J Biomol NMR (2011) 50:331–346
DOI 10.1007/s10858-011-9524-2

ARTICLE

Structure-based prediction of methyl chemical shifts in proteins

Aleksandr B. Sahakyan · Wim F. Vranken · Andrea Cavalli · Michele Vendruscolo

Received: 18 March 2011 / Accepted: 17 May 2011 / Published online: 12 July 2011

Abstract Protein methyl groups have recently been the subject of much attention in NMR spectroscopy because of the opportunities that they provide to obtain information about the structure and dynamics of proteins and protein complexes. With the advent of selective labeling schemes, methyl groups are particularly interesting in the context of chemical shift based protein structure determination, an approach that to date has exploited primarily the mapping between protein structures and backbone chemical shifts. In order to extend the scope of chemical shifts for structure determination, we present here the *CH3Shift* method of performing structure-based predictions of methyl chemical shifts. The terms considered in the predictions take account of ring current, magnetic anisotropy, electric field, rotameric type, and dihedral angle effects, which are considered in conjunction with polynomial functions of interatomic distances. We show that the *CH3Shift* method achieves an accuracy in the predictions that ranges from 0.133 to 0.198 ppm for ^{1}H chemical shifts for Ala, Thr, Val, Leu and Ile methyl groups. We illustrate the use of the method by assessing the accuracy of side-chain structures in structural ensembles representing the dynamics of proteins.

Keywords Protein side-chains · Methyl groups · Chemical shift prediction · Random coil

Electronic supplementary material The online version of this article (doi:10.1007/s10858-011-9524-2) contains supplementary material, which is available to authorized users.

A. B. Sahakyan · A. Cavalli · M. Vendruscolo (✉)
Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
e-mail: mv245@cam.ac.uk

W. F. Vranken
European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

Present Address:
W. F. Vranken
Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

Introduction

Despite the fact that chemical shifts are the most readily and accurately measurable observables in protein NMR spectroscopy, their complex dependence on a myriad of molecular and environmental factors (Oldfield 1995; Jameson 1996) has represented a major obstacle for their direct use in protein structure determination. Recent advances in experimental and computational techniques, however, are starting to make it possible to use them to obtain structures of proteins (Cavalli et al. 2007; Shen and et al. 2008; Raman et al. 2010; Korzhnev et al. 2010) and protein complexes (Montalvao et al. 2008; Das et al. 2009), both in solution and in the solid states (Robustelli et al. 2008; Shen et al. 2009). As the protocols that have been introduced so far for using chemical shifts in structure determination (Cavalli et al. 2007; Shen and et al. 2008; Wishart 2011) require the ability of predicting them based on protein structures, a number of methods for performing such predictions have been developed in the last several years (Wishart et al. 1997; Xu and Case 2001; Meiler 2003; Neal et al. 2003; Shen and Bax 2007; Kohlhoff et al. 2009; Lehtivarjo et al. 2009). Although these methods have so far been mainly concerned with backbone chemical shifts, further progress can be expected in establishing fully reliable methods for protein structure determination using side-chain chemical shifts as well. This idea has been

supported by a series of recent studies that reported quantitative relationships between the rotameric states of side-chain methyl groups and the corresponding chemical shift values (Mulder 2009; Hansen et al. 2010). These developments are particularly interesting since proteins are rich in methyl-bearing amino acids and therefore methyl chemical shifts provide excellent opportunities to probe their structures and dynamics (Tugarinov et al. 2005; Gelis et al. 2007; Hsu et al. 2009; Sheppard et al. 2010; Baldwin et al. 2010). Furthermore, optimized NMR experiments to measure chemical shifts and new schemes for efficient and highly-specific isotope labeling of side-chain methyl groups (Goto and Kay 2000; Tugarinov et al. 2006; Kainosho et al. 2006; Otten et al. 2010) are enabling their use to characterise the structure and dynamics of large protein complexes, and are making methyl chemical shifts an ever-growing component in the Biological Magnetic Resonance Data Bank (BMRB) (Ulrich 2007). In order to exploit the potential of methyl chemical shifts for protein structure determination, we developed the *CH3Shift* method for performing their structure-based prediction. We designed the *CH3Shift* method to be based on differentiable functions of the atomic coordinates of the proteins, because, as we have recently demonstrated in the case of backbone chemical shifts (Kohlhoff et al. 2009; Robustelli et al. 2010) this feature makes it possible to incorporate chemical shift information as restraints in molecular dynamics simulations.

Methods

Structure-based prediction of methyl chemical shifts

Most of the current state-of-the-art methods for performing structure-based predictions of chemical shifts (Wishart et al. 1997; Xu and Case 2001; Meiler 2003; Neal et al. 2003; Shen and Bax, 2007; Kohlhoff et al. 2009; Lehtivarjo et al. 2009) are based on the use of a combination of many factors (Jameson 1996), including ring current (Haigh and Mallion 1972; Haigh and Mallion 1980), magnetic anisotropy (McConnell 1957) and electric field (Buckingham 1960; Buckingham and Pople 1963) effects. In addition, it has also been shown recently that predictions of similar accuracy can be obtained by expressions that capture the relationship between structures and chemical shifts by writing formally the chemical shifts as simple functions of atomic coordinates (Kohlhoff et al. 2009). Although this approach provides less insight into the physical effects that determine the chemical shifts, it has the advantage of being computationally efficient and of generating structural restraints to be used in molecular dynamics simulations because the functions that give the chemical shifts are readily calculable and differentiable.

In order to extend this approach to the chemical shifts of methyl groups, in this work we introduce the *CH3Shift* method, which expresses the chemical shift δ of a given nucleus as a combination of phenomenological terms and distance-based terms

$$\delta = \delta^{rc}_{rot} + \Delta\delta_{dih} + \Delta\delta_{ring} + \Delta\delta_{ma} + \Delta\delta_{EF} + \Delta\delta_{dist} \quad (1)$$

where $\delta^{rc}_{rot}, \Delta\delta_{dih}, \Delta\delta_{ring}, \Delta\delta_{ma}, \Delta\delta_{EF}$ and $\Delta\delta_{dist}$ are, respectively, the rotameric, dihedral, ring current, magnetic anisotropy, electric field and the distance-based contributions. For fitting the parameters in these various terms we use a database of experimental methyl chemical shifts and of corresponding high-resolution X-ray structures (see next section). For defining the distance-based terms, we considered atoms in the region between a smaller sphere of 1.8 Å radius and a larger sphere of 6.5 Å radius around each of the methyl groups, centred on the methyl carbon nucleus (Fig. 1). The smaller sphere includes the methyl group itself and the preceding carbon or sulphure (for methionine) atoms since the arrangement within that region can be considered constant regardless of the structural environment and the side-chain conformation.

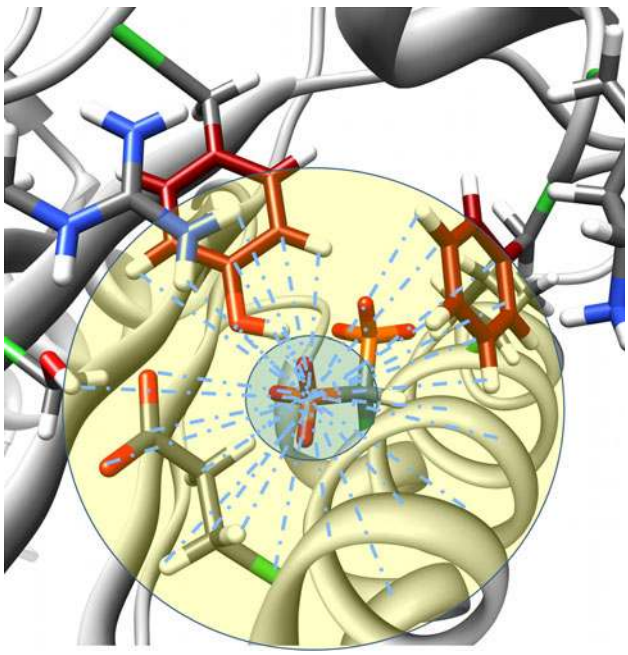

Fig. 1 Illustration of a methyl bearing side-chain with a representation of the active (*yellow*) and neutral (*blue*) regions defined by 6.5 and 1.8 Å cutoff radii, respectively, from the methyl carbon nucleus. Some of the side-chains having significant contributions to the methyl group chemical shifts are explicitly indicated

Database analysis and filtering criteria

In order to parametrize the *CH3Shift* method, we constructed the CH3Shift-DB database by retrieving the chemical shift information from the BMRB (Ulrich 2007) and converting it into CCPN projects (Vranken et al. 2005; Vranken and Rieping 2009). The referencing of the chemical shifts was then corrected, when required, using VASCO (Rieping and Vranken 2010), a method to correct and validate protein chemical shift values in relation to the coordinates of the corresponding nuclei. By an initial filtering, we included only chemical shift entries with stereospecific assignment for Val and Leu residues. Cases for which chemical shifts were flagged as stereospecifically assigned but no difference between the two methyl chemical shifts were discarded. When multiple BMRB records were present, the median of the chemical shift values were taken from all the entries corresponding to the same nuclei in the same protein. This type of averaging ensures that outlying data entries, which can be attributed to various types of artifacts that can arise in the experiments or in the spectra interpretation, have minimal impact on the final compilation of the data. Only the chemical shift entries corresponding to structures determined by X-ray crystallography were considered. Of the total 750 protein structures, each with a unique PDB (Protein Data Bank, Berman et al. 2000) identifier of an X-ray structure, 26 structures were discarded since they were related to protein-nucleic acid complexes; in this way we decreased the possibility of the chemical shift data being modulated by non-protein contacts and ring current effects. 43 other structures were discarded for containing porphyrinic rings, iron or cobalt atoms, in order to filter out any non-standard ring current and paramagnetic effects. The above mentioned filtering criteria resulted in the removal of 1,558 chemical shift entries out of the initial 19,431. The compiled data set thus contained 17,873 residue-specific chemical shift records, which are distributed over the amino acid residue types as 5,965 for Ala, 3,147 for Thr, 2,243 for Val, 2,750 for Leu, 3,126 for Ile, and 642 for Met residues (Fig. 2).

