Chapter 5
Quantum Calculation of Protein NMR Chemical Shifts Based on the Automated Fragmentation Method

Tong Zhu, John Z.H. Zhang and Xiao He

Abstract The performance of quantum mechanical methods on the calculation of protein NMR chemical shifts is reviewed based on the recently developed automatic fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) approach. By using the Poisson-Boltzmann (PB) model and first solvation water molecules, the influence of solvent effect is also discussed. Benefiting from the fragmentation algorithm, the AF-QM/MM approach is computationally efficient, linear-scaling with a low pre-factor, and thus can be applied to routinely calculate the *ab initio* NMR chemical shifts for proteins of any size. The results calculated using Density Functional Theory (DFT) show that when the solvent effect is included, this method can accurately reproduce the experimental ^{1}H NMR chemical shifts, while the ^{13}C NMR chemical shifts are less affected by the solvent. However, although the inclusion of solvent effect shows significant improvement for ^{15}N chemical shifts, the calculated values still have large deviations from the experimental observations. Our study further demonstrates that AF-QM/MM calculated results accurately reflect the dependence of $^{13}C_{\alpha}$ NMR chemical shifts on the secondary structure of proteins, and the calculated ^{1}H chemical shift can be utilized to discriminate the native structure of proteins from decoys.

Keywords Automated fragmentation QM/MM · Linear scaling · Protein NMR chemical shift · Ab initio · Solvent effect

T. Zhu · J.Z.H. Zhang · X. He (✉)
State Key Laboratory of Precision Spectroscopy and Department of Physics,
Institute of Theoretical and Computational Science, East China Normal University,
Shanghai, China
e-mail: xiaohe@phy.ecnu.edu.cn

J.Z.H. Zhang
Department of Chemistry, New York University, New York, NY 10003, USA

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in
Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_5

5.1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is an invaluable and widely used technique in areas of chemistry, biology and medicine [1, 2]. In proteins, the chemical shift tensors are key parameters in the NMR experiment, enabling signals from different nuclei of any given type in a molecule to be distinguished. During the past decades, there has been significant progress in using chemical shift information for characterization of protein structure and dynamics [3–11].

Although the chemical shifts are probably the most precise parameters that can be obtained for biomolecules, the inherently complex dependency on geometric, dynamic and electronic properties has made accurate calculation of chemical shifts of protein a significant challenge [12–14]. There are mainly two widely used methods to calculate protein NMR chemical shifts: the empirical approach based on the experimental database and the *ab initio* approach based on quantum mechanical (QM) calculations. Empirical methods rely on statistical data derived from a limited set of high-quality 3D structures and make use of empirical or semi-empirical equations to account for the non-sequential environment [15–19]. These methods are usually quite successful in predicting backbone chemical shifts, which are primarily determined by the local secondary structure, but they are not so well suited to handle proteins with nonstandard residues, metal cofactors, or protein-ligand complexes.

Over the past decade, QM methods have become increasingly useful for NMR chemical shift studies. Following the pioneering work of de Dios et al. [20–22], a number of quantum calculations have been carried out for chemical shifts in proteins and peptides [23–33]. However, due to the poor scaling of *ab initio* and DFT methods, it has not been practical to apply standard all-electron quantum chemistry methods to realistic macromolecules. In fact, full quantum mechanical computations on structures with 1,000 atoms or more are currently not routinely feasible. Fortunately, many previous studies have proven that there is no need to include all atoms in the QM NMR calculation because the nuclear shielding is fundamentally a local physical property. Cui and Karplus proposed a method for calculating chemical shifts in the QM/MM framework, and concluded that the QM/MM method can provide good descriptions of the environmental effect on chemical shifts [34]. Frank et al. calculated the chemical shifts using the fragment based adjustable density matrix assembler (ADMA) method [35–37]. Gao et al. also reported a fragment molecular orbital (FMO) method for NMR chemical shift calculations at the Hartree-Fock level [38, 39]. In our previous studies [40, 41, 49], a more efficient automated fragmentation quantum mechanics/molecular mechanics approach (AF-QM/MM) was shown to be applicable to routine *ab initio* NMR chemical shift calculation for proteins of any size. In this approach, the entire protein is divided into individual fragments, and residues within a certain buffer region surrounding each fragment are included in the QM calculation to preserve the chemical environment of the divided fragment. The remainder of the system outside the buffer regions is described by the MM method. The AF-QM/MM calculated NMR chemical shifts of several proteins are in good agreement with the experimental measurement [40, 41].

Since most NMR measurements are performed on liquid samples, the NMR parameters (in particular NMR chemical shifts) are highly sensitive to the molecular environment, and especially the solvent effect. The effect of solvent on nuclear magnetic shielding parameters derived from NMR spectroscopy has been of great interest for a long time [42–47]. Several empirical approaches have been formulated to evaluate the solvent effects on nuclear shieldings, however, the development of *ab initio* calculation of NMR properties of proteins in solution has only recently received attention, and most of the studies were focused on small molecular structures or model peptides. In this review, we mainly discuss the influence of solvent effects on the QM calculation of protein NMR chemical shifts, by including both the implicit and explicit solvent model based on our previous works [40, 41, 48, 49].

5.2 Automated Fragmentation QM/MM Method

5.2.1 Fragmentation Criteria

The basic fragmentation scheme in the AF-QM/MM approach is shown in Fig. 5.1. In this approach, the entire protein system is divided into non-overlapping fragments termed core regions. The residues within a certain range from the core region are assigned as the buffer region. Both the core region and its buffer region are treated by QM, whereas the rest of the system is described by an empirical point-charge model. The purpose of the buffer area is to include the local QM

