J Biomol NMR (2010) 48:13–22
DOI 10.1007/s10858-010-9433-9

ARTICLE

SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network

Yang Shen · Ad Bax

Received: 4 June 2010 / Accepted: 30 June 2010 / Published online: 14 July 2010

Abstract NMR chemical shifts provide important local structural information for proteins and are key in recently described protein structure generation protocols. We describe a new chemical shift prediction program, SPARTA+, which is based on artificial neural networking. The neural network is trained on a large carefully pruned database, containing 580 proteins for which high-resolution X-ray structures and nearly complete backbone and $^{13}C^{\beta}$ chemical shifts are available. The neural network is trained to establish quantitative relations between chemical shifts and protein structures, including backbone and side-chain conformation, H-bonding, electric fields and ring-current effects. The trained neural network yields rapid chemical shift prediction for backbone and $^{13}C^{\beta}$ atoms, with standard deviations of 2.45, 1.09, 0.94, 1.14, 0.25 and 0.49 ppm for $\delta^{15}N$, $\delta^{13}C'$, $\delta^{13}C^{\alpha}$, $\delta^{13}C^{\beta}$, $\delta^{1}H^{\alpha}$ and $\delta^{1}H^{N}$, respectively, between the SPARTA+ predicted and experimental shifts for a set of eleven validation proteins. These results represent a modest but consistent improvement (2–10%) over the best programs available to date, and appear to be approaching the limit at which empirical approaches can predict chemical shifts.

Keywords Electric field · Hydrogen bonding · Torsion angles · SHIFTX · Structure database · Camshift · SPARTA

Electronic supplementary material The online version of this article (doi:10.1007/s10858-010-9433-9) contains supplementary material, which is available to authorized users.

Y. Shen · A. Bax (✉)
Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Building 5, room 126, NIH, Bethesda, MD 20892-0520, USA
e-mail: bax@nih.gov

Introduction

NMR chemical shifts have long been recognized as important sources of protein structural information (Saito 1986; Spera and Bax 1991; Wishart et al. 1991; Iwadate et al. 1999; Wishart and Case 2001). During protein structure calculations, chemical shift derived backbone ϕ/ψ torsion angles (Luginbühl et al. 1995; Cornilescu et al. 1999; Shen et al. 2009) are often used as empirical restraints, complementing the more traditional restraints derived from NOEs, J couplings and RDCs. More recently, several approaches for generating protein structures have been developed which rely on backbone chemical shifts as the only source of experimental input information (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008). The success of these methods hinges on the accuracy at which chemical shifts can be related to protein structure. Although chemical shifts can be computed for known structures by *de novo* computational methods (de Dios et al. 1993; Xu and Case 2001; Vila et al. 2008; Vila et al. 2009), database-derived empirically optimized methods yield lower root-mean-square (rms) differences between observed and predicted values. Recent programs of this latter class include ShiftX (Neal et al. 2003), SPARTA (Shen and Bax 2007), and Camshift (Kohlhoff et al. 2009), and these are the chemical shift prediction methods used in chemical shift based structure prediction efforts.

The ShiftX program actually derives predicted ^{1}H, ^{13}C, and ^{15}N chemical shifts from atomic coordinates using a hybrid approach which employs a pre-calculated, database-derived chemical shift hypersurface in combination with classical or semi-classical equations for ring current, electric field, hydrogen bonding and solvent effects. SPARTA is an empirical method which searches a database of assigned proteins of known structure for triplets of

residues that most closely match structural and sequence characteristics of any triplet of residues in the query protein. Camshift is a recently introduced program which predicts chemical shifts using an empirically derived complex polynomial function to correlate interatomic distances with chemical shifts. A neural network based method, known as PROSHIFT (Meiler 2003), also makes good chemical shift predictions, albeit at an accuracy slightly below those of the more recent programs.

In this work we introduce the program SPARTA+, also based on the artificial neural network protocol, to predict chemical shifts for backbone and $^{13}C^{\beta}$ atoms in proteins. Compared to PROSHIFT, SPARTA+ uses an approximately two-fold larger protein database, recently developed for the program TALOS+, which establishes the inverse correlation, i.e., predicts backbone torsion angles from experimental chemical shifts (Shen et al. 2009). As described below, the input parameters for the neural network training procedure differ from those of PROSHIFT, and are more similar to those used by the program SPARTA, hence the naming of the new program.

SPARTA+ employs a well-trained neural network algorithm to make rapid chemical shift prediction on the basis of known structure. Validation on proteins not included in the training set shows modestly improved agreement between the experimental chemical shifts and the SPARTA+ predicted chemical shifts, over chemical shifts predicted by the original SPARTA, Camshift, and ShiftX methods.

