J Biomol NMR (2009) 45:397–411
DOI 10.1007/s10858-009-9379-y

ARTICLE

Protein structure calculation with data imputation: the use of substitute restraints

Carolina Cano · Konrad Brunner · Kumaran Baskaran · Ralph Elsner · Claudia E. Munte · Hans Robert Kalbitzer

Received: 9 August 2009 / Accepted: 17 September 2009 / Published online: 17 October 2009

Abstract The amount of experimental restraints e.g., NOEs is often too small for calculating high quality three-dimensional structures by restrained molecular dynamics. Considering this as a typical missing value problem we propose here a model based data imputation technique that should lead to an improved estimation of the correct structure. The novel automated method implemented in AUREMOL makes a more efficient use of the experimental information to obtain NMR structures with higher accuracy. It creates a large set of substitute restraints that are used either alone or together with the experimental restraints. The new approach was successfully tested on three examples: firstly, the Ras-binding domain of Byr2 from *Schizosaccharomyces pombe*, the mutant HPr (H15A) from *Staphylococcus aureus*, and a X-ray structure of human ubiquitin. In all three examples, the quality of the resulting final bundles was improved considerably by the use of additional substitute restraints, as assessed quantitatively by the calculation of RMSD values to the "true" structure and NMR *R*-factors directly calculated from the original NOESY spectra or the published diffraction data.

Keywords Protein structure determination · Missing values · Substitute restraints · Data imputation

Abbreviations

bb	Backbone
HPr	Histidine-containing phosphocarrier Protein
MD	Molecular dynamics
MDRA	Molecular dynamics result analysis
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
PDB	Protein data bank Brookhaven
RBD	Ras binding domain
RMSD	Root mean square deviation
sc	Sidechain

Introduction

Nowadays, structural biology remains a challenging field since the gap between the number of solved structures and the number of known protein sequences is still huge. In any structure determination process of a biological macromolecule, the general goal is to obtain a structure as accurate as possible from the available experimental data (mainly from X-ray crystallography (Ilari and Savino 2008) and solution NMR spectroscopy (Wüthrich 1990)). Moreover, the structure determination process has to be as fast as possible, demanding that only a minimal set of experimental data is recorded. The common method for biomolecular structure determination by NMR spectroscopy relies on the identification of a dense network of interproton distance restraints. These distances can be obtained from nuclear Overhauser enhancement (NOE), which give rise to cross-peaks in NOE experiments. The structural information contained in NOEs reports on pairwise distances between specific protons and can thus provide unequivocal information about the relative spatial locations of different residues in a protein sequence (Wüthrich 1986). Other experimental information derived from *J*-couplings (Pardi et al. 1984; Kim and Prestegard 1990; Torda et al. 1993; Garrett et al. 1994), chemical shifts (Cavalli et al. 2007; Shen et al. 2008), and residual dipolar couplings (Tolman

C. Cano · K. Brunner · K. Baskaran · R. Elsner · C. E. Munte · H. R. Kalbitzer (✉)
Institut für Biophysik und physikalische Biochemie, University of Regensburg, Universitätstr. 31, 93053 Regensburg, Germany
e-mail: hans-robert.kalbitzer@biologie.uni-regensburg.de

et al. 2001; Qu et al. 2004; Rathinavelan and Im 2008) has also been used to further improve the quality of NMR structures. Despite these new types of experimental data, distance restraints have remained the single most valuable source of information for the elucidation of high-resolution solution structures by NMR spectroscopy, and therefore, traditional NMR structure determination programs such as CNS (Brunger et al. 1998; Brunger 2007) or CYANA (Guntert 2004) require a large number of redundant NOE restraints—typically 15–20 NOE restraints per residue—to obtain a high resolution structure.

Although tremendous advances in both NMR hardware and software have taken place during the past decade, obtaining three-dimensional macromolecular structures by NMR techniques is still a time-consuming task and the availability of a fast and reliable method able to provide a molecular model based on few experimental restraints is still an ambitious goal. As an alternative and complementary approach, protein structure prediction with a limited number of distance restraints using computational tools holds great promise (Smith-Brown et al. 1993; Aszodi et al. 1995; Skolnick et al. 1997; Kolinski and Skolnick 1998; Standley et al. 1999; Bailey-Kellogg et al. 2000; Bowers et al. 2000; Sikorski et al. 2002; Herrmann et al. 2002; Li et al. 2003; Alexandrescu 2004; Gronwald et al. 2004; Fuentes et al. 2005; Tang and Clore 2006; Latek et al. 2007; Rieping et al. 2007; Angyan et al. 2008).

Despite all these many approaches to accelerate structure calculation, capable of yielding a protein structural model of acceptable quality by the use of automated or semi-automated methods, routine structure prediction of new folds is still a challenging task for computational biology, not only in the proper determination of overall fold but also in building models of acceptable resolution, useful for modelling the drug interactions and protein–protein complexes (Wishart 2005). Two interesting computational approaches were developed in our group to accelerate the process of protein determination: the program PERMOL which extracts the structural information from 3D-structures and translates it into a network of conformational restraints to be employed in torsion angle dynamics calculations (Möglich et al. 2005); and a second approach, based on the combination of data from different sources, such as NMR, X-ray or homology modelling, using the module ISIC (Brunner et al. 2006). Both PERMOL and ISIC are part of the larger AUREMOL software package for automated NMR spectrum evaluation and protein structure determination (Gronwald et al. 2004).

Traditional methods for the calculation of high-quality NMR structures rely primarily on the redundancy and completeness of the experimental restraints, and they do not perform satisfactorily when only sparse experimental data are available. In this paper we propose and test a novel approach to protein structure calculation from sparse data that uses the available structural information more efficiently. It is based on well-known data imputation techniques (Rubin 1976, 1981; Schafer and Graham 2002) applicable to incomplete data sets. In our implementation it consists of the automated generation of a large set of substitute restraints by PERMOL which substitute/replace the primary experimental restraints and indirectly add missing information for an optimal convergence of the structure calculation.

This substitute restraints method was successfully tested on two representative globular proteins for which the required NMR data already exist: the Ras-binding domain of Byr2 from *Schizosaccharomyces pombe* (Gronwald et al. 2001) and a mutant of the histidine-containing phosphocarrier protein, HPr (H15A), from *Staphylococcus aureus* (C. E. Munte et al., to be published). A third approach was also tested on a X-ray structure of the model protein ubiquitin (Vijay-kumar et al. 1987) used for the creation pseudo NOE restraints (Table 1). The refinement of all these structures calculated from limited sets of NOE restraints by the use of a network of substitute restraints has proved a good agreement with the experimental data. Modelled structures were quantitatively compared to their respective target structures by calculating RMSD and *R*-factor values.

Materials and methods

NMR spectroscopy and structures

The sequential assignments of the NMR signals of Byr2 from *Schizosaccharomyces pombe* (residues 71–165 here referred as residues 1–95) and the corresponding experimental details have been described in (Gronwald et al.

Table 1 Test proteins

	RBD-Byr2	HPr(H15A)	Ubiquitin
PDB ID	1I35	2KP9	1UBQ
Organism	*Schizosaccharomyces pombe*	*Staphylococcus aureus*	*Homo sapiens*
Method	Solution NMR (10 structures)	Solution NMR (10 structures)	X-ray
Resolution	–	–	1.8 Å
Reference	Gronwald et al. (2001)	Munte et al., to be published	Vijay-kumar et al. (1987)

2001). The NMR structure is deposited in the Protein Data Bank under the PDB ID: 1I35. Details of the structure of the point mutant HPr (H15A) from *Staphylococcus aureus* (PDB ID 2KP9) will be published elsewhere. For the structure validation 2D ^{1}H NOESY spectra were recorded at 800 MHz with mixing times and relaxation delays of 0.1 and 1.41 s for Byr2, and 0.1 and 1.18 s for HPr(H15A), respectively. Spectra were recorded in 90% H_2O/10% D_2O (v/v) at 298 and 303 K, respectively. NMR data were processed with the programs XWINNMR and TopSpin (Bruker Biospin) and were evaluated with the program AUREMOLv.2.3.1 (Gronwald et al. 2004).

