Journal of Biomolecular NMR, **19:** 49–62, 2001.
KLUWER/ESCOM

Variability in automated assignment of NOESY spectra and three-dimensional structure determination: A test case on three small disulfide-bonded proteins

Philippe Savarin, Sophie Zinn-Justin & Bernard Gilquin*
Département d'Ingénierie et d'Etudes des Protéines (Bât. 152), CEA-Saclay, F-91191 Gif-sur-Yvette Cedex, France

Received 12 July 2000; Accepted 19 October 2000

Key words: protein structure determination, NOESY automated assignment, variability of the procedure

Abstract

Three independent runs of automatic assignment and structure calculations were performed on three small proteins, calcicludine from the venom of the green mamba *Dendroaspis angusticeps*, κ-conotoxin PVIIA from the purple cone *Conus purpurascens* and HsTX1, a short scorpion toxin from the venom of *Heterometrus spinnifer*. At the end of all the runs, the number of cross peaks which remained unassigned (0.6%, 1.4% and 2% for calcicludine, κ-conotoxin and HsTX1, respectively), as well as the number of constraints which were rejected as producing systematic violations (2.7%, 1.0%, and 1.4% for calcicludine, κ-conotoxin and HsTX1, respectively) were low. The conformation of the initial model used in the procedure (linear model or constructed by homology) has no influence on the final structures. Mainly two parameters control the procedure: the chemical shift tolerance and the cut-off distance. Independent runs of structure calculations, using the same parameters, yield structures for which the rmsd between averaged structures and the rmsd around each averaged structure were of the same order of magnitude. A different cut-off distance and a different chemical shift tolerance yield rmsd values on final average structures which did not differ more than 0.5 Å compared to the rmsd obtained around the averaged structure for each calculation. These results show that the procedure is robust when applied to such a small disulfide-bonded protein.

Abbreviations: NOE, nuclear Overhauser effect; TOCSY, total scalar coupling correlated spectroscopy; NOESY, nuclear Overhauser effect correlated spectroscopy; COSY, scalar coupling correlated spectroscopy; rmsd, root mean square deviation.

Introduction

In order to carry out more efficient structure calculations from NMR spectroscopy data, recent efforts have been made to automate several steps of NMR data analysis, e.g. peak-picking (Koradi et al., 1998), resonance assignments (Bartels et al., 1997; Buchler et al., 1997), NOE assignments and structure determination (for review, see Moseley et al., 1999). One of the most time-consuming steps in the determination of the three-dimensional structure is the assignment of NOESY spectra. Manual assignment of NOESY data is tedious, time-consuming, and necessitates choices in the interpretation of the NMR data. These choices may have a strong influence on the resulting protein structure, either on its global fold or on local conformations. An automatic assignment procedure has the advantage of quickly producing unbiased structures, which by definition is the image of the precision of experimental data. As numerous factors contribute to the NOESY intensity, e.g. spin diffusion, internal dynamics (Bonvin et al., 1994; Brüschweiler et al., 1994; James, 1991), the distances derived from the

*To whom correspondence should be addressed. E-mail: bgilquin@cea.fr

NMR data are rather imprecise. This imprecision of the NMR data is one of the reasons why ensembles of structures are generated. Several authors, comparing the NMR ensemble with ensembles of structures issued from molecular dynamics simulations, found that both ensembles are correlated (Clore et al., 1993; Zhao et al., 1994). Automatic assignment procedures open a way to obtain such reliable ensembles of NMR structures without any bias. In summary, an automated assignment procedure will accelerate the structure calculations and has the advantage of replacing partially arbitrary choices made manually by experts by rational and unbiased processing of NMR data.

Many efforts have been made to automatically assign NOESY spectra from the assignment of the chemical shifts, e.g. NOAH (Mumenthaler et al., 1995, 1997), ARIA (Nilges, 1995; Nilges et al., 1997). These methods generally perform simultaneously the NOESY assignment and the structure calculation by applying an iterative approach during which the number of correctly assigned cross peaks increases while the structure calculation converges. The assignments are deduced from the structure calculation and vice versa.

Until recently, an NMR ensemble of structures used to be obtained from a unique assignment of NOESY data. In fact, many cross peaks may give ambiguous assignments (Nilges, 1995) and could correspond to multiple assignments. As in an automated procedure the assignment is based on the structures, small differences in the ensemble of structures produce differences in the assignments and vice versa. The variability of automatic NOESY assignment routines can be estimated by performing independent runs of assignment and structure calculation and by comparing their results.

In the present paper we performed three independent runs of automatic assignment and structure calculations on three small proteins. In the first part, we briefly present the method which has been developed. Next, the results are presented in order to answer the following questions: What are the differences between the NOESY cross-peak assignments generated by several independent automatic assignments? What are the differences between the resulting structures? Lastly we discuss the correlation existing between assignment variability and the differences between the averaged final structures.

Methods

NMR spectroscopy and resonance assignment

Proton 2D COSY (Aue et al., 1976), DQF-COSY (Rance et al., 1983), TOCSY (Braunschweiler et al., 1983), and NOESY (Kumar et al., 1980) were recorded at 500 MHz in H_2O or pure D_2O on a Bruker (Rheinstetten, Germany) DRX500 spectrometer. The experiments were made at 303 K, 288 K and 308 K respectively and at pH 4.5, 5.3 and 4.0 respectively for calcicludine, κ-conotoxin PVIIA and HsTX1. The concentrations were 6.5, 3.4 and 3.0 mM respectively. The spectra were recorded with 512 t_1 × 1024 t_2 points (1024 t_1 × 4096 t_2 for the DQF-COSY). Mixing times of 50, 75, 100 and 150 ms were used during the NOESY experiments in H_2O or D_2O. Chemical shifts were measured referenced to internal 3-(trimethylsilyl)[2,2,3,3-$^2H_4$] propionate (see Savarin et al., 1998, 1999; Gilquin et al., 1999, for details).

Overview of the automated assignment procedure

Based on the suggestions of Nilges et al. (1997), a semi-automated iterative assignment procedure (Gilquin et al., 1999) has been developed. This procedure, written in C-shell, is interfaced to X-PLOR3.1 (Brünger, 1992).

Data

The input data are composed of proton chemical shifts, build-up rates of NOE cross peaks, angular restraints, and distance constraints.

- Chemical shift assignment was achieved by analysing the COSY, NOESY and TOCSY spectra (see Methods in Savarin et al., 1998, 1999; Gilquin et al., 1999).
- The list of NOE build-up rates contains the peak numbers, the two chemical shift positions and the rates of the corresponding build-up curves. For each peak, the volume of the cross peaks at different mixing times was integrated from NOESY spectra, and a build-up curve was constructed by fitting the experimental volumes to the following function of the mixing time: $f(\tau_m) = a \times \tau_m + b \times \tau_m^2$. In order to eliminate artefact cross peaks from the assignment procedure, build-up rates with a good fit were selected ($\chi^2 < 1$ and $a > K/5.3^6$ where K is a calibration constant). To calibrate the distance, different references were used depending on the structure. For the calcicludine and HsTX1, a proton pair of one tyrosine residue (calcicludine: Tyr6; HsTX1: Tyr21), which corresponds to

a distance of 2.48 Å, was used. For the κ-conotoxin, as there is no Tyr residue in the sequence, 10 geminal Hβ build-up rates (corresponding to a distance of 1.8 Å) were used. These calibrations were checked versus the known dαN distances in a regular secondary structure (2.2 Å and 3.4 Å in β-sheets and α-helices, respectively; Wüthrich, 1986). The errors made on the distance were evaluated as follows: for each proton, the root mean square deviation (rmsd) between the one, two, three or four corresponding distances measured on both sides of the diagonal and in both solvents was calculated. An error of +/− 25% on the distance values for all proton pairs was applied, except for those for which twice the rmsd was found to be higher than 25% of the distance. For these proton pairs, an error equal to twice their rmsd was used.