Methods

Preparation of the NMR database

This work utilizes the protein structure and chemical shift database, originally used to develop the TALOS program (Cornilescu et al. 1999), and subsequently expanded to 200 proteins for the SPARTA and TALOS+ programs (Shen and Bax 2007; Shen et al. 2009), and most recently expanded further to 580 proteins for developing an empirical relation between chemical shifts and the cis or trans conformation of Xxx-Pro peptide bonds by the program Promega (Shen and Bax 2010). Nearly complete backbone NMR chemical shifts ($\delta^{15}N$, $\delta^{13}C'$, $\delta^{13}C^{\alpha}$, $\delta^{13}C^{\beta}$, $\delta^{1}H^{\alpha}$ and $\delta^{1}H^{N}$) for these proteins are taken from the BMRB (Doreleijers et al. 1999; Doreleijers et al. 2005), with atomic coordinates taken from the corresponding high-resolution X-ray structures in the PDB (Berman et al. 2000). Residues containing two or less assigned chemical shifts were removed from the database. To minimize the influence of chemical shift outliers, chemical shift values that deviate by more than five standard deviations from the SPARTA-predicted values were also removed from the database. Details regarding the preparation of the database, including calibration of reference frequencies, correction of ^{2}H isotope effects on $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$, identification of H-bonds, etc., have been described previously (Shen and Bax 2007; Shen et al. 2009).

Neural network architecture and training

A single-level feed-forward multilayer artificial neural network (ANN) is used in this work to identify the dependence of ^{15}N, $^{13}C'$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$ and $^{1}H^{N}$ chemical shifts on the local structural and dynamic information as well as amino acid type, and those of its immediate neighbors.

This single-level neural network has an architecture very similar to that of the first level neural network used by TALOS+ (Shen et al. 2009). The input signals to the first layer consist of tri-peptide structural parameter sets derived from the above described protein structural database. For predicting the chemical shifts of any given residue by SPARTA+, the structural input parameters include: (1) the backbone and side-chain torsion angles of this residue and its two immediate neighbors, (2) information on interactions such as H-bonding, ring-current effects, and electric field effects, and (3) predicted backbone flexibility (Fig. 1a; Table 1, column "Full"). Specifically, each tripeptide is represented by up to 113 nodes, which include for each residue the twenty amino acid type similarity scores, ten numbers representing $\phi/\psi/\chi_1/\chi_2$ torsion angles of each tripeptide (the ϕ value of the first and ψ value of the last residue of the tripeptides are not used), three numbers for the structure-derived predicted N–H order parameter S^2 (Zhang and Brüschweiler 2002) of each residue, and twenty numbers representing the H-bonding pattern for the tripeptide (Fig. 1b). As was done for the TALOS+ program, amino acid type similarity scores are taken from the 20×20 BLOSUM62 matrix, commonly used for calculating sequence alignment (see http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=sef.figgrp.194). Considering the periodic nature of the torsion angles, each of the $\phi/\psi/\chi_1/\chi_2$ torsion angles is represented by its sine and cosine values, thereby avoiding problems associated with the numerical discontinuities that exist when defining torsion angles in the $-180°$ to $+180°$ range (Meiler 2003). For each of the side-chain χ_1/χ_2 torsion angles, an additional Boolean number [1 or 0] is used to indicate whether a χ_1 or χ_2 torsion angle is defined for any given residue. For example, [sin(χ), cos(χ), 1] denotes a valid χ_1 or χ_2 torsion angle; [0, 0, 0] is used for residues lacking χ_1 or χ_2 torsion angles ($\chi_{2,1}$ torsion angles are used for χ_2 of Thr, Val and Ile). The H-bonding input information of each tripeptide is limited to the $H^{N}/H^{\alpha}/O$ backbone atoms of the center

