J Biomol NMR (2012) 54:257–265
DOI 10.1007/s10858-012-9668-8

ARTICLE

PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles

Da-Wei Li · Rafael Brüschweiler

Received: 19 June 2012 / Accepted: 31 August 2012 / Published online: 13 September 2012
© Springer Science+Business Media B.V. 2012

Abstract The combination of the wide availability of protein backbone and side-chain NMR chemical shifts with advances in understanding of their relationship to protein structure makes these parameters useful for the assessment of structural-dynamic protein models. A new chemical shift predictor (PPM) is introduced, which is solely based on physical–chemical contributions to the chemical shifts for both the protein backbone and methyl-bearing amino-acid side chains. To explicitly account for the effects of protein dynamics on chemical shifts, PPM was directly refined against 100 ns long molecular dynamics (MD) simulations of 35 proteins with known experimental NMR chemical shifts. It is found that the prediction of methyl-proton chemical shifts by PPM from MD ensembles is improved over other methods, while backbone Cα, Cβ, C′, N, and H^N chemical shifts are predicted at an accuracy comparable to the latest generation of chemical shift prediction programs. PPM is particularly suitable for the rapid evaluation of large protein conformational ensembles on their consistency with experimental NMR data and the possible improvement of protein force fields from chemical shifts.

Keywords NMR chemical shift prediction · Side-chain methyl groups · Protein backbone

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9668-8) contains supplementary material, which is available to authorized users.

D.-W. Li · R. Brüschweiler (✉)
Chemical Sciences Laboratory, Department of Chemistry and Biochemistry and National High Magnetic Field Laboratory, Florida State University, Tallahassee, FL 32306, USA
e-mail: bruschweiler@magnet.fsu.edu

Introduction

Chemical shifts represent the most accurate and most ubiquitous NMR information of proteins. For many proteins the resonance assignments, and hence the assigned chemical shifts, are the sole source of NMR information readily available via the BioMagResDataBank (BMRB) (Ulrich et al. 2008), which presently includes chemical shifts of over 5,000 proteins. In spite of their complex dependence on protein structure, significant progress has been made over the years in the prediction of chemical shifts from protein structures using a variety of strategies implemented in software, such as Shifts (Xu and Case 2001, 2002), ShiftX+/ShiftX2 (Neal et al. 2003; Han et al. 2011), SPARTA/SPARTA+ (Shen and Bax 2007, 2010), CamShift (Kohlhoff et al. 2009), CH3shift (Sahakyan et al. 2011) and 4DSPOT (Lehtivarjo et al. 2009, 2012). These programs can be used for the validation and refinement of protein structures or the determination of average protein structures from chemical shifts alone (Cavalli et al. 2007; Shen et al. 2008, 2009; Rosato et al. 2012).

At room temperature the experimental chemical shift of a given nucleus reflects the Boltzmann-weighted average of the 'instantaneous' chemical shifts of a large number of conformational substates that interconvert on the ms timescale or faster. Chemical shift information has been used to extract site-specific order parameters as a measure of local dynamics (Berjanskii and Wishart 2005; Korzhnev et al. 2010). Another application is the quantitative assessment of conformational protein ensembles generated by molecular dynamics (MD) computer simulations for the assessment and comparison of molecular mechanics force fields (Li and Brüschweiler 2010b), the analysis of the performance of enhanced sampling algorithms (Markwick et al. 2010), the comparison of protein structure and

dynamics in solution and in crystals (Robustelli et al. 2012), and the prediction of rotating frame relaxation data (Xue et al. 2012).

The quantum-chemical origin of chemical shifts together with their multifaceted dependence on electronic and structural factors makes their accurate prediction a formidable challenge, especially for macromolecules such as proteins. The traditional approach to chemical shift prediction is through quantum-chemical calculations and approximate analytical relationships based on physical–chemical theory, which form the basis for the Shifts program (Xu and Case 2001, 2002). As a consequence of the rapid expansions of both the BMRB (Ulrich et al. 2008) and PDB (Berman et al. 2000), empirical approaches have been developed in recent years that parametrize chemical shift hyper-surfaces that relate protein structures to experimental chemical shifts (Neal et al. 2003; Shen and Bax 2007, 2010; Kohlhoff et al. 2009; Lehtivarjo et al. 2009, 2012; Sahakyan et al. 2011). While most of the work has focused on chemical shifts of backbone nuclei, amino-acid side-chain chemical shifts are increasingly used for the analysis of protein structure and dynamics. The sensitivity of methyl group chemical shifts on protein structure and their favorable spectroscopic properties even in very large proteins and macromolecular complexes (Ruschak et al. 2010) make their accurate prediction a desirable goal. Several software packages are currently available for the prediction of methyl side-chain ^{1}H chemical shifts (Xu and Case 2001, 2002; Neal et al. 2003; Lehtivarjo et al. 2009; Sahakyan et al. 2011).

