Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins

Andrea Frank, Ionut Onila, Heiko M. Möller, and Thomas E. Exner*

Department of Chemistry and Zukunftskolleg, University of Konstanz, Konstanz D-78457, Germany

ABSTRACT

Despite the many protein structures solved successfully by nuclear magnetic resonance (NMR) spectroscopy, quality control of NMR structures is still by far not as well established and standardized as in crystallography. Therefore, there is still the need for new, independent, and unbiased evaluation tools to identify problematic parts and in the best case also to give guidelines that how to fix them. We present here, quantum chemical calculations of NMR chemical shifts for many proteins based on our fragment-based quantum chemical method: the adjustable density matrix assembler (ADMA). These results show that ^{13}C chemical shifts of reasonable accuracy can be obtained that can already provide a powerful measure for the structure validation. ^1H and even more ^{15}N chemical shifts deviate more strongly from experiment due to the insufficient treatment of solvent effects and conformational averaging.

Key words: protein structure; nuclear magnetic resonance chemical shift calculation; quantum mechanics; fragment-based approach; evaluation.

INTRODUCTION

Knowledge about the three-dimensional (3D) structure (in atomic resolution) of proteins as well as protein–protein and protein–ligand complexes is a precondition for the indepth understanding of biological processes and rational manipulations of these. 3D structure determination by nuclear magnetic resonance (NMR) is well established and an emerging alternative to X-ray crystallography even for larger proteins. Advantages are that sometimes very time-consuming search for optimal crystallization conditions can be skipped and that structures in solution and not in an “artificial” crystal environment are obtained. For protein structure calculation, distance restraints derived from nuclear Overhauser enhancement (NOE) data still play the key role. Additionally, angle restraints derived from experimental and/or statistical data as well as orientational information from residual dipolar couplings (RDC), chemical shift anisotropy (CSA), or heteronuclear relaxation parameters can be used.^{1,2} The task to find the best structure or the best structures consistent with the derived restraints is highly supported by computational methods in which a global optimization procedure is applied using an objective function to score different structures against each other. Objective functions as implemented, for example, in the programs, DYANA,³ CYANA,⁴ or XPLOR,⁵ comprise terms that score the different restraint violations as well as terms that reflect a reasonable covalent geometry and are usually optimized by simulated annealing by molecular dynamics calculations in either Cartesian⁵ or torsion-angle space.^{3–5} These methods usually have to deal with ambiguous and/or erroneous data. Especially, the assignment of NOE crosspeaks often results in highly ambiguous distance constraints. Thus, 3D protein structure determination is a very complex optimization problem in which the atomic positions as well as the NOE assignments have to be determined simultaneously.

Despite the many successfully solved protein structures described in the literature and obtainable from the Protein Data Bank,⁶ there is still debate about the reliability of these models as in some cases wrong folds were proposed as highlighted in a recent paper.⁷ Even if in this publication extreme cases are described, smaller errors in local regions are probably present in many other published structures resulting from insufficient experimental data, as a consequence of an insufficient optimization procedure, or because of human bias during chemical shift assignment and/or structure calculation. As subtle changes can have a strong influence, for example, on protein–protein or protein–ligand interactions, wrong conclusions could be drawn from these almost correct structures. However, quality control of NMR structures is by far not as well established

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

Grant sponsors: Konstanz Research School Chemical Biology (KoRS-CB), Zukunftskolleg of the Universität Konstanz, Junior-professoren-Programm of the state Baden-Württemberg, ERASMUS program of the European Union.

*Correspondence to: Thomas E. Exner, Department of Chemistry and Zukunftskolleg, University of Konstanz, Konstanz D-78457, Germany. E-mail: thomas.exner@uni-konstanz.de

and standardized as in crystallography not to the least because NMR structures are inherently underdetermined such that Rfree-calculations are not possible with similar statistical relevance as for X-ray structures.⁸ Therefore, there is still the need for new, independent, and unbiased evaluation tools to identify problematic parts and in the best case also to give guidelines that how to fix them. NMR chemical shifts are predestined for this purpose as usually these are only indirectly used during structure determination. The comparison of experimental with theoretically calculated values, as an independent quality measure, can be used to identify inconsistency in the protein models. The advantages of chemical shifts are that they are very sensitive to structural changes and that they not only give an evaluation of the global structure but also give information on the spatial distribution of good and bad agreement.

Chemical shifts have served as input for the structure calculation and the validation since many years; however, established methods rely on statistical data derived from a limited set of high-quality 3D structures (TALOS,⁹ TALOS+¹⁰ PREDITOR¹¹) and, in part, make use of empirical or semiempirical equations to account for the nonsequential environment (SHIFTX,¹² SHIFTX2,¹³ SHIFTS,^{14,15} and SPARTA+¹⁶). These methods are usually quite successful in predicting backbone chemical shifts, which are nicely sensitive to the local secondary structure but are not so well suited to predict effects of the tertiary structure to assess the 3D packing and distinguish solvent exposed from core regions.

