J Biomol NMR (2012) 52:257–267
DOI 10.1007/s10858-012-9609-6

ARTICLE

Combining NMR ensembles and molecular dynamics simulations provides more realistic models of protein structures in solution and leads to better chemical shift prediction

Juuso Lehtivarjo · Kari Tuppurainen · Tommi Hassinen · Reino Laatikainen · Mikael Peräkylä

Received: 28 October 2011 / Accepted: 16 January 2012 / Published online: 8 February 2012

Abstract While chemical shifts are invaluable for obtaining structural information from proteins, they also offer one of the rare ways to obtain information about protein dynamics. A necessary tool in transforming chemical shifts into structural and dynamic information is chemical shift prediction. In our previous work we developed a method for 4D prediction of protein ^{1}H chemical shifts in which molecular motions, the 4th dimension, were modeled using molecular dynamics (MD) simulations. Although the approach clearly improved the prediction, the X-ray structures and single NMR conformers used in the model cannot be considered fully realistic models of protein in solution. In this work, NMR ensembles (NMRE) were used to expand the conformational space of proteins (e.g. side chains, flexible loops, termini), followed by MD simulations for each conformer to map the local fluctuations. Compared with the non-dynamic model, the NMRE+MD model gave 6–17% lower root-mean-square (RMS) errors for different backbone nuclei. The improved prediction indicates that NMR ensembles with MD simulations can be used to obtain a more realistic picture of protein structures in solutions and moreover underlines the importance of short and long time-scale dynamics for the prediction. The RMS errors of the NMRE+MD model were 0.24, 0.43, 0.98, 1.03, 1.16 and 2.39 ppm for ^{1}Hα, ^{1}HN, ^{13}Cα, ^{13}Cβ, ^{13}CO and backbone ^{15}N chemical shifts, respectively. The model is implemented in the prediction program 4DSPOT, available at http://www.uef.fi/4dspot.

Keywords Protein · Chemical shift · Prediction · Molecular dynamics · NMR ensembles

Introduction

Chemical shifts are strongly dependent on the tertiary structure of the protein. This structural information has been exploited in determining protein structures using chemical shifts and sequence-based modeling (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008). Recently, restraints for molecular dynamics simulations have been created based on chemical shift prediction, allowing the use of chemical shifts in the model calculation in a more straightforward way (Robustelli et al. 2010).

Over the last decade, a number of empirical protein chemical shift prediction methods have been presented. The growth of databases makes it possible to yield accurate predictions with *homology-based* methods such as the SHIFTY+ part of SHIFTX2 (Han et al. 2011), but the inclusion of sequence similarity-based data does not provide explicit structural information and is less useful when the function of the protein is considered. On the other hand, in recent *structure-based* approaches (Kohlhoff et al. 2009; Liu et al. 2011; Neal et al. 2003; Shen and Bax 2007, 2010; Xu and Case 2001), the prediction accuracy seems to be facing the barrier of the protein structure resolution. In these methods, the teaching data usually consists of X-ray structures or single conformers of NMR structure

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9609-6) contains supplementary material, which is available to authorized users.

J. Lehtivarjo (✉) · K. Tuppurainen · T. Hassinen · R. Laatikainen
School of Pharmacy, University of Eastern Finland,
P.O. Box 1627, 70211 Kuopio, Finland
e-mail: juuso.lehtivarjo@uef.fi

M. Peräkylä
Institute of Biomedicine, University of Eastern Finland,
P.O. Box 1627, 70211 Kuopio, Finland

ensembles, although chemical shifts are measured in solvent, where protein structures are dynamic. Recently, the importance of dynamics in chemical shift prediction has been recognized in the literature (Baskaran et al. 2010; Lehtivarjo et al. 2009; Li and Brueschweiler 2010; Markwick et al. 2010).

Protein structures determined by NMR are published mostly as ensembles in which usually 20 conformers are used to represent the structure in solvent. Flexibility in protein moieties reduces the number of restraints available for model calculation, allowing different conformations for side chains and random coils to be created. Together these conformations then resemble long time-scale motions in the ensembles. Baskaran et al. (2010) showed that by averaging the prediction results of SHIFTX (Neal et al. 2003) or SHIFTS (Xu and Case 2001) over the conformations of the NMR ensembles, about a 9% improvement in prediction accuracy was gained compared with the prediction using only the lowest energy conformations. On the other hand, short MD simulations, mapping local fluctuations, gave a 6–7% benefit for ^{1}H shifts (Lehtivarjo et al. 2009). In order to gain more extensively mapped protein dynamics for the chemical shift prediction, we decided to expand the 4th dimension (conformational space) of our prediction model by combining both approaches. Briefly, this is done by performing MD simulations for all the conformers of the NMR ensembles. As in our previous study, the prediction model itself is also built from dynamic protein models. This study also examined the prediction efficiency of NMR structures, which have been often ignored in earlier studies. The computer program 4DSPOT (4–Dimensional Shift Prediction: averaged Over Time) was updated to use this new model, now also predicting ^{13}C and ^{15}N chemical shifts.

Methods

Database

Altogether 94 NMR structure ensembles, containing a total of 1,809 conformers, (Supplementary material Table S1) were downloaded from the PDB (Protein Data Bank, Berman et al. 2000) and the corresponding observed chemical shifts from the BMRB (Biological Magnetic Resonance Bank Ulrich et al. 2008). Observed chemical shifts were re-referenced using LACS (Wang and Markley 2009) to correct possible biases caused by non-standard sample conditions or misreferencing. A total of 49,939 ^{1}H shifts, 36,135 ^{13}C shifts and 9,492 ^{15}N shifts were imported to the teaching database. The proteins of the teaching database are ligand-free monomers of sizes varying from 46 to 202 amino acid residues, determined with standard protein NMR methods. To achieve the prediction accuracy stated in this study, the query proteins must fulfil the above criteria. As in the previous version of the 4DSPOT program, most of the side chain shifts can be predicted. Included are the side chain methyl groups, whose importance has recently been emphasized (Sahakyan et al. 2011). However, due to the low number of data points, several atom types (see Supplementary material text) are not predicted. For the same reason, the ^{15}N side chain shift prediction is still at an inadequate level, and thus omitted from the results.