Molecular dynamics calculations

Structure calculations were performed using the molecular dynamics program CNS v.1.1. (Crystallography and NMR System for crystallographic and NMR structure determination) (Brunger et al. 1998; Brunger 2007) employing the substitute restraints in a simulated annealing protocol for extended-strand starting structures. High-temperature torsional angle dynamics were run at 50,000 K for 3,000 steps with a time step of 5 fs. The high number of restraints required a threefold reduction of the time step for the integration of the equation of motion to 5 fs and a reduction of the ceiling value to 15 for around 30 restraints per residue for the NOE-energies (the default value is 30 for typically 16 restraints per residue). In the first cooling stage, torsional angle dynamics were used for 3,000 steps with a starting temperature of 50,000 K and a time step of 5 fs. The second cooling stage was performed with 3,000 steps of Cartesian dynamics with a time step of 5 fs and a starting temperature of 3,000 K. In the final stage, 2,000 steps of energy minimization were performed. In the case of the Byr2, the final 10 conformers were refined in explicit water using the CNS protocol re_h2o.inp (Linge et al. 2003) including the NOE distance restraints, H-bond distance restraints, dihedral angle restraints and residual dipolar couplings. Dipolar couplings were introduced in the water refinement using the SANI protocol (Tjandra et al. 1997) where different values of the force constant were tested to obtain the best refinement**.**

Structure validation

The program PROCHECK_NMR (Laskowski et al. 1996) was employed to check the stereochemical quality by calculating Ramachandran plots. The program MOLMOL was used to display the structures and to calculate the RMSD-values (Koradi et al. 1996). NMR *R*-factors were calculated with AUREMOL according to Gronwald et al. (2000). The agreement of the obtained structural bundles with the obtained NOESY-spectra was checked by calculating the NMR *R*-factor of the bundles directly from the corresponding experimental 2D-NOESY spectra. As recommended by Gronwald et al. (2000) the regions from 6.0 to −1.0 ppm for HPr (H15A) and from 4.8 to −1.0 ppm for Byr2 were not considered for the calculation, having 2,462 (HPr(H15A)) and 2,671 (Byr2) experimental peaks automatically assigned for the NMR *R*-factor calculation. The program REFMAC for macromolecular refinement (Murshudov et al. 1997) is included in the CCP4 software package and was used to calculate the total *R*-factor and free *R*-factor for assessing the agreement between the atomic model and X-ray data. The program requires the input files with the coordinates of the model (in PDB format) and structure factors (in mmCIF or MTZ format) and runs completely automatically to give both crystallographic *R*-factors. The agreement between measured residual dipolar couplings and residual dipolar coupling calculated for a certain structure can be estimated by Cornilescu Q-value (Cornilescu et al. 1998) using the program Pales (Zweckstetter 2008).

Implementation overview

The MDRA (Molecular Dynamics Results Analysis) tool in AUREMOL was developed in order to facilitate the analysis of the obtained structures and was used to determine the number of NOEs restraints with a violation >0.05 nm. A second tool was also included to deal with data from X-ray sources and calculate crystallographic *R*-factor, by converting NMR output pdb files to X-ray format file. It is based on the fitting of the NMR model to the target X-ray structure using a rotation matrix and translation vector to have the correct orientation, giving also the RMSD value to the “true” structure; moreover, the hydrogens are removed and the original crystallographic information such as space group and cell dimensions is also included in the final X-ray format output pdb file. Both tools are fully incorporated in the software package AUREMOL [http://www.auremol.de]. A new tool for the automated calculation of substitute restraints was also implemented in AUREMOL.

Theoretical considerations and general strategy

NMR structure determination is a still improving process that relies on two different factors (1) the search for additional information sources and (2) the optimization of the usage of the available experimental as well as a priori information. In recent years most of the efforts have focused on the first problem by developing new experimental methods to gain additional NMR derived information such as the measurements of residual dipole couplings (Tolman et al. 2001) and the use of chemical shift information (Shen et al. 2008) or by using additional (a priori) non-NMR information such as the information from the data base of known protein structures (Brunner et al. 2006). However, there are indications that the typical simulated annealing protocols do not make optimal use of the available NMR information, example are reports that the distance information for small distances is clearly not used satisfactorily (Gronwald et al. 2000) or that application of ISD (Inferential Structure Determination) (Rieping et al. 2005) results in better defined structures. There are two sources that may impede the convergence to an optimal structure, errors in the input data, a too small number of restraints and weak points in the optimization procedure itself. Typical errors are wrong assignments of some cross peaks (especially when automated procedures are used) or the assumption of too small or too large error limits of the NOE intensities measured (because of the non-linear averaging of NOEs, spin diffusion effects, etc.). In addition, a too small number of NOE restraints usually leads to insufficient convergence of the simulated annealing procedure to the optimal ("true") structure. The availability of a too small number of experimental restraints represents a typical missing value problem of statistics (Rubin 1976, 1981; Epron 1979; Schafer and Graham 2002) where the available experimental data are not sufficient to accurately predict properties of the system with standard methods. When the MAR (missing at random) condition (Rubin 1976) is fulfilled and the data themselves are rather sparse, model based data imputation techniques are powerful means to substitute the missing values.

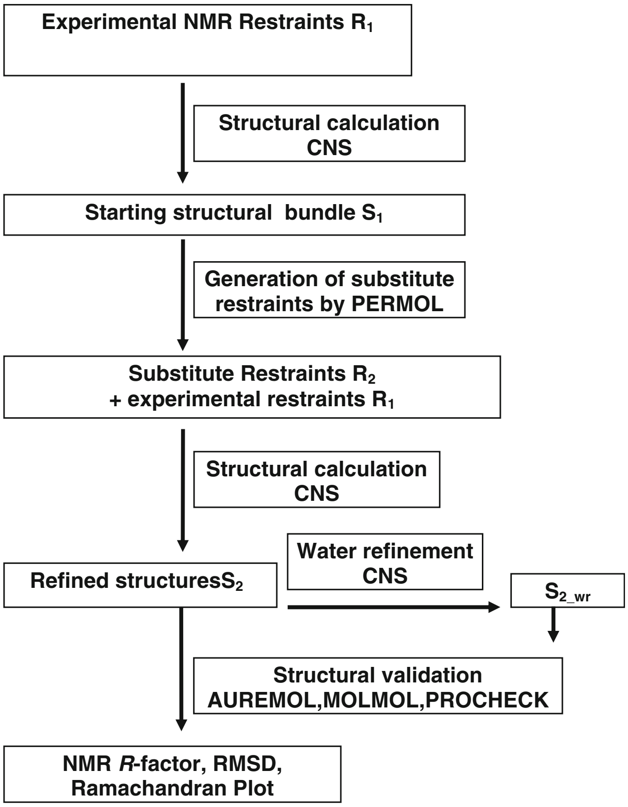
Experimental NMR Restraints R_1
Structural calculation CNS
Starting structural bundle S_1
Generation of substitute restraints by PERMOL
Substitute Restraints R_2 + experimental restraints R_1
Structural calculation CNS
Refined structuresS_2
Water refinement CNS
S_{2_wr}
Structural validation AUREMOL,MOLMOL,PROCHECK
NMR *R*-factor, RMSD, Ramachandran Plot

Fig. 1 Schematic description of the substitute approach for the improvement of NMR structures. In general, the use of the substitute restraints together with the experimental restraints is usually recommended. Since erroneous experimental restraints sometimes lead to suboptimal results (as measured by the *R*-factor), as an option only substitute restraints can be used (see *below*). Refinement of the structures in explicit water after the use of substitute restraints is a strongly recommended option that leads to an additional improvement of the structures (see *below*)

Applied to the problem of structure determination with sparse experimental data but with a large number of missing data (e.g., additional distance and dihedral angle restraints), we propose a fast and reliable method based on a traditional statistical approach, a model based on *mean substitution* of missing data (replacing all missing data in a variable by the mean of that variable) which may accurately predict missing information by producing "internally consistent" sets of results ("true" correlation matrices). From a bundle of N molecular models based on the available experimental restraints, additional conformational restraints (distances, hydrogen bonds and dihedral angles) are estimated to substitute the initial missing information by calculation of the weighted mean values and corresponding standard deviations for selected parameters. The general procedure is schematically depicted in Fig. 1. For calculation of the missing parameters (substitute restraints) we can use algorithms that are implemented in PERMOL (Möglich et al. 2005) and were used originally for a molecular dynamics based structure prediction. Here, the means (expectation values) and error limits are calculated from a bundle of model structures based on a Gaussian approximation tested by Kolmogoroff–Smirnov statistics. The obtained substitute restraints should faithfully represent the accessible conformational space defined by the experimental data and the physical model of the protein and simultaneously guide the optimization procedure to the global minimum. Experience shows that for this aim the additional substitute restraints should form a rather dense network of uniformly distributed restraints. They have to be internally consistent and consistent with the external, experimental restraints, a feature that is automatically granted by their calculation from structural bundles.

Structure calculation protocol

The structure calculation procedure can be described in the following steps: (1) calculate a structural bundle from the original experimental dataset by any of the methods described in literature e.g., by simulated annealing, (2) calculate a set of substitute restraints, (3) perform a restrained molecular dynamics simulation using the substitute restraints and the experimental restraints including optionally a refinement in explicit water, and (4) validate the quality of the structure.