• The angular restraint file contains the ϕ and χ_1 dihedral angle restraints. The ϕ angle restraints were determined from J coupling on the basis of the Karplus relation (Pardi et al., 1984); χ_1 dihedral constraints were determined using the method of Hyberts et al., 1987. The boundaries of intervals were set to $\pm$ 25°, $\pm$ 35°, $\pm$ 50° for ϕ, depending on its value, and were fixed to $\pm$ 45° for χ_1.

• The distance constraint file contains the distance constraints that did not correspond to NOE contacts, like disulfide bridges or hydrogen bond restraints. All the structures considered in the present paper have disulfide bridges (calcicludine and κ-conotoxin have three disulfide bridges and HsTX1 has four), which yield one distance constraint per disulfide bridge. Only for HsTX1, hydrogen bond distance constraints in the β-sheet were used (Savarin et al., 1999).

• To start the cross peak assignment from the beginning, the procedure needs an initial structure. This initial structure is either a model constructed by homology modelling or a linear structure.

Parameters

Essentially four parameters control the assignment procedure:

• A chemical shift tolerance used to compare the values of the chemical shift table with the chemical shift positions of the NOESY cross peaks.

• A cut-off distance that is the maximal distance for selecting proton–proton proximity in the structures.

• A relative peak intensity threshold p_{th}. For each peak the possible assignments were sorted, in decreasing order, by the magnitude of their contributing intensity. The contributing intensities Ii were calculated as the inverse sixth power of individual proton–proton distances d_i. Each intensity value was compared to the previous one. All the possible assignments were considered as long as the contributing intensity was more than p times the previous one:

distance	$d_1\ d_2\ d_3$ .....d_i ...
intensity	$I_1\ I_2\ I_3$ .....Ii ... $I_i = 1/d_i^6$
relative peak intensity	$p_2\ p_3$ .....p_i ... $p_{i+1} = I_{i+1}/I_i$ ($i \in \mathbf{N}^*$)

While $p_i > p_{th}$, the assignments were considered. The following ones were not considered (even though p_j values are less than p_{th}, $j > i$).

• The maximum number of ambiguous assignments contributing to a cross peak. If this number is higher than the maximum allowed, the distance constraint associated with the cross peak is not generated.

General principle

At each iteration of the procedure, NOE cross-peak assignments were calculated based on the best energy structures obtained at the previous iteration and then new structures were calculated with X-PLOR. The assignments were made on the basis of the chemical shift list and the averaged distance between protons in the eight best structures (i.e., lowest energy) using the distance cut-off and the chemical shift tolerance parameters. The NOE cross peaks which could not be assigned unambiguously were converted to ambiguous distance constraints (Nilges, 1995). The ambiguities were selected depending on the cut-off distance, the relative peak intensity threshold p_{th} and the maximum number of assignments. At each iteration, the distance constraints violated by more than 0.5 Å in more than three (or four) of the 10 best structures were added to a file which contains the cross-peak number and its assignment. At the end of the assignment procedure, peaks which appeared with the same assignment in this file and in the assigned list were suppressed from the constraints file.

After proceeding to the assignment of the cross peaks, a set of structures (50 for the first three or four iterations and 100 for the next iterations) was calculated with X-PLOR 3.1. The 10 best structures (i.e. lowest energy) of the previous iteration were used as starting structures in a simulated annealing protocol. The simulated annealing protocol (Nilges, web personal communication) had three steps which mainly consisted of randomisation of the ϕ, ψ of the starting structure, a high dynamics simulation (T = 2000 K for 6500 steps), and cooling (3500 steps). Floating assignments for prochiral groups and swapping

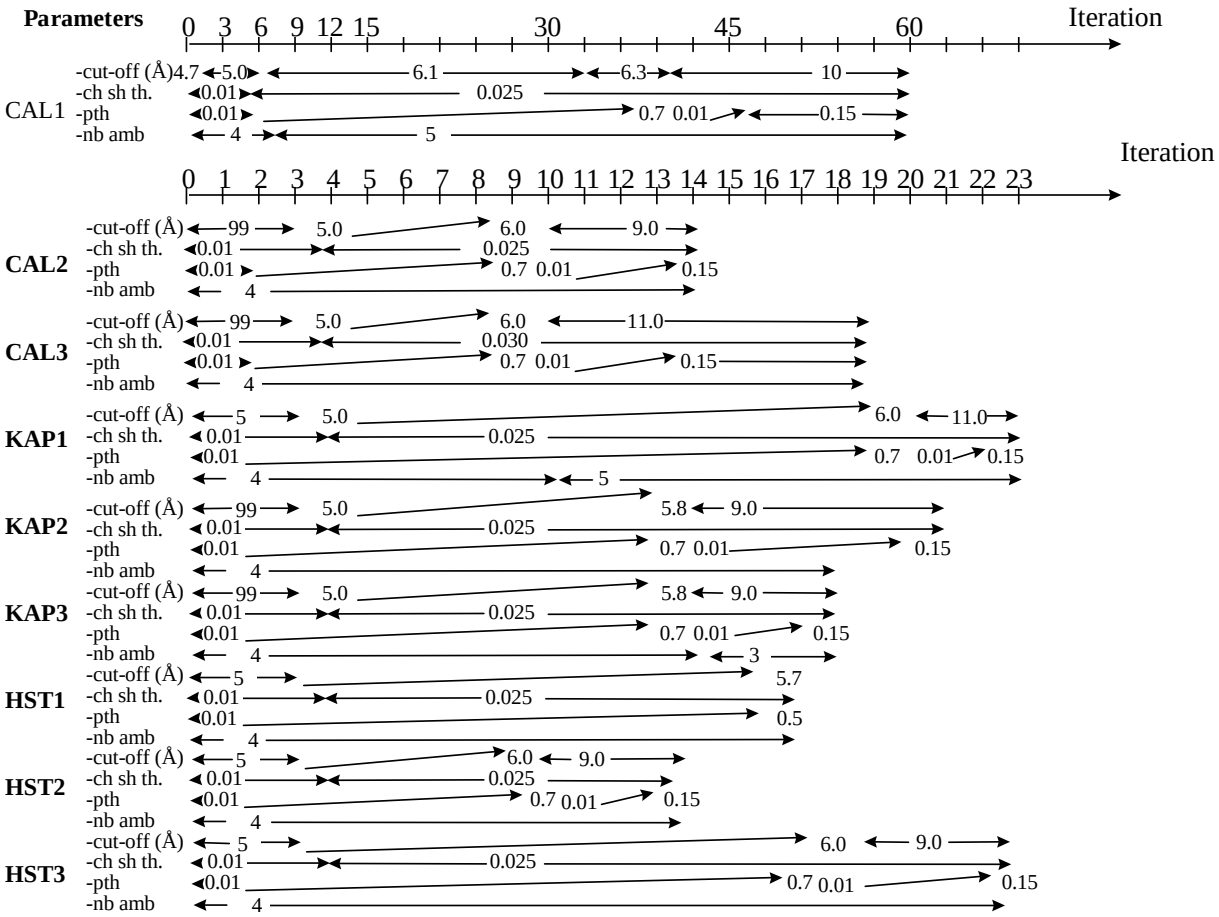


Figure 1. Variations of the parameters during the runs: cut-off, chemical shift threshold (ch sh th.), relative peak intensity threshold (p_{th}), and maximum number of ambiguous assignments contributing to one cross peak (nb amb).

atoms following an evaluation of the NOE term were used (Folmer et al., 1997). A force field adapted to NMR structure calculation (files toppalhdg.pro and parallhdg.pro) was used.

