J Biomol NMR (2009) 45:413–426
DOI 10.1007/s10858-009-9384-1

ARTICLE

4D prediction of protein ^{1}H chemical shifts

Juuso Lehtivarjo · Tommi Hassinen · Samuli-Petrus Korhonen · Mikael Peräkylä · Reino Laatikainen

Received: 3 July 2009 / Accepted: 9 October 2009 / Published online: 30 October 2009

Abstract A 4D approach for protein ^{1}H chemical shift prediction was explored. The 4th dimension is the molecular flexibility, mapped using molecular dynamics simulations. The chemical shifts were predicted with a principal component model based on atom coordinates from a database of 40 protein structures. When compared to the corresponding non-dynamic (3D) model, the 4th dimension improved prediction by 6–7%. The prediction method achieved RMS errors of 0.29 and 0.50 ppm for Hα and HN shifts, respectively. However, for individual proteins the RMS errors were 0.17–0.34 and 0.34–0.65 ppm for the Hα and HN shifts, respectively. X-ray structures gave better predictions than the corresponding NMR structures, indicating that chemical shifts contain invaluable information about local structures. The ^{1}H chemical shift prediction tool 4DSPOT is available from http://www.uku.fi/kemia/4dspot.

Keywords Protein · Proton · Chemical shift · Prediction · Molecular dynamics

Electronic supplementary material The online version of this article (doi:10.1007/s10858-009-9384-1) contains supplementary material, which is available to authorized users.

J. Lehtivarjo (✉) · T. Hassinen · R. Laatikainen
Department of Biosciences, Laboratory of Chemistry, University of Kuopio, P.O.Box 1627, 70211 Kuopio, Finland
e-mail: juuso.lehtivarjo@uku.fi

M. Peräkylä
Department of Biosciences, Laboratory of Biochemistry, University of Kuopio, P.O.Box 1627, 70211 Kuopio, Finland

S.-P. Korhonen
Perch Solutions Ltd., Hyrräkatu 3 A 1, 70500 Kuopio, Finland

Introduction

Traditionally, NMR studies of protein structures have been based on NOE and coupling constant information. Nowadays the dipolar interactions have also been added to the tool box. However, also the chemical shifts contain a large amount of structural information, often used for determining dihedral angle restraints for the structure calculations. The correlations between chemical shifts and protein structures have been intensively studied (for a review, see Wishart and Case 2001). Recently, this has lead to methods for determining whole protein structures using chemical shift information and sequence-based modeling methods only (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008).

The difference between the observed protein chemical shifts and the corresponding random coil shifts is called the *secondary shift*, which is a result of several structural effects. Primarily, the backbone ^{1}H chemical shifts are dependent on the secondary structure and the backbone torsion angles Φ and Ψ (Neal et al. 2003; Wang 2004; Ösapay and Case 1991). For the Hα shifts, the contribution of the secondary shift effects to the total variation is estimated to be approximately 75% (Wishart and Case 2001). For the HN shifts it is almost 100% (Wishart and Case 2001), including a strong contribution from the hydrogen bonding effects (de Dios et al. 1993; Moon and Case 2007; Parker et al. 2006). The largest single effect to all ^{1}H shifts arises from the aromatic ring currents, but they affect only 10–15% of all protons (Wishart and Case 2001). Solvent exposure has also a significant effect to the ^{1}H shifts (Avbelj et al. 2004; Vranken and Rieping 2009). Due to the flexibility of proteins in the liquid phase, conformational averaging affects the chemicals shifts, especially in the coil regions. Smaller effects have been proposed to arise from

side chain orientation, local charges, and experimental conditions like sample pH, temperature and concentration (Neal et al. 2003; Wishart and Case 2001). For the side chain shifts, contribution of folding to shift variations is smaller, mostly arising from the aromatic ring currents.

Recently, several different approaches to predict protein ^{1}H chemical shifts have been proposed. SHIFTX (Neal et al. 2003) is a widely used program, which uses chemical shift hypersurfaces to describe the structural effects such as torsion angles, and classical equations to calculate the physical effects. Quite a similar approach is adapted in the PRSI program (Wang 2004), which uses hypersurfaces separately for different secondary structures and relies on larger database. In the SPARTA program (Shen and Bax 2007) the torsion angle effects are combined with sequence homology search. ^{1}HN shifts are modeled with density functional calculations in the SHIFTS approach (Moon and Case 2007), and PROSHIFT (Meiler 2003) uses artificial neural networks to predict the chemical shifts from empirical structural information. In the CamShift program, shifts are calculated from interatomic distances (Kohlhoff et al. 2009).

Many applications exploiting the protein chemical shift prediction exist, mostly focusing on protein structure elucidation. Direct chemical shift refinement protocols (Kuszewski et al. 1995) has been implemented in the XPLOR-NIH (Schwieters et al. 2006) and the AMBER (Case et al. 2006) programs. In the program SimShiftDB, a large synthetic database containing protein folds and their predicted chemical shifts is used to derive local conformational restraints (Ginzinger and Coles 2009). Recently, protocols generating protein structures using the chemical shifts only, mostly depending on homology modeling or ROSETTA (Simons et al. 1997) de novo protein modeling, have been published. These methods include CHESHIRE (Cavalli et al. 2007), CS-ROSETTA (Shen et al. 2008) and CS23D (Wishart et al. 2008), all using the chemical shift prediction in different ways to score the predicted protein folds. CS-ROSETTA was later modified to work with incomplete shift assignments (Shen et al. 2009). In the future, novel applications may be found in ligand binding studies. For the protein–protein complexes this has already been done, by combining docking simulations and the CHESHIRE method (Montalvao et al. 2008).

Proteins are not rigid structures and their internal motions, the 4th dimension of their structure, have obviously a major role in their action and function (Klepeis et al. 2009; Smock and Gierasch 2009; Saarela et al. 2002). Therefore, the chemical shifts can be expected to contain invaluable information about the protein 4D-structures (Berjanskii and Wishart 2008). The objective of this work was to develop chemical shift prediction that is based on 4D structures and then explore and characterize its properties. For utilizing the method, a computer program 4DSPOT (**4**-**D**imensional **S**hift **P**rediction: averaged **O**ver **T**ime) was written.

Methods

Overview

The prediction protocol is outlined in Fig. 1. In the first phase of the 4D prediction, molecular dynamics (MD) simulations are performed for protein models with the AMBER 9 program (Case et al. 2006). After that, the 3D protein models in the PDB (Protein Data Bank) format and molecular dynamics trajectories in the AMBER format are input into the 4DSPOT main program. In the first phase, another program named 4DLIB calculates a large number of geometric parameters (dihedral angles, interatomic distances, dipolar terms, etc.). These parameters are averaged over the conformational space of the trajectory and written into library files. The chemical shift prediction takes place in the 4DSPOT main program, where the actual *chemical shift descriptors* (see "Descriptors") are created from the library files and the chemical shifts are calculated.

