Journal of Biomolecular NMR, **22:** 249–263, 2002.
KLUWER/ESCOM

Automated assignment and 3D structure calculations using combinations of 2D homonuclear and 3D heteronuclear NOESY spectra *

Numan Oezguen[a], Larisa Adamian[b], Yuan Xu[a], Krishna Rajarathnam[a] & Werner Braun[a,*]
[a]*Department of Human Biological Chemistry and Genetics and Sealy Center for Structural Biology, University of Texas Medical Branch, Galveston, TX 77555-1157, U.S.A.;* [b]*Department of Bioengineering, SEO, MC 063, University of Illinois at Chicago, 851 S. Morgan Street, Chicago, IL 60607-7052, U.S.A.*

Received 16 October 2001; Accepted 14 January 2002

Key words: automatic NOESY assignment, automatic structure determination, chemokine, DIAMOD, global fold, MPIF, NOAH

Abstract

The NOAH/DIAMOD suite uses feedback filtering and self-correcting distance geometry to generate 3D structures from unassigned NOESY spectra. In this study we determined the minimum set of experiments needed to generate a high quality structure bundle. Different combinations of 3D ^{15}N-edited, ^{13}C-edited HSQC-NOESY and 2D homonuclear ^{1}H-^{1}H NOESY spectra of the 77 amino acid protein, myeloid progenitor inhibitory factor-1 (MPIF-1) were used as input for NOAH/DIAMOD calculations. The quality of the assignments of NOESY cross peaks and the accuracy of the automatically generated 3D structures were compared to those obtained with a conventional manual procedure. Combining data from two types of experiments synergistically increased the number of peaks assigned unambiguously in both individual spectra. As a general trend for the accuracy of the structures we observed structural variations in the backbone fold of the final structures of about 2 Å for single spectral data, of 1 Å to 1.5 Å for double spectral data, and of 0.6 Å for triple spectral data sets. The quality of the assignments and 3D structures from the optimal data using all three spectra were similar to those obtained from traditional assignment methods with structural variations within the bundle of 0.6 Å and 1.3 Å for backbone and heavy atoms, respectively. Almost all constraints (97%) of the automatic NOESY cross peak assignments were cross compatible with the structures from the conventional manual assignment procedure, and an even larger proportion (99%) of the manually derived constraints were compatible with the automatically determined 3D structures. The two mean structures determined by both methods differed only by 1.3 Å rmsd for the backbone atoms in the well-defined regions of the protein. Thus NOAD/DIAMOD analysis of spectra from labeled proteins provides a reliable method for high throughput analysis of genomic targets.

Abbreviations: NOAH, program that analyzes NMR-spectra and structures to generate new and evaluate existing assignments; DIAMOD, distance geometry program that generate 3D structures; MPIF-1, myeloid progenitor inhibitory factor-1; rmsd, root mean square displacement of atom positions in optimally superimposed structures; drms, root mean square of differences of all distances between corresponding pairs of $C_\alpha$atoms in 2 or more different structures, drms $=\sqrt{\frac{\sum_{i}^{N}(d_{ik}-d_{il})^2}{N}}$, d_{ik} and d_{il} are corresponding distances of the C_α-atoms in structure *k* and *l* respectivly; TAL, test assignments list; AAL, ambiguously assignments list; UAL, unambiguously assignments list; n, ambiguity parameter that determines the maximal number of possible different assignments of a peak to be taken into the TAL; L0, L1, L2, filter numbers in percent of loaded structures used in evaluation of assignments; Pvio, number in percent of loaded structures violating a given assignment; NV(Li), number of assignments for a given peak violated by less than Li percent of the loaded structures; Δtol_i, tolerance for the chemical shifts in dimension i.

*To whom correspondence should be addressed. E-mail: werner@newton.utmb.edu

Introduction

High-throughput structure analysis of proteins from NMR data sets requires the development of efficient protocols to determine the global fold of proteins with minimal data sets. The two areas where substantial improvement in efficiency can be made is in automating the assignment of spectra and in reducing the amount of data needed to completely define the 3D structure (Bailey-Kellogg et al., 2000; Christendat et al., 2000; Cort et al., 1999; Fowler et al., 2000; Kozlov et al., 2000; Montelione et al., 2000). We and others have demonstrated that automatic interpretation of spectra is faster and at least as accurate as manual methods (Moseley et al., 1999; Nilges, 1997; Xu et al., 1999a). Semi automatic and automatic protocols were used to assign NOESY cross peaks and generate 3D structures simultaneously (Cierpicki et al., 2000; Civera et al., 1999; Duggan et al., 2001; Fraternali et al., 1999; Hare et al., 1999; Kovacs et al., 2001; Nilges et al., 1997; Pascual et al., 1997; Xu et al., 1999b, 2001), but there is no established procedure to determine the 3D fold of a protein in a high-throughput mode. It is also not known what minimal NMR data set could uniquely define the global fold with high reliability.

Our NOAD/DIAMOD suite (Mumenthaler et al., 1995), when supplied with a reasonably complete list of chemical shifts, simultaneously assigns NOESY cross peaks and generates 3D structures by using feedback filtering and self-correcting distance geometry. While most methods for NOESY interpretation require substantial user input, NOAH/DIAMOD generates high quality structures in a completely automatic fashion. In previous tests, this suite automatically assigned simulated and experimental homo- and heteronuclear edited NOESY spectra and determined 3D structures in a reliable fashion (Mumenthaler et al., 1997; Xu et al., 1999b). We recently determined the 3D structure of the 60 residues protein, neurotoxin CsE-v5 from the New World scorpion *Centruroides sculpturatus Ewing*. The NOAD/DIAMOD structure, which was calculated from an automatically picked peak list, agreed in many details with that obtained independently in N. Rama Krishna's group using conventional manual assignment of NOESY peaks and calculation with the XPLOR program (Xu et al., 2001). The root-mean-square deviation (rmsd) between the automatically and manually determined structures was less than 1 Å for the well-defined regions.

Table 1. Experimental data used in NOAH/DIAMOD calculations

Data type	Number of entries
Spin chemical shift data	478
Peaks from ^{15}N-edited 3D HSQC-NOESY	588
Peaks from ^{13}C-edited 3D HSQC-NOESY	578
Peaks from ^{1}H-^{1}H 2D NOESY in D_2O	200
Stereospecific assignment	18
Disulfide bond constraints	3
Hydrogen bond constraints	20
^{3}J-coupling constants	30

Compared to previous calculations with NOAH/-DIAMOD we reduced the number of user supplied parameters and optimized the parameters for combinations of 2D and 3D spectra. In this paper we study the effect of combinations of 2D ^{1}H-^{1}H NOESY spectra, 3D ^{15}N-edited and ^{13}C-edited NOESY heteronuclear NMR spectra on the quality of 3D structures of myeloid progenitor inhibitory factor-1 (MPIF-1). We especially examine what is the minimal set of spectra needed to get an accurate backbone fold, since the structural information provided by multiple NMR spectra are intrinsically redundant. We check the quality of our automated assignments and structures using assignments and structures derived by a traditional manual procedure (Rajarathnam et al., 2001) from the same spectra. We find excellent agreement in assignments and 3D structures, and our calculations also suggest that it might be possible to determine the global fold of a large number of genomic targets without ^{13}C labeling of proteins.

