0) = \frac{P(A_i, a)P(V_0|A_i, a)}{\sum_{i=1}^{N_{ab}} P(A_i, a)P(V_0|A_i, a)}. \qquad (4)$$

The most simple case occurs if only one assignment A_1 is possible from the chemical shifts. Here, the a priori probability for the assignment $P(A_1, a) = 1$ and $P(A_i, a, i > 1) = 0$ leads to

$$P(A_i, a|V_0) = 1.$$

In case that based on the chemical shifts for a given cross peak only two assignment possibilities A_1 and A_2 exist, the probabilities $P(A_i, a)$ and

$P(V_0|A_i, a)$ must be calculated prior to the calculation of $P(A_i, a|V_0)$. Since no other assignments are assumed possible, the a priori probability for $i > 2$ is given by

$$P(A_i, a) = 0 \qquad (i > 2). \tag{5}$$

For the non-trivial cases i=1 and i=2 $P(A_i, a)$ can be approximated by

$$P(A_1, a) = P(A_2, a) = 0.5\,\mathrm{c_s} \qquad \text{with } 0 \leq \mathrm{c_s} \leq 1 \tag{6}$$

if the expected volumes of the two classes show the same probability distribution. The constant c_s is a normalization constant depending of the form of the probability distribution which is cancelled during the calculation of $P(A_i, a|V_0)$. A more general expression that is going to be used within KNOWNOE can be derived as

$$P(A_1, a) = \int\limits_{V_0=0}^{\infty} \int\limits_{V_1=aV_0}^{V_0} p_1(V_1)p_2(V_0 - V_1)dV_1dV_0 \tag{7}$$

and

$$P(A_2, a) = \int\limits_{V_0=0}^{\infty} \int\limits_{V_2=aV_0}^{V_0} p_1(V_0 - V_2)p_2(V_2)dV_2dV_0 \tag{8}$$

with $p_1(V)$ and $p_2(V)$ the normalized probability densities for finding a volume V for pairs of atoms with the assignments A_1 and A_2, respectively. The probabilities defined by eq. 7 and 8 above are properly normalized when the distributions $p_1(V)$ and $p_2(V)$ are normalized. For two possible assignments the probabilities $P(V_0|A_i, a)$ can be obtained by

$$P(V_0|A_1, a) = \int\limits_{V_1=aV_0}^{V_0} p_1(V_1)p_2(V_0 - V_1)dV_1 \tag{9}$$

and

$$P(V_0|A_2, a) = \int\limits_{V_2=aV_0}^{V_0} p_1(V_0 - V_2)p_2(V_2)dV_2. \tag{10}$$

In case of three assignment possibilities for a cross peak, $P(A_i, a)$ can be defined analogously to equations (7) and (8)

$$P(A_i, a) = 0 \qquad (\mathrm{i} > 3) \tag{11}$$

$$P(A_i, a) = c_s/3 \quad \text{with } 0 \leq c_s \leq 1 \text{ for } 1 \leq \mathrm{i} \leq 3.$$

For $P(V_0|A_i, a)$ one obtains

$$P(V_0|A_1, a) = \tag{12}$$

$$\int\limits_{V_1=aV_0}^{V_0} \int\limits_{V_2=0}^{V_0-V_1} p_1(V_1)p_2(V_2)p_3(V_0 - V_1 - V_2)dV_2dV_1.$$

The probabilities for finding a volume in the interval $[aV_0, V_0]$ for the assignments A_2 and A_3 are calculated in a similar fashion that is

$$P(V_0|A_2, a) = \tag{13}$$

$$\int\limits_{V_2=aV_0}^{V_0} \int\limits_{V_1=0}^{V_0-V_2} p_1(V_1)p_2(V_2)p_3(V_0 - V_1 - V_2)dV_1dV_2.$$

and

$$P(V_0|A_3, a) = \tag{14}$$

$$\int\limits_{V_3=aV_0}^{V_0} \int\limits_{V_1=0}^{V_0-V_3} p_1(V_1)p_2(V_0 - V_1 - V_3)p_3(V_3)dV_1dV_3.$$

For more than three assignment possibilities eq. 7 to 14 can easily be generalized which should not be done explicitly here.

Scaling of experimental volumes

For performing the calculations mentioned above the experimental volumes are scaled to adjust to the expected probability distributions of the volumes. We use in the actual version of KNOWNOE a manual approach where the user has to specify a reference volume that corresponds to a certain distance.

Calculation of the probability distributions

An estimate of the probability distributions p_i of the peak volumes for the assignments A_i is required for the calculation of $P(V_0|A_i, a)$. Although it is possible to formulate a priori assumptions on these distributions, the better way is the extraction of statistical data from known protein structures. For obtaining meaningful distributions one has to classify the specific assignments A_i of pairs of atoms to obtain a sufficiently high number of class members for the statistical analysis. A powerful way is to extract the information independently of the absolute positions in the sequence S_i and $S_{\mathrm{i}'}$. With the knowledge of the absolute position S_i of one amino acid of the pair of atoms considered, the pairwise interaction of any atoms in the protein can be described by the separation in the sequence $\Delta S_\mathrm{i} \;\; = S_{\mathrm{i}'} - S_\mathrm{i}$ (without

Table 2. Assignment classes used for the calculation of probability distributions[a]

