Journal of Magnetic Resonance **136,** 76–85 (1999)
Article ID jmre.1998.1616, available online at http://www.idealibrary.com on IDEAL®

Automated 2D NOESY Assignment and Structure Calculation of Crambin(S22/I25) with the Self-Correcting Distance Geometry Based NOAH/DIAMOD Programs

Yuan Xu,* Jian Wu,† David Gorenstein,* and Werner Braun*,[1]

*Sealy Center for Structural Biology and Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, Texas 77555-1157; and †Rohm & Haas Company, P.O. Box 219, Bristol, Pennsylvania 19007

E-mail: werner@newton.utmb.edu.

Received June 8, 1998; revised September 15, 1998

The NOAH/DIAMOD program suite was used to automatically assign an experimental 2D NOESY spectrum of the 46 residue protein crambin(S22/I25), using feedback filtering and self-correcting distance geometry (SECODG). Automatically picked NOESY cross peaks were combined with 157 manually assigned peaks to start NOAH/DIAMOD calculations. At each cycle, DIAMOD was used to calculate an ensemble of 40 structures from these NOE distance constraints and random starting structures. The 10 structures with smallest target function values were analyzed by the structure-based filter, NOAH, and a new set of possible assignments was automatically generated based on chemical shifts and distance constraints violations. After 60 iterations and final energy minimization, the 10 structures with smallest target functions converged to 1.48 Å for backbone atoms. Despite several missing chemical shifts, 426 of 613 NOE peaks were unambiguously assigned; 59 peaks were ambiguously assigned. The remaining 128 peaks picked automatically by FELIX are probably primarily noise peaks, with a few real peaks that were not assigned by NOAH due to the incomplete proton chemical shifts list.

Key Words: **automated NMR spectra assignment; self-correcting distance geometry; crambin; NOAH; DIAMOD.**

INTRODUCTION

The assignment of cross peaks in NOESY spectra is a crucial step in protein structure determination by NMR. As manual interpretation of NMR spectra is time consuming, tedious, and error-prone, advanced iterative approaches have been suggested to automate the assignment of NOESY peaks and 3D-structure calculation (*1–18*). We have developed the NOAH/DIAMOD program suite (*9, 10*) based on feedback filtering and self-correcting distance geometry (SECODG) (*19–22*), which performed well in tests for assigning both simulated and experimental protein NOESY spectra and determining 3D structures. On average, more than 80% of NOESY peaks can be assigned within the given chemical shift tolerance and 95–99% of those peaks were correctly assigned. The structures calculated via NOAH/DIAMOD iterations gave pairwise RMSD ranging from 0.8 to 2.0 Å compared with the target structures in the nonloop regions with simulated data sets (*9*). Recent application of our approach to experimental 2D and 3D homo- and heteronuclear NOESY spectra of six proteins yielded similar structures to those determined previously from manually assigned cross peaks (*10*).

Here we describe the first *ab initio* application of our automatic method to raw spectral NMR data for crambin(Ser22, Ile25). Crambin, isolated from the seeds of *Crambe abysinica,* is a 46-residue protein with unusually high solubility in ethanol and organic solvents. This unusual solubility of crambin, as well as its homology to membrane active plant toxins (such as purothionins (*23*)), has excited much interest in the structure/function relationship of crambin. The recent expression of crambin as a fusion protein in *Escherichia coli* (*24*) means that many mutants of crambin should soon be available for structural analysis, suggesting an immediate use for a good automated method for data assignment and interpretation.

Crambin isolated from seed is a mixture of two nearly identical proteins. A high-resolution structure of one form (Pro22, Leu25) had been determined by both NMR and X-ray crystallography (*25–28*); after completion of this work the X-ray structure of the second isomer was published (*29*). As we were able to directly compare our structure with high-resolution X-ray structures, this was a useful model to demonstrate the speed and accuracy of the SECODG-based method for interpreting previously unassigned NOESY spectra. The experience showed how useful the approach is, as the time for structure determination could be cut from months to several weeks. Structural analysis and comparison allowed us to fine-tune the approach and suggest changes to allow completely automated assignment and structure calculation.

MATERIALS AND METHODS

Isolation of the Protein and NMR Methodology

A mixture of crambin(Pro-22/Leu-25) and (Ser-22/Ile-25) forms, isolated from seeds of *Crambe absynica* as previously

[1] To whom correspondence should be addressed. Fax: (409)747-6850.

1090-7807/99 $30.00

described (*30*), was separated by HPLC using a linear gradient from 20% CH_3CN/H_2O (v/v) to 50% CH_3CN/H_2O containing 1% trichloroacetic acid. The protein samples were hydrolyzed in 6 N HCl at 100°C for 48 h and the amino acid composition determined by HPLC. The NMR sample of 2.5 mM crambin-(Ser/Ile) form was prepared by dissolving about 8.2 mg purified protein in 0.7 ml of 75% d_6-acetone/20%H_2O/5%D_2O. The protein solution was then transferred to a 5-mm NMR tube. 1D and 2D ^1H NMR spectra were obtained on a Varian VXR-500 or a Varian VXR-600 NMR spectrometer. The 2D TOCSY, NOESY (200 ms mixing time), and DQF-COSY NMR spectra were acquired with 512 complex points in the t_1 dimension and 2048 points in the t_2 dimension using a sweep width of 6500 Hz (500 MHz) or 7000 Hz (600 MHz). Water suppression was obtained by irradiation of the HDO signal. Zero-filling ($\times 2$) was applied in the t_1 dimension. The data was processed with an 85° shifted sine-bell function. The final digital resolution was ~3.3 Hz/point (500 MHz) or 3.5 Hz/point (600 MHz) in both dimensions. In a DQF-COSY experiment, zero-filling was also applied to the t_2 dimension to give a final digital resolution of 1.6 Hz/point in the t_2 dimension.

Spin System Assignment

The spin systems of each type of amino acids were manually identified by 2D TOCSY NMR spectra with short (30 ms) and long (120 ms) mixing times. The *J*-connectivities between backbone NH proton and its $C_\alpha H$, $C_\beta H$ and $C_\gamma H$ were readily observed in the 120 ms mixing TOCSY spectrum, permitting assignment of these side chain protons and those of the unique spin systems Phe13, Tyr29, and Tyr44, whose aromatic resonances were located in the region of 6.5–7.6 ppm.

Sequence Specific Assignment

About 80% of the sequential connectivities could be assigned. The two α-helices, Ile7-Leu18 and Glu23-Thr30, were determined by their characteristic NH_i–NH_{i+1}, NH_i–NH_{i+3}, $C_\alpha H_i NH_{i+3}$, and $C_\beta H_i NH_{i+1}$ NOE connectivities. The antiparallel β-sheet formed by Thr1-Cys4 and Cys32-Ile35 was identified by the observation of Thr2H_α-Ile35NH, Thr2H_α-Ile34H_α, Cys4C_αH-Cys32C_αH, and Cys3C_βH-Ile33C_αH NOEs. Their slow NH-exchange rates confirmed the presence of both α-helix and β-sheet structures.