The crystallographic Rfree factor was not used in the filtering procedure because 125 of the 681 PDB files in our database did not include a Rfree value and the values that were available had an average of 0.243, first quartile of 0.222 and third quartile of 0.266, indicating that there are only small variations in these values. It would therefore be difficult to use Rfree values for protein structure selection. We also did not use information about sequence homology for filtering. Indeed for the development of chemical shift predictors the inclusion of similar sequences (and structures) in the database is likely to be advantageous to some extent. Since chemical shift values are very sensitive to the local environment, small changes in homologous

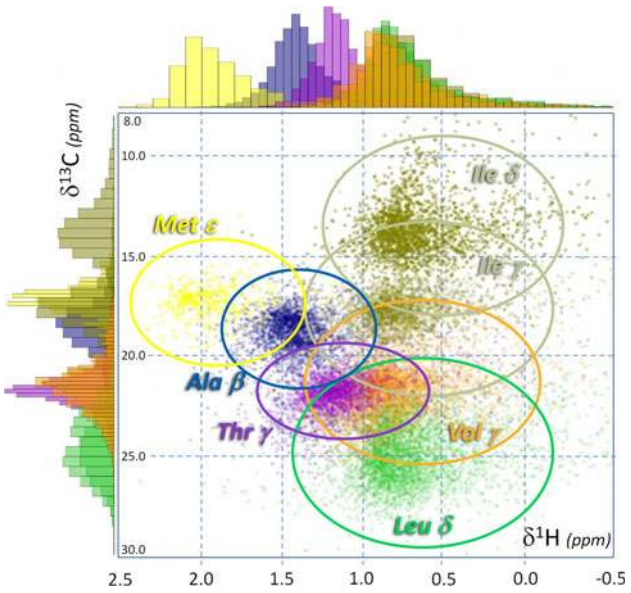

Fig. 2 HSQC-like correlation graph of the methyl group ^{13}C and ^{1}H chemical shift distributions in the CH3Shift-DB database, which shows the different chemical shift propensities for different types of residues. The *ellipsoids* indicate the substantial overlap between the chemical shifts of different methyl group types

structures can result in relatively large differences in actual chemical shift values. For completeness, we calculated the homology between the PDB entries used for generating our database using the PISCES server (Wang and Dunbrack 2003) to generate a list of non-redundant PDB entries from an input list of PDB IDs. A total of 218 entries had a sequence identity of more than 25% with one of the non-redundant entries. Upon increasing the cutoff, the numbers were: 91 entries at 40%, 72 entries at 50%, 55 entries at 60%, 39 entries at 70%, 35 entries at 80% and 31 entries at 90%; thus very similar sequences (more than 80%) only account for about 5% of the total number of entries.

The X-ray structures were preprocessed by the addition of hydrogen atoms followed by 1000 steps of hydrogen-only geometry optimization, using the *Almost* all-atom molecular simulations toolkit (http://open-almost.org, accessed in April, 2010) and the Amber03 force field (Duan et al. 2003). Finally, the database was further optimised by considering only the chemical shifts falling within a window of 2.5 standard deviations for each specific nucleus and residue type, and for which an X-ray structure at 2.0 Å resolution or better was present. The removal of the most uncommon experimental chemical shift values was necessary to avoid the presence of erroneous data or data from measurements in non-standard conditions. This procedure was also useful to avoid the complications associated with considering chemical shifts strongly affected by the close vicinity of aromatic rings or

charged groups, which are highly sensitive to the dynamics and the exact geometric arrangement of the source nuclei and the strong affector moieties.

Rotameric terms

Since effects from the spatial neighbourhood and the conformation of the residue that holds the methyl group alter the chemical shifts of the methyl nuclei from the value determined by the covalently linked local environment, we separated the neighbourhood-independent core component of the chemical shift from the rest. This was done for Ala by allowing the fitting procedure to generate an intercept along with the optimized parameters for the other factors discussed below. For the other residue types, the observation of significant differences between the average chemical shifts in different rotameric states (see Supplementary material S1) suggested the possibility to also account for the rotamer-specific shifts through the intercept. Therefore, for the residue types with a side-chain χ_1 dihedral angle, we considered the expression

$$\delta^{rc}_{rot} = k_1 R_1 + k_2 R_2 + k_3 R_3 \tag{2}$$

where the R_1, R_2 and R_3 factors classify the rotameric state and are equal to 1 for $-120 < \chi_1 \leq 0$, $0 < \chi_1 \leq 120$ and $(120 < \chi_1 \leq 180) \cup (-180 \leq \chi_1 \leq -120)$ conditions for R_1, R_2 and R_3 correspondingly, with 0 values otherwise. The mentioned windows of χ_1 angle well separate the most common three χ_1-based rotameric states and allow treating different rotameric classes separately.

Dihedral angle terms

In these terms we included the backbone ϕ, ψ dihedral angles and all the available side-chain χ_i (with $i = 1, \ldots, 5$) dihedral angles. The effects from each of those angles (if present) were modeled via four polynomial and ten cosine terms (see Supplementary material S2). The ten cosine terms were selected from the analysis of about hundred cosine, sine and mixed terms. We calculated all the geometric terms from the existing dihedral angles in the database of structures. Further, a cross correlation matrix was calculated for the geometric terms along all the functions to identify those correlated with each other. A Pearson correlation coefficient value of 0.7 was used to eliminate strongly correlated functions. The final ten functions were then chosen from the remaining ones according to their simplicity. Different sets of functions were tried, but our results indicated that as long as there is a sufficiently large number of geometric terms that are not strongly correlated (in this case ten cosine functions and four polynomials), the fitting procedure for the coefficient optimization finds values for the coefficients resulting in models of comparable performance.

Ring current terms

Ring current effects on chemical shifts arising from the aromatic rings of Phe, Tyr, His, Trp-5 and Trp-6 (5 and 6-membered tryptophan rings) residues are accounted by the inclusion of $G(\vec{\mathbf{r}})$ geometric factors from the model by Haigh and Mallion (Haigh and Mallion 1972; Haigh and Mallion 1980)

$$\Delta\delta_{ring} = k_{ring} G(\vec{\mathbf{r}}) = k_{ring} \sum_{ij} S_{ij} \left(\frac{1}{r_i^3} + \frac{1}{r_j^3} \right) \tag{3}$$

where S_{ij} is the algebraic (signed) triangle area formed by the O' projection of the query point O onto the ring plane and the ring atoms i and j. Defining $\mathbf{T}_{O'i}$ and $\mathbf{T}_{ij}$ as vectors joining O' to the ring atom i and ring atom i to j respectively, the sign of the triangle is positive if the vector product $\mathbf{T}_{O'i} \times \mathbf{T}_{ij}$ has the same direction as the ring normal with ring atoms counted in $i \rightarrow j$ direction. r_i and r_j are the distances between O and atoms i and j respectively. k_{ring} is a proportionality constant. The summation goes over all the adjacent ij atom pairs forming the ring, that is over the number of bonds in the conjugated ring.

All the aromatic rings that have at least two of their non-hydrogen atoms in the vicinity of the methyl carbon nucleus within the active region are included. For tryptophan residues, if one of the two rings satisfy the above mentioned criterion, the second ring is included as well. The 6.5 Å cutoff radius was chosen because the ring current effects are negligible at distances longer than approximately 5.5 Å (Case 1995). As a query point O, the methyl carbon and the geometric centre of the three methyl hydrogens are taken for ^{13}C and ^{1}H chemical shifts, respectively.

Magnetic anisotropy terms

Magnetic anisotropy effects are incorporated into the calculations by following the method used to account the peptide group anisotropy effects on backbone ^{1}H chemical shifts by Case et al. (Ösapay and Case 1991). The method uses the McConnell formulation (McConnell 1957) of the magnetic anisotropy contribution to the chemical shifts, reduced by an assumption of axial symmetry for the source of the anisotropy. In this case, the distant group magnetic anisotropy contribution to the chemical shift value can be approximated as

$$\Delta\delta_{ma} = \frac{\Delta\chi}{3N_A} \times \frac{3\cos^2\theta - 1}{r^3} \tag{4}$$

where $\Delta\chi$ is the magnetic susceptibility anisotropy, N_A is the Avogadro number, r is the distance between the nucleus and a point defined in the anisotropic moiety, θ is the angle between the **r** vector and the normal of the plane of that group. The second factor in Eq. (4) can be considered as a geometric term for the magnetic anisotropy effects and be included in the modeling of the chemical shifts.

Protein backbone peptide groups, as well as the carboxylic, amide and guanidinium moieties of Asp, Asn, Glu, Gln, and Arg side-chains are considered as sources of magnetic anisotropy. In case of peptide moieties, the optimal placement of the origin on the plane for calculation of **r** is approximately at the center of the OCN group (Ösapay and Case 1991). By generalizing this finding, the geometric centres of the OCO and OCN atoms were used as origins for the carboxylic and amide planes respectively. For arginine side-chains, the carbon centre of the guanidinium group was used.