Materials and methods

Input data for NOAH/DIAMOD

The 2D homonuclear ^{1}H-^{1}H NOESY in D_2O, 3D heteronuclear ^{15}N-edited and ^{13}C-edited HSQC-NOESY spectra used here to automatically determine the structure of MPIF-1 by NOAH/DIAMOD were used previously for the manual structure calculation of MPIF-1 (Rajarathnam et al., 2001). The NMR spectra were processed and NOESY cross peaks were semi-automatically picked using the nmrPipe program suite (Delaglio et al., 1995) and the output reformatted for input to NOAH/DIAMOD. The chemical shifts of the main-chain NH and N resonances were obtained from the assignment of HNCACB (Grzesiek

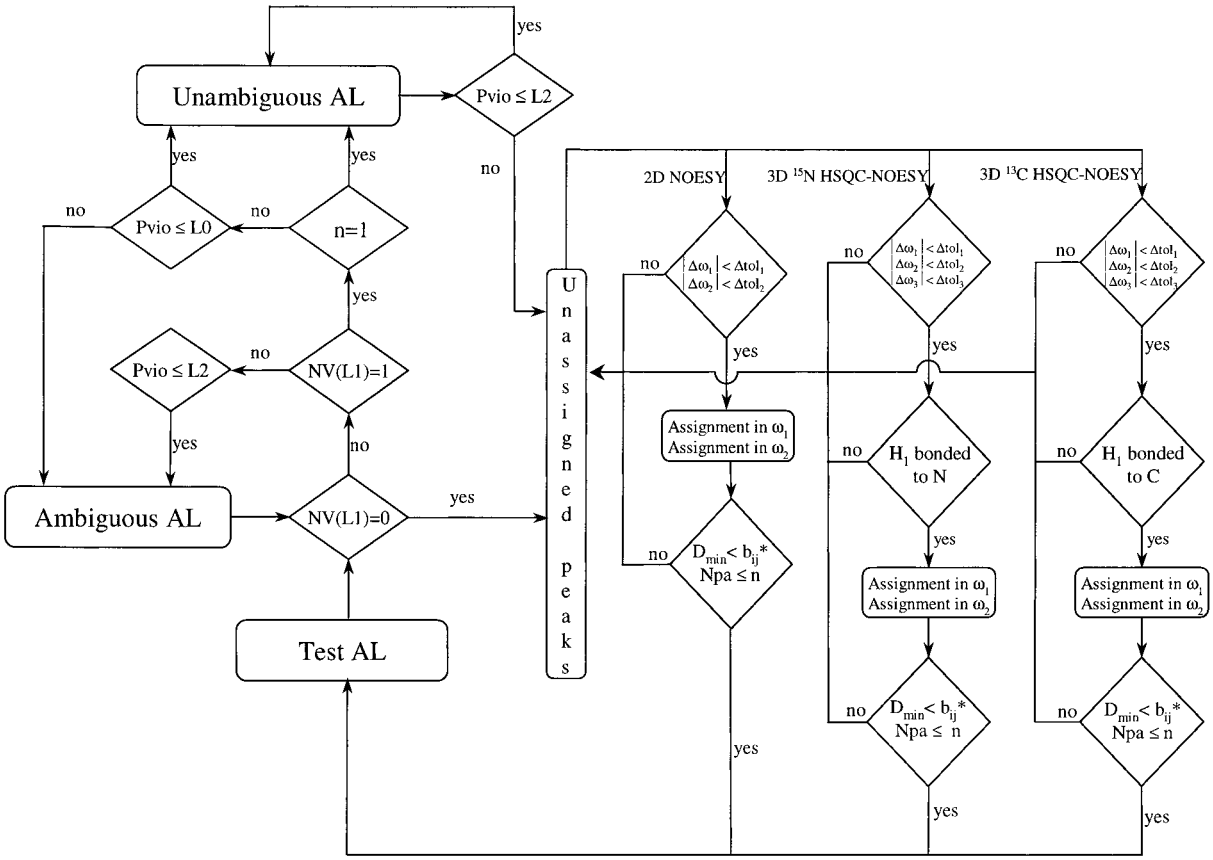

Figure 1. Flowchart of NOAH: Assignments for the NOESY cross peaks are classified in three levels: unambiguous assignments (UAL), ambiguous assignments (AAL) and test assignments (TAL), if they match the criteria defined in the block diamonds. A cross peak at position $(\tilde{\omega}_1, \tilde{\omega}_2, \tilde{\omega}_3)$ is assigned to spin chemical shifts $(\omega_1, \omega_2, \omega_3)$ if the peak falls within the interval $([\omega_1 \pm \Delta tol_1], [\omega_2 \pm \Delta tol_2], [\omega_3 \pm \Delta tol_3])$, where $(\Delta\ tol_i)$ is a user specified tolerance for each dimension. N_{pa} is the number of possible assignments for a cross peak. A peak is considered for assignment if N_{pa} is less than or equal to a user defined threshold (typically n = 2..4). The minimal distance D_{min} of an assignment is calculated from the ensemble of the 10 best structures of the previous NOAH/DIAMOD cycle, and is compared to the limit $b^*_{ij} = b_{ij} + d_{tol}$, the distance calculated from the peak intensity b_{ij} plus a tolerance parameter d_{tol}. The number of structures, in percent of the 10 best, violated by an assignment is given by P_{vio}. NV(L1) counts the number of assignments for a given peak violated by less than L1 percent of the structures. The filter parameters L0, L1 and L2 are user defined.

et al., 1992; Muhandiram et al., 1994) and CBCA (CO)NH (Muhandiram et al., 1994; Wittekind et al., 1993) spectra. The chemical shifts of the side-chain atoms were assigned from ^{15}N-edited total correlation spectroscopy (TOCSY) (Zhang et al., 1994) and HCCH-TOCSY (Kay et al., 1993) experiments. High-resolution 2D ^{1}H-^{1}H NOESY, TOCSY and DQF-COSY experiments were performed to assign the aromatic protons. Tolerances for chemical shift differences in the different spectra were estimated for a few sample cross peaks, and used as guidelines for NOAH/DIAMOD parameters.

Stereo specific assignments of the β-protons and χ^1 restraints were obtained by the analysis of the ^{3}J-coupling constants from an HACAHB experiment and the relative intensities of the NOE's from the NH and the C_αH to C_βH protons in NOESY spectra. Stereo specific assignments of leucine methyl protons were made based on the relative NOE intensity of the C_αH to the CH_3 protons after establishing the χ^1 angle.

The 30 ϕ angular restraints were experimentally obtained from an HNHA experiment (Kuboniwa et al., 1995). Additional dihedral angular constraints for the remaining amino acid residues were based on the empirical distribution of the ϕ, ψ and χ^1 angles for the individual residue types (Abagyan et al., 1994). The hydrogen bond constraints were proposed from the observation of amide protons in a series of slow exchange 2D ^{1}H-^{15}N HSQC experiments with upper (2.3 Å) and lower (1.7 Å) HN-O hydrogen bond constraints.

Table 2. Input parameter for NOAH

Calculation set	Spectrum	L0	L1	L2	Δtol_1	Δtol_2	Δtol_3	d_{tol}
Set 0	2D ^{1}H-^{1}H	10	70	40	0.020	0.020	–	0.8 drms
(no H-bond)	3D ^{15}N-edited	10	70	40	0.025	0.025	0.400	0.8 drms
	3D ^{13}C-edited	10	70	40	0.030	0.030	0.500	0.8 drms
Set 1	2D ^{1}H-^{1}H	10	70	50	0.025	0.025	–	0.8 drms
(with H-bond)	3D ^{15}N-edited	10	70	50	0.030	0.030	0.400	0.8 drms
	3D ^{13}C-edited	10	70	50	0.030	0.030	0.500	0.8 drms

Δtol_1 and Δtol_2 are chemical shift tolerances for protons in dimension 1 and 2.
Δtol_3 is the chemical shift tolerance in the third dimension for ^{15}N or ^{13}C.
L0, L1, L2 are filter thresholds for evaluation of existing and new assignments.
drms is a number calculated by NOAH as a measure for the convergence of the loaded bundle.
d_{tol} + bij = bij* is the maximum distance violation from the loaded bundle of a potential assignment to be taken to the TAL in NOAH.