Molecular dynamics

Local fluctuations of the protein conformers were mapped with 100 ps MD simulations performed using the AMBER 10 program (Case et al. 2008) and the ff99SB force field (Hornak et al. 2006). Protein conformers were solvated by TIP3P water molecules in periodic solvent boxes extending at least 11 Å from the protein atoms. To neutralize the total charge of the simulation systems, an adequate number of Na^{+} or Cl^{-} ions were added. The equilibration protocol of the MD simulations is similar as before (Lehtivarjo et al. 2009). In the production simulations of 100 ps the electrostatics were treated using the particle mesh Ewald method. A time step of 1.5 fs was used and bonds to hydrogen atoms were constrained to their equilibrium lengths using the SHAKE algorithm. During the production simulations structures were saved every 1 ps, yielding 100 snapshots for every conformer. For each protein, the 100 ps trajectories of the individual conformers are then read into the prediction program to create a combined trajectory of length of 1–4 ns, depending on the conformer count.

To assess the effects of the dynamics of different timescales for the prediction results, four different prediction models were created. (1) The non-dynamic model was built from single conformers, without any dynamics, from the "most representative conformers" of the PDB files, thus setting up the baseline for the evaluation. (2) The NMR ensemble (NMRE) model was built using all the conformers of the NMR ensembles as the dynamic data, but without the MD simulations. (3) In the molecular dynamics (MD) model the effect of molecular dynamics is evaluated by carrying out MD simulations for the most representative conformers only. (4) Ultimately, the NMRE+MD model includes the dynamics of both the NMR ensembles and the MD simulations. The schematic presentation of these models as well as the total number of conformations used for each model is shown in Fig. 1.

Prediction protocol

The prediction protocol, for the most part, is the same as that presented in Lehtivarjo et al. (2009). Briefly, the

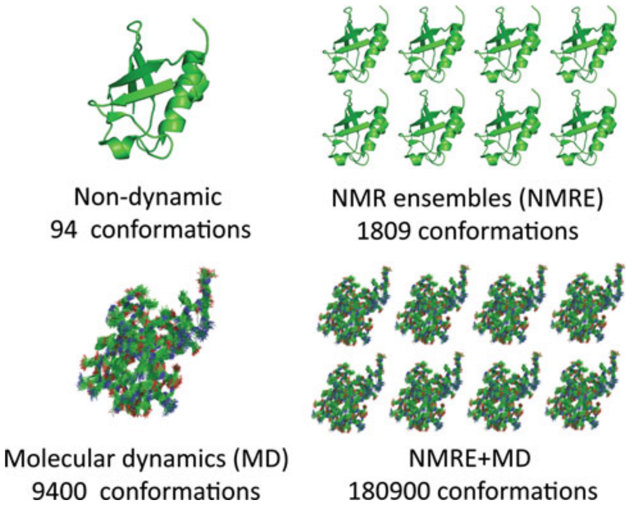

Fig. 1 Schematic presentation of different prediction models used in this study with a number of conformations used for building each model. A database of 94 proteins was used in all models

method derives *molecular descriptors* (Supplementary material Table S2) from protein structures which are then averaged over the combined conformational space of the NMR ensembles and the MD simulations. 194, 142 and 95 descriptors are used for ^{1}H, ^{13}C and ^{15}N models, respectively. Principal component regression (PCR) is then applied to solve the weight factors P for the averaged descriptors $\langle X \rangle$. Chemical shifts can then be calculated by Eq. 1,

$$\delta_n = \delta_n{}^{\circ} + \sum \mathrm{P_i} \langle X_i \rangle \tag{1}$$

in which $\delta_n{}^{\circ}$ is the base value of the chemical shift. The prediction work flow is also similar as before, including the phases of (1) calculating an initial prediction result, (2) based on the initial result, removing the worst 10% of data points from the teaching set for the sake of data certainty, (3) creating and applying correlation parameters X_iX_j and (4) applying PCR locally to individual chemical shift classes (Table 1). The new protocol differs from the previous by introducing a random forest regression (Breiman 2001) protocol as the final adjustment of the prediction. The random forest regression provides some advance to the prediction by searching strong correlations between descriptors. The method will be discussed in more detail elsewhere. The previously unpublished ^{13}C and ^{15}N predictions of 4DSPOT follow the same protocol as ^{1}H shifts, only with different set of descriptors (Table S2). The modelling of solvent effects is improved from the previous descriptor set: in the new method, the solvent molecules are imported all the way to the predictor and explicit solvent descriptors are then created.

The importance of implementing dynamics in the prediction model must be emphasized (Fig. 2). In principle, the dynamic information can be exploited by predicting MD or NMR ensemble conformers one-by-one and then averaging the results, as has been done in the studies of Baskaran et al. (2010) and Markwick et al. (2010). In these studies, however, the prediction models are built from non-dynamic X-ray structures (Neal et al. 2003; Xu and Case 2001). As it is shown that single conformations cannot describe the chemical shifts of a protein correctly (Baskaran et al. 2010; Lehtivarjo et al. 2009; Li and Brueschweiler 2010; Markwick et al. 2010), the teaching data should also be dynamically averaged to gain the most realistic structure. In this study, the prediction model is built from molecular descriptors averaged over the conformations of NMR ensembles and MD simulations. When this approach is used, the query proteins are treated similarly and the prediction is performed in one phase for the whole conformational space.

Table 1 Chemical shift classes used for ^{1}H, ^{13}C and ^{15}N prediction

^{1}H	^{13}C	^{15}N
Hα	Cα	Backbone N
HN	CO	Side chain N
Hβ	Cβ	
Side chain CH_3	Side chain CH_3	
Side chain CH_2	Side chain CH_2	
Proline CH_2	sp^2 C	
Side chain CH		
XH		
Aromatic H		

Results and discussion

All the following results are obtained using the leave-one-out cross validation protocol in which the protein shifts are predicted by excluding the currently predicted protein from the teaching database.

