PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# MOTOR: Model assisted software for NMR structure determination

Ulrich Schieborr,[*,#,○,+] Sridhar Sreeramulu[+], Bettina Elshorst[+], Marcus Maurer[+], Krishna Saxena[#,○,+], Tanja Stehle[+], Denis Kudlinzki[#,○,+], Santosh Lakshmi Gande[#,○,+], and Harald Schwalbe[*,#,○,+]

[+] Johann Wolfgang Goethe-University Frankfurt, Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance, Max-von-Laue-Str. 7, 60438, Frankfurt am Main, Germany

[#] German Cancer Consortium (DKTK), Heidelberg, Germany

[○] German Cancer Research Center (DKFZ), Heidelberg, Germany

ABSTRACT

**Eukaryotic proteins with important biological function can be partially unstructured, conformational flexible, or heterogenic. Crystallization trials often fail for such proteins. In NMR spectroscopy, parts of the polypeptide chain undergoing dynamics in unfavorable time regimes cannot be observed. *De novo* NMR structure determination is seriously hampered when missing signals lead to an incomplete chemical shift assignment resulting in an information content of the NOE data insufficient to determine the structure *ab initio*. We developed a new protein structure determination strategy for such cases based on a novel NOE assignment strategy utilizing a number of model structures but no explicit reference structure as it is used for bootstrapping like algorithms. The software distinguishes in detail between consistent and mutually exclusive pairs of possible NOE assignments on the basis of different precision levels of measured chemical shifts searching for a set of maximum number of consistent NOE assignments in agreement with 3D space. Validation of the method using the structure of the low molecular-weight-protein tyrosine phosphatase A (MptpA) showed robust results utilizing protein structures with 30–45% sequence identity and 70% of the chemical shift assignments. About 60% of the resonance assignments are sufficient to identify those structural models with highest conformational similarity to the real structure. The software was benchmarked by *de novo* solution structures of fibroblast growth factor 21 (FGF21) and the extracellular fibroblast growth factor receptor domain FGFR4 D2, which both failed in crystallization trials and in classical NMR structure determination.**

Proteins 2013; 81:2007–2022.


**Key words: software; protein; NOE; assignment; sparse data; fibroblast growth factor; fibroblast growth factor receptor; algorithm.**

## INTRODUCTION

The development of new NMR and computational methods to speed up NMR structure determination of proteins is still on-going in order to match the requirements of high-throughput protein structure determination projects. One particular important aspect is structure-driven drug design.[1] In structural genomics projects, the detection of a complete set of signals in an initial $^{1}H/^{15}N$ correlation spectrum often becomes a prime criterion for the decision on prioritization of candidate sequences submitted to structure determination pipelines since current NMR structure determination software packages require that almost all nuclei of a protein show signals which can also be assigned.

A variety of promising novel NMR-based methods exists for protein structure determination. They are mostly based on residual dipolar couplings or the correlation between chemical shifts and structure.[2–5] However, to determine high-quality NMR structures, the "classical" protein structure determination approach based on the analysis of NOESY spectra is still indispensable. The essential steps of an NOE-based structure

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Ulrich Schieborr, Johann Wolfgang Goethe-University Frankfurt, Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance, Max-von-Laue-Str. 7, D-60438 Frankfurt am Main, Germany. E-mail: schieborr@nmr.uni-frankfurt.de and Harald Schwalbe, Johann Wolfgang Goethe-University Frankfurt, Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance, Max-von-Laue-Str. 7, D-60438 Frankfurt am Main, Germany; E-mail:schwalbe@nmr.uni-frankfurt.de

Received 26 November 2012; Revised 26 June 2013; Accepted 28 June 2013
Published online 15 July 2013 in Wiley Online Library (wileyonlinelibrary. com).
DOI: 10.1002/prot.24361

determination by NMR are data acquisition, signal identification, chemical shift assignment, NOE assignment, and structure calculation.[4] Current software like UNIO[6,7] or FLYA[8,9] performs the computational steps in an almost fully automated manner. A recent study shows a promising prototype of a protein structure determination procedure that even includes a flexible consecutively optimized data acquisition into the computational procedure without any user intervention.[10] Nevertheless, fully automated structure determination is still demanding, because all steps are potentially error prone, and such errors can accumulate and lead to failure of the entire process.[4]

In this contribution, we focus on the NOE assignment, which is an important bottleneck in NMR structure determination.[2] NOE cross peaks contain implicit information about distances up to about 5–6 Å maximally between two protons in the biomolecular structure. An NOE is described by its intensity and a chemical shift value for each dimension of the spectrum. Together with a previously performed resonance assignment, an NOE can be transferred to a $^{1}H/^{1}H$ distance restraint in structure calculations. This NOE assignment process is a major bottleneck in NMR structure determination[2] and is in practice part of an iterative assignment and structure calculation process.[5]

One central problem of the automated NOE assignment is spectral overlap and limited precision in peak positions making the transfer of an NOE to a distance restraint an ambiguous process for most peaks. Several programs have been developed for the automated NOE assignment[11–21] differing in their way of accounting for ambiguity. Mumenthaler and Brown developed NOAH[18,19] which iteratively calculates structures on the basis of given unambiguous NOEs and a set of selected ambiguous NOEs. Correct assignments are identified from the list of ambiguous NOEs by analyses of preliminary structural ensembles. ARIA[17,20,21] by Nilges *et al.* is based on the definition of ambiguous distance restraints, which allow for assignment of NOEs to more than one proton pair during the structure calculation.[22] The statistical information of known structures to rank ambiguous NOE assignment possibilities was first utilized by Kalbitzer *et al.* They developed KNOWNOE,[11] a Bayesian algorithm based on volume probability distributions derived from a statistical database analysis. ASDP and its predecessor AutoStructure[14] from the Montelione group use a graph theory-based distance network analysis to describe the NOE assignment problem and a statistical analysis[23] to measure the agreement of a specific structure to the NOE data. PASD[16] is an algorithm developed by Kuszewski and Clore designed to be tolerant against ambiguity-related initial NOE assignment errors. In an iterative simulated annealing process, no explicit information of the resulting structure but only likelihoods of specific assignments are transferred to the next iteration step.

Cyana[12] from Güntert and UNIO[6,7] from Herrmann are widely used programs for NMR structure determination. Both programs are based on the NOE assignment strategy CANDID[13] which uses a hybrid scoring function to rank the reliability of any specific NOE assignment consisting of (i) chemical shift differences to the resonance assignment, (ii) the presence of a symmetry-related cross peak, (iii) the compatibility with the covalent topology of the protein, (iv) the compatibility with the structures from the previous iteration cycle, and (v) the "network-anchoring," which is a measure of the fitting of an NOE into the self-consistent NOE network.[13] The different assignment strategies differ in the maximum ambiguity that is tolerated. Common to all assignment strategies, boot-strapping approaches rely on the consistence of local NOE networks in 3D space as a prime criterion for the identification of the structural solution. The strength of these approaches decreases when only an incomplete resonance assignment is available. For CANDID and AutoStructure, it has been shown that at least 85–90% of the resonances of the core region have to be assigned as prerequisite for a reliable NOE assignment leading to a defined protein structure,[24–27] (maximum pair wise backbone RMSD of about $<$3.5 Å).

Our approach presented here was motivated by our experience that a near to complete resonance assignment cannot be reached for many eukaryotic proteins. In particular, we have previously shown that conformational dynamics, for example, of kinases, but also of membrane proteins can be severe and can be different for different parts of the polypeptide chain.[28,29] In addition, a substantial part of all eukaryotic proteins despite playing key roles in biological processes are at least partially intrinsically unstructured.[30] These internal dynamics can lead to a loss of signals in NMR spectra resulting in both, incomplete resonance assignment and NOE distance information.

Therefore, we have developed a computational protocol for the NMR structure determination of proteins with incomplete resonance assignment. The information required to define a protein structure depends on the conditions of the calculation procedure. For example, an algorithm that does not use bond lengths, bond angles, or the geometry of peptide bonds would need much more data than NOEs to converge in one structure. In the other extreme, a pure modeling approach like ROSETTA based on known structures does not need any measured data. MOTOR is an approach somewhere between a conventional NMR structure determination and a modeled structure. It presumes that structural elements of the protein of interest can be predicted correctly in the structural models. In detail, the software combines three strategies to break down the complexity of structure determination by NOE data. (i) The initial ambiguity of the NOE assignment is reduced by the use of homology models derived from known NMR and X-ray structures. (ii) The detailed aspects of different

levels of precision in chemical shift data of NOE peaklists and resonance assignment lists are taken into account to distinguish between consistent and mutually exclusive possible pairs of NOE assignments. (iii) The search for local structural elements with a high number of self-consistent NOE restraints is not an explicit criterion of the algorithm. Instead, the software searches globally for a set of maximum number of NOE assignments that are not mutually exclusive. Cheshire-yapp[31,32] as well as CS-DP-Rosetta[33] also utilizes unassigned NOE lists to find correct structures. But in contrast to our software, the NOE data are used to rank structures that were generated before on the basis of structural libraries. In our method, the information content of the NOE data actively steers the search for the structural solution. We validated our software on the low molecular-weight-protein tyrosine phosphatase A (MptpA), whose structure has been recently solved in our group[34]. About 70% of the resonance assignment is sufficient to solve the structure with a backbone RMSD 1.8 Å to the X-ray structure, when structural models of related proteins are used. Already 60% of the assignment is sufficient to identify those structural models with highest similarity to the correct structure. The developed program was then applied to the structure determination of the fibroblast growth factor FGF21 and the extracellular domain D2 of the fibroblast growth factor receptor 4 (FGFR4 D2), where crystallization tries failed and the achieved assignment was not sufficient to solve the structure by other approaches. We propose to call the new approach "MOdel assisTed sOftwaRe for NMR Structure determination" (MOTOR).