All experimental data used in the automated structure determination of MPIF-1 by NOAH/DIAMOD are summarized in Table 1.

Distance constraints were individually calibrated for each spectral data set. Cross peak intensities were converted to upper distance constraints according to the equation: $I_{i,j} = Ar_{i,j}^{-6}$. The coefficient A was calculated from the assumption that the distance $r_{i,j}$ corresponding to the strongest peak in each spectrum is equal to 2.2 Å.

The NOAH/DIAMOD program suite

The current version of NOAH program can process 2D ^{1}H-^{1}H NOESY, 3D ^{15}N-edited and ^{13}C-edited NOESY spectra. The distance constraints can be generated from a single NOESY spectrum or from a combination of several NOESY spectra in each NOAH/DIAMOD cycle. Figure 1 illustrates the flowchart of the current version of NOAH. Possible assignments satisfying the condition in the block diamonds on the right hand side of Figure 1 are stored in the test assignment list (TAL). NOAH uses the NV(L1) filter (see Figure 1 caption for definition) to further classify the assignments as ambiguous (AAL) or unambiguous (UAL) based on the compatibility of the assignments with the structures calculated in previous NOAH/DIAMOD cycles. If NV(L1) is zero, the corresponding peak is removed from the test assignment list to the pool of unassigned peaks. If NV(L1) = 1 and the number of possible assignments in the test assignment list (TAL) or ambiguous assignment list (AAL) is equal to one (n = 1), then the assignment is added to the unambiguous assignment list (UAL). If n $\geq$ 2, then the assignment is added to UAL, if it violates less than L0 percent of the structures, otherwise it is added to AAL. Assignments in the UAL are removed if they are violated in more than L2 percent of the structures in a given cycle. For peaks with more than one assignment (NV(L1) $\geq$ 2), NOAH checks the structure compatibility. If any of the assignments are violated in less than L2 percent of the structures (Pvio $\leq$ L2), the cross peak assignment is transferred to AAL for use in the next cycle of structure calculation.

Several NOESY spectra can be processed sequentially in every NOAH/DIAMOD cycle; the output files are merged into one list and submitted to the distance geometry program DIAMOD to generate structures for the next cycle. In DIAMOD calculations, distance constraints derived from the UAL are weighted 5 times higher than those derived from AAL and TAL. Fixed unambiguous constraints, such as those from known disulfide or hydrogen bonds, are weighted nine times higher than those from AAL and TAL. Experimental angular constraints are weighted five times higher, while those based on statistical distribution are weighted the same as constraints from AAL or TAL.

NOAH parameters used in calculations

The design of the NOAH/DIAMOD program suite allows the user great flexibility in adapting the set of parameters to the data sets. They may be altered during each cycle of structure calculation. The conservatively chosen NOAH parameters used in this study are listed in Table 2. The chemical shift tolerances Δtol_i are introduced to account for the experimental uncertainty in peak position and chemical shift determination. The d_{tol} value is crucial for the convergence of the structures and is coupled to the spread of the structures

Table 3. Input data sets used in NOAH/DIAMOD calculations[a)]

Run	3D HSQC-NOESY		^{1}H-^{1}H 2D NOESY	Hydrogen bond constraints
	^{15}N-edited	^{13}C-edited		
N0	+	−	−	−
NC0	+	+	−	−
NH0	+	−	+	−
NCH0	+	+	+	−
C0	−	+	−	−
H0	−	−	+	−
CH0	−	+	+	−
N1	+	−	−	+
NC1	+	+	−	+
NH1	+	−	+	+
NCH1	+	+	+	+
C1	−	+	−	+
H1	−	−	+	+
CH1	−	+	+	+

N, C, H in the calculation name indicates usage of 3D 15N-edited, 13C-edited HSQC-NOESY and 2D 1H-NOESY peak lists. '0' and '1' refer to exclusion or inclusion of the 20 hydrogen bond constraints.
[a)] Disulfide constraints were used in all calculations.

as measured by the mean difference of the distances between C_α-atoms in the loaded bundle of structures (drms). Note that drms is different from the usual rmsd. In the initial cycles the spread of the structures is large (high drms), so the structure based filters should be relaxed (large d_{tol}). In this work $d_{tol} = 0.8$ drms was used. A similar strategy was used to calculate structures of crambin (Xu et al., 1999) and scorpion neurotoxin (Xu et al., 2001).

The parameter L1 together with d_{tol} controls the compatibility of the assignments with the calculated structures of the current NOAH/DIAMOD cycle. A low L1-value means that the evaluated assignments have to be in close agreement with a large fraction of the calculated structures to be promoted from the test assignments list. This potentially can lead to a bias towards a wrong fold. As a preventive measure, we chose a high value of L1 = 70 and kept it constant during all calculations. The parameter L0 determines how restrictive the filters are for a promotion to UAL. We kept L0 = 10 for all calculations. Parameter L2 controls whether assignments in AAL and UAL are in agreement with the bundle of structures after every cycle of DIAMOD calculations. Assignments are kept in their category if they violate less than L2% of the input structures.

DIAMOD calculations in every fifth cycle were run with constraints exclusively from the UAL. In our experience, this strategy promotes a faster convergence of the calculated structures (Xu et al., 2001). The ambiguity parameter *n* was set to 1 in the first cycle and was increased by 1 in cycles 10 (n = 2) and 25 (n = 3). For spectra with high peak overlap we do not recommend this strategy. In such cases NOAH would treat all overlapping peaks as one. As these combined peaks would have multiple assignments in agreement with the structures α, they would never be promoted to the UAL. Omitting these assignments would weaken the quality of resulting structures significantly.

Calculation of structures from manually assignment procedure

We compared the quality of structures obtained by automated NOAH/DIAMOD approach with those calculated from manually assigned NOESY spectra (Rajarathnam et al., 2001). In the manual assignment procedure, NOE cross peaks were classified as strong, medium, weak and very weak peak corresponding to 2.8, 3.5, 4.0 and 5.0 Å distance constraints. Manual assignment resulted in 320 intraresidual, 178 short (sequential), 84 medium and 132 long-range distance constraints (714 in total). In addition, 82 dihedral angular constraints and 36 hydrogen-bonding constraints were also used. This peak list was then used to calculate a bundle of 30 structures (MAN structures)

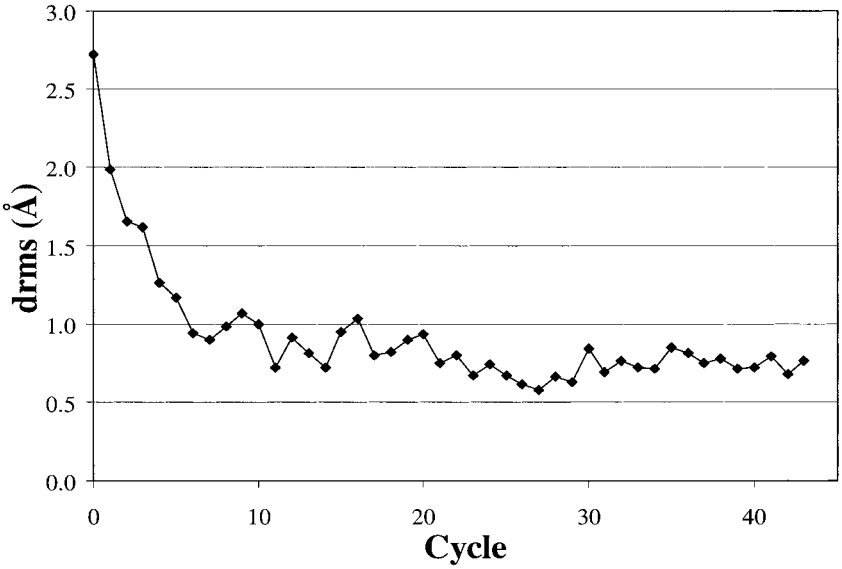

Figure 2. The mean of differences of C_α-distances among the 10 best structures (drms values) is plotted *versus* NOAH/DIAMOD cycles for the NCH1 calculation with input from the 3D ^{15}N edited HSQC-NOESY, 3D ^{13}C edited HSQC-NOESY and 2D ^{1}H-NOESY spectra.

with the combined distance geometry-dynamical simulated annealing method implemented in the XPLOR program (Brünger, 1993).