The first round of assignments was made with a long cut-off distance (99 Å), a small chemical shift tolerance (0.01 ppm), a maximum number of ambiguities equal to 10, and a p_{th} value of zero. In order to avoid assignment errors, the unambiguous long-range NOEs (at least i,i+3) were reassigned with a larger chemical shift tolerance (0.025 or 0.03 ppm). This step assures the assignment of the unambiguous long-range NOEs, which are crucial for the folding of the protein. Then 50 structures were calculated using a simulated annealing protocol. The 10 best energy structures were kept and the constraints violated by more than 0.5 Å in more than three structures were excluded. Fifty or 100 structures were recalculated, the 10 best structures were kept and the constraints violated by more than 0.5 Å in more than three structures were excluded.

Then, based on the eight best (i.e., lowest energy) structures, the NOE cross-peak list was completely reassigned using a short cut-off distance (5 Å), a small chemical shift tolerance (0.012 ppm), a maximum number of ambiguities equal to 6, and a p_{th} of 0.01. For the following iterations, the parameter p_{th} was progressively increased from 0.01 to 0.7 to decrease the number of ambiguous cross peaks. In order to increase the number of assigned NOE rates, the threshold for the chemical shift tolerance and the cut-off distance were progressively increased from 0.012 to 0.025 (or 0.03) ppm and from 5 to 6 Å, respectively. After these iterations, several cross peaks were still not assigned. In order to take into account the pairs of protons that were distant by more than 6 Å, the distance threshold was increased up to 9 or 11 Å. The number of

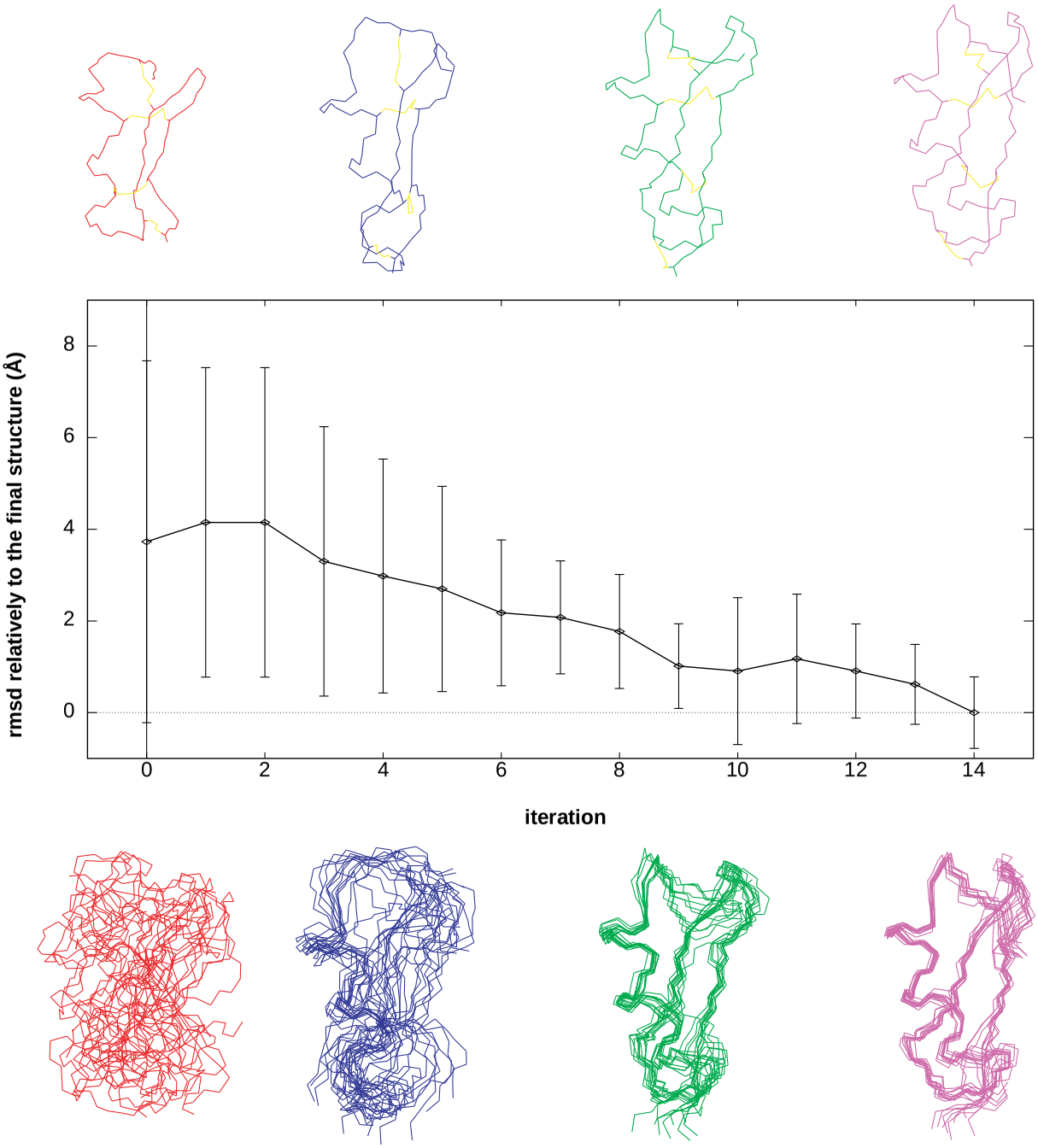


Figure 2. Evolution of the HstX1 structure during the automatic assignment and structure calculation (run HST2).

assigned NOEs increased with this step. At the end of the procedure, several iterations with the same parameters were performed. All these iterations provide a set of structures with low energy, few violations, good Ramachandran plot and low rms differences with the average structure, showing that convergence of the procedure is reached. The 10 structures with the lowest number of violations were kept as the final structures.

These 10 final structures were refined. A restrained molecular dynamics run at 600 K and a slow cooling were carried out with a standard energy function (files topallh22x.pro and parallh22x.pro in X-PLOR3.1) comprising an electrostatic term. The electrostatic term was calculated with no net charge on the side chain atoms and with a distance-dependent dielectric constant.

Description of the protocols and the data used for the three proteins

Three proteins were studied: calcicludine (Schweitz et al., 1994) from the venom of the green mamba *Dendroaspis angusticeps*, κ-conotoxin PVIIA (Terlau et al., 1996) from the purple cone *Conus purpurascens* and HsTX1 (Lebrun et al., 1997), a short scorpion toxin from the venom of *Heterometrus spinnifer*. For each protein, three independent runs of automatic assignment were performed, named CAL1 to CAL3, HST1 to HST3 and KAP1 to KAP3 for calcicludine, HsTX1 and κ-conotoxin, respectively. For each protein, the same chemical shift assignment and list of NOESY build-up rates (Savarin et al., 1998, 1999; Gilquin et al., 1999) were used for the three runs. The same angular restraints file and distance constraints file (constraints corresponding to the disulfide and hydrogen bonds) were also used for the three runs.