Selection of test data sets

The final test criterium of the proposed imputation technique is that the structural ensemble obtained using the proposed imputation technique is closer to the target ("true") ensemble than that obtained by traditional methods. The first test case used here is the Ras-binding domain of Byr2 from *Schizosaccharomyces pombe*. Its NMR structure derived from a limited set of restraints has been already published and has been deposited in the PDB-data base (Gronwald et al. 2001). New NMR data and an extensive data analysis led to higher number of experimental restraints and to a significantly improved solution structure (Elsner 2006). Here, we can test the method proposed on two sets of real NMR experimental data. Ideally, the structures calculated with the smaller set of restraints using data imputation would be of the same quality as the structures calculated from the larger set of restraints in a conventional manner.

As a second test case, a highly resolved NMR structure of mutant HPr (H15A) from *Staphylococcus aureus* (Munte et al., to be published) was selected to test our approach because this available good quality set of experimental NOE restraints could be used to artificially and randomly remove restraints from the original NOE distance restraint list and study how the decreasing number of restraints could affect to the 3D structure of the protein by obtaining increasingly disordered structures and how well we could overcome this lack of information by the use of substitute restraints.

The third test was done on the 1.8 Å X-ray structure of ubiquitin protein (Vijay-kumar et al. 1987). Pseudo NOE distance restraints (classified as intraresidual, sequential, medium and long range restraints like in the experimental datasets) with an upper distance limit of 0.5 nm were extracted automatically from this target structure by PERMOL; using reduced distance restraint lists by deleting systematically restraints from the four distance classes, low resolution structural bundles were calculated which were improved by the substitute restraints.

Generation of substitute restraints

In its application to homology modelling the program PERMOL uses a combination of three types of restraints that showed to be optimal for the prediction of the three-dimensional structure, namely the combination of dihedral angle restraints, hydrogen bond restraints and distance restraints. The same principal types of restraints were used to calculate substitute restraints in this paper and to substitute the missing values but details had of course to be adapted to the new problem.

In PERMOL and in this application local structural restraints are mainly coded by a weighted average of the backbone dihedral angles. Their expectation values and standard deviations are calculated with the algorithm proposed by Döker et al. (1999). Conserved hydrogen bonds are also used to generate distance restraints between the atoms involved in forming the bond (Möglich et al. 2005). The global fold is determined by distance restraints, the selection of atoms used in our application is not trivial, since the number of all pairwise distances is too large to be handled by the available molecular dynamics programs. Therefore, a reduced set of distance restraints has to be defined that represents the structure sufficiently well and creates a energy hyperplane for the structure calculation as smooth as possible. Since the data imputation should not restrict too much the available conformational space, for the error limits used in the molecular dynamics calculations a rather high confidence level of 99.9% (error probability <0.1%) based on a *t*-test was selected for the calculations. The same small error probabilities were also used for the angle restraints and hydrogen bond restraints.

PERMOL allows an arbitrary choice of restraints by extracting the information for selected residues of a given model to create an artificial set of structural restraints; restraints files for two different molecular dynamics programs (CNS and CYANA) are generated automatically and can then be combined with other restraint files. In our approach, from the experimental restraints a structural bundle is calculated by a simulated annealing protocol and a set of structural restraints is calculated. Besides the main chain angles ϕ and ψ all side chain angles χ and the conserved hydrogen bonds were included. For determining the optimal selection of distance restraints several combinations were tested. It turned out that the inclusion of a larger number of atoms that were separated by large distances lead to problems with the convergence of the procedure. Therefore, the upper mean distance between atoms to be considered was limited to smaller value. In one class all average pairwise H^{α} distances in the distance range between 0.18 and 1.5 nm are considered. Intraresidual or sequential contacts are excluded. In the second class distances between all other protons are included provided their average distance is smaller than 0.6 nm and the atoms considered belong to different amino acids.

Results and discussion

Structure improvement of the Ras-binding domain of Byr2

As a first example for testing the effect of the data imputation on the quality of the obtained structure we selected an experimental NMR data set. The NMR structure of the Ras-binding domain of Byr2 from *Schizosaccharomyces*

Table 2 Selection of the substitute restraints

Distance restraints		
Selected atoms	H^{α}, $H^{\alpha 2}$, $H^{\alpha 3}$	H^{N} and all side chain hydrogens
Distance range (nm)	0.18–1.5	0.18–0.6
Confidence level (%)	99.9	99.9
Residue difference	≥2	≥1
Dihedral angles		
Selected angles	Main chain ϕ and ψ, all side chain χ angles of single bonds	
Hydrogen bonds		
Donators	All possible donators in main and side chains	
Acceptors	All possible acceptors in main and side chains	

Atom and dihedral angle nomenclature corresponds to the IUPAC recommendations (Markley et al. 1998). For the definition of confidence levels see paragraph "Generation of substitute restraints"

pombe (residues 71–165 here referred as residues 1–95, PDB ID: 1I35) has been published by Gronwald et al. (2001) and a new experimental structure calculated with a much larger set of experimental restraints is available (Table 3). The original experimental data set R_{1_Byr2} contains 822 distance restraints, 88 dihedral restraints, 29 hydrogen bond distances and 28 amide residual dipolar couplings (967 structural restraints). Employing this set of restraints, 500 structures were calculated by the molecular dynamic program CNS as described in "Material and methods". The 10 best structures in terms of lowest total energy are selected by AUREMOL to define the starting Byr2 structural bundle S_{1_Byr2}. These structures were refined in explicit water and resulted in the structural bundle $S_{1\_Byr2\_wr}$. Using the parameters given in Table 2, 6,237 distance restraints, 351 dihedral angles restraints and 31 H-bonds restraints were created (R_{2_Byr2}). In this example, the original set of experimental RDC was included to show the quality of the substitute data set obtained. With the molecular dynamics program CNS calculations 500 structures were obtained and the 10 lowest energy structures define the improved final bundle (S_{2_Byr2}). These structures were again refined in explicit water to give the bundle $S_{2\_Byr2\_wr}$. In order to compare the result to the best NMR structure derived from a larger set of experimental restraints, a third bundle of structures S_{3_Byr2} was also calculated employing 1,804 experimental distance restraints and the same set of dihedral angle and H-bonds restraints (R_{3_Byr2}) used for S_{1_Byr2} (1,949 structural restraints) (Fig. 2). After water refinement the structural bundle $S_{3\_Byr2\_wr}$ was obtained.

The quality of the resulting structures was compared to that of the original one calculating the RMSD to the mean structure of the obtained bundle, the RMSD to the lowest energy structure of the structural bundle using the larger sets of NOEs after water refinement ($S_{3\_Byr2\_wr}$), the angle distribution in the Ramachandran plots, the Cornilescu Q-value and the NMR *R*-factor (Table 3).

The most important parameters are the RMSD of the structural bundle to the lowest energy structure of the water refined structural bundle calculated with the larger number of structural restraints that is assumed to be closest to the "true" structure. Here, the use of substitute restraints

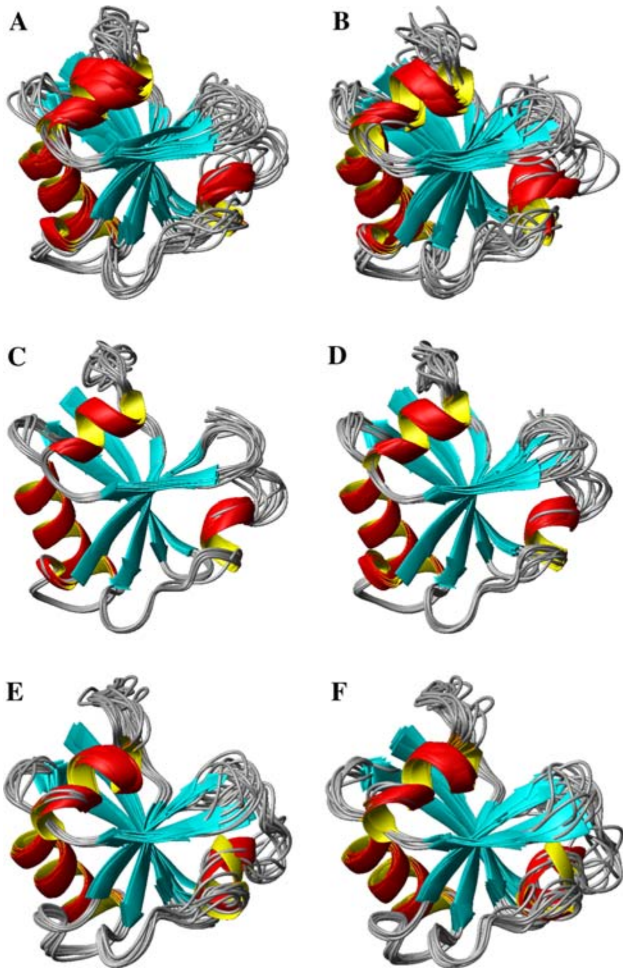

Fig. 2 Improvement of the solution structure of Byr2-RBD by the use of substitute restraints. **a** Starting Byr2 bundle S_{1_Byr2} (822 experimental distance restraints). **b** Starting Byr2 bundle after water refinement $S_{1\_Byr2\_wr}$. **c** PERMOL Byr2 bundle S_{2_Byr2} (6,237 distance restraints). **d** Bundle of NMR Byr2 structures after water refinement $S_{2\_Byr2\_wr}$. **e** Bundle of NMR Byr2 structures S_{3_Byr2} (1,804 experimental distance restraints). **f** Bundle of NMR Byr2 structures S_{3_Byr2} after water refinement $S_{3\_Byr2\_wr}$