Results and discussion

Convergence properties

All possible combinations of peak lists from 2D homonuclear NOESY in D_2O, 3D heteronuclear ^{15}N and ^{13}C edited HSQC-NOESY experiments, and hydrogen bond constraints, were used for NOAH/DIAMOD calculations (Table 3). In all 14 calculations, stereo specific assignments, disulfide bond constraints and dihedral angular constraints derived from experimentally determined ^{3}J-coupling constants and the empirical distribution of the ϕ, ψ and χ^1 angles for the individual residue types were used. Calculations are named according to the specific combination of input data used. Thus the NCH1 included peak lists from 3D ^{15}N edited HSQC-NOESY (N), 3D ^{13}C edited HSQC-NOESY (C), 2D ^{1}H-NOESY (H) spectra and the hydrogen bond constraints (1). We ran all calculations uniformly for 44 NOAH/DIAMOD cycles, although the number of unambiguous assignments and the ensemble of structures converged much earlier. The mean difference of the distances between C_α-atoms in the bundle of structures (drms), a measure of convergence, for the NCH1 calculation is shown in Figure 2. The bundle converges within the first 15 cycles. In the last 20 cycles, the spread of the structures fluctuates between 0.6 Å and 0.8 Å. Figure 3 shows the progression of the number of unambiguous assignments of the NCH1 calculation for the three spectra. In the first 20 cycles approximately 70% of all 3 spectra are unambiguously assigned. In the following cycles, the numbers increased slightly.

Extent of the assignments for the different spectra

Tables 4a and 4b show the final distribution of the assignments obtained from each peak list for all 14 calculations, listed individually also for the three ranges, intraresidual, short range (i.e., assignments consisting of sequential residues), medium range (i.e., assignments with residue differences of two to four) and long range assignments.

In general, the total number of unambiguous assignments is higher if the data set consists of two or three spectra rather than only one type of spectrum. For example, the number of unambiguous assignments in the proton spectrum (1H) increases from 71 in the H0 calculation to 135 in the NCH0 calculation. As the structure bundle converges, many peaks will be moved from the AAL to the unambiguous list, as alternative wrong assignments are now incompatible with the narrower bundle. Using several different spectra tends to increase the number of unambiguously assigned peaks and improves structural convergence.

Table 4. Distribution of the assignments generated by NOAH from each spectrum

(a) Set 0

		NCH0			NC0		NH0		CH0		N0	C0	H0
		^{15}N	^{13}C	^{1}H	^{15}N	^{13}C	^{15}N	^{1}H	^{13}C	^{1}H	^{15}N	^{13}C	^{1}H
Total	UAL	408	411	135	413	430	409	135	416	137	397	412	71
Intraresidual	UAL	151	215	94	150	219	145	90	223	94	149	228	65
Short range[a]	UAL	151	74	6	156	78	154	4	74	7	151	81	4
Medium range[b]	UAL	48	40	10	45	49	50	10	43	12	48	41	1
Long range[c]	UAL	58	82	25	62	84	60	31	76	24	49	62	1
Total	All	499	591	231	492	585	491	228	598	239	541	607	224
Intraresidual	All	173	289	136	171	285	173	134	283	133	171	282	86
Short range[a]	All	184	110	12	186	106	185	17	112	19	197	111	15
Medium range[b]	All	61	74	21	57	83	57	16	72	29	71	73	28
Long range[c]	All	81	118	62	78	111	76	61	131	58	102	141	95

(b) Set 1

		NCH1			NC1		NH1		CH1		N1	C1	H1
		^{15}N	^{13}C	^{1}H	^{15}N	^{13}C	^{15}N	^{1}H	^{13}C	^{1}H	^{15}N	^{13}C	^{1}H
Total	UAL	419	416	145	416	423	411	137	432	145	417	421	135
Intraresidual	UAL	150	208	94	152	207	148	91	215	91	146	218	89
Short range[a]	UAL	159	72	8	158	76	156	4	78	6	161	79	6
Medium range[b]	UAL	49	51	12	45	51	47	13	47	12	49	44	19
Long range[c]	UAL	61	85	31	61	89	60	29	92	36	61	80	21
Total	All	479	585	212	503	605	491	222	595	243	521	613	249
Intraresidual	All	167	288	132	168	291	169	136	282	133	166	279	125
Short range[a]	All	182	106	14	196	109	189	11	109	17	193	114	16
Medium range[b]	All	60	75	17	62	89	59	18	75	23	77	78	39
Long range[c]	All	70	116	49	77	116	74	57	129	70	85	142	69

[a] ΔResidue = 1.
[b] ΔResidue = 2, 3, 4.
[c] ΔResidue $\geq$ 5.

Some peaks with multiple assignments (AAL) in one calculation became uniquely (and correctly) assigned in a calculation with more precisely defined structural bundle.

However, there are exceptions and simply adding more experiments to the analysis does not always increase the number of unambiguously assigned peaks in each spectrum. A single NOESY cross peak in the list may in reality consist of two or more overlapping cross peaks. If the convergence is low, one assignment would be compatible with the bundle of structures, and the peak would be erroneously assigned unique. As the structure bundle converges, the parts of the protein corresponding to the other assignments could become structured. The alternative assignments could then also be compatible with the structures reducing the number of uniquely assigned peaks. Another possibility for a decrease in the number of unambiguous assignments is the higher requirement for structural compatibility when data sets from several spectra are combined. With more distance constraints from different sources, the structural compatibility is more restrictive, and assignments deemed unambiguous in the low converging case fall below that threshold for inclusion in the UAL at higher convergence. Structural flexibility can also contribute to differences in the UAL for different calculations. We compared the 19 medium range UAL assignments of the ^{1}H spectrum in the H1 calculation with 12 corresponding assignments from the NCH1 calculation. Most of the differences in the assignments

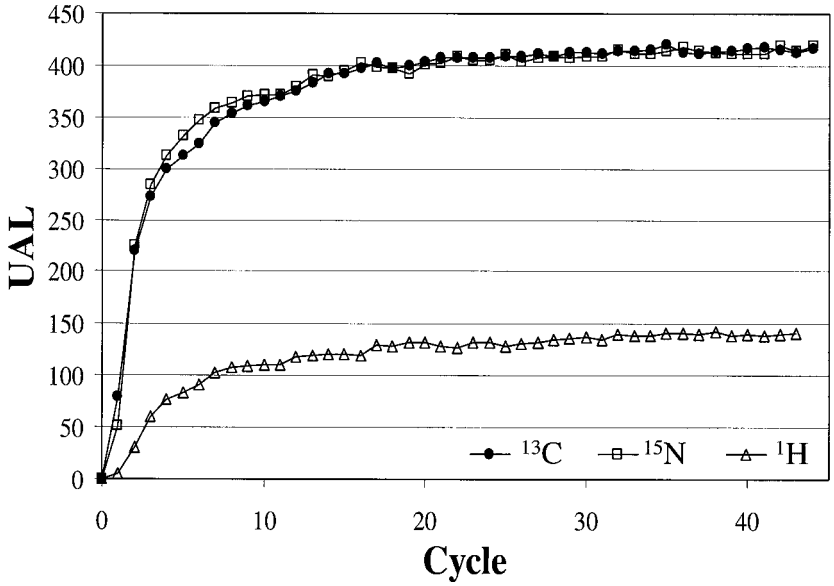

Figure 3. Number of unambiguously assigned cross peaks (UAL) for each spectrum in the NCH1 calculation.

are seen in segments around residues 25, 44 and 65, which are flexible according to the automatic structure determination with the complete data set NCH1 and in the manual assignment procedure.

