NMR Spectroscopic Methods

DOI: 10.1002/anie.200603213

Fast High-Resolution Protein Structure Determination by Using Unassigned NMR Data**

*Jegannath Korukottu, Monika Bayrhuber, Pierre Montaville, Vinesh Vijayan, Young-Sang Jung, Stefan Becker, and Markus Zweckstetter**

NMR spectroscopy provides high-resolution structural information of biomolecules in near-physiological conditions. Although significant improvements were achieved in NMR spectroscopy in the last 20 years,[1] the increase in genome sequencing data has created a need for rapid and efficient methods of NMR-based structure determination.[2,3] NMR data acquisition can be accelerated significantly when sensitive spectrometers are combined with new methods for sampling chemical shifts in multidimensional NMR experiments.[4] Therefore, data analysis and in particular the requirement to assign side-chain chemical shifts to specific atoms is the major bottleneck of rapid NMR-based structure determination. Herein, we present a method, termed FastNMR (**fa**st **st**ructure determination by **NMR**) that enables automatic, high-resolution NMR structure determination of one-domain-sized proteins from unassigned NMR data. By using FastNMR, the de novo structure of the 65-residue cone snail neurotoxin conkunitzin-S2 was determined automatically.

Although good progress has been made towards prediction of 3D protein structures from amino acid sequences, the quality of the predictions is still limited.[5] These problems may be overcome when a limited number of easily accessible NMR spectroscopic data is combined with ab initio methods.[6–8] To obtain high-resolution protein structures, experimental distance information is required. The distance information can be extracted from NOE spectra with little manual intervention when assignment of NOE peaks and structure calculation are performed iteratively.[9–13] To determine the correct structure, however, a nearly complete and error-free manual assignment of chemical shifts is essential.[14] Alternatively, if excellent, unambiguously identified NOESY peak lists are available, it may be possible to obtain a 3D protein structure in the absence of any chemical-shift assignments from the distance information provided in NOESY spectra.[15,16] FastNMR differs from these approaches in that it starts from unassigned chemical shifts, NOEs, and residual dipolar couplings (RDCs), avoids wrong structures by cross-validation, works for experimental data, requires only a limited number of NMR spectra, and produces high-resolution (<1 Å) structures.

The strategy of FastNMR is based on an approach that has proven to be robust in manual structure determination. This includes usage of information from triple-resonance experiments for sequential backbone assignment, use of iterative NOE assignment and structure calculation, and structure refinement by using RDCs (see the Supporting Information). The key to the success of FastNMR is, however, the simultaneous determination of the backbone assignment and the protein fold prior to analysis of NOE data. This is achieved by iterative RDC-enhanced backbone assignment and fold determination by using the assignment program Mars and the ab initio program Rosetta.[6,8] The next step is to get from a protein backbone to a 3D structure including side chains. For this aim, side chains are built onto the protein backbone and proton and carbon chemical shifts are predicted from the ensemble of the 20 lowest-energy Rosetta-NMR structures by using empirical formulas and artificial neural networks.[17,18] Experimental chemical shifts are then matched to the predicted values and assigned based on the minimal chemical-shift difference. The structure of the protein backbone and the assignment of backbone and side-chain chemical shifts are subsequently used for automated NOE assignment by using the program Cyana.[9] Cyana, however, is not only used for NOE assignment. All distance constraints involving proton chemical shifts, which could not be unambiguously assigned by comparison with predicted values, are treated as ambiguous NOEs. In this way, the assignment of side-chain chemical shifts is partially done as part of the automated NOE assignment. By using the NOE-based 3D structure, the prediction of chemical shifts is improved and a second round of automated NOE assignment is performed. Finally, all experimental data are combined and a high-resolution structure is obtained (see the Supporting Information). FastNMR is automatically performed; that is, FastNMR takes lists of unassigned NMR data as the input and outputs a high-resolution 3D structure.