The variations of the values of the parameters during the run and their value in the final iteration are indicated respectively in Figure 1 and in Table 1. For κ-conotoxin three runs with the same parameters were realised; these parameters were also used for HST2, HST3 and CAL2. For calcicludine each run was done with different parameters. For CAL1, the distance cut-off for the last iteration was 10 Å. For CAL3, the distance cut-off and the chemical shift tolerance were larger (11 Å and 0.03 ppm, respectively). The runs and the structures obtained at the end of CAL1 and KAP1 correspond to those published previously (Savarin et al., 1998; Gilquin et al., 1999). To calculate the structure of HsTX1, four constraints were added (Lys23.HN-Lys30.O, Lys30.HN-Lys23.O, Lys23.N-Lys30.O, Lys30.N-Lys23.O). These constraints bring together the two amides of the β-sheet structure still observable after 5 h in D_2O and their corresponding oxygen atoms.

At the end of the assignment procedures of the previously published structures (runs CAL1 and KAP1), a careful examination of the constraints and structures was done. For CAL1, one unambiguous assignment Gly14.HA-Gly41.HA was resolved manually. This assignment was essential for the loop conformations. The same manual assignment was done at the beginning of the procedure for CAL2 and CAL3. For the structure calculations HST1, HST2, HST3, KAP2, and KAP3, no manual intervention was done.

Results

We present the results of three independent automatic assignment processes on three proteins: calcicludine, κ-conotoxin and HsTX1. Calcicludine, κ-conotoxin and HsTX1 are respectively composed of 60, 27 and 34 residues. The experimental NOE data (chemical shifts, rate of the NOESY build-up curves and J-coupling constants) were those obtained previously for calcicludine (Gilquin et al., 1999), κ-conotoxin (Savarin et al., 1998) and HsTX1 (Savarin et al., 1999). For these three proteins, the number of build-up rates is 1881, 873 and 658, respectively. The solution structures of the three proteins were obtained by the authors in applying the present procedure. For calcicludine and κ-conotoxin, the authors have used an initial model constructed by homology modelling and for HsTX1 a linear one. In order to test the influence of the initial model on the assignment procedure and to compare the results of independent assignments on the same protein, new runs of structure calculations were performed. Three runs were performed for each protein: one starting from a model constructed by homology and two from a linear structure (see Methods).

Figure 2 shows the evolution of the structure of calcicludine during one run of structure calculation starting from a linear structure. In the upper part, the averaged structure is represented and in the lower part the 10 best energy structures. From the first iterations, the right topology for the fold is obtained and is preserved during the following iterations. As the iteration number increases, the obtained structures are better defined.

For the three proteins, all the structure calculation runs give well-defined structures (Figure 3, Table 2).

Table 1. Parameters used for the automatic assignment procedure

Protein	Run	No. of iterations	Chemical shift threshold (ppm)	Cut-off (Å)	Relative peak intensity p_{th}	Max no. of ambiguous assignments for one cross peak
Calcicludine	CAL1	60	0.025	10	0.15	5
	CAL2	14	0.025	9	0.15	4
	CAL3	19	0.030	11	0.15	4
κ-conotoxin PVIIA	KAP1	25	0.025	9	0.15	5
	KAP2	21	0.025	9	0.15	4
	KAP3	18	0.025	9	0.15	3
HsTX1	HST1	17	0.025	5.7	0.15	4
	HST2	14	0.025	9	0.15	4
	HST3	23	0.025	9	0.15	4

Table 2. Rmsd (Å) around the averaged structure for backbone atoms and heavy atoms

Protein	Run	Backbone	Heavy atom	No. cons/res
CAL	CAL1	0.61 ± 0.15	1.10 ± 0.20	16.5
	CAL2	0.68 ± 0.08	1.11 ± 0.09	16.2
	CAL3	0.73 ± 0.16	1.20 ± 0.18	16.2
KAP	KAP1	0.59 ± 0.14	1.36 ± 0.21	16.1
	KAP2	0.61 ± 0.16	1.35 ± 0.22	16.0
	KAP3	0.56 ± 0.18	1.32 ± 0.15	15.8
HST	HST1	0.80 ± 0.10	1.50 ± 0.16	10.6
	HST2	0.84 ± 0.18	1.57 ± 0.21	10.6
	HST3	1.08 ± 0.35	1.77 ± 0.44	10.5

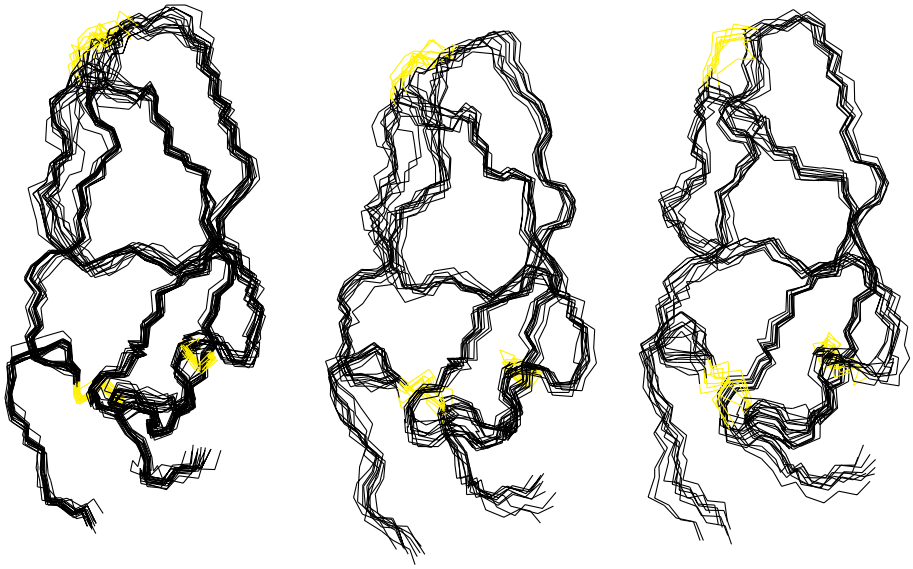
All the rmsd values on the refined structures are below 1.1 Å for the backbone and 1.8 Å for the heavy atoms. The structures have an acceptable covalent geometry, as evidenced by the low rmsd for bond lengths, valence angles and improper dihedral angles. The van der Waals values are small, ruling out unfavourable nonbonded contacts (Table 3). The Ramachandran plot confirms the good quality of the structures (Table 3). A few violations (distance violation larger than 0.50 Å and dihedral violation larger than 10°) are still present at the end of the procedure (Table 3). Except for HsTX1, the violations are different between two runs.

Table 2 shows that for HsTX1, the apparent precision (rmsd around the averaged structures) is not as good as for calcicludine or κ-conotoxine. It is correlated to the number of constraints per residue (Table 2), which is significantly lower for HsTX1 than for the other two proteins. For this protein, the low number of constraints per residue is probably due to the higher temperature of the experiments (see Methods). It is well known that a low number of constraints per residue yields less precise structures. For calcicludine and κ-conotoxin, the number of constraints by residue is roughly the same, but the precision of the structure is lower for calcicludine. These two proteins differ by their respective size. As already noted for a comparable number of constraints by residue, the precision of the structure is lower for larger proteins (Liu et al., 1992).