Both the programs are operated from the 4DSPOT main program. Any graphical molecular modeling software capable of saving protein models in the PDB format can be used in preparing the input structures for the programs. The

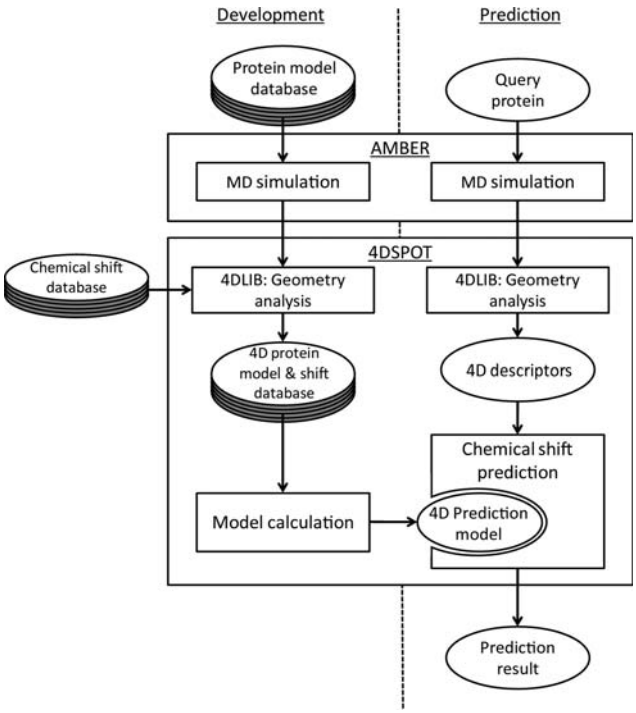

Fig. 1 Information flowchart

prediction results can be output to a text file in 4DSPOT or BMRB (Biological Magnetic Resonance Bank) format, or in a comma separated values (csv) file compatible with most spreadsheet programs. Additionally, the structures and shifts may be exported to PERCH software (www.PerchSolutions.com). For a protein of 100 residues, prediction takes less than one minute on a standard desktop computer (excluding the time needed for MD). More detailed information about the programs can be found in the operating manual of the 4DSPOT package.

Database

A database of 40 proteins (Supplementary material Table S1) was built up using the following criteria: (1) no ligands or paramagnetic atoms are present, (2) chemical shifts are referenced using DSS or TSP, and (3) the NMR structures were recently published as they are expected to be more accurate. The chemical shifts were obtained from BMRB (Ulrich et al. 2008) and the corresponding structure models were downloaded from PDB (Berman et al. 2000). Northeast structural genomics consortium (NESG) was the origin of many (totally 17) recent NMR structures used. The average number of residues in our proteins was 100. Opposite to the other recent approaches, our database was built up mostly from NMR structures (32 NMR vs. 8 X-ray). The possible methodological differences between the NMR structures were not considered as a problem, as all the structures were homogenized during the prediction procedure with the AMBER force field (see "Molecular dynamics"). To achieve the prediction accuracy stated in the results section, the query protein must fulfill the same criteria and conditions that were used in selecting structures to the database. In practice, this means that structures should be monomers of 50–150 amino acid residues, and have no ligands, unnatural amino acids or post-translational modifications.

From the NMR structure ensembles, the conformer used was the "best representative conformer" given in PDB files. Missing atoms (especially in X-ray structures) were added by the leap program of AMBER. The PDB files were analyzed with 4DLIB to create the 4D geometric parameter libraries (LIB files). The chemical shifts were added to the LIB files from the BMRB files. Only the shifts with BMRB's Chemical Shift Ambiguity Index 1 (Uniquely assigned) or 2 (Ambiguity of geminal atoms or methyl proton groups) were approved. Shifts with Ambiguity Index 2 were subjected to an interchange protocol (see "Prediction procedure") during prediction. The observed backbone chemical shifts were re-referenced using the LACS program (Wang and Markley 2009) to prevent biasing of the database by misreferenced shifts. However, this had no significant effects on the results. In addition, pH values of the NMR samples were included in the files.

Molecular dynamics

The MD simulations were done using the AMBER 9 program (Case et al. 2006) and the ff99 force field (Cornell et al. 1995; Wang et al. 2000) augmented with the corrections of Hornak et al. (2006). Protein molecules were solvated by TIP3P water molecules in periodic solvent boxes extending at least 11 Å from the protein atoms. To neutralize the total charge of the simulation systems Na^+ or Cl^- ions were added. The water molecules and hydrogen atoms of proteins were first energy minimized for 500 steps and MD simulation of 11.25 ps at 300 K and at constant volume were done with position restraints of 0.5 kcal/mol-Å^2 on protein heavy atoms. After that the systems were minimized for 500 steps and 11.25 ps MD simulations at 300 K and constant pressure conditions were done with 0.5 kcal/mol-Å^2 position restraints on the backbone atoms. Production simulations of 150 ps were then started. In the simulations the electrostatics were treated using the particle-mesh Ewald method. A timestep of 1.5 fs was used and bonds to hydrogen atoms were constrained to their equilibrium lengths using the SHAKE algorithm. During the 150 ps simulations structures were saved every 0.375 ps. The last 266 of these snapshots were used in averaging the molecular descriptors. The protocol for the 1.0 ns MD simulations was similar to the 150 ps runs except that the lengths of the short initial MD simulations with position restraints were 30 ps and the structures were collected every 2.5 ps for analysis. To model the solvent effects on HN and Hα atoms the average number of water molecules within 5.0 Å (second solvent shell) of the hydrogen atoms was calculated.

Prediction algorithm

In the prediction model, the chemical shift δ_n is expressed by the equation

$$\delta_n = \delta_n^o + \sum P_i < X_i > \quad (1)$$

where δ_n^o is the base value of the chemical shift. These values contain all covalent effects needed, and in fact, they represent the average shifts of each of the 124 observable *proton types* (all the different protons among the 20 natural amino acids). The term $P_i < X_i >$ adds the contribution of the descriptor i ($< X_i >$ = the numerical value of the descriptor averaged over the conformational space, P_i = the weight factor) to the chemical shift. The descriptors can be divided into two groups: (1) terms describing physical effects e.g. Coulombic or van der Waals interactions and magnetic anisotropy, and (2) empirical terms, e.g. torsion angles and solvent effects. Total of 163 3D descriptors are used.

Descriptors

Coulombic and van der Waals—effects

The through-space interactions between atoms are described with 29 descriptors, which are proportional to $1/r^n$ (r = interatomic distance, n = 1–6) and, the descriptors describing Coulombic interactions, also to the atomic charges. Five of the terms are devoted to describe lone electron pair effects. The terms are defined separately for the CH, NH and aromatic protons and, for example, in calculating the proton–carbon interactions, the carbons are divided into aliphatic, carbonyl, carboxylate and aromatic carbons. A cutoff value of 7.5 Å is used for the interactions. The atomic charges are calculated by Allen's method (Allen 1989).

Magnetic anisotropy

Descriptors for magnetic anisotropy are defined for bonds and aromatic rings. The bond anisotropies are described by the dipolar expansion

$$B\frac{1-\cos^2 3\theta_x}{r^3}+A_i\frac{1-\cos^2 3\theta_x}{r_m^3}+A_j\frac{1-\cos^2 3\theta_x}{r_n^3} \qquad (2)$$

where the first B term is the term of bond-type (for C–C, C–H, etc.) and the two A terms define 'atomic corrections to the B term' (i, j = atom type and r_m, r_n = the distance of the atoms m and n from the proton). The angle θ_x is the 'dipolar angle'. The A terms make, for example, the C–O dipolar function asymmetrical. The corresponding expansions for z and y anisotropies are defined for C=O bond, for aromatic C and N, and for peptide C–N bond.

In addition to the contributions of the above bond anisotropy terms, the aromatic ring anisotropy is described by the function

$$S\frac{1-\cos^2 3\theta_z}{R^3}+\frac{T}{R^3} \qquad (3)$$

where the angle θ_z is the dipolar angle between the ring and proton, R is the ring-proton distance and the parameters S and T are defined for 6- and 5-membered rings separately. Total of 68 anisotropic descriptors are included. Cutoff values of 5.0 and 15.0 Å are used for the bond and aromatic ring interactions, respectively.