Data Processing

A NOESY spectrum at 200 ms mixing time was processed using the FELIX E-Z 2D transform protocol (20% DC offset, sine square 90° window function, without solvent suppression and baseline correction) and the matrix was rephased several times in both dimensions. The FELIX automatic peak-picking routine "pick all peaks" was used to obtain all peaks at a contour level of 0.03, and the FELIX peak filter functions were used to remove diagonal and unsymmetrical peaks with a uniform tolerance of two data points. Most of the water peaks and other artifact peaks were removed via those filter functions. FELIX volume integration and optimization with the Lorentzian lineshape algorithm were used to obtain cross peak intensities. We also manually picked some very weak peaks at lower contour levels than the level used for automated peak picking. We picked isolated, well-defined peaks down to a contour level that corresponds to cross peak intensities of 0.001 of our strongest peak. We selected 73 such isolated peaks with line shapes similar to other NOESY cross peaks. An in-house FORTRAN program was used to collect symmetric cross peak pairs and calculate the average peak volumes. This program generates an output peak file with the chemical shift data in a format suitable for the NOAH program input file.

Initial Peak Assignments

As the sequence specific assignment identified some NOE cross peaks, these were used as input in the NOAH/DIAMOD calculation (case 1 calculation). This calculation started from an input peak list file with 613 NOE cross peak intensities, of which 157 peaks (79 inter- and 78 intraresidue) were manually assigned. Six cross peaks were long-range NOESY peaks connecting the two antiparallel β-strands. As a test we treated the 157 manually assigned peaks as not assigned in a case 2 calculation. The input proton list of 204 chemical shifts for this study is listed in Table 1. Pseudo-atoms were used when the chemical shifts of protons were not stereospecifically assigned (*31*). Although crambin has only 46 residues, 15 proton chemical shifts (C_αHs and C_βHs) were missing.

NOAH can also directly read the coupling constants and translates them as angular constraints. Angular constraints from *J* coupling constants were used, along with three pairs of disulfide bridge constraints. There were 33 ϕ angle constraints and 7 χ_1 angle constraints. The cutoffs for ϕ angles are $-90° \leq \phi \geq -40°$ if $J_{HN\alpha} < 5.5$ Hz; $-160° \leq \phi \geq -80°$ if 8.0 Hz $< J_{HN\alpha} <$ 10 Hz; and $-140° \leq \phi \geq -100°$ if $J_{HN\alpha} > 10$ Hz.

NOAH/DIAMOD Assignments and Structure Calculations

Table 2 lists the NOAH parameters used (for a detailed explanation of these parameters and NOAH/DIAMOD flow charts, see reference (*9*)). The following input data were used by NOAH/DIAMOD (*9*):

(a) A nearly complete list of the chemical shifts of protons and the tolerance shift (or chemical shift fluctuation) of the protons, which is used for chemical shift based assignments.

(b) A list of experimental coupling constants used for dihedral angle constraints during the structure calculations.

(c) A 2D (or ND) experimental NOESY cross peak list.

(d) Disulfide bridge constraints for the three disulfide bridges Cys3-Cys40, Cys4-Cys32, and Cys16-Cys26.

Manual assignments, while not essential if the primary data set is complete, facilitate the assignment of automatically generated NOESY peaks. For each NOE peak, an integrated inten-

TABLE 1
Proton Chemical Shifts

Residue	HN	CαH	CβH	Others
THR1		4.225	4.115	γCH_3 1.134
THR2	8.616	5.225	3.731	γCH_3 0.869
CYS3	9.080	5.005	β_2 2.542, β_3 4.602	
CYS4	9.041	5.432	2.897[c]	
PRO5		*a*	2.023, 1.930	CγH 2.887[c], CδH (3.997, 3.794)
SER6	7.000	4.779	4.032[c]	
ILE7	9.226	4.122	1.982	γCH_2 (1.724, 1.340), γCH_3 1.041, $C_\delta H$ 0.972
VAL8	7.680	3.769	2.039	γCH_3 (1.076, 0.987)
ALA9	8.083	4.493	1.700[c]	
ARG10	7.794	4.609	2.034, 1.714	CδH 3.430, CεH 9.690, NH_2 (6.640[c], 7.061[c])
SER11	8.411	4.064[b]	4.093[c]	
ASN12	8.569	4.540	β_2 3.181, β_3 2.710	δNH_2 (6.721, 7.555)
PHE13	9.305	3.960	3.831, 3.557	CδH 7.220[c], CεH 7.461[c], CζH 7.330
ASN14	8.748	*a*	β_2 2.774, β_3 3.314	δNH_2 (7.064, 7.817)
VAL15	8.252	3.698	2.227[c]	γCH_3 (1.156, 0.990)
CYS16	9.302	3.823	2.600, 2.465	
ARG17	7.718	4.057	1.854, 1.703	CγH (1.267, 1.273), CδH (3.250, 2.600), εNH 7.420
LEU18	7.639	4.211	2.078, 1.635	δCH_3 (1.010, 0.930)
PRO19		*a*	*a*	CδH (4.070, 3.963)
GLY20	8.200	3.486, 3.700		
THR21	6.930	4.030	3.860	γCH_3 1.338
SER22	8.202	4.064	3.539[c]	
GLU23	9.685	3.422	β_2 2.021, β_3 1.767	CγH 2.882[c]
ALA24	8.597	4.104	1.470[c]	
ILE25	7.442	3.794	2.050[c]	γCH_2 (1.180, 1.100), γCH_3 0.817
CYS26	8.327	4.674	2.766, 2.492	
ALA27	9.431	4.105	1.556[c]	
THR28	7.676	3.989	*a*	γCH_3 1.130
TYR29	7.921	4.435[b]	3.238, 3.050	CδH 7252[c], $C_\varepsilon H$ 6.764[c]
THR30	7.592	4.645	4.744	γCH_3 1.434
GLY31	8.038	3.960, 5.563		
CYS32	7.759	5.192	β_2 2.871, β_3 2.497	
ILE33	9.047	4.770	1.616[c]	γCH_2 (1.095, 0.817), γCH_3 0.606, δCH_3 0.160
ILE34	8.160	4.738	1.636[c]	γCH_2 (1.373, 1.100), γCH_3 0.775
ILE35	8.490	4.998	2.040[c]	γCH_2 (1.467, 0.965), γCH_3 .816, δCH_3 0.773
PRO36		4.603	*a*	CδH 3.795[c]
GLY37	7.963	4.080[c]		
ALA38	8.472	*a*	1.446[c]	
THR39	7.716	4.552	3.959[c]	γCH_3 1.192
CYS40	8.762	4.880	β_2 2.634, β_3 3.446	
PRO41		4.612	2.420, 2.226	CδH (3.813, 3.676)
GLY42	8.837	3.852[c]		
ASP43	8.389	4.683	3.039, 2.846	
TYR44	8.105	4.475[b]	β_2 2.408, β_3 2.951	CδH 6.835[c], CεH 6.918[c]
ALA45	7.668	4.486	1.370[c]	
ASN46	8.078	4.679	2.550, 1.913	δNH_2 (6.701, 6.987)

[a] Missing chemical shift data (the missing CαH and CβH shift data are marked).
[b] Chemical shifts discovered during the NOAH/DIAMOD calculation.
[c] Pseudo-atom was used in the NOAH/DIAMOD calculation.

sity and two chemical shifts are needed to generate distance constraints.