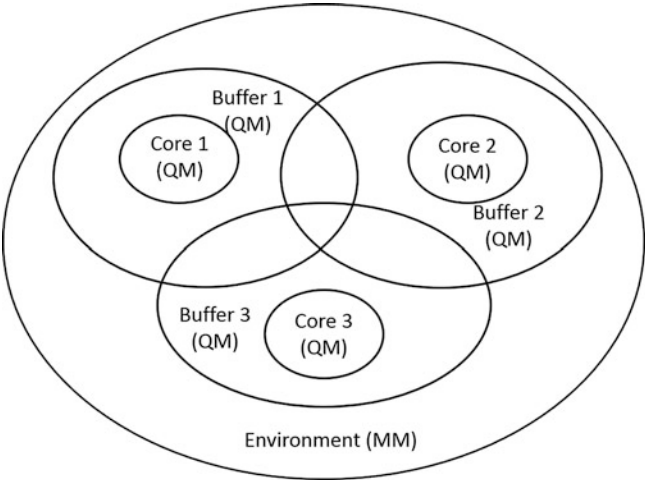

Fig. 5.1 Subsetting scheme for the AF-QM/MM approach

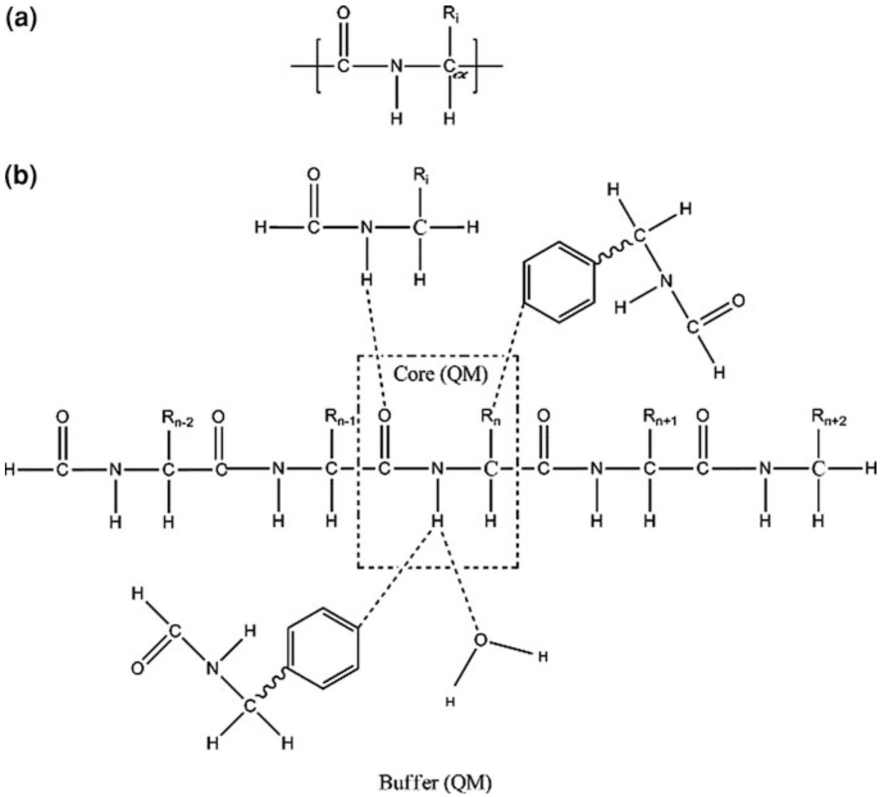

Fig. 5.2 **a** Definition of the residue unit used in this work. **b** *N*th amino acid is the core region. Sequentially connected ($n - 2$)th, ($n - 1$)th, ($n + 1$)th and ($n + 2$)th residues are included in the buffer region. In addition, the residues in spatial contact with the *n*th residue are also assigned to the buffer region (see text for further details)

effects on the chemical shifts. Each fragment-centric QM/MM calculation is carried out separately. Only the shielding constants of the atoms in the core region are extracted from the individual QM/MM calculations. A more detailed illustration of the automated fragmentation scheme is presented in Fig. 5.2.

For proteins discussed in this work, each residue is taken as the core region. A different definition of the residue that consists of the –CO–NH–CHR– is adopted to preserve the electron delocalization across the peptide bond (Fig. 5.2a). A generalized molecular cap was also introduced to take into account the QM polarization effect and charge transfer within the first shell from the residue of interest, as shown in Fig. 5.2b. In this and all our previous studies, we adopt the following distance-dependent criteria to include residues within the buffer region of each core residue: (1) if one atom of the residue outside the core region is less than 4 Å away from any atom in the core region and at least one of the two atoms is a non-hydrogen atom; (2) if the distance between one hydrogen atom in the core region and the other hydrogen atom outside the core region is less than 3 Å, or (3) if a heavy atom on an aromatic ring is within 5 Å from any atom in the core region. Of course, other distance-dependent criteria could be used to further optimize the choice of the buffer region. The non-neighboring residues in the buffer region are simply capped by hydrogen atoms to construct the closed-shell fragment.

The remaining atoms beyond the buffer region are treated by MM method. A point-charge model is employed to account for the empirical electrostatic field outside the QM region. We use the full point charges for those junction atoms that are replaced by hydrogen atoms. Because a buffer region is added to smoothly link the core region and MM environment, atoms on the boundary between the QM and MM regions are relatively far from the core region and their influence is attenuated. By using a general criterion to assign a buffer zone to each residue, we can reduce the size of each fragment in order to make the QM calculation as small as possible until we strike a compromise between the desired accuracy and the computational cost. Although the total number of residue pairs is proportional to the square of the number of residues, the size of each fragment is independent of the overall protein size because each residue can have only a limited number of residues in its vicinity. Hence, the largest fragment normally contains less than 250 atoms consisting of C, H, O, N, and S, which is an affordable calculation at the HF and DFT levels. In this work, all the QM calculation were performed using Gaussian09 program [50].

5.2.2 Solvent Effects

The main obstacle of including solvent effects in QM/MM NMR calculation is the determination of solvent positions around the biomolecules. It is known that the interaction of the biomolecule with solvent is not well represented by the coordinates present in the experimental structures. There are no water molecules in the protein structure obtained by NMR experiment, and even some "crystallographic" waters are present in the X-ray structure, they represent only a fraction of the waters surrounding the biomolecule. In addition, the static positions of water molecules are probably not representative of the environment seen by the atoms of the solvated biomolecule. Therefore, in most of the calculations, the implicit continuum solvation model was used.

In continuum solvation model, the solute (protein) is represented by a charge distribution ρ(r) embedded in a cavity surrounded by a polarizable medium with dielectric constant ε. The solute charge distribution polarizes the dielectric medium and creates a reaction field which acts back to polarize the solute until equilibrium is reached. The reaction field acting on the solute can be effectively represented by that of induced charges on the cavity surface according to the classical electrostatic theory. In the current approach, we use the DivCon [51] program which combines the linear-scaling divide-and-conquer semi-empirical algorithm with the Poisson-Boltzmann (PB) equation to perform the self-consistent reaction field (SCRF)

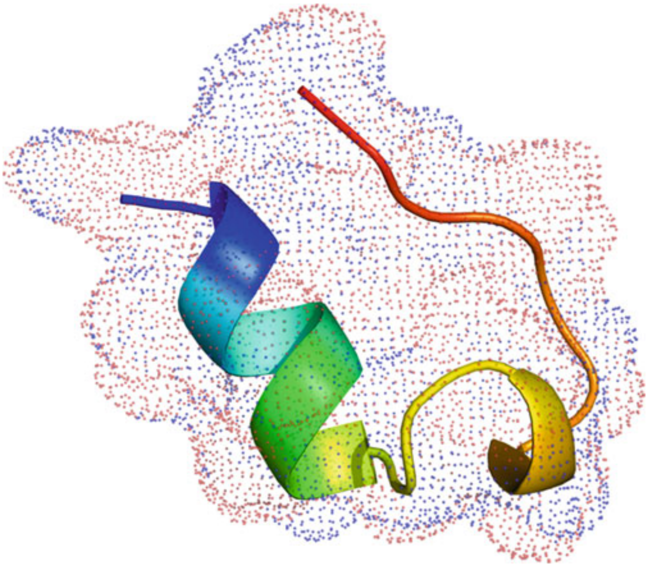

Fig. 5.3 NMR structure of Trp cage (PDB entry: 1L2Y) together with the surface charges calculated by DivCon (*red* and *blue dots* represent the positive and negative charges, respectively)

calculation. The CM2 charges for the atoms of proteins in conjunction with the PM3 methods were derived since the PM3/CM2 is one of the best polarizable charge models for NMR chemical shift calculations, as observed previously for HF/6-31G** and B3LYP/6-31G** calculations [41]. Then the set of point charges of the MM environment and on the molecular surface which represents the reaction field is used as the background charges in the QM calculation. The effective surface charges representing the solvent effects are shown in Fig. 5.3.

5.3 Applications

5.3.1 Comparison with the Full System Quantum Chemistry Calculations

Firstly, the AF-QM/MM method with the solvation model was used to compute the ^{1}H, ^{13}C and ^{15}N absolute chemical shielding tensors of a small protein Trp-cage (20 residues, PDB entry: 1L2Y). And the results are compared with the conventional full system calculations as shown in Fig. 5.4.