Comparison of non-dynamic and dynamic prediction results

RMS errors of the four different prediction models (see "Methods") are compared in Fig. 3. For the R correlation coefficients and mean errors of the same data, see Supplementary material Tables S3 and S4. As previously shown (Baskaran et al. 2010; Lehtivarjo et al. 2009; Li and Brueschweiler 2010; Markwick et al. 2010), including any dynamics to the prediction model improves the prediction results. As anticipated, the NMRE+MD model yielded the lowest RMS errors. More notably, the improvements from dynamics seen in the combined model are equal or larger

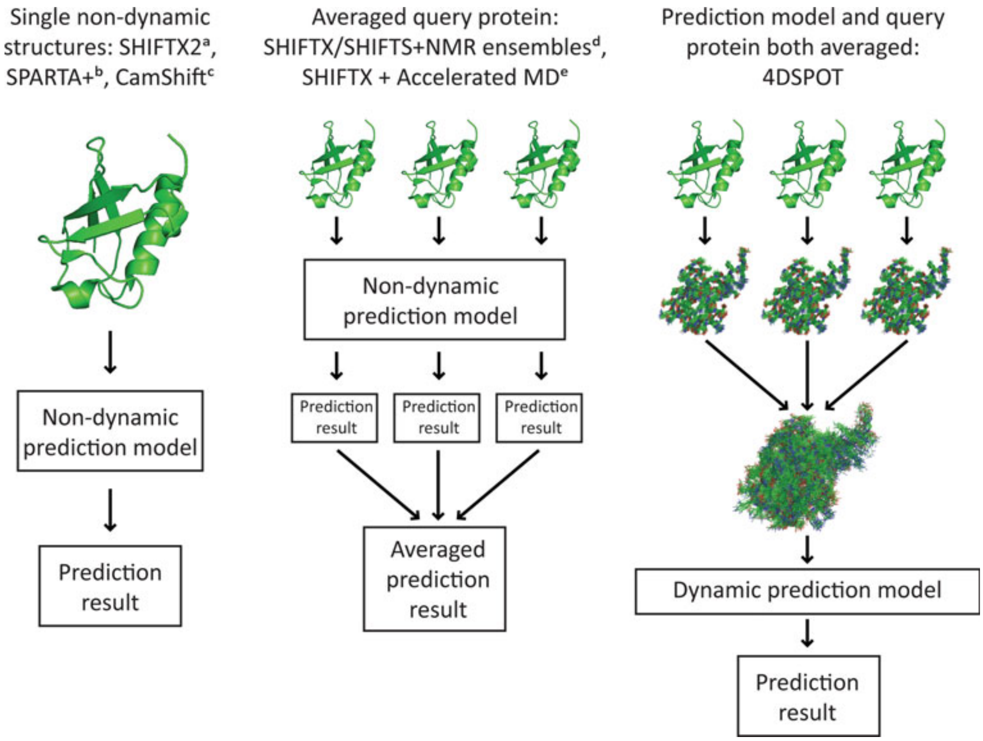
Single non-dynamic structures: SHIFTX2ᵃ, SPARTA+ᵇ, CamShiftᶜ
Non-dynamic prediction model
Prediction result
Averaged query protein: SHIFTX/SHIFTS+NMR ensemblesᵈ, SHIFTX + Accelerated MDᵉ
Non-dynamic prediction model
Prediction result
Prediction result
Prediction result
Averaged prediction result
Prediction model and query protein both averaged: 4DSPOT
Dynamic prediction model
Prediction result

Fig. 2 Schematic presentation of the dynamic approaches of the recent prediction studies. [a] Han et al. (2011), [b] Shen and Bax (2010), [c] Kohlhoff et al. (2009), [d] Baskaran et al. (2010), [e] Markwick et al. (2010)

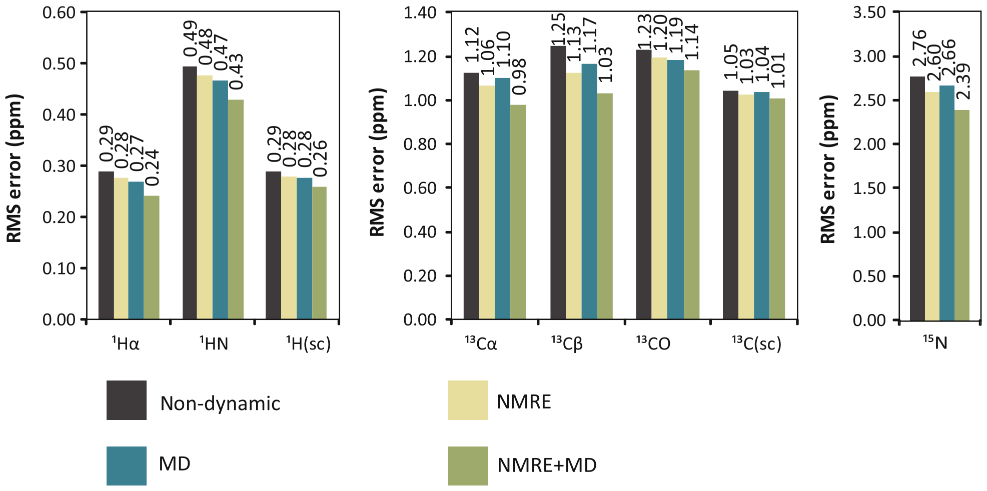

Fig. 3 Prediction RMS errors for backbone and side chain (sc) nuclei predicted with four different models

than the cumulative improvements of its parts (NMRE model + MD model) for all nuclei except ^{13}CO, with synergistic benefits up to 5% (Table 2). This confirms that the NMR ensembles and short MD simulations truly map the dynamics of different time-scales. The protein-specific prediction results of the NMRE+MD model are shown in the Supplementary material Table S5.

Compared with the non-dynamic model, the improvement gained with the NMRE+MD model was on average 13% for backbone nuclei, being largest for ^{13}Cβ and ^{1}Hα shifts (17.1 and 16.4%, respectively). For ^{1}Hα shifts, this improvement was almost three times larger than in our previous study with 40 protein database of single NMR conformers and X-ray structures and motions mapped with 150 ps MD simulations. Furthermore, the ^{1}H results are notably better than in our previous work, in which they were 0.29, 0.50 and 0.28 for ^{1}Hα, ^{1}HN and side chain ^{1}H shifts. For side chains, the dynamic effect is smaller than for backbone shifts. This is mostly due to low sensitivity of side chain shifts to structural effects. However, also in side chain prediction the NMRE+MD model yielded the best results.

