J|A|C|S COMMUNICATIONS

Published on Web 09/09/2009

Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances

Kai J. Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo*

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, U.K.

Received May 8, 2009; E-mail: mv245@cam.ac.uk

Using chemical shifts for protein structure determination has been a long-standing goal in structural biology, since these NMR observables are measurable under very general conditions and with great accuracy.[1,2] One major obstacle, however, has been the difficulty to understand in sufficient detail the complicated conformational dependencies of the chemical shifts. Since the early recognition that chemical shifts can be closely associated with secondary structure elements,[3] accurate methods have been developed to use them to define the values of backbone dihedral angles.[4] To extend these predictions to obtain complete tertiary structures, it is key to first be able to solve the inverse problem - the prediction of the chemical shifts corresponding to a given structure. This subject has recently been studied intensively, and several methods have become available for this purpose.[5−8] With such tools it has become possible to search the conformational space of proteins to find structures whose predicted chemical shifts closely match the experimentally measured ones. These developments have led to a series of methods that enable the determination of the structures of proteins and of protein complexes from chemical shifts at a resolution often comparable to that provided by more standard NMR methods.[9−14] The current expectation is that further advances in chemical shift based structure determination could be made by increasing the accuracy and speed of the predictions of the chemical shifts.

In this work we present a method, CamShift, in which the complex conformational dependence of the chemical shifts is approximated formally as a polynomial expansion of the interatomic distances defining the structure of the protein. The chemical shift of a given atom a is thus expressed in terms of a set of distances between atom pairs (Figure 1).

$$\delta_a^{pred} = \delta_a^{rc} + \sum_{b,c} \alpha_{bc} d_{bc}^{\beta_{bc}} \qquad (1)$$

In eq 1, δ_a^{pred} is the predicted chemical shift of atom a, δ_a^{rc} is its random coil chemical shift, and d_{bc} is the distance between atoms b and c; the sum is extended to a series of atom pairs in the vicinity of atom a, including atom a itself. The α_{bc} and β_{bc} parameters depend on the atom and residue types; the full list of these parameters and their numerical values are provided as Supporting Information (SI, Table S6), together with the list of atom pairs over which the sum in eq 1 is carried out. The atom types include the atomic species (H, C, O, N, and S), the type within the residue (Cα, Cβ, etc.), the residue type (Ala, Val, etc.), and the hybridization state. We considered two types of distances, depending on whether atoms are covalently bonded or not. In the former case, the β_{bc} parameters are set to 1; in the latter we use two separate terms, with the β_{bc} parameters set to 1 and −3, respectively.

CamShift can also optionally consider three further specific contributions to the chemical shifts: backbone dihedral angles, H-bonding, and aromatic ring currents. The H-bonding term was implemented using an approach by Baker and co-workers[15] (see SI), and for ring currents we used the point−dipole method[16] (see SI).