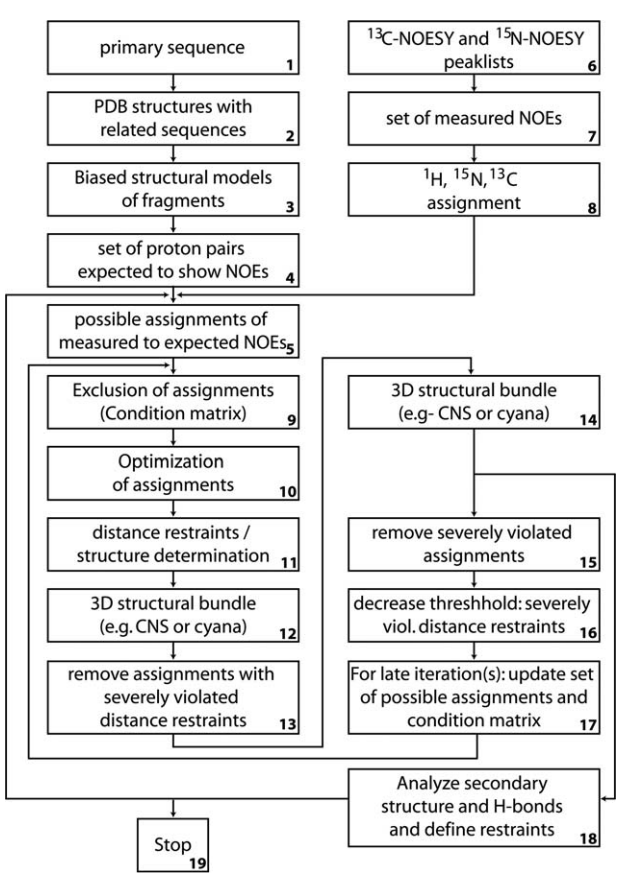



**Figure 1**

Flowchart of the structure determination process by MOTOR based on NMR data and structural models. As input, the primary sequence, peaklists of NOESY spectra and a partial assignment are required.

## MATERIALS AND METHODS

An overview of the structure determination process in MOTOR is schematically depicted in Figure 1. Initial stages (Fig. 1, boxes 1–3) in the structure calculation process are performed with external software. BLAST[35] is used to identify structures deposited in the pdb database[36] with primary sequence similarity to the protein of interest. Structures with a corresponding BLAST bit score higher than ~50 (*E*-values were $\leq 10^{-8}$) are retrieved. Structural models are generated with Modeler[37] by insertion or deletion of amino acids in the primary sequences and exchanging the side chains of the homologous structure to match the primary sequence of the target protein. Structural models, primary sequence, resonance assignment, and integrated peak lists of the measured NOESY spectra (Fig. 1, boxes 6 and 7) are the basis of the MOTOR-steered NOE assignment and structure determination process.

First, NOEs expected from the model structures (in the further text called "expected NOEs") are defined. Proton pairs with a distance up to 6 Å in at least one of the structural models are identified. We set the threshold to 6 Å, because it is about the maximum distance between two protons showing an NOE. This value balances also the required computational time and the completeness of the considered expected peaks for which also a measured peak exists. At the beginning of the calculation, only these hydrogen pairs are considered to show expected NOEs (Fig. 1, box 4). The chemical shift assignment (Fig. 1, box 8) is used to fix those chemical shift ranges in the NOESY spectra, where resonances caused by each proton and heavy atom could be found. In the applications shown here, the corresponding chemical shifts of peaks in the NOESY spectra are restricted to a range of the previously assigned nucleus chemical shift +/−0.6 ppm for $^{15}$N and $^{13}$C resonances and +/−0.06 ppm for $^{1}$H resonances.

The self-consistent assignment of the measured NOEs to expected NOEs solves the structural problem. Depending on the kind of the NOESY experiments, each of the acquired spectra leads to peaks that reflect the distances between members of a specific subset of nuclei. For example, the 3D-$^{1}$H-$^{15}$N-HSQC-NOESY on a fully protonated

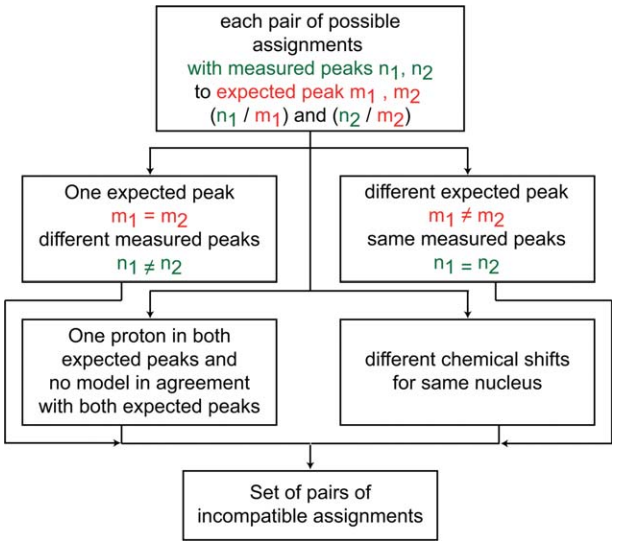



**Figure 2**

Definition of mutual exclusive pairs of assignment. The assignment of a measured peak *n* to the expected peak *m* excludes (i) all assignments of other measured peaks than *n* to the same expected peak *m* and (ii) all assignments of other expected peaks than *m* to the same measured peak *n*. Additionally, (iii) assignments of expected peaks are excluded, which involve at least one nucleus which is also affected by expected peak *m* and the difference in the chemical shift values of this nucleus is larger than a certain threshold. The values for the thresholds depending on the nuclei types, spectra and dimensions are summarized in Table I.

sample leads to distances from an amide proton to any other proton of the protein. A 3D-$^{1}$H-$^{13}$C-HSQC-NOESY spectrum on a fully protonated sample in $D_2O$ shows peaks for distances from an aliphatic proton to any non-acidic proton. In contrast, a 3D-$^{13}$C-NOESY on a perdeuterated and specifically protonated sample does not show peaks for deuterated residues. Therefore, the measured counterpart of an expected NOE can only be found in the corresponding NOESY(s) (Fig. 1, box 5). Additionally, a measured peak can be assigned to an expected NOE, only if the chemical shift of the measured peak is within the shift range of the expected peak and the measured peak volume (relative to the corresponding diagonal peak volume) does not imply a significant shorter distance than the minimal distance in the structural models.

In a next step, mutually exclusive assignments are identified (Fig. 1, box 9). As summarized in Figure 2, two different assignments can be mutually exclusive for different reasons. An expected peak can be assigned to at most one measured peak and one measured peak can be assigned to at most one expected peak. Peak overlap can be an issue in that context. In the current MOTOR version, overlapping undistinguishable diagonal peaks have to be picked twice (or even more often, when more diagonal peaks are on the same position). An additional criterion can be used for cases, when the conformations in the model structures are very diverse and assignments could be made that are mutually exclusive because they come from totally different local conformations in the model structures. When the structural models are fragmental so that none of the models includes the entire environment of a proton, this criterion should be switched off, in later calculations, when the overall structure is already clear to reach a maximum number of assignments. The criterion for chemical shift consistency is different depending on whether the two peaks are found in the same spectrum and if the common nucleus is represented in the same dimension, because the precision of the peak position is dominated in decreasing order of accuracy either by the digital resolution, the accuracy of the peak picker in the indirect dimension or the comparability of different experiments or samples. The used values for consistency of chemical shifts are summarized in Table I. For each pair of an expected peak and a measured peak, one variable is defined which can either be 1 for "assigned" or 0 for "not assigned." We presume that the correct NOE assignment is correlated with a maximal number of assignments that are in agreement with the structure in 3D space (Fig. 1, box 10). In a first step, an arbitrary "allowed," without mutual exclusions, assignment is generated which is then altered to maximize the total number of assignments. Although this mathematical problem is a linear optimization, the large number of both, variables and conditions ($\sim 10^4$ to $10^5$) makes it practically unfeasible to find the exact solution due to limitations of computational resources. Additionally, it is not worth to search for the exact solution, since the compatibility of the assignments in 3D space is not completely taken into account. Therefore, a fast algorithm was applied that iteratively maximizes the total number of NOE assignments. The software searches for subsets of current assignments that, if they would all be removed, would allow making a higher number of other assignments. When such a pattern is found, the exchange of the assignments leads to a higher total number of NOE assignments and is therefore one step forward to the optimum. In a first step, the search is performed removing only one assignment. If the removal of this assignment allows the assignment of

**Table I**

Maximal Difference between the Chemical Shift of Two NOEs Involving the Same Nucleus Depending on Dimensions and Spectra

| | Same spectrum | | |
|---|---|---|---|
| | Same dimension (ppm) | Different dimension (ppm) | Different spectra (ppm) |
| $^{1}$H | 0.02 | 0.05 | 0.05 |
| $^{13}$C | 0.5 | – | 0.7 |
| $^{15}$N | 0.5 | – | 0.7 |

The corresponding difference to the "correct" resonance of the nucleus is half of these values.

**Table II**
Parameters of Each Simulation Comprising Box 9–17 in Figure 1

| Cycle index | Dihedral angle restraints Secondary structure | Distance restraints H-bonds | Distance restraints Unassigned loops | Model Structures |
|---|---|---|---|---|
| #1 | Talos+ | – | – | Pdb |
| #2 | Results: #1 | – | – | Pdb |
| #3 | Results: #2 | H-bonds #2 | – | Pdb |
| #4 | Results: #3 | H-bonds #3 | – | Pdb |
| #5 | Results: #4 | H-bonds #4 | Loop models | Pdb/results #4 |
| #6 | Results: #5 | H-bonds #5 | Loop models | Pdb/results #4 |

two or more other assignments, then the exchange is performed. If not, the program tries the same with the next variable and so on until all assignments were checked. Then the same procedure is performed with two assignments at the same time, implying that if any removal of two assignments allows for more than two new assignments, then the assignments are exchanged. The computational time required to find any exchange pattern increases with the number of assignments that are exchanged in one step. Therefore, the maximum number of assignments that are removed is restricted. With a current personal computer (Intel® Xeon®, quad core, 8 CPUs, 2.33 GHz running Linux), we found a maximum number of eight assignments for removal as reasonable balancing the required computational time and the completeness of the optimization. The algorithm could effectively be implemented on parallel processors.