Assignment	ΔS_i	Z_j	T_i	$Z_\mathrm{j'}$	$T_\mathrm{i'}$
Intraresidual ($T_\mathrm{i} = T_\mathrm{i'}$)	0	HN ($Z_\mathrm{j} = 1$)[b]	1,...,20	$Z_\mathrm{j} > 1$[b]	1,...,20
	0	$2 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]	ring protons	aromatics[e]
Sequential	1	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	HN ($Z_\mathrm{j} = 1$)[b]	averaged[b]
	1	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	Hα($Z_\mathrm{j} = 1$)[b]	averaged
	1	$2 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]	ring protons	aromatics[e]
	1	ring protons	aromatics[e]	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]
Medium range	2	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	HN ($Z_\mathrm{j} = 1$)[b]	averaged[c]
	2	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	Hα($Z_\mathrm{j} = 1$)[b]	averaged[c]
	2	$2 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]	ring protons	aromatics[e]
	2	ring protons	aromatics[e]	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]
	3	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	HN ($Z_\mathrm{j} = 1$)[b]	averaged[c]
	3	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	Hα ($Z_\mathrm{j} = 1$)[b]	averaged[c]
	3	$2 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]	ring protons	aromatics[e]
	3	ring protons	aromatics[e]	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]
	4	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	HN ($Z_\mathrm{j} = 1$)[b]	averaged[c]
	4	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	Hα ($Z_\mathrm{j} = 1$)[b]	averaged[c]
	4	$2 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]	ring protons	aromatics[e]
	4	ring protons	aromatics[e]	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]
Long range	> 4 averaged	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	HN ($Z_\mathrm{j} = 1$)[b]	averaged[c]
	> 4 averaged	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged (no ring protons)[d]	Hα ($Z_\mathrm{j} = 1$)[b]	averaged[c]
	> 4 averaged	$2 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]	ring protons	aromatics[e]
	> 4 averaged	ring protons	aromatics[e]	$1 \leq Z_\mathrm{j} \leq Z_\mathrm{max}$[b]	averaged[c]

[a] Probability tables were calculated from 326 protein structures from the PDB data base (for a list of the structures used see Moussa, 2001).

[b] For each given value of Z a separate class was calculated.

[c] Data were averaged over corresponding atoms in all 20 amino acids.

[d] Data were averaged over corresponding atoms in all 20 amino acids but ring protons of the aromatic residues His, Trp, Phe, and Tyr were excluded.

[e] aromatic residues His, Trp, Phe, and Tyr.

Table 3. Iterative assignment of NOEs and automated structure calculation of *Tm*Csp[a]

Iteration #	U_0-list	$P_\mathrm{signal-min}$	Δ_1/ppm	Δ_2/ppm	D_max/nm	a	$P(V_0 \mid A_i, a)_\mathrm{min}$	U-list	A-list	RMSD (nm)
1	916	0.4	0.01	0.01	∞	0.9	0.8	5868	283	0.32
									254	0.14
2		0.4	0.01	0.01	1.25	0.9	0.8	3150	474	0.21
									393	0.13
3		0.4	0.01	0.01	1.00	0.9	0.8	2614	530	0.11
									452	0.10
4		0.4	0.01	0.01	0.75	0.9	0.8	2018	591	0.08
									521	0.05
5		0.4	0.01	0.01	0.50	0.9	0.8	1460	599+86[b]	0.05
									645	0.03

[a] U_0-list, number of cross peaks identified at a probability threshold $P_\mathrm{signal-min}$ (symmetry related cross peaks are not counted); Δ_1 and Δ_2 are the chemical shift tolerances in δ_1 and δ_2-dimension; D_max the maximum interatomic distance where assignments are accepted; $P(V_0 \mid A_i, a)_\mathrm{min}$ is the probability that more than a-times of the peak volume is explained by a single assignment; the U-list contains the number of possible ambiguous assignments, the A-list the number of unambiguous assignments used for the calculation. The first row in each iteration cycle shows the number of the primarily used NOEs and the second row the number of unique assignments after removal of conflicting NOEs.

[b] 86 additional restraints were obtained from the analysis of the ambiguous NOEs.