Table 3 Number of restraints and quality values for bundles of Byr2 NMR structures

	$S^{a}_{1\_Byr2}$	$S^{a}_{1\_Byr2\_wr}$	$S^{a}_{2\_Byr2}$	$S^{a}_{2\_Byr2\_wr}$	$S^{a}_{3\_Byr2}$	$S^{a}_{3\_Byr2\_wr}$
NOE distance restraints	822	822	6,237	6,237	1,804	1,804
H-bonds restraints	29	29	31	31	29	29
Dihedral angle restraints	88	88	351	351	88	88
Residual dipolar couplings	28	28	28	28	28	28
RMSD bb(nm) to mean[b]	0.129	0.136	0.059	0.061	0.009	0.009
RMSD bb(nm) to the "true" structure $S^{c}_{3\_Byr2\_wr}$	0.217	0.215	0.204	0.199	0.138	0.120
Procheck Ramach mf+a (%)[d]	85.1	88.5	89.6	89.7	83.9	85.0
AUREMOL *R*-factor	0.477	0.475	0.470	0.453	0.474	0.464
MDRA violated NOE restraints (%)[e]	4.88	4.51	4.76	4.87	0.00	0.08
Cornilescu Q-value[f]	0.009	0.011	0.007	0.007	0.012	0.012

[a] S_{1_Byr2}, initial structural bundle calculated from 967 experimental restraints in a conventional manner; $S_{1\_Byr2\_wr}$, structural bundle obtained after water refinement of S_{1_Byr2}; S_{2_Byr2}, Byr2 bundle calculated with substitute restraints only; $S_{2\_Byr2\_wr}$, Byr2 bundle after water refinement of S_{2_Byr2}; S_{3_Byr2}, structural bundle calculated with 1,949 experimental restraints; $S_{3\_Byr2\_wr}$, structural bundle obtained after water refinement of S_{3_Byr2}; water refinement SANI force constants: $S_{1\_Byr2\_wr}$ and $S_{3\_Byr2\_wr} = 5.5$, $S_{2\_Byr2\_wr} = 3.5$

[b] RMSD values of the backbone atoms (N, C^{α}, C) of the 10 structures to the mean structure calculated by the program MOLMOL

[c] Average pairwise RMSD values of the backbone atoms (N, C^{α}, C) of each 10 structures bundle to the lowest energy structure of $S_{3\_Byr2\_wr}$ assumed to be close to the true structure

[d] Ramachandran Plot percentages of residues in most favoured and allowed regions

[e] The percentage of violated NOEs from the corresponding experimental set R_{3_Byr2} with violation >0.05 nm calculated by Molecular Dynamics Results Analysis tool (MDRA) included in AUREMOL

[f] Cornilescu Q-value was calculated by Pales program from a set of 28 experimental residual dipolar coupling and $Da = -18.11$, $R = 0.3$

followed by water refinement has the largest effect, the RMSD to the "true" structure of the backbone atoms (N, C^{α}, C) of the bundle decreases from 0.217 nm for S_{1_Byr2} to 0.199 nm for $S_{2\_Byr2\_wr}$. The other, equally important parameter is the NMR *R*-factor. Note that it is always calculated directly from the experimental NOESY spectrum and is thus not directly dependent on the NOEs used for the calculation of the structures but only on the quality of the structural bundle. Also the NMR *R*-factor decreases significantly by almost 5% indicating that the use of substitute restraints together with water refinements results in clearly better structures. Finally, the agreement of the experimental residual dipolar couplings measured by the Cornilescu Q-factor decreases somewhat with use of substitute restraints together with the amide residual dipolar couplings.

The other factors listed in Table 3 are more indirect quality measures of the obtained structures. As to be expected the RMSD values of the backbone atoms to the mean averaged structure for the newly calculated bundles decreases when the number of restraints increases. Compared to the input NMR structure S_{1_Byr2} (0.129 nm) the RMSD to the mean structure of the bundle decreases to 0.059 nm and is even lower than the value for the third bundle obtained with a higher number of experimental restraints S_{3_Byr2} (0.086 nm). When water refinement was performed, the values are slightly higher but the tendency is the same: $S_{1\_Byr2\_wr}$ 0.136 nm, $S_{2\_Byr2\_wr}$ 0.067 and $S_{3\_Byr2\_wr}$ 0.090. This clearly shows the positive influence of the well defined restraints created by PERMOL on the structural calculation. In addition, the stereochemical quality of the models measured by the number of φ- and ψ-torsional angles in the energetically most favoured and allowed regions of the Ramachandran plot increases for the final set $S_{2\_Byr2\_wr}$ (89.7%) compared to the input structures of set S_{1_Byr2} (85.1%). However, also the water refinement procedure alone has a strong effect on the stereochemical quality of the structures. Besides, we examined the percentage of violated NOE restraints using the tool Molecular Dynamics Results Analysis included in AUREMOL. This tool gives the percentage of violated NOE restraints whose violation is higher than 0.05 nm of the examined structures after CNS calculation, compared to the NOE restraint file used to calculate the target structure S_{3_Byr2}. These values also are in line with an improvement of the refined final structure S_{2_Byr2} (4.76%), slightly lower than the initial S_{1_Byr2} (4.88%). The final water refinement leads to a small increase of the NOE violations.

In conclusion, the quality of the structures obtained from 967 structural restraints (NOEs, dihedral angle restraints, hydrogen bonds, and amide residual dipolar couplings) is strongly improved when substitute restraints are used. In fact, with respect to the NMR *R*-factor the structural bundle obtained with substitute restraints $S_{2\_Byr2\_wr}$ (0.453) was better than the bundle obtained with the higher number of 1,949 experimental NOE restraints in a conventional manner $S_{3\_Byr2\_wr}$ (0.464).

Structure refinement of HPr (H15A)

Another way to test the performance of the proposed method is to select a well-resolved NMR structure (the target structure) and create new structural bundles with lower resolution by reducing the number of experimental restraints. For obtaining realistic conditions the cross peaks corresponding to randomly selected atoms were removed. This corresponds to the situation when the assignment of the spectra gets more and more incomplete. The structures calculated from the reduced sets of restraints using substitute restraints should be closer to the target structure than those calculated in the conventional way. The structure of HPr (H15A) (S_{1_HPr}) from *Staphylococcus aureus* (Munte et al., to be published) has been solved by multidimensional NMR spectroscopy from a set of 2,325 distance restraints and 69 3-bond J coupling restraints (R_{1_HPr}) (Table 4). After randomly removing a part of the experimental NOE-restraints (5, 15, 25,…, 85% of the total number of restraints) these new, reduced sets of restraints (R_{2_HPr}) together with the remaining 3-bond J-coupling restraints were used to calculate 500 new structures by CNS molecular dynamics calculation. The 10 lowest energy structures were selected by AUREMOL to define the starting bundles (S_{2_HPr}) for the refinement with substitute restraints. From these bundles new sets of substitute restraints (distance restraints, dihedral angles and hydrogen bond restraints) were calculated (R_{3_HPr}) and were employed alone or together with the corresponding, reduced lists of original restraints (R_{2_HPr}) in the CNS molecular dynamics calculations (Fig. 3). The obtained substitute restraints were used alone (structures S_{3_HPr}), together with the corresponding experimental NOE-restraints (structures S_{4_HPr}), and together with the