After the near to optimal solution has been found, the given assignment is transformed to 3D structures by conventional structure determination software with a fixed set of distance restraints defined by the NOE assignment (Fig. 1, box 11,12). In the examples reported here, the basic tool of Cyana[12] with fixed distance restraints was used. The distances in the calculated 3D structures are then compared to the restraint distance values. Assignments corresponding to severely violated distances in the structures are removed from the list of possible assignments (Fig. 1, box 13). Another structure calculation without the severely violated distances is performed which delivers a new set of distances (Fig. 1, box 14). Again, severely violated distance restraints are removed. The next iteration circle (Fig. 1, box 9) starts with the current assignment for further optimization. The threshold for the definition of severely violated distances is decreased from 6 Å in the first iteration to 1 Å in the seventh iteration step. The cycle of boxes 9–17 in Figure 1 is run through seven times. In the reported applications here, five (MptpA) to six (FGF21 and FGFR4 D2) of such calculations were performed, each based on results of the previous calculation. In each simulation, additionally to the NOE based distance restraints, hydrogen bond and dihedral angle restraints were introduced as summarized in Table II. In the first calculation, dihedral angle restraints were applied originating from the software Talos+.[38] It predicts secondary structure elements from the chemical shift values of the resonance assignment. For residues predicted by Talos+ to form a β-strand conformation, the backbone dihedral angles were restrained to $-169° < \varphi < -109°$ and $105° < \psi < 165°$. Residues with predicted α-helical conformation were restrained to $-90° < \varphi < -40°$ and $-70° < \psi < 10°$. In the following calculations, dihedral angle restraints were applied, when a α-helical or a β-strand conformation was confirmed in the previous calculation result. A secondary structure element in one specific residue was defined as confirmed when either at least 50% of the structural ensemble show $\varphi/\psi$ values located in the corresponding most favored area of the Ramachandran plot or all structures lie in the favored area. Distance restraints justified by hydrogen bonds found in the previous structures are introduced starting from the third calculation. A hydrogen bond was defined as formed when it is present in at least 50% of the structural ensemble and in at least one of the model structures. In the fifth and sixth calculation for FGF21 and FGFR4 D2, additional restraints were applied to stabilize the conformation of unassigned loops as described below.

## MptpA calculations

Forty-six experimental structures (Supporting Information Table S1) from the pdb database with the highest similarity to MptpA in their primary sequence were used to form structural models of MptpA using the software MODELLER[37] [Fig. 3(c)]. The experimental MptpA structures were excluded from the set of structural models.

Structure calculations were performed using 100, 90, 80, 70, and 60% of the resonance assignment. This was achieved by removal of the respective ratio of assignments of entire residues equally distributed over the primary sequence. Each calculation was performed in five consecutive calculation steps with parameters summarized in Table III. Calculation one to four corresponds to step 1–4 in Table II. In a fifth calculation, only the structures resulting from the previous calculation and the model structures with lowest RMSD to those structures are used as structural models for the NOE assignment.

Three additional calculations were performed as control experiments. To show the influence of sequentially missing resonance assignments in whole loops of the protein, MOTOR was run after removal of five loop regions each up to 12 residues long leading to a resonance assignment of 72%. As further control, we used a set of model structures which are all uncorrelated to the real structure of MptpA. In a last calculation, we used one reasonable model structure (backbone RMSD to MptpA X-ray structures 1.9 Å, NMR structures 2.1–2.3 Å) in addition to the models of the former calculation to

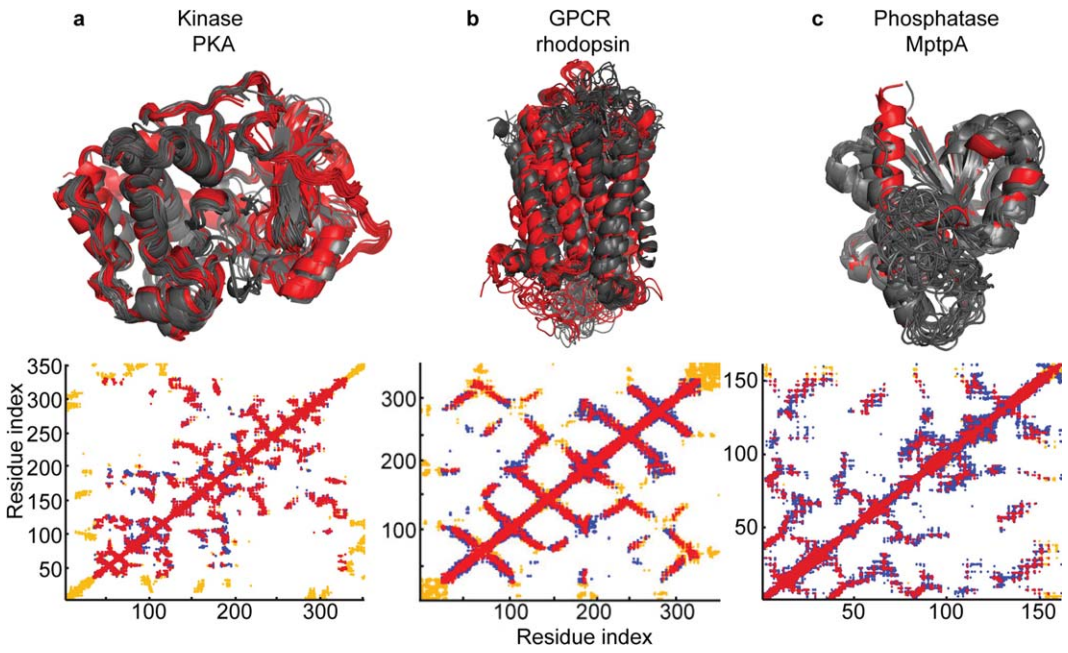



**Figure 3**

Top: Alignment of published pdb structures (red) to MODELLER-generated homology models of these proteins. Bottom: Contact maps of pairs of protons with a distance lower than 6 Å. Red dots mean that the contact is found in both, real structures and models. Blue dots indicate contacts found in at least one model structure but in none of the real structures. Contacts found in at least one real structure but none of the model structures are symbolized as orange dots. Pdb files with a sequence similarity higher than 95% were chosen as real structures irrespectively if they are apo or complex structures. (**a**) Application to human cyclic AMP dependent protein kinase A (PKA). The model structures have sequence identity of 38–50%. (**b**) Application to Rhodopsin with 33 real structures and 60 model structures with 21–27% sequence identity. (**c**) Application to MptpA with two X-ray structures and 12 model structures with 30–43% sequence identity.

check if MOTOR is able to distinguish between reasonable and non reasonable models.

### *De Novo* MOTOR structures

The structure determination of both proteins was performed on the basis of $^{15}$N- and $^{13}$C-3D-NOESY spectra. An additional 2D-NOESY for aromatic protons was available for FGF21. NOE cross-peaks and non-diagonal peaks were obtained by manual peak-picking in all NOESY spectra. The detailed parameters of the structure calculation are summarized in Table III. In the respective first calculation steps, 44 (FGFR4 D2) and 54 (FGF21) Talos+[38] based $\varphi/\psi$ dihedral angle restraints were introduced. In the following calculations, dihedral angle restraints were only applied, when a secondary structure element could be confirmed in the previous simulation. In the sixth simulation step 60 $\varphi/\psi$ dihedral angle restraints could be confirmed for FGFR4 D2 and 80 for FGF21.

Beginning from the third calculation, H-bonds present in at least one model structure are restrained, when they were present in more than 50% of the structures in the previous calculation. In the last calculation, 21 H-bonds were found for FGFR4 D2 and 40 for FGF21. In the fifth and Sixth calculations, large unassigned parts were modeled based on homologue structures. For FGFR4 D2, the loop from Ile 200 to Trp 207 is modeled by the introduction of four H-bonds, which are present in this region for FGFR1, FGFR2, and FGFR3, and 8 $\varphi/\psi$ restraints based on the values in the related receptors. For FGF21 the loop between Pro 147 and Ala 157 is completely unassigned due to severe line broadening, while residues before and after this loop show resolved NOESY peaks indicating a defined structure for those regions. The models for this region of FGF21 are significantly different depending on using either FGF19 or FGF23 the two most related molecules as a template. For FGF19 a X-ray structure (PDB ID: 1PWA) exists that even lacks structural information for this region. Anyhow, the spatial position of the start and the end of the loop is about the same for FGF19 and FGF23. For the last two refinement calculations, we decided to model the loop (Pro 147–Gly 160) with weak restraints to the local conformation observed for FGF23 (PDB ID: 2P39). The distances between N, $C_\alpha$, and $C_\beta$ atoms within the loop were determined in the structural model derived from FGF23. These distances were restrained with a tolerance of 1 Å in the last two cycles. We do not have experimental evidence for this region. The loop could also look like the

**Table III**
Calculation Parameters

| General parameters | |
|---|---|
| Number of NOE assignment cycles | 7 |
| Maximal NOE distance | 6 Å |
| Reference to structures of previous cycle | Only in cycle 5 |
| H2/H3 proton resonance assignment swapping | All cycles |
| Conditions for considered NOE assignments | |
| Max(minimal model distance - intensity distance) | Cycle 1: 1 Å<br>Cycle 5: $\infty$ |
| Max \|assignment chem. shift - NOE chem. shift\| | 0.06 ($^{1}$H), 0.6 ($^{15}$N), 0.6 ($^{13}$C) |
| NOE assignment optimization | |
| Maximum number of removed assignments in one substitution | 8 |
| Number of CPUs used | 8 |
| Structure generation | |
| Upper limit weighting (for each cycle) | 0.3, 0.5, 0.6, 0.8, 0.9, 1 |
| Number of structures for NOE violation | 20 |
| Max NOE violation to be an accepted NOE in a structure (for each cycle) | (6, 4, 3, 2, 1.5, 1) Å |
| Minimum number of structures with unaccepted NOE to remove NOE in following cycle | 6 |

homologue in FGF19 or even completely different. Therefore, we left out Pro 147–Gly 160 for the analysis. For further refinement, the resulting structures of FGF21 and FGFR4 D2 were analyzed using the software package Cing.[39] The found secondary structure elements were checked for residues which were marked as suspect by Cing, If the suspect local conformation was probable to be induced by a wrong dihedral angle of hydrogen bond restraint, the corresponding restraint was removed or altered. In a trial and error approach, MOTOR calculations were performed using the modified set of restraints until no further improvement of the structure was possible and Cing reported of high quality (green) structures.