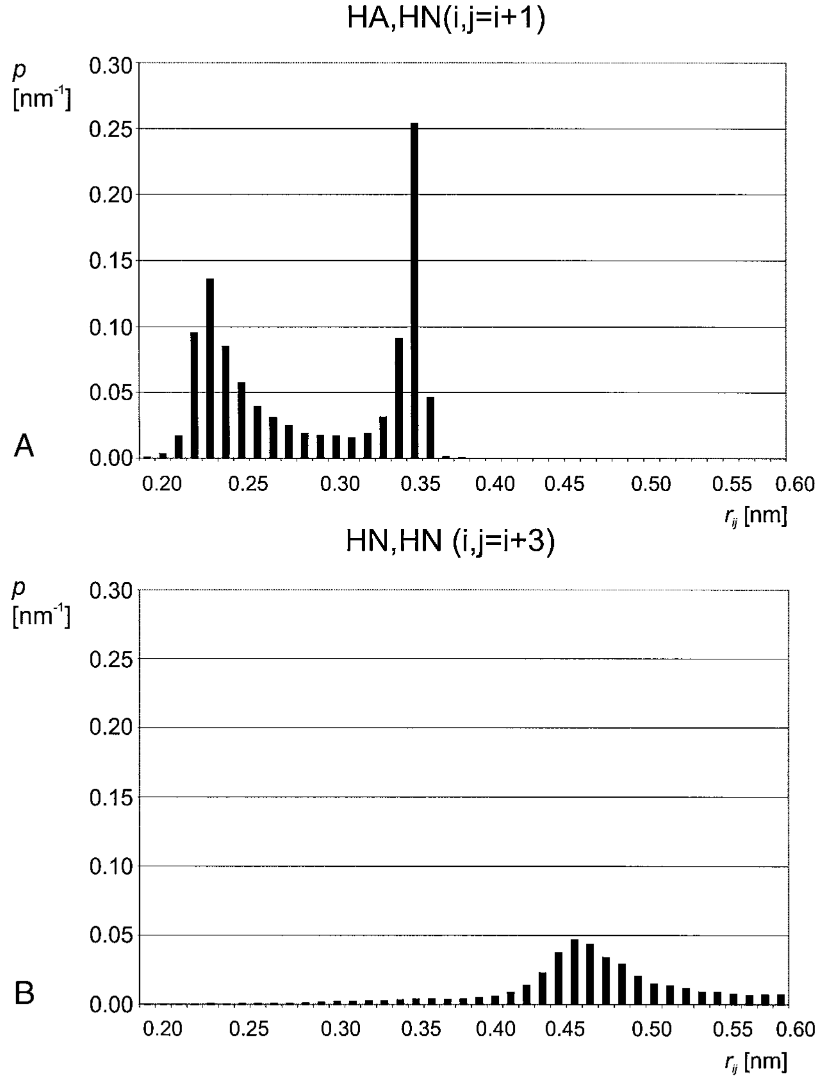

Figure 3. Probability distribution of the distances. (A) Probability density p of the distances r between Hα of a residue in position i and an HN in position i+1. (B) Probability density p of the distances r between HN of a residue in position i and an HN in position i+3.

restricting the generality we assume in the following $S_i \leq S_{i'}$), and by the atom numbers Z_j and $Z_{j'}$. The total sequence information can be coded if in addition the residue types T_i and $T_{i'}$ of amino acids are stored. An assignment A_k can be stored as a vector $A_k = (S_i, \Delta S_i, Z_j, Z_{j'}, T_i, T_{i'})$.

If we create sequence independent classes C_l (l $= 1, \ldots$,L) defined by the sequence independent information (ΔS_i, Z_j, $Z_{j'}$, T_i, $T_{i'}$), A_k can be written in the reduced form $A_k = (S_i, C_l)$. In our case the probability distribution p_k(V) of the volume V of a possible assignment A_k can be approximated by the probability distribution $\tilde{p}_k(V)$ of the corresponding class C_l. The same notation can be used for other purposes as well as it has been done for example in the knowledge based structure prediction published by (Subramaniam et al., 1996).

Results

Calculation of probability tables

As data basis for the statistics 326 protein structures from the PDB databank were taken (for a list of the structures used see Moussa, 2001). Only NMR structures of water soluble proteins containing no paramag-

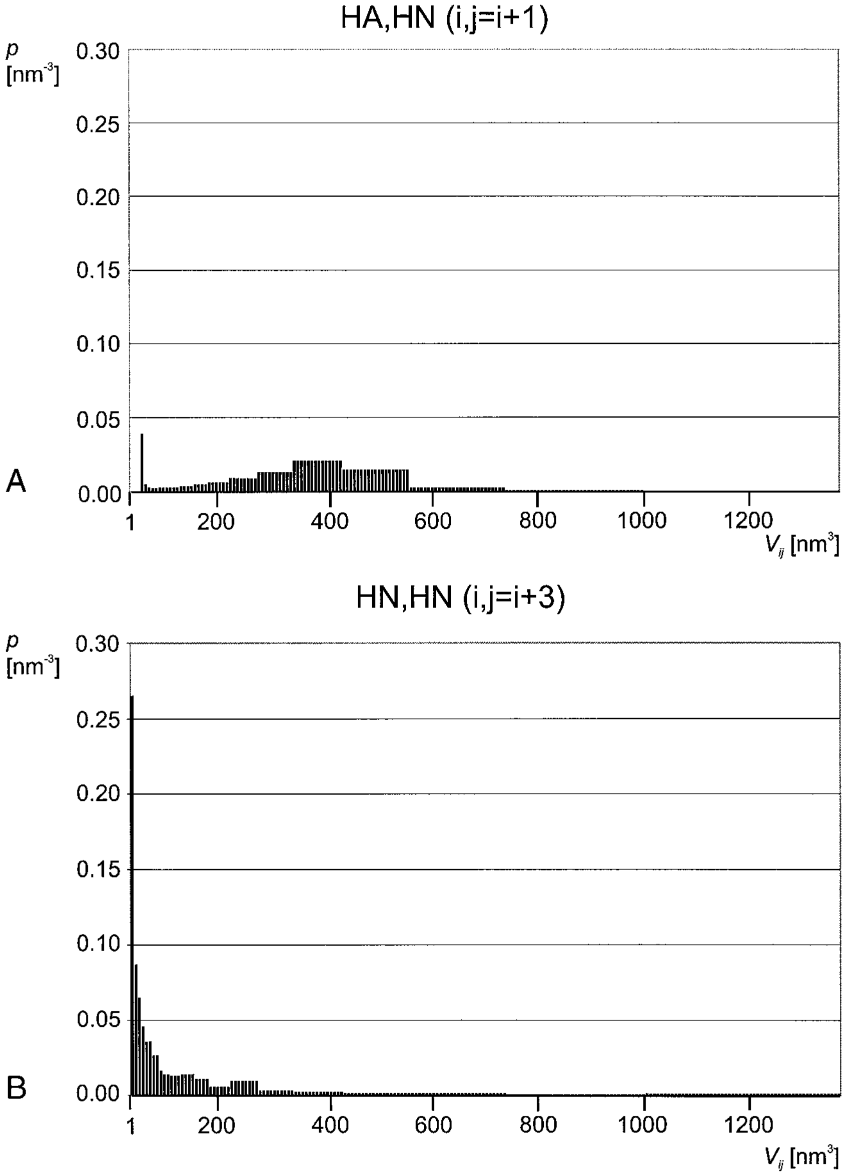

Figure 4. Probability distribution of the volumes. (A) Probability density p of the expected normalized volumes V of a cross peak between Hα of a residue in position i and an HN in position i+1. (B) Probability density p of the expected normalized volumes V of a cross peak between HN of a residue in position i and an HN in position i+3. Note that in this case the volume distribution was calculated directly from the distance contribution, leading to a volume dependent resolution in the volume space.