FastNMR was tested on the 60-residue conkunitzin-S1 (Conk-S1) and the 76-residue protein ubiquitin. For both proteins, 3D structures as well as chemical-shift assignments are known allowing evaluation of FastNMR.[22,27] Furthermore, the high-resolution structure of the 65-residue toxin conkunitzin-S2 (72% sequence identity to Conk-S1) was determined by FastNMR. Neither NMR data nor a 3D

[*] J. Korukottu, M. Bayrhuber, Dr. P. Montaville, V. Vijayan, Dr. Y.-S. Jung, Dr. S. Becker, Dr. M. Zweckstetter
Department of NMR-Based Structural Biology
Max Planck Institute for Biophysical Chemistry
Am Fassberg 11, 37077 Göttingen (Germany)
Fax: (+49) 551-201-2202
E-mail: mzwecks@gwdg.de

[**] We thank Christian Griesinger for useful discussions, Karin Giller and Kamila Sabagh for technical assistance, and Baldomero M. Olivera for the cDNA clones of Conk-S1 and Conk-S2. This work was supported by the Max Planck Society. M.Z. is the recipient of a DFG Emmy Noether fellowship (ZW 71/1-5).

Supporting information for this article is available on the WWW under http://www.angewandte.org or from the author.

structure were previously available for Conk-S2. Figure 1 and Table 1 show that, for all three proteins, FastNMR calculated high-resolution 3D structures from unassigned NMR data. The spread in the ensemble of 20 lowest-energy structures was below 0.7 Å for the backbone and below 1.4 Å for all heavy atoms. The manually and automatically determined structures were of similar energy. The FastNMR structures of Conk-S1 and ubiquitin deviate by 0.4 Å and 0.6 Å, respectively, from the conventionally determined structures.[22,27] The FastNMR calculation of each protein was completed in less than 24 h.

In FastNMR, the assignment of side-chain resonances is performed automatically by comparison with values predicted from protein backbones established early in the FastNMR calculation. Tests show that the root-mean-square deviation (rmsd) between predicted and experimental chemical shifts is 0.19 ppm for protons and 1.1 ppm for carbons visible in HCCONH- and CCONH-TOCSY spectra (see the Supporting Information).[17,18] Accordingly, assignments are only considered when the difference between predicted and measured chemical shift is less than 0.3 ppm for protons and 1.3 ppm for carbons. In addition, when two experimentally observed ^{1}H chemical shifts belonging to the same residue (as established by HCCONH and CCONH-TOCSY spectra) differ by less than 0.3 ppm, then all NOE signals owing to either of the two shifts are considered as ambiguous during the automated NOE assignment. By using this approach, all experimentally observed carbon chemical shifts of Conk-S1, Conk-S2, and ubiquitin were assigned unambiguously. ^{1}H chemical shifts, however, are often degenerate and about 10% of the measured side-chain ^{1}H chemical shifts could not be assigned unambiguously (see the Supporting Information).

Previously, it was suggested that for successful automated NOE assignment at least 90% of all proton chemical shifts have to be assigned.[14] FastNMR in its current implementation, however, only uses 3D CCONH- and HCCONH-TOCSY NMR experiments and only approximately 60% of all protons were assigned by FastNMR prior to starting the NOE analysis (see Table S2 in the Supporting Information). For all the protons for which no experimental chemical shifts are available, FastNMR uses predicted chemical shifts for automated NOE assignment. As the predicted chemical shifts are not very accurate, the window size that is used for matching NOEs to ^{1}H chemical shifts was increased from 0.05 ppm to 0.3 ppm. Furthermore, NOE distance restraints assigned to protons with predicted chemical shifts are used in the final structure refinement only if the same proton

Table 1: Structural statistical data of the investigated peptides.[a]