## RESULTS

### Structural models

One aspect of MOTOR is to ease the NOE assignment by a decrease of its ambiguity due to the use of homology models. In an academic context, structure determination is most attractive, when no structural homologous exist. On the other hand, high resolution structures, including detailed side chain information, are important in structure-driven drug development programs. Figure 3 (top) illustrates known pharmacologic interesting experimental protein structures (red) and their homology models (gray) using sequence-related pdb structures generated by the program MODELLER.[37] The cyclic AMP-dependent protein kinase PKA (a), rhodopsin (b) and the low-molecular-weight protein tyrosine phosphatase MptpA (c) are depicted as examples for pharmacological relevant target structures. The sequence identity of the selected models to the target structure is between 21 and 50%. The positional RMSD of the backbone atoms of the models to the real structures is 1.1–7.5 Å for PKA, 1.8–7.8 Å for rhodopsin, and 1.9–5.3 Å for MptpA, respectively. The GDT-TS values (using only the $C_\alpha$-positions and the average over the thresholds 1, 2, 4, and 8 Å) between the models and the real structures were between 0.51 and 0.91 for PKA, 0.22 and 0.79 for rhodopsin, and between 0.38 and 0.78 for MptpA. On the bottom of Figure 3, the respective contact maps of pairs of protons with a distance lower than 6 Å are depicted. Despite the high positional RMSD values between the models and the experimental structures, 87.4% (PKA), 76.2% (rhodopsin), and 92.7% (MptpA) of all contacts of the different experimental structures are found also in at least one model structure. However, the ratio of proton/proton contacts identified in the experimental structures relative to all (by permutation) possible proton pairs is only 2.5% (PKA), 6.6% (rhodopsin), and 4.6% (MptpA). The ratio of the considered contacts which are also present in the experimental structure relative to all considered contacts for the NOE assignment procedure can be substantially increased by incorporation of homology models. We find that the ratio is enriched by a factor of 33.4 (PKA), 11.9 (rhodopsin), and 11.1 (MptpA) when homology models are used.

We statistically checked whether models have already been used in NMR structure determination. Figure 4(a) shows a statistical analysis of molecules deposited in the pdb database[36] that are solved by both methods NMR and X-ray. The number of molecules relative to the number of all analyzed molecules is shown as a function of the backbone RMSD between the respective NMR and X-ray structures. RMSD values lower than 0.9 Å are almost exclusively observed for molecules that were first solved by X-ray (blue bars). The NMR data are in agreement with the X-ray structure for these molecules. Despite an agreement between experimental data and the correct structure, deposited NMR structures can be significantly different, since additional structural models can also be in agreement with the NMR data. In the absence of a gold standard, the selection between possible models only based on NMR can be difficult. We conclude that structural templates help to interpret the NMR data and templates have already been used in the NMR community-probably not always in a transparent and reproducible way. As kind of the same idea in the opposite direction, the usefulness of high quality NMR structures as templates for molecular replacement with X-ray data has already been shown.[40] Interestingly, NMR structures that fail to provide good enough models for

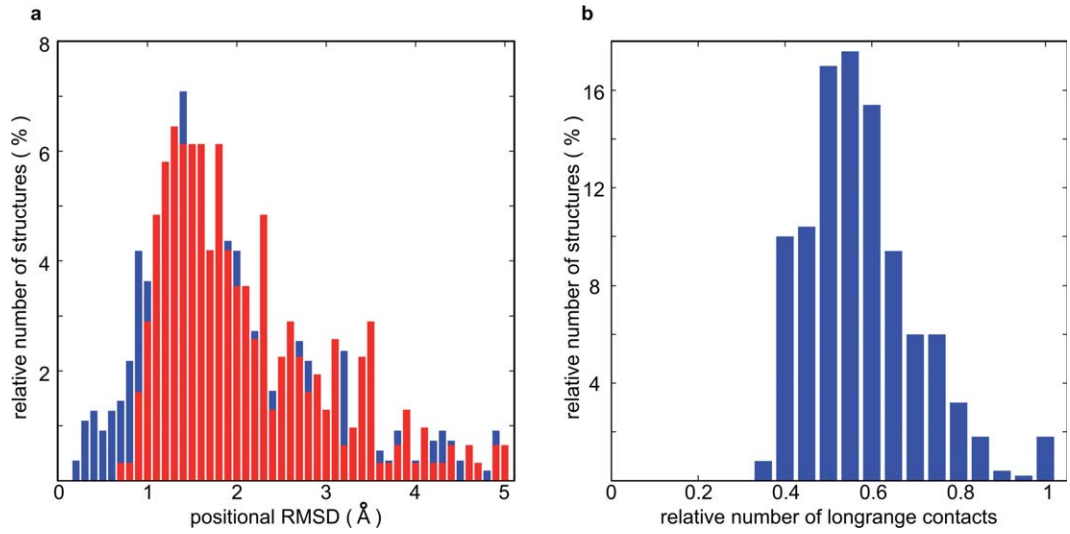



**Figure 4**

Statistics over the molecules deposited in the pdb database solved by both NMR and X-ray. (**a**) The percentage of molecules relative to all analyzed molecules is depicted as a function of the backbone RMSD between the NMR and the X-ray structures of each molecule. RMSD ranges of 0.1 Å are taken together in one bar, respectively. The distribution for molecules that were first solved by X-ray (primary sequence similarity higher than 95%) is depicted in *blue*. Structures first solved by NMR are depicted in red. (**b**) Percentage of molecules relative to all analyzed molecules as a function of the completeness of expected long-range NOEs. The number of low distance (<3 Å) long-range (primary sequence distance ≥5) $^{1}H/^{1}H$-contacts showing NOEs relative to all of these low distance long-range contacts expected from the published NMR structures is used as a measure for the completeness of the expected NOEs. Only protons with known chemical shifts are considered.

(X-ray) molecular replacement can be improved by a Rosetta refinement of the NMR structures,[40,41] which implicitly utilizes X-ray data, because a structural library is used that is dominated by X-ray structures.

## MptpA: validation of MOTOR

Our newly developed program MOTOR was validated on MptpA, for which experimental X-ray[42] and NMR[34] structures as well as almost the complete resonance assignment are available. The calculations were performed using 100, 90, 80, 70, and 60% of the resonance assignment. Figure 5(a) shows a comparison of the published X-ray and NMR structures (sand) and the result of the MOTOR calculation (green) using the full resonance assignment. Residues of published structures are shown in *white,* when the B-factor in the corresponding $C_\alpha$ atom is higher than 20 $Å^2$ in the X-ray structure 1U2P. Residues Leu 100 to Arg 106 as well as Arg 116 to Ala 121 and Tyr 129 to Ser 133 show a deviation of the backbone conformation between the published X-ray and NMR structures. Interestingly, the loops Gly 118 to His 120 and Tyr 129 to Ser 133 show increased B-values in the backbone conformation which might be a hint for local flexibility. Those residues were left out for further RMSD calculations. Figure 5(b) shows the MOTOR structural ensemble grayscale coded by the number of assigned NOEs per residue. A high number of NOEs per residue all over the protein indicates a well defined molecule and high reliability of the calculated structures.

Figure 6(a,b) illustrates the precision of the calculations by showing the positional RMSDs of the backbone (a) and all heavy atoms (b) of the resulting MOTOR structures, when only a partial resonance assignment is used. With model-unassisted NOE assignment strategies like CANDID or AutoStructure, no defined structure [like depicted in Fig. 8(a)] can be observed with resonance assignment less than ~85 to 90%. With MOTOR, the structure is retained down to an RMSD of 1.8 Å for the backbone and 2.7 Å for all heavy atoms even with 70% resonance assignment indicating high precision even with sparse assignment data. The seemingly precision reflected by the bundle RMSD of an ensemble as outcome of one calculation is even higher than the precision that results from comparison to the structures with complete assignment. But, artificial high precision in NMR structural ensembles is a known and general phenomenon of NMR-based structure determination.[5,43,44] During an NOE assignment process, the current assignment corresponds to a preliminary structure which is frequently updated. Since the search for additional NOEs is influenced by this structure, additional assignments further fix the preliminary structure the more NOEs have already been found. This iterative process can lead to seemingly high precision by an over-interpretation of the NOE data content. However, we expect that this effect is

