J|A|C|S
ARTICLES
Published on Web 09/19/2006

Automated Protein Structure Determination from NMR Spectra

Blanca López-Méndez† and Peter Güntert*

Contribution from the Tatsuo Miyazawa Memorial Program, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

Received February 23, 2006; E-mail: guentert@gsc.riken.jp

Abstract: Fully automated structure determination of proteins in solution (FLYA) yields, without human intervention, three-dimensional protein structures starting from a set of multidimensional NMR spectra. Integrating existing and new software, automated peak picking over all spectra is followed by peak list filtering, the generation of an ensemble of initial chemical shift assignments, the determination of consensus chemical shift assignments for all ^{1}H, ^{13}C, and ^{15}N nuclei, the assignment of NOESY cross-peaks, the generation of distance restraints, and the calculation of the three-dimensional structure by torsion angle dynamics. The resulting, preliminary structure serves as additional input to the second stage of the procedure, in which a new ensemble of chemical shift assignments and a refined structure are calculated. The three-dimensional structures of three 12−16 kDa proteins computed with the FLYA algorithm coincided closely with the conventionally determined structures. Deviations were below 0.95 Å for the backbone atom positions, excluding the flexible chain termini. 96−97% of all backbone and side-chain chemical shifts in the structured regions were assigned to the correct residues. The purely computational FLYA method is suitable for substituting all manual spectra analysis and thus overcomes a main efficiency limitation of the NMR method for protein structure determination.

Introduction

NMR spectroscopy has by now been used to determine the three-dimensional (3D) structures in solution of more than 5000 proteins in the Protein Data Bank (PDB).[1] However, limitations of efficiency, molecular size, and objectivity continue to curb its potential for protein structure analysis. The laborious and often difficult interpretation of the NMR spectra requires (too) much human time and expertise. Structures of proteins above 30 kDa are difficult to solve by NMR. The spectrum interpretation relies on subjective human decisions, which impedes exact reproducibility of NMR protein structures. Furthermore, a truly objective measure for the agreement between a protein structure and the raw NMR data that does not rely on such subjective intermediate results as peak lists or chemical shift assignments is still lacking.

The NMR structure determination of a protein commonly involves the preparation of uniformly ^{13}C/^{15}N-labeled, soluble protein, the acquisition of a set of 2D and 3D NMR experiments, NMR data processing, peak picking, chemical shift assignment, NOE assignment and collection of conformational restraints, structure calculation, refinement, and validation.[2] A variety of computational approaches have been introduced either to support the interactive analysis by visualization and book-keeping or to provide automation for specific parts of an NMR structure determination.[3−6] A recent review[5] documents close to 100 such algorithms and programs. Automated procedures are now widely accepted for the assignment of NOE distance restraints and the structure calculations.[5−10] The automation of the preceding steps of peak picking and resonance assignment has also been the subject of intensive research[3−5] although manual or semiautomated approaches still prevail, especially for the assignment of the sidechain chemical shifts. Most of the automated approaches concentrate on the assignment of the backbone and C^{β} resonances on the basis of a specific set of triple resonance experiments by either exhaustive, heuristic, or database searches or by using Monte Carlo or simulated annealing methods.[5] Also nonclassical approaches that do not rely on sequence-specific resonance assignments[11−13] and methods using residual dipolar couplings to determine the backbone structure without the need for side-chain assignments[14−16] have been proposed.

† Current address: Centro de Investigaciones Biológicas CIB-CSIC, C/Ramiro de Maeztu 9, 28040 Madrid, Spain.

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(2) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986.
(3) Baran, M. C.; Huang, Y. J.; Moseley, H. N. B.; Montelione, G. T. *Chem. Rev.* **2004**, *104*, 3541−3555.
(4) Altieri, A. S.; Byrd, R. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 547−553.
(5) Gronwald, W.; Kalbitzer, H. R. *Prog. NMR Spectrosc.* **2004**, *44*, 33−96.
(6) Güntert, P. *Prog. NMR Spectrosc.* **2003**, *43*, 105−125.
(7) Nilges, M.; Macias, M. J.; O'Donoghue, S. I.; Oschkinat, H. *J. Mol. Biol.* **1997**, *269*, 408−422.
(8) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319*, 209−227.
(9) Huang, Y. J.; Moseley, H. N. B.; Baran, M. C.; Arrowsmith, C.; Powers, R.; Tejero, R.; Szyperski, T.; Montelione, G. T. *Methods Enzymol.* **2005**, *394*, 111−141.
(10) Kuszewski, J.; Schwieters, C. D.; Garrett, D. S.; Byrd, R. A.; Tjandra, N.; Clore, G. M. *J. Am. Chem. Soc.* **2004**, *126*, 6258−6273.
(11) Kraulis, P. J. *J. Mol. Biol.* **1994**, *243*, 696−718.
(12) Atkinson, R. A.; Saudek, V. *FEBS Lett.* **2002**, *510*, 1−4.
(13) Grishaev, A.; Llinás, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6707−6712.
(14) Meiler, J.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 15404−15409.

10.1021/ja061136l CCC: $33.50

Complete automation of the NMR structure determination process, however, has not yet been realized in practice. Fully automated NMR structure determination is more demanding than automating one part of NMR structure determination because the cumulative effect of imperfections at successive steps can easily render the overall process unsuccessful. For example, it has been demonstrated recently that reliable automated NOE assignment and structure calculation require around 90% completeness of the chemical shift assignment,[8,17] which is not straightforward to achieve by unattended automated peak picking and automated resonance assignments. Present systems designed to handle the whole process therefore generally require certain human interventions.[5,9] The interactive validation of peaks and assignments, however, still constitutes a time-consuming obstacle for high-throughput NMR protein structure determination.

The crucial indicator for a fully automated NMR structure determination method is the accuracy of the resulting 3D structures when real experimental input data is used and any human interventions at intermediate steps are avoided. Even "small" manual corrections, or the use of idealized input data, can lead to substantially altered conclusions and prejudice the assessment of different methods.

Here we present a fully automated computational approach that is capable of solving 3D protein structures using as experimental input data only the amino acid sequence and a set of multidimensional NMR spectra. In analogy to automated crystallographic systems for high-throughput macromolecular structure determination,[18−20] we have achieved complete automation of NMR protein structure determination by combining and extending software packages that carry out individual steps in NMR structure determination into an integrated system.

Algorithm

Fully Automated NMR Structure Determination Algorithm (FLYA). Any combination of the commonly used hetero- and homonuclear two-, three-, and four-dimensional NMR spectra can be used as input for the FLYA algorithm, provided that it affords sufficient information (see Reliability Indicators below) for the assignment of the backbone and side-chain chemical shifts and for the collection of conformational restraints. Seven purely computational steps (Figure 1) are applied in three successive stages. Stage I comprises steps 1−6. Stage II (steps 3−6) and stage III (steps 3−7) differ from the initial stage I only in that the bundle of conformers obtained at the end of the preceding stage is used as additional input for the ensemble chemical shift assignment. The complete procedure runs without human intervention, driven by the NMR structure calculation program CYANA.[6,21] In addition to the 3D structure of the protein, FLYA yields backbone and side-chain chemical shift assignments and cross-peak assignments for all spectra.

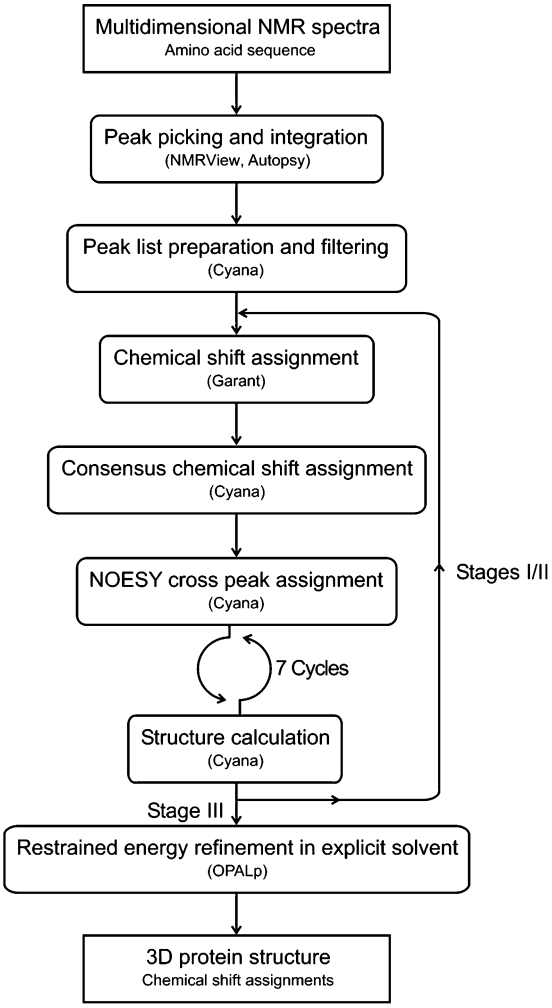

Figure 1. Flowchart of the fully automated structure determination algorithm, FLYA.