Chemical shift prediction programs compute the chemical shift $\delta^{(k)}$ of a nucleus of type k (e.g. Hδ1 of Ile) from a given protein structure represented by the 3 M-dimensional cartesian vector $\mathbf{r}$, where M is the number of protein atoms:an ensemble average of the protein mostly

$$\delta^{(k)}_{predict} = \sum_j a^{(k)}_j f^{(k)}_j(\mathbf{r}) + \delta^{(k)}_0 \quad (1)$$

where $\delta^{(k)}_0$ is a conformation-independent chemical shift offset and the functions $f^{(k)}_j(\mathbf{r})$ describe the geometric dependence of various contributions j to the chemical shift stemming from ring currents, electric fields, hydrogen bonds, magnetic anisotropies, dihedral angles, etc. Sometimes Eq. (1) also comprises additional empirical contributions, including artificial neural network based expressions (Moon and Case 2007; Shen and Bax 2010). The prefactors $a^{(k)}_j$ are typically optimized by minimizing the root-mean-square difference between experimental chemical shifts $\delta^{(k)}_{exp}$ calculated from average X-ray crystal structures $\mathbf{r}_{\mathrm{Xray}}$ by a linear-least squares fit:

$$\delta^{(k)}_{exp} = \sum_j a^{(k)}_{j,Xray} f^{(k)}_j(\mathbf{r}_{Xray}) + \delta^{(k)}_0 \quad (2)$$

While X-ray crystal coordinates $\mathbf{r}_{\mathrm{Xray}}$ represent the average structure of the protein in a crystalline environment, sometimes at cryogenic temperature, the $\delta^{(k)}_{exp}$ correspond to an ensemble average of the protein in solution at ambient temperatures. Hence, in order to bridge these two different conditions the fitted $a^{(k)}_{j,Xray}$ parameters inherently include a certain amount of motional averaging. This generally leads to a reduction of the (absolute) values of $a^{(k)}_{j,Xray}$. When using the $a^{(k)}_{j,Xray}$ parameters for the prediction of chemical shifts by Eq. (1) from an average protein structure, the motional averaging encoded in the $a^{(k)}_{j,Xray}$ parameters helps improve prediction accuracy.

By contrast, when calculating chemical shifts from a conformational ensemble, such as a MD trajectory, the use of the 'pre-averaged' $a^{(k)}_{j,Xray}$ values is problematic because dynamic averaging effects would be counted twice: once in $a^{(k)}_{j,Xray}$ and once by the explicit representation of the dynamics in the form of multiple conformers. This inconsistency can be resolved, in principle, by the use of 'static' $a^{(k)}_j$ parameters when back-calculating chemical shifts from conformational ensembles. The static $a^{(k)}_j$ parameters can be determined from quantum-chemical chemical shift calculations of static protein fragments as originally used in the Shifts program. Alternatively, the $a^{(k)}_j$ parameters can be fitted against an entire conformational ensemble:

$$\delta^{(k)}_{exp} = \sum_j a^{(k)}_{j,MD} \left\langle f^{(k)}_j(\mathbf{r}_n) \right\rangle_{MD} + \delta^{(k)}_0 \quad (3)$$

where the angular brackets indicate averaging over a canonical ensemble represented by snapshots $\mathbf{r}_n$ (n = 1, …,N). It should be noted that since $\left\langle f^{(k)}_j(\mathbf{r}_n) \right\rangle_{MD} \neq f^{(k)}_j(\langle \mathbf{r}_n \rangle_{MD})$ the fitting parameters obtained from Eqs. (2) and (3) are not equivalent, $a^{(k)}_{j,MD} \neq a^{(k)}_{j,Xray}$, even if the *average* protein structure during the MD simulation is identical to the X-ray crystal structure $\langle \mathbf{r}_n \rangle_{MD} = \mathbf{r}_{Xray}$. This situation is fully analogous to the parametrization of other NMR observables, such as scalar ^{3}J-couplings via Karplus relationships and the interpretation of residual dipolar couplings where the parametrization against average protein structures using experimental data collected under physiological conditions leads to the absorbtion of dynamic properties by the parametrization constants (Brüschweiler and Case 1994; Meiler et al. 2001; Lindorff-Larsen et al. 2005; Vogeli et al. 2007; Markwick et al. 2009).

The strategy underlying Eq. (3) was used by Lehtivarjo et al. (2009) for the parametrization of ^{1}H chemical shifts by performing 150 ps–1 ns MD simulations with the resulting chemical shift predictor implemented in the 4DSPOT software. This work was subsequently extended for the prediction of other backbone chemical shifts from NMR ensembles by performing short 100 ps MD trajectories of each NMR structural model (Lehtivarjo et al. 2012).