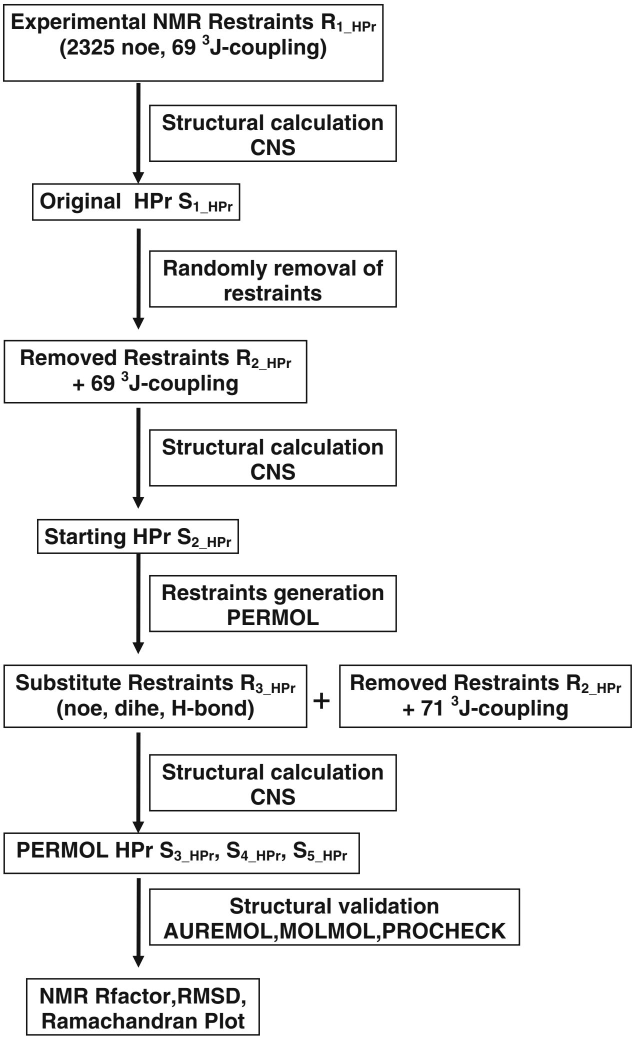

Fig. 3 Schematic description of the substitute approach for the improvement of HPr(H15A)

Table 4 Number of restraints used in HPr(H15A) test

Experimental restraints for HPr(H15A) (R_{1_HPr})										
Total number of NOEs	1,984									
Intraresidual NOEs	783 (39.5%)									
Sequential NOEs (i, $i + 1$)	449 (22.6%)									
Medium-range NOEs (i, $i + j$; $1 < j \leq 4$)	294 (14.8%)									
Long range NOEs (i, $i + j$; $j > 4$)	458 (23.1%)									
^{3}J-coupling constants (not Gly)	69 observed from 80 (86%)									
Randomly removed sets of NOE distance restraints (R_{2_HPr})										
% removed	0	5	15	25	35	45	55	65	75	85
NOE	2,325	2,209	1,976	1,744	1,511	1,279	1,046	814	581	349
PERMOL restraints (R_{3_HPr})										
NOE	–	6,752	6,712	6,716	6,611	6,544	6,458	6,357	5,512	4,845
H-bond	–	36	39	39	38	28	27	24	14	3
Dihedral	–	358	340	346	350	344	332	336	302	292

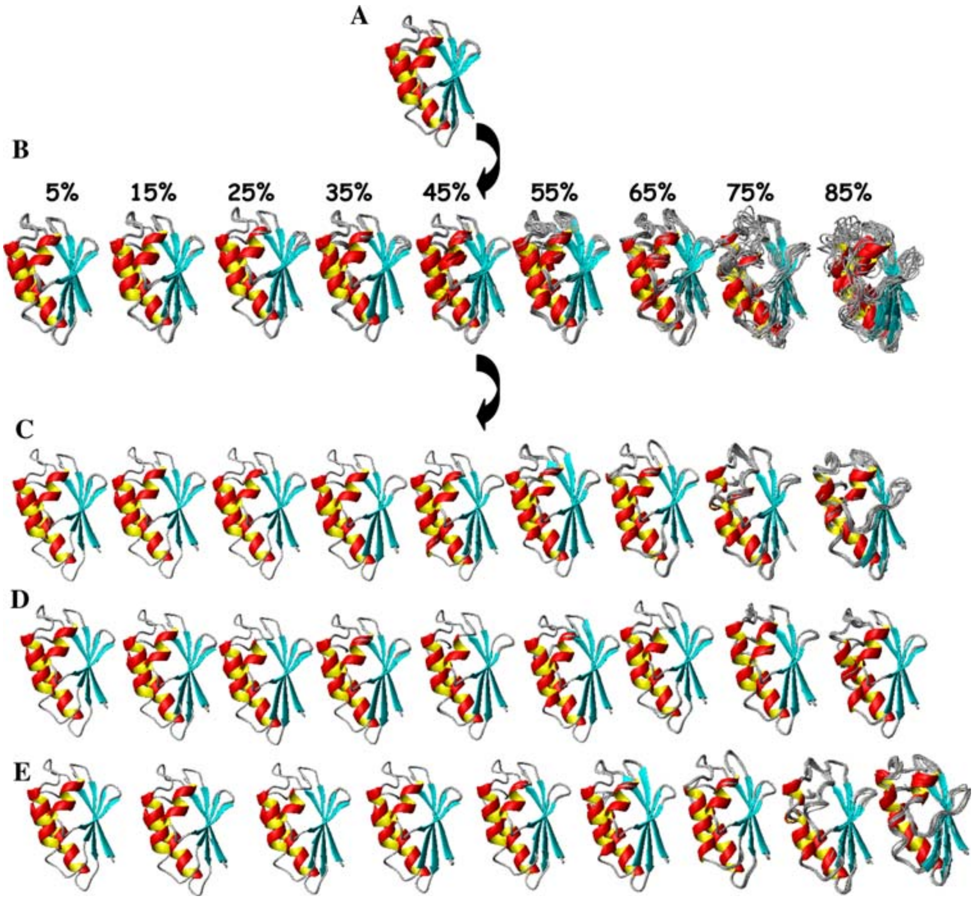

Fig. 4 Improvement of the solution structure of HPr(H15A) by the use of substitute restraints: **a** Starting structural bundle HPr (H15A) S_{1_HPr} calculated from 2,325 distance restraints and 69 3-bond J-coupling restraints (R_{1_HPr}), **b** structural bundle S_{2_HPr} calculated with reduced sets of restraints (R_{2_HPr}), **c** bundle S_{3_HPr} calculated with substitute restraints, **d** bundle S_{4_HPr} calculated with substitute restraints and the corresponding NOE data set, **e** bundle S_{5_HPr} calculated with substitute restraints and the corresponding sets of experimental NOEs and dihedral angles

corresponding experimental NOE and $^3J_{HN-H\alpha}$ coupling restraints (structures S_{5_HPr}). The obtained structural bundles are shown in Fig. 4.

The most important parameter, the RMSD of the structural bundle to the lowest energy structure of the initial bundle calculated with all NOEs decreases when substitute restraints are used (Table 5; Fig. 5). In general, the best results are obtained, when substitute restraints are used together with the experimental NOE restraints. When only 15% of the initial experimental NOE were used, still reasonable structures are obtained using substitute restraints. The RMSD to the optimal structure is 0.22 nm (about 24% smaller than that obtained with the experimental restraints only). The NMR *R*-factor follows this trend, using substitute restraints results always in a lower NMR-*R*-factor. This is also true for the NOE-violations that are only calculated for the NOE set used for the actual structure calculation. The additional use of the experimental data does not lead always to better results, probably since there are always some inconsistencies in the experimental data.

The factors that generally describe the quality of the structures independent of the experimental data also get better when using the substitute restraints (Table 5). The initial structure if our test protein HPr(H15A) (88 residues) is calculated with a quite high number of experimental restraints, the NOE restraints and the 3-bond J-coupling restraints add up to 23 restraints per residue. As to be expected, the reduction of the number of structural restraints is paralleled by a reduction of the quality of the structures (Fig. 5). However, the quality of the structures (measured by the RMSD to the "true" structure) initially decreases only slowly and only after removing 75% of the experimental restraints a strong deterioration of the structural quality can be observed. However, the number of experimental restraints per residue now is seven restraints per residue. The standard simulated annealing protocol does not find a unique tertiary structure when more than 90% of the experimental restraints are removed. In general, the structures significantly improve by the use of substitute restraints, especially when including also the corresponding original NOE restraints and excluding the available 3-bond J coupling.