Assignment

At the end of the procedure, the number of peaks which remains without any assignment is low (on average 0.6%, 1.4% and 2% for calcicludine, κ-conotoxin and HsTX1, respectively; Table 4). The unassigned cross peaks could be explained by the unassigned proton resonances, multiple minor conformations of the protein or noise peaks. They represent less than 2% of the number of peaks for the three proteins. Respectively, 97%, 92% and 57% of these unassigned cross peaks were identical in the three runs for each protein.

The number of constraints which are rejected as producing systematic violations is low (an average of 2.7%, 1.0%, and 1.4% of the number of constraints for calcicludine, κ-conotoxin and HsTX1, respectively (Table 4)). The rejected constraints are between 17% and 45% identical in the three runs (two runs for HST, because HST1 has no rejected constraint). It should be noted that in all cases, these rejected constraints correspond to constraints violated (between 0.5 and 5 Å) in the final structure. The error should not be



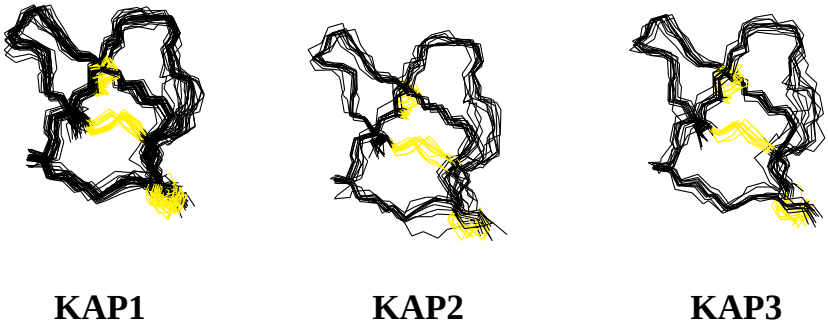


Figure 3. Superimposition of the best 10 energy structures of calcicludine, HsTX1 and κ-conotoxin. For each run, the structures are superimposed independently.

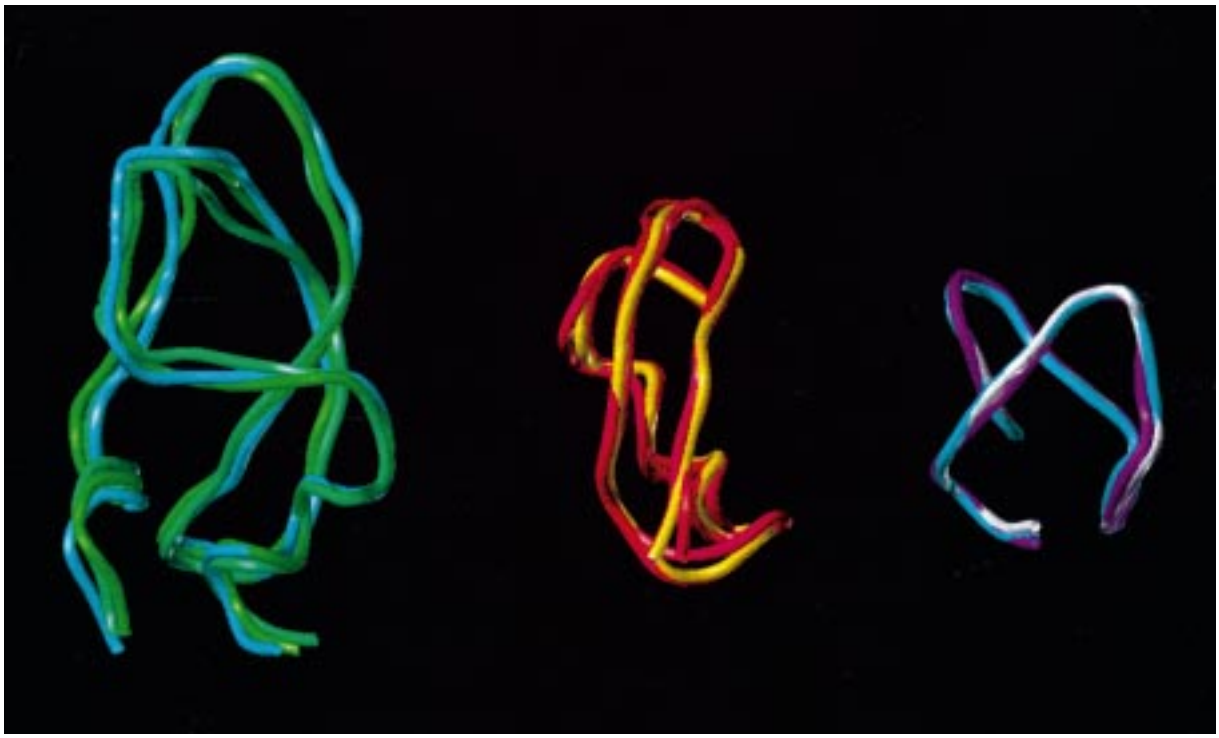


Figure 4. Superimposition of the three averaged structures of calcicludine, HsTX1 and κ-conotoxin (from left to right) obtained by independent runs of the automatic assignment and structure determination procedure.

in the assignment but could be due to the measured cross-peak volume.

This low value of non assigned cross peaks and incompatible constraints is due to the use of manual peak picking, in which the peaks were carefully examined and the quality of the fitting was analysed. This procedure in selecting the good fitting build-up curves eliminates the majority of the cross-peak artefacts: for calcicludine, κ-conotoxin and HsTX1, respectively 20%, 29% and 29% of the peaks were eliminated. Furthermore, build-up curves provide more reliable distances as spin diffusion is taken into account.

Comparison of the cross-peak assignments and constraint files

Table 5 gives the results of the comparison of the three independent runs performed on the three proteins. For the majority of cross peaks the assignments are identical. At the end of the assignment process, between 5 and 30% of the cross peaks remain ambiguously assigned (Table 4). For each cross peak, the assignment has two parts. Each part corresponds to one dimension of the spectra. Each part of the assignment is composed of a set of atoms. To compare the assignments of a cross peak in two different runs, each part of both assignments has to be analysed. Each part could have all, several or no atoms in common with the corresponding part of the other assignment. If both parts of the two assignments have atoms in common, we consider the two assignments to be closely related. The numbers of identical or related assignments represent more than 95%, 99% and 98% of all peaks for calcicludine, κ-conotoxin and HsTX1, respectively (Table 5). Only very few cross peaks were assigned in only one run.