netic center or larger cofactor were selected. No RNA and DNA structures or complex structures of proteins with RNA or DNA were considered. Using these structures 1577 different assignment class probability tables were calculated containing the corresponding distance distributions. The probability tables were normalized to an integral of 1. Two examples of distance probability distributions (DPDs) are shown in Figure 3. Volumes V_{ij} were calculated from the distances r_{ij} between atoms i and j by the relation $V_{\mathrm{ij}} = c_{\mathrm{V}} r_{\mathrm{ij}}^{-6}$ with c_{V} arbitrarily set to 0.047 nm^9 so that the detection limit of 0.6 nm corresponds to a volume of 1.0 nm^3 and the closest approach of two protons of 0.18 nm to a relative volume of 1382 nm^3. In principle one can recalculate the volume probability distributions (VPDs) from the DPDs as we did for this study. The result is depicted in Figure 4 for the same pairs of atoms shown in Figure 3. The disadvantage of this

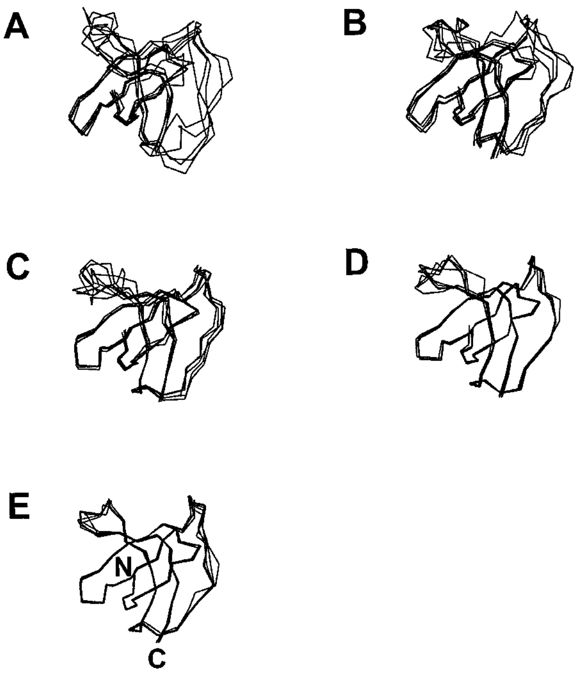

Figure 5. Structural bundles obtained during the iterative structure calculation. The 5 lowest energy structures of *Tm*Csp obtained in different phases of the automated structure calculation are superposed. Only the Cα-atoms are traced. First iteration (A), second iteration (B), third iteration (C), fourth iteration (D), fifth iteration (E). For each iteration the structures after removal of violated NOEs are shown.

procedure is that the resolution (the basic widths of the volume classes) is now dependent on the volume. It is obviously more appropriate to calculate the volume distributions directly from the basic data sets as will be done in the next implementation of KNOWNOE. Nevertheless, it is clear that the two DPDs and VDPs strongly distinguish between the two possible assignments. Probability tables were only calculated for interproton contacts since only these are detected in standard NOESY spectra of proteins. For reducing the number of classes some more general assignment classes were defined (Table 2).

As an example we have used the calculated probability distributions and have estimated $P(A_i, a|V_0)$ for a volume V_0 which coresponds to an distance of 2.3 nm with a cutoff a of 0.9 and have allowed three different assignments A_i for the cross peak. If A_1 corresponds to the sequential NOE between Hα amino acid i and HN of amino acid i+1, A_2 to the intermediate range NOE between Hα amino acid i and HN of amino acid i+3, and A_3 to the long range NOE between Hα amino acid i and HN of amino acid i+7 than the probability that more than 90% of the cross peak volume is explained by assignment A_1 is 0.999. This example clearly shows that often unambiguous results can be obtained by the use of the knowledge based approach.

Application of KNOWNOE to proteins

In the first application of KNOWNOE to proteins it was investigated whether it was possible to automatically determine the solution structure of *Tm*Csp at 303 K without a model structure and using only an experimental 2D NOESY spectrum of *Tm*CSP in H_2O. As described above the experimental spectrum was automatically peak-picked and artifacts and noise were removed as described in the algorithm section. Primarily 7341 peaks were obtained in the region from 10.5 ppm to 3.7 ppm in F2 and from 10.5 ppm to −0.1 ppm in F1, discarding the region containing sidechain-sidechain NOE contacts. This selection takes into account that in an experimental 2D NOESY spectrum the region of the sidechain-sidechain NOE contacts is usually too crowded to be analyzed. In addition, a region between 4.6 ppm and 5.0 ppm in F2 enclosing the water signal was excluded from the peak search. After the Bayesian analysis and the removal of symmetry related signals 916 cross peaks remained with a probability $P_{\text{signal}}(\text{i})$ larger than the probability threshold $P_{\text{signal-min-0}}$ of 0.4 to be a true signal and not noise or artifact. These peaks were transferred to the U_0-list which primarily contains the peak coordinates in ppm, the volumes, and the signal probabilities $P_{\text{signal}}(\text{i})$.