Structure improvement of ubiquitin

As a third test system an X-ray structure of a protein was used since X-ray structure are often thought to be superior to NMR-structures. We used the structure of human

Table 5 Refinement of NMR structures of HPr(H15A)

% removed	0	5	15	25	35	45	55	65	75	85
NOE Restraints	2,325	2,209	1,976	1,744	1,511	1,279	1,046	814	581	349
RMSD of the backbone N atoms to the mean of each bundle (nm)										
S_{2_HPr}	0.028	0.028	0.030	0.027	0.040	0.039	0.064	0.065	0.123	0.176
S_{3_HPr}	0.006	0.013	0.009	0.006	0.009	0.011	0.015	0.018	0.032	0.046
S_{4_HPr}	0.008	0.006	0.005	0.008	0.008	0.011	0.014	0.021	0.039	0.044
S_{5_HPr}	0.008	0.008	0.009	0.009	0.012	0.015	0.030	0.023	0.033	0.060
Average pairwise RMSD of the N, C^{α}, C atoms to the original structure (nm)[a]										
S_{2_HPr}	0.046	0.049	0.050	0.058	0.082	0.075	0.124	0.108	0.245	0.287
S_{3_HPr}	0.039	0.043	0.044	0.056	0.070	0.063	0.093	0.092	0.210	0.224
S_{4_HPr}	0.039	0.042	0.042	0.055	0.068	0.061	0.086	0.086	0.204	0.216
S_{5_HPr}	0.039	0.046	0.047	0.057	0.082	0.074	0.144	0.102	0.223	0.253
Ramachandran plot analysis of φ, ψ (%)										
S_{2_HPr}										
Most favoured	89.7	89.7	91.0	84.6	85.9	88.5	79.5	66.7	46.2	44.9
Additional allowed	10.3	10.3	7.7	14.1	10.6	11.5	16.7	29.9	32.1	42.3
Generously allowed	0.0	0.0	0.0	1.3	2.6	0.0	1.3	3.8	11.5	9.0
Disallowed	0.0	0.0	1.3	0.0	1.3	0.0	2.6	2.6	10.6	3.8
S_{3_HPr}										
Most favoured	92.3	93.4	92.3	89.7	91.0	87.2	92.7	79.5	70.5	65.4
Additional allowed	7.7	6.4	7.7	10.3	9.0	12.8	7.7	16.7	23.1	28.2
Generously allowed	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.8	6.4
Disallowed	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.8	2.6	0.0
S_{4_HPr}										
Most favoured	91.0	92.3	92.3	89.7	91.0	87.2	88.5	80.8	76.9	62.8
Additional allowed	9.0	7.7	7.7	10.3	9.0	12.8	11.5	12.8	17.9	30.8
Generously allowed	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.1	3.8	5.1
Disallowed	0.0	0.0	0.0	0.0	0.0	0.0	1.3	1.3	1.3	1.3
S_{5_HPr}										
Most favoured	92.3	92.3	92.3	91.0	88.5	88.5	80.8	78.2	71.8	61.5
Additional allowed	7.7	7.7	7.7	9.0	9.0	11.5	17.9	15.4	21.8	32.1
Generously allowed	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.8	5.1	5.1
Disallowed	0.0	0.0	0.0	0.0	0.0	0.0	1.3	2.6	1.3	1.3
NMR *R*-factor										
S_{2_HPr}	0.182	0.195	0.199	0.205	0.213	0.211	0.249	0.251	0.299	0.389
S_{3_HPr}	0.181	0.194	0.195	0.205	0.191	0.206	0.226	0.246	0.278	0.304
S_{4_HPr}	0.187	0.184	0.187	0.201	0.194	0.203	0.223	0.216	0.268	0.310
S_{5_HPr}	0.189	0.197	0.183	0.196	0.191	0.202	0.243	0.230	0.272	0.325
Violated NOE restraints (%)[b]										
S_{2_HPr}	1.15	0.49	0.73	1.14	1.93	1.90	3.47	4.44	9.49	12.78
S_{3_HPr}	0.53	0.74	0.86	1.28	1.74	1.84	2.75	3.44	7.46	8.87
S_{4_HPr}	0.30	0.41	0.56	1.16	1.52	1.39	2.56	3.31	7.11	8.21
S_{5_HPr}	0.33	0.50	0.65	1.14	1.76	1.72	3.22	3.35	7.71	9.26

S_{1_HPr}, NMR structural bundle calculated from 69 ^{3}J H^{N}–H^{α} coupling constants and 2,325 NOE distance restraints (R_{1_HPr}); S_{2_HPr}, NMR bundles calculated from sets of restraints (R_{2_HPr}) obtained after random removal of a given percentage of the original NOE restraints; S_{3_HPr}, NMR bundles calculated with the substitute restraints (R_{3_HPr}) extracted from their corresponding bundles S_{2_HPr}; S_{4_HPr}, structures calculated from R_{3_HPr} and the corresponding NOE restraints R_{2_HPr}; S_{5_HPr}, structures calculated from R_{3_HPr}, the corresponding NOE restraints R_{2_HPr} and the original ^{3}J H^{N}–H^{α} coupling

[a] Lowest energy structure in the original structural bundle of HPr(H15A) is considered as the reference structure to fit and calculate the corresponding RMSD values of the new sets of structures by MOLMOL

[b] Percentage of violated NOEs from the corresponding experimental set R_{2_HPr} with violations >0.05 nm

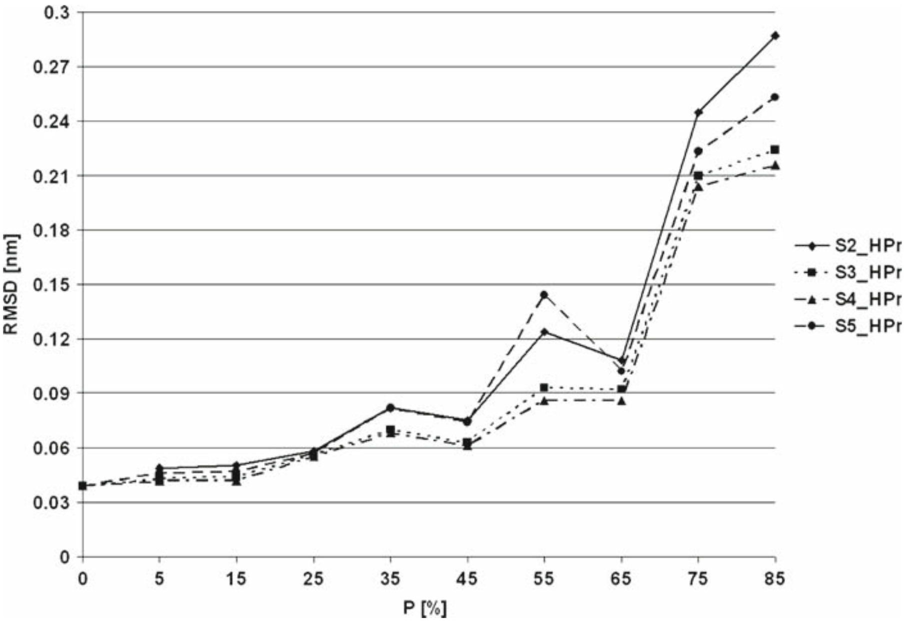

Fig. 5 Deviation of the calculated structure from the true structure of HPr(H15A). The pairwise RMSD values of the obtained structures to the lowest energy structure of the original data set is plotted as a function of the percentage P of removed restraints: (S_{2_HPr}, *solid line*) conventional calculation, (S_{3_HPr}, *dashed line*) calculation using substitute restraints only, (S_{4_HPr}, *dotted line*) calculation using substitute restraints together with experimental NOE restraints, (S_{5_HPr}, *dash-dot line*) calculation using substitute restraints together with all available experimental restraints

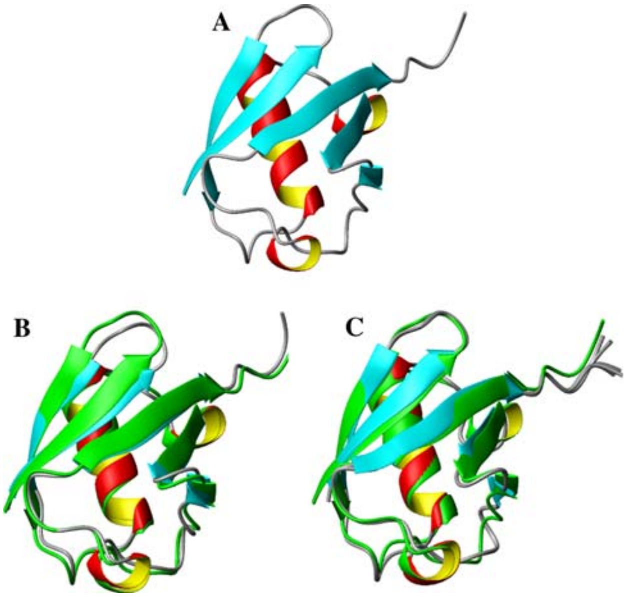

Fig. 6 The different structural bundles shown were calculated with the restraints described in Table 6. **a** X-ray structure 1UBQ, **b** Overlay of 1UBQ in green and ubiquitin bundle $S_{3\_UBQ\_1000}$ using additional substitute restraints, **c** Overlay of 1UBQ in green and ubiquitin bundle $S_{3\_UBQ\_498}$ using additional substitute restraints

erythrocytic ubiquitin (76 residues) at 1.8 Å resolution (PDB ID: 1UBQ, Vijay-kumar et al. 1987) as target structure. As described in Fig. 6 an artificial "experimental" set of restraints was generated from the structure. The set of artificial NOE restraints contained intraresidual, sequential (i, $i+1$), medium range (i, $i+j$; $1<j\leq 4$) and long range (i, $i+j$; $4<j$) restraints, considering an upper distance limit as 0.6 nm (Table 6). In addition,