On the one hand, NMR calculations for small molecules are nowadays done on a routine basis. For good agreement with experiment, large basis sets and high levels of theory are needed. On the other hand, calculations for proteins are still out of reach for standard methods due to their sheer size. Even if the first trials to calculate chemical shifts of proteins using quantum mechanical methods on small model systems started relatively early,^{17–19} only the more recent publications are really relevant for the work proposed here. Scheraga and co-workers^{20–24} developed a method for protein NMR structure determination, refinement, and validation based on quantum chemical ¹³C_α chemical shifts calculated for a large number of different conformations of the central residues in tripeptides. These shifts are mainly dependent on the secondary structure so that the inclusion of additional parts of the molecule is not that important. Jacob and Visscher²⁵ calculated NMR chemical shifts with the frozen-density embedding (FDE) scheme originally introduced by Wesolowski and Warshel.²⁶ FDE is based on a partitioning into separate subsystems, which are independently calculated, with an effective embedding potential modeling the effects of the other, frozen subsystems. Linear combinations of the relevant derivatives of the energies of smaller fragments are used in the approach of Lee and Bettens.²⁷ Johnson and DiLabio applied their

mixed quantum mechanical/molecular mechanical (QM/MM) algorithm (B3LYP density functional, aug-cc-pVTZ basis set, and quantum capping potentials) to calculate chemical shifts of the single amino acid gly39 of the fungal dockerin domain.²⁸ Even though they get good agreement with the full quantum chemical calculations of this small protein, the results are very far to the experiment. But as only four values were compared, the statistical significance is questionable. In contrast, He *et al.*²⁹ published a very similar method, which they called automated fragmentation quantum mechanics/molecular mechanics (AF-QM/MM), which was able to give excellent agreement with the full calculations and also a good correlation with experimental values. As the name already implies, the fragmentation is done automatically, which is a great advantage for out-of-the-box usage. Especially, the success of this latter method gave us confidence that it may be very useful to apply our knowledge also in fragment-based quantum chemistry to NMR calculations. Therefore, we will present here quantum chemical calculations of NMR chemical shifts for many proteins based on our fragment-based quantum chemical method: the adjustable density matrix assembler (ADMA).^{30–36} These results show that ¹³C chemical shifts of reasonable accuracy can be obtained that can already provide a powerful measure for the structure validation. ¹H and even more ¹⁵N chemical shifts deviate more strongly from experiment for reasons that will be discussed. Future directions and developments will be highlighted.

MATERIALS AND METHODS

Ab initio calculations of chemical shifts, mainly done for small molecules, are based on the chemical shielding tensor that describes the relative change in the local magnetic field at the nucleus position relative to the external magnetic field.³⁷ The components of the tensor are given as the mixed second derivative of the total electronic energy of the molecule with respect to the external magnetic field *B* and the magnetic moment of the nucleus of interest *μ*.

$$\sigma_{\alpha\beta} = \frac{\partial^2 E}{\partial \mu_\alpha \partial B_\beta} \quad (1)$$

Using perturbation theory, these values can be calculated with a large number of theories including Hartree-Fock (HF) and post-HF methods as well as density functional theory (DFT). One problem occurring because only a finite basis can be used in the expansion of the wave function or electron density is the lack of gauge invariance of the calculated chemical shieldings, the so-called gauge problem. Different approaches have been developed to overcome this problem. For a description of these, we refer to Refs. 37, 66 and references therein. In

this work, we will only use the GIAO (gauge invariant or including atomic orbitals) approach.

The chemical shift tensor δ is then defined by

$$\delta = 1\sigma_{\text{iso}} - \sigma \quad (2)$$

where σ is the chemical shielding tensor, 1 is the unit matrix, and σ_{iso} is the isotropic value or trace of the chemical shielding of the standard reference used in the NMR experiment, which is obtained from a separate calculation of the reference, for example, tetramethylsilane. The isotropic chemical shift comparable with the experimental spectra is given as the trace of the chemical shift tensor.

To do such calculations for proteins, the system has, as outlined in the "Introduction" section and the publications cited therein, to be subdivided into smaller parts that resemble the complete protein as good as possible but can be treated by standard quantum chemical calculations in reasonable time. In recent years, we developed and optimized a fragment-based quantum chemical approach, called ADMA,^{30–36} which is exactly designed to combine these two features. The theory of ADMA and the enhanced field-adapted ADMA (FA-ADMA) and applications to different physicochemical properties are described elsewhere^{30–36}, and we will give only a short summary here. An ADMA calculation is started with the subdivision of the target molecule into a set of m mutually exclusive families of nuclei defining the molecular fragments. For each of these, a parent molecule is constructed by surrounding the fragments by additional regions, called surroundings in the following discussion, with the same local nuclear geometry as in the macromolecule up to a specified distance d . The accuracy of ADMA is solely determined by this distance. Putting point charges on all atoms not explicitly included in the QM calculations, as done in FA-ADMA, resulted in an extreme increase in the accuracy for highly polar and formally charged molecules. Standard quantum chemical calculations are performed for the parent molecules. Fragment electron density matrices are extracted from these calculations and combined to get an approximation of the complete density matrix of the macromolecule, from which the total energy of the target molecule can be calculated.

To calculate NMR chemical shifts, derivatives of the ADMA energies would have to be calculated, which is impossible for larger proteins. However, the chemical shifts can also be directly taken from the fragment calculations. This is a reasonable approximation, as chemical shifts are mainly influenced by the local surroundings in contrast to the total energy. Thus, for the studies presented here, we took the fragmentation algorithm from the ADMA method and did NMR calculations for the resulting parent molecules as described above. The fragmentation procedure is depicted in Figure 1. A specific

Figure 1

Schematic 2D representation of the fragmentation procedure. (a) Spheres with a radius d are placed on each of the atoms of a molecular fragment (red part of the molecule). Each atom inside of one of these spheres is selected (blue parts). (b) To get reasonable molecules, the complete side-chain and/or backbone of an amino acid are included in the surroundings if at least one atom is selected in the previous step (blue and green parts). Additionally, the CO group of the previous and the NH group of the following amino acid are also added (magenta parts). (c) The rest of the molecule is cut off and the broken bonds are saturated by hydrogen capping atoms resulting in the parent molecule is shown.