Electric field terms

Electric fields alter the chemical shifts by polarizing the local electronic distributions. For an atom X that is connected only to another atom Y, this dependence was shown to be approximated by the chemical shift polarizability constant multiplied by the electric field projection along the X-Y axis (Buckingham 1960; Buckingham and Pople 1963). Here, the electric field effect was accounted for by following Coulomb's law and reducing the electrostatic effects of the atoms to the simple electric monopole interactions. Amber03 charges (Duan et al. 2003) were used and only the atoms within the active region were considered. The electric field along the local symmetry axis of the methyl group was calculated, i.e. along the H_3C–C or H_3C–S (for methionine) bond. Thus, the implemented electric field term is

$$\Delta\delta_{EF} = k_{EF}\sum_i \frac{q_i cos\theta}{r_i^2} \quad (5)$$

where q_i is the partial charge of the ith atom in the active region, θ is the angle between the local symmetry axis of the methyl group and the vector **r** with length r_i that joins the methyl nucleus with the ith atom. k_{EF} is the proportionality constant for the electric field term.

Distance-based terms

The distance-based terms used in *CH3Shift* are modified from the scheme implemented in the CamShift method for the backbone nuclei (Kohlhoff et al. 2009). Here we used fewer types of distances, but they were included in a greater number of polynomial terms

$$\Delta\delta_{dist} = \sum_{i\in\{-1,1,3,6\}} k_i r^{-i} \quad (6)$$

Besides the r and r^{-3} terms, which are used for all the atoms, r^{-1} and r^{-6} terms are also added. The inclusion of the r^{-6} term has been implemented in chemical shift predictors for small molecules to treat the weak interaction between atoms (Abraham et al. 2001). The combination of the r, r^{-1} and r^{-3} terms effectively takes into account the electrostatic interactions, given the presence of screening effects that can alter the dielectric constant of the surrounding medium with the strength linearly proportional to the distance from the NMR active nucleus. Furthermore, besides the backbone N, C, H, C_α, H_α and C_β atoms, which are essentially always present in the proximity of the side-chain methyl groups and allow parameter fitting with high statistical significance, the rest of the distances are treated jointly. We used a procedure in which distances are merged, i.e. they are summed after the corresponding power operation. The list of distances treated in this way includes those between the given nucleus and (a) sp^3 hybridized carbons, (b) hydrogen atoms attached to a sp^3 hybridized carbon, (c) sp^2 hybridized carbons (in aromatic rings), (d) hydrogens attached to a sp^2 hybridized carbon, (e) sulphure atoms, (f) hydroxylic oxygens, (g) hydroxylic and thiolic hydrogens, (h) other carbons (side-chain carboxylic, amide), (i) other hydrogen atoms (imino, amino, guanidinium), (j) other oxygen atoms (side-chain carboxylic, amide) and (k) other nitrogen atoms (heterocyclic, amide, guanidinium, lysine amino). The optimal types of merged distances and terms were found by multiple trials, paying a particular attention to measures for avoiding overfitting.

Since accounting for the correct protonation state is very challenging, in the current parameterization we enforced the most common protonation states for all the relevant amino acids during hydrogen addition to the structures in the database. All acidic residues were considered as deprotonated, lysine and cysteine as protonated, and histidine as protonated only at the δ positions. The importance of considering explicitly in the parametrisation the exact protonation states is decreased by the joint treatment of the distances, which we adopted to avoid overfitting problems because the database that we used includes a relatively low number of instances of any particular type of internuclear distance. An accurate assessment of the effects stemming from the different protonation states should become possible with the growth of the amount of structures and associated chemical shift data.

Parameter fitting, optimization and overfitting control

We used the least squares fitting procedure to determine the coefficients in Eq. (1). All the calculations as well as data filtering and manipulations were done in the R statistical

programming language (R Development Core Team, *R: a Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2011).

In order to decrease the number of parameters and increase the statistical significance of the predictions, the model optimization was done by a Monte Carlo procedure in the space of the possible combinations of terms in Eq. (1). In this approach, all the terms were set as adjustable (i.e. present or absent), except the ring current and magnetic anisotropy terms, as they were statistically significant even when the full model was used for fitting. For each nucleus and residue types, 70000 Monte Carlo steps were performed; at each step a randomly selected term was switched on or off with an acceptance probability defined by the Metropolis criterion. As the pseudo-energy in the Monte Carlo procedure, the fitting quality from the leave-one-out tests after each fitting step was used. The temperature factor was defined to obtain about 60–70% acceptance rate, and thus sample the parameter space efficiently. The final model was selected as the one resulting in the best agreement between the predicted and experimental chemical shifts from the leave-one-out tests (see Table 1). Further evidence that the procedure that we followed did not suffer from over-fitting in a significant manner is provided by the observation that different Monte Carlo runs for optimizing some of the empirical geometric terms resulted in slightly different models having between one and four different terms; these models, however, exhibited negligible differences in performance. In addition, the best ten models from each optimization had a quite similar performance. The resulting coefficients can be obtained from the authors as R data objects upon request.

As typical of phenomenological approaches, there is an overlap between different terms in the procedure that we followed here, which can account for a given effect in more than one way. For instance, the anisotropy and ring current effects are modeled by both special geometric factors and the distances joining the atoms of the aromatic rings or magnetically anisotropic molecular moieties to the methyl nuclei. Electric field effects, which is included as a direct evaluation based on partial charges, is also covered by the distance terms. This double-counting makes it difficult to provide a physical interpretation of the individual coefficients resulting from the fitting procedure. Therefore we performed extensive tests of consistency of the prediction performance, looking for possible abrupt changes in the prediction qualities from one trial to another, or from one compilation of the training data to another, which would have suggested the presence of an over-fitting problem. We performed two types of tests to assess the quality of the fits. The first was the standard leave-one-out test, in which any single prediction is done while that particular chemical shift entry with the corresponding structural parameters is excluded from the training set used to

Table 1 Summary of the results of the *CH3Shift* model optimization

Res.	Nucl.	offs.	rot.	F_{EF}	r	$1/r$	$1/r^3$	$1/r^6$	ϕ	ψ	χ_1	χ_2	θ	θ^2	θ^3	θ^4	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6	Ω_7	Ω_8	Ω_9	Ω_{10}	SD_{train}/SE_{pred}
Ala	$^{13}C_\beta$	+	−	−	+	+	+	+	+	+	−	−	+	+	+	+	+	−	+	−	−	−	+	−	+	−	1.873
Ala	$^{1}H_\beta$	+	−	+	+	−	+	+	+	+	+	+	−	+	−	−	+	+	+	−	+	−	−	+	−	−	1.545
Thr	$^{1}H_{\gamma 2}$	+	+	+	+	+	+	+	+	−	−	−	−	+	−	−	+	+	−	−	−	+	−	−	+	+	1.375
Val	$^{1}H_{\gamma 1}$	+	−	+	+	+	+	+	+	+	−	+	−	−	+	−	−	+	−	−	−	−	−	−	−	−	1.362
Val	$^{1}H_{\gamma 2}$	+	+	−	−	+	−	+	−	+	−	+	−	−	−	−	−	+	−	−	−	+	−	+	−	−	1.433
Leu	$^{1}H_{\delta 1}$	+	+	−	−	−	+	+	−	+	−	+	−	−	−	+	−	+	−	−	−	−	−	−	−	−	1.252
Leu	$^{1}H_{\delta 2}$	+	−	+	−	+	+	+	+	−	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	1.421
Ile	$^{1}H_{\gamma 2}$	+	+	+	−	+	+	+	−	−	+	−	+	+	−	+	+	−	−	−	+	+	+	+	−	−	1.413
Ile	$^{1}H_{\delta 1}$	+	−	+	+	+	−	+	+	−	−	−	−	−	+	−	+	+	−	−	−	−	−	−	+	−	1.496

The ratios of the standard deviation of experimental chemical shifts used for the model fitting and the standard error of the predictions in the fitted data (not from the leave-one-out test) are shown. All optimized models have offsets in their equations; the offsets for Thr-$^{1}H_{\gamma 2}$, Val-$^{1}H_{\gamma 2}$, Leu-$^{1}H_{\delta 1}$ and Ile-$^{1}H_{\gamma 2}$ nuclei are rotamer-specific. All the Ω_i terms, which denote the ten cosine functions that we used (see Supplementary Material S2), as well as the θ^i terms, operate on each of the four dihedral angles ϕ, ψ, χ_1 and χ_2. Therefore the absence (−) of any of them results in the reduction of the number of parameters by four. Likewise, the absence of any of the ϕ, ψ, χ_1 or χ_2 terms in the final model means a reduction of the number of parameters by 14 (four for θ^i and ten for Ω_i). All the models also include the terms for ring current and magnetic anisotropy effects from conjugated rings, peptide moieties and anisotropic side-chain moieties, which were always set present and non-adjustable

optimize the coefficients. For the second test, the compiled chemical shift data with the associated structural factors were randomly split into training and test sets with the percentage of data in the test set varying from 5 to 30% of the whole set. The calculations were run for each of the residue and nucleus types separately, and, each of the random splitting of the data were replicated 250 times. The fitting quality is assessed by examining the dependence of the standard errors of prediction in the training and test sets (with all the 250 trials) against the percentage of the whole data used to optimize the coefficients. The cases of over-fitting are characterised by an artificial improvement in the quality of the predictions in the training set associated by a decrease in the quality in the test set, when the percentage of data used for training was decreased (for an example, see Supplementary Material S3). The cases that we report in this work are those for which we found no behaviour characteristic of over-fitting. In other cases, however, e.g. for methionine ^{1}H and ^{13}C chemical shifts, we could not eliminate over-fitting, a result mainly determined by the fewer amount of currently available experimental chemical shift data for methionine residues. Therefore the chemical shifts of methyl groups of methionine side-chains will only be predicted in future versions of the *CH3Shift* method, which will be reparametrized when it will be possible to increase the size of the CH3Shift-DB database.