As Tables 4a and 4b show, for the most part combining spectra has a synergistic effect on peak assignment. The total number of peaks in the ^{1}H spectrum that could be assigned unambiguously almost doubled when data from the ^{15}N or ^{13}C 3D spectra were combined, as seen in CH0, NH0 and NCH0. The effect is most pronounced between the calculations with one and two spectra. The synergetic effect is also observed for the 3D spectra, but the effect is much smaller, e.g., the number of unambiguous assignments for the ^{15}N peak list increased from 397 in N0 to 413 in NC0, and for the ^{13}C peak list from 412 in C0 to 430 in NC0. This increase is observed almost uniformly in each category for assignments.

Hydrogen bond constraints increase the extent of the assignments. The most profound influence was found on the extent of the assignments in the calculation with the 2D ^{1}H-^{1}H spectrum alone (H0 and H1). The structure bundle in the H0 calculation did not converge, while adding hydrogen bond constraints to this data (H1) yielded converged structure bundle and peak assignments that were nearly as complete as those obtained in the NCH1 calculation (135 versus 145).

Comparison of the accuracy of the structures

The quality of the structures was assessed by the structural variation of the ensemble of structures. We compared the average rmsd of the 10 best structures from NOAH/DIAMOD to their mean structure and to the mean of the structures based on manual assignments (MAN) (Rajarathnam et al., 2001) (Table 5). As a general trend for the accuracy of the structures we observe structural variations in the backbone fold of the final structures of about 2 Å for single spectral data, of 1 Å to 1.5 Å for double spectral data, and of 0.6 Å for the triple spectral data sets. An exception is the H0 calculation, which did not converge. The deviations within the NCH0 and NCH1 structures closely resemble the variations within the manually derived structures (rmsd values of 0.6 Å and 1.3 Å for backbone and heavy atoms, respectively). We discuss the deviations between the mean structures of NCH1 and MAN in some details below. Figure 4 illustrates the variations of the structures in a stereo view of the bundle of the 10 best NCH1 structures, showing the well-defined protein core of residues 11–66 and the unstructured N- and C-terminus.

To obtain high quality 3D structures constraints from all three spectra are necessary. However, if the goal is to obtain the global fold of a protein with minimal data sets, calculations with NH0 and NH1 data sets are surprisingly good. The deviations of the mean structures of these calculations to the mean structure

Table 5. RMSD (Å) of the bundles of 10 best structures to their mean and RMSD of the mean structures to the mean of the manually assigned structures

(a) Set 0

		MAN	NCH0	NC0	NH0	CH0	N0	C0	H0
Bundle of 10 best structures to their mean structure	bb	0.6	0.6	1.3	0.8	1.5	2.0	1.8	4.6
	heavy	1.1	1.3	2.0	1.5	2.2	2.7	2.7	5.2
Mean structure of 10 best to the mean structure of MAN	bb	–	1.3	1.6	1.2	2.4	2.0	3.1	6.2
	heavy	–	1.9	2.2	1.8	3.4	2.7	3.6	7.0

(b) Set 1

		MAN	NCH1	NC1	NH1	CH1	N1	C1	H1
Bundle of 10 best structures to their mean structure	bb	0.6	0.6	1.3	1.1	1.2	1.9	1.8	1.6
	heavy	1.1	1.3	1.9	1.7	1.9	2.7	2.5	2.3
Mean structure of 10 best to the mean structure of MAN	bb	–	1.3	1.5	1.8	2.2	2.1	2.4	2.4
	heavy	–	1.9	1.9	2.6	2.9	2.8	2.9	3.4

The rmsd values are calculated using the well-structured core of the protein (residues 11–66).

of the MAN calculation are only a few tenths of an Angstrom larger than by comparing the mean structure of NCH1 (1.2 Å to 1.8 Å versus 1.3 Å).

The hydrogen bond constraints had little effect on the final structures except in the single spectrum calculation with homonuclear 2D ^{1}H-^{1}H spectrum. The rmsd for NCH calculations is the same in both sets, but the addition of hydrogen bond constraints slightly improved the rmsd for the NC series of calculations. In general, hydrogen bond constraints are not necessary to obtain a well-defined protein core. NH0 and NH1 calculations have approximately the same number and distribution of unambiguous assignments. The deviation of the mean structure from the NH0 bundle to the mean structure of the MAN structures is smaller than the deviation of the mean of the NH1 structures to the mean of MAN. This discrepancy might be due to a difference in H-bond constraints in NOAH/DIAMOD and XPLOR calculations, detected in the final analysis. The automatic calculation included a H-bond constraint between the amide hydrogen of VAL 59 and the carbonyl oxygen of SER 55 located at the end of the helical region that was not included in the final round of calculations of the MAN structures. This hydrogen bond constraint is violated by about 3.3 Å in the MAN structures. However both the automatic and MAN structures are consistent with the input constraint lists.

In both sets, the structures obtained from the calculations NH are better than CH and structures from the calculation N are better than C, even though the number of long range constraints derived from the ^{13}C spectra are about 25–50% higher than those from the ^{15}N spectra.

The 3D structures derived from the ^{15}N spectrum combined with one other experiment, i.e., NC0, NC1, NH0 and NH1, are comparable in quality to those derived from combining all 3 spectra, NCH0 and NCH1. The deviations of the mean structures of these calculations to the mean structure of the MAN calculation are only a few tenths of an Angstrom larger than by comparing the mean structure of NCH1 (1.2 Å to 1.8 Å versus 1.3 Å). These results suggest that well-defined 3D NMR structures can be automatically determined without ^{13}C spectral data.

We suggest that constraints from the ^{15}N spectrum restrict the polypeptide backbone more than those derived from the ^{13}C-edited spectrum. At least one of the constrained atoms in the former is a backbone amide proton, whereas a large number of constraints derived from the ^{13}C spectrum are between side chain atoms only.

Table 6. Distribution of assignments violating at least 5 of 10 structures probed with different combinations. Only constraint for residues 11–66 are taken into account

	> 0.5 Å		> 1.0 Å		> 2.0		> 3.0Å		< 3.0Å[d]
MAN Str. – MAN Const.[a]	160		63		13		2		100%
MAN Str. – NCH1 Const.[b]	203	142[c]	129	88[c]	46	36[c]	17	15[c]	97%
NCH1 Str. – MAN Constr.[a]	162		86		22		6		99%
NCH1 Str. – NCH1 Constr.[b]	107	65[c]	32	23[c]	0	0[c]	0	0[c]	100%

[a]Total number of manually determined nonredundent assignments is 632.
[b]Total number of automatically accumulated nonredundent assignments is 694. 509 of them were assigned unambiguously. For this comparison only the unambiguously assignments are used.
[c]Number of violations remaining after removing HN involving constraints with ^{15}N origin.
[d]Percentage of constraints consistent with the structures.