In the full system calculation, the protein is computed as an intact molecule with the presence of the same set of surface charges. As one can see from Fig. 5.4, the root mean square errors (RMSEs) for the ^{1}H, ^{13}C and ^{15}N are only 0.06, 0.22 and 0.55 ppm, respectively. All these errors are very small as all of them are less

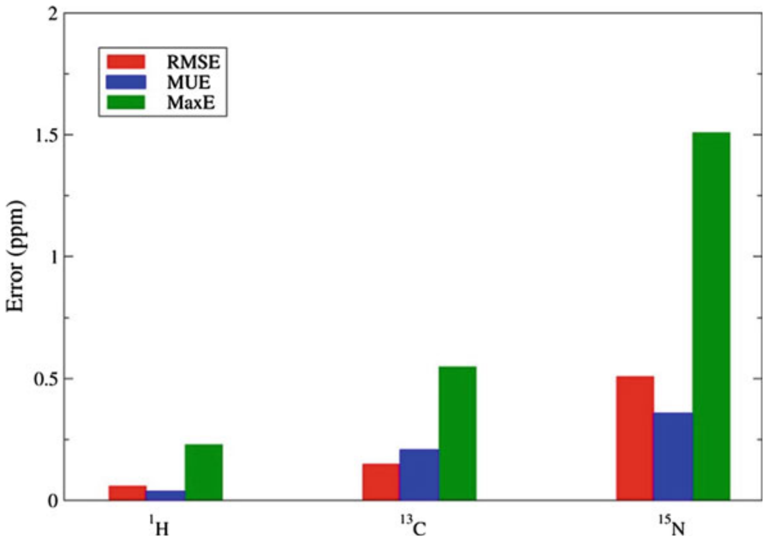

Fig. 5.4 Root mean square error (*RMSE*), mean unsigned error (*MUE*) and maximum error (*MaxE*) of AF-QM/MM with respect to the full system calculated ^{1}H, ^{13}C and ^{15}N chemical shifts in Trp-cage at the B3LYP/6-31G** level

than 1 % of the absolute chemical shielding tensor. The result clearly demonstrates that, as expected, the AF-QM/MM calculated chemical shifts can well reproduce the full quantum mechanical calculations for proteins.

5.3.2 ^{1}H Chemical Shifts in Proteins

Proton chemical shifts are the most important and most studied output of NMR experiments. In proteins, the proton atoms can be divided into two categories. One is called non-polar ^{1}H atoms which usually form covalent bonds with C atoms; the other group is called polar ^{1}H atoms which usually from covalent bonds with polar N, S or P atoms, and most of them are involved in hydrogen bonding interactions. The measured chemical shifts of ^{1}H atoms for these two groups are quite different. In this section, we first compare the non-polar ^{1}H chemical shift of Trp-cage calculated by the AF-QM/MM method with the experimental values. For the hydrogen atoms, calculations in both gas phase and solution phase give excellent agreement with the experimental value as shown in Fig. 5.5a.

The RMSE, MUE, correlation coefficient and the fitted function are given in Table 5.1. Although the calculated results for trp-cage in the gas phase are pretty well, the inclusion of the solvent effects still improves the correlation between the theoretical and experimental values from 0.977 to 0.986. The RMSE also decreased from 0.39 to 0.29 ppm, and the slop of the correlation function is closer to 1.

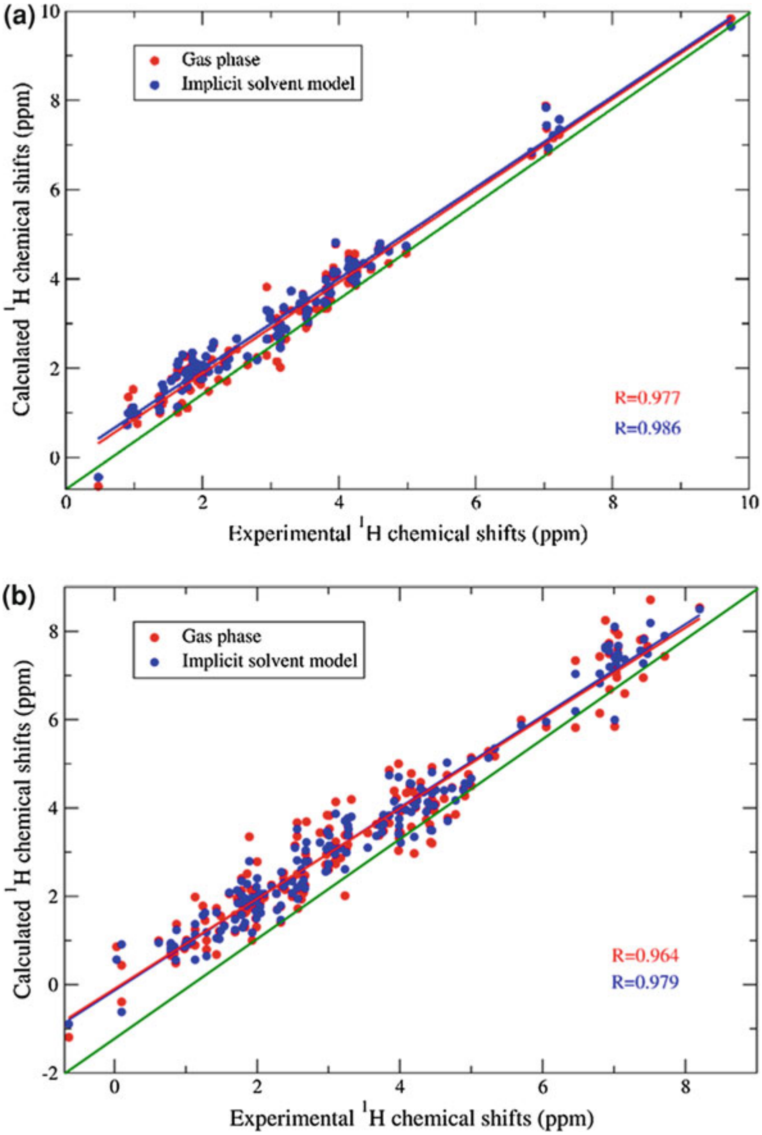

Fig. 5.5 Correlation between experimental and calculated ^{1}H NMR chemical shifts. **a** Trp-cage, **b** Pin1 WW domain. The exchangeable protons were excluded

The results here show that the solvent effects are important and calculated NMR chemical shifts with the solvation model for ^{1}H atom clearly improve the agreement between theory and experiment.

We also calculated the non-polar ^{1}H chemical shifts of Pin1 WW domain (PDB entry: 1PIN) which mainly consists of β-sheets. The comparison of our calculated

Table 5.1 Comparison of AF-QM/MM and experimental chemical shifts for the ^{1}H atoms in Trp-cage and Pin1 WW domain

		RMSE	MUE	R	Correlation function
Trp-Cage	G.	0.39	0.30	0.977	1.024 x −0.17
	S.	0.29	0.23	0.986	1.018 x −0.06
Pin1 WW domain	G.	0.57	0.44	0.964	1.036 x −0.13
	S.	0.43	0.33	0.979	1.023 x +0.09
GB3	G.	0.86	0.39	0.925	0.976 x −0.02
	S.	0.53	0.29	0.983	0.991 x −0.02

G. gas phase; *S.* in solution. The exchangeable protons were excluded

chemical shifts with the experimental values is given in Fig. 5.5b and Table 5.1. Here, the similar result as that of trp-cage is observed. The theoretical ^{1}H chemical shift in solvation is better correlated with the experimental values than results from gas-phase calculations. The RMSE using the solvent model is 0.42 ppm, which is smaller than the gas phase result of 0.57 ppm and the correlation between theoretical and experimental values also improved from 0.964 to 0.979. Thus, the inclusion of the solvent effects clearly improves the theoretical result. In our previous study [40], we also performed calculations on a large protein with mixed α-helical and β-sheet secondary structures, GB3 (PDB entry: 1IGD, 61 residues). The comparison between theoretical and experimental result is also shown in Table 5.1. Again, better correlation with experiment is seen for non-polar ^{1}H chemical shift with the inclusion of solvation.

Furthermore, we check the performance of the AF-QM/MM method on the polar hydrogen especially for protein amide H atoms. The ^{1}H$_N$ chemical shift is one of the most precise NMR parameters that can be measured, which plays key roles in peak assignments. Thus, a QM model that can accurately predict their chemical shift is in demand. Previous studies have found that the main reason for the inaccuracy in computed amide H chemical shifts arises from the improper treatment of the solvation effect, especially the specific solvent-solute hydrogen bond effect. To include these effects in the calculation, explicit inclusion of solvent molecules is required. In our previous study [49], we used a 3D reference interaction site model (3D-RISM) to correct the distribution of explicit solvent molecules. The algorithm of 3D-RISM method is based on statistical mechanics and has been shown to accurately reproduce water distributions at a reduced computational cost. The PLACEVENT [52, 53] program developed by Hirata and co-workers was utilized to translate the continuous distributions to explicit water molecules. In the calculation, only the water molecules in the first and second solvation shell (within 6.0 Å from any atom in the protein) are regarded as part of the entire system. While the implicit solvent model was used to represent the bulk solvent effect beyond the second solvent shell as shown in Fig. 5.6 [49].