Recent advances in computer hardware have made molecular dynamics trajectories into the submicrosecond range a routine task (Klepeis et al. 2009). Moreover, the development of protein force fields, such as AMBER ff99SB (Hornak et al. 2006), AMBER ff03 (Duan et al. 2003), and CHARMM CMAP (Buck et al. 2006), stimulated the wider use of experimental NMR data for the quantitative certification of MD simulations. These include a wide range of solution NMR data of proteins and peptides, such as residual dipolar couplings (RDCs) (Showalter et al. 2007; Lange et al. 2010; Long et al. 2011), scalar J-couplings (Markwick et al. 2009; Wickstrom et al. 2009; Lange et al. 2010), spin relaxation order parameters (Markwick et al. 2007; Showalter and Brüschweiler 2007; Trbovic et al. 2008) and chemical shifts (Li and Brüschweiler 2010b; Markwick et al. 2010; Robustelli et al. 2012). Furthermore, chemical shift data can be used for the direct improvement of molecular mechanics force fields of proteins (Li and Brüschweiler 2010a, 2011).

The primary goal of the present work is the development of a methyl-side chain and protein backbone atom chemical shift predictor based on Eq. (3). This predictor, termed PPM, is specifically designed for the validation of large conformational ensembles and the improvement of protein force fields from chemical shifts. Hence, both computational efficiency and accuracy are critical. For this purpose we have assembled a library of 35 different proteins with known backbone and side-chain chemical shift assignments and performed 100 ns molecular dynamics simulations for each protein starting from medium-to-high resolution X-ray crystal structures.

Methods

The PDB and BMRB codes of the 35 different proteins used in this work are listed in Table S1 of the Supporting Information. Some of these proteins were used in our previous studies (Li and Brüschweiler 2010a, 2011) or taken from the RefDB database (Zhang et al. 2003). All these proteins share the availability of (i) methyl-proton chemical shifts and (ii) a PDB structure solved by X-ray crystallography with a resolution better than 2.0 Å, with one exception, 2RNJ, which was solved by NMR. All protonation states correspond to pH 7. All MD simulations were performed using the Gromacs 4 program (Berendsen et al. 1995; Lindahl et al. 2001; van der Spoel et al. 2005; Hess et al. 2008). Water molecules were included explicitly using the TIP3P model (Jorgensen et al. 1983). All bond lengths involving hydrogen atoms were constrained by the SETTLE algorithm and a 2 fs time step was used. All van der Waals interactions were cut off at 10 Å and electrostatic interactions were cut off at 8 Å. The long-range electrostatic interactions were calculated using the PME algorithm with 1.2 Å spacing. The final production runs were at constant temperature and pressure (NPT ensemble) of 300 K and 1 atm, respectively. All MD simulations were run for 100 ns and coordinates were saved every 100 ps, which yielded MD ensembles consisting of 1000 conformers for each protein. All MD simulations were stable with an average root-mean-square-deviation (RMSD) between the MD ensembles and their X-ray crystal structures of 1.50 Å.

The recently developed protein force field ff99SB_$\varphi\psi$(g24;CS) (Li and Brüschweiler 2011) was used for all simulations, in which the backbone φ, ψ dihedral angle potentials had been improved by the inclusion of cross terms through the addition of 24 bivariate Gaussian potential functions of variable depth, width, and tilt angle. This force field, which reproduces a wide range of experimental NMR parameters of full-length proteins, was further enhanced by the addition of dihedral angle corrections of ILDN side chains (Lindorff-Larsen et al. 2010). The combination of the NMR-optimized force field ff99SBnmr1 (Li and Brüschweiler 2010a) with the ILDN correction has been shown recently to reproduce experimental NMR parameters with remarkably high accuracy (Long et al. 2011; Beauchamp et al. 2012). A recent comparison of the performance of 12 of the latest protein force fields (Beauchamp et al. 2012) showed that for a benchmark set of 524 protein NMR parameters the ff99SBnmr1-ILDN force field (Li and Brüschweiler 2010a; Lindorff-Larsen et al. 2010; Long et al. 2011) has the highest accuracy among the force fields tested. Since the backbone potential of ff99SB_$\varphi\psi$(g24;CS) was improved over ff99SBnmr1, ff99SB_$\varphi\psi$(g24;CS)+ILDN is expected to be at least as accurate as ff99SBnmr1+ILDN.

The chemical shift prediction is based on Eq. (3) where index k denotes each type of chemical shift that was parameterized. This includes 9 methyl side-chain proton chemical shifts, namely Hβ of Ala, Hε of Met, Hγ2 of Thr, Hγ1 and Hγ2 of Val, Hδ1 and Hδ2 or Leu, Hγ2 and Hδ of Ile, and the 5 backbone chemical shifts of the Cα, Cβ, C$'$, N, and H^N atoms of all amino acids except Cys. The angular brackets in Eq. (3) refer to the ensemble average over the 1000 MD snapshots per trajectory. All parameters were determined by a linear least-squares regression over

all 35 proteins. Besides amino-acid type specific chemical shift offset values, the following descriptors were included: ring current effects, magnetic anisotropy effects, electric field effects, dihedral angle effects, hydrogen bond effects, and amino-acid sequence effects, such as the characteristic effect of a proline following the residue of interest (see Supporting Information for details).