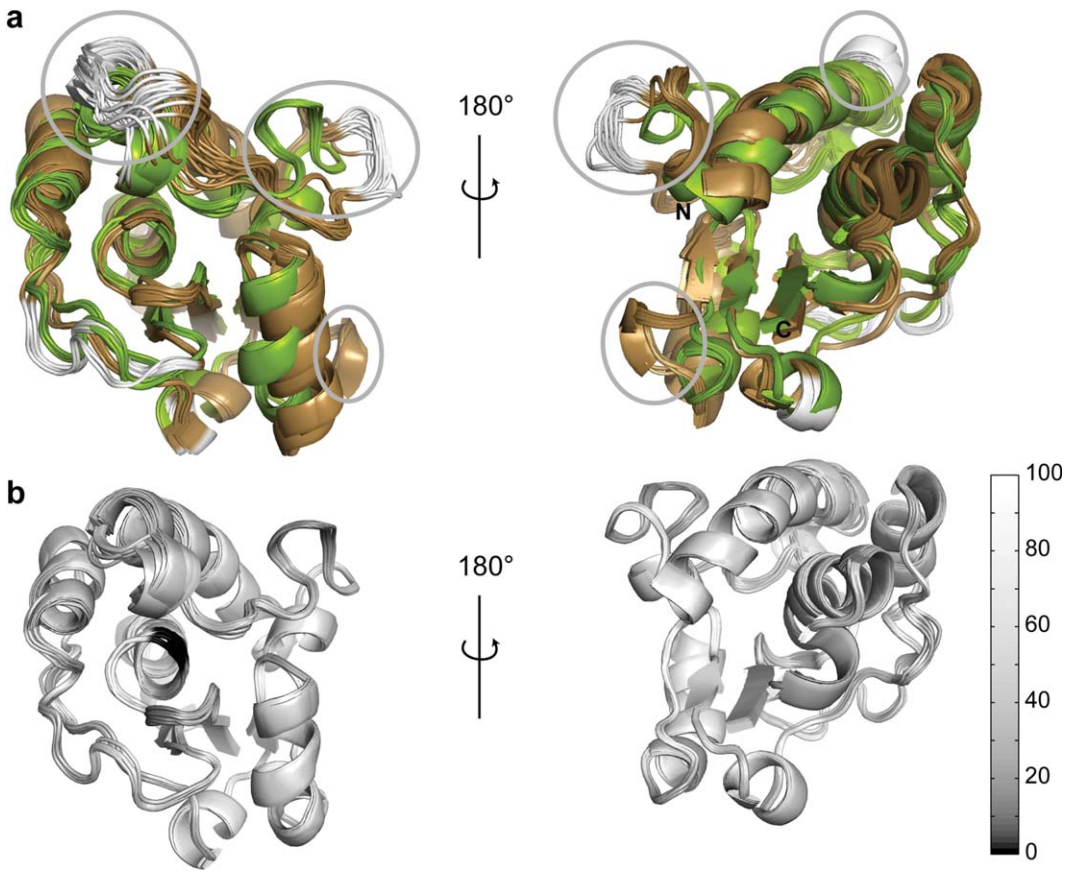



**Figure 5**

MOTOR structure results on MptpA. (**a**) MOTOR structural ensemble with full resonance assignment (green) in comparison to the published structures solved by X-ray and NMR spectroscopy (sand, PDB ID:s 1U2P, 1U2Q, and 2LUO). Residues of published structures with higher B-factors than 20 $Å^2$ in the $C_\alpha$ carbon atom of 1U2P are colored in white. Areas excluded from the RMSD calculations are marked with gray circles. (**b**) MOTOR structural ensemble grayscale coded by the number of NOE assignments per residue.

not pronounced with MOTOR compared to NOE assignment strategies that typically rely on the local structure, because in MOTOR the exchange of sets of assignments during the NOE optimization phases is not steered by the local NOE networks or by a specific reference structure but only by the mutual exclusions.

Figure 6(c) illustrates the correctness of MptpA MOTOR structures by low positional RMSDs of the best MOTOR structure to published structures. Backbone RMSDs of 1.6 Å (X-ray) and 1.7 Å (NMR) and a heavy atom RMSD of 2.3 and 2.4 Å indicate that MOTOR structures are in agreement with published data which utilize a broader set of data for their structure determination processes. Also with MOTOR, any kind of additional restraint can in principle be incorporated during the structure calculation steps which would lead to a further precision and correctness of the structure. Interestingly, the RMSDs from some of the structural models to the X-ray structure are even lower than the RMSD of the published NMR structure to the X-ray structure. It remains speculative if the similarities between structures based on the same technique are due to the same environments (NMR: liquid or X-ray: crystal) or due to similar technical artifacts. Therefore, it is difficult to set one of the experimental structures as the gold standard or real structure.

The comparison to the best MOTOR structures [Fig. 6(a,b)] reports about the precision of the calculation, the comparison to the published structures reports about the correctness [Fig. 6(c)].

Figure 6(d) illustrates the similarity of the MOTOR structures using different extents of the resonance assignment to the model structures used for these calculations, the X-ray structures and the best ranked MOTOR structure using the complete resonance assignment. Though, some model structures have even lower RMSD values to the X-ray structure than the published NMR structure, the MOTOR structures do not simply adapt to one of the model structures. For 70–100% assignment, the RMSDs of the MOTOR structures to the best ranked

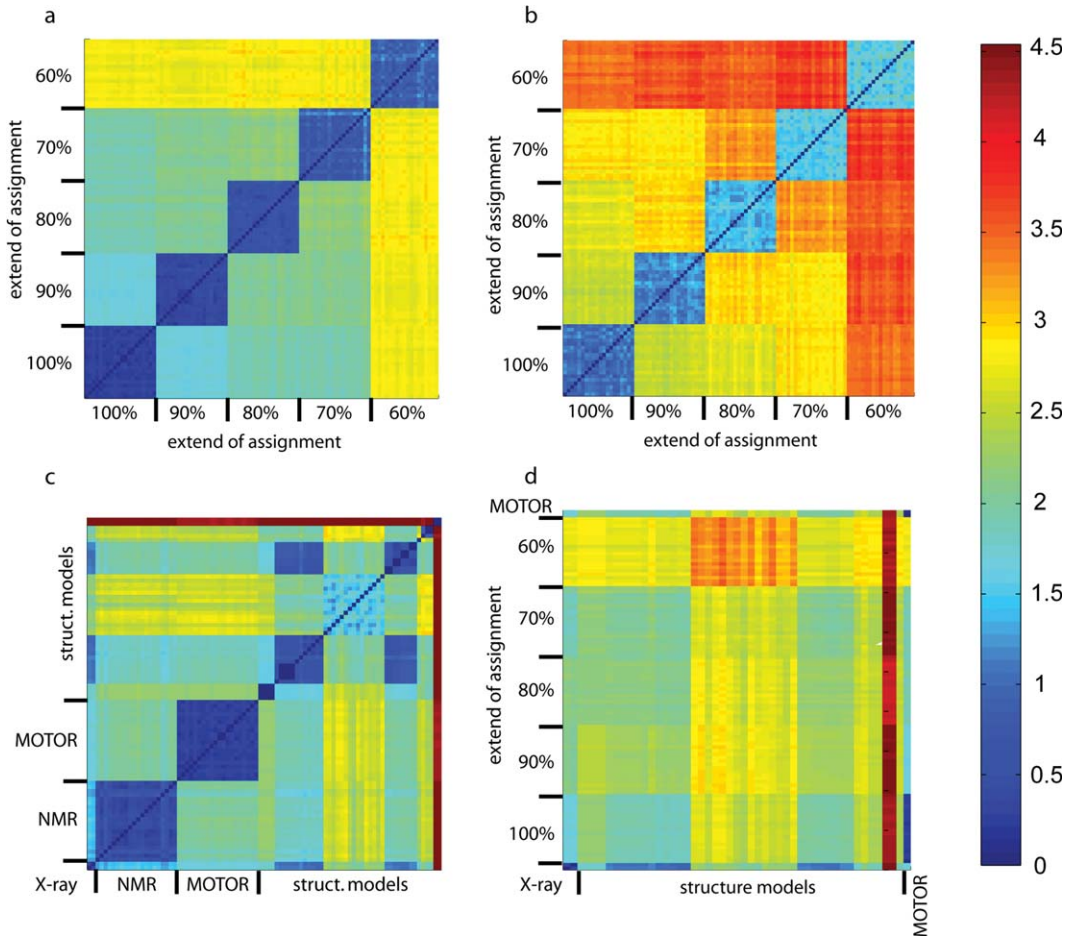
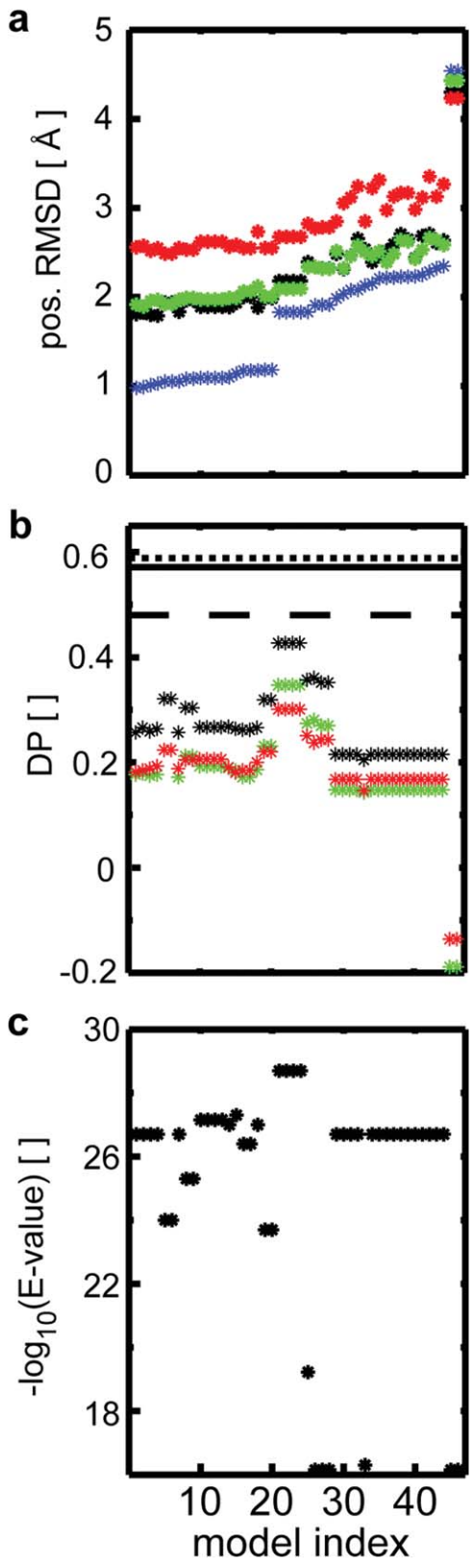



**Figure 6**

Comparison of MptpA structures derived from different data and algorithms as color-coded by positional RMSDs. (**a**) Positional backbone and (**b**) heavy atom RMSDs between MOTOR structures generated with 60–100% of the resonance assignment. (**c**) Published X-ray (PDB IDs: 1U2P and 1U2Q) and NMR (PDB ID: 2LUO) structures compared to the structural models and the MOTOR structure using the full resonance assignment. (**d**) Comparison of the utilized structural models, the X-ray structures and the "best" MOTOR structure to the MOTOR structures using 60–100% of the resonance assignment. Models with an RMSD higher than 4.5 Å to the X-ray structures are not depicted in d.