In fact, the data used in X-PLOR3.1 for the structure calculation are the constraints generated from the assigned peaks. One constraint could correspond to more than one peak if peaks are found on both sides of the diagonal of the NOESY in D_2O and H_2O with the same assignment. Differences in the constraints files concern cross peaks ambiguously assigned which are not found with the same assignment on both sides of the diagonal of the NOESY map. This is due to the difference in the spectral resolution in the two dimensions. On the contrary, the unambiguously assigned peaks are generally found with the same assignment on both sides of the diagonal and in the two experiments. More than 88%, 98% and 89% of the con-

Table 3. Structural statistics for the three independent runs of automatic assignment and structure calculation of κ conotoxin, HsTX1 and calcicludine

		Energy (kcal/mol)	Electrostatic[a] (kcal/mol)	Quality index[b] (%)	rmsd from restraints[c]		rmsd from ideal values			Number of violations >0.5 Å
					Distance (Å)	Dihedral (°)	Bond (Å)	Angle (°)	Improper (°)	
KAP	1	−223 (±19)	−572 (±15)	92.8	0.052 (±0.006)	0.3 (±0.43)	0.02 (±4e–4)	3.4 (±0.1)	2.7 (±0.5)	0.2 (±0.4)
	2	−217 (±18)	−566 (±10)	94.6	0.055 (±0.006)	0.70 (±0.51)	0.02 (±4e–4)	3.4 (±0.1)	2.7 (±0.4)	0.1 (±0.3)
	3	−214 (±18)	−566 (±9)	94.1	0.057 (±0.008)	0.5 (±0.54)	0.02 (±4e–4)	3.4 (±0.1)	2.5 (±0.6)	0.4 (±0.5)
HST	1	−188 (±15)	−586 (±14)	97.4	0.049 (±0.005)	1.2 (±0.38)	0.02 (±2e–4)	3.3 (±0.1)	2.0 (±0.3)	0.6 (±0.8)
	2	−96 (±32)	−588 (±16)	93.1	0.058 (±0.006)	5.3 (±0.5)	0.02 (±4e–4)	3.6 (±0.1)	2.5 (±0.4)	4.1 (±1.4)
	3	−80 (±42)	−589 (±15)	88.6	0.06 (±0.007)	5.0 (±0.5)	0.02 (±4e–4)	3.7 (±0.1)	2.5 (±0.4)	4.1 (±1.1)
CAL	1	335 (±14)	−575 (±11)	97.4	0.07 (±0.004)	1.3 (±0.2)	0.02 (±1e–4)	3.5 (±0.1)	3.1 (±0.5)	0.6 (±0.5)
	2	294 (±34)	−560 (±21)	97.5	0.07 (±0.004)	0.9 (±0.4)	0.02 (±2e–4)	3.4 (±0.1)	2.3 (±0.2)	2 (±0.5)
	3	404 (±14)	−566 (±21)	96.5	0.08 (±0.005)	1.6 (±0.7)	0.02 (±2e–4)	3.6 (±0.1)	2.4 (±0.3)	2 (±1.2)

[a]The electrostatic energy is calculated with a switch function, CHARMM22 parameters, no net charge on side-chain atoms, and a distance-gated dielectric constant.
[b]Percentage of residues found in most favored and additional allowed regions of the Ramachandran plot, as calculated by Procheck-NMR (Laskowski et al., 1996).
[c]The values of the square-well NOE and dihedral angle potentials are calculated with a constant force of 20 kcal/mol Å^2 and 50 kcal/mol rad^2.

Table 4. Result of the assignment of the NOE cross peaks of the three small proteins

Protein	Run	Cross peaks				Constraints		
		Total	Assigned	Unassigned	Rejected	Total	Unambiguous	Rejected
CAL	CAL1	1881	1871	10	18	987	715	17
	CAL2	1881	1870	11	47	972	754	27
	CAL3	1881	1871	10	62	974	680	33
KAP	KAP1	873	860	13	2	434	368	1
	KAP2	873	861	12	9	432	364	4
	KAP3	873	861	12	11	427	355	6
HST	HST1	658	645	14	0	360	343	0
	HST2	658	644	14	7	363	291	5
	HST3	658	646	12	19	359	290	10

straints files for calcicludine, κ-conotoxin and HsTX1, respectively (Table 5) are identical or related.

Differences between the averaged structures

The last two columns of Table 5 present the values of the rms deviation between the three averaged structures of the three proteins. Figure 4 shows the superimposition of the backbones of the three averaged structures for each of the three proteins. For the three proteins, the three runs of assignment and structure calculation give essentially the same result. For calcicludine, the deviations on the backbone atoms between the averaged structures are between 1 and 1.3 Å rmsd. The averaged distances between Cα atoms of CAL1/2, CAL1/3 and CAL2/3 have been considered. The distances higher than 1.5 Å are found for residues of the N terminus, for residues 14 and 19–21 (loop 1), for residue 28 (turn), residue 38 (loop 2) and residues of the C terminus. These residues are responsible for about 50% of the differences observed in the constraints files and are located in the less well defined part of the structure. For κ-conotoxin, the three values of backbone rmsd between the averaged structures are less than 0.5 Å and the differences in the constraints file are low (<2%). For HsTX1, HST1 is very close to the published NMR structure. The rmsd on backbone atoms between the two averaged structures is 0.69 Å and less than 4% of the constraints are different between the two runs. The rmsd values on backbone atoms between the pair of structures

Table 5. Comparison of the cross-peak assignments, constraints files and averaged structures

Protein	Run/run	Comparison of the cross-peak assignments					Comparison of the constraints files						Structure rmsd (Å)	
		Assigned	Iden.	Com	Diff	Diff (%)	Constraints	Iden.	Com	Diff.	Diff (%)	Diff + com(%)	Back	Heavy
CAL	CAL1/2	1815/1806	1499	265	46	2.6	987/972	794	113	72	7	18	1.07	1.66
	CAL1/3	1815/1806	1313	377	120	6.7	987/964	682	158	135	14	30	1.29	1.88
	CAL2/3	1806/1806	1326	378	102	5.7	972/964	695	151	122	13	30	1.21	1.82
KAP	KAP1/2	861/860	811	48	1	0.3	434/432	403	23	7	2	7	0.31	0.57
	KAP1/3	861/861	801	58	2	0.3	434/427	396	27	7	2	8	0.31	0.67
	KAP2/3	861/860	789	69	3	0.4	432/427	388	33	8	2	9	0.21	0.73
HST	HST1/2	644/645	550	87	8	1.2	360/363	294	54	13	4	18	1.25	1.95
	HST1/3	644/646	553	86	5	1.0	360/359	291	52	16	5	19	1.23	2.05
	HST2/3	645/646	573	65	7	1.2	363/359	319	35	7	2	12	0.37	1.08

Iden.: number of assignments, or constraints, that are identical; Com: number of assignments, or constraints, for which in both parts of the assignment common atoms are found; Diff: number of assignments, or constraints, in which at least one part of the assignment in one run has no common atom in the other run; Diff(%) indicates the averaged percentage in which at least one part of the assignment in one run has no common atom in the other run; Diff +com indicates the averaged percentage of constraints which is assigned differently in both runs. For structure comparison, the rmsd was calculated by superimposing backbone (col name 'back') or heavy atoms (col. name 'heavy').

HST1/2, HST1/3 and HST2/3 lie between 0.4 and 2 Å. Averaged distances between Cα atoms of averaged structures of HST1, HST2 and HST3 higher than 1.5 Å are found for residue 10 (first residue of the helix), residue 20 (turn), residue 26 (turn) and the C terminus (residue 34). These residues make up about 20% of the differences observed in the constraints files.

These rmsd values between the averaged structures have to be compared to the rmsd around each averaged structure. For calcicludine and HsTX1, the rmsd around the averaged structures (on average 0.7 Å and 0.9 Å) is lower than the rmsd between the three averaged structures (on average 1.2 Å for KAP and 1.2 Å between HST1 and HST2 or HST3). For κ-conotoxin, the order of magnitude is the same between the rmsd around the structure (0.5 Å) and the rmsd between the averaged structure (0.3 Å). For this protein, differences between the constraints remain lower than 10% while for CAL and HST1 with HST2/3, they are higher than 18%. Therefore, differences in the constraints files give differences between the final averaged structures.