We found that for the methyl proton side-chain chemical shifts only the first 3 terms, which are described below in more detail, contribute significantly. Hence, their parametrization was based on:

$$\delta_{exp}^{(k)} = \delta_{ring-current}^{(k)} + \delta_{magn-aniso}^{(k)} + \delta_{0}^{(k)} \quad (4)$$

using least-squares fitting to determine $\delta_0^{(k)}$ and the prefactors $a_{j,MD}^{(k)}$ of the terms for $\delta_{ring-current}^{(k)}$ and $\delta_{magn-aniso}^{(k)}$ according to Eq. (3). For both the ring current and magnetic anisotropy contributions, the prefactors $a_{j,MD}^{(k)}$ do not depend on the type of the nucleus k. To monitor and prevent overfitting, 20 % of the data points were randomly excluded during fitting and only used for validation. This process was repeated 1,000 times and the average RMS error of the fitting set and validation set were compared (see also Table 1). For the parametrization of backbone chemical shifts additional terms were included (see Supporting Information).

Ring current effects $\delta_{ring-current}^{(k)}$ can arise from 5 different aromatic amino acid rings, namely the ones in Phe, Tyr, His, and the 5-ring and 6-ring of Trp (Trp-5, Trp-6). This geometric descriptor is defined as (Haigh and Mallion 1972, 1979; Osapay and Case 1991; Sahakyan et al. 2011)

$$f_{ring-\ current} = \sum_{p,q} S_{pq}\left(\frac{1}{r_p^3} + \frac{1}{r_q^3}\right) \quad (5)$$

where the sum includes all adjacent atom pairs in the ring. r_p and r_q are the distances between neighboring ring atoms p and q to the proton and S_{ij} is the area of the triangle formed by atom i, atom j and the projection of the proton on the aromatic ring (denoted as "o"). The sign of S_{ij} is determined whether the vector product $\mathbf{t}_{oi} \times \mathbf{t}_{ij}$ is parallel (positive sign) or antiparallel (negative sign) to the ring normal defined by $\mathbf{t}_{12} \times \mathbf{t}_{23}$, where $\mathbf{t}_{oi}$ is the vector pointing from o to atom i and and $\mathbf{t}_{ij}$ as the vector pointing from atom i to atom j.

Magnetic anisotropy effects $\delta_{magn-aniso}^{(k)}$ were calculated using the axially symmetry model by Case (Osapay and Case 1991) following McConnell's formulation of anisotropy effects of peptide groups (McConnell 1957):

$$f_{magn-aniso} = \frac{1}{r^3}\left(3\cos^2\theta - 1\right) \quad (6)$$

where r is the distance from the proton to the peptide amide group (containing the OC′N backbone atoms) and θ is the angle between the vector joining the proton to the amide group and the amide group normal. Following Sahakyan et al., an analogous treatment was employed for the OCN side-chain groups of residues Asn and Gln, for the OCO side-chain groups of Glu and Asp, and for the NCN side-chain group of Arg.

Results and discussion

We tested the importance of the different terms in Eq. (3) on the methyl-proton chemical shifts by systematically excluding individual terms. It turned out that the dihedral angle terms had no effect on the prediction accuracy. Although in principle electric field effects can have a sizeable effect on proton chemical shifts, inclusion of this term in the parametrization did not improve the performance either. The combination of ring current and magnetic anisotropic effects [Eq. (4)] provided the best methyl-

Table 1 RMSDs (in units of ppm) of methyl-proton chemical shift prediction (values in parentheses are explained in the main text)

Software input	PPM		CH3Shift		Shifts		4DSPOT[a] MD
	PDB	MD	PDB	MD	PDB	MD	
Ala Hβ	0.17	0.15 (0.14)	0.21 (0.18)	0.21	0.18	0.17 (0.16)	0.18
Val Hγ1	0.20	0.13 (0.13)	0.19 (0.17)	0.18	0.19	0.15 (0.15)	0.17
Val Hγ2	0.18	0.15 (0.14)	0.17 (0.15)	0.18	0.20	0.17 (0.15)	0.18
Leu Hδ1	0.23	0.22 (0.13)	0.30 (0.16)	0.35	0.22	0.22 (0.13)	0.24
Leu Hδ2	0.18	0.16 (0.13)	0.23 (0.18)	0.23	0.23	0.17 (0.14)	0.19
Ile Hγ2	0.21	0.19 (0.17)	0.23 (0.22)	0.23	0.24	0.22 (0.20)	0.19
Ile Hδ	0.22	0.22 (0.17)	0.23 (0.17)	0.25	0.23	0.22 (0.18)	0.24
Thr Hγ2	0.14	0.13 (0.12)	0.15 (0.13)	0.16	0.18	0.15 (0.14)	0.16
Met Hε	0.28	0.19	NA	NA	0.26	0.20	0.25