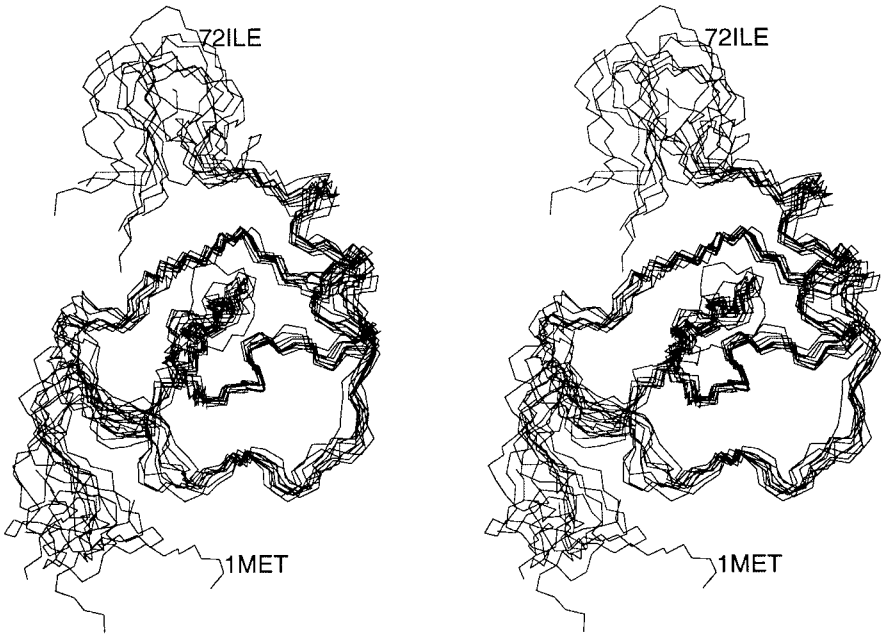

Figure 4. Stereo view of the backbone atoms for the 10 best energy refined structures in the NCH1 calculation. The C_α-atoms of the core region from residues 11 to 66 are optimally superposed.

Comparison of the structures from NCH1 with the manually derived structure

The NCH series of NOAH/DIAMOD calculations produced 3D structures of similar accuracy as the structures calculated by XPLOR from the manually derived constraints. Here we compare the well-defined core regions (residues 11–66) of NCH1 with MAN bundles of structures in more detail. The 10 best NCH1 structures were energy minimized using the program FANTOM with the ECEPP/2 force field and the DIAMOD constraints of the last cycle (Schaumann et al., 1990). PROCHECK (Laskoswski et al., 1993) confirmed the high quality of the NCH1 and MAN structures. In both cases 80% of all ϕ and ψ angles were in the most favored regions and there were no residues in the disallowed regions.

Figure 5 shows the non-redundant number of UAL per residue from the NCH1 calculation and from the manual assignment procedure. For the well-defined core region (residues 11–66), the total number of automatically determined non-redundant assignments by NOAH in the NCH1 calculation is 694, where 509 assignments were classified as unambiguous. There were 632 constraints derived by the manual assignment. The distribution of the number of constraints per residue is strikingly similar in both methods, e.g., both methods found only few constraints for the loop re-

Table 7. More than 3 Å violated assignments in 5 or more of 10 structures

	Assignment					No. of violations	Range of violation
MAN str. – MAN constr.	40	VAL QG2	28	TYR	QD	6	(2.92...3.40)
	52	ALA QB	20	ILE	QD1	7	(0.00...4.00)
MAN str. – NCH1 constr.	15	TYR HD1	52	ALA	HN	5	(0.00...3.77)
	15	TYR HE1	39	GLY	HN	6	(1.11...4.94)
	18	ARG+ QG	20	ILE	HG22	5	(2.03...4.84)
	21	PRO HA	24	LEU	HD21	9	(2.36...4.51)
	27	SER HN	43	LEU	HD11	6	(2.75...4.22)
	28	TYR HN	66	LEU	HD11	5	(1.12...4.35)
	28	TYR HB2	40	VAL	HG21	6	(2.09...3.49)
	28	TYR HD1	66	LEU	HD11	7	(0.84...5.57)
	28	TYR HE1	66	LEU	HN	9	(0.93...4.46)
	29	PHE HE1	31	THR	HG21	5	(0.00...5.08)
	40	VAL HG11	63	MET	QE	5	(0.00...5.51)
	40	VAL HG21	63	MET	QE	5	(0.00...6.53)
	43	LEU HD21	49	ARG+	HN	7	(1.85...4.28)
	50	PHE HD1	52	ALA	HA	5	(0.19...4.88)
	50	PHE HD1	52	ALA	HB3	5	(0.09...5.72)
	50	PHE HE1	52	ALA	HB3	5	(0.35...5.70)
	53	ASN HB2	58	GLN	HN	6	(2.68...3.97)
NCH1 str. – MAN constr.	13	ILE QG2	14	SER	QB	6	(1.75...4.90)
	24	LEU QB	20	ILE	QG2	10	(3.95...6.36)
	24	LEU QQD	20	ILE	HA	8	(2.80...5.09)
	35	CYSS HN	32	ASN	QB	8	(0.00...5.25)
	44	THR HN	49	ARG+	QB	6	(1.40...3.51)
	52	ALA QB	59	VAL	QG2	10	(3.16...3.89)

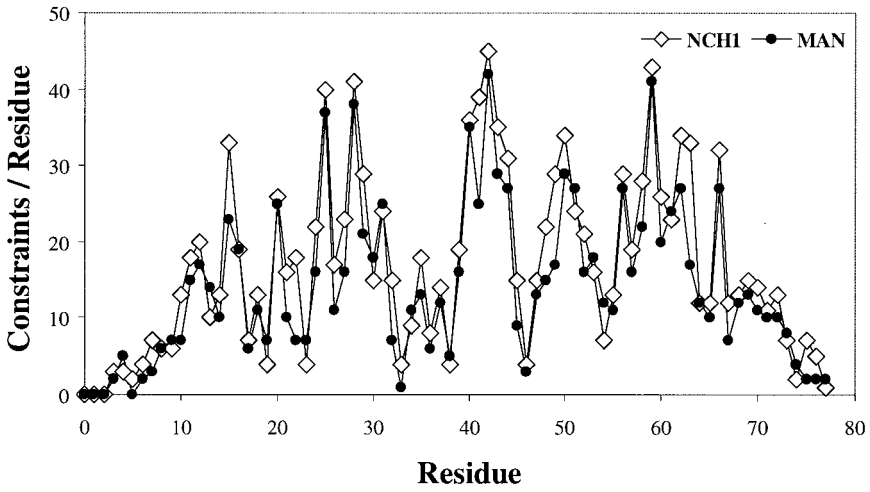

Figure 5. Distribution of the non-redundant constraints per residue of the automatic NOAH/DIAMOD assignment (NCH1 calculation) compared to the conventional manual assignment procedure (Rajarathnam et al., 2001).