The *CH3Shift* software program and web server

The structure-based chemical shift predictor for the methyl groups in proteins that we describe in this work is available as a software program. Besides the stand-alone implementation, we created a *CH3Shift* web server. Given the structure file of a protein in PDB format, the program returns the predicted methyl group ^{1}H and ^{13}C chemical shifts. In addition, it has multiple functionalities, such as comparison of the results to the experimental data, re-referencing of the results based on the provided experimental chemical shifts via a least squares optimization and various plotting options. The program is available through the http://vendruscolo.ch.cam.ac.uk/software.html web address. The GUI is developed via the Rwui, a web application to create user friendly interfaces for R scripts (R. Newton and L. Wernisch, *Rwui: A Web Application to Create User Friendly Web Interfaces for R Scripts*, http://rwui.cryst.bbk.ac.uk, 2010).

Results and discussion

CH3Shift-DB, a database of methyl chemical shifts

We created the CH3Shift-DB database of methyl group chemical shifts by filtering and re-referencing the side-chain methyl chemical shifts available from the BMRB database. The CH3Shift-DB database reflects the chemical shift distributions of ^{1}H and ^{13}C atoms for each of the residue and methyl type (see Fig. 2). The significant overlap in the methyl chemical shifts represents the main obstacle in the efficient assignment of the experimental spectra of the methyl group region. The representation in Fig. 2 clearly illustrates the importance of the recent advances in the assignment of the NMR spectra, in particular for large protein complexes (Sprangers and Kay 2007; Sheppard et al. 2009; Xu et al. 2009; Ruschak and Kay 2010).

Analysis of the differences in the methyl group chemical shifts of Val, Leu and Ile residues

The differences of the ^{13}C chemical shifts of the two methyl groups in Val, Leu and Ile residues have recently been shown to be useful for deriving structural information (London et al. 2008; Mulder 2009; Hong et al. 2009). These chemical shift differences depend on the rotameric states of the side-chains, an observation strengthened by the finding that ^{13}C chemical shifts and vicinal J-couplings are correlated (Mulder, 2009). The initial analysis of the CH3Shift-DB database outlines an interdependence of some types of chemical shifts from different methyl groups of Val, Leu and Ile residues (Fig. 3). A significant correlation is present between the two ^{1}H chemical shifts of Val and Leu residues regardless of the rotameric states of the residue (Fig. 3). The reason for the correlations observed among ^{1}H nuclei, but not among ^{13}C nuclei, can be the more pronounced sensitivity of proton chemical shifts on the long-range environmental interactions that are correlated at the two methyl sites of the same residue. These results demonstrate that the magnitude of the chemical shift alterations from the non-bonded interactions are approximately of the same order at two methyl sites of the same residue. On the contrary, the ^{13}C chemical shifts, besides the sensitivity towards the non-bonded effects, are also sensitive to the core effects as supported by the observation of their strong dependence on the dihedral angles defining side-chain conformation (Pearson et al. 1997). Hence, taking the difference of carbon chemical shifts minimizes the contribution from the long-range effects, leaving only the core effects which clearly correlate with the χ dihedral angles.

Challenges in the structure-based predictions of methyl chemical shifts

Despite the recent advances in the structure-based predictions of backbone chemical shifts (Xu and Case 2001; Meiler

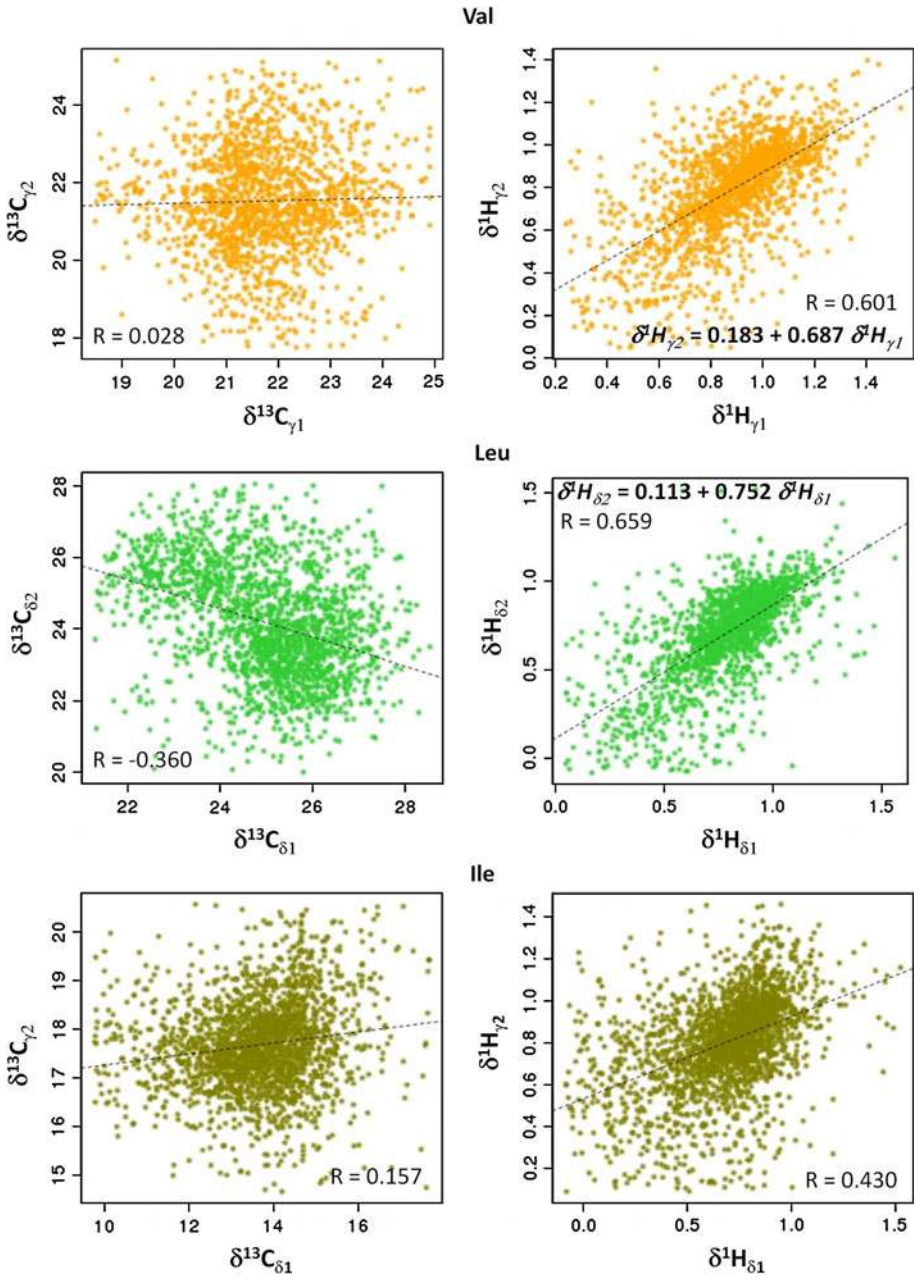

Fig. 3 Correlation between the methyl chemical shifts of the amino acid residues in the CH3Shift-DB database that contain two methyl groups. The correlation coefficients and the linear equations are shown

2003; Neal et al. 2003; Shen and Bax 2007; Kohlhoff et al. 2009; Lehtivarjo et al. 2009; Wishart 2011), the extension of these methods to side-chains has been very challenging for a series of reasons. The first is that the number of methyl chemical shift records in the BMRB is still small when compared to the number of entries for protein backbone nuclei. Thus, the fitting of the parameters for methyl chemical shift predictors can be done based on just a few thousands of experimental data for each methyl type, as opposed to tens of thousand experimental chemical shift entries available for each backbone nucleus. This scarcity of experimental data restricts the number of factors that can be included in the model in order to avoid over-fitting.

The second reason is that our current knowledge of the structure and dynamics of the side-chains, for which methyl group chemical shifts are measured, is often limited. Protein side-chains tend to be rather dynamic, and their positions can be variable because of rotameric jumps. Furthermore, even small uncertainties in the determined average χ_i dihedral angles for the residues, where the methyl is joined to the backbone by a longer chain, result in a more substantial distortion of the methyl group position from its average value. These uncertainties are especially relevant for methyl groups close to aromatic rings, because the geometric factor for describing ring current effects is very sensitive to small fluctuations in the

geometry. The dynamics of the methyl groups have been shown to be comparable in solid and solution states of proteins (Reif et al. 2006; Agarwal et al. 2008), and are expected to be non-negligible (DeGortari et al. 2010). Moreover, solvent-exposed methyl groups, which are likely to be even more dynamic than buried ones, comprise a substantial proportion of the filtered database, since the high quality NMR and X-ray investigations are mostly done on relatively smaller proteins for which the ratio of the surface and core methyl groups is greater than the average. Therefore, in the CH3Shift-DB database, the average structures of the methyl groups from the X-ray studies can vary from the solution state and can negatively affect the quality of the predictions. In an attempt to avoid these problems, we filtered out the surface methyl groups from the training database. The solvent accessible surface area was calculated for each methyl carbon in the database, and the corresponding residue was classified as buried if all its methyl carbons had zero solvent accessible surface area. The percentages of the solvent exposed residues in the database were 73.6% for Ala-β, 86.5% for Thr-γ2, 44.2% for Val-γ1, 43.0% for Val-γ2, 39.0% for Ile-γ2, 38.2% for Ile-δ1, 39.4% for Leu-γ1, 38.3% for Leu-γ2, 66.0% for Met-ϵ. The reduction of the number of entries, however, led to over-fitting problems and thus this approach was not implemented. Furthermore, the existing predictor, which is trained on the database with both buried and exposed residues, did not show an improvement of the performance when only the buried residues were used in leave-one-out tests. On the contrary, a slight decrease of performance was noted for all the tested nuclei, pointing out that, overall, the high-resolution protein structures used in the fitting procedure resulted in a model that is close to the maximum possible performance one can expect from the current state of the database and the difference between the buried and exposed residues can be accounted only after having a substantial improvement of the quality and quantity of data in the CH3Shift-DB database.