Table 2 Synergistic benefits of NMRE+MD model for backbone and side chain (sc) shifts

Nuclei	Expected RMS error of NMRE+MD model (ppm)[a]	RMS error, NMRE+MD model (ppm)	Synergistic benefit (%)
^{1}Hα	0.25	0.24	3.8
^{1}HN	0.44	0.43	3.4
^{13}Cα	1.03	0.98	5.1
^{13}Cβ	1.03	1.03	0.1
^{13}CO	1.15	1.14	1.0
^{15}N	2.49	2.39	4.1
^{1}H(sc)	0.27	0.26	3.5
^{13}C(sc)	1.01	1.01	0.8

[a] Calculated by multiplying the result of non-dynamic model with the dynamic improvements of the NMR ensemble model and MD model

The two dynamic parts of the combined model, NMR ensembles and MD simulations, have dissimilar effects on backbone nuclei. Namely, the NMRE model gives somewhat lower RMS errors than the MD model for ^{13}Cα, ^{13}Cβ and backbone ^{15}N nuclei, whose chemical shifts are dominated by torsion angle effects (Wishart and Case 2001). These effects are evidently mapped better with dynamics of longer time-scale that NMR ensembles mimic. ^{13}CO and ^{1}HN nuclei, which are known to be sensitive to local structure and dynamics due to r^{-3} dependency of hydrogen bonding (Dedios and Oldfield 1994; Moon and Case 2007; Parker et al. 2006), and ^{1}Hα nuclei, which are exposed to aromatic ring currents (Wishart and Case 2001), are slightly better predicted with the MD model.

For most nuclei, the plot of prediction error versus observed shift is biased: both the low and high values of shifts are poorly predicted. This behavior is best seen in ^{1}HN shifts (Fig. 4), in which the variance between random coil shifts of different residues is minimal: a strong correlation between observed chemical shift and prediction error is found. The bias is described by the R of the correlation: the smaller the correlation, the better. When we analyzed the Cα shift prediction data provided by the study of Liu et al. (2011), and the data from our own tests with SPARTA+ (Shen and Bax 2010), similar trends were observed. This means that the prediction, or molecular models, or both are not capable of properly describing the strong effects, such as aromatic ring currents (low values) and hydrogen bonding (high values) in the case of ^{1}HN shifts. Very recently, the bias was reported also in the paper of shAIC prediction method (Nielsen et al. 2012). Overall, the correlation offers a sensitive statistic for the quality of the prediction model.

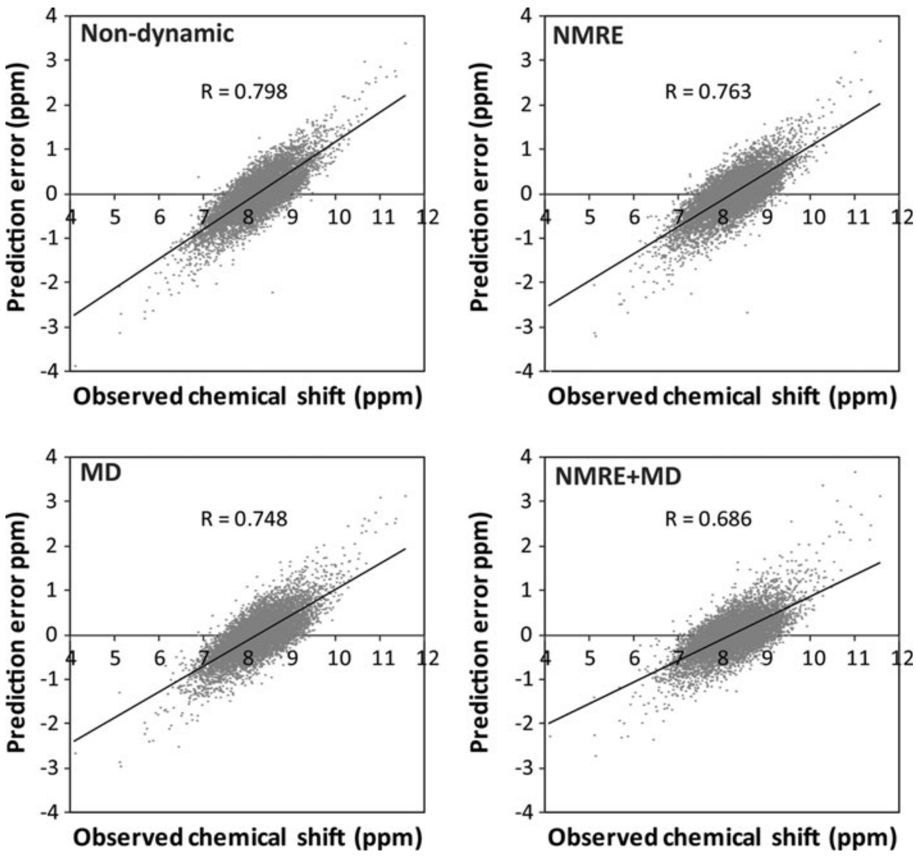

Fig. 4 Correlation of observed chemical shift versus prediction error for ^{1}HN shifts in different prediction models (n = 8,603). The R values are the Pearson correlation coefficients. The RMS errors of the ^{1}HN prediction are 0.49, 0.48, 0.47 and 0.43 ppm for non-dynamic, NMRE, MD and NMRE+MD models, respectively

The freedom caused by the inclusion of 4th dimension could be expected to loosen the NH–aromatic and NH–O=C contacts and to decrease their contribution to the shifts. However, it decreases the correlation (Fig. 4) and improves the prediction (RMS error). Our explanation to this dilemma is that only the real aromatic and hydrogen bonded contacts are preserved in the dynamic protein model while the more or less accidental contacts are loosened. In other words, when the protein models become more realistic, the prediction model is also improved so that the abovementioned contributions are strengthened.