Figure 2

Average (solid fill) and maximum (texture fill) errors of the fragment-base calculations using HF theory for the ^{13}C and ^1H chemical shifts of the trp-cage miniprotein compared with full QM.

fragment is composed of the backbone or the side chain atoms of one amino acid. On each of these atoms, a sphere with a radius corresponding to the distance criterion d is placed. All atoms lying inside of at least one of these spheres are added to the surroundings. In a second step, complete side chains or backbones of single amino acids, with at least one atom (side-chain or backbone atom, respectively) already part of the surroundings from the first round, are included. Additionally, the CO group of the previous and the NH group of the following amino acid are also added to each continuous backbone piece. All other parts are removed from the parent molecule of the specific fragment and the broken bonds are saturated with capping hydrogen atoms. Please note that the distance criterion is calculated not only along the bonding network but also through space as the Euclidean distance to the closest fragment atom. In this way, parent molecules can be composed of multiple disconnected molecular parts. This is especially important for NMR calculations as aromatic systems have a large impact on the chemical shifts even if they are separated in sequence but close in space.

For the parent molecules, NMR calculations are performed using the GIAO method^{38–42} within the Gaussian03 program package.⁴³ If not otherwise stated, DFT with the B3LYP functional⁴⁴ and the 6-31g(d) basis set^{34,45–52} were used. The partial charges representing the additional part of the molecule in FA-ADMA were generated with the Gasteiger–Hückel method.^{53–56} To include polarization effects of the surrounding water, some calculations were repeated with an implicit solvent model (IEF-PCM^{57–59}) based on the self-consistent reaction field (SCRF). Tetramethylsilane for ^1H and ^{13}C and ammonia for ^{15}N calculated at the same level of theory were taken as reference. From each parent molecule calculation, the chemical shifts of the atoms of the central fragment are collected. As each atom is included in just one fragment, exactly one value is obtained for each atom, which can then be compared with the calcu-

lations carried out with the entire protein (if these full calculations are possible as in small peptides) and with experiments.

RESULTS AND DISCUSSION

Trp-cage miniprotein

The first calculations presented here are performed on the artificially designed trp-cage miniprotein (PDB entry 2JOF).⁶⁰ This eicosapeptide (DAYAQWLKDGPPSSGR PPPS) adopts (despite its small size) a well-defined 3D structure, which was determined by NMR spectroscopy. Therefore, it is the perfect test example as full quantum chemical calculations are possible, which can be compared with the fragment-based calculations as well as to the experiment, and was already used in the study by He *et al.*²⁹

For this small peptide, first a calculation of the full protein and then fragment calculations with varying distance criteria from 3 to 7 Å with the same level of theory (HF/6-31g(d)) were performed. The results are summarized in Figure 2. These results clearly demonstrate that, as expected, with increasing radius of the surroundings better and better agreement with the quantum chemical calculation of the full protein is obtained. Especially, the maximum error for ^{13}C chemical shifts drops dramatically when going from 5 to 6 Å to just above 2 ppm. Thus, we can conclude here that surroundings of 6 Å should be sufficient to get reliable results at least for ^{13}C chemical shift. In contrast, even if the absolute errors for ^1H chemical shifts are smaller, the maximum error of up to 0.7 and 0.4 ppm for 6 and 7 Å, respectively, may be too large for the validation of protein structures as these correspond to almost 10% of the total range of ^1H chemical shifts (in comparison to a maximum relative error of around 1% for the ^{13}C chemical shifts).

For the trp-cage miniprotein, only experimental ^1H chemical shifts are available from the Biological Magnetic

Figure 3

Average and maximum errors of calculated ^1H chemical shifts of the trp-cage miniprotein compared with experiment: Results for the full QM calculation using HF theory (labeled HF) and the fragment-based calculations with different distance criteria (3.0–7.0 HF) are shown. Additionally, the results using DFT without (B3LYP) and with (solv.impl) an implicit solvent model for the full protein are given. As hydrogens bound to nitrogen give very large errors due to the missing hydrogen network with water molecules, the values are also shown excluding these hydrogens.

Resonance Data Bank⁶¹ (BMRB, Entry 15169) and, thus, we can only compare our calculations to these. In Figure 3, the average and maximum errors are given. The comparison reveals unexpectedly high deviations from experimental values. However, as the full calculations have similar large errors, the failure cannot be explained by the fragment-based approach. Having a closer look, hydrogen atoms bonded to nitrogen can be identified as causing the major problems. This can easily be explained: the chemical shift of these atoms is highly dependent on the presence of the hydrogen-bonding network and on the degree of the individual H-bridge being formed. Yet our calculations are performed in vacuo in contrast to the experimental measurement in aqueous solution. As no explicit solvent is present, this network cannot be correctly predicted, and we neglect these chemical shifts in the following. For the remaining hydrogens, a reasonable agreement between experiment and calculation can be obtained (Fig. 4). It is interesting to note that increasing the distance criterion and even doing the calculations on the complete protein does not reduce the deviation to the experiment. Thus, different reasons than the fragment-based approach are causing the remaining errors.

One possible cause could be the fact that we used HF theory up to now, which neglects electron correlation. Therefore, we switched to DFT using the B3LYP functional. Additionally, polarization effects of the solvents can be approximated by an implicit solvent model. We used here the IEF-PCM model^{57–59} based on the SCRF approach. The average and maximum errors for the full protein calculations are also given in Figure 3 and show a reduction of the maximum error when including correlation effects and a significant improvement of the average errors when using the implicit solvent model. The spatial distribution of errors in the implicit solvent calculation mapped onto the 3D structure of trp-cage is

shown in Figure 5. The largest errors are located in a few very confined regions. Most of them are solvent exposed so that these are probably caused by insufficient representation of solvent effects. However, one region is centered on the aromatic system (Tyr3) in the upper central part of the figure. As the effect of the ring current in aromatic systems is strongly influencing the chemical shifts and is very position dependent, we tested if slightly changing the orientation of Tyr phenol ring would improve the calculations. For doing so, the ring was

Figure 4

Correlation between calculated (HF, full QM) and experimental ^1H chemical shifts of the trp-cage miniprotein (hydrogen atoms bound to nitrogen are not shown). [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://www.interscience.wiley.com).]