Torsion angles

Total of 40 terms are used to describe the protein torsion angles. 16 of these are cosine functions ($\cos\phi$ and $\cos^2\phi$) of proton vicinal torsion angles (H-X-X-X). Terms are determined for different bond paths, e.g. H–C_{sp3}–C_{sp3}–C_{sp3}, H–C_{sp3}–C_{sp3}–H or H–C_{sp3}–C_{sp3}–C(=O). Thus, these terms include also the important backbone torsion angles Ψ and Φ for the HN and Hα protons. The next 23 descriptors include sums of cosine functions of four and five bond paths. One side-chain angular term is given for the CH protons.

Neighboring residues

Since HN protons are slightly affected by the residue type of the preceding neighbor (Wang and Jardetzky 2002), 13 terms are devoted to describe such interactions. Similar to proton type descriptors, neighboring residue descriptors are dichotomic truth-values of neighboring amino acids. In order to reduce the number of the terms, some amino acids were grouped together: Ile, Leu, Met and Val form a hydrophobic group; Phe, Tyr and Trp form an aromatic group; Cys and Ser form their own group and for Ala the term is zero.

Solvation

Solvation was modeled in two ways, implicitly and explicitly. Six terms describe the solvation in implicit manner by calculating of the sums $\Sigma 1/r_n$ and $\Sigma 1/r_n^2$, where $1/r_n$ is the distance of the proton from protein atom n and tells about the proton's location in the protein: small values correspond to lateral protons and large values to central protons. These terms are defined for HN, Hα, and side chain protons separately. Other two solvent terms are explicit solvent shell descriptors (see MD methods), containing average amount of water molecules within two solvent shells (5.0 Å), determined only for Hα and HN protons.

Other

Two terms are used to describe the flexibility of proton site. The values are averaged deviations from its torsion angle average value over the conformational space. Other two terms describe CH_2 and CH bond angle strain by deviations from average. There is also one descriptor for pH value. For complete list of descriptors, see Supplementary material Tables S2A–S2F.

Prediction procedure

The weight factors in Eq. 1 are solved using principal component regression (PCR). The prediction model is built up in four stages. In the first phase, the principal components and the calculated shifts are solved straight from the original shift and descriptor matrix, with all the shifts included. However, as no perfect protein structure model exists (Joosten et al. 2009), the most poorly predicted shifts

were assumed to be structural or assignment errors and ignored when creating the model. Therefore, in the second phase, the worst 10% of predictions, calculated separately for each proton type, were removed from the matrix. In the third phase, in order to treat the non-linearity of the model, correlation terms X_iX_j were formed. In PCR, 35 of these new terms were found to be significant and thus remained in the final model. The fourth phase is called *local PCR*. In this phase the model is build up for each *proton class* separately, simply by giving small weight factors for the data from the other proton classes. This yields individual models for each proton class, but still the information about most explicit effects is incorporated into the models from the other classes. Protons are classified to nine groups, following the typical Hα-, Hβ- and HN classification. The rest of the side chain protons are classified to CH_3, CH_2, CH and aromatic groups. The side chain protons bonded to heteroatoms are assigned to XH class and the proline 5-ring CH_2 protons to their own. In the third and fourth phase, the interchange protocol is applied, allowing the predictor to interchange the observed geminal shifts with BMRB Ambiguity Index 2, if the prediction errors of both geminal shifts are consequently decreased. In the end of the four phases, the prediction model contained total of 322 descriptors: 124 for proton types, 163 actual 4D descriptors and 35 correlation terms.

Results and discussion

All the results presented are obtained by using the Leave-One-Out cross validation protocol where the protein shifts are predicted one by one excluding the current protein from the teaching database. A common manner to present results is to omit predicted shifts with an error larger than three standard deviations from the mean (Neal et al. 2003; Shen and Bax 2007). In the following, these values are presented in parentheses. The extraction was done separately for each proton class. All results were calculated using the 150 ps 4D prediction model, unless otherwise stated. Pearson correlation coefficient R was used in describing correlation between the observed and predicted shifts.

General trends and statistics

The overall RMS error for all the observable protons in the proteins was 0.33 (0.29) ppm and the value of the correlation coefficient R was 0.992 (0.994). More diagnostic statistics, given separately for each proton class, are shown in Fig. 2. It is also illustrative to evaluate results by noting that 61% of all the shifts are predicted below 0.20 ppm error and 75% below 0.30 ppm error.

The total correlation coefficient R reflects mainly the large range of chemical shifts but when the R is calculated for each Hα and HN separated by residue type, we get a measure showing how a large part of the secondary shift variation is explained by the model. The R values for Hα and HN are given in Table 1. The average R values are 0.770 (0.792) for Hα and 0.644 (0.677) for HN. These values represent the real efficiency of the prediction method. As expected, HN shift results do not suffer from this evaluation (Table 1 vs. Fig. 2), as they are already almost residue independent. Correlation coefficients of Hα shifts are slightly decreased. The worst correlation coefficient, 0.248, is obtained for glycine Hα, where the $H\alpha^2/H\alpha^3$ assignment is ambiguous in 66% of database shifts, as indicated by an ambiguity index of 2 in BMRB files. Indeed, correlation coefficient for glycine Hα shifts with unambiguous assignments is 0.600, and for those with ambiguity, it is 0.108. However, this does not affect to the prediction model, as the interchange protocol is used. Another problematic residue is proline with Hα shift correlation coefficient of 0.578, probably because of its special location in 5-ring. Among the HN shifts, the residue yielding the weakest results is histidine with correlation coefficient of 0.509, probably arising from the small number of shifts in the database.

Basically, the Hα shifts correlate strongly with torsion angle effects, which make their prediction relatively easy: the RMS error was 0.29 (0.26) ppm with the correlation coefficient of 0.834 (0.855). The largest errors were found in those parts of the proteins where the ring currents cause a strong upfield or downfield shift to the observed shifts and where the 3D structure is also poorly determined. Typically predicted shifts were ca. 1 ppm too large.

As seen from their large RMS error of 0.50 (0.45) ppm and the poor correlation coefficient of 0.655 (0.687), the prediction of HN shifts poses the greatest challenge, as also noticed before (Neal et al. 2003; Shen and Bax 2007). A large part of the HN secondary shift arises from the hydrogen bonding effects, related to hydrogen bond length, on which the HN shifts are r^{-3} dependent (Wagner et al. 1983). Therefore, small inaccuracies in the hydrogen bond lengths are enough to distort the prediction results. More complex hydrogen-bond contributions, like those arising from the hydrogen bond angles (Moon and Case 2007) and cooperative hydrogen bonding (Parker et al. 2006), can be simulated with ab initio methods, but they may not be properly accounted for in molecular mechanics used in this work. The largest HN prediction errors are found in the atoms where the hydrogen bonds to backbone or side chain carboxyl groups are missing in the molecular models, causing the predicted shifts to be at least 1 ppm too small. This is seen in the scatter plot of HN nuclei (Fig. 2).