	Conk-S1		Ubiquitin		Conk-S2
	2CA7	FastNMR	1D3Z[b]	FastNMR	FastNMR
Number of NOEs[c]	551	464	1744[d]	635	570
long-range	113	72	731	119	160
medium-range	73	57	291	106	79
short-range	365	335	722	410	331
Number of dihedral angles	126	126	98	127	167
Violations $>5°$	3 ± 1	3 ± 1	0	1 ± 1	2 ± 1
Number of RDCs	190	190	200	200	138
RDC types[e]	1,2,3,4	1,2,3,4	1,2,3	1,2,3	1,2,3
Energy [kcal mol^{-1}]	−1267.2	−1387.9	−2767.9	−2247.6	−1025.0
Ramachandran plot [%]					
Most favored	88.2	87.5	95.0	95.5	84.5
Disallowed	2.0	1.0	0.0	0.0	1.9
Coordinate precision [Å][f]					
Backbone atoms	0.6	0.7	0.3	0.4	0.6
All heavy atoms	1.2	1.4	0.9	1.1	1.2

[a] Statistics for ensembles of 20 structures. [b] Structure recalculated based on experimental restraints of 1D3Z. [c] None of the structures exhibited distance violations greater than 0.5 Å. [d] Only 58% of the long-range NOEs are nonredundant. [e] 1, 2, 3, 4 refer to the RDCs $^1D_{N\text{-}H}$, $^1D_{C\text{-}N}$, $^1D_{C\alpha\text{-}C}$, $^1D_{C\alpha\text{-}H\alpha}$, respectively. [f] Defined as the average rmsd difference between the final 20 FastNMR structures and the mean coordinates for residues 2–72 (ubiquitin), 3–60 (Conk-S1), and 5–60 (Conk-S2).

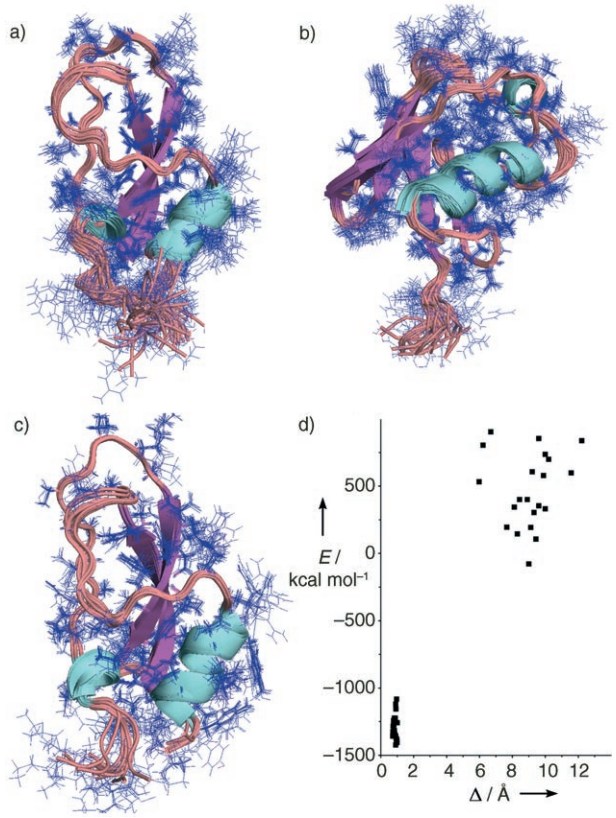

Figure 1. FastNMR 3D structure of a) Conk-S2, b) ubiquitin, and c) Conk-S1. d) Comparison of the total energy with the deviation from the native structure of Conk-S1 in FastNMR stability tests (see the Supporting Information).

(predicted chemical shift) is assigned to two or more NOE peaks and the experimental chemical shift of the two NOE peaks differ by less than 0.1 ppm. In combination with the backbone conformation that is already established prior to the NOE analysis, this allows the determination of high-resolution protein structures.

FastNMR was applied to NMR data of three proteins with all the inherent difficulties of peak overlap, missing backbone resonances, noise peaks, and multiple conformations. NOE peaks with multiple chemical-shift assignments are fully taken into account by the use of ambiguous distance constraints. In addition, we tested the impact of reduced data quality and incorrect backbone assignments (see the Supporting Information). Despite these complications, FastNMR produced high-resolution structures, including the de novo structure of Conk-S2. These results demonstrate that FastNMR is highly robust.