Step 1: Automated Peak Picking and Peak Integration. The present version of the FLYA algorithm identifies the frequency position of signals in the multidimensional spectra using the automated algorithm in the program NMRView,[22,23] or optionally the program AUTOPSY.[24] Peak picking is performed over all spectra using automated scripts. The only important parameter for the peak picking algorithm[23] is the intensity threshold for peak identification, which can be set relative to the noise level of each spectrum. Peak integrals for NOESY cross-peaks are determined simultaneously. Since no manual corrections are applied, the resulting raw peak lists may contain, in addition to the entries representing true signals, a significant number of artifacts. The following steps of the fully automated structure determination algorithm can tolerate the presence of such artifacts, as long as the majority of the true peaks have been identified.

Step 2: Peak List Preparation and Filtering. Peak lists that conform to the requirements of the subsequent chemical shift assignment and NOESY assignment steps are prepared by the program CYANA from the raw peak lists of step 1. This may include combining data from multiple spectra, e.g., if data for aromatic and other carbons have been recorded separately, the unfolding of aliased peaks, possible systematic corrections

(15) Jung, Y. S.; Sharma, M.; Zweckstetter, M. *Angew. Chem., Int. Ed.* **2004**, *43*, 3479−3481.
(16) Prestegard, J. H.; Mayer, K. L.; Valafar, H.; Benison, G. C. *Methods Enzymol.* **2005**, *394*, 175−209.
(17) Jee, J.; Güntert, P. *J. Struct. Funct. Genomics* **2003**, *4*, 179−189.
(18) Terwilliger, T. C.; Berendzen, J. *Acta Crystallogr.* **1999**, *D55*, 849−861.
(19) Brunzelle, J. S.; Shafaee, P.; Yang, X.; Weigand, S.; Ren, Z.; Anderson, W. F. *Acta Crystallogr.* **2003**, *D59*, 1138−1144.
(20) Ness, S. R.; de Graaff, R. A. G.; Abrahams, J. P.; Pannu, N. S. *Structure* **2004**, *12*, 1753−1761.
(21) Güntert, P.; Mumenthaler, C.; Wüthrich, K. *J. Mol. Biol.* **1997**, *273*, 283−298.
(22) Johnson, B. A.; Blevins, R. A. *J. Biomol. NMR* **1994**, *4*, 603−614.
(23) Johnson, B. A. *Methods Mol. Biol.* **2002**, *278*, 313−352.
(24) Koradi, R.; Billeter, M.; Engeli, M.; Güntert, P.; Wüthrich, K. *J. Magn. Reson.* **1998**, *135*, 288−297.

of chemical shift referencing, the removal of peaks near the diagonal or water line, conversion to XEASY[25] peak list format, which is used in subsequent steps, and consistent naming of the peak lists according to the experiment type. In the present version of the algorithm, the peak lists resulting from this step remain invariant throughout the rest of the procedure.

Step 3: Ensemble Chemical Shift Assignment. In analogy to NMR structure calculation in which not a single structure but an ensemble of conformers is calculated using identical input data except for different randomized start conformers,[2] the initial chemical shift assignment produces an ensemble rather than a single chemical shift value for each ^{1}H, ^{13}C, and ^{15}N nucleus. The ensemble of chemical shift assignments is obtained from 20 independent runs[26] of the GARANT algorithm,[27,28] each starting from the same peak lists but using a different random number generator seed value for the memetic algorithm that optimizes the matching between the peaks expected from combined knowledge of the primary structure and the magnetization transfer pathways in the spectra used and the peaks observed experimentally. The algorithm combines an evolutionary algorithm with a local optimization routine and uses a scoring scheme to distinguish between correct and incorrect resonance assignments. The score[27] captures the essential features of a correct resonance assignment, i.e., the presence of expected peaks, the positional alignment of peaks that originate from the same atoms, and the normality of the assigned resonance frequencies with respect to a chemical shift database that was compiled from the known resonance assignments of many proteins.[29] A main advantage of the GARANT algorithm is its ability to analyze the peak lists from all available spectra simultaneously. The input peak lists and additional information on their type and formatting required by GARANT are prepared automatically within the program CYANA, which is also used for the parallelization of the ensemble chemical shift assignment calculations. Tolerances of 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N chemical shifts were used in all FLYA calculations of this paper for the matching of peaks with identical assignments. The original GARANT algorithm[27] was extended for additional types of NMR experiments and by a new treatment of NOESY spectra that makes quantitative use of peak intensities and preliminary 3D structures in stages II and III of the FLYA algorithm.

Step 4: Consensus Chemical Shift Assignment. The most highly populated chemical shift value in the ensemble is computed for each ^{1}H, ^{13}C, and ^{15}N spin by CYANA and selected as the consensus chemical shift value that will be used for the subsequent automated assignment of NOESY peaks. The consensus chemical shift for a given nucleus is the value ω that maximizes the function $\mu(\omega) = \Sigma_j \exp(-(\omega - \omega_j)^2/2\Delta\omega^2)$, where the sum runs over all chemical shift values ω_j for the given nucleus in the ensemble of raw chemical shift assignments from step 3, and $\Delta\omega$ denotes the chemical shift tolerance, which has the same values as those in step 3. The absence of stereospecific assignments for diastereotopic groups is taken into account when computing the consensus chemical shift values by calculating the two consensus chemical shifts ω and ω' of a diastereotopic pair from the values in the two sets $\{\min(\omega_j, \omega_j')\}$ and $\{\max(\omega_j,\omega_j')\}$, where j runs over all pairs of chemical shift values for the given diastereotopic pair of nuclei in the ensemble of raw chemical shift assignments.

Step 5: NOESY Cross-Peak Assignment. NOESY cross-peaks are assigned automatically[8] on the basis of the consensus chemical shift assignments and the same peak lists and chemical shift tolerance values used already for the chemical shift assignment in step 3. The automated NOE assignment algorithm of the program CYANA, version 2.1, is used. The overall probability for the correctness of possible NOE assignments is calculated as the product of three probabilities that reflect the agreement between the chemical shift values and the peak position, the consistency with a preliminary 3D structure,[30] and network-anchoring,[8] i.e., the extent of embedding in the network formed by other NOEs. Restraints with multiple possible assignments are represented by ambiguous distance restraints.[31]

Step 6: Structure Calculation. The structure calculation is performed using the standard protocol of the program CYANA 2.1, which is based on simulated annealing driven by molecular dynamics simulation in torsion angle space.[21] Seven cycles of combined automated NOESY assignment (step 5) and structure calculation are followed by a final structure calculation. Constraint combination[8] is applied in the first two cycles to all NOE distance restraints spanning at least three residues in order to minimize distortions of the structures by erroneous distance restraints that may result from spurious entries in the peak lists and/or incorrect chemical shift assignments.

Step 7: Restrained Energy Refinement in Explicit Solvent. The 20 final CYANA conformers with the lowest target function values obtained in stage III are subjected to restrained energy minimization in explicit solvent against the AMBER force field,[32] using the program OPALp.[33,34]

Reliability Indicators. The performance of the FLYA algorithm is monitored at different steps of the procedure by quality measures that can be computed without referring to external reference assignments or structures.

1. Peak picking extent: The extent to which the peak picking yielded the expected number of cross-peaks, measured by the average over all spectra of the percentage of the number of observed cross-peaks relative to the number of expected cross-peaks, is evaluated in Step 2. In the case of NOESY spectra, the number of expected cross-peaks is estimated to be twice the number of short-range NOEs expected from the amino acid sequence.[27]

2. Peak assignment completeness: The average over all spectra of the percentage of assigned observed cross-peaks relative to the number of expected cross-peaks is evaluated in Step 3.