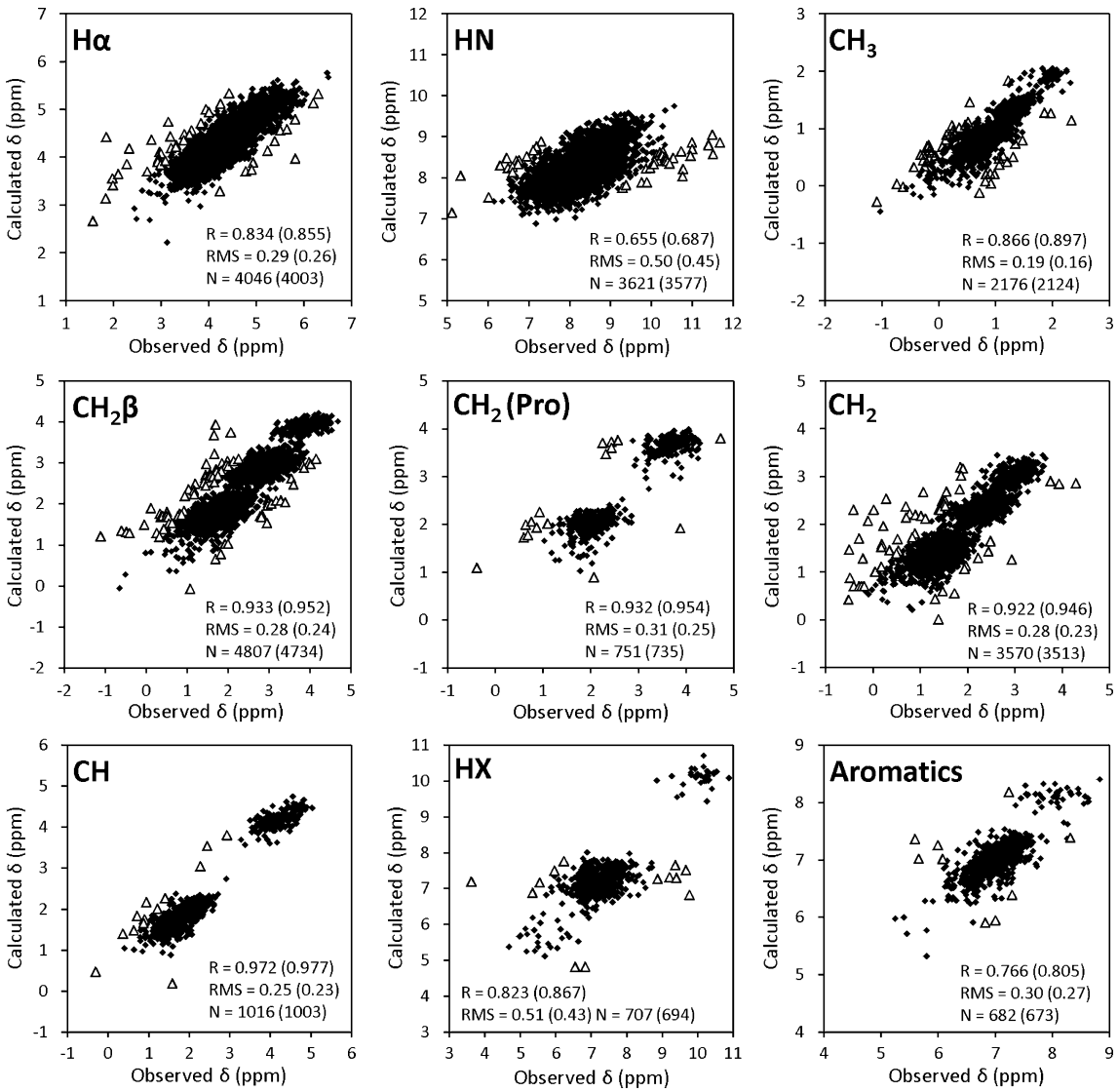

Fig. 2 Prediction scatter plots for each proton class. The values in *parentheses* are calculated with prediction errors more than three standard deviations omitted. In the plots, those predictions are marked with *open triangles*

Side chains yield mainly good prediction results. The combined side-chain RMS error was 0.29 (0.24) ppm with the correlation coefficient of 0.987 (0.991). Mostly this reflects the facts that side chains do not suffer from strained torsion angles, as the backbone shifts, and that their shifts do not differ much from the random coil shifts. Again, largest deviations occur for nuclei under aromatic ring currents. Prediction accuracy for some nuclei in the HX and aromatic classes suffer from shortage of data points, and possibly also from misassigned shifts.

Structural 3D trends

The prediction results for each protein Hα and HN chemical shifts are shown in Fig. 3. The results for different proteins differed notably. The worst protein models yielded RMS errors two times larger than the best one in both Hα and HN shifts. If we assume that the differences between proteins arise from structure qualities, not from prediction, the actual prediction result could be in principle defined as the result for the best protein structure in the database. The standard deviations of the RMS errors are 0.05 and 0.07 ppm for Hα and HN shifts, respectively. To prevent the bad structures impairing the prediction model, 10% of the worst shifts were excluded from the prediction model (see "Prediction procedure"). The percentage was applied for the whole database, not for individual proteins.

The predictions for the Hα and HN shifts were grouped on the basis of their secondary structure type obtained from

Table 1 *R* correlation coefficients for Hα and HN shift for each residue type

	Hα	HN
Ala	0.850 (0.879)	0.687 (0.700)
Cys	0.657 (0.772)	0.669 (0.706)
Asp	0.725 (0.746)	0.611 (0.683)
Glu	0.828 (0.848)	0.631 (0.643)
Phe	0.864 (0.872)	0.687 (0.707)
Gly	0.248 (0.400)	0.607 (0.624)
His	0.841 (0.839)	0.509 (0.573)
Ile	0.866 (0.877)	0.725 (0.755)
Lys	0.872 (0.876)	0.649 (0.669)
Leu	0.874 (0.879)	0.674 (0.698)
Met	0.774 (0.775)	0.760 (0.760)
Asn	0.712 (0.739)	0.600 (0.614)
Pro	0.578 (0.533)	–
Gln	0.844 (0.839)	0.648 (0.648)
Arg	0.868 (0.885)	0.623 (0.665)
Ser	0.808 (0.808)	0.662 (0.709)
Thr	0.812 (0.819)	0.576 (0.600)
Val	0.844 (0.855)	0.663 (0.710)
Trp	0.673 (0.721)	0.594 (0.714)
Tyr	0.862 (0.878)	0.663 (0.697)
Average	0.770 (0.792)	0.644 (0.678)

The values in parentheses are calculated with the prediction errors larger than three standard deviations omitted

the PDB data. HN shift predictions behaved as expected: in the regular secondary structure elements, RMS errors were notably smaller (0.46 and 0.48 ppm for α-helix and β-sheet, respectively) than in the random coil areas (0.55 ppm). For the Hα shifts, however, the β-sheet regions yielded inferior results (0.34 ppm) compared to those of α-helices (0.24 ppm) and random coils (0.29 ppm). Although the β-sheet data is somewhat smaller than that of α-helices, a more probable explanation for poor prediction result is that the variation of the Hα proton orientations is larger in β-sheets than in α-helices. Moreover, the shift dispersion of Hα protons is smaller in regions more exposed to the solvent (Vranken and Rieping 2009), which also facilitates the shift prediction of random coils, usually located on the surface of the protein.

Some weak correlations (correlation coefficients *R* between 0.55 and 0.60) were found between prediction results and miscellaneous protein structure quality indicators. First, as expected, larger structures tend to give less accurate prediction (Supplementary material Fig. S1A). More correlations were found between all shift RMS errors versus percentage of residues within "the most favored" Ramachandran plot region (Fig. S1B), calculated with PROCHECK (Laskowski et al. 1996), and "Packing quality" and "Backbone conformation" *Z*-scores (Fig. S1C, S1D) calculated with WHAT_CHECK (Hooft et al. 1996).