The protein GB3 is taken as the initial geometry. Besides the crystallographic water, 678 more water molecules were added by the PLACEVENT program to

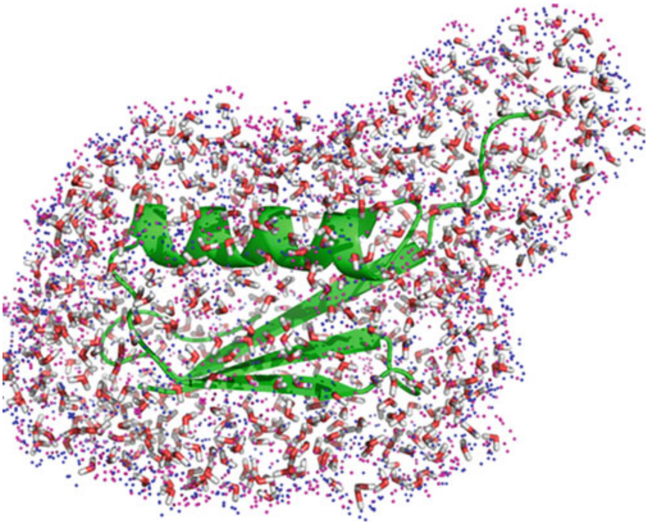

Fig. 5.6 Graphical representation of GB3 (PDB entry: 2IGD) together with the first, second solvation shells and surface charges calculated by DivCon program [51]. (*Red* and *blue dots* represent the positive and negative surface charges, respectively)

mimic the first and second solvent shell. Calculated ^1H_N chemical shifts using both the explicit and implicit solvent models are compared in Fig. 5.7. As can be seen, the inclusion of explicit water molecules gives considerably better agreement with experiment over the implicit solvent model. The correlation coefficient (R) between the theoretical and experimental values is improved from 0.673 to 0.835. The RMSE is also decreased from 1.19 to 0.86 ppm. Table 5.2 lists those residues which have amide protons forming hydrogen bonds (H-bonds) with water molecules. It can be seen that those calculated ^1H_N chemical shifts using the pure implicit solvent model show large upfield shifts as compared to experimental values.

When the explicit solvents were included in the fragment QM calculations, the results show significant improvement. It clearly indicates that hydrogen bonding has large electronic polarization effect on the ^1H_N chemical shift (up to 2–3 ppm). The water molecule which forms direct H-bond with the amide proton in proteins should be treated quantum mechanically to accurately reproduce the experimental ^1H_N chemical shifts.

As shown in Fig. 5.7, although the inclusion of explicit water molecules improves the results, the calculated ^1H_N chemical shifts with the explicit solvent model are systematically underestimated by about 0.5 ppm. Previous studies on some model systems have illustrated that the cooperative hydrogen bonding effect has a non-negligible influence on ^1H_N chemical shifts by affecting the primary hydrogen bond geometry and polarizing the electron density around the amide proton. Therefore, we further explored the cooperative hydrogen bond effect on the protein ^1H_N chemical shifts. For simplicity, we took the N-methylacetamide (NMA) as the central fragment, the cooperative hydrogen bonding effects caused

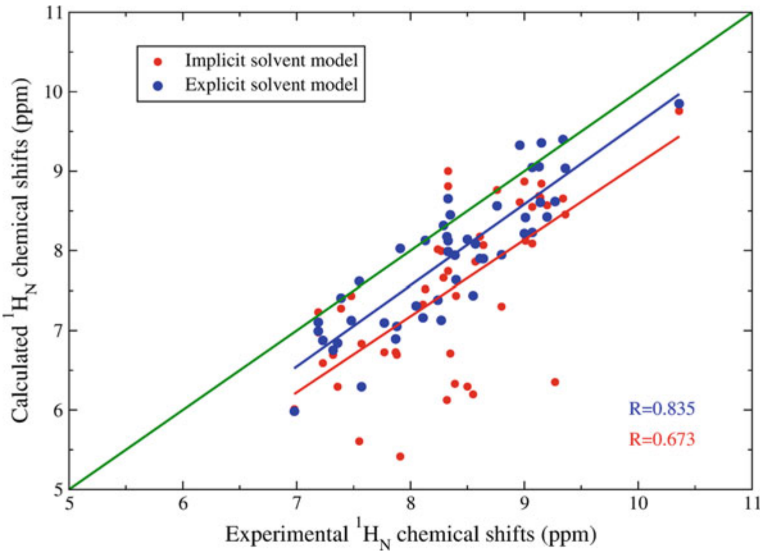

Fig. 5.7 Correlation between the experimental and calculated ^1H_N chemical shifts of GB3 using the AF-QM/MM method (the QM level is at B3LYP/6-31G**). (*red circle* ^1H_N chemical shifts calculated using the implicit solvent model; *blue circle* ^1H_N chemical shift calculated using the explicit solvent model.)

Table 5.2 Comparison of the experimental and AF-QM/MM calculated ^1H_N chemical shifts (in ppm) of GB3 for residues which form hydrogen bonds with water molecules using the explicit and implicit solvent models, respectively

Residue	LEU12	VAL21	ALA23	GLU24	GLY41	TRP43	THR45
Implicit solvation	5.61	6.30	6.13	6.71	5.42	6.35	6.20
Explicit solvation	7.62	8.14	8.18	8.45	8.03	8.62	7.94
Experiment	7.55	8.50	8.32	8.35	7.91	9.27	8.55

The QM level is at B3LYP/6-31G**

by both water and NMA molecules were investigated. As shown in Fig. 5.8, when the cooperative hydrogen bond was formed, the chemical shifts of the ^1H_N atom in the central residue are downfielded by around 0.3–0.5 ppm as opposed to the case of single H-bond. Therefore, we expand our definition of the buffer region to include the secondary hydrogen bond acceptor (the whole residue or water molecule) in the QM region. As depicted in Fig. 5.9, if the ^1H_N chemical shift in the core residue (A) is to be calculated and there is a cooperative hydrogen bond across the peptide bonds of residues: A, B (primary H-bond acceptor) and C (secondary H-bond acceptor), we also include residue C in the buffer region.

The other factors that may govern the accuracy of calculated ^1H_N chemical shifts include the density functional and the size of basis set chosen in our calculation. Previous studies on small organic molecules have demonstrated that, at

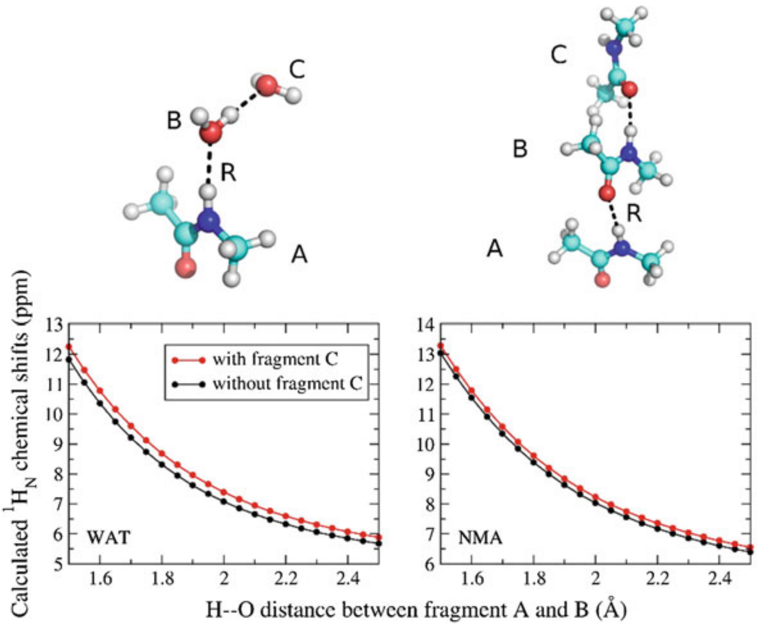

Fig. 5.8 The ^1H_N chemical shift of the central fragment (A) as a function of the ^1H_N-O distance between fragment A and B calculated at the B3LYP/6-311++G** level. *Left panel* both the primary and secondary hydrogen bond acceptors are water molecules; *right panel* both the primary and secondary hydrogen bond acceptors are N-methylacetamides (NMAs). The H-bond length between fragment B and C are fixed at the original optimized structure at the B3LYP/6-31G** level (1.98 Å for WAT-WAT and 2.09 Å for NMA-NMA, respectively.)