MOTOR structure are lower than the RMSDs of any of the models used to the best MOTOR structure. Hence, the NOE assignment process does not simply adapt the assignment to one of the models but finds an independent solution. The analysis with GDT-TS values shows the same. The maximal GDT-TS value of the MOTOR structures to the MOTOR structures with full assignment decreases from 1 to 0.8 with 70% resonance assignment. This is still slightly more precise than the used model structures, which have a maximum GDT-TS value of 0.77 to the MOTOR structures with full assignment. Utilizing 60% of the assignment, the RMSD of the MOTOR structures and the structural models are comparable [Fig. 6(d)]. The GDT-TS value even decreases to 0.64, which is significantly less precise than the best model structure. Nevertheless, the RMSDs between 60% level MOTOR structures and the used structural models are minimal for those structural models that show the lowest RMSD to the NMR structure with full assignment. Therefore, even with 60% assignment, it is possible to distinguish between models that are closer or far away from the real structure. Figure 7 illustrates the strength of MOTOR to identify good structural models by the comparison of different indicators as measure for the model quality. Figure 7(a) shows the mean positional RMSD of the structural models to the X-ray and the MOTOR structures, respectively. The correlation of these two sets of RMSDs is 0.98 with 100% assignment 0.96 with 70% assignment and 0.95 with 60% assignment. In principle, the identification of good models can also be achieved by the use of RPF scores.[23] Figure 7(b) illustrates the DP scores from an RPF-analysis of the structural models and the given NOE data leading to significant lower correlation (0.73 for 100% assignment) to the RMSD of the models to the X-ray structure. The corresponding DP values of the X-ray structure (0.48), the published NMR




structure (0.59), and the MOTOR structure (0.57) are depicted as black lines. As expected, the NMR and X-ray structures show significantly higher DP scores than all the model structures. The DP score is lower for the X-ray structure than it is for the MOTOR structure, which is a hint for the significance of differences between the crystal and the MOTOR determined solution structure. Figure 7(c) shows the blast *E*-values of the model structures, which correlate poorly [0.44 to $\log_{10}$(*E*-value)]. Therefore, the use of the model derived from a structure with highest primary sequence similarity also does not guarantee a good model structure. Additionally, MOTOR delivers valuable information also at the 60% level of assignment. The detailed interpretation of the NOE data can serve as the basis for further NMR experiments like complex formation or protein/ligand binding experiments.

MOTOR takes differential tolerances into account, between the NOEs and their listed resonance chemical shifts as well as between two NOEs depending on spectra and dimensions of the NOEs. Other software takes only a match tolerance value between NOEs and the assigned resonance chemical shifts into account. But the tolerance in the measure of the chemical shift values within the NOESY spectra is smaller than the tolerance of the NOE positions to the resonance chemical shifts from the assignment spectra. As a consequence also pairs of NOE assignments within the match tolerance are still allowed, although they are mutual exclusive, when only one tolerance value to the resonance assignment is used. Nevertheless, in practice the use of RPF scores to preselect models for MOTOR calculations would probably accelerate the process at this stage of assignment.

As a negative control, a MOTOR calculation was performed exclusively with model structures uncorrelated to the real structure of MptpA. It resulted in a conformation with defined secondary structure elements. The relative orientation of these elements, however, is different to that in the real structure. For the reliability of MOTOR structures, a criterion is required that can distinguish between reasonable results and structures that are likely to be wrong. In principle, RPF values[23] could be used for that, but a DP score of 0.41 for the given example of a wrong structure was in the same range or even higher than the good model structures. This is probably due to

**Figure 7**

Measure of the quality of each model by different indicators. The structural models of MptpA are ordered in increasing order of positional RMSD to the X-ray structure on the *x*-axis. (**a**) Positional backbone RMSD of the X-ray structure (blue) and the MOTOR structures to the structural models. MOTOR structures derived from 100% resonance assignment are visualized in black, from 70% assignment in green and 60% in red. (**b**) DP scores from an RPF-analysis of the structural models and the NOE data. For comparison, the corresponding DP scores of the X-ray structure (dashed), the MOTOR structure (solid), and the published NMR structure (dotted) are depicted as black lines. (**c**) *E*-values of the primary sequences of the pdb-sources of the model structures compared to the MptpA primary sequence.

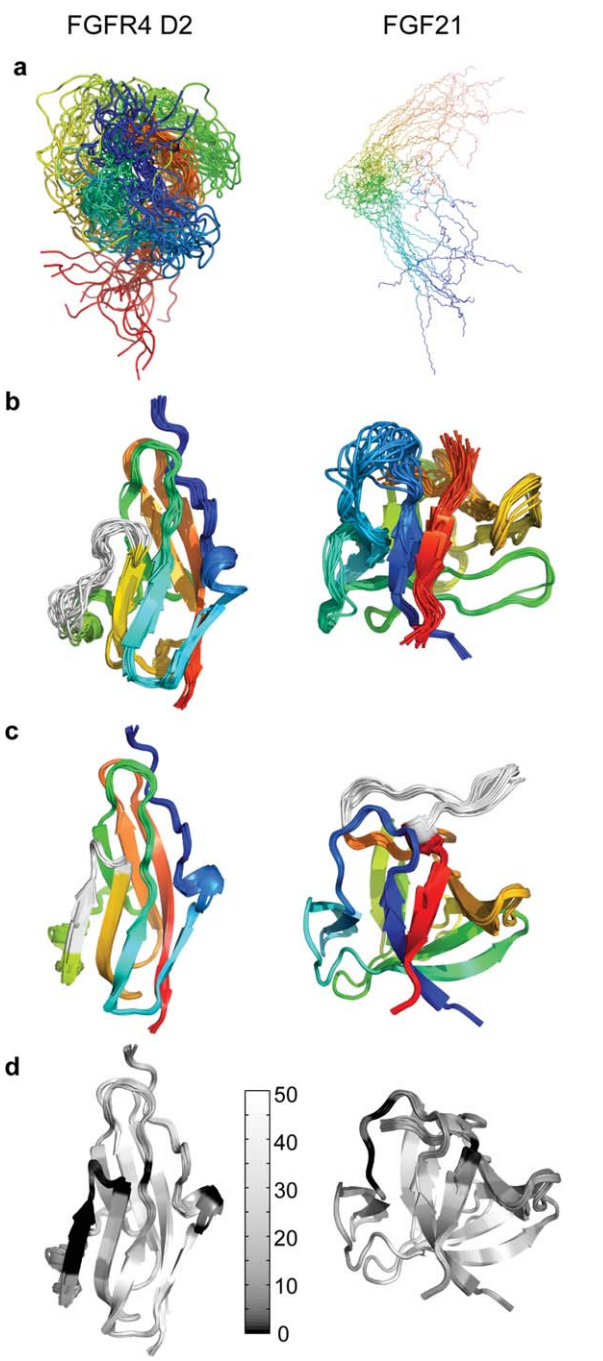



**Figure 8**

Cartoon representation of an ensemble of 20 conformations representing the structure of FGFR4 D2 (Ser 143–Leu 242, left) and FGF21 (Gly 36–Pro 178, right). The color in (**a**), (**b**), and (**c**) represents the chain course from the N-terminus in blue to the C-terminus in red. (a) Result of CANDID as automated NOE assignment strategy using Cyana. (b) MOTOR after first iteration (c) and (d) MOTOR after sixth iteration. The modeled parts Pro 147–Gly 160 for FGF21 and Ile 200–Trp 207 for FGFR4 D2 are colored white in b and c. The gray-scale coding in d represents the number of NOE assignments per residue as defined by the grayscale bar. The modeled loop of FGF21 Pro 147 to Gly 160 is not depicted in b and d.

different levels of peak amplitudes in difficult proteins which require peak picking close to the noise. That brings also erroneous picked noise spikes into the peak list and makes it difficult to determine the number of false positives. Instead we used a value similar to the "Recall" value in the RPF method which does not need the number of false positive peaks. It describes the reproducibility of the tertiary structure by the NOE assignment. Secondary structures are built also with wrong model structures, because the local secondary structure already defines the distances between protons close in the residue chain. These distances are independent of the relative orientation of the secondary structure elements and therewith the tertiary structure. Therefore, the long-range NOE assignment is important for the tertiary structure check. Long-range NOEs (at least five residues distance in primary sequence) define the tertiary structure by fixing the relative orientation of structural elements from different parts of the polypeptide chain. Potential long-range contacts are also found in wrong structures but consistency of the correct fold with the NOE data should result in a higher amount of long-range NOEs. Therefore, the BMRB database was analyzed with respect to the ratio of the set of assigned long-range NOEs relative to the number of long-range low distance $^{1}H/^{1}H$ distances in the corresponding published NMR structures. Those NMR structures of Figure 4(a) are used that are primarily defined by NOEs and have a backbone RMSD to the X-ray structure lower than 2.5 Å. The results are depicted in Figure 4(b). The ratio of identified expected long-range contacts varies but for almost all structures it is over 0.35 and for none of the structures under 0.3. The MptpA structures with 70–100% assignment had ratios between 0.35 and 0.41, which is in agreement with the published structures. In contrast, the calculation with only non reasonable models results in an NOE ratio of 0.24 indicating a wrong structure.