Cross peak intensities were converted to upper distance constraints by the equation $I = Ar^{-6}$ (*9*). An upper distance limit of 2.2 Å was assigned to the strongest NOESY cross peak intensity (range 1.8–2.2 Å) to determine the constant *A*. The rest of upper distance limits were then calculated by the inverse sixth power law. The van der Waals distance was used as the lower distance limit. We modified the NOAH/DIAMOD peak weighting in this study to take account of the incomplete chemical shift data (*9*). Originally, distance constraints for a cross peak assigned unambiguously according to the NOAH criteria were weighted equally with manually assigned NOE peaks during the DIAMOD structure calculation, while ambig-

TABLE 2
NOAH Parameters Used for Crambin Assignments

Cycle	L_1%	L_2%	N_{pa}	Δ_{tol} (ppm)	d_{tol} (Å)[a]	W_M	W_{Uamb}	$W_{Amb+test}$
0	—	—	2	0.03	—	9	5	1
1	60	70	2	0.03	15	9	5	1
2	50	70	2	0.03	8.0	9	5	1
3	40	70	2	0.03	7.0	9	5	1
4	40	70	2	0.03	6.0	9	5	1
5	40	70	2	0.03	—	9	5	1
6–9	40	70	2	0.03	5.0	9	5	1
10	40	70	2	0.03	—	9	5	1
11–13	40	70	2	0.03	4.0	9	5	1
14	40	70	2	0.03	3.5	9	5	1
15	40	70	2	0.03	—	9	5	1
16	40	70	2	0.03	3.5	9	5	1
17–18	40	70	2	0.03	2.0	9	5	1
19–22	40	70	2	0.03	1.5	9	5	1
23–24	40	70	2	0.03	1.0	9	5	1
25	40	70	2	0.03	—	9	5	1
26–29	40	70	2	0.03	0.5	9	5	1
30	40	70	2	0.03	—	9	5	1
31	40	70	2	0.03	0.5	9	5	1
32–34	40	70	2	0.03	0.4	9	5	1
35–39	40	70	2	0.03	0.3	9	5	1
40–46	30	70	3	0.03	0.2	9	5	1
47–50	30	50	3	0.03	0.2	9	5	1
51–55	30	40	3	0.03	0.2	9	5	1
56–58	25	40	3	0.03	0.2	9	5	1
59–60	20	40	4	0.03	0.2	9	5	1

Note: Cycle: NOAH/DIAMOD iteration cycles; L_1, L_2 are used by NOAH for structure based peak assignment. If an assigned distance constraint violates less than L_1% of the bundle structures, then the assignment is treated as correct assignment. If the assignment violates more than L_2% of the structures, the assignment is discarded. The number of structures in the bundle is 10 in this study; N_{pa} is the upper limit for possible assignments of each peak; Δ_{tol} chemical shift tolerance; d_{tol} distance tolerance; and W's are weights for assigned distance constraints in DIAMOD structure calculation.

[a] At cycle 5, 10, 15, 20, 25, 30, only unambiguous and ambiguous distance constraints were applied to structure calculation.

uous distance constraints received one-fifth weighting. Here, the weighting ratio of manual, unambiguous, and ambiguous was 9:5:1. This modification improved the convergence of the bundle structures calculated by DIAMOD significantly. The weighting for angular constraints was the same as that for manually assigned NOE distance constraints.

At cycle 0, the NOAH program converts manually assigned peaks to upper distance constraints which are kept fixed and highly weighted in all cycles. NOAH then searches for possible assignments of every unassigned peak within the chemical shift tolerances in both dimensions. If the number of possible assignments is less than a user-defined threshold (N_{pa}), these peaks are selected, and all possible assignments of these peaks are converted to low-weighted, upper distance constraints and referred to as test assignments (*9*). These test assignments are then promoted to unambiguous or ambiguous assignments, or retained as test assignments or unassigned peaks, depending on the consistency of the distance constraints with the calculated bundle structures at the next cycle. For each cycle 40 structures are calculated by DIAMOD starting from random structures. The 10 best structures are fed back to NOAH and analyzed to improve the assignments.

To achieve convergence we had to increase the number of NOAH/DIAMOD cycles from 25 to 60 (~60 CRAY J90 hours) with this experimental data set. The number of iterations in each cycle during the minimization of the target function was also increased (*32*). Floating assignments were used for diastereotopic methylene protons.

RESULTS

Convergence Similarity with (Case 1) and without (Case 2) Manual Assignments

We tested the NOAH/DIAMOD method with two different input data sets to determine its ability to accurately and automatically calculate structures from spectral data. The peak list for case 1 included 157 manual assignments which were kept fixed during the calculation. The number of peaks assigned by NOAH as a function of NOAH/DIAMOD cycles (Fig. 1a) reaches a plateau between 50 and 60 cycles. After 60 iterative cycles, NOAH assigned 269 NOESY peaks unambiguously and 59 peaks ambiguously in addition to the 157 manual assignments. The RMSDs of the 10 best structures to their