Discussion

The use of an automated procedure has the advantage of speeding up the structure determination, avoiding manual errors in data interpretation. Three independent runs of automatic assignment and structure calculation were performed for three small proteins: calcicludine, κ-conotoxin and HsTX1. The differences in the resulting NOE assignment and structure were compared for the three proteins.

Assignment procedure

The present semi-automated iterative assignment procedure of NOE cross peaks is based on the use of ambiguous distance restraints, as in the ARIA (Nilges et al., 1997) procedure. In NOAH (Mumenthaler et al., 1995), the cross peaks for which the assignment is not unique are not treated by ambiguous distance restraints. Instead, NOAH uses the principle of 'self correcting distance geometry'. The use of ambiguous distance restraints for cross peaks which correspond to several proton chemical shifts seems efficient and reliable for an automated assignment procedure. In the present study, no manual assignment is used and the first iterations are especially designed to obtain the right fold. In the first step, the cross peaks unambiguously assigned to long range are reassigned with a larger tolerance. The few remaining unambiguous constraints are sufficient to get the right fold for the proteins presently studied. At the opposite, the use of a large tolerance for all cross peaks did not allow convergence. Generally, authors using ARIA start from numerous manual assignments. After obtaining the fold, the cross peaks are slowly progressively assigned in order to avoid misassignments. In our procedure, we attempted to assign cross peaks with a low chemical shift tolerance and a short cut-off. As the iterative procedure progressed, the chemical shift tolerance and the cut-off were increased in order to assign all the

cross peaks. In this way, during the procedure, all the intermediate structures presented few violations.

Rejected and unassigned cross peaks

In all the runs, the total number of unassigned and rejected cross peaks represents less than 5% of the total number of cross peaks (Table 4). Therefore between 96 and 98% of the cross peaks were assigned. This present result was obtained from a hand-picked NOE list in which only the good fitting NOESY build-up curves were selected. The importance of build-up curves will be discussed elsewhere (Gilquin, personal communication). This peak selection seems to efficiently exclude a large majority of the artefact cross peaks and therefore reduce the number of unassigned or rejected cross peaks. Manual peak-picking may also limit the number of excluded constraints. From a hand-picked NOESY list, ARIA seems to exclude less than 5% of the peaks (Nilges et al., 1998). NOAH, in a model calculation in which a small chemical shift threshold was used, assigned 70–90% of all the cross peaks (Mumenthaler et al., 1995, 1997). Our present results are in agreement with these studies.

The peaks incompatible with a 3D structure can simply be a consequence of erroneous bonds set too narrowly (in particular resulting from underestimated internal dynamics; Schneider et al., 1999). It has been shown by Chalaoux et al. (1999) that for more than 12% of the restraints, the distance in a reference structure lay outside $\pm$25% of the distances determined from spectral densities from molecular dynamics.

Differences between the final structures

The differences between averaged final structures and differences between the assignments appear correlated (Table 5). The larger the differences of constraints employed are, the more important are the differences between the final structures. When 30% of the constraints employed are different (CAL1-2/3), the rmsd between the averaged structures is 0.6 Å higher than the precision of the structures. When 20% are different (HST1/3, HST2/3, CAL1/2), the rmsd values between the averaged final structures are 1.25 Å, 1.23 Å and 1.07 Å, respectively, i.e. about 0.4 Å more than the precision of the structures. With less than 10% of differences in the constraints, the rmsd obtained between the averaged structures is of the same order as the precision of the structures (KAP1/2).

(a) Role of the parameters: Same parameters

Mainly two parameters influence the variations in the assignments: the chemical shift tolerance and the distance cut-off. When the same parameters are used (HST2/3 and the three κ-conotoxin runs), less than 10% of the constraints are different. Between the runs HST2 and HST3, neither the cut-off nor the chemical shift tolerance is changed and the rmsd obtained between the averaged structures (0.34 Å) is inferior to the rmsd obtained in a single calculation (about 0.9 Å). For the three KAP runs, the number of ambiguous constraints is low (less than 17%). In the comparison between HST2 and HST3, the number of ambiguous constraints is higher (more than 18%). Nevertheless, differences in the constraints files were limited (11%) and did not induce differences between the averaged structures larger than the rmsd around the structures.

It should be noted that the only difference between KAP1 and KAP2 is the initial structure. Therefore the initial model (linear or constructed by homology) has no influence on the final structures. This conclusion was pointed out in other studies (Fraternali et al., 1999). Structures obtained with more or less manual assignments in using the same parameters were already compared in the ARIA or NOAH procedures (Liu et al., 1999; Xu et al., 1999): changes in the starting conditions cause perturbations of 1 Å or less to the backbone of the protein.

The results obtained by comparing two following iterations at the end of any of the runs are analogous: the differences between the averaged structure (less than 0.5 Å) and the constraint file (less than 11% difference) are small. This is independent of the number of ambiguous constraints.

(b) Role of the parameters: Different parameters

Changing the distance cut-off and/or the chemical shift tolerance generates differences in the constraints. Between CAL1 and CAL2, the cut-off distance decreases from 10.0 to 9.0 Å and the differences of constraints between the two runs are 18%. The same result is observed between HST1 and HST2 or HST3 (18% different constraints are observed). Changing the chemical shift tolerance from 0.025 to 0.03 ppm (comparison of CAL2 with CAL3) and the cut-off from 9 to 11 Å induces 30% differences in the constraints. The majority of the differences in the constraints files concerns ambiguous constraints (58% for CAL1/2/3 and HST1-HST2/3). These differences in the constraints files generate differences between the final structures (on average 1.19 Å for CAL, 1.24 Å for HST1-HST2/3).

These structural differences are higher than the precision of a single calculation (0.67 Å for CAL, 0.91 Å for HST). Nevertheless, the differences are less than 0.5 Å higher than the rmsd of a single calculation, indicating that the use of different parameters gives the same structures. The magnitude of this difference between final averaged structures provided by independent runs of structure calculation seems correlated to the percentage of ambiguous constraints combined with the number of constraints by residue.

To test the effect of increasing cut-off and chemical shift tolerance, an additional run for KAP was performed: the same parameters as for KAP2 were used but the cut-off and the chemical shift tolerance were increased to 10 Å and 0.035 ppm, respectively. A total of 444 constraints were obtained; 345 were unambiguous. The percentage of constraints differences compared to KAP2 was 22%. The precision around the structures decreased: an rmsd of 0.86 Å was obtained on the final structures on the backbone. The rmsd on the backbone between the final averaged structures was 0.82 Å. Increasing the number of ambiguous constraints decreased the precision of the structure and increased the differences between the final averaged structures. As for CAL2/3, larger chemical shift tolerances induce large changes in the constraints, with an increasing number of ambiguous constraints. These larger chemical shift tolerances (more than 0.025 ppm) may induce wrongly assigned peaks, as already demonstrated by Mumenthaler et al. (1995).