Many of the geometric factors in Eq. (1) are very sensitive to the dynamics of the methyl groups and the surrounding residues. Moreover, the dependence is not linear, thus short and long-range structural fluctuations are crucial in determining the actual values of the structural factors. Ideally, instead of using a single structure for each of the selected proteins, an ensemble of conformations should be analysed to retrieve and average out all the structural factors. However, although feasible for protein backbone atoms (Lehtivarjo et al. 2009), the ensemble version of the *CH3Shift* parametrization is yet to benefit from the increasing quality of molecular mechanics force fields for side-chains (Lindorff-Larsen et al. 2010). The complex effects that the dynamics has on the chemical shifts are also indicated by the result that the changes in the absolute errors in the ^{1}H chemical shift predictions calculated from the X-ray structure were not correlated with the S^2 order parameter over different methyl groups in ubiquitin (Supplementary Information S5). Although a special attention is paid to the processing and filtering steps (see section "Methods"), some remaining uncertainties in referencing and stereospecific assignment can still be an issue in the compiled chemical shift data. The fraction of those uncertainties will certainly be reduced with time, owing to increasingly standardized experiments and efficient stereospecific assignment techniques.

Finally, perhaps the biggest problem in developing a protein methyl chemical shift predictor is the small variance of the experimental chemical shift values observed in methyl ^{1}H and ^{13}C chemical shifts, as compared to the variance of the chemical shifts of backbone nuclei. Thus, for an acceptable predictive power, the model here is required to produce results that have much smaller standard errors as compared to the backbone chemical shift predictors, for the errors to be smaller than the already small standard deviations of the corresponding experimental chemical shift values in BMRB.

Random coil methyl chemical shifts

As noted above, methyl chemical shifts of proteins tend to have a small variance compared to other types of chemical shifts, as clearly indicated by the BMRB statistics (Ulrich 2007). This observation can be explained by the dynamical nature of the methyl group bearing side-chains and the absence of specific interactions, such as hydrogen bonding, that involve or are close to the sites of the side-chain methyl groups. A smaller electronic polarizability at the methyl sites in comparison to that at the diatomic moieties of the protein backbone can also be the reason for the smaller methyl chemical shift variance, as the electron distribution at the methyl sites and the corresponding nuclear shieldings are expected to be less affected by environmental and non-bonded effects. Thus, methyl chemical shifts are expected to be fairly close to their random coil values. For a quantitative investigation of this phenomenon, we further analysed the extracted and re-referenced chemical shift data to derive random coil values for the methyl ^{13}C and ^{1}H chemical shifts. Here, for a given type of nucleus and amino acid, the random coil chemical shift is defined as the average value of all the recorded experimental chemical shifts that come from solvent accessible residues which, along with the adjacent two residues, have ϕ/ψ dihedral angle combinations characteristic to either turns or coils. This definition is analogous to that used in the CamCoil method, which has been shown to provide accurate predictions of backbone random coil

Table 2 Comparison of the random coil chemical shifts for the ^{13}C and ^{1}H nuclei of the protein side-chain methyl groups with the corresponding average chemical shift values for the α-helical and β-strand structures

	Ala-β	Thr-γ2	Val-γ1	Val-γ2	Leu-δ1	Leu-δ2	Ile-γ2	Ile-δ1	Met-ϵ
^{13}C									
$\overline{\delta}_{rc}$	19.015	21.673	21.231	20.955	24.684	23.794	17.567	13.457	17.285
SD_{rc}	1.341	0.638	0.895	1.191	1.326	1.300	0.844	1.305	0.906
N_{rc}	721	367	134	95	177	125	126	128	37
$\overline{\delta}_{\alpha}$	18.199	21.695	22.115	22.372	24.785	24.015	17.599	13.663	17.010
SD_{α}	0.927	0.759	1.051	1.205	1.389	1.535	0.923	1.247	0.789
N_{α}	1520	271	341	308	641	509	439	445	128
$\overline{\delta}_{\beta}$	21.552	21.565	21.499	21.281	24.957	24.832	17.825	13.878	17.317
SD_{β}	1.660	0.860	0.960	1.287	1.549	1.517	0.961	1.296	1.014
N_{β}	494	339	532	375	394	267	537	529	58
^{1}H									
$\overline{\delta}_{rc}$	1.356	1.177	0.903	0.834	0.844	0.742	0.846	0.748	1.911
SD_{rc}	0.163	0.152	0.165	0.216	0.180	0.242	0.216	0.244	0.299
N_{rc}	515	496	136	102	171	141	165	152	52
$\overline{\delta}_{\alpha}$	1.439	1.190	0.949	0.835	0.783	0.707	0.790	0.676	1.827
SD_{α}	0.189	0.155	0.206	0.257	0.220	0.249	0.231	0.260	0.283
N_{α}	954	332	338	306	599	501	505	509	150
$\overline{\delta}_{\beta}$	1.272	1.078	0.823	0.732	0.760	0.631	0.758	0.660	1.820
SD_{β}	0.200	0.162	0.208	0.230	0.223	0.270	0.235	0.237	0.341
N_{β}	338	443	528	429	366	285	645	595	75

The standard deviations (SD) and the number of entries (N) in the corresponding data sets are shown

chemical shifts (DeSimone et al. 2009). The resulting values are summarized in Table 2 along with the standard deviation (SD) and the number (N) of chemical shift entries that fulfilled the above mentioned filtering criteria. For the comparison of the derived random coil values and the associated statistical data with those from structured regions of proteins, a similar filtering of data was done to derive average α-helical and β-strand chemical shift values. We found that chemical shifts from the structured regions do not differ much from their random coil values (Table 2). The only exception is for alanine residues, for which the methyl group is of C_{β} type, thus is strongly influenced by the backbone conformation. Overall, the data indicate that the development of a protein methyl chemical shift predictor concerns relatively small deviations from random coil chemical shift values.

Performance of the *CH3Shift* method

In order to assess the performance of the *CH3Shift* predictor, we report the correlations between the predicted and experimental chemical shifts with standard errors, which are defined as the standard deviation of the prediction errors (in ppm), and correlation coefficients indicated on the plots (Fig. 4, left). The correlation is obtained from leave-one-out tests, so that the tested data were not used in the parametrization of the method for that particular prediction. The corresponding distributions of the prediction errors are presented in Fig. 4, right. Only those nuclei and residue types are presented and discussed herein for which the prediction accuracy is substantial.

Except for alanine residues, predictions for ^{13}C nuclei do not provide a significant improvement over those based on the average values derived from the BMRB database (Supplementary Information S4). The reason for this situation is most probably the neglect of the strong isotope effects on ^{13}C nuclei caused by the immediately attached hydrogen. It will perhaps become possible to account for these effects in the parametrization step by considering a database that includes additional information about the isotopic state of the attached hydrogen atoms (–CD3, –CHD2, –CH2D, –CH3).

We then considered the standard errors of the *CH3Shift* chemical shift predictions (Fig. 5, green bars), and compared them with the standard deviations of the corresponding chemical shifts in the BMRB repository. Overall, the prediction quality is the best for alanine residues (Figs. 4, 5). We also found, not unexpectedly, a