Table 6 PERMOL parameters to generate NOE distance restraints in Ubiquitin test

NOE distances Ubiquitin			
Selected atoms	H^{α}, H^{N} and all sidechain hydrogens		
Distance range (nm)	0.18–0.6 nm		
Confidence level (%)	99.9		
	R_{1_UBQ}	$R_{2\_UBQ\_1000}$	$R_{2\_UBQ\_498}$
Number of restraints			
Intraresidual	1,692	450	225
Sequential (i, $i+1$)	1,586	225	112
Medium range(i, $i+j$; $1<j\leq 4$)	1,632	100	50
Long range (i, $i+j$; $4<j$)	3,845	225	112
Total	8,755	1,000	498

dihedral angle and H-bond restraints were extracted and used as additional "experimental" restraints. This set of restraints R_{1_UBQ} was employed to calculate the starting bundle S_{1_UBQ} by CNS. As already done for HPr(H15A) from this NOE distance restraint list (containing 8,755 distance restraints), a given number of restraints were randomly deleted in such a way that the restraints were still distributed as typical in the different distance classes (45% intra-residual, 22.5% sequential, 10% medium range, and 22.5% long range). Thus, one set of "experimental" restraints $R_{2\_UBQ\_1000}$ containing 1000 "experimental" NOE distance restraints (15 NOE restraints per residue to have a good structure in CNS calculations) and a second set $R_{2\_UBQ\_498}$ containing 498 "experimental" NOE distance restraints were employed together with the complete set of

Table 7 Number of restraints used in CNS MD calculations and quality values for Ubiquitin test

	1UBQ	S_{1_UBQ}	$S_{2\_UBQ\_1000}$	$S_{3\_UBQ\_1000}$	$S_{2\_UBQ\_498}$	$S_{3\_UBQ\_498}$
NOE restraints	–	8,755	1,000	1,000 + 9,223	498	498 + 8,418
H-bond restraints	–	53	53	37	53	26
Dihedral angle restraints	–	350	350	334	350	336
RMSD of N-atoms to the mean (nm)[a]	–	0.018	0.015	0.000	0.022	0.015
RMSD of bb atoms to the X-ray structure (nm)[b]	–	0.080	0.098	0.094	0.123	0.118
Ramachandran plot[c]	95.5%	87.9%	90.1%	89.4%	86.4%	86.4%
	4.5%	12.1%	9.1%	10.4%	13.6%	13.6%
Crystallographic *R*-factor[d]						
R-factor (w + t)	0.194	0.221	0.254	0.233	0.291	0.287
R-factor (w)	0.191	0.218	0.250	0.230	0.286	0.283
Free *R*-factor	0.251	0.277	0.330	0.307	0.398	0.376

[a] RMSD values of the backbone N atoms to the mean averaged structure of the 10 lowest energy structures calculated by the program MOLMOL

[b] Average pairwise RMSD values of the backbone atoms (N, C^{α}, C) of the 10 lowest energy structures to the X-ray structure (1UBQ as reference) calculated by the program AUREMOL

[c] Percentages of residues in the most favoured and allowed regions of the Ramachandran plot calculated by the program Procheck

[d] Refmac tool from the CCP4 software package calculates the three crystallographic *R*-factors: *R*-factor (working + test set), *R*-factor (working set) and free *R*-factor; The first factor is defined as $\Sigma|F_{\mathrm{obs}} - F_{\mathrm{calc}}|/\Sigma F_{\mathrm{obs}}$, ($F_{\mathrm{obs}}$, experimental structure factor and F_{calc}, structure factor calculated from the model), refining against the complete dataset (all F_{obs}). The free *R*-factor (R-free) (Brunger 1992) is calculated for a random subset (4.7%) of the dataset that is set aside and labelled the *test set*. The remaining 95.3% of the dataset (*working set*) is used to form the target function for refinement and to compute the traditional crystallographic *R*-factor. All the *R*-factor values refer to the first structure of the bundle

dihedral angle and H-bond restraints to calculate 500 structures using a simulated annealing protocol. The 10 lowest energy structures were selected to define the starting bundles (S_{2_UBQ}). From these bundles, a set of substitute distance, dihedral angle and hydrogen bond restraints (R_{3_UBQ}) was created. Employing these substitute restraints together with the corresponding "experimental" NOE distance restraints from the corresponding R_{2_UBQ} in CNS molecular dynamics calculation, 500 structures were calculated, selecting the 10 lowest energy ones to define the final improved bundle (S_{3_UBQ}).

The deviation of the backbone atom positions of the recalculated structural bundles from the X-ray structure is always smaller when substitute restraints are used (Table 7). However, the effect is smaller then that observed for Byr2 or HPr(H15A), most probably because still a large number of "experimental" dihedral angles and hydrogen bonds were retained in the calculation. Since our target structure is a X-ray structure, X-ray *R*-factors can be used to compare the quality of the resulting structures. The calculation of the crystallographic *R*-factor from the CNS structures was supported by a new tool in AUREMOL that converts NMR output pdb-files to the correct format of X-ray structures including all crystallographic information and the correct orientation in the unit cell. The advantage of the use of X-ray *R*-factors is that they are better defined than NMR *R*-factors because the diffraction data are essentially free of noise and artifacts. In addition, free *R*-factors can also be calculated more reliably since the number of diffraction signals is much larger than NOESY signals in NMR. As observed for the other examples studied here the *R*-factors (and especially the free *R*-factor) decreases when substitute restraints are used. The same is true for the data independent quality measures (Table 7); they improve when substitute restraints are used.

The advantage of data imputation

Compared to X-ray crystallography, NMR data are always incomplete and are not sufficient to obtain structures with the same precision without additional information. Therefore, a physical model is always required for the structural calculation and the obtained structures depend on the parametrisation and approximation implemented in the given program. Experience show that structural bundles obtained with different MD programs (e.g., CNS used here and CYANA) give different results. This effect can easily be seen in the our structure calculations performed with ubiquitin: although an almost ideal "experimental" data set with 8,755 NOEs, 53 hydrogen bonds, 350 dihedral angles together with a physical model (part of the molecular dynamics program) was used, the RMSD of the obtained bundle to the original structure was still 0.08 nm. Although procedures have been introduced for automated interpretation of the thousands of cross peaks in such NOE spectra, their success depends on the quality and quantity of the spectral data. Obtaining 115 NOEs per residue is far away from the real situation. This paper proposes a data imputation technique to substitute/replace missing experimental data (that partly cannot be obtained experimentally) by combining the well-developed molecular dynamics calculations (simulated annealing procedures optionally combined with refinement in explicit water, performed in our case with CNS) with additional data extracted from a set of structural models that are in agreement with the available experimental data. As common in imputation techniques mean values and error distributions are calculated in the high-dimensional space of conformational restraints (distances, dihedral angles and hydrogen-bond distance restraints). We show that the method improves the quality of the structural bundle in the three different examples studied as verified by a closer RMSD from the "true" structure. In addition, generally a better agreement with the experimental data used to calculate the "true" structure is obtained as can be verified on the basis of the NMR or X-ray *R*-factors.

Selection of data to be imputed

In principle, the selection of the substitute restraints used for the structural calculation will also influence the outcome of the method. The use of all possible restraints such as all pairwise distances will lead to a too large number of restraints that cannot be handled by the existing MD programs successfully. Therefore, one has to restrict to a smaller set of substitute restraints. We tested a number of plausible combinations, the selection used here proved most successful. The use of dihedral angle restraints together with hydrogen bonds for defining the local structures together with distance restraints between all hydrogen atoms in a sphere of 0.6 nm corresponds closely to the situation found in excellent NMR-data. However, intraresidual contacts were omitted since they contain not much additional information. In addition, long range distance restraints were allowed for all pairs of H^{α}-atoms in a distance range smaller than 1.5 nm, information that cannot be obtained by NOEs but has similarities with that obtained by paramagnetic relaxation enhancement measurements. It could be worth to introduce also information on directions in an internal coordinate system, similar to that obtained from residual dipolar couplings. However, this is outside of the scope of the present paper.

Validity of the MAR-condition

The validity of the MAR (missing-by-random) condition increases the probability that data imputation techniques can be used successful but it is in general not required when assumptions about the mechanism of the incomplete sampling can be made (Rubin 1976). However, when we start with the consideration that in principle a NOESY-spectrum of a well-folded protein represents all proton distances existing but that only a subset is really assigned or measurable because of the signal-to-noise ratio or spectral artifacts, we can assume that the MAR condition is rather well-fulfilled. In our test cases we removed randomly distance restraints, here clearly the MAR condition is fulfilled. A practical case where the MAR condition is not fulfilled strictly would be the case where in a part of the protein the resonances are exchange broadened and therefore not visible. Here, data imputation as we propose it has simply no effect because the conformational space is not restricted by our restraint definition. This would be different when much larger error probabilities than 0.1% were accepted.