4D effects

Table 2 contains the RMS errors of the non-dynamic (3D) and 4D predictions. For the backbone chemical shifts, the 4D descriptors yielded about 6–7% better RMS errors compared to the 3D model. Increasing the simulation length from 150 ps to 1 ns did not improve the predictions. As the side chains and loop regions do not usually adopt new conformations even in 1 ns simulations, benefits from the time-averaged prediction seems to mostly arise from the mapping of local vibrations, which evens out the effect of anomalous conformations and fixes, among other things, bad aromatic packing and strained folds. Protein-specific comparison of the 3D and 4D predictions is shown in Fig. 4. In almost all the cases, for both the Hα and HN shifts, the prediction RMS error decreases or stays unchanged. However, in some cases, notable improvements, up to 0.11 ppm in both the Hα and HN shift RMS errors (28 and 23%, respectively), are observed. The RMS error improvement with the 4D model did not correlate with protein size, initial RMS error (those of non-dynamic structures) or whether the protein was NMR or X-ray structure.

An example of remarkable HN shift prediction improvement is displayed in Fig. 5, where NH shift RMS errors for the 3D and 4D predictions for protein BlaI (PDB 1P6R) are shown as the function of the protein sequence. Noticeably, most of the prediction errors are negative, meaning too small predicted shifts, usually caused by erroneous hydrogen bonds. When this case was further analyzed, it was noticed that the whole original structure seems to be too tightly packed. As the structure is loosened during MD simulation, the prediction errors decrease. From the eight HN nuclei where prediction improvement was more than 0.4 ppm, seven explicit hydrogen bond breakages, in residues E13, V17, I18, T36, W39, F66 and I72, were observed. In residue L49, the hydrogen bond is not fully broken, and improvements probably arose from torsion angle changes.

Large deterioration of HN shift prediction was found in XC975 protein (PDB 1XS3, Fig. 4). When this case was investigated, one large error, accounting for most of the poor RMS error, was found near the flexible N-terminal of XC975, arising from the hydrogen bond formed during MD simulation. Another problematic area was the flexible loop region between V53 and P57, which contains two more new errors arising from MD. Sometimes these kind of errors are found when large conformational changes, larger than usual local vibration, take place during MD simulation and 150 ps is not long enough to properly map

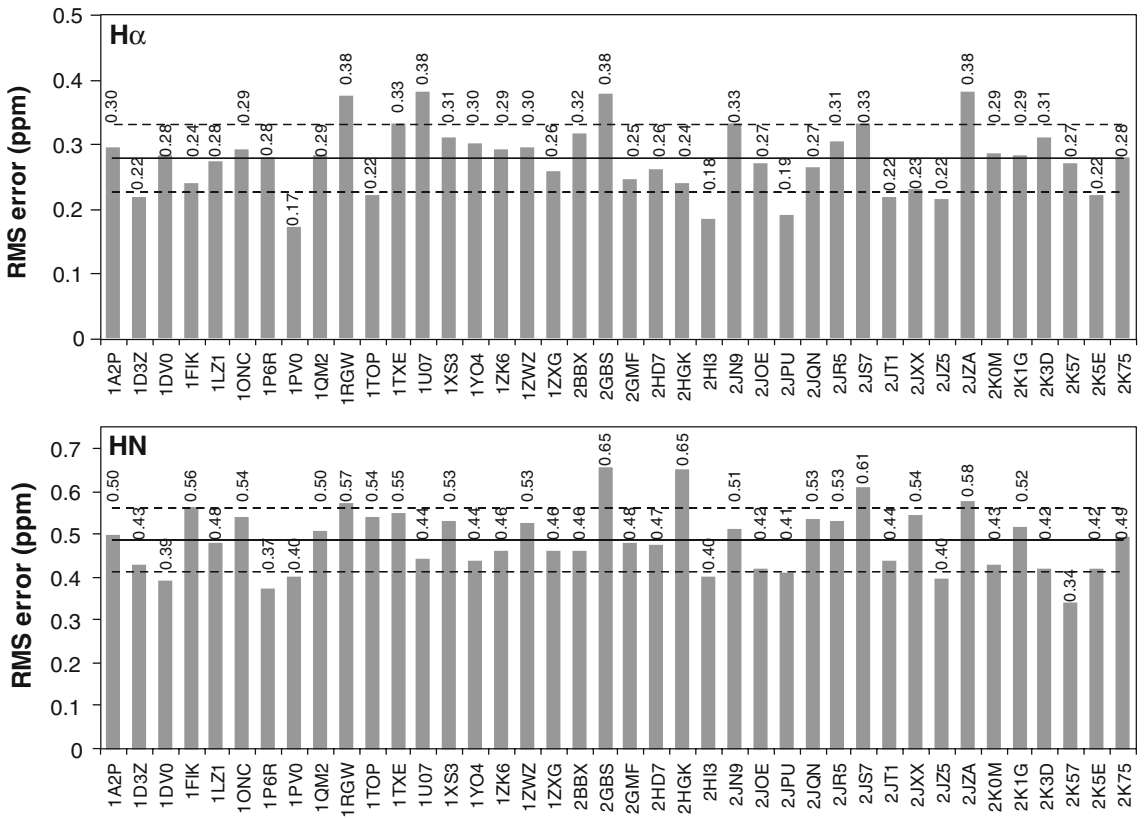

Fig. 3 Prediction RMS errors of Hα and HN shifts in different proteins, with all the shifts included. The *solid line* is the average RMS error and the *broken lines* are the standard deviations of RMS errors

Table 2 RMS errors of 3D and 4D predictions

Prediction model	Hα	HN	Side chain
3D (non-dynamic)	0.31 (0.28)	0.53 (0.48)	0.29 (0.24)
4D (0.15 ns)	0.29 (0.26)	0.50 (0.45)	0.29 (0.24)
4D (1.0 ns)	0.29 (0.26)	0.50 (0.45)	0.28 (0.23)

The values in parentheses are calculated with the prediction errors larger than three standard deviations omitted

conformations of these regions. Thus, if these kind of regions (with large RMSD from initial structure) are found, the results should be taken with due caution, or longer MD runs calculated. Indeed, when predicted from 1 ns MD simulation, XC975 yields 0.03 ppm smaller HN shift RMS error than from the 150 ps database.

Due to the fact the MD protocol samples only a limited part of the conformational space it gives rise to an uncertainty to the 4D prediction. To evaluate the uncertainties arising from MD simulation itself, five parallel 150 ps simulations were performed for two proteins: ubiquitin (PDB 1D3Z), which is a small, tightly folded and well-known protein structure and twinfilin (PDB 2HD7), which is larger and less rigid structure. For the individual shifts, average standard deviations between calculations for ubiquitin were 0.05 and 0.07 ppm for Hα and HN shifts, respectively. For twinfilin, they were 0.07 ppm (Hα) and 0.11 ppm (HN). However, considering total RMS errors, standard deviations were below 0.01 ppm for ubiquitin and 0.01 ppm (Hα) and 0.02 ppm (HN) for twinfilin. Assuming that the total uncertainty S^2_{tot} is composed of two parts

$$S^2_{\text{tot}} = S^2_{\text{model}} + S^2_{\text{MD}} \tag{4}$$

where S^2_{model} is the variance related to the prediction model and S^2_{MD} is the variance arising from the 4D calculation. For ubiquitin the Hα we get $0.220^2 = S^2_{\text{model}} + 0.05^2$ and, thus, $S_{\text{model}} = 0.214$, which means that S^2_{MD} has no major significance in total statistics. For comparison of 3D and 4D models we may assume that the total uncertainty of the 4D model is composed of three parts

$$S^2_{4D} = S^2_{3D} - \left(S^2_{\text{model}} - S^2_{\text{MD}}\right) \tag{5}$$

where S^2_{3D} is the variance of non-dynamic model. For ubiquitin $S_{\text{3D}} = 0.27$ ppm and using the above values we get $S_{\text{4D}} = 0.18$ ppm. This gives an estimate of the contribution of the 4th dimension to the shifts. For ubiquitin HN shifts, the corresponding value is 0.16 ppm and for twinfilin, the values are 0.15 and 0.14 ppm for Hα and HN shifts, respectively.