Figure 5

Spatial distribution of the errors in the ^1H chemical shifts calculated using DFT (B3LYP) and an implicit solvent model: the values are coded by the size of the spheres representing the atoms (larger spheres correspond to larger errors). [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://www.interscience.wiley.com).]

rotated by 5° in three different directions. The calculated chemical shifts are depicted in Figure 6. Even by this simplistic approach, the deviation of the atom with the largest error ($\text{H}_{\beta 2}$ of Tyr3) could be lowered from 1.203 to 1.080. However, some deviations increase as a result of this "conformational search." Thus, this simple approach of changing the conformation is not able to improve the overall agreement but a chemical-shift-guided exhaustive optimization of the structure could be advantageous and will be followed soon.

Concluding this first part, we can say that, even if there is space for improvement, the accuracy of the quantum chemical calculations is advancing into areas interesting for practical use. To prove that this performance can also be achieved for larger proteins, we next describe two well-investigated proteins from recent studies in the following sections: YuaF⁶² and p63 (Enthart *et al.*, unpublished data).

YuaF

The first larger example is a transmembrane protein from *Bacillus subtilis*, member of the NfeD-like clan with a potential role in maintaining membrane integrity during cellular stress.⁶² Its soluble C-terminal domain consists of 84 residues and was solved by Walker *et al.*⁶² using solution NMR spectroscopy (PDB entry 2K14, BMRB entry 15475). Because of its size, full QM calculations are not feasible in a reasonable amount of time

and, thus, we only compare our fragment-based calculations to the experimental chemical shift. To increase the speed of the calculations further, mixed basis sets, also called locally dense basis sets in the scope of NMR calculations, were applied with the 6-31g(d) and the 3-21g(d) basis set for the fragment atoms and the surroundings, respectively. Stereo-specific assignments were manually checked and corrected if necessary. The average and maximum errors using HF theory are given in Table I. Here, the same effect as for the trp-cage miniprotein can be observed, that is, hydrogens involved in hydrogen bonds have relative large errors and the results are more or less independent of the size of the fragmentation distance (surroundings d) used. If the hydrogens bound to nitrogens are excluded, again a reasonable agreement with experiment can be seen (Fig. 7). For the ^{13}C , an almost perfect correlation is obtained with a correlation coefficient $R^2 = 0.9958$ and a slope of almost 1.

If the level of theory is switched to DFT, the average and maximum errors listed in Table II are obtained. For hydrogen, very comparable results are obtained when including or excluding correlation effects. In contrast, the results for carbons are not as good with the B3LYP functional compared with HF. This can also be seen in the correlations shown in Figure 8. Even if the correlation coefficient is similar, the slope is significantly larger than 1. This means that for the carbonyl carbons with their down-field shifted chemical shifts (high ppm values) much smaller values are calculated than the experimentally obtained values. If these were excluded, the slope would be much closer to 1 and the results would again be comparable to HF. Therefore, we conclude here that DFT and especially the B3LYP functional leads to systematic errors for carbonyl chemical shifts and probably also other groups with double or triple bonds. This phenomenon was also observed in benchmark calculations on small model systems.⁶³

By the visualization of the protein with the errors mapped on the corresponding atoms (see Fig. 9) as done for the hydrogens in the trp-cage, it is again easily possible to analyze the spatial distribution. In the figure, two large deviations are observed, which are attributed to C_α of Arg16 and Arg21 residues (for the third Arg23 in YuaF no experimental chemical shift value for this atom is available). Thus, C_α seems to be always badly predicted and the problem with multibonded atoms already observed in carbonyl groups appears to be a general limitation of the B3LYP functional. The other errors are located in highly flexible, solvent-exposed regions for which conformational averaging and solvent effects are relevant. In the Supporting Information, correlations for the solvent-exposed and core carbons are given separately (Supporting Information Fig. S2), showing better agreement for the latter. The almost perfect reproduction of the experimental values in the structure-determining core

Figure 6

Changes in the errors in the local region centered on Tyr3 of the trp-cage miniprotein when slightly modifying the position of the aromatic system: (a) original structure from the PDB and (b)–(d) three conformations where the aromatic system is rotated by 5° around different axes. For comparison, the original conformation of the aromatic system is shown in green in all pictures.

is, to our opinion, a very strong support for the published structure.

strates the very good overall agreement of the calculated and experimental chemical shifts.

p63 DNA-binding domain

p63 DNA-binding domain (sequence: 113–345, PDB entry 2RMN) used as second example was solved in the group of Kessler and coworkers (Enthart *et al.*, unpublished data). This human tumor protein acts as a sequence-specific DNA-binding transcriptional activator or repressor. The isoforms contain a varying set of transactivation and autoregulating, transactivation inhibiting domains leading to isoform-specific activation or repression. The chemical shifts were calculated as described above with the B3LYP functional and the mixed basis set and then compared with the experimental values (Kessler, private communication). Only calculations with surroundings of $d = 5 \text{ \AA}$ were performed to limit the computational demand. As in YuaF, all calculated ^{13}C NMR chemical shifts are in good agreement with the experiment except carbonyl carbon shifts that show deviations of about 20 ppm for reasons already discussed above (see Fig. 10). If carbonyl carbons are neglected, the slope of the correlation line is again almost 1, which, combined with the high correlation coefficient of 0.948, demon-

Tfb1/p53 complex

To show that intermolecular interactions are also well described, the Tfb1/p53 complex was taken as the next example. The interaction between the transactivation domain (TAD) of p53 and TFIID is directly correlated with the ability of p53 to activate both transcription initiation and elongation. Di Lello *et al.*⁶⁴ identified a specific region in p53 TAD interacting with the pleckstrin homology domain of the Tfb1 subunit of yeast TFIID and solved the complex structure by NMR spectroscopy (PDB entry 2GS0, BMRB entry 6225).