In a next step, the sequential assignments of *Tm*Csp (Harrieder et al., 1998; Kremer et al., 2001) were automatically adapted to the actual NOESY spectrum. With the strategy presented in Figure 2 five iteration cycles were performed. The relevant parameters are shown in Table 3. Starting with an extended structure, a maximum allowed distance D_{max} of 30 nm (greater than the dimensions of the unfolded protein), a signal probability $P_{\text{signal-min}} = P_{\text{signal-min-0}}$ and chemical shift tolerances Δ_1 and Δ_2 of 0.01 ppm the U-list was generated. 166 peaks could be assigned only to one set of chemical shifts and therefore were transferred directly to the A-list, the list of the unambiguously assignable peaks. Of these were 49 intraresidual, 45 sequential, 17 medium range, and 55 long range signals. Using an a-value of 0.9 and a lower limit of 0.8 of the probability $P(V_0|A_i, a)$ for an unambiguous assignment further 117 peaks could be assigned unambiguously and be transferred to the

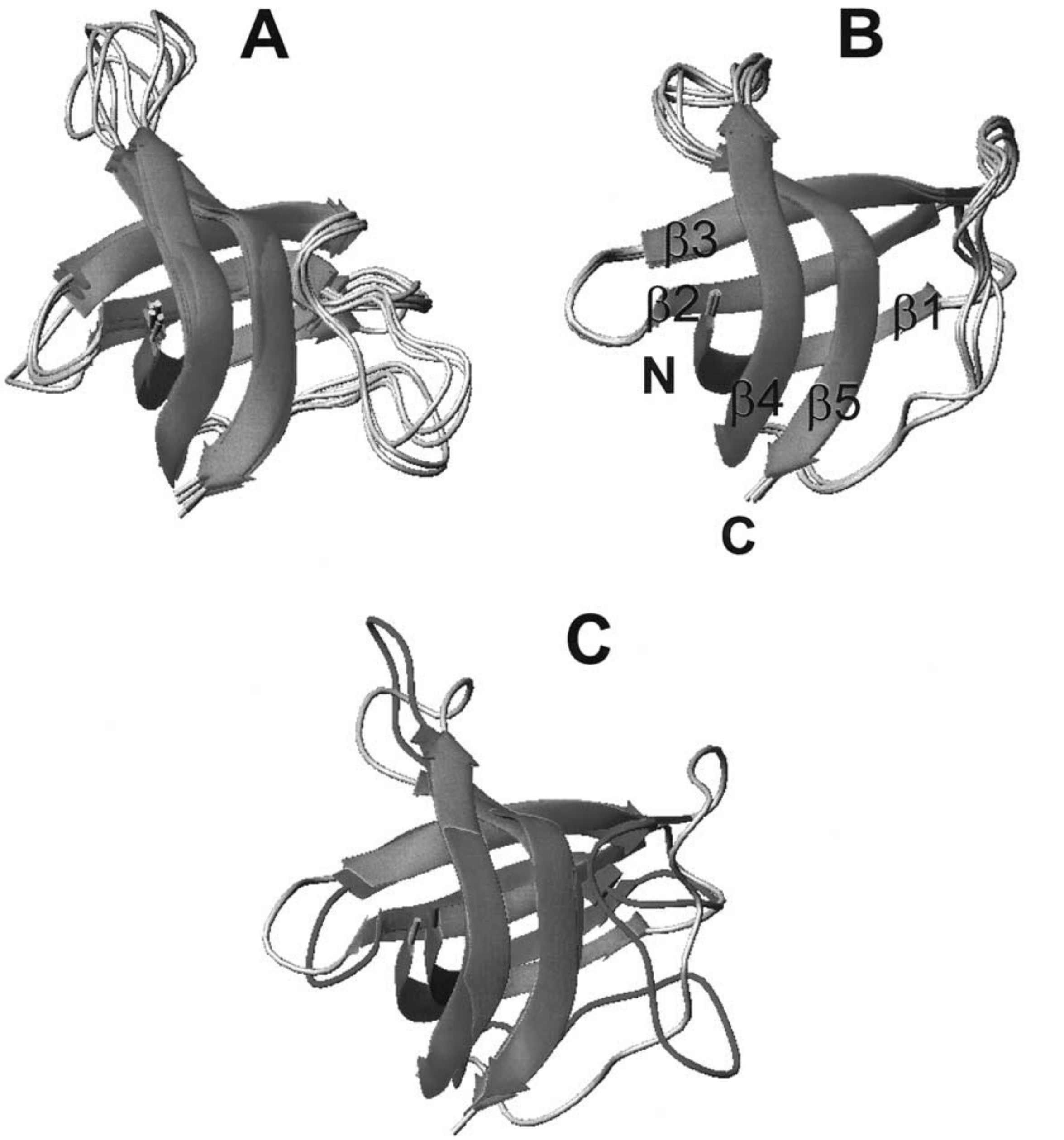

Figure 6. Comparison of the structures obtained by manual and automated NOE assignment. Superposition of the 5 lowest energy structures calculated from automated (A) and manual (B) assignment of NOEs. (C) Superposition of the lowest energy structure obtained by manual and automated assignment.