Effect of dynamics on the prediction model

In this study, the benefit from dynamics is expected to arise from two sources: the mapping of conformational space of the query protein and the improvement of the prediction model itself. It has been assumed here that building a dynamic prediction model instead of just averaging the results of single conformations predicted with non-dynamic prediction model would yield better results (see "Methods"). To test this assumption, all snapshots from NMRE+MD ensembles of five proteins were predicted one by one with the non-dynamic model. The results were then simply averaged over the number of conformers (1,000–2,000) of each protein. Using this approach, the conformational space of the query protein is similar to the actual NMRE+MD model, and the effect of the prediction model can be estimated. Table 3 shows the results of the snapshot approach compared with the non-dynamic model and the NMRE+MD model. For each protein individually, the results are shown in Supplementary material Table S6.

On average, the fully dynamic prediction yields 18% lower RMS errors than the snapshot approach. It is also notable that including only query protein dynamics may even impair the results, e.g. in the case of ubiquitin (PDB 1D3Z) which is a rigid and already well defined structure. In addition, as the non-dynamic model is here built with original models not homogenized with the ff99SB force field, uncertainty between the query protein and the model is created. The results with snapshot approach fall mostly between the fully non-dynamic and fully dynamic models (Table S5). For the backbone nuclei ^{1}Hα, ^{1}HN, ^{13}Cα and ^{13}Cβ, 65–80% of the dynamic benefit originates from the improved prediction model, and the rest from the dynamics of the query protein (Table 3). In the case of backbone ^{15}N nuclei, prediction model and query protein dynamics were equally important. In this evaluation, only two proteins with observed ^{13}CO shifts were present. For those, all improvement came from the better defined prediction model. ^{13}CO shifts are extremely sensitive to local structure (Dedios and Oldfield 1994), so adding larger motions to query protein but not to the prediction model slightly impairs the results.

Table 3 Effect of dynamics in prediction model using five proteins for the evaluation

Nuclei	Prediction RMS error (ppm)		
	Non-dynamic	Snapshot approach[a]	NMRE+MD
^{1}Hα	0.28	0.24	0.18
^{1}HN	0.45	0.42	0.35
^{13}Cα	0.97	0.92	0.73
^{13}Cβ	1.36	1.22	0.98
^{13}CO[b]	1.12	1.14	0.93
^{15}N	2.71	2.48	2.23

[a] Conformational space of query proteins modelled as in NMRE+MD model, but predicted with non-dynamic model one snapshot at a time

[b] ^{13}CO shifts were present in only two out of five selected proteins

Effect of force field homogenization

Using teaching data from many different sources leads to uncertainty stemming from different structure determination methods, e.g. force fields. Therefore, it is probable that some of the dynamic benefit reported above is due to the homogenization of the structures with the ff99SB force field (Hornak et al. 2006). To evaluate the significance of this effect, another non-dynamic single conformer model was built so that each protein was geometry optimized by 100 steps using the ff99SB force field. The RMS errors of this model were 0.29, 0.49, 1.11, 1.22, 1.22 and 2.73 ppm for ^{1}Hα, ^{1}HN, ^{13}Cα, ^{13}Cβ, ^{13}CO and backbone ^{15}N chemical shifts, respectively. Compared with the non-dynamic model of raw PDB data, the RMS errors are only 1–2% smaller, thus explaining only a small part of the overall dynamic benefit.

Origins of the dynamic effects on the prediction results

The protein-specific comparison of the dynamic benefits, averaged over the six backbone nuclei types, is presented in Fig. 5. The figure shows, that in almost every case the use of NMRE+MD model improves the prediction result and, as indicated by the error bars, the improvement is uniform among the backbone nuclei. The same data for each backbone nuclei separately can be found from Supplementary material Figure S1. In some cases, remarkable improvements up to 52% are seen. Quite often these improvements originate from one very poorly predicted shift, which is then corrected with the dynamic approach. Next, selected examples how dynamic prediction improves the prediction are described.

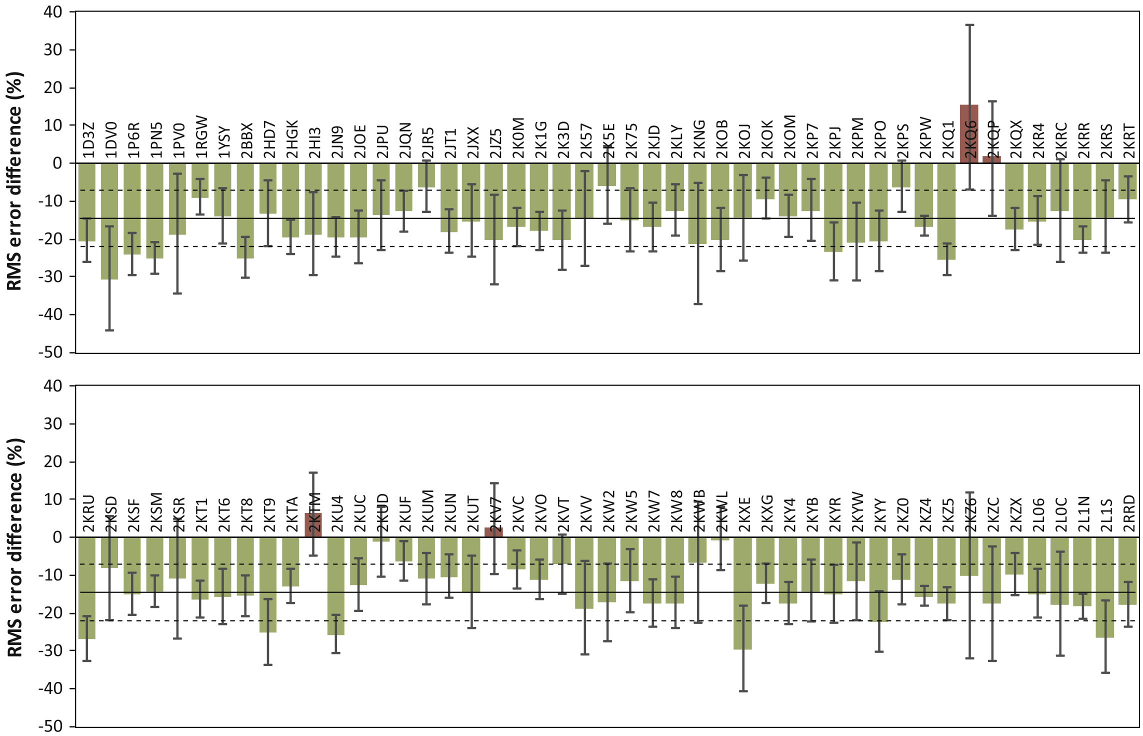

Fig. 5 Protein-specific comparison of non-dynamic and NMRE+MD prediction models. The *bars* indicate average backbone nuclei RMS error changes (in %) from the non-dynamic model (zero level) to the NMRE+MD model. The *error bars* are standard deviations of backbone nuclei RMS error changes. *Solid line* is the average RMS error change and the *broken lines* are standard deviations of the protein RMS error changes