[a] Using the "MD" parameter set

proton chemical shift prediction. The prefactors $a^{(k)}_{ring-current,MD}$ for the ring current effects from Phe, Tyr, His, Trp-5 and Trp-6 are 5.817, 4.887, 4.860, 5.566, and 5.961 ppm Å, respectively. These parameters are similar to previous work (Osapay and Case 1991) (5.455, 4.582, 4.910, 5.673, and 5.564 ppm Å, respectively). The prefactors $a^{(k)}_{magn-aniso,MD}$ for the magnetic anisotropic effect of backbone OCN groups, side-chain OCN groups of Asn and Gln, OCO groups of Glu and Asp, NCN group of Arg are −4.479, −0.926, −1.692, and −0.408 ppm Å^3, respectively. The chemical shift offsets of Hβ of Ala, Hγ2 of Thr, Hε of Met, Hγ1 and Hγ2 of Val, Hδ1 and Hδ2 or Leu, Hγ2 and Hδ of Ile are 1.3876, 1.2671, 2.0763, 1.0064, 0.9611, 0.9102, 0.8890, 0.9976, and 0.8584 ppm, respectively). However, as pointed out previously (Osapay and Case 1991), the backbone peptide group will also contribute to the chemical shift offset in a way that makes them hard to separate. Therefore, the fitted prefactor $a^{(k)}_{magn-aniso,MD}$ for the magnetic anisotropic effect and chemical shift offsets cannot be directly compared with theoretical estimates. The previous work by Osapay and Case resulted in a prefactor for the backbone OCN groups of −4.37 ppm Å^3, which is similar to the one in this work (note that the equation and the units in the original work by Osapay and Case slightly differ from the ones used here). The magnetic anisotropic effects from side chain groups, which have a larger fitting uncertainty, have only a minor effect on the overall fit quality. Since the number of fitting parameters (18) is small compared to the number of data points (1,544) used in this work, the fits were very stable, i.e. overfitting was not an issue. In Table 1, the prediction accuracy for the 9 types of methyl-proton group chemical shifts are listed for PPM, in comparison with the results obtained from the chemical shift prediction programs CH3Shift, Shifts, and 4DSPOT (for 4DSPOT, the "MD" parameter set was selected with other parameter sets affecting the accuracy only minimally). Correlation coefficients and slopes of the linear regression analysis for PPM are also provided in Table 2. Comparison between prediction and experiment is depicted in Fig. 1. CH3Shift was parametrized using only experimental chemical shifts that deviated from the average by less than 2.5 standard deviations, citing possible experimental errors for the chemical shifts that were excluded in this way (Sahakyan et al. 2011). While PPM was parametrized including all chemical shifts, for the comparison with CH3Shift we used the same subset (the comparison for all chemical shifts is given in parentheses in Table 1). Since 4DSPOT was parameterized based on conformational ensembles, only ensemble-based predictions are reported. Predictions based on both static PDB structures and MD ensembles are reported for PPM, Shifts, and CH3Shift. Consistent with our previous observation (Li and Brüschweiler 2010b), the prediction accuracy of Shifts improved significantly when ensemble averaging was employed instead of an average PDB structure. On the other hand, for CH3Shift, which is based on a knowledge-based parameterization against static PDB structures, MD-ensemble averaging showed no improvement.

Overall, PPM provides better agreement with experiment than any of the other programs tested. Despite the fact that PPM uses only a subset of the structural descriptors used in CH3Shift, it achieved better prediction accuracy for all 9 methyl proton sites (note that CH3Shift did not report chemical shifts of Met Hε). This fact highlights the importance of the inclusion of realistic protein dynamics, as provided by the 100 ns MD trajectories, for the parametrization of chemical shifts. Despite the improvements of PPM over other predictors, the achieved chemical shift RMSDs still significantly exceed the experimental errors, leaving room for further improvement. Chemical shifts are inherent quantum-chemical quantities and hence sensitive to electronic effects. However, parameterization of chemical shifts solely as a function of atomic positions cannot do full justice to effects arising from the delocalized nature of the electron density. Hence, it seems likely that further improvement of the chemical shift prediction accuracy will require the explicit inclusion of the quantum-chemical nature of the electronic effects. However, this is not a straightforward task as the accuracy of quantum-

Table 2 Linear regression results of PPM

Methyl site	RMSD[a]	Slope[b]	Intercept[b]	Correlation coefficient[b]
Ala Hβ	0.15	0.69	0.42	0.85
Val Hγ1	0.13	0.66	0.27	0.82
Val Hγ2	0.15	0.57	0.33	0.77
Leu Hδ1	0.22	0.61	0.28	0.79
Leu Hδ2	0.16	0.76	0.16	0.87
Ile Hγ2	0.19	0.65	0.26	0.68
Ile Hδ	0.22	0.58	0.27	0.78
Thr Hγ2	0.13	0.70	0.35	0.85
Met Hε	0.19	0.67	0.59	0.91