A-list. The *A*-list now contained 283 automatically assigned NOE signals that were used to automatically create a restraint file adapted to the formats of the molecular dynamics program CNS. The expectation values r_{exp} of the interatomic distances were calculated on the basis of the initial slope approximation from the experimental data. The lower error bounds Δ^- were set to the proton van der Waals distance ($\Delta^- = r_{\mathrm{exp}} - 0.18$ nm) (Wüthrich, 1986) while the upper error bounds Δ^+ were set to 0.125 r_{exp}^2 as proposed by Nilges and O'Donoghue (1998). In addition to the NOE restraints 28 hydrogen-bonds and 42 ϕ-angle restraints were used in the following structure calculations. Using CNS $N_s = 50$ structures were

Table 4. Comparison of relevant structural parameters of the manually and automatically determined structures of *Tm*Csp

	Manual	Automated
NOEs	933	645
Intraresidual (i, i)	358	294
Sequential ($i, i + 1$)	283	172
Backbone-backbone	125	76
Backbone-side chain	153	91
Side chain-side chain	5	5
Intermediate range ($i, i + n; n \leq 5$)	47	26
Backbone-backbone	8	6
Backbone-side chain	29	15
Side chain-side chain	10	5
Long range ($i, j; j > i + 5$)	245	153
Backbone-backbone	89	36
Backbone-side chain	98	79
Side chain-side chain	58	38
ϕ-Angle constraints	28	28
Hydrogen bonds	54	54
Energies (kJ mol^{-1})		
E_{total}	8948.2 ± 277.8	1277.9 ± 18.6
E_{NOE}	4364.2 ± 207.7	276.1 ± 11.4
E_{dihed}	199.2 ± 23.5	18.9 ± 2.8
E_{bond}	518.6 ± 30.7	60.9 ± 2.0
E_{angle}	1844.0 ± 190.1	484.5 ± 8.8
E_{vdW}	1640.5 ± 86.1	276.1 ± 11.4
E_{imp}	381.8 ± 60.1	83.8 ± 1.8
rmsd (nm) Residues 1 – 66		
Backbone atoms (N, Cα, C′)	0.100	0.032
Cα	0.104	0.034
all heavy atoms	0.167	0.077

calculated using a standard simulated annealing protocol and the five best structures in terms of total energy were selected for further analysis. The selected structures were then automatically screened for NOE violations. All restraints that were violated by more than $\Delta_{\text{viol}} = 0.02$ nm (see Equation 2) in at least two of the selected structures were automatically removed from the restraint file. It should be noted that the violated restraints of the previous step were only removed from the restraint file but not from the U_0-list. Therefore, the corresponding signals could be reassigned in the next iteration. Here, 29 restraints were removed and 50 new structures were calculated with the updated set of restraints. Again the 5 best structures in terms of total energy were selected and superimposed. The rmsd value for the Cα atoms of the selected structures to the mean structure was 0.14 nm. The automatically obtained structures are already showing the correct fold after the first iteration (Figure 5).

From the obtained structures the one possessing the lowest total energy was selected for (re)assigning the NOEs in the next iteration cycle. The maximum distance D_{max} was set to 1.25 nm. KNOWNOE assigned 474 NOEs automatically and structures were calculated and analyzed as described for the previous steps. Due to NOE violations 81 out of the 424 restraints were removed and 50 new structures were calculated. The rmsd value for the Cα atoms to the mean structure for the five best structures in terms of total energy was 0.13 nm.

In the third iteration the procedures of iteration 2 were repeated with the difference that a distance cutoff D_{max} of 1.00 nm was used. 530 restraints were automatically assigned of which 78 were automatically removed and the rmsd value for the Cα atoms of the five selected structures was 0.10 nm.

In the fourth iteration a distance cutoff D_{max} of 0.75 nm was used, yielding 591 restraints of which 70 were removed due to NOE violations and the rmsd value for the Cα atoms of the five selected structures was 0.05 nm.

The final iteration consists of the refinement procedure described above using a distance cutoff D_{max} of 0.50 nm. The 599 assignments contained in the final A-list were retained and in addition ambiguous assignments of the U-list were also converted into restraints. Please note that in this test case a D_{max} value of 0.50 nm was used for the ambiguous assignments as well. To obtain these restraints the volume of each of these peaks was distributed according to the corresponding distances of the corresponding ambiguous assignments in the best structure of the previous iteration. This procedure yielded an additional 86 restraints. As described above structures were calculated and due to NOE violations 40 restraints were removed. The removed restraints included 21 intraresidual, 13 sequential, 0 medium range, and 6 long range restraints. Using this final set of 645 NOE restraints structures were calculated and the rmsd for the five best structures was 0.034 nm. Note that in Table 3 now the number of assignments are listed whereas for the first iterations only the number of peaks were counted.

For judging the quality of the assignments obtained by KNOWNOE the manually determined structure of *Tm*Csp was recalculated using the original manually determined NOE assignments but the same dihedral- and hydrogen-bond restraints files and the same CNS protocol as for the automatically determined structures. The inter atomic distances and upper and lower bonds were automatically calculated from the same NOESY-spectrum as used before. Since the original structure (Kremer et al., 2001) was calculated with the aid of a full set of two-dimensional and three-dimensional NOEs in H_2O and D_2O a number of NOEs could not be assigned in the single NOESY spectrum in H_2O. However, also for these NOEs upper and lower bounds were calculated as described before. As for the automatically determined structures 50 structures were calculated and the best five in terms of total energy were selected for further analysis. Figure 6 shows five of the manually determined recalculated structures of *Tm*Csp at 303 K together with a superposition of the five final structures obtained by automated assignment. The superposition of the lowest-energy structure from automated assignment and manual assignment shows that for the regular secondary structure elements virtually the same structures are obtained by the two methods.