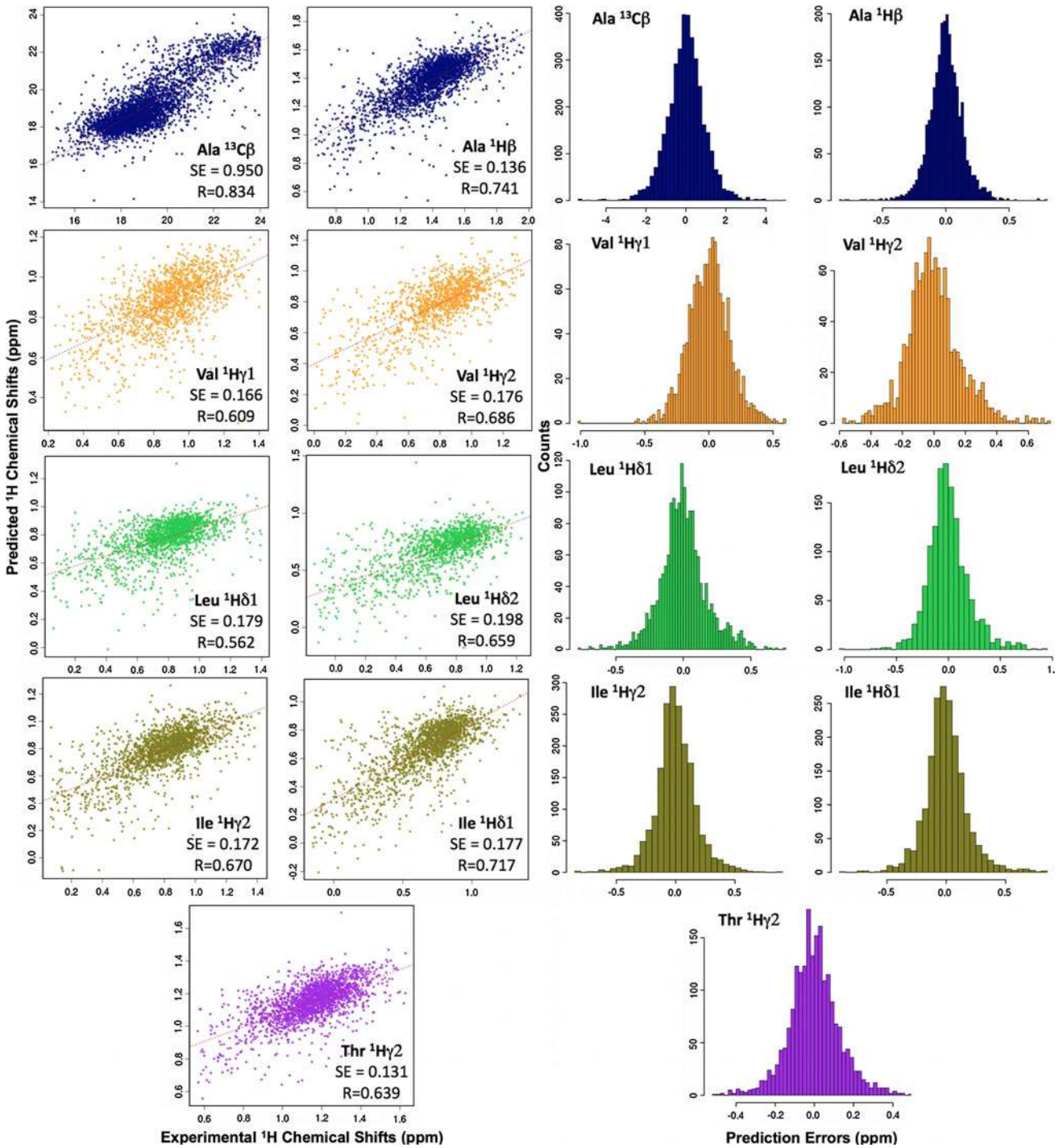

Fig. 4 Correlation between predicted and experimental chemical shifts for all the types of methyl ^{1}H and Ala ^{13}C nuclei (*left*) in the CH3Shift-DB database. Predictions are obtained from leave-one-out tests, with standard errors given in ppm; the Pearson correlation coefficients are also shown. The histograms of the error distributions for each of the discussed nucleus and residue types are shown at the *right side*

decay of the performance of predictor as the side-chain length grows (Fig. 5). This effect can be attributed to the structural and dynamical uncertainties associated with the increase in the number of dihedral angles defining the system.

An assessment of the applicability of the *CH3Shift* method for protein structure determination

The *CH3Shift* method was designed to provide methyl chemical shift predictions that can be incorporated in

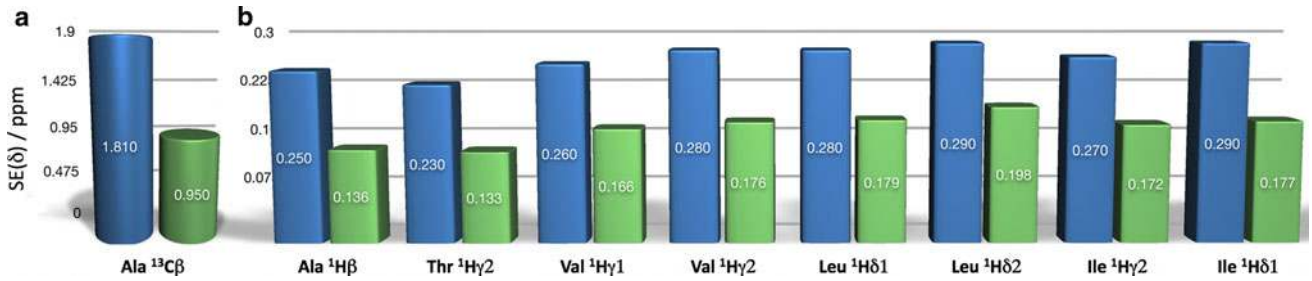

Fig. 5 Histogram of the standard errors (in ppm) of the methyl chemical shift predictions in the different types of protein side-chain methyl groups for which a good accuracy is achieved. The *green bars* show the standard errors of the *CH3Shift* predictor, the *blue bars* show the standard deviations of the corresponding chemical shifts as inferred from BMRB

protein structure determination methods. In this sense, the *CH3Shift* method extends to methyl-bearing side-chains the strategy that we recently proposed for backbone chemical shifts using the CamShift method (Kohlhoff et al. 2009; Robustelli et al. 2010). Our initial tests indicated that, despite the associated errors in predictions of the methyl chemical shifts in the current implementation of the *CH3Shift* method, such predictions can be used to correctly rank protein structures in terms of their overall distance from the reference conformation of the protein, for which we took a high-resolution X-ray structure (Vila and Scheraga 2009). To test the possibility for such usage of the *CH3Shift* predictor, we analysed with *CH3Shift* the 2NR2 dynamical ensemble of ubiquitin (Richter et al. 2007). The chemical shifts were calculated for the methyl group nuclei for each of the 144 conformers of the ensemble. The outcome of this trial demonstrates that for a given methyl group the structures that result in better predictions have local environments closer to that in the reference X-ray structure (1UBQ, (Vijay-Kumar et al. 1987)) of ubiquitin (Fig. 6). The green model corresponds to the X-ray structure of ubiquitin, whereas the blue and red models to the structures with the best and worst agreement, respectively, of the methyl group chemical shift prediction results with the experimental values. For each of the methyl groups, the best local structure is selected from 144 conformations as the one with the best predicted ^{1}H chemical shifts and the ^{13}C predictions in the top ten. This scheme reduces the importance of the carbon chemical shifts, because of the current overall lower prediction quality for methyl carbons. For Ala-46 (Fig. 6), although the neighbouring phenylalanine ring position of the worst agreement structure is closer to that in the X-ray one, the methyl group is shifted with a significant deviation of its position relative to the ring. On the contrary, the structure of best agreement, which is altered by the loop movement, keeps the relation between the side-chain positions close to the arrangement in the X-ray structure. For Thr-66, an excellent match between the best-agreement and X-ray structures is found, whereas the structure

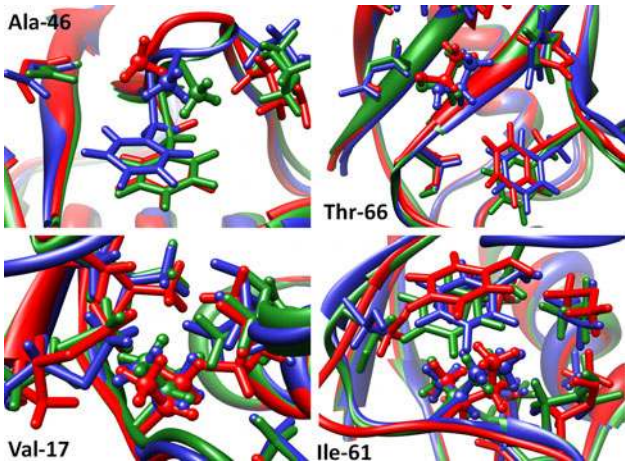

Fig. 6 Methyl chemical shift analysis of the 2NR2 dynamical ensemble of ubiquitin. The X-ray structure (*green*) is compared with the best (*blue*) and the worst (*red*) structures in the 2NR2 ensemble in terms of agreement between experimental and calculated methyl chemical shifts. The methyl containing target residues are highlighted as *ball-and-stick* representations, and the notable residues in vicinity are shown as *stick* representations

of worst agreement suffers from significantly distorted phenilalanine and histidine ring positions. For Val-16, the overall positions of all the influential moieties around the methyl groups are closer between the X-ray and best-agreement structures. An interesting case is that of Ile-61, for which not only the tyrosine ring is substantially distorted in the worst-agreement structure, but also the rotameric type of the isoleucine side-chain itself is different. These results thus indicate that refinement strategies based on methyl chemical shifts have the potential of increasing the accuracy of the side-chain positions.