Conclusions

The application of data imputation techniques to NMR structure determination appears logical when we consider the fact that NMR data are always sparse when compared to X-ray crystallography. Traditionally, the missing data are partly substituted by using a more or less complex physical model for the structure calculation. The most powerful physical model is provided by including explicit water in the calculation (Linge et al. 2003). We propose here a method that can (and should be used) in addition to already existing method, a model based data imputation method. The substitute restraints calculated here are used together with the original data. As in most bootstrapping methods different sets of data are generated and analysed. In our case we characterise these multidimensional sets of data (structural restraints) by their means and their variations.

Data imputation techniques do not grant an improvement of the parameter estimation (in our case the three-dimensional structure) but only lead to an improvement in the majority of the experimental data sets. We could show in our examples that the obtained structures almost always improve when considering the deviation from the "true" structure and the *R*-factors. In fact, in our test cases an improvement is always observed when exclusively the substitute restraints are used. However, the inclusion of the experimental data usually leads to better results.

The use of substitute restraints can also suppress possible inconsistencies in the experimental data because the number of substitute restraints usually is much larger than the experimental restraints and may dominate a single inconsistent restraint. In one case the inclusion of dihedral angle restraints from 3-bond J-couplings gave non-optimal results in HPr(H15A), probably because of conflicting experimental data. Wrong experimental dihedral angle restraints cannot be cured by the torsional angle substitute restraints since their numbers are almost equal. Although one could draw the conclusions that the experimental dihedral angle restraints should not be included in the calculation, a general strategy would include all experimental restraints unless one can directly show that a restraint is erroneous.

Since data imputation does not grant a more correct solution, a critical analysis of the results is recommended in literature. Applied to the actual case, the quality of the obtained structures has to be checked as it is done in classical structure determination. Here, the NMR-*R*-factor calculated directly from the NOESY spectra is an important parameter since of course the deviation from the "true" structure cannot be used in practical cases.

Acknowledgments This work was supported by the Alexander von Humboldt Foundation, the Fonds of the Chemical Industry (FCI), the German Research Foundation (DFG), Bruker Biospin, and the Ministry for Education and Research (BMBF). We thank Dr. Ingrid Vetter for the help with calculation of the crystallographic *R*-factors.

References

Alexandrescu AT (2004) Strategy for supplementing structure calculations using limited data with hydrophobic distance restraints. Proteins 56:117–129

Angyan AF, Perczel A, Pongor S, Gaspari Z (2008) Fast protein fold estimation from NMR-derived distance restraints. Bioinformatics 24:272–275

Aszodi A, Gradwell MJ, Taylor WR (1995) Global fold determination from a small number of distance restraints. J Mol Biol 251:308–326

Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR (2000) The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. J Comput Biol 7:537–558

Bowers PM, Strauss CE, Baker D (2000) De novo protein structure determination using sparse NMR data. J Biomol NMR 18:311–318

Brunger AT (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355:472–475

Brunger AT (2007) Version 1.2 of the crystallography and NMR system. Nat Protoc 2:2728–2733

Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS et al (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921

Brunner K, Gronwald W, Trenner JM, Neidig KP, Kalbitzer HR (2006) A general method for the unbiased improvement of solution NMR structures by the use of related X-ray data, the AUREMOL-ISIC algorithm. BMC Struct Biol 6:14

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104:9615–9620

Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 120:6836–6837

Döker R, Maurer T, Kremer W, Neidig K, Kalbitzer HR (1999) Determination of mean and standard deviation of dihedral angles. Biochem Biophys Res Commun 257:348–350

Elsner R (2006) NMR-basierte Aufklärung der Strukturen von Ras-bindedomänen und ihrer Wechselwirkungen mit den kleinen GTPasen H-Ras adn Rap1A. Dissertation, University of Regensburgs

Epron B (1979) Bootstrap methods: another look at the jackknife. Ann Statist 7:1–26

Fuentes G, Nederveen AJ, Kaptein R, Boelens R, Bonvin AM (2005) Describing partially unfolded states of proteins from sparse NMR data. J Biomol NMR 33:175–186

Garrett DS, Kuszewski J, Hancock TJ, Lodi PJ, Vuister GW, Gronenborn AM, Clore GM (1994) The impact of direct refinement against three-bond HN-C alpha H coupling constants on protein structure determination by NMR. J Magn Reson B 104:99–103

Gronwald W, Kirchhofer R, Gorler A, Kremer W, Ganslmeier B, Neidig KP, Kalbitzer HR (2000) RFAC, a program for automated NMR R-factor estimation. J Biomol NMR 17:137–151

Gronwald W, Huber F, Grunewald P, Sporner M, Wohlgemuth S, Herrmann C, Kalbitzer HR (2001) Solution structure of the Ras binding domain of the protein kinase Byr2 from *Schizosaccharomyces pombe*. Structure 9:1029–1041

Gronwald W, Brunner K, Kirchhöfer R, Nasser A, Trenner J, Ganslmeier B, Riepl H, Ried A, Scheiber J, Elsner R et al (2004) AUREMOL, a new program for the automated structure elucidation of biological macromolecules. Bruker Rep 154(155): 11–14

Guntert P (2004) Automated NMR structure calculation with CYANA. Methods Mol Biol 278:353–378

Herrmann T, Guntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227

Ilari A, Savino C (2008) Protein structure determination by X-ray crystallography. Methods Mol Biol 452:63–87

Kim Y, Prestegard JH (1990) Refinement of the NMR structures for acyl carrier protein with scalar coupling data. Proteins 8:377–385

Kolinski A, Skolnick J (1998) Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins 32:475–494

Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14(51–55):29–32

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486

Latek D, Ekonomiuk D, Kolinski A (2007) Protein structure prediction: combining de novo modeling with sparse experimental data. J Comput Chem 28:1668–1676

Li W, Zhang Y, Kihara D, Huang YJ, Zheng D, Montelione GT, Kolinski A, Skolnick J (2003) TOUCHSTONEX: protein structure prediction with sparse NMR data. Proteins 53:290–306

Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M (2003) Refinement of protein structures in explicit solvent. Proteins 50:496–506

Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes BD, Wright PE, Wüthrich K (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids—IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. Eur J Biochem 256:1–15

Möglich A, Weinfurtner D, Gronwald W, Maurer T, Kalbitzer HR (2005) A restraint molecular dynamics and simulated annealing approach for protein homology modeling utilizing mean angles. BMC Bioinform 6:91–104

Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53:240–255

Pardi A, Billeter M, Wüthrich K (1984) Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3JHN alpha, in a globular protein. Use of 3JHN alpha for identification of helical secondary structure. J Mol Biol 180:741–751

Qu Y, Guo JT, Olman V, Xu Y (2004) Protein fold recognition through application of residual dipolar coupling data. Pac Symp Biocomput 2004:459–470

Rathinavelan T, Im W (2008) A novel strategy to determine protein structures using exclusively residual dipolar coupling. J Comput Chem 29:1640–1649

Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. Science 309:303–306

Rieping W, Habeck M, Bardiaux B, Bernard A, Malliavin TE, Nilges M (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 23: 381–382

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Rubin DB (1981) The Bayesian bootstrap. Ann Statist 9:130–134

Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7:147–177

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A et al (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690

Sikorski A, Kolinski A, Skolnick J (2002) Computer simulations of protein folding with a small number of distance restraints. Acta Biochim Pol 49:683–692

Skolnick J, Kolinski A, Ortiz AR (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 265:217–241

Smith-Brown MJ, Kominos D, Levy RM (1993) Global folding of proteins using a limited number of distance constraints. Protein Eng 6:605–614

Standley DM, Eyrich VA, Felts AK, Friesner RA, McDermott AE (1999) A branch and bound algorithm for protein structure refinement from sparse NMR data sets. J Mol Biol 285:1691–1710

Tang C, Clore GM (2006) A simple and reliable approach to docking protein–protein complexes from very sparse NOE-derived intermolecular distance restraints. J Biomol NMR 36:37–44

Tjandra N, Omichinski JG, Gronenborn AM, Clore GM, Bax A (1997) Use of dipolar ^{1}H–^{15}N and ^{1}H–^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution. Nat Struct Biol 4:732–738

Tolman JR, Al-Hashimi HM, Kay LE, Prestegard JH (2001) Structural and dynamic analysis of residual dipolar coupling data for proteins. J Am Chem Soc 123:1416–1424

Torda AE, Brunne RM, Huber T, Kessler H, van Gunsteren WF (1993) Structure refinement using time-averaged J-coupling constant restraints. J Biomol NMR 3:55–66

Vijay-kumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol 194:531–544

Wishart D (2005) NMR spectroscopy and protein structure determination: applications to drug discovery and development. Curr Pharm Biotechnol 6:105–120

Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York

Wüthrich K (1990) Protein structure determination in solution by NMR spectroscopy. J Biol Chem 265:22059–22062

Zweckstetter M (2008) NMR: prediction of molecular alignment from structure using the PALES software. Nat Protoc 3:679–690