(25) Bartels, C.; Xia, T.; Billeter, M.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **1995**, *6*, 1−10.
(26) Malmodin, D.; Papavoine, C. H. M.; Billeter, M. *J. Biomol. NMR* **2003**, *27*, 69−79.
(27) Bartels, C.; Güntert, P.; Billeter, M.; Wüthrich, K. *J. Comput. Chem.* **1997**, *43*, 139−149.
(28) Bartels, C.; Billeter, M.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **1996**, *7*, 207−213.
(29) Seavey, B. R.; Farr, E. A.; Westler, W. M.; Markley, J. L. *J. Biomol. NMR* **1991**, *1*, 217−236.
(30) Güntert, P.; Berndt, K. D.; Wüthrich, K. *J. Biomol. NMR* **1993**, *3*, 601−606.
(31) Nilges, M. *J. Mol. Biol.* **1995**, *245*, 645−660.
(32) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(33) Koradi, R.; Billeter, M.; Güntert, P. *Comput. Phys. Commun.* **2000**, *124*, 139−147.
(34) Luginbühl, P.; Güntert, P.; Billeter, M.; Wüthrich, K. *J. Biomol. NMR* **1996**, *8*, 136−146.

3. Chemical shift assignment redundancy: The average over all assigned nuclei of the number of non-NOESY cross-peaks assigned to a given nucleus is evaluated in Step 3.

4. Chemical shift ensemble self-consistency: The percentage of assigned nuclei for which more than 80% of the chemical shift values in the ensemble agree within the given chemical shift tolerances with the consensus value is evaluated in Step 4.

5. Long-range distance restraints per residue: The density of the network of tertiary structure defining NOEs, measured by the average number of long-range (spanning at least five residues) distance restraints per residue, is evaluated in Step 5.

6. Initial fold precision: The precision of the initial global fold, measured by the RMSD value relative to the average coordinates for the backbone atoms N, C^{α}, and C′ in the structured regions of the conformers obtained in the first cycle of combined automated NOE assignment and structure calculation with CYANA,[8] is evaluated in Step 6.

7. Packing quality: The average intraprotein Lennard-Jones energy per residue according to the AMBER force field[32] is evaluated in Step 7.

Materials and Methods

Proteins. The FLYA algorithm was applied for the NMR structure determination of three proteins: the ENTH-VHS domain At3g16270-(9−135) from *Arabidopsis thaliana* (ENTH),[35] the rhodanese homology domain At4g01050(175−295) from *Arabidopsis thaliana* (RHO),[36,37] and the Src homology domain 2 from the human feline sarcoma oncogene Fes (SH2).[38,39] High-quality NMR solution structures of these three proteins have been solved recently using conventional techniques, i.e., manual assignment of the chemical shifts, manual NOESY peak picking, and combined NOE assignment and structure calculation with CYANA followed by restrained energy minimization in explicit solvent.

The ENTH construct that had been used for the NMR measurements is a protein of 140 amino acid residues comprising the 127 residues of the natural sequence of At3g16270(9−135) that form a seven-α-helix superhelical fold and 13 non-native terminal residues.[35] RHO is a 134-residue α-*β* rhodanese domain with a central five-stranded parallel *β*-sheet flanked by four α-helices and two small 3_{10}-helices.[37] SH2 comprises 114 residues and has the canonical Src homology 2 domain fold with a central three-stranded antiparallel *β*-sheet flanked on either side by an α-helix and three short antiparallel *β*-strands that pack against the second α-helix.[39] The 7 N-terminal and 6 C-terminal residues of all three proteins are non-native flanking regions related to the expression and purification system that are flexibly disordered in solution.[35,37,39] The structured regions of the proteins consist of the residues 11−130 for ENTH, 7−125 for RHO, and 8−108 for SH2. The coordinates of the 3D structures and the chemical shift assignments determined earlier by conventional methods are available from the Protein Data Bank with accession codes 1VDY for ENTH, 1VEE for RHO, and 1WQU for SH2 and from the BioMagResBank with accession numbers 5928 for ENTH, 5929 for RHO, and 6331 for SH2, respectively. Results obtained with FLYA were assessed against these reference structures and reference assignments.

NMR Spectroscopy. The raw data from the NMR measurements[35,36,38] that had been performed for the previous, conventional structure determinations were again used for the fully automated approach. NMR experiments had been collected at 25 °C on Bruker DRX 600 spectrometers operating at a proton frequency of 600 MHz, except for the 3D NOESY experiments that had been recorded on Bruker AV 800 spectrometers operating at a proton frequency of 800 MHz. Table 1 shows a summary of the spectra that had been acquired for the three proteins. For each protein the NMR measurements had been carried out with a single, uniformly ^{13}C- and ^{15}N-labeled sample, containing 1.1−1.2 mM protein dissolved in 90% H_2O/10% D_2O (v/v), 20 mM Tris buffer, pH 7.5, for ENTH, 20 mM phosphate buffer, pH 6.0, for RHO, or 20 mM Tris-HCl buffer, pH 7.0, for SH2. In addition, all samples contained 100 mM NaCl, 1 mM dithiothreitol, and 0.02% NaN_3.[35,36,38]

NMR Spectra Processing. The NMRPipe[40] software was used to retransform the original time-domain NMR data into frequency-domain spectra. At least 2-fold zero-filling in each spectral dimension, linear prediction in one indirect dimension, apodization by 60°−90° shifted sine-squared window functions, and baseline correction along all dimensions were applied. Whenever possible, the same parameters were used for corresponding dimensions of all experiments for a given protein (Table 1). Particular attention was paid to accurate and consistent chemical shift referencing in the direct and indirect dimensions in order to enable the use of small tolerances for the matching of peak positions from different spectra during the fully automated analysis. Spectral files in the formats of the programs NMRView[22] and XEASY[25] were produced.

FLYA Software. FLYA calculations were run by the program CYANA,[6,21] using the independent programs NMRView[23] (www.onemoonscientific.com) or AUTOPSY[24] and GARANT,[27,28] as additional plug-ins. These softwares are available (see www.tmmp.gsc.riken.jp for details). CYANA and the plug-in algorithms refer as much as possible to the same databases on amino acid geometry,[41] IUPAC nomenclature,[42] and other properties. Internal data conversions are performed automatically by CYANA, as needed.

Automated Peak Picking and Peak Integration. Peak picking and NOESY peak integration over all spectra were performed by the automated algorithms of NMRView[23] or AUTOPSY.[24] In the generic version of the FLYA algorithm, peak identification and NOESY peak integration were performed using an automated Tcl/Tk script for NMRView. Intensity thresholds for peak identification were set according to the noise level for each spectrum. In all cases, the complete spectrum was used. No spectral regions or individual peaks were interactively excluded from peak picking.

In addition, alternative peak picking methods were evaluated for ENTH. Automated peak picking of the entire spectra was performed also with the program AUTOPSY.[24] AUTOPSY first segmented the spectra, given in the format of the program XEASY,[25] into connected regions with a minimal size of two data points in all dimensions and intensity at least 2.7 or 1.8 times the local noise level in 2D or 3D spectra, respectively. Peaks with an intensity at least 4.0 times above the local noise level were identified in the 2D spectra. In the 3D spectra, peaks that were at least 2.5 times above the local noise level were identified first, and their line shapes subsequently were used to find further peaks at least 2.0 times above the local noise level. In line with earlier applications of AUTOPSY,[26] the results were found to be insensitive to moderate variations of the peak picking parameter values.

As a reference, peaks were also picked by interactive visual inspection of all ENTH spectra using NMRView for spectral display, bookkeeping, and NOESY peak integration. To avoid a possible bias

(35) López-Méndez, B., et al. *J. Biomol. NMR* **2004**, *29*, 205−206.
(36) Pantoja-Uceda, D., et al. *J. Biomol. NMR* **2004**, *29*, 207−208.
(37) Pantoja-Uceda, D.; López-Méndez, B.; Koshiba, S.; Inoue, M.; Kigawa, T.; Terada, T.; Shirouzu, M.; Tanaka, A.; Seki, M.; Shinozaki, K.; Yokoyama, S.; Güntert, P. *Protein Sci.* **2005**, *14*, 224−230.
(38) Scott, A.; Pantoja-Uceda, D.; Koshiba, S.; Inoue, M.; Kigawa, T.; Terada, T.; Shirouzu, M.; Tanaka, A.; Sugano, S.; Yokoyama, S.; Güntert, P. *J. Biomol. NMR* **2004**, *30*, 463−464.
(39) Scott, A.; Pantoja-Uceda, D.; Koshiba, S.; Inoue, M.; Kigawa, T.; Terada, T.; Shirouzu, M.; Tanaka, A.; Sugano, S.; Yokoyama, S.; Güntert, P. *J. Biomol. NMR* **2005**, *31*, 357−361.
(40) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277−293.
(41) Engh, R. A.; Huber R. *Acta Crystallogr. A* **1991**, *47*, 392−400.
(42) Markley, J. L.; Bax, A.; Arata, Y.; Hilbers, C. W.; Kaptein, R.; Sykes, B. D.; Wright, P. E.; Wüthrich, K. *J. Mol. Biol.* **1998**, *280*, 933−952.