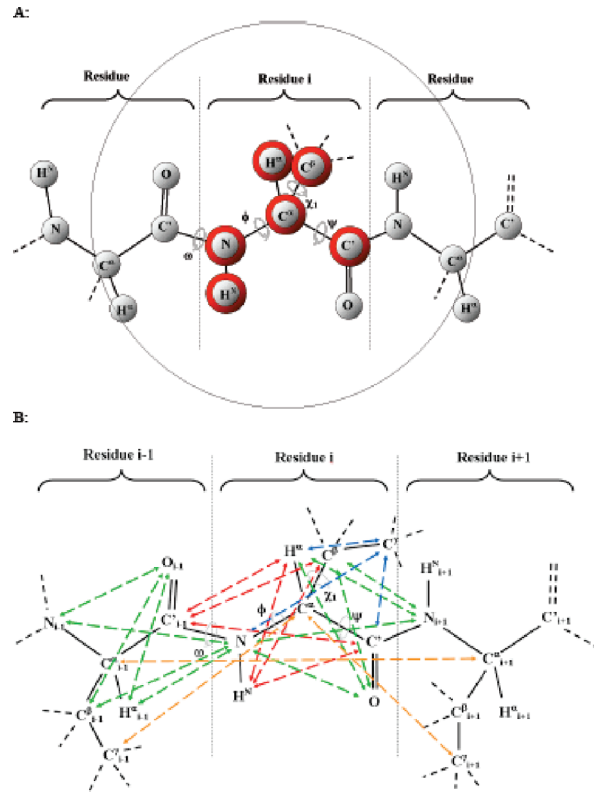


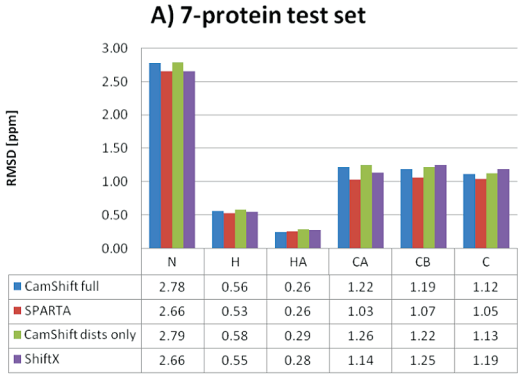
Figure 1. Illustration of the distances used in the CamShift predictions. We show the amino acid triplet centered around residue i, which contains the query atom (in this case the Cα atom); red circles indicate the atoms for which CamShift can currently perform predictions. (A) Interatomic distances are considered from the query atom to the backbone atoms shown, as well as to the side-chain atoms of residue i. In addition, nonbonded interactions are included for a series of atoms within a sphere of a 5 Å radius, indicated by the thin black circle around the query atom. (B) A set of additional distances is included independently from the query atom to better capture the φ, ψ, and χ_1 dihedral angle dependencies and side-chain orientations. CamShift is freely available by uploading structures in PDB format to a web server (http://www-vendruscolo.ch.cam.ac.uk/software.html).

The parameters in eq 1 and those for dihedral angles, ring currents, and H-bonding were fitted by maximizing the agreement between predicted and experimental chemical shifts for a set of proteins for which both structures and chemical shifts are known experimentally. We used the RefDB[17] database of chemical shifts and corresponding Protein Data Bank (PDB) structures, from which we extracted a total of 224 036 chemical shifts for Hα, Cα, Cβ, C′, H_N, and N backbone atoms. In creating the database we considered only structures derived from X-ray procedures with a resolution of 2.3 Å or better. As most of the X-ray structures in the PDB do not contain the positions of hydrogen atoms, these were added using the all-atom molecular simulations package

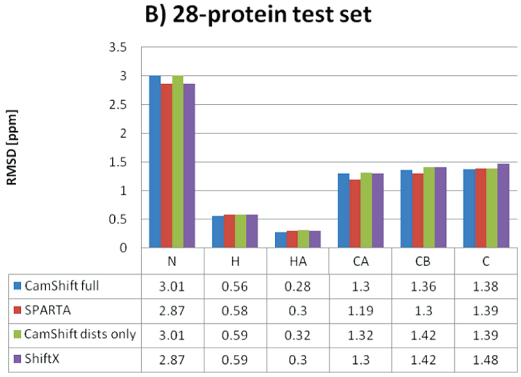
10.1021/ja903772t CCC: $40.75

almost[9] (available for download at http://open-almost.org), in accordance with the CHARMM22 topology file.[18] We left out most distances with very narrow distributions around their mean values, which could lead to numerical instabilities in the prediction of extreme outliers (see SI). We also left out distances to atoms that are unlikely to contain accurate structural information because of their dynamics, such as those involving methyl and hydroxyl groups or amino H-atoms of side chains. Distances to atoms for which different stereochemical conventions might result in inconsistencies between different force fields, such as the branched γ carbons in Val residues or the branched δ carbon atoms in Leu residues, were fitted together with a single term with average distances.

The results of the chemical shift predictions for Hα, Cα, Cβ, C′, H_N, and N atoms are summarized in Figure 2, where we compared the distance-based predictions provided by eq 1 with those obtained by also using the contributions from disulfide bridges, dihedral angles, ring currents, and H-bonding. We found that the inclusion of these further terms improved slightly the quality of the CamShift predictions. We also present comparisons with SHIFTX[6] and SPARTA[8], which are two state-of-the-art chemical shift predictors. All predictors were tested on two test sets. The first set consists of seven structures previously used to compare SHIFTX and SPARTA[8]. We excluded two of the nine structures that were used in the original study,[8] because no BMRB record was defined (GB3) or an almost identical structure is contained in the CamShift and SPARTA training databases (3CBS and 1CBS with 0.4 Å backbone rmsd). Specific comparisons for each atom type in each protein are reported in Table S3. The second test comprises 28 structures from the RefDB database that were not used in the CamShift fit and are not homologues (according to the ASTRAL SCOP classification[19]) to any of the structures in the CamShift, SHIFTX, or SPARTA training data sets. The 28-protein test set therefore reduces the relative contributions from structural homology. The results (Figure 2) show that, considering both the 7-protein and the 28-protein test sets, CamShift and SPARTA provide an overall similar accuracy, although SPARTA seems to provide better predictions for C and N atoms, and CamShift for H atoms. Both methods provide a marginally better accuracy than that of SHIFTX. For both test sets, the accuracy achieved by the distance-only version of CamShift is closer to that of SHIFTX than that of SPARTA.