Next, we analysed the 2K39 (Lange et al. 2008) ensemble and the 1D3Z (Cornilescu et al. 1998) structures in comparison to the 2NR2 ensemble and the 1UBQ X-ray structure. Unlike 1D3Z, which contains 10 structures that fit to the NOE, J-coupling and RDC data individually, the 2K39 and 2NR2 ensembles (with 116 and 144 structures respectively) are the results of a treatment of NMR data

aimed at reflecting the dynamics of the protein. A recent model free analysis (MFA) of the NMR restraints for the ubiquitin methyl side-chains has shown (Fares et al. 2009) that the 2NR2 ensemble agrees best with the RDCs derived from spherical harmonics according to the Pearson correlation coefficient, but the 2K39 ensemble exhibits a better RMSD (in ppm). Therefore, additional comparisons of these two ensembles using different approaches can be important for a further assessment of the methodologies to derive protein dynamics from NMR data. We assessed the quality of the back-calculated *CH3Shift* chemical shifts for methyl ^{1}H nuclei of the ubiquitin various ensembles in representing the experimental values. Average RMSDs (in ppm) of the methyl ^{1}H chemical shift prediction errors in 2K39 (116 structures, red), 2NR2 (144 structures, blue) and 1D3Z (10 structures, grey) ensembles, as compared to the prediction errors from the 1UBQ X-ray structure of ubiquitin (green) are shown in Fig. 7. If the residue contains two methyl groups, the data from both methyl moieties are used for the RMSD calculations. The whiskers indicate the standard deviation of RMSDs over the constituent conformers. The worse RMSDs are not directly related to the solvent accessibility of the residue, as can be seen from the colour-coded band at the bottom of the figure. The observed large RMSDs for Ala-46 and Leu-50 are likely to be connected to the effects of the Phe-45 and Tyr-59

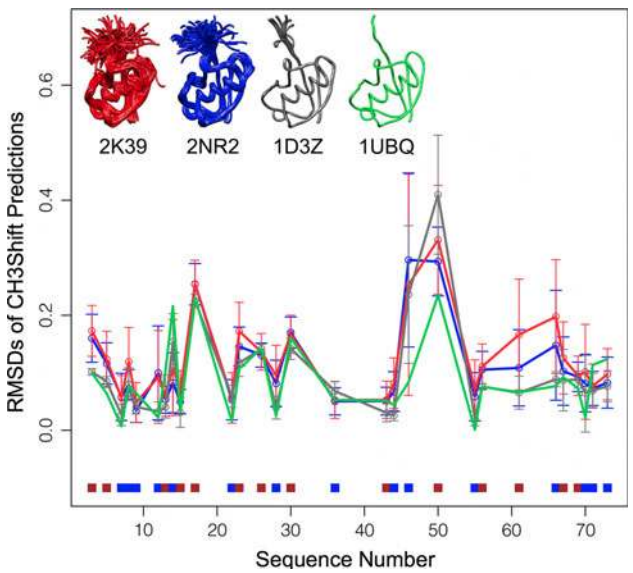

Fig. 7 Average RMSDs (in ppm) in the *CH3Shift* predictions of methyl ^{1}H chemical shifts for the 2K39 (116 structures, red), 2NR2 (144 structures, *blue*) and 1D3Z (10 structures, *gray*) ensembles. For comparision, the corresponding RMSDs are shown for an X-ray structure of ubiquitin (1UBQ, *green*). Standard deviations of the RMSD values over the conformers are shown as *whiskers*. The colour-coded band at the *bottom* indicates the residue-specific solvent accessibility with the *blue colour* for the solvent-exposed methyl groups and *brown colour* for the buried ones

aromatic rings at the vicinity. For a clearer view of the correspondence between the calculated and experimental chemical shifts, the individual correlation plots are shown in Fig. 8. The best agreement is found for the X-ray structure (Figs. 7, 8). Although this result could simply be a consequence of the fact that only X-ray structures of proteins were used to parametrize the *CH3Shift* predictor, it may be also possible that the NMR ensembles, which were derived using other NMR parameters (S^2 order parameters and RDCs), may not represent very accurately the specific population weights that would result in better estimates of the chemical shifts.

As a further assessment of the quality of the ensembles, the leucine ^{13}C chemical shift differences were estimated via the equation (Mulder 2009) $\Delta\delta^{13}C(\delta_1 - \delta_2) = -5 + 10p_{tr}$ and compared to the experimental values. The p_{tr} is the fraction of the leucine side-chain trans (by χ_2) rotamer during the course of the dynamics and is estimated here based on all the constituent conformers in each of the ubiquitin ensembles. The results are summarized in Fig. 9. The results from 1D3Z should be interpreted considering that this ensemble is not meant to represent the dynamics of the protein, but rather to provide a high-resolution representation of its average structure. It should also be noted that, in the case of the structural ensembles considered here, the overall correspondence between the experimental ^{13}C chemical shift difference for leucine and the corresponding values predicted through Mulder's equation is comparable to that of the standard deviation of the experimental chemical shifts (1.59 ppm for $C_{\delta 1}$ and 1.68 ppm for $C_{\delta 2}$). The examination of the χ_1/χ_2 rotamer distribution for the 2NR2 ensemble indicates a strong correlation of the two side-chain dihedral angles with a prevalent population of two rotameric states in most of the cases. This result, although is in contrast to the similar examination of the 2K39 ensemble, is in a good agreement with previous observations on the usual behaviour of leucine side-chains (London et al. 2008; Mulder 2009; Hansen et al. 2010).

In principle, one could expect an improvement in the predictions of the chemical shifts from the inclusion of time and ensemble averaging (DeGortari et al. 2010; Jensen et al. 2010). It is therefore of great relevance to develop methods of the type that we present here to enable the chemical-shift based refinement of side-chain conformations and dynamics.

Conclusions

We have presented the *CH3Shift* method for the structure-based prediction of protein methyl chemical shifts. The predictions are performed by using a combination of

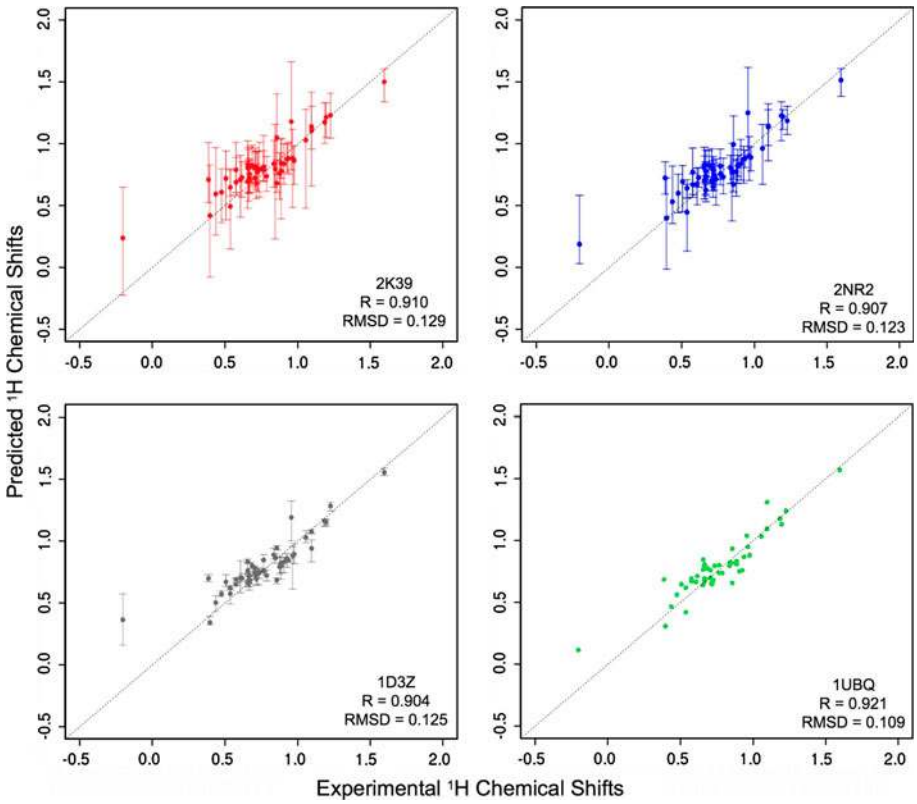

Fig. 8 Correlation between the predicted and experimental ^{1}H chemical shifts for the methyl groups in three ubiquitin ensembles (2NR2, 2K39, 1D3Z) and one X-ray structure (1UBQ). The *whiskers* show the range of the predicted chemical shifts over the multiple conformers where available. The Pearson correlation coefficients are shown

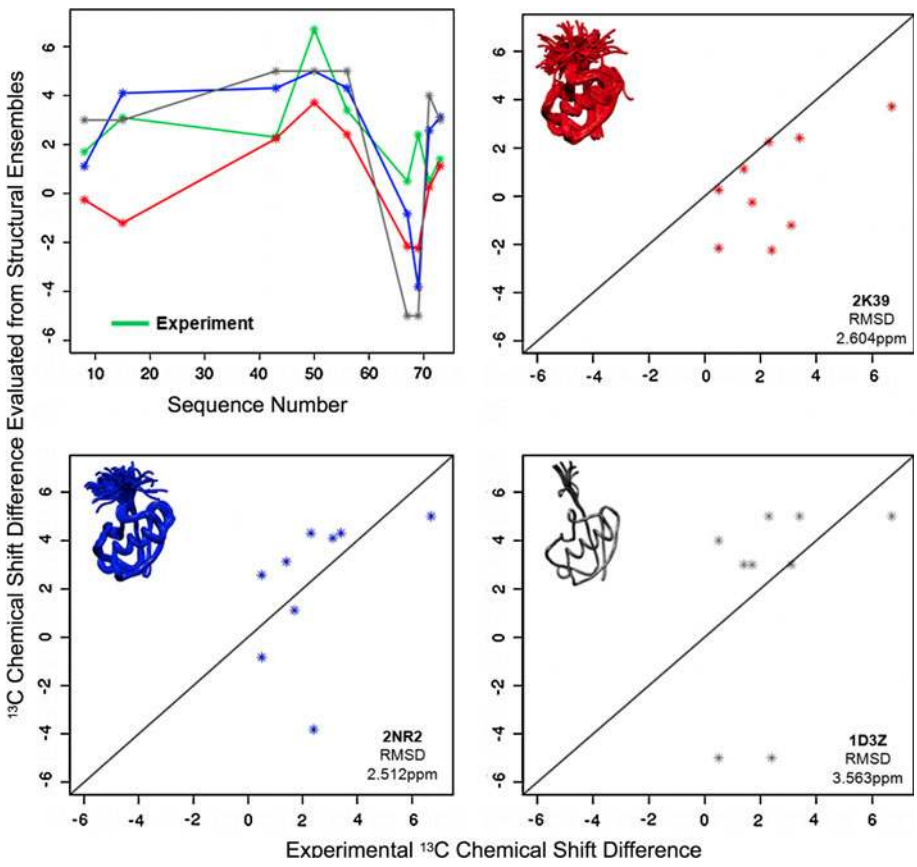

Fig. 9 Differences (in ppm) in the methyl chemical shifts of leucine side-chains in three ubiquitin ensembles (2K39—*red*, 2NR2—*blue* and 1D3Z—*gray*) as predicted through the formula proposed by Mulder (Mulder 2009). Residue-specific predictions are compared with the corresponding experimental values (*green*)

polynomial functions of interatomic distances with well-characterised phenomenological terms that describe effects of ring currents, magnetic anisotropies, electric fields, rotameric types, and dihedral angles. We have shown that the performance of the *CH3Shift* method for Ala, Thr, Val, Leu and Ile methyl groups provides an opportunity for the use of the *CH3Shift* method to assess the quality of protein structures. Furthermore, we anticipate that it will be possible to continuously improve the quality of the predictions with the growth in the number of methyl chemical shift data deposited in the BMRB, and the development of molecular mechanics force fields optimized for side-chain atoms.