Table 1. Multidimensional NMR Spectra for ENTH, RHO, and SH2[a]

	ENTH			RHO			SH2		
spectrum	points	widths (kHz)	peaks	points	widths (kHz)	peaks	points	widths (kHz)	peaks
				2D:					
^{15}N-HSQC	512 × 128	11.2, 1.8	149	512 × 64	11.2, 2.7	148	512 × 46	11.2, 2.7	123
^{13}C-HSQC[b]	512 × 128	11.2, 8.7	667	512 × 128	11.2, 15.1	616	512 × 40	11.2, 7.9	544
				512 × 64	5.4, 4.8				
				3D for Backbone Assignment:					
HNCO	27 × 70	8.4, 1.4, 3.3	148	46 × 50	8.4, 2.0, 3.3	146	46 × 50	8.4, 2.0, 3.3	121
HN(CA)CO	27 × 70	8.4, 1.4, 3.3	282	46 × 50	8.4, 2.0, 3.3	272	46 × 50	8.4, 2.0, 3.3	226
HNCA	29 × 70	8.4, 1.4, 4.8	283	46 × 50	8.4, 2.0, 4.8	274	46 × 50	8.4, 2.0, 4.8	228
HN(CO)CA	27 × 70	8.4, 1.4, 4.8	148	46 × 50	8.4, 2.0, 4.8	146	46 × 50	8.4, 2.0, 4.8	121
CBCANH	32 × 75	8.4, 1.4, 11.3	548	46 × 64	8.4, 2.0, 11.3	518	46 × 64	8.4, 2.0, 11.3	433
CBCA(CO)NH	32 × 70	8.4, 1.4, 11.3	288	46 × 64	8.4, 2.0, 11.3	279	46 × 64	8.4, 2.0, 11.3	232
				3D for Side-Chain Assignment:					
HBHA(CO)NH	26 × 60	8.4, 1.4, 6.8	411	46 × 64	8.4, 2.0, 7.5	401	46 × 64	8.4, 2.0, 8.4	350
(H)CC(CO)NH	24 × 60	8.4, 1.4, 11.3	451	46 × 64	8.4, 2.0, 11.3	423	46 × 64	8.4, 2.0, 11.3	370
H(CCCO)NH	27 × 77	8.4, 1.4, 6.8	664	46 × 64	8.4, 2.0, 7.5	619	46 × 64	8.4, 2.0, 6.7	540
HCCH-COSY[b]	32 × 85	7.8, 6.5, 6.8	1756				50 × 100	8.4, 11.3, 8.4	1352
	17 × 85	5.2, 4.0, 5.2		16 × 80	5.4, 3.9, 5.4	40	16 × 80	6.1, 4.6, 6.1	
(H)CCH-TOCSY	44 × 80	8.4, 6.5, 13.9	1636						
HCCH-TOCSY	32 × 120	7.8, 6.5, 6.8	2812	64 × 100	8.4, 11.3, 8.4	2644	64 × 100	8.4, 11.3, 8.4	2144
				3D NOESY for Assignment and Restraint Collection:					
^{15}N-edited NOESY	36 × 128	11.2, 1.8, 10.1	1624	46 × 128	11.2, 2.7, 11.2	1597	46 × 128	11.2, 2.7, 11.2	1340
^{13}C-edited NOESY[b]	46 × 150	11.2, 8.7, 8.8	6590	34 × 116	11.2, 7.7, 11.2	6205	40 × 150	11.2, 8.0, 11.2	5672
				32 × 128	11.2, 5.1, 11.2				

[a] Points: complex time domain data points in the indirect dimensions. The number for the first indirect dimension refers to ^{15}N, if present, or ^{13}C otherwise. The second number refers to ^{1}H, if present, or ^{13}C otherwise. In all 3D spectra, 512 complex time domain data points were recorded in the directly detected ^{1}H dimension. Widths: spectral widths in the directly detected dimension and in the indirectly detected dimension(s). Peaks: number of cross-peaks expected under ideal conditions, based on the knowledge of the magnetization transfer pathways for each experiment. In the case of NOESY spectra the expected peaks correspond to ^{1}H−^{1}H distances shorter than 4.5 Å in the reference structures. [b] The two sets of values refer to the two spectra recorded separately for the aliphatic and aromatic carbon region, respectively.

that could arise from using manually picked peak lists that had been subsequently refined in the light of known assignments, we did not use the peak lists that had resulted from the conventional structure determination but prepared a new set of peak lists by visual inspection of each individual spectrum.

Automated Chemical Shift Assignment. The input data files for GARANT[27,28] consisted of the protein sequence and lists of (unassigned) peaks for the spectra listed in Table 1. GARANT was run 20 times with different random number generator seed values to produce an ensemble of 20 raw chemical shift assignments for the complete proteins. Using CYANA for parallelization, the GARANT calculations were performed simultaneously on 20 processors of a Linux cluster system with Intel Xeon 3.06 GHz processors and 2 gigabytes of memory per two-processor node. The standard optimization macro that combines a genetic algorithm with local optimization[27] was used. The population size for one generation of resonance assignments in the genetic algorithm was 100. The peak position tolerance was set to 0.03 ppm for the ^{1}H dimensions and to 0.4 ppm for the ^{13}C and ^{15}N dimensions.

At stage I, in the absence of an input 3D structure, only the commonly expected intraresidual and sequential NOEs[27] are taken into account. In stages II and III the expected NOESY cross-peaks are generated for all ^{1}H−^{1}H pairs for which the corresponding minimal distance $d_{\min}$ within the conformers of a structure bundle is shorter than 5.5 Å. In the case of methyl groups, nonstereoassigned pairs, or equivalent aromatic ring protons, the appropriate pseudo atom corrections are applied. In this way, the tertiary structure and its expected long-range NOEs are introduced into the ensemble chemical shift assignment. The initial probability for actually observing an expected NOE is set to $P_{\mathrm{NOE}}^{(0)} = 1.0$ if the corresponding distance is shorter than 4 Å in all conformers, to $P_{\mathrm{NOE}}^{(0)} = 0.2$ if it is shorter than 5.5 Å in the majority of the conformers, and to $P_{\mathrm{NOE}}^{(0)} = 0.1$ otherwise, regardless of the peak volume. The overall probability for an expected NOESY cross-peak is then computed as $P_{\mathrm{NOE}} = P_{\mathrm{NOE}}^{(0)} \exp(-(\max(d_{\min} - u,0)/\sigma)^2/2)$, with $\sigma = 0.5$ Å.

The upper distance bound u is computed for each NOESY cross-peak using the automated NOE calibration method implemented in CYANA, which employs a $V = Au^{-6}$ relationship between the peak volume, V, and the corresponding upper distance bound, u. To avoid a possible bias from the presence of a significant number of artifact peaks with a generally small volume that may have resulted from the automated peak picking, the constant A is calculated in two steps. (Other artifacts associated with intense solvent and diagonal peaks have already been removed in the preceding step 2.) First, A is initialized such that the median of all peak volumes in a given NOESY spectrum corresponds to a distance of 4.0 Å. Second, A is recalculated in the same way but including into the median calculation only the peaks for which the initial value of A yielded an upper distance bound of 5.5 Å or less.

The output of the ensemble chemical shift assignment comprised 20 chemical shift lists, which are used in the subsequent step of consensus chemical shift determination. For each spectrum, 20 peak lists with peak assignments are produced as informational output that is not used further within the FLYA algorithm.

Consensus Chemical Shift Assignment. The correctness of the consensus chemical shifts was assessed by comparison with the chemical shifts obtained previously by interactive methods.[35,36,38] A consensus chemical shift assignment was considered to be in agreement with the interactively determined reference if the two corresponding chemical shift values differed by less than 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C or ^{15}N. Such assignments were classified as *equal*. Chemical shifts that deviated by more than these tolerances from the interactively assigned value were classified as *different*. These chemical shifts are further classified as *wrong* if their value does not match the interactively assigned chemical shift value of any atom within the same residue. Only this latter type of assignment error can potentially lead to a serious distortion of the resulting structure. Note that the percentages of equal and different peaks do not necessarily sum up to 100% because only the nuclei that were assigned by both methods can be classified in this way.