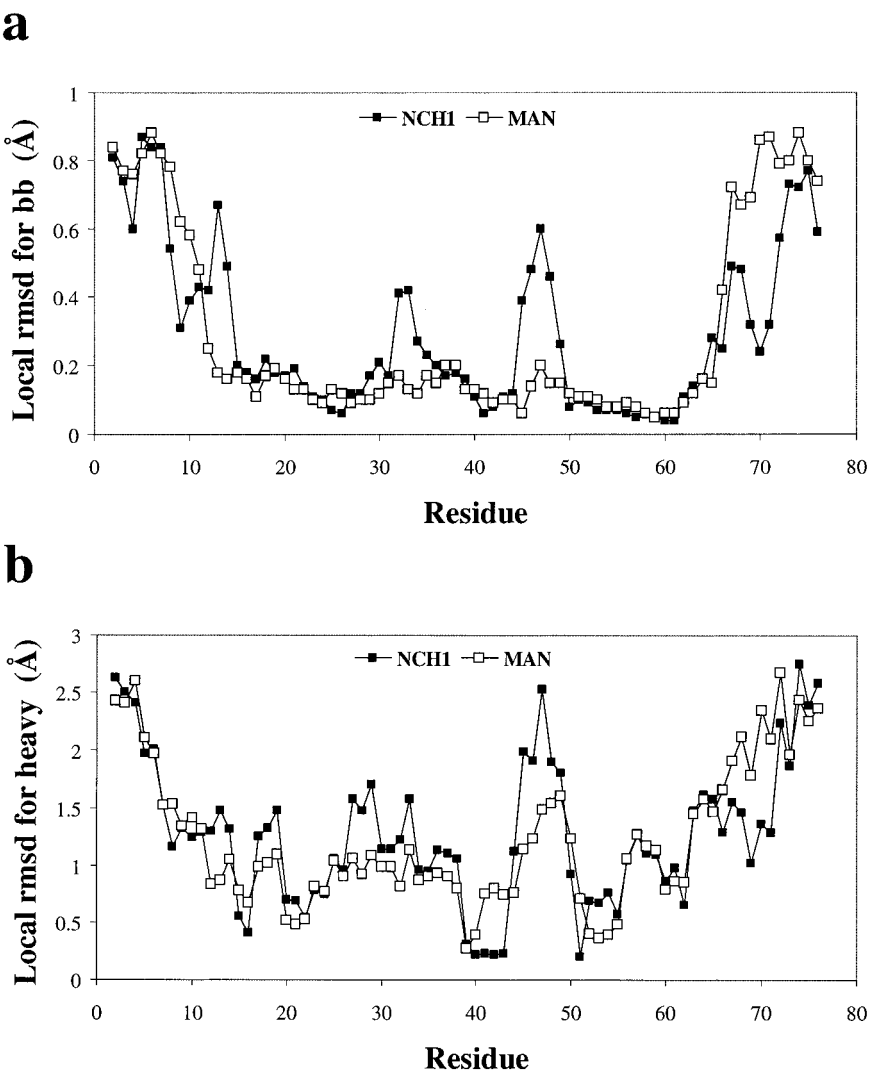

Figure 6. Distribution of the local rmsd values for backbone (a) and heavy atoms (b) of the energy refined structures from the NCH1 calculation and the manual assignment procedure (MAN). The local rmsd values are calculated by MOLMOL (Koradi et al., 1996) as the mean values of rmsd values of segments of 3 residues to the corresponding segments in the mean structure. Regions around residue SER 33 and THR 46 are not well structured loop regions.

gions around residues 33 and 46 and the local maxima in both distributions are almost identical.

A similar pattern of local variations is also seen in the local rmsd values for backbone (Figure 6a) and heavy atoms (Figure 6b). The local rmsd values calculated by MOLMOL (Koradi et al., 1996) are the average of rmsd values after locally fitting 3 residue segments of the 10 best structures to their mean. The local rmsd values for the structures calculated by both methods are 0.2 Å for backbone and 0.5 to 1.0 Å for heavy atoms, for the core region residues 11–66 exclusive of the loop regions around residues 33 and 46, for which few constraints were obtained. The N- and C-terminal regions are unstructured according to both methods due to the low number of constraints per residue.

The global fold of the mean structure of NCH1 is identical to the mean MAN structure (Figure 7). The main elements of the secondary structure, which include a three-strand β-sheet with an α-helix on top if it, superimposes very closely. Slight deviations are seen only in the loop regions around SER 33 and THR 46. Side chain packing in the core region is also consistent in most regions, as demonstrated by the low local rmsd values for heavy atoms. As an example of side chain packing we show the orientation of the pheny-

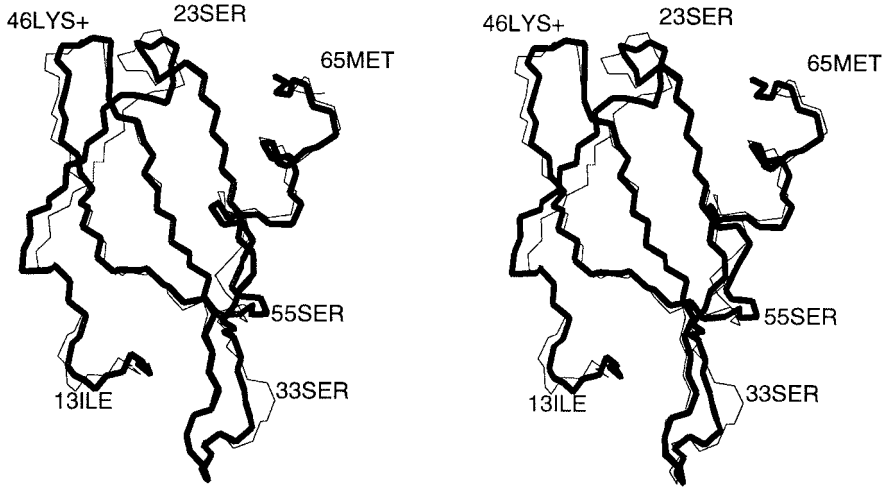

Figure 7. Stereo view of the mean structure of the 10 best energy refined structures from the NCH1 calculation (thick line) as compared to the mean structure from the manual assignment procedure (Rajarathnam et al., 2001) (thin line).

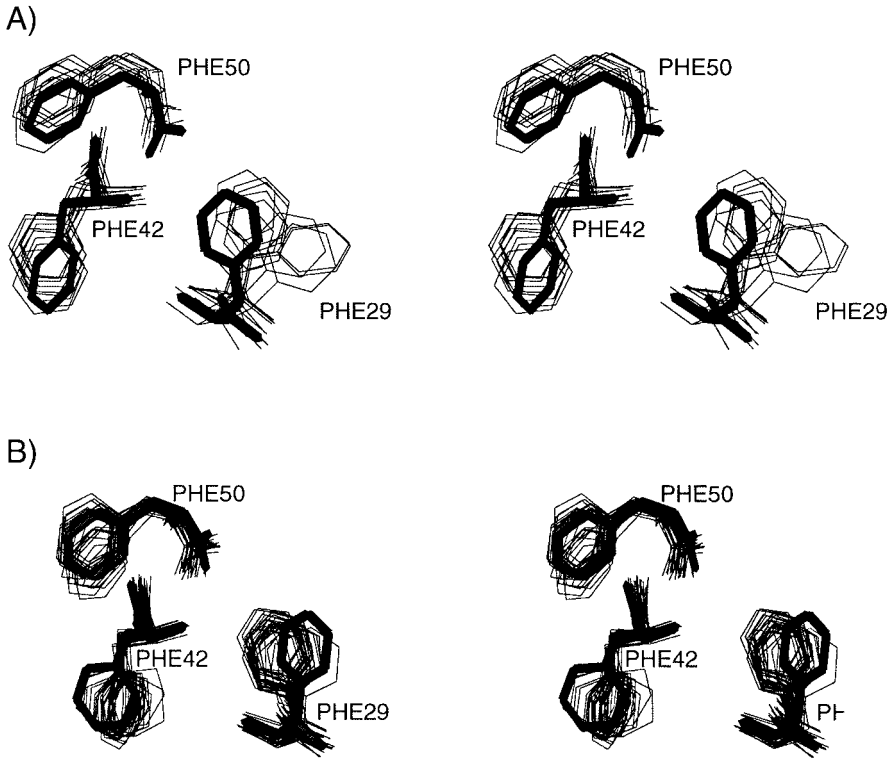

Figure 8. Side chain conformations of the 3 PHE residues in the bundle of the NCH1 structures (a) and in the bundle of the MAN structures (b).

lalanines in the bundle of the 10 NCH1 structures compared with the best MAN structure (Figure 8a) and in the bundle of the 30 MAN structures with the best NCH1 structure (Figure 8b). The best MAN structure drawn in Figure 8a is the closest to the mean of the 30 MAN structures. Similarly the structure drawn in Figure 8b is the closest to the mean of the 10 energy minimized NCH1 structures. The orientation of side chain PHE 50 is well defined in both NCH1 and MAN bundles and is identical in all structures of NCH1 and the best MAN structure. The χ^1 angle of residue PHE 42 is variable in both the NCH1 and the MAN bundle of structures. For side chains of PHE 29 in the NCH1 structures, we found two conformations. The most populated conformation was identical to that of PHE 29 in the MAN structures.