Discussion

The results clearly show that for TmCsp the correct fold was obtained using KNOWNOE and CNS in a multistep procedure. A detailed comparison of the manually recalculated (see results section for details) and the automatically determined structures of *Tm*Csp (in the following referred to as the manual and automated case, respectively) shows that the regions comprising the regular secondary structure elements i.e., the five stranded β-barrel, are virtually identical. Minor differences can be observed for the loop regions, e.g., for the long surface loop (Ser30 - Glu42) connecting β-strands three and four and for the loop connecting β-strands four and five. It should be noted that these are the regions for which the greatest variability in the manually determined structures was observed. Comparing the numbers of automatically and manually assigned NOEs (Table 4) it becomes clear that roughly 30% more NOEs were manually assigned. That is mainly due to the fact that in the manual case two homonuclear 2D NOESY spectra measured in H_2O and D_2O and one ^{15}N edited 3D NOESY spectrum measured in H_2O were used, while in the automated case only a homonuclear 2D NOESY measured in H_2O was used. The general distribution of NOEs, e.g., long-range NOEs to medium-range, to sequential and to intraresidual NOEs is in both cases similar. However, when comparing the corresponding rmsd values, substantial lower values are obtained for the automated case (0.034 nm in the automated case versus 0.104 nm in the manual case) indicating that a sufficient number of NOE restraints has been obtained in this case to allow a precise structure determination. When comparing the energies of the final five structures obtained each in the manual and automated case (Table 4) it is found that in the automated case in general smaller energies were obtained. The overall and NOE energies obtained in the automated case are roughly 86% and 94% smaller than the energies obtained in the manual case indicating a probable connection between the number of assignments and the obtained energies. The calculation of R-factors is a reliable method to judge how well the obtained structures fit the experimental data (Gronwald et al., 2000). For *Tm*CSP we have calculated R-factors for

Table 5. R-factors for *Tm*CSP

	KNOWNOE[a]	Manual_1[b]	Manual_2[c]
R-factor	0.35	0.35	0.36

[a] Structures were calculated based on the NOE assignments obtained by KNOWNOE.
[b] Recalculated structures obtained using the manually determined NOE lists and the same CNS protocol as for the automatically determined structures.
[c] One of the originally published structures.
In all cases the same peak list and the sequential assignments that were used for KNOWNOE were used in the R-factor calculation. We used the global R-factor that is based on the long-range and the non assigned signals, since this R-factor definition is most sensitive to structural changes (Gronwald et al., 2000).

the automated case, the manual case, and for the NMR structure submitted to the protein data bank (Kremer et al., 2001). In each case the same peak list from the automated assignment was used. For the comparisons in this paper we calculated in each test case the global R-factor for the structure possessing the lowest total energy including the nonassigned signals of the peak-list since this R-factor is most sensitive to global structural changes (Gronwald et al., 2000). The results (Table 5) show that for all cases very similar R-factors were obtained indicating that all structures fit the experimental data equally well. To analyze the structural differences visible in the loop regions of the manually and automatically determined structures we performed a more detailed residue by residue R-factor analysis, that included for each residue all corresponding experimental and simulated NOEs (data not shown). This analysis showed that especially in the long loop region from Ser30 to Glu42 lower R-factors were obtained for the structures solved by KNOWNOE indicating a possible missassignment in the manually determined structures in this region. It should be noted that in this loop region the number of structure relevant long range NOEs is relatively small compared to the number of intraresidual and sequential NOEs. Therefore, a small change in R-factors might reflect a large structural change in this region. The manually determined structures show to some degree better local R-factors in the regions of the regular secondary structure elements. It can be assumed that due to the larger number of used restraints obtained from additional NMR experiments these regions are still better defined in the manual case, although the regions of the regular secondary structure elements are very similar in both cases. Since in these regions the number of the structural relevant long range NOEs is relatively high compared to the number of sequential and intraresidual NOEs, a small structural change might be indicated by a substantial change in R-factors. In summary we can say that for *Tm*Csp the automatically determined structures are better defined in the loop regions while the definition of the regular secondary structure elements is still a bit better in the manually determined structures.

Here it is important to note that the structures obtained using KNOWNOE were solved in a fraction of the time required for the manual structure determination with a minimal set of structurally relevant spectra.

In the tests of KNOWNOE described above the solution structure of the molecule was determined using no model structure of, e.g., a homologous protein, which is quite a challenging test for the program. In this regard it should be noted that KNOWNOE works best using a complete or almost complete sequential assignment where all main chain and side chain resonances have been assigned. In addition, it is preferable that the sequential assignment fits the spectra in use as good as possible, enabling the user to use small chemical shift tolerance values. This is especially important for the first iteration where no model structure is available. In this case large chemical shift tolerance values usually result in many assignment possibilities for a given signal. However, if the number of possible assignments exceeds three, this signal will be excluded by KNOWNOE leading to a relatively small number of NOE restraints. Therefore, in unfortunate cases it is possible that for large chemical shift tolerance values not enough NOE restraints might be obtained to allow proper folding of the molecule.

In case that a model structure is available the number of possible assignments will be limited depending on the used distance cut-off, enabling the user to use increased chemical shift tolerance values.