[a] Chemical shift RMSDs of PPM when applied to 100 ns trajectories (3rd column of Table 1)

[b] Linear regression results expressed in terms of slope (predicted vs. experimental chemical shifts), intercept, and Pearson correlation coefficient

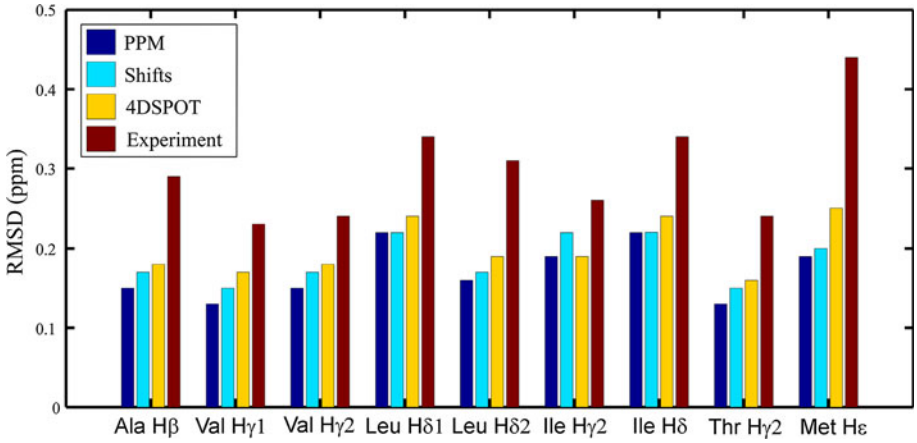

Fig. 1 Performance of the prediction of methyl-proton chemical shifts by PPM (*blue bars*) compared with other programs. The methyl chemical shifts of 35 proteins were used (see text). The *red bars* indicate the standard deviations of the experimental chemical shifts for reference

chemistry based chemical shift calculations of proteins reported for several backbone nuclei was below the one of the empirical methods (Vila et al. 2009).

While PPM and 4DSPOT both use conformational ensembles instead of an average static structure for chemical shift parameterization and prediction, they follow a different philosophy. In 4DSPOT parameterization was done based on either an ensemble generated by a very short (100 ps) MD run, where MD helps to locally relax the initial structure and explore the vibrational motions around it, or an NMR ensemble, where all conformers are treated with equal weights (Lehtivarjo et al. 2009, 2012). However, an NMR ensemble generally does not represent a Boltzmann-weighted ensemble and conformational heterogeneity in some parts of a protein does not necessarily reflect conformational flexibility, as it can also be caused by the lack of NMR constraints. By contrast, PPM is based on the assumption that current molecular dynamics simulations are increasingly realistic in describing protein dynamics in solution, including methyl side-chain dynamics (Showalter et al. 2007; Long et al. 2011). On the one hand, slower timescales sampled by the 100 ns trajectories, probing a larger portion of conformational space, can be used to better parametrize chemical shifts. This effect can be seen when limiting the MD trajectories to shorter durations. For example, the RMSDs of the chemical shift predictions increase on average by 3 % when the MD trajectories for training and prediction are limited to only 10 ns lengths. On the other hand, discrepancies between back-calculated and experimental chemical shifts of increasingly long MD trajectories will also help uncover deficiencies in the force fields themselves. Hence, the approach taken here can be considered as the starting point of a strategy with goal to iteratively improve the prediction of chemical shifts, or other biophysical observables, based on MD trajectories, which in turn help optimize molecular mechanics force fields until a self-consistent predictor and force field are obtained. It remains to be seen, however, whether and how rapidly such an approach will converge.

We extended the application range of PPM to the prediction of backbone Cα, Cβ, C′, N and H^N atoms using the same 35 proteins used for the methyl chemical shifts plus another 12 proteins (12 bottom entries in Table S1). Only physical–chemical contributions to the chemical shift were included (see SI for more information). The RMS errors calculated from both the fitting set and validation set are listed in Table 2. During fitting, only 80 % of randomly selected chemical shift data were used, while the remaining 20 % of the data points were used for validation only. This process was repeated 1000 times with the mean values reported in Table 3.