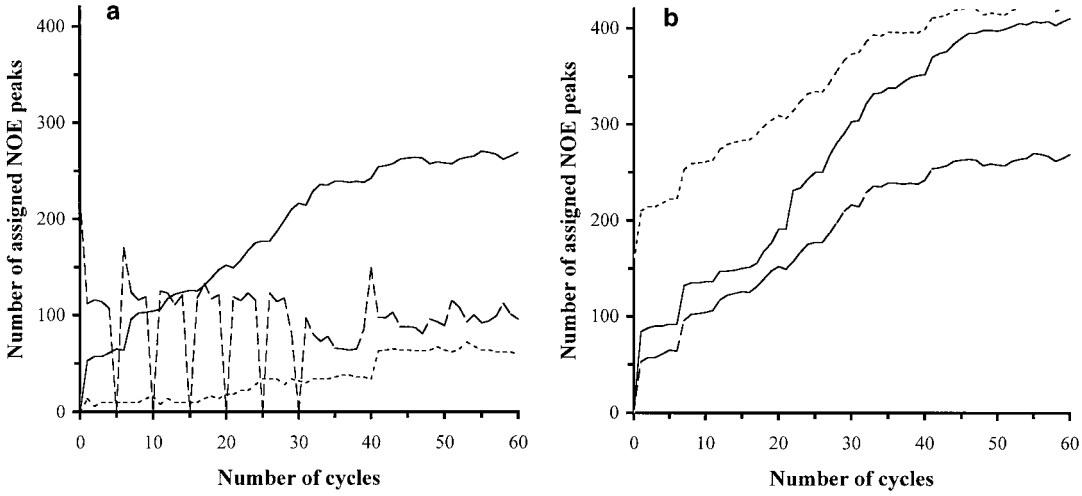

FIG. 1. (a) Number of NOESY peak assignments vs NOAH/DIAMOD cycles for case 1. Solid line: unambiguous assignments; dotted line: ambiguous assignments; dashed line: test assignments. The test assignments were excluded at cycles 5, 10, 15, 20, 25, and 30 in creating upper distance constraints for DIAMOD calculations. (b) Number of unambiguous NOESY peak assignments vs NOAH/DIAMOD cycles for the cases 1 and 2. Dashed line: number of unambiguous assignments made by NOAH for case 1 (excluding manual assignments); solid line: number of unambiguous assignments for case 2; dotted line: number of unambiguous assignments (as indicated by dashed line) plus 157 manual assignments. At the final NOAH/DIAMOD cycles, almost all of the manually assigned 157 NOESY peaks had also been assigned by NOAH in case 2.

mean structure are, respectively, 1.12 Å (global backbone), 0.67 Å (residue 2–18), 0.58 Å (residue 26–40), and 0.84 Å (residues 2–18 and 26–40) before energy minimization.

We also tested the method's ability to work in a completely automated fashion using only the peaks picked by the FELIX procedure without any manual assignments (case 2, Fig. 1b). NOAH automatically determined most of these manual assignments of case 1 during the calculation in case 2. After 60 NOAH/DIAMOD iterations using the same parameter set used for case 1, 410 peaks were assigned unambiguously and 85 ambiguously. The RMSDs of the 10 best structures to their mean structure are, respectively, 1.42 Å (global backbone), 0.77 Å (residue 2–18), 0.81 Å (residue 26–40), and 1.0 Å (residues 2–18 and 26–40) before energy minimization.

The spread of the two bundles of 10 best structures, as measured by the distance root-mean-square deviations (DRMSD), is higher in case 2 at the initial cycles, as expected, but reaches similar levels at the end of the calculation (Fig. 2a). The 3D structures are also similar in both cases. The heavy atom RMSDs of the mean structures from case 1 and 2 are 1.68 Å for backbone, 1.63 Å from residue 2 to 18, 0.61 Å from residue 26 to 40, and 1.31 Å for the nonloop regions. This shows that NOAH/DIAMOD could also make successful automated assignments and structure calculations without manual assignments.

Figure 2b shows the convergence of the mean structures at intermediate stages of the NOAH/DIAMOD calculation toward the mean structure of the final cycle in case 1. At the first three cycles (panels A, B, and C), the secondary structures are not yet completely formed, and the deviation of the global fold from the final structure with an RMSD value of about 5 Å is relatively high. However, already at around cycle 30 the mean structure is very similar to the mean structure at the final cycles with RMSD values of about 1.5 Å. This observation suggests that using the mean structure of intermediate stages as a filter for identifying correct NOESY cross peak assignments might speed up our procedure.

At cycle 0 (which started with random structures), NOAH unambiguously assigned 23 NOESY peaks with a tolerance of 0.03 ppm based on chemical shifts alone for case 1 and 33 unambiguous assignments for case 2. The latter assignments included all 23 assigned peaks of case 1 plus 9 of the manual assignments and 1 peak which was unambiguously assigned in later cycles in case 1. These unambiguous assignments are not sufficient to define the global fold of crambin, but convergence is achieved at early stages if ambiguous and test assignments are included.

Excluding test assignments from every fifth cycle (9) (at cycle 5, 10, 15, 20, 25, 30) improved the number of unambiguous assignments and the bundle's DRMSD. When test assignments were not periodically excluded, only 90 peaks were assigned unambiguously at cycle 10, and 242 by cycle 60. Although the bundle of structures all had similar DRMSD (1.19 Å vs 1.17 Å if test assignments were included in all cycles), the first α-helix was distorted in the calculation that includes test assignments at all cycles.

Table 3 lists a few of the 59 ambiguously assigned peaks from case 1 with their possible assignments suggested by NOAH at the end of calculation. These 59 ambiguously assigned peaks can in practice be further examined by the experimentalist. NOAH assigned 69.5% ((157 + 269)/613) unambiguously; 80% of the peaks were assigned if the ambiguous peaks are included. Of the unassigned peaks, many resulted from our deliberate inclusion of 73 very weak manually picked