least a triple-zeta basis set with the diffuse basis function should be utilized to accurately reproduce the experimental amide hydrogen chemical shift. However, the computational cost is very demanding to apply large basis sets on the entire QM region consisting of normally 150–300 atoms, which is the normal size of each fragment (core + buffer region) using the current definition of the buffer region. Hence, the use of locally dense basis sets, i.e. the combination of two basis sets where the larger one is used for the atoms of interest and the smaller one for all the other atoms, is adopted. The 6-311++G** basis set was employed on the –CO–NH– atoms in both the core residue and other residues involved in the primary and secondary H-bonds (as illustrated in Fig. 5.9). If the H-bond acceptor is a water molecule, the entire water molecule is treated with the 6-311++G** basis set, while the rest atoms in the QM region are set to a smaller basis set. In this work, the 4-31G* basis set has been utilized and the result is shown in Fig. 5.10.

As can be seen, the inclusion of cooperative hydrogen bond effect and applying the locally dense basis set give remarkable improvement for the ^1H_N chemical shifts (compare Fig. 5.10 with Fig. 5.7). The calculation with the B3LYP/6-311++G**/4-31G* method decreases the RMSE from 0.86 to 0.49 ppm. In our previous study [49], we found that the increase of the lower basis set from 4-31G* to 6-31G*

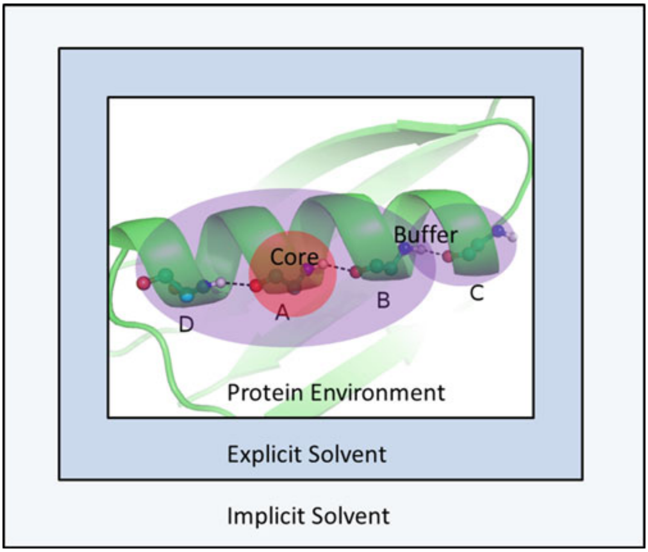

Fig. 5.9 Subsetting scheme for the AF-QM/MM-PB approach with the explicit solvent model. The *red* and *blue* region represents the core and buffer region, respectively. On top of the original definition of the buffer region described in Ref. [41], this study adds one additional criterion which is including the secondary hydrogen bond acceptor (residue C) in the buffer region to take cooperative hydrogen bonding effect into account. The rest of the protein and explicit solvent molecules are described by point charges. The bulk solvent effect is described by the classical electrostatic potential induced by the point charges on the cavity surface calculated using the PB model

or 6-311G** does not reduce the overall RMSE for GB3. Hence, we conclude that the B3LYP functional with the mixed basis set of 6-311++G**/4-31G* strikes a compromise between the computational cost and attained accuracy.

5.3.3 ^{13}C and ^{15}N Chemical Shifts in Proteins

Taking protein GB3 as an example, we also show the influence of solvation effects on the NMR chemical shifts of ^{13}C. The comparison between theoretical and experimental result is shown in Fig. 5.11.

It is not surprising that, as shown in Fig. 5.11, the chemical shift of ^{13}C atom is not significantly influenced by including the solvent effects, with the correlation coefficient of 0.994 in gas phase and 0.998 in implicit solvent. This is mainly because the ^{13}C chemical shifts span a large range from aliphatic region (15–35 ppm) to the carbonyl region (170–180 ppm). To further analyze the result in more details, we also plot the chemical shift of $^{13}C_{\alpha}$ in Fig. 5.12.

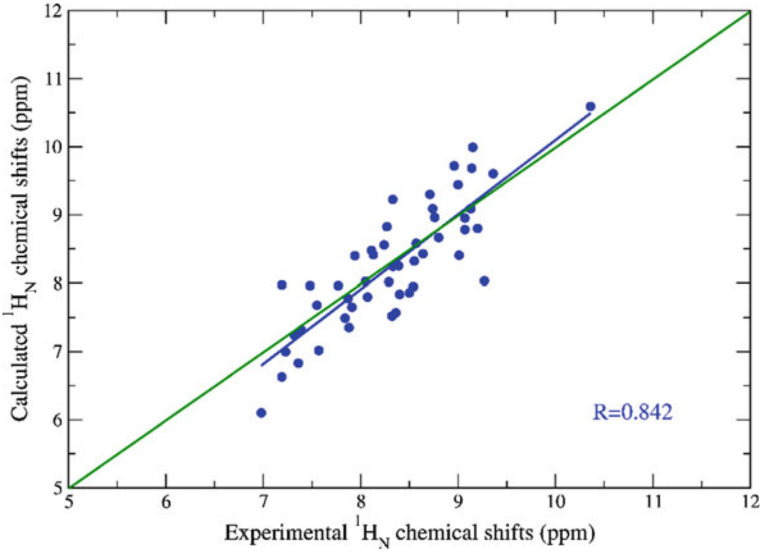

Fig. 5.10 Correlation between experimental and calculated ^1H_N chemical shifts of GB3 using the AF-QM/MM method (the QM level is at B3LYP/6-311++G**/4-31G*)

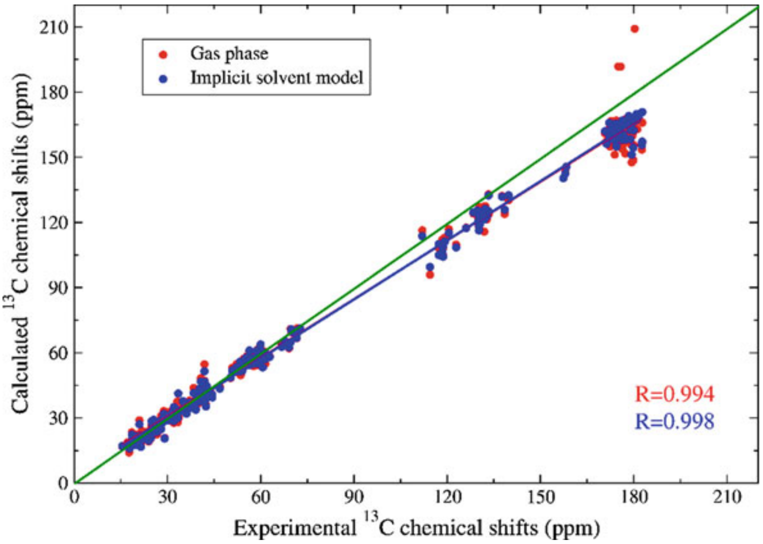

Fig. 5.11 Correlation between the experimental and calculated ^{13}C chemical shifts of GB3 using the AF-QM/MM method (the QM level is at B3LYP/6-31G**) (*red circle* ^{13}C chemical shifts calculated in the gas phase; *blue circle* ^{13}C chemical shifts calculated using the implicit solvent model)