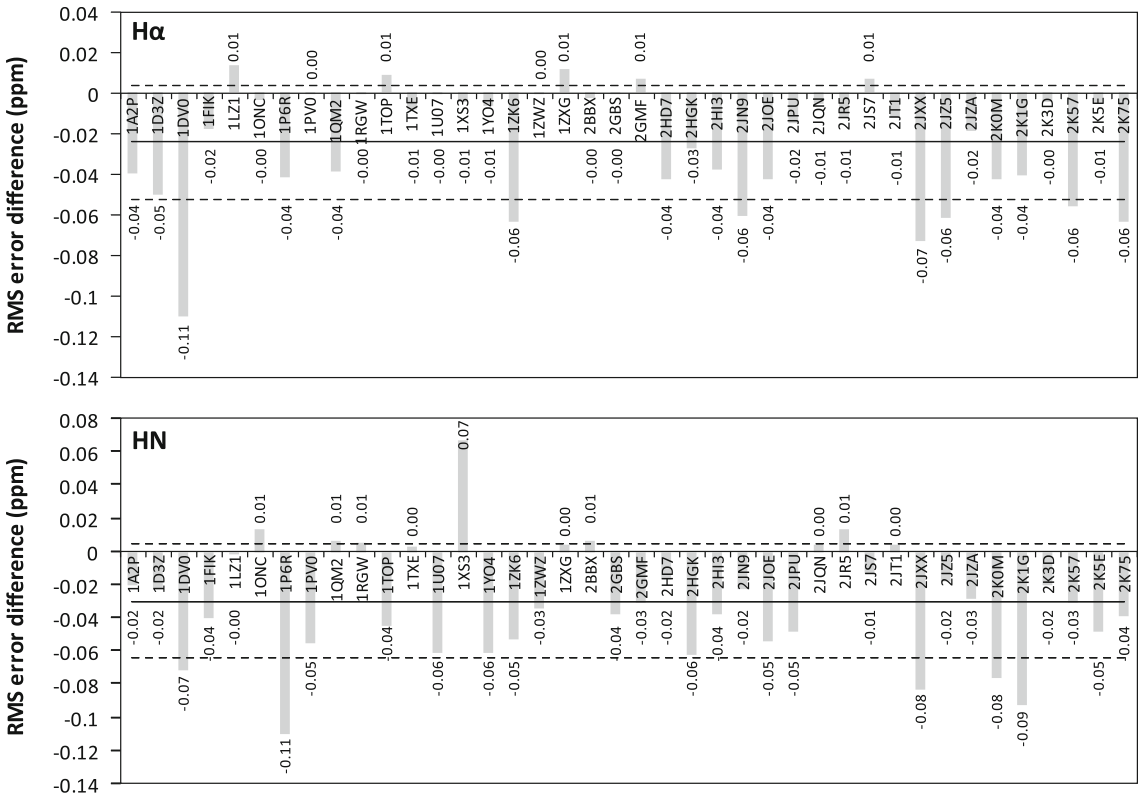

Fig. 4 Protein-specific comparison of non-dynamic 3D and 150 ps 4D models for Hα and HN shifts. The *bars* indicate 150 ps 4D model RMS errors compared to the 3D model prediction. All the shifts were included in this comparison. The *solid line* is average RMS error and the *broken lines* show the standard deviation of the RMS errors

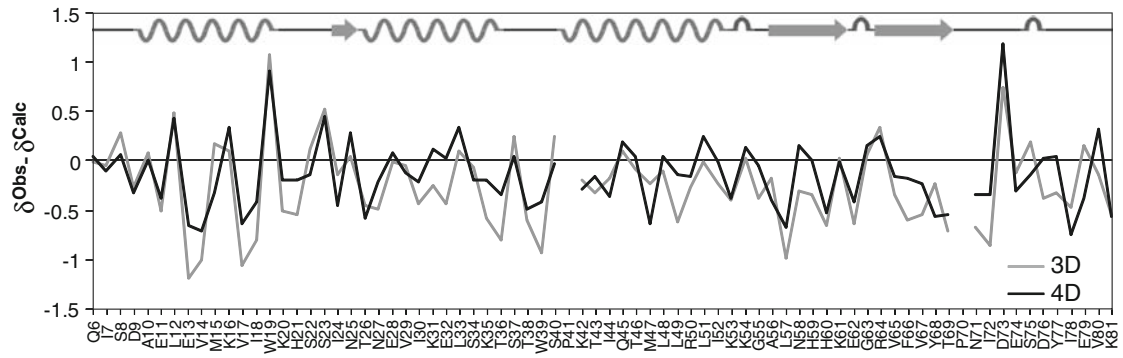

Fig. 5 3D and 4D prediction errors ($\delta^{obs} - \delta^{calc}$) for BlaI HN shifts shown as functions of the sequence. Secondary structure scheme, showing helices, sheets and turns, is the PDB SEQRES sequence

Prediction was cross-tested by predicting 3D structures with 4D database and vice versa. For example, ubiquitin yields HN shift RMS errors of 0.46 ppm when predicting the 150 ps structure with the 3D prediction model, and 0.57 ppm when predicting the non-dynamic structure with the 150 ps model, compared to the original results of 0.43 ppm for the 150 ps structure and 0.45 ppm for the non-dynamic structure. Similar results were observed with other proteins, meaning that the prediction model used in prediction should correspond to the MD method used with the query protein. Especially, non-dynamic structures should not be predicted with the dynamic databases, as this leads to extrapolation problems. The present 4DSPOT software package provides separate models for non-dynamic, 150 ps and 1 ns predictions.

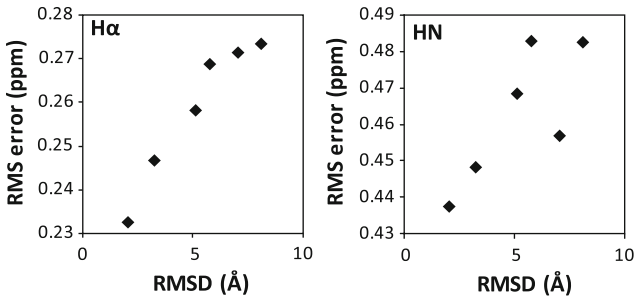

Fig. 6 RMS error of ubiquitin Hα and HN shift prediction versus RMSD from initial conformation during thermal denaturation simulation

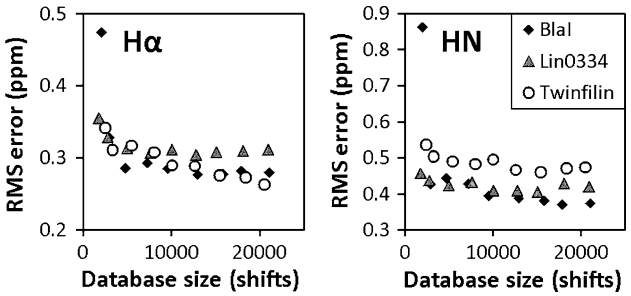

Fig. 7 Plots of Hα and HN prediction RMS errors of three proteins versus teaching database size (all shifts)

Properties of 4D prediction

Extrapolation to zero RMSD

To explore the sensitivity of the prediction RMS error to the quality of the structure, we performed a protein denaturation simulation. Ubiquitin (PDB 1D3Z) was first simulated in 300 K for 150 ps and then heated from 300 to 500 K in 750 ps. Trajectory was then divided to 150 ps fragments and average RMSD (Root Mean Square Distance), compared to the initial conformation, was calculated for each fragment. After that, the chemical shifts for each fragment were predicted. Plot of the Hα shift prediction RMS error versus RMSD (Fig. 6) shows how the RMS error grows when the native protein structure is broken. The HN shifts behaved in a similar way, however, with one outlier, caused by a momentarily better conformation during denaturation.