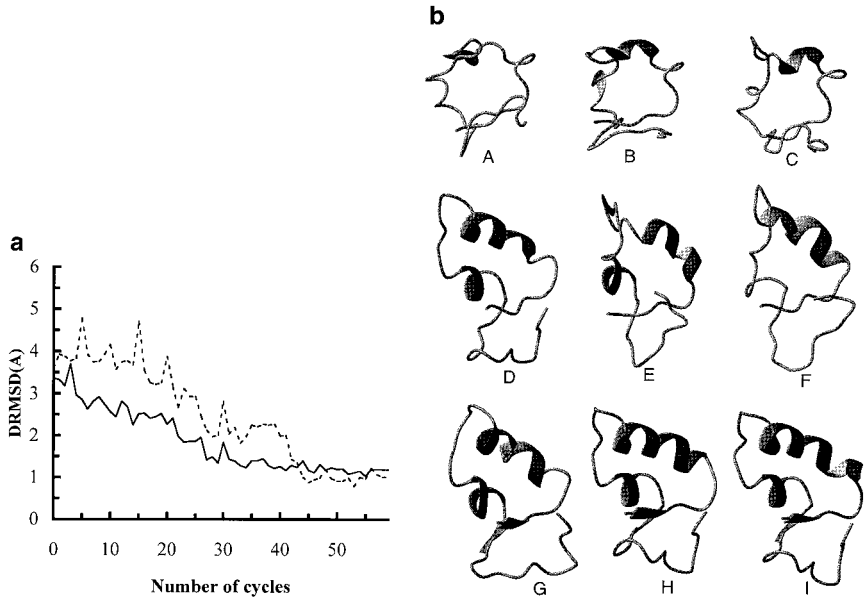

FIG. 2. (a) The convergence of the bundle structures vs NOAH/DIAMOD cycles. The 10 structures with smallest DIAMOD target function values were selected from a total of 40 structures at every cycle to compute the average DRMSD (distance RMSD) values. Solid line: case 1. Dotted line: case 2. (b) Mean structures of the 10 structures that have smallest DIAMOD target function values at intermediate stages of the NOAH/DIAMOD calculation in case 1. The cycle numbers and the root-mean-square deviations of the mean structure of a given cycle to the mean structure of the final cycle (cycle 60) are as follows for the panels A to I. A: cycle 1, 5.66 Å; B: cycle 2, 4.55 Å; C: cycle 3, 5.06 Å; D: cycle 5, 1.14 Å; E: cycle 10, 3.17 Å; F: cycle 20, 2.83 Å; G: cycle 30, 1.52 Å; H: cycle 50, 1.04 Å; I: cycle 60, 0.0 Å.

peaks. Many of these peaks were located in the water peak region or along the diagonal and have intensities 1000 times smaller than the strongest cross peak. A large fraction of these peaks are probably noise.

TABLE 3
Ambiguously Assigned Peaks

NOE peak	Proton 1	Proton 2	P_{vio} (%)	d (Å)
110	16 HB3	17 HG2	50	5.50
	16 HB3	17 HG3	30	5.50
429	9 QB	5 HD3	20	4.88
	9 QB	8 HA	100	4.88
565	44 HD1	4 HB2	10	5.50
	44 HD1	43 HB3	20	5.50
	44 HD1	5 HG2	50	5.50
602	16 HA	26 HB2	10	5.50
	25 HA	26 HB2	30	5.50
607	40 HB3	44 HB2	30	5.50

Note: Peaks with assignments that violated less than L_2 ($\leq$40%) (at the end of NOAH/DIAMOD iterations. Total number of such ambiguous assignments are 59. For definition of P_{vio} see (*11*); d is distance constraint (the upper limit) for the given assignment. For instance, peak 602 has two possible assignments. The assignment 16CHα-26CHβ2 violates the distance limit (b_{ij} + d_{tol}; see text for their definitions) in one structure out of 10 (10%). The assignment 25IHα-26CHβ2 violates the distance limit in 3 out of 10. This type of assignment should be analyzed carefully by users to make final decision. Such types of constraints usually make a very small contribution to the structure calculation.

Reliability of the NOAH Assignments

The reliability distance (RD), meaning the distance a residue must be moved to fulfill an alternate assignment (*9*), was computed for each peak (Table 4). A large RD value is a strong indication that the assignment is correct. Despite the lack of chemical shift data and the low number of manual assignments, fewer than 30% of the assignments had an RD $<$ 1 Å. At the moment we have no indices to evaluate the reliability of these peaks quantitatively. From previous experience with simulated data sets approximately half of these assignments are correct.

In case 2, the NOAH/DIAMOD procedure unambiguously and correctly found 127 out of the 157 manual assignments. For another 24 cross peaks (15%), the manual assignment was found as one of several possibilities. For the remaining six cross peaks, the assignments of pairs was reversed within the

TABLE 4
Reliability Distance Distribution[a]

RD (Å)	0	1	2	3	4	5	>5
N_{pk}	111	129	39	30	18	14	85

[a] Reliability Distance Distribution (RD) of 426 unambiguous assignments at the end of NOAH/DIAMOD iterations in case 1 calculation. A high RD for an assignment is a strong indication that it is correct. 69% of NOE peaks were unambiguously assigned and 269 peaks were assigned by NOAH. The distribution of RD in this case is similar to the theoretical RD distribution in the simulation study (*11*).

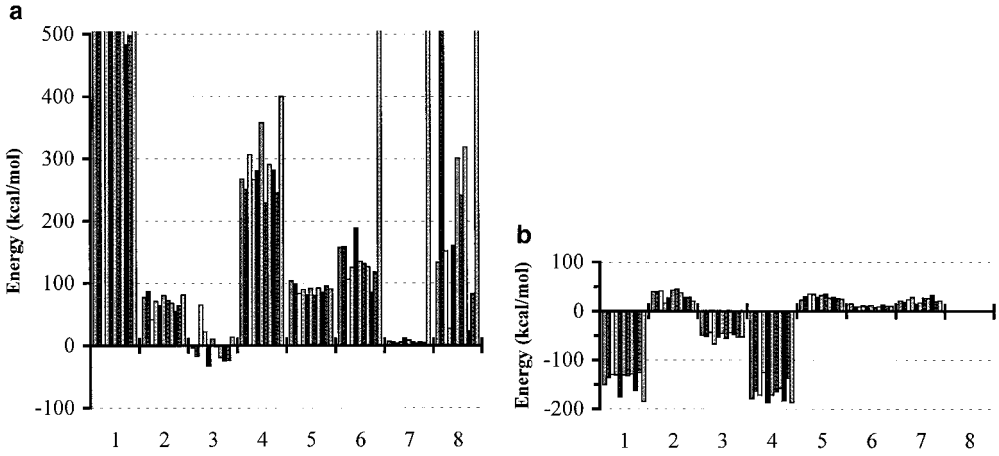

FIG. 3. (a) The energy distributions of the 10 structures with smallest DIAMOD target function values and their mean structure before the FANTOM energy minimization. (b) The energy distributions after the minimization. 1: The total energy; 2: electric energy; 3: HB energy; 4: LJ energy; 5: torsional energy; 6: S—S bond energy; 7: NOE energy; 8: dihedral energy.

chemical shift tolerance. For example, the proton pair assignments for peaks 91 and 92 with similar volumes, Ser11C_αH (4.07 ppm)-Asn12HN (8.57 ppm) and 11C_βH (4.08 ppm)-12HN (8.57 ppm), were reversed by NOAH/DIAMOD as compared to the manual assignment. This difference does not strongly influence the list of distance constraints or the final structures.

The Effect of Different Parameter Settings on NOAH/DIAMOD Calculations

We made four test runs with the number of possible assignments N_{pa} = 1, 2, 3, and 4, respectively, for the first 40 cycles, and gradually increasing N_{pa} to 4 in the final cycle. The parameter setting N_{pa} = 1 and 2 in the early cycles gave best results. At N_{pa} = 1, NOAH made 271 unambiguous assignments, at N_{pa} = 2, 269, while only 256 peaks were assigned unambiguously for N_{pa} = 3 and 4. At N_{pa} = 4, the structure bundle had a high DRMSD of 1.9 Å at the end of the 60 cycles.