In this test of KNOWNOE the aliphatic region of the 2D-NOESY spectrum corresponding to sidechain-

sidechain, e.g., Hβ-Hβ contacts was discarded since in 2D-spectra this region is usually too crowded to be analyzed. To make use of these contacts it is recommended to use ^{13}C edited 3D-NOESY spectra which can be automatically analyzed by the presented algorithm as well. However, aim of this paper was to show that with a minimal set of spectra (a 2D-NOESY) and unlabeled material a decent NMR structure can be obtained.

KNOWNOE was also successfully tested on mixed α/β proteins (data not shown). Pure α-helical proteins were not investigated so far. Since the chemical shift dispersion is usually limited in these types of proteins it may be necessary to use ^{15}N and/or ^{13}C edited NOESY spectra in these cases.

In summary it can be stated that KNOWNOE is a useful tool for the automated assignment of protein NOESY NMR spectra.

Acknowledgements

We like to thank S. Harrieder for their help with the assignment of *Tm*Csp and T. Maurer for helpful discussions. We thank B. Schuler and R. Jaenicke for providing *Tm*Csp. This work was supported by the Deutsche Forschungsgemeinschaft and the Biotechnology Program of the European Union.

References

Antz, C., Neidig, K.-P. and Kalbitzer, H.R. (1995) *J. Biomol. NMR*, **5**, 287–296.

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) *J. Comp. Chem.*, **18**, 139–149.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) *Acta Cryst.*, **D54**, 905–924.

Clore, G.M. and Gronenborn, A.M. (1991) *Science*, **252**, 1390–1399.

Cornfield, J. (1967) *Rev. Int. Statist. Inst.*, **35**, 34–49.

Cornfield, J. (1969) *Biometrics*, **25**, 617–642.

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

Ferrige, A.G. and Lindon, J.C. (1978) *J. Magn. Reson.*, **31**, 337–340.

Fesik, S.W. and Zuiderweg, E.R.P. (1988) *J. Magn. Reson.*, **78**, 588–593.

Geyer, M., Herrmann, C., Wohlgemuth, S., Wittinghofer, A. and Kalbitzer, H.R. (1997) *Nat. Struct. Biol.*, **4**, 694–699.

Geyer, M., Neidig, K.-P. and Kalbitzer, H.R. (1995) *J. Magn. Reson. B*, **109**, 31–38.

Glaser, S. and Kalbitzer, H.R. (1986) *J. Magn. Reson.*, **68**, 350–354.

Görler, A. and Kalbitzer, H.R. (1997) *J. Magn. Reson.*, **124**, 177–188.

Görler, A., Gronwald, W., Neidig, K.-P. and Kalbitzer, H.R. (1999) *J. Magn. Reson.*, **137**, 39–45.

Gronwald, W., Kirchhöfer, R., Görler, A., Kremer, W., Ganslmeier, B., Neidig, K.-P. and Kalbitzer, H.R. (2000) *J. Biomol. NMR*, **17**, 137–151.

Harrieder, S. (1998) Diploma Thesis, University of Regensburg, Regensburg.

Huang, L., Weng, X.W., Hofer, F., Martin, G.S. and Kim, S.H. (1997) *Nat. Struct Biol.*, **4**, 609–615.

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

Koradi, R., Billeter, M., Engeli, M., Güntert, P. and Wüthrich, K. (1998) *J. Magn. Reson.*, **135**, 288–297.

Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graphics*, **14**, 51–55.

Kremer, W., Harrieder, S., Geyer, M., Gronwald, W., Welker, C., Schuler, B., Jaenicke, R. and Kalbitzer, H.R. (2001) *Eur. J. Biochem.*, **268**, 2527–2539.

Laskowski, R.A., Rullmann, J.A., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996). *J. Biomol. NMR*, **8**, 477–486.

Marion, D. and Wüthrich, K. (1983) *Biochem. Biophys. Res. Comm.*, **113**, 967–974.

Mitschang, L., Neidig, K.-P. and Kalbitzer, H.R. (1990) *J. Magn. Reson.*, **90**, 359–362.

Moussa, S. (2001) Ph.D. Thesis, University of Regensburg, Regensburg.

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

Nagayama, K. and Wüthrich, K. (1981) *Eur. J. Biochem.*, **114**, 365–374.

Neidig, K.-P., Geyer, M., Görler, A., Antz, C., Saffrich, R., Beneicke, W. and Kalbitzer, H.R. (1995) *J. Biomol. NMR*, **6**, 255–270.

Neidig, K.-P., Saffrich, R., Lorenz, M. and Kalbitzer, H.R. (1990) *J. Magn. Reson.*, **89**, 543–552.

Nilges, M. (1995) *J. Mol. Biol.*, **245**, 645–660.

Nilges, M. and O'Donoghue, S.I. (1998) *Prog. NMR Spectrosc.*, **32**, 107–139.

Saffrich, R., Beneicke, W., Neidig, K.-P. and Kalbitzer, H.R. (1993) *J. Magn. Reson. B*, **101**, 304–308.

Schulte, A.C., Görler, A., Antz, C., Neidig, K.-P. and Kalbitzer, H.R. (1997) *J. Magn. Reson.*, **129**, 165–172.

Subramaniam, S., Teheng, D.K. and Fenton, J.M. (1996) *ISMB*, **96**, 218–229.

Wagner, G. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 347–366.

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

Wüthrich, K., Wider, G., Wagner, G. and Braun, W. (1982) *J. Mol. Biol.*, **155**, 311–319.

Xu, Y., Wu, J., Gorenstein, D. and Braun, W. (1999) *J. Magn. Reson.*, **136**, 76–85.