Table I

Average and Maximum Error (ppm) of the Fragment-Based Calculation Compared to Experiment for YuaF Using Hartree-Fock Theory

$d \text{ (\AA)}$	^1H (all)		^1H (no NH)		^{13}C	
	Aver.	Max.	Aver.	Max.	Aver.	Max.
3.0	0.84	5.04	0.50	2.70	4.00	20.20
4.0	0.85	5.12	0.51	2.69	4.07	20.16
5.0	0.85	5.14	0.53	2.44	4.05	19.11
6.0	0.85	4.98	0.52	2.44	4.04	19.11
7.0	0.83	5.07	0.51	2.45	4.05	19.15

Figure 7

Correlation between calculated (HF, $d = 7 \text{ \AA}$) and experimental ^1H (left) and ^{13}C (right) chemical shifts of YuaF. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The results of our calculations in comparison to experimental ^1H , ^{13}C , and additionally ^{15}N chemical shifts are shown in Figure 11. If the hydrogens bound to nitrogens are excluded, the same reasonable and good correlations as for the examples described above are seen for ^1H and ^{13}C . In the figure, the results of two empirical approaches are also shown. Both, SPARTA+¹⁶ and SHIFTX2,¹³ show excellent agreement for backbone atoms and are at the moment still the methods of choice for these atoms. But for side chain atoms, this picture changes in favor of the quantum mechanical method. SPARTA+ only gives results up to the C_β atoms and the maximum error of SHIFTX2 for ^1H chemical shifts (2.05 ppm) is comparable with the one of the ADMA method (if hydrogens bound to nitrogen are neglected). The average and maximum errors are given in the Supporting Information. But for ^{15}N , a bad correlation with a very flat slope much different from 1 is obtained. It is very interesting that the range of the calculated values is much too broad (87–153 ppm) than the experimental ones (105–130 ppm). When looking at the spatial distribution, it becomes evident that the chemical shifts of the solvent-exposed nitrogens are underestimated, whereas the ones of the core nitrogens are overestimated (see Fig. 12). The use of implicit solvent results in a slight improvement for the solvent-exposed nitrogens but the values for the core nitrogens do not change significantly (Figs. 11 and 12). Thus, the underestimation can be explained by missing or insufficiently represented solvent effects. But reasons for the overestimation of the core chemical shifts have still to be found.

The second thing we tested on the Tfb1/p53 complex was if some specific structures of the NMR ensemble in

the file from the Protein Data Bank agree better with the experiment than others. The correlations of the ^{13}C chemical shifts for 11 of these structures are shown in the Supporting Information. No significant differences in the correlation coefficients are found. Thus, all structures have to be considered as equally good with respect to their agreement with the NOE distance constraints and our chemical shift calculations. Nevertheless, the largest deviations are seen for different chemical shifts in each model so that local regions are better described in some structures than in others. Combining these better fitting regions could be used to generate a consensus model based on the NMR chemical shifts.

The next step will be to calculate the chemical shift perturbations at amino acids in the binding site of Tfb1 occurring on the binding of the activation domain of p53. If the experimental changes can be reproduced by the difference between the chemical shifts calculated for the free Tfb1 and the complex, our method could probably also be used for the validation of complex structures. Corresponding studies are ongoing.

Table II

Average and Maximum Error (ppm) of the Fragment-Based Calculation Compared to Experiment for YuaF Using Density Functional Theory (B3LYP)

$d \text{ (\AA)}$	^1H (all)		^1H (no NH)		^{13}C	
	Aver.	Max.	Aver.	Max.	Aver.	Max.
3.0	0.86	5.13	0.48	2.76	6.63	27.20
4.0	0.89	6.54	0.52	6.54	6.77	27.17
5.0	0.87	5.23	0.50	2.82	6.76	34.79
6.0	0.88	5.07	0.51	4.98	6.73	34.67
7.0	0.83	5.15	0.48	2.68	6.67	30.22

Figure 8

Correlation between calculated (B3LYP, $d = 7 \text{ \AA}$) and experimental ^1H (left) and ^{13}C (right) chemical shifts of YuaF. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Dynein light chain 2A

As described in the "Introduction" section, there are some known cases where wrong folds have been proposed by NMR spectroscopy, deposited in the PDB, and that were replaced by corrected versions only after some years and after new experimental information became available. One extreme example is the light chain 2A of the protein dynein, which was misinterpreted to be a monomeric protein domain (obsolete PDB entry 1TGQ)

during the first structure determination. This first structure was replaced by the structure of the dimer 17 months later (PDB entry 2B95; Liu *et al.*, unpublished data). Nabuurs *et al.*⁷ demonstrated that the publication of the wrong structure could have been avoided if a combination of validation criteria had been applied even if no single criterion was able to detect the error on its own. Therefore, we will try here as the last example if

Figure 9

Spatial distribution of the errors in the ^{13}C chemical shifts of YuaF calculated using DFT (B3LYP): the values are coded by the size of the spheres representing the atoms (larger spheres correspond to larger errors). Red and blue represent overestimations and underestimations of the calculated shifts relative to the experimental ones, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 10

Correlation between calculated (B3LYP, $d = 5 \text{ \AA}$) and experimental ^{13}C chemical shifts of p63. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 11

Correlation between calculated (B3LYP, $d = 5 \text{ \AA}$) and experimental ^1H (upper), ^{13}C (middle), and ^{15}N (lower) chemical shifts of the Tf1/p53 complex calculated with the ADMA approach without (left) and with (right) an implicit solvent model is shown in blue. Additionally, the results of two empirical programs, SPARTA+ (red) and SHIFTX2 (yellow), are given.