Aromatic ring currents

Aromatic ring currents cause the most intensive effects on ^{1}H shifts (Wishart and Case 2001). In the DNA repair protein HHR23A (PDB 1DV0), the K28Hα nucleus is located close to aromatic side chain rings of Y23 and F24 (Fig. 6a). This leads to a very unusual observed chemical shift of 1.67 ppm, as the random coil shift for lysine ^{1}Hα is 4.32 ppm (Wishart et al. 1995). The non-dynamic prediction for the best conformer of the NMR ensemble for this nucleus is 3.82 ppm. In this conformer the aromatic rings are not close enough and not orthogonally oriented towards the K28Hα nucleus. It is notable that neither the NMRE model (predicted shift 3.49 ppm) nor the MD model (2.89 ppm) alone can map the ring orientations correctly, but with the NMRE+MD model the predicted shift of 2.03 ppm is rather close to the experimental value of 1.67 ppm. Moreover, the observation that result-averaged approach (all 100 ps trajectories predicted with the MD model) also gives an inadequate prediction of 2.86 ppm indicates that the improvement is not coming only from more realistic structure of query protein, but also that the prediction model itself is improved as more realistic aromatic anisotropy model is obtained. Evidently, in non-dynamic models, aromatic anisotropy terms become too weak due to lack of resolution.

Dynamic averaging

Proteins with very flexible regions often achieve great improvements from dynamic averaging. This is confirmed in the work of Markwick et al. (2010), for example, where up to 28% improvements were obtained for flexible ankyrin repeat protein IKBA (PDB 1NFI) with accelerated molecular dynamics as the method to simulate protein motion. Naturally, for small and tightly packed globular proteins such as ubiquitin, the improvements are less notable. This is also seen in our database.

In BH0266 protein (PDB 2KQ1) there is a flexible loop of six residues from Y20 to Y25 (Fig. 6b). In the middle of the loop, there is E23Cα having an unexceptional observed shift of 58.03 ppm, which is poorly predicted (56.21 ppm) using the best representative conformer and the non-dynamic model. Further exploration of the case shows that the non-dynamic predictions of the ensemble conformers vary considerably from each other with a standard deviation of 0.79 ppm, some of them giving precisely the observed shift and some showing prediction errors up to

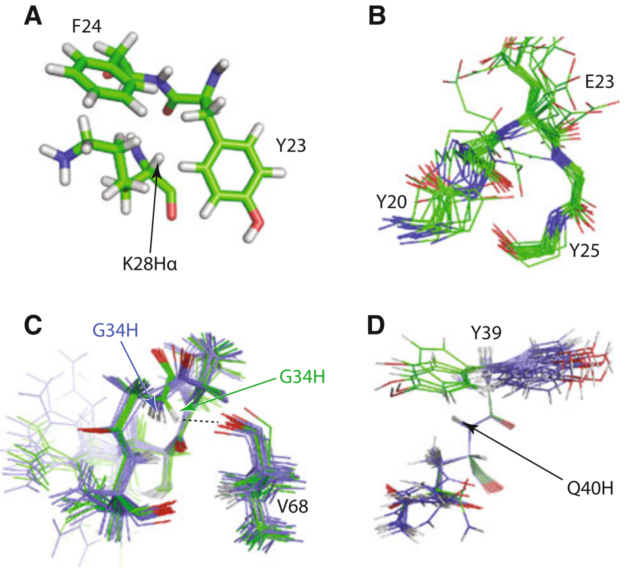

Fig. 6 Selected examples of dynamic effects to the prediction results. **a** Aromatic groups close to K28Hα in the best representative conformer of the NMR ensemble (PDB 1DV0). **b** Alignment of NMR ensemble conformations of the flexible loop region Y20–Y25 (PDB 2KQ1, backbone and residue E23 heavy atoms shown) **c** Alignment of NMR ensemble conformations of β-turn R32–D35 and residue V68 (PDB 2K0M). Hydrogen-bonded conformations shown *green*, others *blue*. **d** Alignment of NMR ensemble conformations of residues Y39 and Q40 (PDB 2JT1). *Gauche+* conformations of Y39 χ^1 angle shown *green*, *trans* conformations *blue*

2.01 ppm. The average predicted shift was 57.3 ppm. Thus, the benefit from the ensemble dynamics is notable, but still leaves the predicted shift rather far from the observed one. When 100 ps MD simulation is applied to the conformers and shifts are predicted with the MD model, the standard deviation of the calculated shifts drops to 0.39 ppm and the average predicted shift becomes 57.81 ppm. This shows how the mapping of local fluctuations smoothes the predictions of this flexible moiety and is almost as essential here as the use of the ensemble dynamics. Finally, when the NMRE+MD model is used, the prediction error drops to 0.02 ppm. Naturally, this final improvement arises from the better prediction model, as the conformational space is the same as in the result-averaging approach.