Calculations with different chemical shift tolerances at N_{pa} = 2 showed that a shift tolerance of 0.02–0.03 ppm was optimal. No convergence was achieved at 0.01 ppm, where only 175 unambiguous assignments were made after 60 cycles (DRMSD ~ 2.0 Å), and only 212 unambiguous assignments were made at 0.04 ppm.

Energy Minimization with FANTOM

As in other NMR structure determinations with distance geometry methods, the final structures were energy refined (*33*) using our FANTOM program (*34, 35*) to determine low-energy conformers. In these restrained energy minimizations we use the distance constraints with a relatively high weight assuming that they are correct. We therefore only used the 426 unambiguous assignments as distance constraints. The average RMSD of the bundle to their mean is larger after the energy minimization because of the smaller number of constraints (it increased from 1.12 to 1.45 Å). FANTOM effectively removed any large violations of the Lennard–Jones, dihedral, torsional, and disulfide bridge energy terms in the initial structures and reduced the total energies from above 500 kcal/mol to ~−150 kcal/mol while only slightly increasing the NOE distance constraint violations (Figs. 3a and 3b). Superposition of the initial mean structures and the 10 final structures after the energy minimization shows that the unambiguously assigned NOE peaks were consistent in both NOAH/DIAMOD and FANTOM calculations (Fig. 4).

Identification of Missing Chemical Shifts

In previous studies, NOAH was able to work from a relatively complete chemical shift list. In this case, as is usual in practice, the chemical shifts for many protons were missing. Some of these shifts can be found during the NOAH/DIAMOD calculation. For

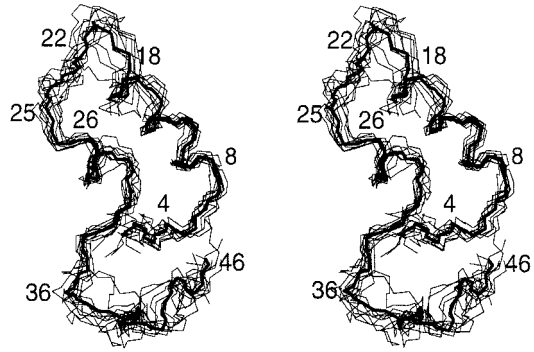

FIG. 4. Side by side stereo view of the backbone of the 10 final FANTOM structures with their energy minimized mean structure (thick line). The secondary structure of the mean is: 2–3 sheet, 7–12 helix, 13 turn, 14–17 helix, 26–29 helix, and 33–34 sheet. The backbone average RMSD values to the mean structure: all residues: 1.48 ± 0.33 Å; residues 2–19 and 25–46: 1.36 ± 0.30 Å; residues 20–25: 1.16 ± 0.19 Å; residues 2–21 (β sheet, α helix): 0.97 ± 0.36 Å; residues 25–46 (β sheet and C terminus): 1.07 ± 0.20 Å.

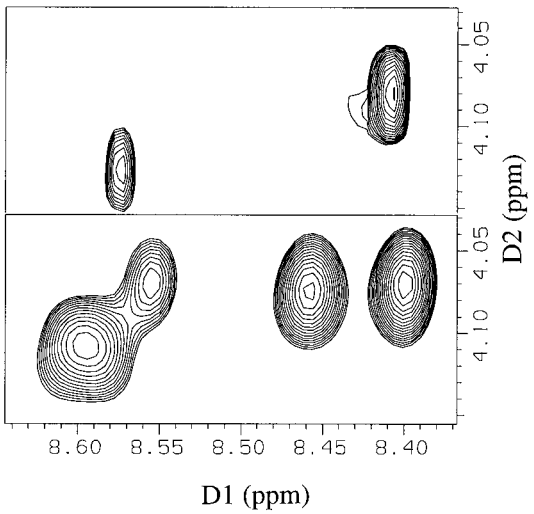

FIG. 5. Identification of the missing chemical shift CαH of residue Ser11. Left panel: a portion of the NOESY spectrum at 200 ms. Right panel: the TOCSY spectrum at the same region at 75 ms. The chemical shift CαH of Ser11 was missing in our first NOAH/DIAMOD calculation. The NOESY peaks at (8.40 ppm, 4.07 ppm) and at (8.56 ppm, 4.07 ppm) were then assigned by NOAH/DIAMOD to NH(11)-CαH(22) and to NH(12)-CαH(22), respectively. The TOCSY spectrum shows an intraresidue cross peak at (8.41 ppm, 4.07 ppm) near the NOESY cross peak. Therefore, we added the chemical shift of 4.07 ppm to CαH of Ser11 (see text).

example, we could identify the missing chemical shift of the $C_\alpha H$ proton of residue Ser11 from our initial NOAH/DIAMOD calculation. Including this chemical shift in the chemical shift list significantly improved the quality of the assignments. Two NOESY peaks at (8.56, 4.07) and (8.40, 4.07) ppm were initially assigned to proton pairs Asn12HN-Ser22C_αH and Ser11HN-Ser22C_αH by NOAH (Fig. 5). As the two peaks were strong, these distance constraints forced residue 22 to fold back to the first α helix, locally strongly distorting the structures. The TOCSY spectrum in this region showed an intraresidue cross peak 11HN-11C_αH at (8.41, 4.07) which overlapped within the tolerance range with one of the two NOESY cross peaks. When the $C_\alpha H$ proton of Ser11 was added to the proton list (note that the $C_\alpha H$ chemical shifts of Ser11 and Ser22 are the same), without making any manual assignments, the two peaks were automatically reassigned to the pairs 11HN-11C_αH and 11C_αH-12HN. Similar analysis yielded the chemical shifts of $C_\alpha H$ (Tyr29) and $C_\alpha H$ (Tyr44). These additional proton shifts greatly improved the structure calculation.

Comparing the NMR Structures of Crambin(S22/I25) and Crambin(P22/L25)

In the manual determination of the NMR structure of crambin(P22/L25) (*26*), 543 NOE constraints (248 intraresidue, 129 sequential, 75 medium-range, and 91 long-range constraints) were used during the structure refinements. We used 581 distance constraints for crambin(S22/I25) (426 unambiguous, 59 ambiguous, and 96 test assignments; 289 are intraresidue, 130 sequential, 72 medium range ($1 < R_{ij} < 6$), and 90 long-range distance constraints). The number of unambiguous assignments (with unique distance constraints) in our crambin(S22/I25) structure determination is less than in the previous study (*26*). However, some of the NOESY cross peaks (159) were only included as qualitative distance constraints in that study, because of difficulties in the interpretation of the build-up curve. We had 41 more intraresidue assignments, but most of these were redundant or had no structural information. On the other hand, we could only set lower and upper limits for 7 χ_1 angles and 29 ϕ angles, compared to 17 χ_1 angle and 31 ϕ angle limits for crambin(P22/L25).