Figure 12

Spatial distribution of the errors in the ^{15}N chemical shifts of the Tfb1/p53 complex calculated using DFT (B3LYP) without (left) and with (right) an implicit solvent model: the values are coded by the size of the spheres representing the atoms (larger spheres correspond to larger errors). Red and blue represent overestimations and underestimations of the calculated shifts relative to the experimental ones, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the comparison of experimental and calculated chemical shifts would be able to identify the correct structure. For PDB entry 2B95, only backbone assignment is available. But for an independently solved dimeric structure of the same protein (PDB entry 1Z09,⁶⁵ BMRB entry 6527), the full set of chemical shifts can be downloaded, and we use these shifts for comparison with the calculations of the monomeric and dimeric structures. For the carbonyl carbons, the same problems described above are observed. Additionally, the C-terminal residue shows extreme deviations from the experiment if the 1TGO structure is used to calculate the chemical shifts. Because these deviations cannot be explained by the misinterpretation as a monomeric protein (the C-terminus is relative flexible and not involved in the dimer formation) and are probably caused by convergence problems of the quantum chemical calculations due to local distortion in the structure, we removed them from the analysis.

The correlation between the calculated and experimental chemical shifts in the range between 0 and 100 ppm is shown in Figure 13. Even if both models (monomeric and dimeric) show relatively large errors, the chemical shifts of the monomer scatter more than the ones of the dimer resulting in a lower correlation coefficient. Looking at the spatial distribution of the errors (see Fig. 14), the major deviations in the dimeric structure are again in solvent-exposed, flexible regions. In contrast, in the monomeric structure, they are located in the core and especially in the regions of the alpha helix strongly distorted to fulfill the NOE constraints intramolecularly, which connect the two monomers in the correct structure. Thus, by inspecting the deviations between calculated and measured shifts, the wrong structure would have been identified on the basis of localized, very strong deviations in regions of a structure that is normally predicted with relatively high accuracy.

As described above, the PDB entry for the monomeric structure was replaced by a dimeric one in PDB entry 2B95. To see if this structure also agrees to the experimental data, the chemical shifts for this structure were also calculated. An overlay of the two structures as well as the correlation with experiment for 2B95 is given in the Supporting Information. Both dimeric structures show the same secondary structure elements but with slight differences in their relative orientation resulting in a C_α root mean square deviation of 2.7 Å. The correlation coefficients for the carbon atoms between 0 and 100

Figure 13

Correlation between calculated (B3LYP, $d = 5$ Å) and experimental ^{13}C chemical shifts of the dynein light chain 2A based on the monomeric (PDB entry 1TGO, red) and dimeric (PDB entry 1Z09, blue) structures.

Figure 14

Spatial distribution of the errors in the ^{13}C chemical shifts of light chain 2A in the monomeric (left) and dimeric (right) structures calculated using DFT (B3LYP): the values are coded by the size of the spheres representing the atoms (larger spheres correspond to larger errors). In the monomeric structure, clusters of large errors can be seen in the core region indicating problems in the structure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

ppm are very comparable for both dimeric structures, so that they, with respect to chemical shifts, conform to the experimental data to the same degree, which is a little bit unexpected regarding the large C_α RMSD. Thus, the calculations prefer the correct dimeric structure over the wrong monomeric one. But perhaps there are still some minor problems in both dimeric structures leading to the relatively weak correlation in comparison with the other test examples of this publication.

CONCLUSIONS

In this study, we performed quantum chemical calculations of NMR chemical shifts for many proteins. These show that our fragment-based ADMA approach is well suited to reproduce experimental values even for large proteins and protein–protein complexes at least for ^{13}C and can be used to evaluate protein structures determined by NMR spectroscopy. However, empirical methods for the prediction of chemical shifts still give better results especially for backbone atoms, and in a more time-efficient way, so that they are still the methods of choice for application where large data sets are available for parameterization. In contrast, quantum chemical cal-

culations will be more important in cases in which an empirical parameterization provides ambiguous results or where it is not possible at all due to the lack of experimental data. For example, due to the multitude of different kinds of interactions and chemical groups involved, the prediction of chemical shift perturbations resulting from the formation of protein–protein or protein–ligand complexes is very challenging. Additionally, the assignment of chemical shifts, if a crystal or a homology structure of the protein is available, could be guided by quantum chemical calculations. The latter is important when additional studies on the dynamics of a protein are to be performed using NMR methods.

Obviously, there is still place for improvement. The B3LYP functional seems to have problems to describe atoms with double and triple bonds such as carbonyl carbons or the C_ϵ of Arg. Thus, different levels of theory (e.g., post-HF methods) or different density functionals in combination with larger basis sets should be tested to see their influence on the calculated values. Additionally, the introduction of partial charges to describe additional parts of the macromolecule in the parent molecule calculations could be beneficial. For ^1H and ^{15}N , the results are not satisfactory at the moment. For this and also for the remaining errors of ^{13}C ,

the missing or insufficient treatment of solvent effects and conformational averaging was identified as possible reasons. An implicit solvent model was able to partly account for the first but some effects, such as the hydrogen-bonding network with water molecules, can only be treated with explicit solvents. And last but not least, a reason for the overestimation of the ^{15}N chemical shifts in the core region of the protein has still to be found. Research in these directions is also on its way.