Erroneous structures and hydrogen bonding

As the NMR structure determination is still based mostly on NOE restraints, it is highly possible that sometimes erroneous folds will enter the published ensembles and even the best representative conformers, because there are not always enough restraints to unambiguously determine all parts of the structures. In the uncharacterized protein from Rhodospillirum rubrum (PDB ID 2K0M), the β–turn of residues R32–D35 is found in two different conformations, having the G34 φ angle either −154° ± 15° (8 conformers) or 155° ± 5° (12 conformers, Fig. 6c). However, the predictions of the G34H nucleus of the static conformers suggest that the former conformer is erroneous, as its average prediction error is 1.66 ± 0.10 ppm, whereas in the latter case the error is 1.02 ± 0.16 ppm. In the former case, a hydrogen bond to V68O, essential to reproduce the upfield observed shift of G34H of 9.47 ppm, cannot be formed. In the latter case, this hydrogen bond is present in 7 out of 8 conformers. Moreover, in 18 out of 20 cases MD simulations convert the fold to a typical βII turn, with average torsion angles of 126° ± 4° and 89° ± 9° for the P33 ψ and G34 φ angles, respectively, indicating strong tension at the original ensemble conformers. This consequently leads to strengthening of the G34H–V68O hydrogen bond and to a decreased average prediction error of 0.59 ppm. Finally, the NMRE+MD model yields the prediction error of 0.34 ppm. This example shows how even short MD simulations can fix erroneous structures and emphasizes the importance of correct hydrogen bond modeling to the prediction, especially for ^{1}HN shifts for which the contribution to the secondary shift is up to 25% (Wishart and Case 2001). In addition, it shows that by using chemical shift information, these kinds of erroneous conformations can be easily found and prevented from entering the final ensemble. The use of back-calculated chemical shifts, among other parameters, has also been suggested before for ensemble conformer selection (Krzeminski et al. 2009) and ensemble assessment (Angyan et al. 2010).

Side chain dynamics

Side chain dynamics often play important role in chemical shift mapping, especially if hydrogen bond donors or acceptors, or aromatic rings are present in the side chains. However, in statistical sense they are less important as the contribution to the backbone secondary shift is estimated to be only 5% (Wishart and Case 2001). Besides the direct effect of the χ_n angles to the shifts of the residue itself, the effects can be seen also in the neighboring residues or other spatially close residues. In the original NMR ensemble of PefI protein (PDB 2JT1) the side chain of Y39 lies in either *trans* (16 out of 20 conformers) or *gauche+* (4 out of 20) conformation (Fig. 6d). Depending on the conformation, the Q40H proton is located inside the aromatic ring current or not. In the *trans* case, the average prediction error of Q40H is −0.52 ± 0.07 ppm while in the *gauche+* case it is 0.04 ± 0.05 ppm, clearly declaring the latter conformation correct. Although the *gauche+* conformation is present only in 4 out of 20 conformers of the ensemble, the MD simulations and dynamic prediction model is capable of decreasing the prediction error down to −0.21 ppm from

−0.60 ppm of the non-dynamic prediction of the best representative conformer (*trans* conformation). The long time-scale side chain dynamics is often achievable from NMR ensembles, but not from the 100 ps MD simulations, where side chain rotations rarely occur.

Impaired prediction results

In 4 out of 94 proteins, the dynamic approach degraded the prediction results (Fig. 5). The largest negative effect was seen on EF-hand domain of polycystin-2 (PDB 2KQ6). In this case the inferior prediction is explained with an erroneous descriptor value caused by an extrapolation problem in N1Cβ. Such problems are rare but are sometimes observed, especially in uncommon structures such as the N-terminal asparagine of this case. Note that the 4DSPOT program warns the user if the predicted shift is over three standard deviations away from BMRB average. Moreover, there is another loop region (D44–T52) with large prediction errors. It is evident that MD simulations cannot always fix flexible regions and may distort them even more, thus leading, as seen in the cases of G47H and T52N, to less accurate predictions. Another example with impaired ^{1}H prediction results is the H2H3 domain of ovine prion protein (PDB 2KTM). It is a rod-shaped protein with a number of aromatic residues and flexible termini, making it very difficult to map the conformational space accurately enough. Furthermore, in the static structure of the acyl carrier protein from Borrelia burgdorferi (PDB 2KWL) there are two large ^{13}Cα shift prediction errors (E37Cα, 3.92 ppm and I61Cα, −3.08 ppm) in the loop regions. When the NMRE+MD model is used, the errors grow larger (5.17 and −4.97 ppm, respectively). The errors are present and the loops are in identical conformations throughout the ensemble. It is therefore evident that the structure must be incorrectly folded or at least too rigid in these parts. It is seen in some cases that existing large errors in non-dynamic structures become even larger after MD if the structure is distorted enough and short MD is unable to fix the structure.

Protein-specific results

The proteins of the teaching database are a cross-section of the recent NMR structures submitted to BMRB, reflecting the present-day level of NMR structure accuracy. In the database structures of varying precision are present, with the lowest RMS errors as low as 0.10 (PDB 2KNG), 0.21 (2KNG), 0.56 (2KRU), 0.47 (2KRU), 0.62 (2KPO) and 1.48 ppm (1PN5) for ^{1}Hα, ^{1}HN, ^{13}Cα, ^{13}Cβ, ^{13}CO and backbone ^{15}N chemical shifts, respectively. The protein-specific prediction results of the NMRE+MD model are shown in Supplementary material Table S5. The three proteins with lowest RMS errors are (1) C-domain of Lsr2 (PDB 2KNG), which is a small and simple protein without error-prone aromatic residues, (2) de novo designed Rossmann 2 × 2 fold protein (PDB 2KPO), which has a compact and rigid structure and (3) PCP_red domain of light-independent protochlorophyllide reductase (PDB 2KRU), which is a small α-helical protein. Common properties of 2KNG and 2KRU are that they are small (55 and 63 aa, respectively) and have flexible termini, allowing the dynamic benefit (21 and 27%, respectively) to be evolved. Moreover, they are both α–helical, which on average yields smaller errors than β-sheets and random coils for ^{1}Hα shifts, as noticed in our previous study. On the other hand, 2KPO is bigger (110 aa) and has a very stable structure without flexible loop regions. Still, 2KPO gets a 20% dynamic improvement, arising from side chain and C-terminus dynamics. As there are no prediction errors larger than 0.8 ppm in backbone ^{1}H shifts and 2.4 ppm in backbone ^{13}C shifts, it is evident that the structure of 2KPO is determined remarkably well, especially as there are four aromatic residues present.