The experiments reflect the incompleteness of the prediction and protein structure models: if we extrapolated the functions to RMSD of zero, the prediction RMS error would still be significant, approximately 0.22 ppm for Hα and 0.42 for HN whereas the values obtained using structure with 10 Å resolution would be 0.28 and 0.50 ppm. However, if sequence-corrected random coil shifts (Schwarzinger et al. 2001) are used instead of predicted shifts, the RMS values would be 0.50 and 0.62 ppm, which simply means that the 10 Å structure still has considerable local structure left.

Database size

Database size is usually considered critical in the chemical shift prediction. Using three proteins models as indicators, prediction was tested with reduced teaching databases, from five proteins to the whole database of 40 proteins. In each test, the proteins in reduced databases were randomly selected. As can be seen from Fig. 7, the prediction cannot be further improved, at least significantly, by increasing the database size. Remembering the large variation between different proteins, it seems that the quality of the structures in the database is more important than the database size. For the tested proteins, the critical number of data points for efficient prediction seems to be about 2,000 for HN shifts and 1,000 for Hα shifts. However, for some rare amino acids like tryptophan, problems may rise earlier. Very small number of data points may cause mathematical problems and then abnormally large errors, as seen in the case of BlaI protein. Compared to sequence homology methods, like SPARTA (Shen and Bax 2007), where 200 protein database is used, PCR seems to leverage to information contained in the database quite effectively.

Contributions of descriptors

To explore the importance of our descriptors, prediction was done selectively with certain descriptor classes omitted. The RMS errors for Hα, HN and side chain protons are shown in Table 3. In the first row, the RMS error is calculated without any other descriptors but the base shift values of the proton types (δ_n^o). For comparison, plotting random coil shifts (Wishart et al. 1995) versus experimental shifts gives very similar results for Hα and HN nuclei: 0.49 and 0.67 ppm, respectively. On the other way round, predicting shifts with all other descriptors but proton type group, the HN shift result was still good, as they are practically independent of residue type. Similarly, Hα results are about 25% worse without the proton types defined, which fits nicely to the estimate that 25% of the Hα shift arises from residue type (Wishart and Case 2001). The side chain shifts are less sensitive to folding, as their prediction result is fair using nothing but proton type descriptors and improves only slightly with additional 3D descriptors.

Bond anisotropy has the most significant effect to Hα and HN. However, using torsion angle terms instead of bond anisotropy yields quite similar results. This is due to the fact that as far as short-range effects are considered they, in practical terms, contain the same information, only

Table 3 Contributions of the shift descriptors

Descriptor classes used						Hα	HN	Side chain
Proton type	Bond anisotropy	Aromatic anisotropy	Torsion angles	Coulombic and vdW	Remainder[a]			
X						0.48	0.66	0.33
	X	X	X	X	X	0.36	0.52	0.38
X	X					0.34	0.52	0.32
X		X				0.44	0.61	0.29
X	X	X				0.31	0.51	0.29
X			X			0.36	0.56	0.33
X				X		0.43	0.53	0.31
X	X	X		X		0.31	0.50	0.29
X	X	X	X	X		0.29	0.50	0.29
X	X	X	X	X	X	0.29	0.50	0.29

The values are prediction RMS errors in ppm with no bad predictions excluded

[a] Contains descriptor groups "neighboring residues", "solvation" and "other"

expressed in different way. For example, within the same residue, the effect of backbone C=O bond to HN proton can be expressed by describing the backbone torsion angles Φ and Ψ, or directly describing the spatial distance and orientation of the carbonyl bond to the HN proton, as done in our anisotropic terms. Aromatic ring and lone pair terms have smaller contribution to backbone shifts, because all protons are not affected by them. The Coulombic and vdW effects greatly improve the amide proton prediction, because hydrogen bond effects, known to be crucial to HN shifts, are largely carried by these descriptors.

Comparison of NMR and X-ray structures

For eight proteins, both NMR and X-ray structures were downloaded and chemical shifts were predicted with the 150 ps 4D database. Results for Hα and HN shifts are shown in Table 4. On average, X-ray structures gave somewhat better results for the both backbone proton classes. However, this is not the case for all the tested proteins and, moreover, the results of Hα and HN shifts do not always correlate, as can be seen from the results of P-30 and Ton-B.

The differences in the total RMS errors of the proteins (Table 4) were not statistically significant (Paired samples *t*-test's 2-tailed *p*-values were 0.150 and 0.051 for Hα and HN shifts, respectively). When calculated from individual shifts, the RMS errors were 0.30 and 0.53 ppm for Hα and HN of X-ray structures versus 0.32 and 0.56 ppm for Hα and HN of NMR structures, with the corresponding *p*-values being <0.001 and 0.009 for Hα and HN shifts, respectively. This indicates that the backbone shift RMS errors of NMR and X-ray structures differ significantly ($p < 0.05$) and confirms that the X-ray structures carry information about the local structures better, despite the fact that some structural differences between NMR and X-ray structures are reported (Andrec et al. 2007).

Comparison to other methods

A set of 10 proteins, seven NMR and three X-ray structures, was predicted with the programs 4DSPOT, SHIFTX

Table 4 Comparison of NMR and X-ray structures

	PDB ID (NMR)	PDB ID (X-ray)	Hα		HN	
			NMR	X-ray	NMR	X-ray
Barnase	1FW7	1A2P	0.31	0.30	0.50	0.50
P-30	1PU3	1ONC	0.30	0.29	0.53	0.54
Profilin	1PFL	1FIK	0.28	0.24	0.62	0.56
Ton-B C-terminal domain	1XX3	1U07	0.34	0.38	0.47	0.44
Troponin C	1BLQ	1TOP	0.27	0.22	0.56	0.54
Ubiquitin	1D3Z	1UBQ	0.22	0.23	0.43	0.44
Pyrophosphokinase (BMRB 4299)	2F63	1HKA	0.33	0.30	0.56	0.52
Ribonuclease SA (BMRB 4259)	1C54	1RGE	0.46	0.38	0.71	0.66
Average			0.31	0.29	0.55	0.52

The values are RMS errors (in ppm) for all the predicted shifts

Table 5 Comparison of 4DSPOT with SHIFTX and SPARTA

Protein	Hα			HN			Side chain	
	4DSPOT	SPARTA	SHIFTX	4DSPOT	SPARTA	SHIFTX	4DSPOT	SHIFTX
A219	0.26	0.32	0.32	0.45	0.66	0.68	0.25	0.41
NESG PsR76A	0.27	0.36	0.32	0.34	0.35	0.44	0.20	0.25
HOP	0.19	0.26	0.30	0.40	0.46	0.55	0.30	0.42
MyD88	0.33	0.34	0.40	0.55	0.56	0.60	0.32	0.37
hGM-CSF[a]	0.25	0.32	0.29	0.37	0.48	0.45	0.31	0.34
Snu13p[a]	0.30	0.31	0.30	0.48	0.47	0.46	0.25	0.32
Atu0742 (PDB 2K54, BMRB 15823)[b]	0.35	0.41	0.43	0.60	0.60	0.59	0.37	0.44
IL-15Ra (PDB 2ERS, BMRB 6882)[b]	0.36	0.45	0.54	0.49	0.72	0.70	0.36	0.45
MSP (PDB 1XHH, BMRB 5565)[b]	0.38	0.39	0.40	0.43	0.54	0.60	0.36	0.39
Dsk2p UBL (PDB 2BWF, BMRB 15769)[a,b]	0.34	0.32	0.36	0.35	0.42	0.39	0.29	0.34
Average (all)	0.30	0.35	0.37	0.45	0.53	0.55	0.30	0.37
Average (NMR)	0.31	0.36	0.39	0.47	0.56	0.59	0.31	0.39