To decrease the number of ambiguous constraints and to try to decrease the differences between the final structures, supplementary iterations with a higher relative peak intensity threshold and a smaller cut-off were realised. CAL1 and CAL3 were chosen because the rmsd between the averaged structures was the highest (1.29 Å). P_{th} was increased to 0.5 and the cut-off was decreased to 9.0 Å. A total of 973 (resp. 948) constraints with 872 (resp. 855) unambiguous constraints was obtained. The number of peaks incompatible with the three-dimensional structure was stable. The percentage of constraints differences between the two additional iterations was 19%. The averaged energy was higher than the energy obtained for CAL1 or CAL3, but the number of violations remained similar. Respectively, rmsd values of 0.88 and 0.44 Å were obtained around the new final averaged structure. The rmsd obtained between the new final averaged structures on the backbone is 1.64 Å. Therefore, the lower number of ambiguous constraints did not give a better convergence of the structure. In fact, the distance between the structures increased in the poorly defined region (residues 1, 2, 12–15, 40–45, which correspond to loops in the structure) and decreased for the β-sheet and for the helices (residues 8 to 10, 21 to 32 and 49 to 60). The decrease of the number of ambiguities yields structural differences in poorly defined regions, but not in regions with well-defined secondary structures. In these poorly defined regions, the small number of constraints due to the variability of the structure could yield different unambiguous constraints. It can be noted that calcicludine has two large, poorly defined loops. This may explain the stronger effect of the number of ambiguities on the calcicludine structure compared to the structure of κ-conotoxin, this toxin having only one small, well-defined loop.

Conclusions

Automatic NOESY assignment and structure calculation have been used to perform several independent runs in order to estimate the variability of the assignment and the consequences for the resulting structures. The precision of the final structure is independent of the initial structure, but depends on the number of constraints per residue, on the size of the protein, and also on the protocol used. Two parameters seem to play a crucial role in the assignments variation: the chemical shift tolerance and the cut-off. The chemical shift tolerance and the cut-off must be sufficiently high to select the correct assignment and not too large, to avoid too many wrong assignment possibilities. A correlation between assignment variability (for example, variation of the number of ambiguous constraints) and differences between the averaged structures was noted. More than 18% differences in the constraints files created an rmsd between the final averaged structures 0.5 Å higher than the rmsd around the averaged structure. These results prove that the procedure is robust when applied to this kind of small disulfide-bonded proteins.

References

Aue, B.P., Bartholi, E. and Ernst, R.R. (1976) *J. Chem. Phys.*, **64**, 2229–2246.

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

Bonvin, A., Boelens, R. and Kaptein, R. (1993) In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental*

Applications, Vol. 2 (Eds., van Gunsteren, W., Weiner, P. and Wilkinson, A.), ESCOM, Leiden, p. 407.
Braunschweiler, L. and Ernst, R.R. (1983) *J. Magn. Reson.*, **B53**, 521–528.
Brünger, A. (1992) *X-PLOR Version 3.1: A System for X-Ray Crystallography and NMR*, Yale University Press, New Haven, CT.
Brüschweiler, R. and Case, D. (1994) *Prog. NMR Spectrosc.*, **26**, 27.
Buchler, N.E., Zuiderweg, E.R., Wang, H. and Goldstein, R.A. (1997) *J. Magn. Reson.*, **125**, 34–42.
Chalaoux, F.R., O'Donoghue, S.I. and Nilges, M. (1999) *Proteins*, **34**, 453–463.
Clore, G.M., Robien, M.A. and Gronenborn, A.M. (1993) *J. Mol. Biol.*, **231**, 82–102.
Folmer, R.H., Hilbers, C.W., Konings, R.N. and Nilges, M. (1997) *J. Biomol. NMR*, **9**, 245–258.
Fraternali, F., Amodeo, P., Musco, G., Nilges, M. and Pastore, A. (1999) *Proteins*, **34**, 484-496.
Gilquin, B., Lecoq, A., Desne, F., Guenneugues, M., Zinn-Justin, S. and Menez, A. (1999) *Proteins*, **34**, 520–532.
Hare, B.J. and Prestegard, J.H. (1994) *J. Biomol. NMR*, **4**, 35–46.
Hyberts, S.G., Marki, W. and Wagner, G. (1987) *Eur. J. Biochem.*, **164**, 625–635.
James, T.L. (1991) *Curr. Opin. Struct. Biol.*, **1**, 1042.
Koradi, R., Billeter, M., Engeli, M., Güntert, P. and Wüthrich, K. (1998) *J. Magn. Reson.*, **135**, 288–297.
Kumar, A., Ernst, R.R. and Wüthrich, K. (1980) *Biochem. Biophys. Res. Commun*, **95**, 1–6.
Laskowski, R.A., Rullmann, J.A., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) *J. Biomol. NMR*, **8**, 477–486.
Lebrun, B., Romi-Lebrun, R., Martin-Eauclaire, M.F., Yasuda, A., Ishiguro, M., Oyama, Y., Pongs, O. and Nakajima, T. (1997) *Biochem. J.*, **328**, 321–327.
Liu, Y., Zhao, D., Altman, R. and Jardetzky, O. (1992) *J. Biomol. NMR*, **2**, 373–388.
Liu, H., Farr-Jones, S., Ulyanov, N.B., Llinas, M., Marqusee, S., Groth, D., Cohen, F.E., Prusiner, S.B. and James, T.L. (1999) *Biochemistry*, **38**, 5362–5377.
Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.
Morelle, N., Brutscher, B., Simorre, J.-P. and Marion, D. (1995) *J. Biomol. NMR*, **5**, 154–160.
Moseley, H.N. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.
Mumenthaler, C. and Braun, W. (1995) *J. Mol. Biol.*, **254**, 465–480.
Mumenthaler, C., Güntert, P., Braun, W. and Wüthrich, K. (1997) *J. Biomol. NMR*, **10**, 351–362.
Nilges, M. (1995) *J. Mol. Biol.*, **245**, 645–660.
Nilges, M., Macias, M.J., O'Donoghue, S.I. and Oschkinat, H. (1997) *J. Mol. Biol.*, **269**, 408–422.
Nilges, M. and O'Donoghue, S.I. (1998) *Prog. NMR Spectrosc.*, **32**, 107–139.
Olson, J.B. Jr. and Markley, J.L. (1994) *J. Biomol. NMR*, **4**, 385–410.
Pardi, A., Billeter, M. and Wüthrich, K. (1984) *J. Mol. Biol.*, **180**, 741–751.
Rance, M., Sørensen, O.W., Bodenhausen, G., Wagner, G., Ernst, R.R. and Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun.*, **117**, 479–485.
Savarin, P., Guenneugues, M., Gilquin, B., Lamthanh, H., Gasparini, S., Zinn-Justin, S. and Menez, A. (1998) *Biochemistry*, **37**, 5407–5416.
Savarin, P., Romi-Lebrun, R., Zinn-Justin, S., Lebrun, B., Nakajima, T., Gilquin, B. and Menez, A. (1999) *Protein Sci.*, **8**, 2672–2685.
Schneider, T.R., Brünger, A.T. and Nilges, M. (1999) *J. Mol. Biol.*, **285**, 727–740.
Schweitz, H., Heurteaux, C., Bois, P., Moinier, D., Romey, G. and Lazdunski, M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 878–882.
Terlau, H., Shon, K.J., Grilley, M., Stocker, M., Stuhmer, W. and Olivera, B.M. (1996) *Nature*, **381**, 148–151.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.
Xu, Y., Wu, J., Gorenstein, D. and Braun, W. (1999) *J. Magn. Reson.*, **136**, 76–85.
Zhao, D. and Jardetzky, O. (1994) *J. Mol. Biol.*, **239**, 601–607.