To evaluate the PPM approach and compare with other software, RMS errors obtained for PPM, SPARTA+, ShiftX, Shifts, CamShift and 4DSPOT are listed in Table S2 for 9 proteins, which were previously used as a validation set for SPARTA+. (Two proteins from the original list are excluded because no medium-to-high resolution X-ray crystal structures were available.) As for the side-chain chemical shifts, we find that MD ensemble averaging of the predictions by Shifts improves the agreement while the same procedure does not help in the case of SPARTA+, ShiftX, and CamShift. Because PPM was parameterized directly against extended MD ensembles, it

Table 3 RMSDs (in units of ppm) of backbone chemical shift prediction

Software	PPM fitting	PPM validation
Input	MD	MD
Cα	1.00	1.06
Cβ	1.16	1.23
C′	1.21	1.32
H^N	0.49	0.53
N	2.75	2.91

is not surprising that its application to static X-ray crystal structures produces RMS errors that are larger than for some of the other programs.

Overall, PPM predicts chemical shifts from conformational ensembles with comparable or better accuracy than other programs, except for SPARTA+. SPARTA+ was trained against a significantly larger set of proteins (580) than the one used here, as it only uses the X-ray crystal structures as input. This allowed the inclusion of next-neighbor amino acid effects (amino acid triples) for improved prediction accuracy. PPM, on the other hand, requires an extended MD trajectory, which naturally limits the number of proteins used for parametrization. Furthermore, SPARTA+ uses a significantly more sophisticated prediction function than PPM. Besides physical–chemical terms, which also includes the contact model to predict local dynamics from the X-ray crystal structure (Zhang and Brüschweiler 2002), SPARTA+ uses a neural network. In this way, SPARTA+ captures certain motional effects without the need of an explicit MD ensemble. This is beneficial for the prediction of chemical shifts from an average structure, but it does not translate into an improved prediction when averaging over a MD ensemble (Table S2).

Due to the simplicity of the expressions used in Eqs. (4)–(6) the computational efficiency of PPM is high. The chemical shift prediction of all backbone and methyl protons for a protein with 100 amino acids from a trajectory of 1,000 snapshots takes only about 30 s on a single processor machine using one core.

Concluding remarks

Why is there a need for another chemical shift prediction program? First, for the proteins tested here, PPM predicts side-chain proton methyl chemical shifts with better accuracy than other current programs (Fig. 1). Second, experimental chemical shifts represent averages over large conformational ensembles, which makes them important probes for the assessment of the quality of canonical protein ensembles generated in silico, such as the ones by explicit-solvent MD simulations. The model for the dependence of the chemical shifts on protein coordinates underlying PPM is solely based on physical–chemical effects parametrized directly against 100 ns MD ensembles of 35 different proteins. In this way, the effect of dynamic averaging during parametrization is limited to the prefactors of the different terms of Eq. (3). This allows the direct assessment of the effect of chemical shift averaging over an explicit conformational ensemble. This is in contrast to machine-learning based predictors where the effect of ensemble averaging can be hidden. Not surprisingly, PPM does not predict backbone chemical shifts from single protein structures as accurately as the latest generation of chemical shift prediction programs. However, when provided with an extended ensemble of conformations, the performance of PPM becomes comparable or better than for other programs (Table S2), with the exception of SPARTA+, which performs better for individual X-ray crystal structures than MD ensembles. As PPM uses analytical physical-chemistry based expressions for the various chemical shift contributions, the generally large drop in the chemical shift RMSD when going from a single structure to an MD ensemble highlights the significance of protein dynamics on chemical shift averaging.

PPM should prove particularly useful for the routine assessment of the quality of Boltzmann-weighted protein ensembles at atomic detail, especially ones generated by long MD simulations, and to improve molecular mechanics force fields from experimental NMR chemical shift data of full-length proteins. As the length of MD trajectories continues to grow, the resulting ensembles will help further optimize chemical shift predictors, such as PPM, and, in turn, the comparison of predicted with experimental shifts should guide the further improvement of molecular mechanics force fields.

Web server availability

The PPM web server is available on http://spin.magnet.fsu.edu. The PPM program is available upon request.

Acknowledgments We thank Art Palmer for stimulating discussions. This work was supported by grant MCB-0918362 of the National Science Foundation.

References

Beauchamp KA, Lin YS, Das R, Pande VS (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. J Chem Theory Comput 8:1409–1414

Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS—a message-passing parallel molecular-dynamics implementation. Comput Phys Commun 91:43–56

Berjanskii MV, Wishart DS (2005) A simple method to predict protein flexibility using secondary chemical shifts. J Am Chem Soc 127:14970–14971

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

Brüschweiler R, Case DA (1994) Adding harmonic motion to the Karplus relation for spin–spin coupling. J Am Chem Soc 116:11199–11200

Buck M, Bouguet-Bonnet S, Pastor RW, MacKerell AD (2006) Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme. Biophys J 90:L36–L38

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

Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012

Haigh CW, Mallion RB (1972) New tables of ring current shielding in proton magnetic-resonance. Org Magn Res 4:203–228

Haigh CW, Mallion RB (1979) Ring current theories in nuclear magnetic-resonance. Prog NMR Spectrosc 13:303–344

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

Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4:435–447

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

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential function for simulating liquid water. J Chem Phys 79:926–935

Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE (2009) Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol 19:120–127