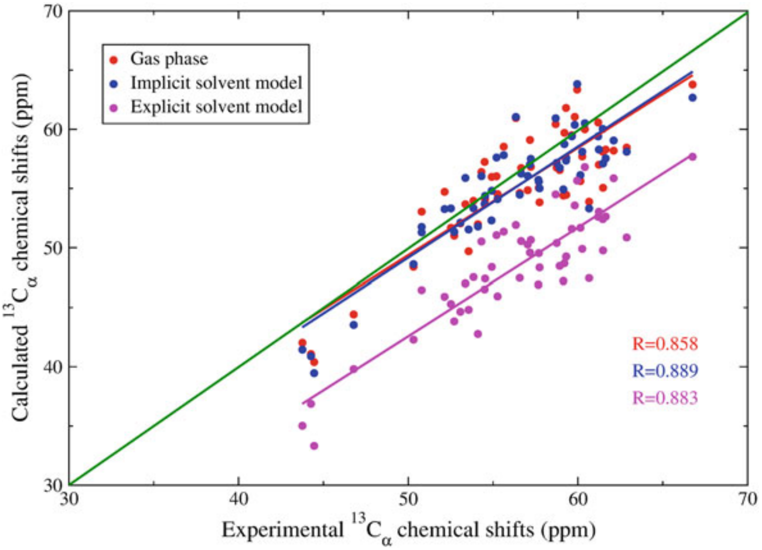

Fig. 5.12 Correlation between the experimental and calculated $^{13}C_\alpha$ chemical shifts of GB3 using the AF-QM/MM method (*red circle* $^{13}C_\alpha$ chemical shifts calculated in the gas phase at the B3LYP/6-31G** level; *blue circle* $^{13}C_\alpha$ chemical shifts calculated using the implicit solvent model at the B3LYP/6-31G** level; *magenta circle* $^{13}C_\alpha$ chemical shifts calculated using the explicit solvent model at the B3LYP/6-311++G**/4-31G* level)

As shown in Fig. 5.12, the inclusion of the implicit solvation model shows some improvement and the overall correlation with experiment increased from 0.858 to 0.889 for GB3, and the RMSE was reduced from 2.89 to 2.41 ppm, but the influence is not very significant. That is mainly because most of the C_α atoms are buried in the hydrophobic core region of the protein and are nonpolar, their chemical shifts are less affected by the solvent. The errors of the calculated $^{13}C_\alpha$ chemical shifts with respect to the experimental values are likely due to the insufficient sampling of the protein structure, since the experimental observed chemical shifts represent the ensemble-averaged values. When the explicit waters were included, the calculated results did not show any improvement and the calculated $^{13}C_\alpha$ chemical shifts using the 6-311++G** basis set are all clearly underestimated with respect to the experimental values. Similar findings have also been concluded in the previous study by Case and co-workers [28].

The ^{15}N results are summarized in Fig. 5.13. It has long been known that the ^{15}N chemical shift in protein presents a challenge for first principle prediction because they are very sensitive to the chemical environment and are influenced by numerous factors. To accurately predict the ^{15}N NMR chemical shifts usually requires high-level electron correlation methods beyond DFT.

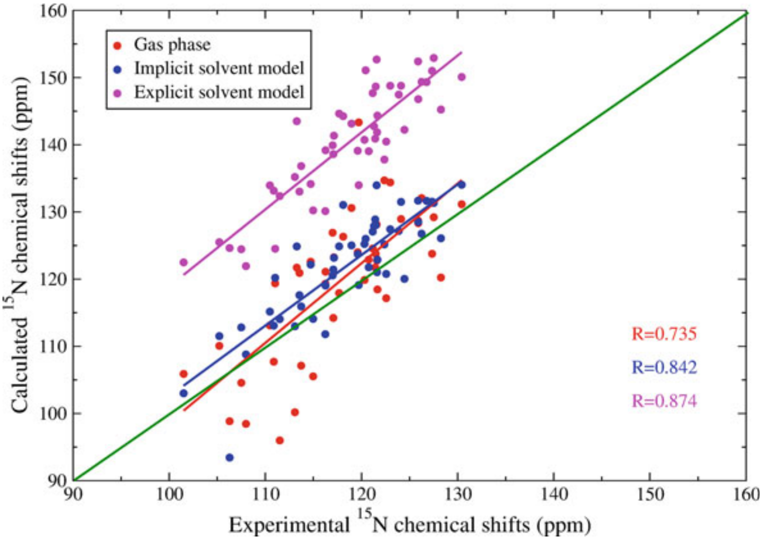

Fig. 5.13 Correlation between experimental and calculated ^{15}N chemical shifts of GB3 using the AF-QM/MM method (*red circle* ^{15}N chemical shifts calculated in the gas phase at the B3LYP/6-31G** level; *blue circle* ^{15}N chemical shifts calculated using the implicit solvent model at the B3LYP/6-31G** level; *magenta circle* ^{15}N chemical shifts calculated using the explicit solvent model at the B3LYP/6-311++G**/4-31G* level)

From the B3LYP/6-31G** calculation, the correlation (*R*) between the calculated and experimental ^{15}N chemical shifts is only 0.735 for GB3 in the gas phase. Although the inclusion of solvent effects shows significant improvement (with the correlation of 0.842 for GB3), it still has large deviations from the experimental values. As one can see from Fig. 5.13, the implicit solvation treatment on the nitrogen atoms improves more significantly than the nonpolar C_α atoms. However, as discussed in our previous study [40], there is a difference between backbone and side chain nitrogen atoms. For backbone amide nitrogen, which is buried in the core region of protein, the solvent effects on the ^{15}N chemical shift are relatively weak, and the calculated shifts are usually larger than the experimentally measured values [40]. In contrast, solvent effects on the nitrogen atoms from the side chain amine groups (mostly exposed to the solvent) are stronger. However, as shown in Fig. 5.13, including explicit water molecules did not give much improvement. The correlation coefficient is marginally increased from 0.842 to 0.874. Besides the solvent effect, there are other factors which may govern the accuracy of theoretical prediction on ^{15}N NMR chemical shifts, such as conformational sampling, the choice of DFT functionals, etc. Research along these lines is currently underway in our laboratory.

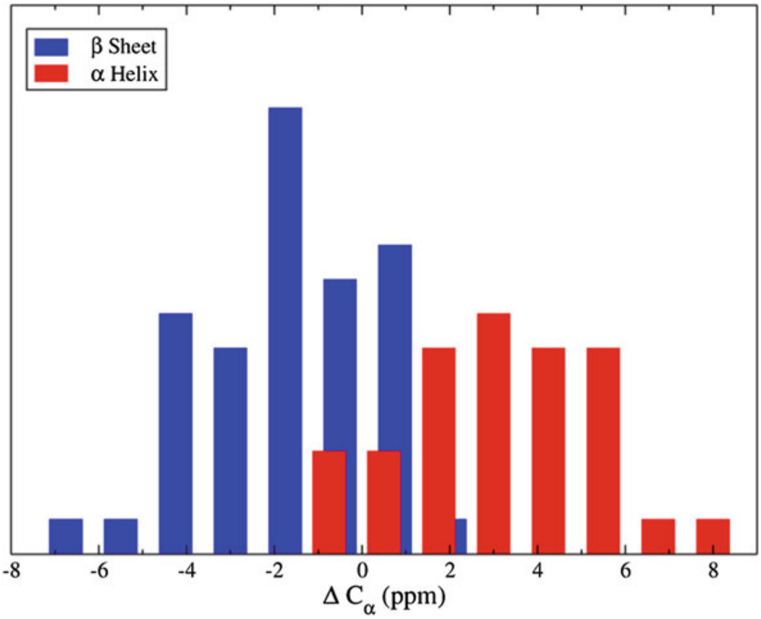

Fig. 5.14 Histograms of secondary shift (the calculated chemical shifts in the native structure minus the random coil values) distribution of α-helical and β-sheet chemical shifts for $^{13}C_\alpha$ in four proteins (Trp-cage, Pin1 WW domain, GB3 and ubiquitin)