Comparison with other methods

As the approach presented here differs vastly from other prediction methods by the use of dynamics, a direct comparison with the non-dynamic structure prediction would be unfair. However, the RMS errors reported here are quite similar to the state-of-the-art structure-based method SPARTA+ (Shen and Bax 2010), which reports the RMS errors of 0.25, 0.49, 0.94, 1.14, 1.09 and 2.45 ppm for ^{1}Hα, ^{1}HN, ^{13}Cα, ^{13}Cβ, ^{13}CO and backbone ^{15}N chemical shifts, respectively, for 11 X-ray structures predicted with a database of 580 protein models. This suggests that the chemical shifts of proteins can also be effectively predicted without the superior resolution of X-ray structures, as the incorporation of the dynamic information from the side chains and random coils carried by NMR ensembles is able to compensate the less accurate local structures. As the proteins are dynamic structures in NMR sample, this sounds reasonable: a single conformation cannot explain all the chemical shift contributing effects of dynamic origin. Indeed, even atomic resolution X-ray structures (below 1.0 Å) are sometimes predicted with rather large errors (Shen and Bax 2010). Moreover, the differences between X-ray and NMR structures, estimated to be on average 1.4 Å in backbone RMSD (Andrec et al. 2007), are not compromising the prediction results.

Conclusions

In this study, a dynamic model combining NMR ensembles and MD simulations (NMRE+MD) for prediction of protein chemical shifts was introduced. The model achieved 13% lower RMS errors for the backbone nuclei, when compared with the non-dynamic model. This paper also presents the largest *dynamic* teaching database available for protein chemical shift prediction, containing 95,566 chemical shifts and a total of 180.9 ns of simulations of 94 proteins. The inclusion of dynamics in the prediction model was found to have clear advantages compared with approach in which only the dynamics of the query proteins were mapped.

In contrast to previous prediction methods, our prediction model uses only NMR determined protein structures in the teaching database. In our previous work, X-ray structures gave better prediction than the corresponding NMR structures. The more accurate local structures of X-ray structures compensate the fact that they do not contain much information about weakly folded structures and side chain conformations on protein surface. These features were neither described well in our previous teaching database, because we used only single NMR conformations and, therefore, the 4th dimension (conformational space) of NMR structures became only partly described. In this work, we used NMR ensembles to expand the 4th dimension. All our present results clearly show that the NMRE+MD model offers a more realistic representation of the protein structures than the previous models.

In spite of the improvement brought by the present approach, the structure-based prediction of protein chemical shifts, especially ^{1}HN, ^{13}CO and ^{15}N, stays far from satisfactory. We propose that the problem is more in the present protein models, which do not yield an accurate picture about non-bonded interactions and motions in proteins. If the function of proteins, in which the dynamics is expected to have a key role, is wanted to be really understood, the understanding of the interactions would be of essential importance—and chemical shifts offer a tool for this. We suggest that the NMRE+MD approach offers the best 4D model achievable with reasonable efforts.

Software availability

The 4DSPOT software package for Windows or Linux, containing manuals and pre-calculated prediction models for single conformations and dynamic structures, can be downloaded from http://www.uef.fi/4dspot/.

Acknowledgments This work was supported by the The National Doctoral Programme in Informational and Structural Biology (ISB). The Finnish IT Center for Science (CSC) is acknowledged for providing the computational resources.

References

Andrec M, Snyder DA, Zhou Z, Young J, Montelione GT, Levy RM (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. Proteins 69:449–465

Angyan AF, Szappanos B, Perczel A, Gaspari Z (2010) CoNSEnsX: an ensemble view of protein structures and NMR-derived experimental data RID D-9861-2011 RID D-9594-2011. BMC Struct Biol 10:39

Baskaran K, Brunner K, Munte CE, Kalbitzer HR (2010) Mapping of protein structural ensembles by chemical shifts. J Biomol NMR 48:71–83

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Breiman L (2001) Random forests. Mach Learn 45:5–32

Case DA, Darden TA, Cheatham TEI, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, Kollman PA (2008) AMBER 10. University of California, San Francisco

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

Dedios AC, Oldfield E (1994) Chemical-shifts of carbonyl carbons in peptides and proteins. J Am Chem Soc 116:11485–11488

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins 65:712–725

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

Krzeminski M, Fuentes G, Boelens R, Bonvin AM (2009) MINOES: a new approach to select a representative ensemble of structures in NMR studies of (partially) unfolded states. Application to Δ25-PYP. Proteins 74:895–904

Lehtivarjo J, Hassinen T, Korhonen SP, Peräkylä M, Laatikainen R (2009) 4D prediction of protein ^{1}H chemical shifts. J Biomol NMR 45:413–426

Li D, Brueschweiler R (2010) Certification of molecular dynamics trajectories with NMR chemical shifts. J Phys Chem Lett 1:246–248

Liu X, Ren Y, Zhou P, Shang Z (2011) Prediction of protein ^{13}Cα NMR chemical shifts using a combination scheme of statistical modeling and quantum-mechanical analysis. J Mol Struct 995:163–172

Markwick PR, Cervantes CF, Abel BL, Komives EA, Blackledge M, McCammon JA (2010) Enhanced conformational space sampling improves the prediction of chemical shifts in proteins. J Am Chem Soc 132:1220–1221

Moon S, Case DA (2007) A new model for chemical shifts of amide hydrogens in proteins. J Biomol NMR 38:139–150

Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. J Biomol NMR 26:215–240

Nielsen JT, Eghbalnia HR, Nielsen NC (2012) Chemical shift prediction for protein structure calculation and quality assessment using an optimally parameterized force field. Prog Nucl Magn Reson Spectrosc 60:1–28

Parker LL, Houk AR, Jensen JH (2006) Cooperative hydrogen bonding effects are key determinants of backbone amide proton chemical shifts in proteins. J Am Chem Soc 128:9863–9872

Robustelli P, Kohlhoff K, Cavalli A, Vendruscolo M (2010) Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. Structure 18:923–933

Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Structure-based prediction of methyl chemical shifts in proteins. J Biomol NMR 50:331–346

Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J Biomol NMR 38:289–302

Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR 48:13–22

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

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408

Wang L, Markley JL (2009) Empirical correlation between protein backbone ^{15}N and ^{13}C secondary chemical shifts and its application to nitrogen chemical shift re-referencing. J Biomol NMR 44:95–99

Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338:3–34

Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD (1995) ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J Biomol NMR 5:67–81

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36:W496–W502

Xu XP, Case DA (2001) Automated prediction of ^{15}N, ^{13}Cα, ^{13}Cβ and ^{13}C' chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333