Figure 6a and 6b show the distributions of unambiguous NOE constraints and the relative RMSD of individual residues to their mean in our study. The largest RMSD to their mean is in the segment from residue 19 to 24 (Fig. 6a) where few constraints (Fig. 6b) could be identified by NOAH. There was only one constraint for Pro19, which has the highest RMSD to the mean structure, because only two chemical shifts (C_δHs) were found (Table 1). The $C_\alpha H$ chemical shift of residue Asn14 is missing; the highest RMSD is in the first α helix. Consistent with previous results (*26*), high RMSDs among the bundle structures are the result of the low number of constraints. The RMSD of our structure bundle is larger than that for the crambin isoform in the reference, probably due to this lack of chemical shift data.

The superposition of our mean structure to the isoform NMR mean structure is shown in Fig. 7. The largest difference is from residue 19 to 28, where the two isoform residues are involved at positions 22 and 25 (Fig. 7). As in the NMR structures of crambin(P22/L25), the 10 NOAH/DIAMOD structures vary greatly in this segment.

Differences in this segment between the two structure bundles could be due to two causes. (1) There may be real differences of the residues at position 22 and 25, as there are large fluctuations among the two sets of structures at this regions. (2) As crambin(P22/L25) forms three CO—HN hydrogen bonds at T21-Q23, P22-L25, and P22-C26, while we found two hydrogen bonds are formed at Q23-I25 and A24-C26 in crambin(S22/I25), the latter may be more flexible than its isoform because of the Pro–Ser exchange at position 22.

Comparing the NMR and Crystal Structures of Crambin(S22/I25)

As mentioned in the introduction, the X-ray structure of crambin(S22/I25) was determined after completion of this work (*29*). While our NMR structure was determined at room temperature, the crystal structure was done at low T (150 K). The RMSD between our NMR and the X-ray structures of isoform S22/I25 shows some differences. The backbone RMSD between X-ray and NMR structures is 2.19 Å for the whole molecule, 1.2 Å for nonloop regions (2–18 and 26–40), and the largest (2.8 Å) at loop regions (19–25, 41–46). In contrast, the RMSDs among two isoforms and mixture form in crystal are very small (0.056 Å). The larger RMSDs between

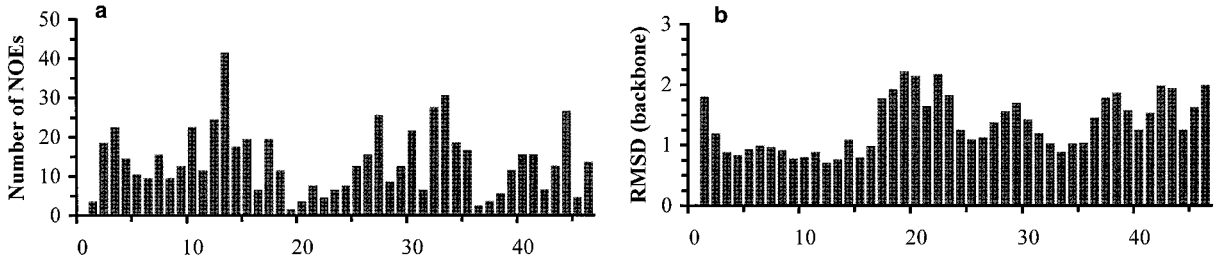

FIG. 6. (a) The number of unambiguous NOE constraints assigned by NOAH at the end of NOAH/DIAMOD cycles as function of the residue number. (b) The backbone RMSD of each residue to its mean structure.

the crystal and the NMR solution structures might be due to increased flexibility in the loop region 19–25 in the solution structures, as both ensembles of NMR structures show large deviations in this region.

DISCUSSION

This study demonstrates that the NOAH/DIAMOD/FANTOM suite can be used to generate structures automatically from previously uninterpreted, real NOE data with little manual interference. Although the overall quality of the structure bundle was better when a few manual assignments were used, the convergence was similar without them and 127 of the 157 manual assignments were made by NOAH/DIAMOD in case 2. Besides the 127 unique assignments that were rediscovered in case 2, 18 of the manually assigned peaks in case 1 were also assigned ambiguously (each peak has two or more assignments) in case 2, which *includes the manual assignments.* Only 12 manually assigned peaks in case 1 were assigned differently in case 2. However, by carefully checking the peak lists we found that although the same proton pair was assigned to two different peaks in case 1 and case 2, the two peaks are almost identical in peak intensities and chemical shifts at both D1 and D2 dimensions. These are either duplicated peaks or overlapped peaks picked by the FELIX program's automated peak picking routines. For instance, peak 124 was manually assigned to proton-pair $26H_{\alpha}$-26HN in case 1, but peak 124 was not assigned in case 2; instead, peak 123 was assigned to $26H_{\alpha}$-26HN. On the other hand, the two peaks are almost identical, i.e., peak 123 has chemical shifts of (4.665 ppm, 8.332 ppm) and peak 124 has chemical shifts of (4.674 ppm, 8.336 ppm). The intensities of peaks 123 and 124 are 1.07×10^7 and 1.06×10^7, respectively. As the chemical shift tolerance of the assignment was set to 0.02 ppm in both dimensions, the assignment made in case 2 by NOAH is the same as manually assigned in case 1.

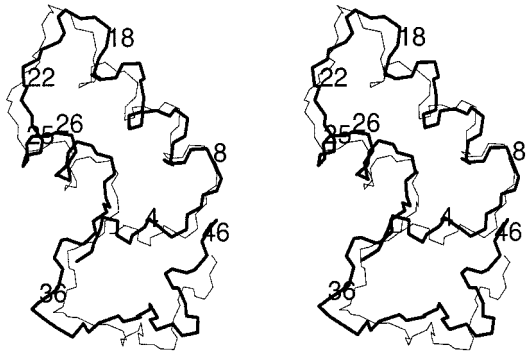

FIG. 7. Stereo view of the mean structure of crambin(Ser22/Ile25) (thin line) superimposed on the NMR structure of the isoform of crambin(Pro22/Leu25) (thick line). The backbone RMSD of the mean to the NMR isoform; all residues: 2.36 Å; residues 2–19, 25–46: 1.62 Å; residues 20–25: 1.43 Å; residues 2–18: 1.02 Å; residues 25–46: 1.36 Å.

Overall, NOAH/DIAMOD unambiguously rediscovered 139 manual assignments out of 157 in case 2. The rest of these manual assignments were found as one possible assignment in ambiguously assigned cross peaks. The structures calculated in the two cases are also quite similar, as the backbone RMSDs of mean structures from case 1 and 2 differ by only 1.3 Å at nonloop regions.