ACKNOWLEDGMENTS

The authors thank the Common Ulm Stuttgart Server (CUSS) and the Baden-Württemberg grid (bwGRiD), which is part of the D-Grid system, for providing the computer resources making the computations possible.

REFERENCES

- Gronwald W, Kalbitzer HR. Automated structure determination of proteins by NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc* 2004;44:33–96.
- Güntert P. Structure calculation of biological macromolecules from NMR data. *Q Rev Biophys* 1998;31:145–237.
- Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* 1997;273:283–298.
- 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.
- Brünger. ATX-PLOR Version 3.1. New Haven/London: Yale University Press; 1992.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
- Nabuurs SB, Spronk CAEM, Vuister GW, Vriend G. Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. *PLoS Comp Biol* 2006;2:71–79.
- Saccanti E, Rosato A. The war of tools: how can NMR spectroscopists detect errors in their structures? *J Biomol NMR* 2008;40:251–261.
- Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 1999;13:289–302.
- 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.
- Berjanskii MV, Neal S, Wishart DS. PREDITOR: a web server for predicting protein torsion angle restraints. *Nucleic Acids Res* 2006;35:W63–W69.
- Neal S, Nip AM, Zhang H, Wishart DS. Rapid and accurate calculation of protein ^1H , ^{13}C and ^{15}N chemical shifts. *J Biomol NMR* 2003;26:215–240.
- Han B, Liu Y, Ginzinger SW, Wishart DS. SHIFTX2: significantly improved protein chemical shift prediction. <http://www.shiftx2.ca/> 2010.
- Ösapay K, Case DA. A new analysis of proton chemical shifts in proteins. *J Am Chem Soc* 1991;113:9436–9444.
- Xu XP, Case DA. Automated prediction of ^{15}N , $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$ and $^{13}\text{C}'$ chemical shifts in proteins using a density functional database. *J Biomol NMR* 2001;21:321–333.
- Shen Y, Bax A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* 2010;48:13–22.
- Laws DD, de Dios AC, Oldfield E. NMR chemical shifts and structure refinement in proteins. *J Biomol NMR* 1993;3:607–612.
- Le H-B, Pearson JG, de Dios AC, Oldfield E. Protein structure refinement and prediction via NMR chemical shifts and quantum chemistry. *J Am Chem Soc* 1995;117:3800–3807.
- Oldfield E. Chemical shifts in amino acids, peptides, and proteins: from quantum chemistry to drug design. *Annu Rev Phys Chem* 2002;53:349–378.
- Vila JA, Villegas ME, Baldoni HA, Scheraga HA. Predicting $^{13}\text{C}_\alpha$ chemical shifts for validation of protein structures. *J Biomol NMR* 2007;38:221–235.
- Vila JA, Ripoll DR, Scheraga HA. Use of $^{13}\text{C}_\alpha$ chemical shifts in protein structure determination. *J Phys Chem B* 2007;111:6577–6585.
- Vila JA, Arnautova YA, Scheraga HA. Use of $^{13}\text{C}_\alpha$ chemical shifts for accurate determination of β -sheet structures in solution. *Proc Natl Acad Sci USA* 2008;105:1891–1896.
- Vila JA, Scheraga HA. Factors affecting the use of $^{13}\text{C}_\alpha$ chemical shifts to determine, refine, and validate protein structures. *Proteins* 2008;71:641–654.
- Vila JA, Aramini JM, Rossi P, Kuzin A, Su M, Seetharaman J, Xiao R, Tong L, Montelino GT, Scheraga HA. Quantum chemical $^{13}\text{C}_\alpha$ chemical shift calculations for protein NMR structure determination, refinement, and validation. *Proc Natl Acad Sci USA* 2008;105:14389–14394.
- Jacob CR, Visscher L. Calculation of nuclear magnetic resonance shieldings using frozen-density embedding. *J Chem Phys* 2006;125:194104.
- Wesolowski TA, Warshel A. Frozen density functional approach for ab initio calculations of solvated molecules. *J Phys Chem* 1993;97:8050–8053.
- Lee AM, Bettens RPA. First principles NMR calculations by fragmentation. *J Phys Chem A* 2007;111:5111–5115.
- Johnson ER, DiLabio GA. Convergence of calculated nuclear magnetic resonance chemical shifts in a protein with respect to quantum mechanical model size. *J Mol Struct Theochem* 2009;898:56–61.
- He X, Wang B, Merz KM, Jr. Protein NMR chemical shift calculations based on the automated fragmentation QM/MM approach. *J Phys Chem B* 2009;113:10380–10388.
- Mezey PG. Functional groups in quantum chemistry. *Adv Quantum Chem* 1996;27:163–222.
- Mezey PG. Macromolecular density matrices and electron densities with adjustable nuclear geometries. *J Math Chem* 1995;18:141–168.
- Mezey PG. Quantum chemistry of macromolecular shape. *Int Rev Phys Chem* 1997;16:361–388.
- Exner TE, Mezey PG. Ab initio quality electrostatic potentials for proteins: an application of the ADMA approach. *J Phys Chem A* 2002;106:11791–11800.
- Exner TE, Mezey PG. Ab initio quality properties for macromolecules using the ADMA approach. *J Comput Chem* 2003;24:1980–1986.
- Exner TE, Mezey PG. The field-adapted ADMA approach: introducing point charges. *J Phys Chem A* 2004;108:4301–4309.
- Exner TE, Mezey PG. Evaluation of the field-adapted ADMA approach: absolute and relative energies of crambin. *Phys Chem Chem Phys* 2005;7:4061–4069.
- Facelli JC. Calculations of chemical shieldings: theory and applications. *Concepts Magn Reson* 2004;20A:42–69.
- London F. The quantum theory of inter-atomic currents in aromatic combinations. *J Phys Radium* 1937;8:397–409.
- McWeeny R. Perturbation theory for Fock–Dirac density matrix. *Phys Rev* 1962;126:1028–1034.
- Ditchfield R. Self-consistent perturbation theory of diamagnetism. I. Gauge-invariant LCAO method for NMR chemical shifts. *Mol Phys* 1974;27:789–807.