With a second control we checked if MOTOR is capable to find the correct fold when the correct fold is only one model under lots of wrong folds. Therefore, a simulation with the same models as the negative control and additionally one reasonable model (backbone RMSD: X-ray structures 1.9 Å, NMR structures 2.1–2.3 Å) was performed. Already after the first iteration of the cycles in Figure 1 boxes 9–17, the resulting structure was close to the calculation with 100% resonance assignment using all models (backbone RMSD to the calculation with all reasonable models 1.8 Å, RMSD to X-ray 2.4–2.7 Å, RMSD to NMR 2.3–2.7 Å), which indicates that MOTOR is capable to identify the correct models out of a set of reasonable and not-reasonable models. In a prospective structure determination, the residual cycles would then be performed only with the reasonable model.

Due to the resonance assignment strategies, the unassigned regions of a protein are often clustered in loop regions. With another control experiment we checked if

**Table IV**
Structural Statistics of FGF21 and FGFR4 D2 (20 Structures)

| | FGFR4 D2 | FGF21 |
|---|---|---|
| *Construct* | Asn138–Leu242 | His29–Gly189 |
| *Sequential assignment statistics* | | |
| N | 76.2% | 76.4% |
| $H_N$ | 82.4% | 84.7% |
| $C_\alpha$ | 90.5% | 84.4% |
| $H_\alpha$ | 88.5% | 81.9% |
| $C_\beta$ | 87.6% | 84.8% |
| $H_\beta$ | 66.3% | 57.3% |
| $C_\gamma$ | 61.1% | 53.4% |
| $H_\gamma$ | 32.8% | 25.3% |
| *NOESY spectra* | $^{13}C$, $^{15}N$ | $^{13}C$, $^{15}N$, $^{13}C$ aromatic |
| *Calculation conditions* | | |
| Model PDB IDs | 1CVS, 1DJS, 1E0O, 1EV2, 1EVT, 1FQ9, 1II4, 1IIL, 1NUN, 1RY7, 1WVZ, 2FDB, 3CAF, 3CU1, 3DAR, 3EUU, 3GRW, 3OJ2, 3OJM, 3OJV | 1AFC, 1BAR, 1BFF, 1E0O, 1FDB, 1FQ9, 1G82, 1IHK, 1IJT, 1K5U, 1NUN, 1PWA, 1QQK, 1RG8, 2J3P, 2P23, 2P39, 2Q9X, 2UUS, 3B9U, 3CQA, 3CRH, 3F1R, 3FJK, 3HAL |
| *Upper/lower distance restraints* | | |
| Disulfide bridge | (3×) 1 | (3×) 1 |
| | 8 $\varphi/\psi$ (Ile200–Trp 207) | |
| Restraints within modeled parts | 3 H-bonds | 459 (Pro 147–Gly 160) |
| *Found structural elements* | | |
| *hydrogen bonds* | 28 | 25 |
| φ angle | 86 | 103 |
| ψ angle | 94 | 111 |
| *NOE assignment results* | | |
| *Upper distance restraints* | | |
| Intra residual ($i - j = 0$) | 496 | 405 |
| sequential ($\|i - j\| = 1$) | 369 | 281 |
| medium range ($1 < \|i - j\| \leq 4$) | 160 | 137 |
| Long-range ($\|i - j\| \geq 5$) | 594 | 481 |
| *Structure generation* | | |
| Cyana target function *f* | | |
| Number of upper limit violations | 152–154 kcal/mol | 83.4–84.8 kcal/mol |
| >1 Å | 30 | 9 |
| Max. upper limit violation | 1.9 Å | 1.5 |
| Number of $\varphi/\psi$ violations > 20° | 14 | Å4 |
| Max. $\varphi/\psi$ violation | | |
| *Close contacts* Heavy atoms | 38° | 35° |
| (<2.2 Å) | 0 | 1 per model |
| Hydrogen (<1.6 Å, all H–O) | 1–2 per model | 1 per model |
| *RMSD from average* | Ser 143–Leu 242 | Gly 36–His 145 Pro 161–Gly169 |
| Backbone | 0.15–0.3 Å | 0.23–0.34 Å |
| Heavy atoms | 0.56–0.85 Å | 0.54–0.62 Å |
| *Ramachandran analysis (20 structures)* | Ser 143–Leu 242 | Gly 36–Gly169 |
| Most favored regions | 84.3% | 86.9% |
| Additional allowed regions | 14.2% | 12.7% |
| Generously allowed regions | 1.5% | 0.4% |
| Disallowed regions | 0.0% | 0.0% |

MOTOR is able to find the correct structure also for these cases. The assignment of residues 1–5, 44–55, 66–72, 115–126, and 157–163 were removed leading to a MOTOR calculation with an assignment ratio of 72%. The resulting structure has a backbone RMSD to the best MOTOR structure with the full assignment of 1.6 Å (heavy atom RMSD 2.2 Å, backbone RMSD to X-ray 2.0–2.3 Å, and RMSD to NMR 2.1–2.7Å), which indicates that the structure is retained also with clustered unassigned loop regions.

## *De Novo* MOTOR structures

We put our MOTOR approach to the test by solving previously unknown structures of two potential drug targets, FGF21 and FGFR4 D2. Here, we report the methodological aspects, while the discussion of the biological implication of the structures will be published elsewhere.

FGFR4 D2 is an important extracellular domain of the tyrosine kinase receptor FGFR4.[45] Up to now, the NMR structure of the D2 domain of FGFR2,[46] and several

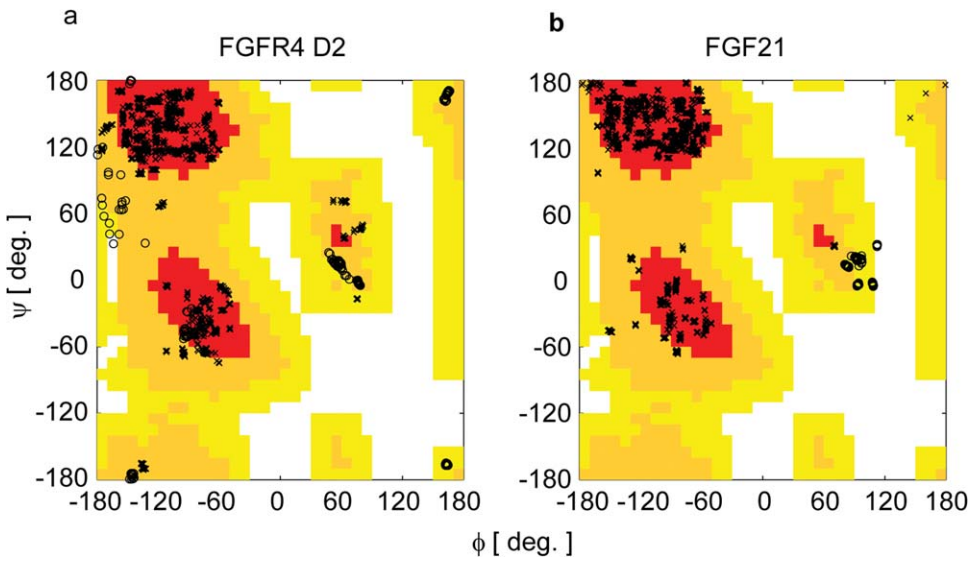



**Figure 9**

$\varphi$,$\psi$-correlation in the calculated FGF21 (left) and FGFR4 D2 (right) structures. Most favored regions are red, favored regions are orange and allowed regions are yellow. Each cross represents the $\varphi/\Psi$-values of one residue in one structure. Glycine residues are represented by circles.

X-ray structures of FGFR1, FGFR2, and FGFR3 including the D2-domain have been published. No structure of an extracellular domain of FGFR4 in known, so far. FGFR4 D2 has 58% primary sequence identity (77% homology) to FGFR1 D2, 65% identity (82% homology) to FGFR2 D2 and 72% identity (81% homology) to FGFR3 D2. FGFR4 is over-expressed in breast cancer and ovarian cancer tissue and could be important in FGF19-mediated tumorgenesis.[47]

FGF21 belongs to the fibroblast growth factor family comprising 22 members divided into six sub-families. The FGF19-sub-family involving also FGF21 and FGF23 are of special interest, because, in contrast to other FGFs, they have a systemic, hormone-like effect. Crystal structures of FGF19 and FGF23 are already known which have both 40% sequence identity and 55 and 62% sequence similarity to FGF21. As depicted in Figure 8(a), the extent of the resonance assignment[48] was insufficient to obtain any of the two solution structures by using current automated procedures like Cyana. In fact, the impossibility to provide a structure determination triggered us to develop the MOTOR approach. The detailed resonance assignment statistics of both proteins are given in Table IV. Figure 8(b,c) shows the MOTOR structural ensemble after the first cycle (b) and after refinement (c) calculation. The backbone RMSD between the structures before and after the refinement steps was 1.7–1.9 Å for FGFR4 D2 and 1.9 to 2.2 Å for FGF21. A unique solution was found already in b. The Ramachandran plot of the "best" 20 structures of the calculations depicted in Figure 9 indicates structures of a good quality. The structural parameters are summarized in Table IV. The surprisingly high number of NOE violations larger than 1 Å in FGFR4 D2 is due to peak picking close to the noise level, which leads to erroneous picked noise spikes. Fourteen of the 30 violated restraints are derived from peak intensities indicating distances longer than 5 Å. A more careful peak picking or the consideration of additional NMR data would probably reduce the number of violated NOEs. The ratio of assigned expected long range NOEs according to Figure 4(b) is 41.2% for FGF21 and 48.9% for FGFR4 D2.