Automated NOE Assignment, Structure Calculation, and Refinement. Combined automated NOE assignment and the structure calculation were performed by the standard procedures of CYANA 2.1, using

as input the consensus chemical shifts from the preceding step and the NOESY peak lists from automated peak picking. Assignments with an overall probability below 10% in cycle 1 or 20% in cycles 2−7 are discarded. In cycles 2−7, the probability for agreement of a given restraint with the preliminary structure from the preceding cycle is computed as the fraction of all conformers in which the restraint is violated by less than a violation cutoff of 1.5, 0.9, 0.6, 0.3, 0.1, and 0.1 Å in cycles 2−7, respectively. To automatically account for slight errors in peak integrals or intensities, the upper distance bound of a restraint that is consistently but only slightly violated is increased in four steps up to maximally 1.25 times the original bound until the resulting, more conservative upper distance limit can be fulfilled by 80% or more of the conformers from the previous cycle. This procedure is applied in the calculation cycles 3−7. The structure calculation is started in each cycle from 100 conformers with random torsion angle values. The 20 conformers with the lowest final CYANA target function values are retained for analysis and passed to the next cycle. The covalent parameters of Engh and Huber[41] are used in conjunction with slightly larger repulsive core radii than those in previous versions of DYANA and CYANA,[43] namely 0.95 Å for amide hydrogens, 1 Å for other hydrogens, 1.5 Å for carbonyl carbons, 1.6 Å for other carbons, 1.45 Å for amide nitrogens, 1.5 Å for other nitrogens, 1.3 Å for oxygens, and 1.8 Å for sulfurs. All repulsive core radii were reduced by 0.1 Å for interactions between atoms separated by three covalent bonds. Restraints that involve degenerate groups of protons, e.g., methyls, accidentally degenerate methylenes, and equivalent aromatic ring protons, are expanded into ambiguous distance restraints between all corresponding pairs of hydrogen atoms.[31] Nondegenerate diastereotopic pairs are periodically swapped for a minimal target function value during simulated annealing in cycles 1−7. Weak restraints on ϕ/ψ torsion angle pairs and on side-chain torsion angles between tetrahedral carbon atoms are applied temporarily during the high-temperature and cooling phases of the simulated annealing schedule in order to favor allowed regions of the Ramachandran plot and staggered rotamer positions, respectively.

To enable easy access to the FLYA results by other structure refinement and validation programs that do not handle ambiguous distance restraints, only unambiguously assigned upper distance bounds are used in the final structure calculation. This is achieved by splitting the volume of the peaks that have multiple assignments in cycle 7 according to the distances observed in the structure bundle from cycle 7 into separate peaks with unique assignment, and by applying pseudo atom corrections and symmetrization to account for the absence of stereospecific assignments.[43]

CYANA was also used to parallelize the restrained energy refinement with OPALp of the 20 final CYANA conformers in stage III. The parameters for the energy refinement were identical to the ones that had been used for obtaining the reference structures. The protein was immersed in a shell of water molecules with a thickness of 8 Å. A maximum of 3000 steps of restrained conjugate gradient minimization were applied, using, in addition to the standard AMBER force field, a pseudo-potential for NOE upper distance bounds that is proportional to the sixth power of the restraint violation. The force constant is chosen such that a restraint violation of 0.1 Å contributes 0.3 kcal/mol to the potential energy. The resulting 20 energy-minimized CYANA conformers that represent the solution structure of the protein are the principal result of the FLYA fully automated NMR structure determination method. The FLYA structures and conformational restraints of ENTH, RHO, and SH2 have been deposited in the PDB database with accession codes 2DCP, 2DCQ, and 2DCR, respectively.

Analysis and Structure Comparison. The program MOLMOL[44] was used to visualize 3D structures. CYANA was used to obtain statistics on target function values, restraint violations, Ramachandran plots according to PROCHECK[45] conventions, etc. RMSD values were calculated with CYANA for superpositions of the backbone atoms N, C^{α}, and C′, or the heavy atoms in the structured regions of the proteins. To obtain the RMSD of a structure represented by a bundle of conformers, all conformers were superimposed on the first one, and the average of the RMSD values between the individual conformers and their average coordinates were computed. The single RMSD value between the two sets of mean coordinates of two structure bundles was used to quantify the deviation of one structure bundle from another.[46] Conformational energies were calculated with OPALp[33,34] using the AMBER[32] force field.

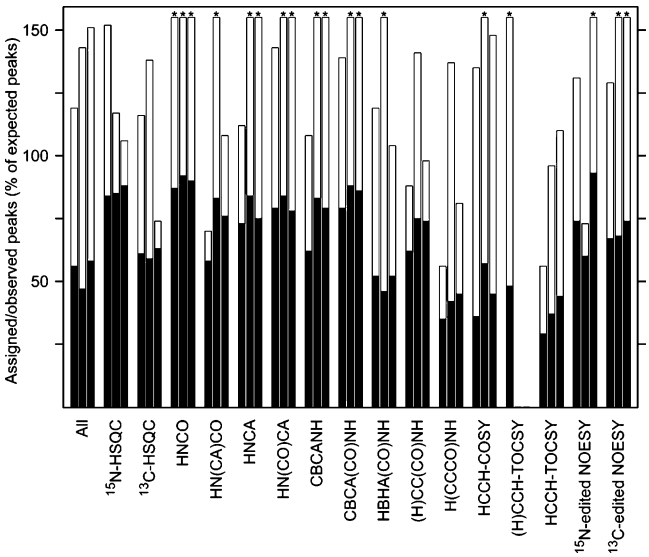

Figure 2. Numbers of observed (open bars) and assigned (black bars) peaks in the fully automated structure determinations of the three proteins ENTH (left bars), RHO (middle bars), and SH2 (right bars), expressed in percent of the corresponding numbers of peaks that are expected based on the amino acid sequence, the ideal magnetization transfer pathways,[27] and, in the case of NOESY spectra, the ^{1}H−^{1}H distances shorter than 4.5 Å in the reference structure. Off-scale values larger than 155% are marked by asterisks.

Results

Fully Automated Structure Determination of Three Proteins. The FLYA algorithm was applied for the NMR structure determination of three proteins of 114−140 amino acid residues: the ENTH-VHS domain At3g16270(9−135) from *Arabidopsis thaliana* (ENTH), the rhodanese homology domain At4g01050(175−295) from *Arabidopsis thaliana* (RHO), and the Src homology domain 2 from the human feline sarcoma oncogene Fes (SH2). High-quality solution structures of these proteins had been determined earlier by conventional techniques.[35−39] Results obtained with FLYA were assessed against these reference structures and reference assignments. A complete FLYA calculation for the ENTH protein required 15 h of computation time for stage I and 25 h each for stage II and stage III.

Peak Identification. The NMR spectra that constituted the input for FLYA (Table 1) were identical with those of the previous conventional structure determination. The number of peaks identified by FLYA ("observed peaks") exceeded the number of ideally expected peaks for most spectra (Figure 2). As a consequence of spectral artifacts and noise, and imperfections of the peak picking algorithm, the experimental peak lists

(43) Güntert, P.; Braun, W.; Wüthrich, K. *J. Mol. Biol.* **1991**, *217*, 517−530.
(44) Koradi, R.; Billeter, M.; Wüthrich, K. *J. Mol. Graph.* **1996**, *14*, 51−55.
(45) Laskowski, R. A.; Rullmann, J. A.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. *J. Biomol. NMR* **1996**, *8*, 477−486.
(46) Güntert, P. *Q. Rev. Biophys.* **1998**, *31*, 145−237.