	N	H	HA	CA	CB	C
CamShift full	2.78	0.56	0.26	1.22	1.19	1.12
SPARTA	2.66	0.53	0.26	1.03	1.07	1.05
CamShift dists only	2.79	0.58	0.29	1.26	1.22	1.13
ShiftX	2.66	0.55	0.28	1.14	1.25	1.19



	N	H	HA	CA	CB	C
CamShift full	3.01	0.56	0.28	1.3	1.36	1.38
SPARTA	2.87	0.58	0.3	1.19	1.3	1.39
CamShift dists only	3.01	0.59	0.32	1.32	1.42	1.39
ShiftX	2.87	0.59	0.3	1.3	1.42	1.48

Figure 2. Comparison between the predictions provided by different methods: SHIFTX[6], SPARTA[8], and two variants of CamShift (with all contributions included and with interatomic distances only). The comparison is made in terms of the root-mean-square deviation (rmsd) between experimental and predicted chemical shifts.

In comparing the performances of the different methods we observe that the results may depend considerably on the particular set of proteins used for validation. We found that by repeating the calculations for ten subsets of seven proteins extracted from the 28-protein test set, the results showed a variability ranging from 7% for H_N atoms to 38% for C′ atoms (Table S4). These rather large variations in the accuracy of the predictions can also be observed from Table S3, which presents detailed results for each protein in the 7-protein test set. We also considered the quality of the predictions in different secondary structure elements, which revealed that there are systematic differences. In all the prediction methods that we considered, chemical shift predictions were better in α helices than in β strands, and predictions in α helices and β strands were much more accurate than those in turns and coil (Table S5).

In summary, we have described the CamShift method for predicting protein chemical shifts, which was introduced to have a prediction procedure based on a differentiable function of the atomic coordinates of a protein. This aspect makes the CamShift predictions very rapid and suitable to define chemical shift restraints in molecular dynamics simulations. We thus anticipate that the use of CamShift will enable the determination of the structures of proteins from chemical shift information in a similar manner in which other standard NMR observables, such as NOEs, scalar couplings, and residual dipolar couplings, are used.

Acknowledgment. This work was supported by grants from Microsoft Research, the Gates Cambridge Trust, the European Union, the Leverhulme Trust, EMBO, and the Royal Society.

Supporting Information Available: Materials and methods. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Wuthrich, K. *Science* **1989**, *243*, 45–50.
(2) Wishart, D. S.; Case, D. A. *Methods Enzymol.* **2001**, *338*, 3–34.
(3) Pastore, A.; Saudek, V. *J. Magn. Reson.* **1990**, *90*, 165–176.
(4) Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A. *J. Biomol. NMR* **2009**, *44*, 213–223.
(5) Xu, X. P.; Case, D. A. *J. Biomol. NMR* **2001**, *21*, 321–333.
(6) Neal, S.; Nip, A. M.; Zhang, H. Y.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215–240.
(7) Meiler, J. *J. Biomol. NMR* **2003**, *26*, 25–37.
(8) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38*, 289–302.
(9) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615–9620.
(10) Shen, Y.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685–4690.
(11) Montalvao, R. W.; Cavalli, A.; Salvatella, X.; Blundell, T. L.; Vendruscolo, M. *J. Am. Chem. Soc.* **2008**, *130*, 15990–15996.
(12) Robustelli, P.; Cavalli, A.; Vendruscolo, M. *Structure* **2008**, *16*, 1764–1769.
(13) Wishart, D. S.; Arndt, D.; Berjanskii, M.; Tang, P.; Zhou, J.; Lin, G. *Nucleic Acids Res.* **2008**, *36*, W496–W502.
(14) Shen, Y.; Vernon, R.; Baker, D.; Bax, A. *J. Biomol. NMR* **2009**, *43*, 63–78.
(15) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6946–6951.
(16) Pople, J. A. *Mol. Phys.* **1958**, *1*, 175–180.
(17) Zhang, H. Y.; Neal, S.; Wishart, D. S. *J. Biomol. NMR* **2003**, *25*, 173–195.
(18) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
(19) Chandonia, J. M.; Walker, N. S.; Conte, L. L.; Koehl, P.; Levitt, M.; Brenner, S. E. *Nucleic Acids Res.* **2002**, *30*, 260–263.

JA903772T