As it was already observed for MptpA, the positional RMSD values of both, the backbone and all heavy atoms are surprisingly low for both proteins. These values do not reflect the precision of the structural ensemble, but only the structural precision of the data interpretation according to its algorithm. Precision and accuracy of NMR structures in general have already been discussed in the literature.[44,49] Following the suggestion in Ref. [50], we prefer local values that reflect the reliability of the local structure. Figure 8(d) shows the number of assigned NOEs per residue mapped onto the MOTOR structures. These illustrations are much more helpful than a global RMSD value for the interpretation of structural details and to prevent over interpretation.

## DISCUSSION

MOTOR is software capable of solving *de novo* protein structures based on sparse NMR data. Typical run times are 6–12 h for each calculation. The structure determination process is influenced by the use of structural models. Even with this introduced bias, we show for the example of MptpA that the software generates structural

ensembles in good agreement (RMSD $< 2$ Å) with the published structures. This solution is found with a higher accuracy than any of the used structural models even when only 70% of the nuclei are assigned. Those structural models with highest similarity to the real structure can be identified even with 60% assignment. Even at this level of assignment, we found that the RMSD of the models to the MOTOR structure correlates with 0.96 to the RMSD of the models to the X-ray structure. The use of structurally unrelated models does not lead to a structure that is in agreement with the NOE data as could be shown by the ratio of the number of assigned to the expected long range NOEs. This parameter is in analogy of the "recall" part of the RPF value. We used it, because it is difficult to measure the number of false positive NOE peaks for difficult proteins with large differences in peak amplitudes. The software was successfully applied to two *de novo* structures, FGF21 and FGFR4 D2, where the NMR data was not sufficient to solve the structural problem with conventional structure determination software and crystallization trials failed.

We consider our approach to be very useful in the finalization of various structural projects that have been cancelled due to a lack of resonance assignment.

MOTOR is software written in MATLAB. Binary versions for Linux or MacOS can be received upon request from US. MOTOR is free of charge for academic institutions without financial interests. Supportive information for installation and use of MOTOR can be found at www.motor-nmr.org. A structure solver like the basic tool of Cyana or CNS has to be installed to run MOTOR. MODELER and BLAST are required for the model structure generation.

## ACKNOWLEDGMENTS

We thank Torsten Herrmann (University Lyon, France) for carefully reading of this manuscript and very helpful comments. Matthias Dreyer and Oliver Boscheinen (both Sanofi, Germany) are acknowledged for fruitful discussions on structure-driven drug discovery. We thank the anonymous referees for a significant improvement of this manuscript by their valuable comments and suggestions.

## REFERENCES

1. Billeter M, Wagner G, Wüthrich K. Solution NMR structure determination of proteins revisited. J Biomol NMR 2008;42:155–158.
2. Donald BR, Martin J. Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints. Prog Nucl Magn Reson Spectrosc 2009;55:101–127.
3. Guerry P, Herrmann T. Advances in automated NMR protein structure determination. Q Rev Biophys 2011;44:257–309.
4. Güntert P. Automated structure determination from NMR spectra. Eur Biophys J 2009;38:129–143.
5. Williamson MP, Craven CJ. Automated protein structure calculation from NMR data. J Biomol NMR 2009;43:131–143.
6. Guerry P, Herrmann T. Comprehensive automation for NMR structure determination of proteins. In: Shekhtman A, Burz DS, editors. Protein NMR technics. New York: Springer Science; 2012. p 429–451.
7. Serrano P, Pedrini B, Mohanty B, Geralt M, Herrmann T, Wuthrich K. The J-UNIO protocol for automated protein structure determination by NMR in solution. J Biomol NMR 2012;53:341–354.
8. López-Méndez B, Güntert P. Automated protein structure determination from NMR spectra. J Am Chem Soc 2006;128:13112–13122.
9. Schmidt E, Güntert P. A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 2012;134:12817–12829.
10. Bahrami A, Tonelli M, Sahu SC, Singarapu KK, Eghbalnia HR, Markley JL. Robust, integrated computational control of NMR experiments to achieve optimal assignment by ADAPT-NMR. PLoS One 2012;7:e33173.
11. Gronwald W, Moussa S, Elsner R, Jung A, Ganslmeier B, Trenner J, Kremer W, Neidig KP, Kalbitzer HR. Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE). J Biomol NMR 2002;23:271–287.
12. Güntert P. Automated NMR structure calculation with CYANA. Methods Mol Biol 2004;278:353–378.
13. Herrmann T, Güntert P, Wüthrich K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 2002;319:209–227.
14. Huang YJ, Tejero R, Powers R, Montelione GT. A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins 2006;62:587–603.
15. Hung LH, Samudrala R. An automated assignment-free Bayesian approach for accurately identifying proton contacts from NOESY data. J Biomol NMR 2006;36:189–198.
16. Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM. Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. J Am Chem Soc 2004;126:6258–6273.
17. Linge JP, Habeck M, Rieping W, Nilges M. ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 2003;19:315–316.
18. Mumenthaler C, Braun W. Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. J Mol Biol 1995;254:465–480.
19. Mumenthaler C, Güntert P, Braun W, Wüthrich K. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. J Biomol NMR 1997;10:351–362.
20. Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H. Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. J Mol Biol 1997;269:408–422.
21. Rieping W, Habeck M, Bardiaux B, Bernard A, Malliavin TE, Nilges M. ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 2007;23:381–382.
22. Nilges M. Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. J Mol Biol 1995;245:645–660.
23. Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 2005;127:1665–1674.
24. Baran MC, Huang YJ, Moseley HN, Montelione GT. Automated analysis of protein NMR assignments and structures. Chem Rev 2004;104:3541–3556.
25. Güntert P. Automated NMR protein structure calculation. Prog Nucl Magn Reson Spectrosc 2003;43:105–125.
26. Jee J, Güntert P. Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. J Struct Funct Genomics 2003;4:179–189.

27. Moseley HN, Montelione GT. Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 1999;9: 635–642.
28. Saxena K, Schieborr U, Anderka O, Duchardt-Ferner E, Elshorst B, Gande SL, Janzon J, Kudlinzki D, Sreeramulu S, Dreyer MK, Wendt KU, Herbert C, Duchaussoy P, Bianciotto M, Driguez PA, Lassalle G, Savi P, Mohammadi M, Bono F, Schwalbe H. Influence of heparin mimetics on assembly of the FGF. FGFR4 signaling complex.J Biol Chem 2010;285:26628–26640.
29. Vogtherr M, Saxena K, Hoelder S, Grimme S, Betz M, Schieborr U, Pescatore B, Robin M, Delarbre L, Langer T, Wendt KU, Schwalbe H. NMR characterization of kinase p38 dynamics in free and ligand-bound forms. Angew Chem Int Ed Engl 2006;45:993–997.
30. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
31. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 2007;104:9615–9620.
32. Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Gutmanas A, Guntert P, He Y, Herrmann T, Huang YJ, Jaravine V, Jonker HR, Kennedy MA, Lange OF, Liu G, Malliavin TE, Mani R, Mao B, Montelione GT, Nilges M, Rossi P, van der Schot G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, Vries S, Vuister GW, Wu B, Yang Y, Bonvin AM. Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 2012;20:227–236.
33. Raman S, Huang YJP, Mao BC, Rossi P, Aramini JM, Liu GH, Montelione GT, Baker D. Accurate automated protein NMR Structure determination using unassigned NOESY data. J Am Chem Soc 2010;132:202–207.
34. Stehle T, Sreeramulu S, Löhr F, Richter C, Saxena K, Jonker HR, Schwalbe H. The apo-structure of the low-molecular-weight protein tyrosine phosphatase A (MptpA) from Mycobacterium tuberculosis allows for better target-specific drug development. J Biol Chem 2012;287:34569–34582.
35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
36. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. Protein Data Bank - Computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
37. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics 2006;Chapter 5:Unit 5 6.
38. Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 2009;44:213–223.
39. Doreleijers JF, Sousa da Silva AW, Krieger E, Nabuurs SB, Spronk CA, Stevens TJ, Vranken WF, Vriend G, Vuister GW. CING: an integrated residue-based structure validation program suite. J Biomol NMR 2012;54:267–283.
40. Mao B, Guan R, Montelione GT. Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure 2011;19:757–766.
41. Ramelot TA, Raman S, Kuzin AP, Xiao R, Ma LC, Acton TB, Hunt JF, Montelione GT, Baker D, Kennedy MA. Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. Proteins 2009;75:147–167.
42. Madhurantakam C, Rajakumara E, Mazumdar PA, Saha B, Mitra D, Wiker HG, Sankaranarayanan R, Das AK. Crystal structure of low-molecular-weight protein tyrosine phosphatase from Mycobacterium tuberculosis at 1.9-A resolution. J Bacteriol 2005;187: 2175–2181.
43. Markwick PR, Malliavin T, Nilges M. Structural biology by NMR: structure, dynamics, and interactions. PLoS Comput Biol 2008;4: e1000168.
44. Snyder DA, Bhattacharya A, Huang YJ, Montelione GT. Assessing precision and accuracy of protein structures derived from NMR data. Proteins 2005;59:655–661.
45. Lee PL, Johnson DE, Cousens LS, Fried VA, Williams LT. Purification and complementary DNA cloning of a receptor for basic fibroblast growth factor. Science 1989;245:57–60.
46. Hung KW, Kumar TK, Kathir KM, Xu P, Ni F, Ji HH, Chen MC, Yang CC, Lin FP, Chiu IM, Yu C. Solution structure of the ligand binding domain of the fibroblast growth factor receptor: role of heparin in the activation of the receptor. Biochemistry 2005;44: 15787–15798.
47. Wu X, Li Y. Understanding the structure-function relationship between FGF19 and its mitogenic and metabolic activities. Adv Exp Med Biol 2012;728:195–213.
48. Elshorst B, Saxena K, Schieborr U, Schwalbe H. (1)H, (13)C and (15)N assignment of D2 domain of human fibroblast growth factor receptor 4. Biomol NMR Assign 2012.
49. Rieping W, Habeck M, Nilges M. Inferential structure determination. Science 2005;309:303–306.
50. Nabuurs SB, Spronk CA, Vuister GW, Vriend G. Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. PLoS Comput Biol 2006;2:e9.