5.3.4 Probing the Secondary Structures of Proteins by the AF-QM/MM Method

In structural biology, the protein chemical shift is a powerful tool for studying the structure and dynamics of the protein. They are often used to predict regions of secondary structure in native and nonnative states of proteins, to aid the refinement of complex structures and characterization of conformational changes. Here we validated the capability of using the $^{13}C_\alpha$ secondary chemical shifts (i.e. the calculated chemical shifts in the native structure minus the random coil values) calculated by AF-QM/MM approach to distinguish the α-helix and β-sheet structures. The $^{13}C_\alpha$ random coil chemical shifts are taken from the CamCoil module [54]. The calculated results are presented in Fig. 5.14. As expected, there is a clear separation between the shieldings of the two secondary structure types. The $^{13}C_\alpha$ chemical shift experiences a downfield shift with an average value of 2.55 ppm (with respect to the random coil value) when in a helical configuration and a comparable upfield shift of −2.38 ppm in average when in β-sheet configuration. It shows that the AF-QM/MM method accurately reflects the influence of the local geometry on the chemical shift calculation.

Recent studies [3, 6] have reported that, in combination with traditional molecular mechanical force field or de novo protein structure sampling techniques,

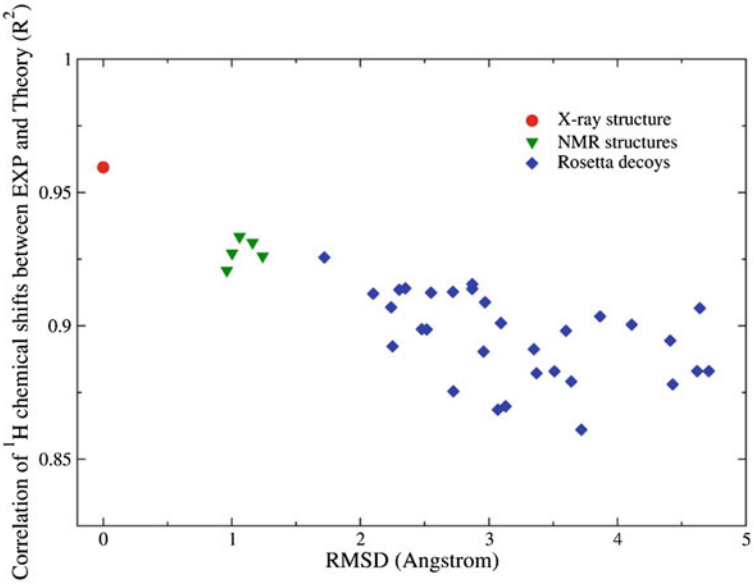

Fig. 5.15 Correlation between the experimental and calculated ^{1}H chemical shifts versus backbone RMSD for Pin1 WW domain (PDB entries for the X-ray structure and NMR structures are 1PIN and 1I6C, respectively)

protein structures can be derived using ^{1}H, ^{13}C and ^{15}N NMR chemical shifts. Hence, we also used the ^{1}H chemical shifts calculated by the AF-QM/MM method in detecting misfolded proteins relative to the natively folded target protein. The X-ray structure of Pin1 WW domain was taken as the native structure and a set of decoy structures for the same amino acid sequence was generated using Rosetta program. Figure 5.15 gives the correlation between calculated and experimental measured chemical shifts versus backbone RMSD with respect to the X-ray structure. As indicated, with the increase of the backbone RMSD, the NMR chemical shift correlation is declined. The RMSD values with the lowest correlations are far from the native structure. The results show that using the chemical shifts to detect the native structure from a collection of decoys is quite remarkable and would have significant potential in this regard.

5.4 Conclusions

In this review, we discussed the performance of QM methods on the calculation of protein chemical shifts based on the recently developed AF-QM/MM approach. By using the PB model and first solvation water molecules, the influence of solvent effect is also explored. Benefit from the fragment algorithm, the AF-QM/MM approach is computationally efficient and linear-scaling with a low pre-factor. The calculation for each residue takes about 2–4 h of computer time using the current definition of the buffer region. The approach is massively parallel and can be applied to routinely calculate the *ab initio* NMR chemical shifts for proteins of any size.

The calculated results also indicate that when the solvent effect is included, the calculated ^{1}H and ^{15}N chemical shifts show remarkable improvement over those from the gas phase calculations, while the nonpolar ^{13}C chemical shifts are less affected by the solvent. In addition, to accurately calculate the $^{1}H_{N}$ chemical shifts, the explicit solvent method should be taken into account. However, although the inclusion of solvent effect shows significant improvement for ^{15}N chemical shifts, they still have large deviations from the experimental values.

Our study also demonstrated that the AF-QM/MM calculated result accurately reflects the dependence of $^{13}C_{\alpha}$ chemical shifts on the secondary structure of proteins, and the use of ^{1}H chemical shift to discriminate the native structure of proteins from decoys is quite remarkable as proton chemical shift is highly influenced by the local chemical environment. The use of *ab initio* calculated chemical shifts is capable of facilitating accurate protein structure refinement and determination.

The AF-QM/MM method can be further utilized to predict other local chemical properties, such as chemical shift tensor anisotropies and J coupling constants. The applications may also be extended to more general biological systems, such as proteins with nonstandard residues, metalloproteins, protein-ligand, protein-DNA/RNA and membrane protein-lipid complexes.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grants No. 10974054, 20933002 and 21303057) and Shanghai PuJiang program (09PJ1404000). X.H. is also supported by the Specialized Research Fund for Doctoral Program of Higher Education (Grant No. 20130076120019) and the Fundamental Research Funds for the Central Universities. We thank the Supercomputer Center of East China Normal University for providing us computational time. X.H. also gratefully acknowledges many helpful discussions with Kenneth Merz, Bing Wang, Ning Liao, David Case and Sishi Tang.

References

1. Bieri M, Kwan AH, Mobli M, King GF, Mackay JP, Gooley PR (2011) Macromolecular NMR spectroscopy for the non-spectroscopist: beyond macromolecular solution structure determination. FEBS J 278:704–715
2. Kwan AH, Mobli M, Gooley PR, King GF, Mackay JP (2011) Macromolecular NMR spectroscopy for the non-spectroscopist. FEBS J 278:687–703
3. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104:9615–9620
4. Lang WH.; Coats JE, Majka J, Hura GL, Lin Y, Rasnik I, McMurray CT (2011) Conformational trapping of mismatch recognition complex MSH2/MSH3 on repair-resistant DNA loops. Proc Natl Acad Sci USA 108:1–8
5. Selvaratnam R, Chowdhury S, VanSchouwen B, Melacini G (2011) Mapping allostery through the covariance analysis of NMR chemical shifts. Proc Natl Acad Sci USA 108:6133–6138

6. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690
7. Wylie BJ, Sperling LJ, Nieuwkoop AJ, Franks WT, Oldfield E, Rienstra CM (2011) Ultrahigh resolution protein structures using NMR chemical shift tensors. Proc Natl Acad Sci USA 108:16974–16979
8. Ulmer TS, Ramirez BE, Delaglio F, Bax A (2003) Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J Am Chem Soc 125:9179–9191
9. De Gortari I, Portella G, Salvatella X, Bajaj VS, van der Wel PC, Yates JR, Segall MD, Pickard CJ, Payne MC, Vendruscolo M (2010) Time averaging of NMR chemical shifts in the MLF peptide in the solid state. J Am Chem Soc 132:5993–6000
10. Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Using side-chain aromatic proton chemical shifts for a quantitative analysis of protein structures. Angew Chem Int Ed 50:9620
11. Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Structure-based prediction of methyl chemical shifts in proteins. J Biomol NMR 50:331–346
12. Helgaker T, Jaszunski M, Ruud K (1999) Ab initio methods for the calculation of NMR shielding and indirect spin-spin coupling constants. Chem Rev 99:293–352
13. Facelli JC (2011) Chemical shift tensors: theory and application to molecular structural problems. Prog Nucl Magn Reson Spectrosc 58:176–201
14. Saito H, Ando I, Ramamoorthy A (2010) Chemical shift tensor—the heart of NMR: insights into biological aspects of proteins. Prog Nucl Magn Reson Spectrosc 57:181–228
15. Moon S, Case DA (2007) A new model for chemical shifts of amide hydrogens in proteins. J Biomol NMR 38:139–150
16. Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131:13894–13895
17. Baskaran K, Brunner K, Munte CE, Kalbitzer HR (2010) Mapping of protein structural ensembles by chemical shifts. J Biomol NMR 48:71–83
18. Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57
19. Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Structure-based prediction of methyl chemical shifts in proteins. J Biomol NMR 50:331
20. De Dios AC, Pearson JG, Oldfield E (1993) Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. Science 260:1491–1496
21. De Dios AC (1996) Ab initio calculations of the NMR chemical shift. Prog Nucl Magn Reson Spectrosc 29:229–278
22. De Dios AC, Pearson JG, Oldfield E (2008) Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. J. Chem. Phys. 260:1491–1496
23. Ochsenfeld C, Kussmann J, Koziol F (2004) Ab initio NMR spectra for molecular systems with a thousand and more atoms: a linear scaling method. Angew Chem Int Ed 43:4485–4489
24. Oldfield E (2002) Chemical shifts in amino acids, peptides, and proteins: from quantum chemistry to drug design. Ann Rev Phys Chem 53:349–378
25. Abildgaard J, Hansen PE, Manalo MN, LiWang A (2009) Deuterium isotope effects on 15 N backbone chemical shifts in proteins. J Biomol NMR 44:119–126
26. Tang S, Case DA (2011) Calculation of chemical shift anisotropy in proteins. J Biomol NMR 51:303
27. Beer M, Kussmann J, Ochsenfeld C (2011) Nuclei-selected NMR shielding calculations: a sublinear-scaling quantum-chemical method. J Chem Phys 134:074102
28. Moon S, Case DA (2006) A comparison of quantum chemical models for calculating NMR shielding parameters in peptides: mixed basis set and ONIOM methods combined with a complete basis set extrapolation. J Comput Chem 27:825–836

29. Komin S, Gossens C, Tavernelli I, Rothlisberger U, Sebastiani D (2007) NMR solvent shifts of adenine in aqueous solution from hybrid QM/MM molecular dynamics simulations. J Phys Chem B 111:5225–5232
30. Hinton JF, Guthrie P, Pulay P, Wolinski K (1992) Ab initio quantum mechanical calculation of the chemical shift anisotropy of the hydrogen atom in the (H_2O) 17 cluster. J Am Chem Soc 114:1604
31. Vila JA, Aramini JM, Rossi P, Kuzin A, Su M, Seetharaman J, Xiao R, Tong L, Montelione GT, Scheraga HA (2008) Quantum chemical C-13(alpha) chemical shift calculations for protein NMR structure determination, refinement, and validation. Proc Natl Acad Sci USA 105:14389–14394
32. Buehl M, Mourik VT (2011) NMR spectroscopy: quantum-chemical calculations, *Wires Comput*. Mol. Sci. 1:634–647
33. Mourik VT (2006) Density functional theory reveals an increase in the amino (1)H chemical shift in guanine due to hydrogen bonding with water. J Chem Phys 125:191101
34. Cui Q, Karplus M (2000) Molecular properties from combined QM/MM methods. 2. Chemical shifts in large molecules. J Phys Chem B 104:3721–3743
35. Frank A, Onila I, Möller HM, Exner TE (2011) Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins. Proteins 79:2189–2202
36. Exner TE, Frank A, Onila I, Moeller HM (2012) Toward the quantum chemical calculation of nmr chemical shifts of proteins. 3. conformational sampling and explicit solvents model. J Chem Theory Comput 8:4818–4827
37. Frank A, Moeller HM, Exner TE (2012) Toward the quantum chemical calculation of NMR chemical shifts of proteins. 2. Level of theory, basis set, and solvents model dependence. J Chem Theory Comput 8:1480–1492
38. Gao Q, Yokojima S, Kohno T, Ishida T, Fedorov DG, Kitaura K, Fujihira M, Nakamura S (2007) Ab initio NMR chemical shift calculations on proteins using fragment molecular orbitals with electrostatic environment. Chem Phys Lett 445:331–339
39. Gao Q, Yokojima S, Fedorov DG, Kitaura K, Sakurai M, Nakamura S (2010) Fragment-molecular-orbital-method-based ab initio NMR chemical-shift calculations for large molecular systems. J Chem Theory Comput 6:1428–1444
40. Zhu T, He Xiao, Zhang JZH (2012) Fragment density functional theory calculation of NMR chemical shifts for proteins with implicit solvation. Phys Chem Chem Phys 14:7837–7845
41. He X, Wang B, Merz KM (2009) Protein NMR chemical shift calculations based on the automated fragmentation QM/MM approach. J Phys Chem B 113:10380–10388
42. Mogelhoj A, Aidas K, Mikkelsen KV, Kongsted J (2008) Solvent effects on the nitrogen NMR shielding and nuclear quadrupole coupling constants in 1-methyltriazoles. Chem Phys Lett 460:129–136
43. Kitevski-LeBlanc JL, Evanics F, Prosser RS (2009) Approaches for the measurement of solvent exposure in proteins by 19F NMR. J Biomol NMR 45:255–264
44. Dracinsky M, Bour P (2010) Computational analysis of solvent effects in NMR spectroscopy. J Chem Theory Comput 6:288–299
45. Witanowski M, Biedrzycka Z, Sicinska W, Grabowski Z (1998) A study of solvent polarity and hydrogen bonding effects on the nitrogen NMR shielding of isomeric tetrazoles and ab initio calculation of the nitrogen shielding of azole systems. J Magn Reson 131:54–60
46. Witanowski M, Sicinska W, Biedrzycka Z, Webb GA (1996) Solvent effects on the nitrogen NMR shieldings of cyanamide and N,N-dimethyl cyanamide. J Mol Struc 380:133
47. Mennucci B, Martinez JM, Tomasi J (2001) Solvent effects on nuclear shieldings: continuum or discrete solvation models to treat hydrogen bond and polarity effects? J Phys Chem A 105:7287–7296
48. Tang M, Sperling LJ, Berthold DA, Schwieters CD, Nesbitt AE, Nieuwkoop AJ, Gennis RB, Rienstra CM (2011) High-resolution membrane protein structure by joint calculations with solid-state NMR and X-ray experimental data. J Biomol NMR 51:227–233

49. Zhu T, Zhang JZH, He X (2013) Automated fragmentation QM/MM calculation of amide proton chemical shifts in proteins with explicit solvent model. J Chem Theory Comput 9:2104–2114
50. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JAJ, 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, Adamao 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 J (2010) Gausian09, revision B.01, Gaussian, Inc., Wallingford, CT
51. Dixon SL, van der Vaart A, Gogonea V, Vincent M, Brothers EN, Suarez D, Westerhoff LM, Jr. Merz KM (1999) DivCon. The Pennsylvania State University, University Park, PA
52. Imai T, Hiraoka R, Kovalenko A, Hirata F (2007) Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. Proteins 66:804–813
53. Yoshida N, Phongphanphanee S, Maruyama Y, Imai T, Hirata F (2006) Selective ion-binding by protein probed with the 3D-RISM theory. J Am Chem Soc 128:12042–12043
54. De Simone A, Cavalli A, Hsu S-TD, Vranken W, Vendruscolo M (2009) Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. J Am Chem Soc 131:16332