Table 2. Results of Automated Chemical Shift Assignment[a]

	ENTH				RHO				SH2			
nuclei	number	equal	different	wrong	number	equal	different	wrong	number	equal	different	wrong
All Residues:												
all	1573	84	11	6	1479	85	13	6	1296	85	11	4
backbone, C^β/H^β	1059	85	8	5	997	91	8	4	851	90	5	4
Structured Region:[b]												
all	1405	90	9	4	1345	88	11	3	1199	89	10	3
backbone, C^β/H^β	914	94	5	2	889	96	4	1	760	97	3	2
other CH/CH_2	275	79	21	9	238	71	28	7	202	76	24	5
CH_3	122	95	5	0	108	87	13	2	128	78	20	2
aromatic[c]	70	79	16	7	70	76	17	6	78	73	24	14
NH_2 of Asn/Gln	18	78	22	22	30	53	47	40	24	96	4	4

[a] Number: number of assigned ^1H, ^{13}C, and ^{15}N nuclei. Equal: percentage of chemical shifts that are, within tolerances of 0.03 ppm for ^1H and 0.4 ppm for $^{13}C/^{15}N$, equal to the corresponding shift from the interactive assignment. Different: percentage of shifts that differ by more than the tolerance from the interactively assigned value. Wrong: like "Different" but without a matching interactively assigned shift value within the same residue. ^{13}C and ^{15}N atoms not bound to ^1H are excluded because their assignment has no direct impact on the NOE assignment and the structure calculation. [b] Residues 11−130 for ENTH, 7−125 for RHO, 8−108 for SH2. [c] All aromatic CH and Trp ϵ1NH.

included a considerable number of spurious entries. For certain spectra fewer peaks were observed than expected because of abundant cross-peak overlap, small signal dispersion, and incomplete magnetization transfer in the TOCSY- and COSY-type spectra for the assignment of aliphatic side-chain resonances. The fact, shown below, that spurious and/or missing peaks in the automatically generated peak lists did not prevent the FLYA algorithm from finding the correct 3D structure indicates that a substantial number of artifacts can be tolerated as long as most of the true peaks are identified, too.

The extent to which the automated peak picking could identify the true peaks is reflected in the percentage of expected peaks that could be assigned to an observed peak by the FLYA algorithm (black bars in Figure 2). This quantity shows smaller fluctuations among the three proteins and the different spectra than the corresponding percentage of observed peaks. The completeness of peak assignments is higher for spectra with less signal overlap or higher sensitivity, e.g., the 3D spectra for the backbone assignment. Likewise, in the pairs of related spectra, CBCA(CO)NH and HBHA(CO)NH, (H)CC(CO)NH and H(CCCO)NH, and (H)CCH-TOCSY and HCCH-TOCSY, the completeness of peak assignments is consistently higher for the spectrum that correlates carbon shifts than for the spectrum that correlates protons. The lowest percentages of assigned peaks were obtained for the HCCH-TOCSY experiments, which are in general also difficult to exhaustively analyze manually.

Sequence-Specific Resonance Assignments. Overall, 84−85% of all consensus chemical shifts obtained in step 4 of the final stage III of FLYA were found to agree with the reference values obtained previously by conventional methods (Table 2). These figures rise to 88−90% if the unstructured chain ends, where many resonances could not be assigned manually, are excluded. Furthermore, many of the remaining differences to the reference assignment resulted from local permutations of resonance assignments within a given residue. Assignment differences that map to atoms of the same residue are less likely to affect the derived structure than assignment differences that map to atoms on different residues. The percentage of assignments that differ from all reference assignments within the same residue is 3−4% for all nuclei in the structured regions (Table 2).

A particularly high degree of 94−97% consistency with the reference assignment was observed for the backbone and C^β/H^β chemical shifts. Only 1−2% of these shifts were not assigned to the correct residue. This reflects the fact that the NMR experiments for the backbone assignment are generally more sensitive and exhibit less peak overlap than those for the side-chain assignment, and that the redundancy among the different backbone assignment spectra allows for extensive cross-checks during the automated procedure.

Methyl groups are crucial for the determination of the 3D structure because they are predominantly located in the hydrophobic core where they give rise to many NOEs. Of the methyl groups in the structured region, 98−100% were assigned to the correct residue. Slightly lower values of 91−95% were observed for the side-chain methylene and methine groups beyond C^β/H^β, presumably because of poor signal dispersion and frequent accidental degeneracies of the chemical shifts of diastereotopic methylene protons that complicate the task of the automated chemical shift assignment algorithm. The assignment of aromatic ring protons relies on NOEs that have an intrinsically higher degree of ambiguity than through-bond connectivities. Of the aromatic resonances, 86−94% were assigned to the correct residue. The correctness of the NOE-based chemical shift assignment of the Asn/Gln side-chain amide resonances varied from 96% for SH2, 78% for ENTH, to 53% for RHO.

The extent of correct chemical shift assignments for the three proteins in Table 2 fulfills the previously established minimal requirements for the successful use of combined automated NOE assignment and structure calculation with CYANA.[8,17]

NOESY Cross-Peak Assignment and Structure Calculation. 3D structures were calculated with CYANA on the basis of the chemical shift assignments of Table 2 and the automatically prepared NOESY peak lists. The structures from FLYA agree well with those from the conventional approach, as shown by the superpositions of Figure 3 and by RMSD values of 0.77−0.94 Å for the backbone and 1.26−1.44 Å for all heavy atoms between the mean structures from both approaches (Table 3). Mainly because of loops on the protein surface, which are poorly defined by the experimental NMR data, these deviations are slightly higher than the corresponding RMSDs within the reference structure bundles of 0.43−0.50 Å for the backbone and 0.84−0.97 Å for all heavy atoms.

The automatically picked NOESY peak lists contained between 1.8 and 3.3 times as many entries as the ones used for the conventional structure determination (Table 3), many of

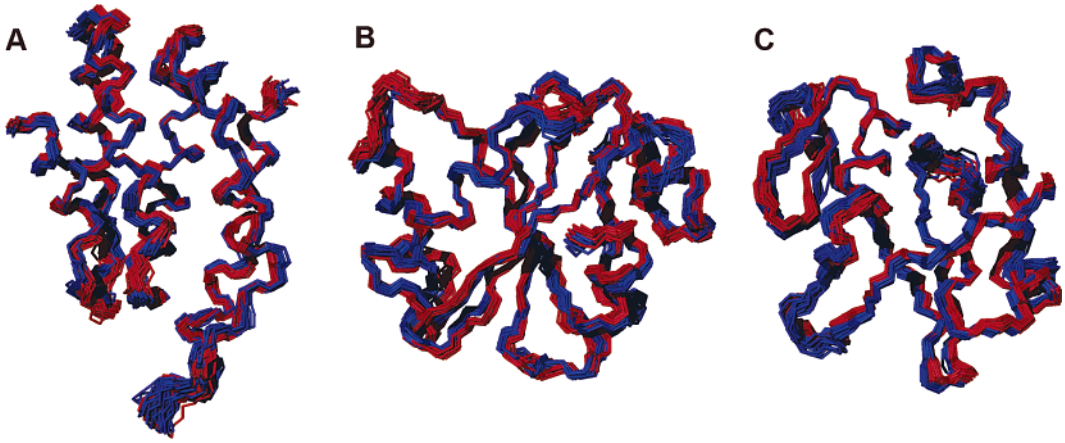

Figure 3. Structures obtained by fully automated structure determination (blue) superimposed on the corresponding NMR structures determined by conventional methods (red). (A) ENTH. (B) RHO. (C) SH2.

Table 3. Statistics of the Structure Determinations Using Either the Fully Automated FLYA Algorithm or the Conventional Approach[a]

	ENTH		RHO		SH2	
	FLYA	1VDY	FLYA	1VEE	FLYA	1WQU
NOESY cross-peaks picked	10 706	5910	14 056	5411	13 894	4238
NOESY cross-peaks assigned	5409	5768	4653	5294	4864	4109
NOE upper distance limits:	3440	3348	2903	3043	3145	2291
short-range, $\|i-j\| \leq 1$	1605	1557	1350	1392	1455	1007
medium-range, $1 < \|i-j\| < 5$	897	915	450	539	492	379
long-range, $\|i-j\| \geq 5$	938	876	1103	1112	1198	905
maximal violation (Å)	0.15	0.14	0.14	0.14	0.14	0.14
violations > 0.2 Å	0	0	0	0	0	0
CYANA target function[b] (Å^2)	0.51	0.48	0.73	0.74	1.54	0.47
AMBER energy (kcal/mol)	−5456	−5506	−4770	−5021	−3858	−4047
Ramachandran plot[c] (%)	86/14/1/0	88/11/0/0	71/27/1/1	79/21/0/0	76/23/1/0	81/18/1/0
RMSD to mean coordinates (Å)[d]	0.42/0.84	0.50/0.97	0.39/0.77	0.43/0.84	0.35/0.72	0.44/0.93
RMSD between mean structures (Å)	0.77/1.26		0.94/1.39		0.94/1.44	

[a] The conventional approach that was employed to determine the structures deposited in the PDB with accession codes 1VDY for ENTH, 1VEE for RHO, and 1WQU for SH2 is based on interactive chemical shift assignment, automated NOESY assignment, and restrained energy minimization against the AMBER force field. When applicable, the value given is the average over the 20 energy-refined CYANA conformers that represent the solution structure. [b] Calculated before restrained energy minimization using the same repulsive core radii as in the original, conventional structure determinations. [c] Percentage of residues in most favored, additionally allowed, generously allowed, and disallowed regions of the Ramachandran plot. [d] RMSD values for the backbone atoms N, C^α, and C′ or for all the heavy atoms, respectively, in the structured regions of residues 11−130 for ENTH, 7−125 for RHO, and 8−108 for SH2.