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

Korzhnev DM, Religa TL, Banachewicz W, Fersht AR, Kay LE (2010) A transient and low-populated protein-folding intermediate at atomic resolution. Science 329:1312–1316

Lange OF, van der Spoel D, de Groot BL (2010) Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. Biophys J 99:647–655

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

Lehtivarjo J, Tuppurainen K, Hassinen T, Laatikainen R, Perakyla M (2012) Combining NMR ensembles and molecular dynamics simulations provides more realistic models of protein structures in solution and leads to better chemical shift prediction. J Biomol NMR 52:257–267

Li DW, Brüschweiler R (2010a) NMR-based protein potentials. Angew Chem 49:6778–6780

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

Li DW, Brüschweiler R (2011) Iterative optimization of molecular mechanics force fields from NMR data of full-length proteins. J Chem Theory Comput 7:1773–1782

Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7:306–317

Lindorff-Larsen K, Best RB, Vendruscolo M (2005) Interpreting dynamically-averaged scalar couplings in proteins. J Biomol NMR 32:273–280

Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78:1950–1958

Long D, Li DW, Walter KFA, Griesinger C, Brüschweiler R (2011) Toward a predictive understanding of slow methyl group dynamics in proteins. Biophys J 101:910–915

Markwick PRL, Bouvignies G, Blackledge M (2007) Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. J Am Chem Soc 129:4724–4730

Markwick PR, Showalter SA, Bouvignies G, Brüschweiler R, Blackledge M (2009) Structural dynamics of protein backbone phi angles: extended molecular dynamics simulations versus experimental (3) J scalar couplings. J Biomol NMR 45:17–21

Markwick PRL, 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

McConnell HM (1957) Theory of nuclear magnetic shielding in molecules. I. Long-range dipolar shielding of protons. J Chem Phys 27:226–229

Meiler J, Prompers JJ, Peti W, Griesinger C, Brüschweiler R (2001) Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins. J Am Chem Soc 123:6098–6107

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 HY, Wishart DS (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. J Biomol NMR 26:215–240

Osapay K, Case DA (1991) A new analysis of proton chemical-shifts in proteins. J Am Chem Soc 113:9436–9444

Robustelli P, Stafford KA, Palmer AG (2012) Interpreting protein structural dynamics from NMR chemical shifts. J Am Chem Soc 134:6365–6374

Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Gutmanas A, Guntert P, He YF, Herrmann T, Huang YPJ, Jaravine V, Jonker HRA, Kennedy MA, Lange OF, Liu GH, Malliavin TE, Mani R, Mao BC, Montelione GT, Nilges M, Rossi P, van dS, G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, de V, S, Vuister GW, Wu B, Yang YH, Bonvin AMJJ (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20:227–236

Ruschak AM, Religa TL, Breuer S, Witt S, Kay LE (2010) The proteasome antechamber maintains substrates in an unfolded state. Nature 467:868–871

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 GH, Eletsky A, Wu YB, 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

Shen Y, Vernon R, Baker D, Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43:63–78

Showalter SA, Brüschweiler R (2007) Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: application to the AMBER99SB force field. J Chem Theory Comput 3:961–975

Showalter SA, Johnson E, Rance M, Brüschweiler R (2007) Toward quantitative interpretation of methyl side-chain dynamics from NMR by molecular dynamics simulations. J Am Chem Soc 129:14146–14147

Trbovic N, Kim B, Friesner RA, Palmer AG (2008) Structural analysis of protein dynamics by MD simulations and NMR spin-relaxation. Proteins 71:684–694

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

Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718

Vila JA, Arnautova YA, Martin OA, Scheraga HA (2009) Quantum-mechanics-derived C-13(alpha) chemical shift server (CheShift) for protein structure validation. Proc Natl Acad Sci USA 106:16972–16977

Vogeli B, Ying JF, Grishaev A, Bax A (2007) Limits on variations in protein backbone dynamics from precise measurements of scalar couplings. J Am Chem Soc 129:9377–9385

Wickstrom L, Okur A, Simmerling C (2009) Evaluating the performance of the ff99SB force field based on NMR scalar coupling data. Biophys J 97:853–856

Xu XP, Case DA (2001) Automated prediction of 15N, 13Calpha, 13Cbeta and 13C′ chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333

Xu XP, Case DA (2002) Probing multiple effects on 15N, 13C alpha, 13C beta, and 13C′ chemical shifts in peptides using density functional theory. Biopolymers 65:408–423

Xue Y, Ward JM, Yuwen TR, Podkorytov IS, Skrynnikov NR (2012) Microsecond time-scale conformational exchange in proteins: using long molecular dynamics trajectory to simulate NMR relaxation dispersion data. J Am Chem Soc 134:2555–2562

Zhang F, Brüschweiler R (2002) Contact model for the prediction of NMR N-H order parameters in globular proteins. J Am Chem Soc 124:12654–12655

Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195