Acknowledgments A.B.S. thanks Herchel Smith Foundation for the generous support. M.V. acknowledges the funding from the Leverhulme Trust, EMBO, the Royal Society and the BBSRC. W.F.V. was supported by the EU FP7 e-NMR grant 213010.

References

Abraham R, Canton M, Griffiths L (2001) Proton chemical shifts in nmr: Part 17. Chemical shifts in alkenes and anisotropic and steric effects of the double bond. Magn Reson Chem 39:421–431

Agarwal V, Xue Y, Reif B, Skrynnikov NR (2008) Protein side-chain dynamics as observed by solution- and solid-state nmr spectroscopy: a similarity revealed. J Am Chem Soc 130:16611–16621

Baldwin AG, Religa TL, Hansen DF, Bouvignies G, Kay LE (2010) 13chd2 methyl group probes of millisecond time scale exchange in proteins by 1h relaxation dispersion: an application to proteasome gating residue dynamics. J Am Chem Soc 132:10992–10995

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

Buckingham AD (1960) Chemical shifts in the nuclear magnetic resonance spectra of molecules containing polar groups. Can J Chem 38:300–307

Buckingham AD, Pople JA (1963) High-resolution n.m.r. spectra in electric fields. Trans Faraday Soc 59:2421–2430

Case DA (1995) Calibration of ring-current effects in proteins and nucleic acids. J Biomol NMR 6:341–346

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

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

Das R, Andre I, Shen Y, Wu YB, Lemak A, Bansal S, Arrowsmith CH, Szyperski T, Baker D (2009) A transient and low-populated protein-folding intermediate at atomic resolution. Proc Natl Acad Sci USA 106:18978–18983

DeGortari I, Portella G, Salvatella X, Bajaj VS, van der Wel PS, 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

DeSimone A, Cavalli A, Hsu STD, 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 131:16332–16333

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 feld for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012

Fares C, Lakomek NA, Walter KFA, Frank BTC, Meiler J, Becker S, Griesinger C (2009) Accessing ns-μs side chain dynamics in ubiquitin with methyl rdcs. J Biomol NMR 45:23–44

Gelis I, Bonvin AM, Keramisanou D, Koukaki M, Gouridis G, Karamanou S, Economou A, Kalodimos CG (2007) Structural basis for signal-sequence recognition by the translocase motor seca as determined by nmr. Cell 131:756–769

Goto NK, Kay LE (2000) New developments in isotope labeling strategies for protein solution nmr spectroscopy. Curr Opin Struct Biol 10:585–592

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

Haigh CW, Mallion RB (1980) Ring current theories in nuclear magnetic resonance. Prog NMR Spectrosc 13:303–344

Hansen DF, Neudecker P, Vallurupalli P, Mulder FAA, Kay LE (2010) Determination of leu side-chain conformations in excited protein states by nmr relaxation dispersion. J Am Chem Soc 132:42–43

Hong M, Mishanina TV, Cady SD (2009) Accurate measurement of methyl ^{13}c chemical shifts by solid-state nmr for the determination of protein side chain conformation: the influenza a m2 transmembrane peptide as an example. J Am Chem Soc 131: 7806–7816

Hsu STD, Cabrita LD, Fucini P, Christodoulou J, Dobson CM (2009) Probing side-chain dynamics of a ribosome-bound nascent chain using methyl nmr spectroscopy. J Am Chem Soc 131:8366–8367

Jameson CJ (1996) Understanding nmr chemical shifts. Annu Rev Phys Chem 47:135–169

Jensen MR, Salmon L, Nodet G, Blackledge M (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. J Am Chem Soc 132:1270–1272

Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei MO, Guntert P (2006) Optimal isotope labeling for nmr protein structure determinations. Nature 440:52–57

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, Lakomek NA, Fares C, Schroder GF, Walter KFA, Becker S, Meiler J, Grubmuller H, Griesinger C, de Groot BL (2008) Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. Science 320:1471–1475

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

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

London RE, Wingad BD, Mueller GA (2008) Dependence of amino acid side chain ^{13}c shifts on dihedral angle: application to conformational analysis. J Am Chem Soc 130:11097–11105

McConnell HM (1957) Theory of nuclear magnetic shielding in molecules. 1. long-range dipolar shielding of protons. J Chem Phys 27:226–229

Meiler J (2003) Proshift: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26:25–37

Montalvao R, Cavalli A, Salvatella X, Blundell TL, Vendruscolo M (2008) Structure determination of protein-protein complexes using nmr chemical shifts: the case of an endonuclease colicin—immunity protein complex. J Am Chem Soc 130:15990–15996

Mulder FAA (2009) Leucine side-chain conformation and dynamics in proteins from ^{13}c nmr chemical shifts. Chem Bio Chem 10:1477–1479

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

Oldfield E (1995) Chemical shifts and 3-dimensional protein structures. J Biomol NMR 5:217–225

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

Otten R, Chu B, Krewulak KD, Vogel HJ, Mulder FA (2010) Comprehensive and cost-effective nmr spectroscopy of methyl groups in large proteins. J Am Chem Soc 132:2952–2960

Pearson JG, Le H, Sanders LK, Godbout N, Havlin RH, Oldfield E (1997) Predicted chemical shifts in proteins: structure refinement of valine residues by using ab initio and empirical geometry optimizations. J Am Chem Soc 119:11941–11950

Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) Nmr structure determination for larger proteins using backbone-only data. Science 327:1014–1018

Reif B, Xue Y, Agarwal V, Pavlova MS, Hologne M, Diehl A, Ryabov YE, Skrynnikov NR (2006) Protein side-chain dynamics observed by solution- and solid-state nmr: comparative analysis of methyl ^{2}h relaxation data. J Am Chem Soc 128:12354–12355

Richter B, Gsponer J, Varnail P, Salvatella X, Vendruscolo M (2007) The mumo (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. J Biomol NMR 37:117–135

Rieping W, Vranken WF (2010) Validation of archived chemical shifts through atomic coordinates (vasco). Proteins 78:2482–2489

Robustelli P, Cavalli A, Vendruscolo M (2008) Determination of protein structures from solid-state nmr chemical shifts. Structure 16:1764–1769

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

Ruschak A, Kay LE (2010) Methyl groups as probes of supramolecular structure, dynamics and function. J Biomol NMR 46:75–87

Shen Y et al (2008) Consistent blind protein structure generation from nmr chemical shift data. Proc Natl Acad Sci USA 105:4685–4690

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

Sheppard D, Guo C, Tugarinov V (2009) 4d ^{1}h - ^{13}c nmr spectroscopy for assignments of alanine methyls in large and complex protein structures. J Am Chem Soc 131:1364–1365

Sheppard D, Sprangers R, Tugarinov V (2010) Experimental approaches for nmr studies of side-chain dynamics in high-molecular-weight proteins. Prog NMR Spectrosc 56:1–45

Sprangers R, Kay L (2007) Quantitative dynamics and binding studies of the 20s proteasome by nmr. Nature 445:618–622

Tugarinov V, Ollerenshaw JE, Kay LE (2005) Probing side chain dynamics in high molecular weight proteins by deuterium nmr spin relaxation: an application to an 82-kda enzyme. J Am Chem Soc 127:8214–8225

Tugarinov V, Kanelis V, Kay LE (2006) Isotope labeling strategies for the study of high-molecular-weight proteins by solution nmr spectroscopy. Nat Protoc 1:749–754

Ulrich EL (2007) Biomagresbank. Nucl Acids Res 36:D402–D408

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

Vila JA, Scheraga HR (2009) Assessing the accuracy of protein structures by quantum mechanical computations of 13c(alpha) chemical shifts. Acc Chem Res 42:1545–1553

Vranken WF, Rieping W (2009) Realtionship between chemical shift value and accessible surface area for all amino acid atoms. BMC Struc Biol 9:20

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The ccpn data model for nmr spectroscopy: development of a software pipeline. Proteins 59:687–696

Wang G, Dunbrack RL (2003) Pisces: a protein sequence culling server. Bioinformatics 19:1589–1591

Wishart DS (2011) Interpreting protein chemical shift data. Prog Nucl Magn Reson Spectrosc 58:62–87

Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) Automated h-1 and c-13 chemical shift prediction using biomagresbank. J Biomol NMR 10:329–336

Xu XP, Case DA (2001) Automated prediction of 15n, 13cα, 13cβ and 13c$'$ chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333

Xu Y, Liu M, Simpson PJ, Isaacson R, Cota E, Marchant J, Yang D, Zhang X, Freemont P, Matthews S (2009) Automated assignment in selectively methyl-labeled proteins. J Am Chem Soc 131:9480–9481