Ambiguous and test assignments, while necessary in the early stages of an automatic assignment method, can interfere with the convergence of the procedure and have to be carefully selected. The number of NOE peak assignments and the structure quality can be improved if initially only test assignments of low ambiguity are included, e.g., by choosing $N_{\mathrm{pa}} = 2$ in the initial cycles. Peaks with higher ambiguity can be tested at later stages in the iterative calculations, as an overwhelming effect on the structure bundle of a wrong constraint is less likely.

The procedure was also fairly robust, as 73 cross peaks with low intensity comparable to typical noise peaks (selected with manual peak picking and visual examinations of the NOESY spectrum at much deeper contour level) were not assigned or assigned ambiguously by NOAH.

We are now working on a new version of the NOAH program to include statistical proton chemical shift data for different types of residues in all NMR structures. This data may be used in NOAH/DIAMOD calculations when no chemical shift data is available (*7*). Manual inspection of the final structures can greatly aid in further structure refinement. Adding additional unambiguous manual distance constraints to those derived by NOAH as input for FANTOM will definitely improve the precision of our bundle structures after energy

minimization. Further refinement of the structures can be expected from relaxation matrix refinement methods.

CONCLUSIONS

The NOAH/DIAMOD/FANTOM suite can automatically calculate 3D protein structures with information routinely obtained from homonuclear NMR spectra for small and medium proteins. Our experience with this experimental data set for crambin can be summarized as follows:

Inclusion of 3–4 manual assignments per residue as in this study, while not necessary for obtaining convergence with the NOAH/DIAMOD suite, improves the quality of the structure bundle. The chemical shift tolerance for NOAH assignment should be within the range of 0.02–0.03 ppm. Peaks with a large number of possible assignments (i.e., with $N_{pa} > 2$) should not be included in initial NOAH/DIAMOD iterations. A nearly complete list of proton chemical shifts can significantly improve the reliability of the automated assignment and structural calculation.

ACKNOWLEDGMENTS

We thank Dr. C. H. Schein for critical reading of the manuscript, Kenneth D. Carlson of the U.S. Department of Agriculture for seeds of *C. abysinica*, and Peiling Wang and Bruce Luxon for initial protein purification and NMR studies. This work was supported by grants (to DG) from the NIH (AI27744) and the Lucille P. Markey Foundation, and (to WB) from the NSF (DBI-9632326) and the DOE (DE-FG03-96ER62267).

REFERENCES

1. P. Güntert, W. Braun, and K. Wüthrich, *J. Mol. Biol.* **217,** 517 (1991).
2. P. Güntert, K. D. Berndt, and K. Wüthrich, *J. Biomol. NMR* **3,** 601 (1993).
3. C. M. Oshiro and I. D. Kuntz, *Biopolymers* **33,** 107 (1993).
4. J. B. J. Olson and J. L. Markley, *J. Biomol. NMR* **4,** 385 (1994).
5. R. P. Meadows, E. T. Olejniczak, and S. W. Fesik, *J. Biomol. NMR* **4,** 79 (1994).
6. B. J. Hare and J. H. Prestegard, *J. Biomol. NMR* **4,** 35 (1994).
7. C. Antz, K. P. Neidig, and H. R. Kalbitzer, *J. Biomol. NMR* **5,** 287 (1995).
8. N. Morelle, B. Brutscher, J. P. Simorre, and D. Marion, *J. Biomol. NMR* **5,** 154 (1995).
9. C. Mumenthaler and W. Braun, *J. Mol. Biol.* **254,** 465 (1995).
10. C. Mumenthaler, P. Guntert, W. Braun, and K. Wüthrich, *J. Biomol. NMR* **10,** 351 (1997).
11. M. Nilges, *J. Mol. Biol.* **245,** 645 (1995).
12. M. Nilges, *Folding & Design* **2** (1997).
13. M. Nilges, *Curr. Opin. Str. Biol.* **6,** 617 (1996).
14. M. Nilges, M. J. Macias, S. I. O'Donoghue, and H. Oschkina, *J. Mol. Biol.* **269,** 408 (1997).
15. J. Pascual, M. Pfuhl, D. Walther, M. Saraste, and M. Nilges, *J. Mol. Biol.* **273,** 740 (1997).
16. D. Zimmermann, C. Kulikowski, L. Wang, B. Lyons, and G. T. Montelione, *J. Biomol. NMR* **4** (1994).
17. D. E. Zimmerman, C. A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G. T. Montelione, *J. Mol. Biol.* **269,** 592 (1997).
18. D. E. Zimmerman and G. T. Montelione, *Curr. Opin. Str. Biol.* **5,** 664 (1995).
19. W. Braun and N. Go, *J. Mol. Biol.* **186,** 611 (1985).
20. W. Braun, *Quart. Rev. Biophys.* **19,** 115 (1987).
21. G. Hänggi and W. Braun, *FEBS Letters* **344** (1994).
22. C. Mumenthaler and W. Braun, *Protein Sci.* **4,** 863 (1995).
23. S. Orru, A. Scaloni, M. Giannattasio, K. Urech, P. Pucci, and G. Schaller, *Biol. Chem.* **378,** 989 (1997).
24. L. Lobb, B. Stec, E. K. Kantrowitz, A. Yamano, V. Stojanoff, O. Markman, and M. M. Teeter, *Protein Engin.* **9,** 1233 (1996).
25. M. Teeter, S. Roe, and N. Heo, *J. Mol. Biol.* **230,** 292 (1993).
26. A. M. Bonvin, R. Boelens, and R. Kaptein, *Biopolymers* **34,** 39 (1994).
27. A. M. Bonvin, J. A. Rullmann, R. M. Lamerichs, R. Boelens, and R. Kaptein, *Proteins* **15,** 385 (1993).
28. A. Yamano and M. M. Teeter, *J. Biol. Chem.* **269,** 13956 (1994).
29. A. Yamano, N. H. Heo, and M. M. Teeter, *J. Biol. Chem.* **272,** 9597 (1997).
30. J. A. W. H. Vermeulin, R. M. J. N. Lamerichs, L. J. Berliner, A. De Marco, M. Linas, R. Boelens, J. Alleman, and R. Kaptein, *FEBS Lett.* **219,** 426 (1987).
31. K. Wüthrich, M. Billeter, and W. Braun, *J. Mol. Biol.* **169,** 949 (1983).
32. P. Güntert, W. Braun, M. Billeter, and K. Wüthrich, *J. Am. Chem. Soc.* **111,** 3997 (1989).
33. C. Spitzfaden, W. Braun, G. Wider, H. Widmer, and K. Wüthrich, *J. Biomol. NMR* **4,** 463 (1994).
34. T. Schaumann, W. Braun, and K. Wüthrich, *Biopolymers* **29,** 679 (1990).
35. B. V. Freyberg and W. Braun, *J. Comp. Chem.* **12,** 1065 (1991).