Cross-validation ensures that no incorrect structures are produced by FastNMR: For signal assignment and protein-backbone-structure determination, only RDCs and chemical shifts are used, whereas during automated NOE assignment RDCs are not used. Thus, in case the initial protein backbone is incorrect, it is unlikely that a sufficient number of NOEs were assigned during automated, structure-based NOE assignment. Even if a large enough number of NOEs are assigned, the NOE-based structure will likely differ significantly from the initial structure of the protein backbone and disagree with the RDCs. Therefore, in the final stage of FastNMR, when all experimental data are combined, convergence to a low-energy structure is not possible. This is can be determined from Figure 1 d and additional stability tests: A low total energy is only obtained by FastNMR for correct, high-resolution structures. In addition, FastNMR structures have to pass the following check points: 1) at the end of each stage, FastNMR structures must have converged to a unique conformation, 2) structural changes during FastNMR must differ by less than 3.5 Å from the initial backbone structure to the high-resolution structure, 3) more than 85 % of the backbone resonances must have been assigned before the automated NOE assignment is started, and 4) FastNMR structures have to pass the standard NMR spectroscopy quality criteria, such as a low number of violations of the experimental restraints (Table 1).

FastNMR in its current implementation is limited to domain-sized proteins. This is mainly because the only experiments that are used for extraction of side-chain chemical shifts are CCONH- and HCCONH-TOCSY experiments. The capabilities of these experiments decreases with increasing molecular weight of the protein and also do not allow access to chemical shifts of aromatic groups. A larger number of chemical shifts will be available when 3D HCCH-COSY and 3D HCCH-TOCSY spectra[19] are incorporated into FastNMR. In addition, aromatic chemical shifts can be obtained from two-dimensional $(H_\beta)C_\beta(C_\gamma C_\delta)H_\delta$ and $(H_\beta)C_\beta$-$(C_\gamma C_\delta C_\varepsilon)H_\varepsilon$ spectra.[20] The incorporation of these experiments into FastNMR is in progress.

In conclusion, we have demonstrated that it is possible to determine high-resolution structures of domain-sized proteins within 24 h of starting from unassigned chemical shifts, RDCs, and NOE peak lists. We have also used this approach to determine the de novo structure of the neurotoxin Conk-S2. FastNMR runs automatically, avoids wrong structures by cross-validation, works for experimental data, requires only a limited number of NMR spectra, and produces high-resolution structures. No manual assignment of chemical shifts or interresidue correlations is required.

Experimental Section

Ubiquitin and Conk-S1 were produced recombinantly as described previously.[21,22] The production and structural details of Conk-S2 will be reported elsewhere. NMR spectra were recorded on Bruker 600-, 700-, or 900-MHz spectrometers according to availability. A detailed list of the spectra and the conditions used can be found in the Supporting Information. Referencing of spectra, peak picking, and peak grouping were performed by using the program Sparky. Torsion angles χ_1 were obtained from 2D ^{15}N-$^{13}C^\gamma$ and ^{13}C-$^{13}C^\gamma$ spin-echo difference experiments.[23] NOE peak volumes were derived from 3D ^{15}N- and ^{13}C-edited NOESY spectra.[24,25] RDCs were measured from interleaved 3D TROSY-HNCO and 3D CBCA(CO)NH spectra.[26] For the calculations, a cluster of ten 3.06-GHz Linux PCs was used; however, most calculation steps use only a single CPU. The structure of Conk-S2 has been deposited in the protein databank (PDB code 2J6D). Software for the use of FastNMR is available from the authors upon request.