The values are RMS errors in ppm

[a] X-ray structure. Resolutions are 2.4, 1.9 and 1.15 Å for hGM-CSF, Snu13p and Dsk2p, respectively

[b] Not belonging to the 40 proteins of the 4DSPOT teaching database

(Neal et al. 2003) version 1.1 and SPARTA (Shen and Bax 2007) version 2008.02.11. None of the predicted proteins were included in SHIFTX or SPARTA teaching databases. The 4DSPOT results were validated with Leave-One-Out protocol, where the protein shifts are predicted one by one excluding the current one from the teaching database and moreover, four of the proteins were never used in the 4DSPOT teaching database. The set of 10 proteins present good, average and poor RMS errors of 4DSPOT prediction. If certain shifts were not predicted in one of the programs (e.g. terminal residue shifts in SPARTA and aromatic side chain protons in SHIFTX), they were left out in other program results too. No bad predictions were excluded from the results, except over 1.5 ppm errors of same sign predicted by each of the programs.

The results for Hα, HN and side chain protons are presented in Table 5. Compared to SPARTA and SHIFTX, Hα and HN RMS errors are ca. 15% smaller in 4DSPOT. As the benefit of 4D prediction is 6–7%, it accounts for about half of this margin. Another half probably arises from the use of PCR and the extensive selection of molecular descriptors, instead of simpler chemical shift hypersurfaces or sequence homology methods. Shen and Bax published RMS errors for NMR structures (0.37 ppm for Hα and 0.54 ppm for HN) very close to our results with SPARTA, indicating validity of this comparison. In the side chain prediction, 4DSPOT was ca. 20% better than SHIFTX for all structures. For the X-ray structures, 4DSPOT gives very similar results compared to the other two programs. As no significant development in prediction results for X-ray structures is recently presented, it is probable that with this accuracy in databases, the results are already as good as they can be.

Conclusions

In this work, an empirical ^{1}H chemical shifts prediction protocol based on protein 4D structure was developed and assessed. With the inclusion of the 4th dimension we expected to obtain a more realistic picture about the protein structures. However, at the same time we created a new source of uncertainty arising from the MD calculations. In our cases, the uncertainty created in this way was 0.05–0.11 ppm for individual shifts, depending from quality of the initial structure and the actual dynamics of the protein. In spite of that, the inclusion of the 4th dimension led on average to 6–7% reduction of total RMS error, which suggests that the 4D contribution to the ^{1}H shifts is ca. 0.16 ppm. A considerable part of this benefit can be accounted for the averaging of aromatic ring conformations. The χ^{1} torsion angle of phenylalanine and tyrosine is rather flexible and the 3D model cannot describe these structures well. Another source of improvement is due to averaging of the hydrogen bonds. Moreover, in some initial structures there seems to be strained regions, which are then released in MD simulation. In general, our PCR protocol for protein ^{1}H chemical shift prediction appeared to be at least as effective as those based on sequence homology (Shen and Bax 2007), chemical shift hypersurfaces (Neal et al. 2003; Wang 2004) and neural networks

(Meiler 2003). The 4DSPOT prediction model is competent even without the 4th dimension.

The observation that the X-ray structures gave better prediction than the corresponding NMR structures confirms that the chemical shifts contain invaluable information about local structures, which are obviously better defined in X-ray structures. The NMR solution tertiary structures which often differ from solid state structures (Andrec et al. 2007) are not well defined by the shifts. The same conclusion was also obtained in the thermal denaturation simulation, where the RMS error increased surprisingly slowly as structure broke up. Due to the local nature of the chemical shift information, they offer a unique way to study local dynamics.

The overall RMS errors (from 0.29 ppm for Hα to 0.50 ppm for NH protons) of all the present prediction methods propose that the models are able to predict only a rather modest fraction of the 3D and 4D effects. When those effects are completely ignored, the RMS values for proteins in the 4DSPOT database are 0.48 and 0.66 ppm, respectively. However, the rather large range of prediction RMS errors for different proteins, from 0.17 to 0.38 ppm for Hα shifts and from 0.34 to 0.65 ppm for NH shifts, suggests that there are large variations in the quality of structures. We are not able to propose any single way to improve these results, but we may only conclude that the present models for both the protein structures and the chemical shifts need to be adjusted. At the positive side is the thought that the chemical shifts offer an obvious way to improve protein models and, in addition, the force fields. Thus, the obvious next step to take is the development of the chemical shift based structure refinement protocols, including the dynamic effects shown to be important in this paper.

Software availability

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

References

Allen LC (1989) Electronegativity is the average one-electron energy of the valence-shell electrons in ground-state free atoms. J Am Chem Soc 111:9003–9014

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

Avbelj F, Kocjan D, Baldwin RL (2004) Protein chemical shifts arising from alpha-helices and beta-sheets depend on solvent exposure. Proc Natl Acad Sci U S A 101:17394–17397

Berjanskii MV, Wishart DS (2008) Application of the random coil index to studying protein flexibility. J Biomol NMR 40:31–48

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

Case DA, Darden TA, Cheatham TE, Simmerling CLI, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong KF, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Monga JN, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, Kollman PA (2006) AMBER 9, University of California, San Francisco

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

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J Am Chem Soc 117:5179–5197

de Dios AC, Pearson JG, Oldfield E (1993) Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. Science 260:1491–1496

Ginzinger SW, Coles M (2009) SimShiftDB; local conformational restraints derived from chemical shift similarity searches on a large synthetic database. J Biomol NMR 43:179–185

Hooft RWW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381:272

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

Joosten RP, Salzemann J, Bloch V, Stockinger H, Berglund A, Blanchet C, Bongcam-Rudloff E, Combet C, Da Costa AL, Deleage G, Diarena M, Fabbretti R, Fettahi G, Flegel V, Gisel A, Kasam V, Kervinen T, Korpelainen E, Mattila K, Pagni M, Reichstadt M, Breton V, Tickle IJ, Vriend G (2009) PDB_REDO: automated re-refinement of X-ray structure models in the PDB. J Appl Crystallogr 42:376–384

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

Kuszewski J, Gronenborn AM, Clore GM (1995) The impact of direct refinement against proton chemical shifts on protein structure determination by NMR. J Magn Reson B 107:293–297

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486

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

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

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

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

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

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

Saarela JTA, Tuppurainen K, Perakyla M, Santa H, Laatikainen R (2002) Correlative motions and memory effects in molecular dynamics simulations of molecules: principal components and rescaled range analysis suggest that the motions of native BPTI are more correlated than those of its mutants. Biophys Chem 95:49–57

Schwarzinger S, Kroon GJA, Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123:2970–2978

Schwieters CD, Kuszewski JJ, Clore GM (2006) Using xplor-NIH for NMR molecular structure determination. Prog Nucl Magn Reson Spectrosc 48:47–62

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, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A 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

Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 268:209–225

Smock RG, Gierasch LM (2009) Sending signals dynamically. Science 324:198–203

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

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

Wagner G, Pardi A, Wuthrich K (1983) Hydrogen bond length and proton NMR chemical shifts in proteins. J Am Chem Soc 105:5948–5949

Wang Y (2004) Secondary structural effects on protein NMR chemical shifts. J Biomol NMR 30:233–244

Wang Y, Jardetzky O (2002) Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc 124:14075–14084

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

Wang JM, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comput Chem 21:1049–1074

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

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

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