Distance constraints produced by both methods were examined for consistency with NCH1 and the MAN structures, giving a self-consistency and a cross-consistency check (Table 6). We consider assignments non-consistent, if they are violated by more then 3 Å in at least 5 of 10 structures. Both methods are 100% self-consistent according to this criterion. The cross consistency of the distance constraints, assessed by calculating the violations of the constraints obtained by manual assignment on NOAH/DIAMOD structures and vice versa, was 97% or better. The largest violation of the automatically derived distance constraints is due to the assignment of the side chain proton HG21 of VAL 40 to the side chain pseudo atom QE of MET 63 (Table 7), but this large violation is only seen in half of the MAN-structures. Tracking down the origin of this assignment shows that the automatic procedure assigned two weak peaks (one from ^{15}N, one from 2D spectrum) that were unassigned in the manual procedure. In fact there are no constraints for the methyl group of side chain of MET 63 in the manual procedure. In the manual procedure the end of the side chain of MET 63 was left totally unconstrained. There are only two cases of consistent violations in ten NCH1 structures, both of which involve pseudo atoms. We propose that these differences are caused to some extent by the flexibility of the protein and cannot be definitely interpreted as structural differences.

Small deviations are more abundant and arise primarily from constraints on amide protons. In the consistency check, where we challenge the MAN structures with the NCH1 constraints, the contribution of constraints with amide protons declines from 30% at violation level 0.5 Å to 10% at violation level 2.0 Å. The calibration scheme for calculating constraints according to the peak intensity in 3D ^{15}N-spectra accounts for most of these differences.

Final remarks

We have demonstrated that the NOAH/DIAMOD suite of programs is a robust and reliable software tool for the NMR community that can process and assign all major 2D and 3D NOESY spectra. The quality of the assignments and structures is similar to those obtained from traditional assignment methods. The program suite is ready for use in experimental laboratories, where it can significantly reduce the time required to determine 3D structures of proteins. The program is available, on request, from the corresponding author (W. Braun).

We have shown the impact of using different combinations of 2D and 3D NOESY data sets on the extent of the assignments and the accuracy of the resulting 3D structures. Our model calculations suggest that the most cost efficient way to determine the global fold for a large portion of target proteins in genomics projects would be to use ^{15}N labeled proteins.

Acknowledgements

We thank Dr Catherine H. Schein for critical reading of the manuscript, and Cynthia Orlea for help in editing the manuscript. This work was supported by grants from the National Science Foundation (DBI 9714937) and the Department of Energy (Grant DE-FG03-00ER63041) to WB.

References

Abagyan, R. and Totrov, M. (1994) *J. Mol. Biol.*, **235**, 983–1002.

Bailey-Kellogg, C., Widge, A., Kelley, J.J., Berardi, M.J., Bushweller, J.H. and Donald, B.R. (2000) *J. Comput. Biol.*, **7**, 537–558.

Brünger, A.T. (1993). *XPLOR Version 3.1 Manual*, Yale University, New Haven, CT.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C.H. and Edwards, A.M. (2000) *Prog. Biophys. Mol. Biol.*, **73**, 339–345.

Cierpicki, T. and Otlewski, J. (2000) *J. Mol. Biol.*, **302**, 1179–1192.

Civera, C., Vazquez, A., Sevilla, J.M., Bruix, M., Gago, F., Garcia, A.G. and Sevilla, P. (1999) *Biochem. Biophys. Res. Commun.*, **254**, 32–35.

Cort, J.R., Koonin, E.V., Bash, P.A. and Kennedy, M.A. (1999) *Nucl. Acids Res.*, **27**, 4018–4027.

Delaglio, F., Gresiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

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

Fowler, C.A., Tian, F., Al-Hashimi, H.M. and Prestegard, J.H. (2000) *J. Mol. Biol.*, **304**, 447–460.

Fraternali, F., Amodeo, P., Musco, G., Nilges, M. and Pastore, A. (1999) *Proteins*, **34**, 484–496.

Grzesiek, S. and Bax, A. (1992) *J. Am. Chem. Soc.*, **114**, 6291–6293.

Hare, B.J. and Wagner, G. (1999) *J. Biomol. NMR*, **15**, 103–113.

Kay, L.E., Xu, G.Y., Singer, A.U., Muhandiram, D.R. and Forman-Kay, J.D. (1993) *J. Magn. Reson.*, **101**, 333–337.

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

Kovacs, H., Comfort, D., Lord, M., Yudkin, M., Campbell, I.D. and Nilges, M. (2001) *J. Biomol. NMR*, **19**, 293–304.

Kozlov, G., Ekiel, I., Beglova, N., Yee, A., Dharamsi, A., Engel, A., Siddiqui, N., Nong, A. and Gehring, K. (2000) *J. Biomol. NMR*, **17**, 187–194.

Kuboniwa, H., Grzesiek, S., Delaglio, F. and Bax, A. (1995) *J. Biomol. NMR.*, **4**, 871–878.

Laskoswski, R.A., MacArthur, M.W., Moss, D.S. and Thronton, J.M. (1993) *J. Appl. Crystallogr.*, **26**, 283–291.

Montelione, G.T., Zheng, D., Huang, Y.J., Gunsalus, K.C. and Szyperski, T. (2000) *Nat. Struct. Biol.*, **7**, 982–985.

Moseley, H.N. and Montelione, G.T. (1999) *Curr. Opin. Str. Biol.*, **9**, 635–642.

Muhandiram, D.R. and Kay, L.E. (1994) *J. Magn. Reson.*, **103**, 203–216.

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

Mumenthaler, C., Guntert, P., Braun, W. and Wuthrich, K. (1997) *J. Biomol. NMR*, **10**, 351–362.

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

Pascual, J., Pfuhl, M., Walther, D., Saraste, M. and Nilges, M. (1997) *J. Mol. Biol.*, **273**, 740–751.

Rajarathnam, K., Li, Y., Rohrer, T. and Gentz, R. (2001) *J. Biol. Chem.*, **276**, 4909–4916.

Schaumann, T., Braun, W. and Wuthrich, K. (1990) *Biopolymers*, **29**, 679–694.

Wittekind, M. and Mueller, L. (1993) *J. Magn. Reson.*, **101**, 201–205.

Xu, Y., Jablonsky, M.J., Jackson, P.L., Braun, W. and Krishna, R. (2001) *J. Magn. Reson.*, **148**, 35–46.

Xu, Y., Schein, C.H. and Braun, W. (1999a) Combined automated assignment of NMR spectra and calculation of three-dimensional protein structures. In *Biological Magnetic Resonance*, Vol. 17, Berliner, L.J. and Rama Krishna, N. (Eds.), Plenum Publishers, New York, pp. 37–39.

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

Zhang, O., Kay, L.E., Olivier, J.P. and Forman-Kay, J.D. (1994) *J. Biomol. NMR*, **4**, 845–858.