which must relate to noise or artifacts that have been mistaken as peaks by the automated peak picking algorithm. Accordingly, 49−67% of the entries in the NOESY peak lists were not assigned and not used for the generation of conformational restraints by CYANA. In contrast, during the conventional structure determination, 97−98% of the entries in the virtually artifact-free NOESY peak lists prepared by careful visual inspection of the spectra had been assigned. However, despite the different sizes and qualities of the input peak lists, the number of assigned NOESY cross-peaks from FLYA differed by only −6% for ENTH, −12% for RHO, and +18% for SH2 from those of the conventional structure determination, and comparable numbers of NOE distance restraints were obtained. This indicates the effectiveness of network-anchoring and constraint combination[8] in the presence of many artifact peaks. Most of the erroneous NOESY peaks were rejected already in the first cycle of automated NOESY cross-peak assignment when no 3D structure was available to validate NOEs. For instance, 5024 out of the finally 5297 unassigned peaks for ENTH were already rejected in the first cycle.

If atoms with correct or wrong chemical shift assignments were used equally often in forming NOE distance restraints, one would expect, for instance in the case of ENTH with $p =$ 96% of the chemical shift assignments to the correct residue, to find $p^2 = 92\%$ of the NOE restraints involving only atoms assigned to the correct residue. However, the latter percentage was 97% for ENTH, and only 5 of the 102 NOE distance restraints that involved chemical shifts assigned to a wrong residue were consistently violated by more than 0.4 Å in the ENTH reference structure. This shows the effectiveness of the FLYA algorithm in minimizing the effect of erroneously assigned chemical shifts on the resulting structure.

The structures from FLYA and from the conventional approach do not differ with regard to the amount or size of residual restraint violations. Favorable conformational energies and Ramachandran plot statistics resulted for both structures. Overall, using the fully automated approach lead only to a marginal decrease of the structural quality measures in Table 3 relative to the conventionally determined reference structures.

Reliability Measures. The acceptable ranges and actual values of the reliability indicators introduced in the Algorithm section to monitor the performance of the FLYA algorithm are given in Table 4. The *peak picking extent* varies considerably among the three proteins. Nevertheless, the algorithm could cope with the different amounts of artifact peaks and yielded structures and assignments of comparable quality for the three proteins, as indicated by similar values of the reliability indicators for the steps of the algorithm that follow peak picking.

Table 4. Reliability Measures for FLYA Calculations

reliability measure[a]	acceptable range	ENTH	RHO	SH2
peak picking extent (%)	65−300	110	264	203
peak assignment completeness (%)	>50	57	67	66
chemical shift assignment redundancy (peaks/shift)	>4	8.8	7.3	8.6
chemical shift ensemble self-consistency (%)	>70	78	81	84
long-range distance restraints per residue	>4	6.7	8.2	10.5
initial fold precision[b] (Å)	<3.0	1.24	1.99	1.27
packing quality (kcal/mol)	<−2.4	−3.3	−3.1	−2.9

[a] Reliability measures are defined in the Algorithm section. [b] RMSD values are calculated for the structured regions comprising residues 11−130 for ENTH, 7−125 for RHO, and 8−108 for SH2.

Table 5. Results of Automated Chemical Shift Assignment and Statistics of Structure Determinations for ENTH Using Different Peak Picking Procedures[a]

	NMRView			interactive
	Stage I	Stage III	AUTOPSY	peak picking
Chemical Shift Assignments:				
all residues:				
all (equal/different/wrong; %)	82/12/6	84/11/6	82/12/6	85/10/5
backbone, C^β/H^β	87/7/5	85/8/5	85/9/5	86/7/5
structured region (residues 11−130):				
all (equal/different/wrong; %)	88/11/5	90/9/4	88/11/4	92/7/3
backbone, C^β/H^β	96/4/2	94/5/2	93/6/2	96/4/1
other CH/CH_2	73/27/9	79/21/9	71/29/12	84/16/7
CH_3	89/11/2	95/5/0	93/7/1	93/7/0
aromatic	69/26/17	79/16/7	89/6/3	80/14/6
NH_2 of Asn/Gln	44/56/56	78/22/22	83/17/17	67/33/33
Structure Determination:				
NOESY cross-peaks picked	10706	10706	8362	5902
NOESY cross-peaks assigned	5160	5409	4592	5285
NOE upper distance limits	3211	3440	2797	3246
long-range NOE restraints, $\|i - j\| \geq 5$	821	938	673	853
AMBER energy (kcal/mol)	−5249	−5456	−5384	−5458
Ramachandran plot statistics (%)	81/18/1/0	86/14/1/0	84/15/1/0	85/14/1/0
RMSD to mean coordinates (Å)	0.48/0.94	0.42/0.84	0.55/1.00	0.61/1.10
RMSD to reference structure (Å)	1.18/1.74	0.77/1.26	0.99/1.57	0.92/1.32

[a] Input peak lists for the fully automated procedure were prepared using either automated peak picking with NMRView or AUTOPSY or interactive peak picking using the program NMRView. Quantities are defined as those in Tables 1 and 2.

The *peak assignment completeness* measure shows that well over half of all theoretically expected peaks were found and assigned. The *chemical shift assignment redundancy* measure indicates that the chemical shift assignments are based on an average of 7−9 peaks for each nucleus in the "through-bond" spectra, which have an intrinsically lower degree of ambiguity than the NOESY spectra. Spurious chemical shift assignments that rely on a single or very small number of non-NOESY peaks are therefore rare. Consequently, the *chemical shift ensemble self-consistency* is high, as indicated by 78−84% of the assigned nuclei having more than 80% of the 20 individual values in the ensemble of chemical shift assignments in agreement with their corresponding consensus chemical shift value. Numbers of 7−10 *long-range distance restraints per residue* are indicative of a dense network of tertiary structure-defining conformational restraints. The *packing quality* of the resulting three-dimensional structures is corroborated by low, negative values of the AMBER Lennard-Jones energy. All FLYA calculations in Tables 2−5 fulfill all the acceptance criteria of Table 4.

Discussion

In this section we discuss several important aspects of fully automated structure determination with the FLYA algorithm, using ENTH as an example.

Alternative Peak Picking Methods. To evaluate the susceptibility of the FLYA algorithm on the peak picking algorithm used, we compared the generic FLYA method that uses the automated peak picking algorithm of NMRView with variants using the program AUTOPSY for automated peak picking or using manual peak identification by visual inspection of the spectra. Despite considerable differences in the numbers of picked peaks, the percentage of assigned peaks varied only from 50% with AUTOPSY, 56% with NMRView, to 58% with manually prepared peak lists (Figure 4), and similar degrees of correct chemical shift assignments were obtained (Table 5): 92%, 90%, and 88% of all chemical shifts in the structured region were assigned correctly using peak lists prepared by visual inspection, NMRView, or AUTOPSY, respectively. The number of chemical shifts that were not assigned to the correct residue varied among the three different peak picking procedures by not more than 1% for all atoms. The fraction of correct chemical shift assignments achieved by the automated procedures was largely equivalent to that from the manually prepared peak lists, even though the automatically prepared peak lists contained many more erroneous entries. Of the atoms in the structured region for which the chemical shifts were not assigned to the correct residue with at least one of the three peak picking methods, 62%, 19%, and 19%, respectively, were misassigned with one, two, or all three peak picking methods simultaneously. This suggests that the contents of the peak lists play a more important role in causing erroneous chemical shift assignments than possibly incomplete convergence or instability of the

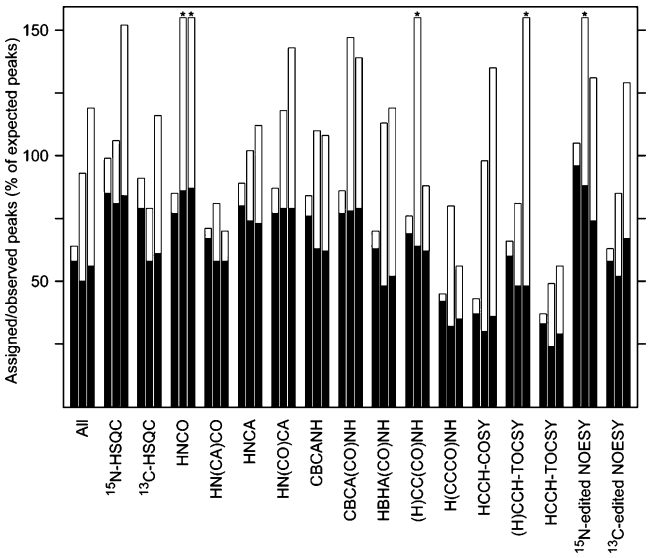

Figure 4. Observed (open bars) and assigned (black bars) peaks in fully automated structure determinations of the protein ENTH using peak picking by interactive visual inspection of the spectra (left bar for each spectrum), by the automated AUTOPSY algorithm[24] (middle bars), or by the automated algorithm in NMRView[23] (right bars). Percentages are defined as those in Figure 2.

algorithm. It is conceivable to improve the FLYA results by the simultaneous use of multiple peak picking algorithms.