Received: August 8, 2006
Published online: January 5, 2007

Keywords: genomics · NMR spectroscopy · protein structures · structure elucidation

[1] K. Wüthrich, *Angew. Chem.* **2003**, *115*, 3462; *Angew. Chem. Int. Ed.* **2003**, *42*, 3340.
[2] A. Abbott, *Nature* **2005**, *435*, 547.
[3] G. T. Montelione, D. Y. Zheng, Y. P. J. Huang, K. C. Gunsalus, T. Szyperski, *Nat. Struct. Biol.* **2000**, *7*, 982.
[4] G. H. Liu, Y. Shen, H. S. Atreya, D. Parish, Y. Shao, D. K. Sukumaran, R. Xiao, A. Yee, A. Lemak, A. Bhattacharya, T. A. Acton, C. H. Arrowsmith, G. T. Montelione, T. Szyperski, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10487.
[5] P. Bradley, K. M. Misura, D. Baker, *Science* **2005**, *309*, 1868.
[6] Y. S. Jung, M. Sharma, M. Zweckstetter, *Angew. Chem.* **2004**, *116*, 3561; *Angew. Chem. Int. Ed.* **2004**, *43*, 3479.
[7] J. Meiler, D. Baker, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 15404.
[8] C. A. Rohl, D. Baker, *J. Am. Chem. Soc.* **2002**, *124*, 2723.
[9] P. Guntert, *Prog. Nucl. Magn. Reson. Spectrosc.* **2003**, *43*, 105.
[10] J. P. Linge, M. Habeck, W. Rieping, M. Nilges, *Bioinformatics* **2003**, *19*, 315.
[11] J. Kuszewski, C. D. Schwieters, D. S. Garrett, R. A. Byrd, N. Tjandra, G. M. Clore, *J. Am. Chem. Soc.* **2004**, *126*, 6258.
[12] W. Gronwald, S. Moussa, R. Elsner, A. Jung, B. Ganslmeier, J. Trenner, W. Kremer, K. P. Neidig, H. R. Kalbitzer, *J. Biomol. NMR* **2002**, *23*, 271.
[13] Y. J. Huang, R. Tejero, R. Powers, G. T. Montelione, *Proteins Struct. Funct. Genet.* **2006**, *62*, 587.
[14] J. Jee, P. Guntert, *J. Struct. Funct. Genomics* **2003**, *4*, 179.
[15] A. Grishaev, M. Llinas, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 10941.
[16] A. Grishaev, C. A. Steren, B. Wu, A. Pineda-Lucena, C. Arrowsmith, M. Llinas, *Proteins Struct. Funct. Genet.* **2005**, *61*, 36.
[17] J. Meiler, *J. Biomol. NMR* **2003**, *26*, 25.
[18] K. Osapay, D. A. Case, *J. Biomol. NMR* **1994**, *4*, 215.

[19] L. E. Kay, G. Y. Xu, A. U. Singer, D. R. Muhandiram, J. D. Forman-Kay, *J. Magn. Reson. Ser. B* **1993**, *101*, 333.
[20] T. Yamazaki, J. D. Forman-Kay, L. E. Kay, *J. Am. Chem. Soc.* **1993**, *115*, 11 054.
[21] G. A. Lazar, J. R. Desjarlais, T. M. Handel, *Protein Sci.* **1997**, *6*, 1167.
[22] M. Bayrhuber, R. Graf, M. Ferber, M. Zweckstetter, J. Imperial, J. E. Garrett, B. M. Olivera, H. Terlau, S. Becker, *Protein Expression Purif.* **2006**, *47*, 640.
[23] A. Bax, G. W. Vuister, S. Grzesiek, F. Delaglio, A. C. Wang, R. Tschudin, G. Zhu, *Methods Enzymol.* **1994**, *239*, 79.
[24] S. Talluri, G. Wagner, *J. Magn. Reson. Ser. B* **1996**, *112*, 200.
[25] Y. Xia, A. Yee, C. H. Arrowsmith, X. Gao, *J. Biomol. NMR* **2003**, *27*, 193.
[26] V. Vijayan, M. Zweckstetter, *J. Magn. Reson.* **2005**, *174*, 245.
[27] G. Cornilescu, J. L. Marquardt, M. Ottiger, A. Bax, *J. Am. Chem. Soc.* **1998**, *120*, 6836.