41. Wolinski K, Hilton JF, Pulay P. Efficient implementation of the gauge-independent atomic orbital method for NMR chemical shift calculations. *J Am Chem Soc* 1990;112:8251–8260.
42. Cheeseman JR, Trucks GW, Keith TA, Frisch MJ. A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J Chem Phys* 1996;104:5497–5509.
43. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 03 (Revision C.2). Wallingford, CT: Gaussian, Inc.; 2004.
44. Becke AD. Density-functional thermochemistry. III. The role of exact exchange. *J Chem Phys* 1993;98:5648–5652.
45. Hariharan PC, Pople JA. Accuracy of AHn equilibrium geometries by single determinant molecular orbital theory. *Mol Phys* 1974;27:209–214.
46. Hariharan PC, Pople JA. Influence of polarization functions on MO hydrogenation energies. *Theor Chim Acta* 1973;28: 213–222.
47. Frandl MM, Pietro WJ, Hehre WJ, Binkley JS, Gordon MS, DeFrees DJ, Pople JA. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J Chem Phys* 1982;77:3654–3665.
48. Blauddau JP, McGrath MP, Curtiss LA, Radom L. Extension of Gaussian-2 (G2) theory to molecules containing third-row atoms K and Ca. *J Chem Phys* 1997;107:5016–5021.
49. Binning RC, Jr, Curtiss LA. Compact contracted basis sets for third-row atoms: gallium-krypton. *J Comput Chem* 1990;11: 1206–1216.
50. Rassolov VA, Ratner MA, Pople JA, Redfern PC, Curtiss LA. 6–31G* basis set for third-row atoms. *J Comput Chem* 2001;22:976–984.
51. Rassolov VA, Pople JA, Ratner MA, Windus TL. 6–31G* basis set for atoms K through Zn. *J Chem Phys* 1998;109:1223–1229.
52. Gordon MS. The isomers of silacyclopentane. *Chem Phys Lett* 1980;76:163–168.
53. Purcell WP, Singer JA. A brief review and table of semiempirical parameters used in the Hückel molecular orbital method. *J Chem Eng Data* 1967;12:235–246.
54. Marsili M, Gasteiger J. π Charge distribution from molecular topology and π orbital electronegativity. *Croat Chem Acta* 1981;53:601–614.
55. Gasteiger J, Marsili M. Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron* 1980;36:3219–3222.
56. Gasteiger J, Marsili M. Prediction of proton magnetic resonance shifts: the dependence on hydrogen charges obtained by iterative partial equalization of orbital electronegativity. *Org Magn Reson* 1981;15:353–360.
57. Cancès E, Mennucci B, Tomasi J. A new integral equation formalism for the polarizable continuum model: theoretical background and applications to isotropic and anisotropic dielectrics. *J Chem Phys* 1997;107:3032–3041.
58. Mennucci B, Tomasi J. Continuum solvation models: a new approach to the problem of solute's charge distribution and cavity boundaries. *J Chem Phys* 1997;106:5151–5158.
59. Cossi M, Barone V, Mennucci B, Tomasi J. Ab initio study of ionic solutions by a polarizable continuum dielectric model. *Chem Phys Lett* 1998;286:253–260.
60. Barua B, Lin JC, Williams VD, Kummeler P, Neidigh JW, Andersen NH. The Trp-cage: optimizing the stability of a globular miniprotein. *Protein Eng Des Sel* 2008;21:171–185.
61. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CR, Tolmie DE, Wenger RK, Yao H, Markley JL. BioMagResBank. *Nucleic Acids Res* 2007;36:D402–D408.
62. Walker CA, Hinderhofer M, Witte DJ, Boos W, Möller HM. Solution structure of the soluble domain of the NfeD protein YuaF from *Bacillus subtilis*. *J Biomol NMR* 2008;42:69–76.
63. Auer A, Gauss J, Stanton JF. Quantitative prediction of gas-phase ^{13}C nuclear magnetic shielding constants. *J Chem Phys* 2003; 118:10407.
64. Di Lello P, Jenkins LM, Jones TN, Nguyen BD, Hara T, Yamaguchi H, Dikeakos JD, Appella E, Lagault P, Omichinski JG. Structure of the Tfb1/p53 complex: insights into the interaction between the p62/Tfb1 subunit of TFIIF and the activation domain of p53. *Mol Cell* 2006;22:731–740.
65. Ilangoan U, Ding W, Zhong Y, Wilson CL, Groppe JC, Trbovich JT, Zuniga J, Demeler B, Tang Q, Gao G, Mulder KM, Hinck AP. Structure and dynamics of the homodimeric dynein light chain km23. *J Mol Biol* 2005;352:338–354.
66. Helgaker T, Jaszuski M, Ruud K. Ab initio methods for the calculation of NMR shielding and indirect spin–spin coupling constants. *Chem Rev* 1999;99:293–352.