No superior structural quality in terms of precision, accuracy, Ramachandran plot statistics, and conformational energy was achieved when peak lists were prepared manually (Table 5). The backbone RMSD to the reference structure was in all cases below 1.0 Å for the structured region. This robustness of the FLYA algorithm with regard to imperfections of the peak picking algorithms is important because a strictly computational peak picking algorithm that could identify true peaks with the same reliability as an experienced spectroscopist remains elusive.

Recently, fully automated sequence-specific resonance assignment of the 14 kDa protein azurin by the combined use of the programs AUTOPSY for peak picking, PICS for peak list recalibration and filtering, and GARANT for resonance assignment was reported.[26] Correct assignments could be determined for 85% of the chemical shifts of azurin, which is in agreement with the 84−85% correct assignments obtained by FLYA for the three proteins ENTH, RHO, and SH2. These results suggest the future feasibility of structure calculations for azurin on the basis of the automated resonance assignments of ref 26.

Consensus Chemical Shifts. Automated chemical shift assignment with FLYA consists of two steps (steps 3 and 4 in Figure 1). First, an ensemble of chemical shift assignments is computed. Second, these raw chemical shift assignments are consolidated into a single consensus chemical shift list, which is then used for the automated assignment of NOEs. We verified the importance of the consensus chemical shift assignment step by performing 20 FLYA test calculations for ENTH, each one using the chemical shifts from a single one of the 20 individual GARANT runs, skipping the consolidation step. None of these 20 simplified calculations could reach the same agreement with the reference assignments and the reference structure as the generic FLYA algorithm that uses consensus chemical shifts. The overall percentage of correct consensus chemical shift assignments from the generic FLYA algorithm was 84% (Table 2), whereas the individual runs reached only 80−83%. The structures resulting from the individual runs without consolidation deviated from the reference structure by RMSDs of 0.88−3.10 Å for the backbone and 1.43−3.38 Å for all atoms, which were always larger than the corresponding values of 0.77 and 1.26 Å from the standard FLYA algorithm that uses consensus chemical shifts. The consolidation of the ensemble of chemical shift assignments from multiple GARANT runs into a single consensus chemical shift list is thus essential for the reliability of the FLYA algorithm and should always be applied, despite the considerably increased computation time that is required for the additional GARANT runs.

Use of Intermediate Structures to Refine the Chemical Shift Assignment. The intermediate 3D structures used in stages II and III of the FLYA algorithm provide additional information to the chemical shift assignment algorithm that is used to supplement the list of expected NOESY peaks by medium- and long-range NOEs and to allow for a refined generation of expected short-range NOEs. The improvement of the quality of the chemical shift assignments and the structure by the iterative use of intermediate structures is illustrated in Table 5 by a comparison of the results from the FLYA stages I and III for ENTH. The extent of correct chemical shift assignments increased from stage I to stage III. In particular, there were remarkable increases of 6−34% in the correctness of the chemical shift assignments for the different groups of side-chain nuclei. Not surprisingly, the most pronounced improvements were observed for the nuclei whose assignment relies on NOEs, the aromatics, from 69% in stage I to 79% in stage III, and the side-chain NH_2 groups, from 44% in stage I to 78% in stage III. The improved chemical shift assignments in stage III resulted in more NOESY cross-peak assignments, a 14% higher number of long-range NOEs, and a more accurate 3D structure (Table 5). The RMSD to the reference structure decreased from 1.18 Å in stage I to 0.77 Å in stage III for the backbone atoms and from 1.74 Å in stage I to 1.26 Å in stage III for all heavy atoms.

Conclusions

The results of the FLYA structure determinations of three proteins show that fully automated NMR structure determination of proteins up to 140 amino acid residues is possible now. The method is purely computational and can cope with the amount of overlap and artifacts present in typical experimental NMR spectra. The FLYA structure determinations in this paper have been performed without any manual intervention. It would be straightforward to further improve the results by interactive improvements of the peak lists, corrections of erroneous chemical shift assignments, and/or additional conformational restraints for torsion angles, hydrogen bonds, residual dipolar couplings, etc. Various extensions of the basic FLYA algorithm can be envisaged. NMR data processing can be incorporated in order to start the procedure from the raw time-domain data from the NMR spectrometer. Alternative peak picking algorithms can be used. The currently static peak lists may be replaced by dynamic peak lists that will be updated continuously on the basis of intermediate results[47] during a FLYA calculation. An optimized resonance assignment algorithm can reduce the computation time and make more sophisticated use of intermediate 3D structures. Additional refinement techniques can

(47) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **2002**, *24*, 171−189.

improve the structures with respect to common quality measures.[48,49] The number of input spectra can be reduced for well-behaved proteins.[50]

Combining FLYA with stereo-array isotope labeling (SAIL)[51] promises to enhance significantly the efficiency, reliability, and size range of applicability of fully automated NMR protein structure determination. Chemical and enzymatic synthesis incorporates deuterium into the protein's constituent amino acids such that each carbon atom will have at most one ^{1}H nucleus bonded to it, the remaining hydrogens having been replaced by ^{2}H. SAIL provides maximal structural information consistent with spectral simplification and isotopic dilution. It decreases spectral crowding, preserves through-bond connectivities for backbone and side chain assignments, eliminates the need for stereospecific assignments, and reduces spin diffusion. Lines are sharpened by eliminating dipolar relaxation pathways and long-range couplings, resulting in 3−7 times higher signal-to-noise ratios. This gain can be exploited for higher quality spectra of larger proteins or for an order of magnitude shorter NMR measurement times with smaller proteins. Automated signal identification can be achieved with higher reliability for the fewer, sharper, and more intense peaks of SAIL proteins. The danger of making erroneous assignments decreases with the number of nuclei and peaks to assign, and less spin diffusion allows NOEs to be interpreted more quantitatively. A SAIL-adapted FLYA algorithm can make use of these features to enable automated NMR structure determination of proteins with a molecular weight above 20 kDa,[51,52] for which the large number of chemical shifts and peaks renders the traditional manual analysis method particularly cumbersome and error-prone. For the future, we expect fully automated NMR protein structure determination to replace most manual and semiautomatic approaches and to produce structures of the same quality as by manual spectrum analysis.

Acknowledgment. This work was supported by the National Project on Protein Structural and Functional Analyses of the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), a Grant-in-Aid for Scientific Research of the Japan Society for the Promotion of Science (JSPS), and the Tatsuo Miyazawa Memorial Program of RIKEN Genomic Sciences Center.

Supporting Information Available: Complete refs 35 and 36. This material is available free of charge via the Internet at http://pubs.acs.org.

JA061136L

(48) Linge, J. P.; Williams, M. A.; Spronk, C. A. E. M.; Bonvin, A. M. J. J.; Nilges, M. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 496−506.
(49) Nederveen, A. J.; Doreleijers, J. F.; Vranken, W.; Miller, Z.; Spronk, C. A. E. M.; Nabuurs, S. B.; Güntert, P.; Livny, M.; Markley, J. L.; Nilges, M.; Ulrich, E. L.; Kaptein, R.; Bonvin, A. M. J. J. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 662−672.
(50) Scott, A.; López-Méndez, B.; Güntert, P. *Magn. Reson. Chem.* **2006**, *44*, S83−S88.
(51) Kainosho, M.; Torizawa, T.; Iwashita, Y.; Terauchi, T.; Ono, A. M.; Güntert, P. *Nature* **2006**, *440*, 52−57.
(52) Ikeya, T.; Terauchi, T.; Güntert, P.; Kainosho, M. *Magn. Reson. Chem.* **2006**, *44*, S152−S157.