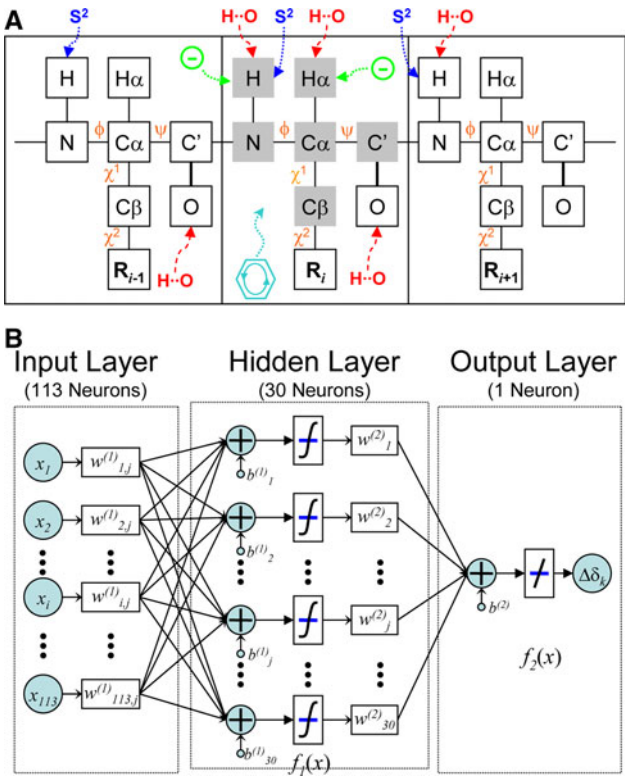

Fig. 1 **a** Illustration of a protein tripeptide chain together with factors that impact the backbone NMR chemical shifts, considered by the SPARTA+ program. Factors used for prediction of the chemical shifts of the center residue ^{15}N, ^{13}C', $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$ and $^{1}H^{N}$ (shaded in *grey*) include the backbone ϕ/ψ and side-chain χ_1/χ_2 torsion angles (colored *orange*), hydrogen bonding (*red*), electric fields (*green*), and ring-current effects (*aqua*), as well as backbone flexibility (*blue*). **b** Flow chart of the artificial neural network used in this work to study the relation between the protein structural and dynamic parameters (input layer) and NMR chemical shift (output layer)

residue, the carbonyl O atom of the first residue, and the H^N atom of the last residue. The H-bond information of each atom is denoted by three geometric parameters (Morozov et al. 2004), representing the distance between the donor hydrogen and the acceptor atom (H…A, d_{HA}), the cosine value of the angle at the acceptor atom (B–A…H, Φ), and the angle at the donor hydrogen (A…H–D, Ψ), plus one additional Boolean number [1 or 0] to indicate whether the atom is H-bonded. So, four numbers [d_{HA},cos(Φ),cos(Ψ),1] are used for each of the potentially H-bonded backbone atoms, and [0,0,0,0] represents the absence of a H-bond.

In the hidden layer of the network, where each node receives the weighted sum of the input layer nodes as a signal, 30 such nodes (or hidden neurons) are used. The output of a hidden layer node is obtained through a nodal transformation function (Fig. 1b).

For the purpose of predicting the NMR chemical shifts from protein structural parameters, the secondary chemical shift $\Delta\delta X$ ($X = {}^{15}$N, ^{13}C', $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$ or $^{1}H^{N}$) of the center residue of each tri-peptide in the database is used as the target of the first level network, after subtracting the contributions from ring-current effects (δX_{ring}) and electric fields effects (δX_{EF}), i.e.,

$$\Delta\delta X = \delta X - \delta X_{rc} - \delta X_{ring} - \delta X_{EF} \quad (1)$$

where δX_{rc} is the random coil chemical shift of nucleus X, δX_{EF} is calculated for $^{1}H^{\alpha}$ and $^{1}H^{N}$ nuclei only, using the Buckingham method (Buckingham 1960) and atom selection criteria analogous to those of the ShiftX program (Neal et al. 2003), δX_{ring} is calculated for all six types nuclei using the Haigh-Mallion model (Haigh and Mallion 1979; Case 1995), in the same way as used by the SPARTA program (Shen and Bax 2007). Note that chemical shift corrections from the neighboring residues, as used by the TALOS, SPARTA, and TALOS+ methods, are not included here when calculating the secondary chemical shifts, $\Delta\delta X$, because the neural network optimally accounts for those effects after training of the network on the database. Each output value has one node with a linear activation function ($f_2(x) = x$; Eq. 2). The empirical relationship between the NMR secondary chemical shift and the protein structural and sequence data, received by the network (Fig. 1b), is given by

$$\Delta\delta_{1\times1} = f_2\left(f_1\left(X_{1\times113}\times W^{(1)}_{113\times30} + b^{(1)}_{1\times30}\right)\times W^{(2)}_{30\times1} + b^{(2)}_{1\times1}\right) \quad (2)$$

with $f_1(x) = (1 - e^{-2x})/(1 + e^{-2x})$, and $f_2(x) = x$. $X_{1\times113}$ is the input data vector consisting of 113 elements; $\boldsymbol{W}^{(1)}$ and $\boldsymbol{b}^{(1)}$ are the weight matrix and bias, respectively, for the connection between the nodes in the input and the hidden layer; $\boldsymbol{W}^{(2)}$ and $\boldsymbol{b}^{(2)}$are the weight matrix and bias, for the connection between the nodes in the hidden and output layer; $\Delta\delta_{1\times1}$ is the training target or the output vector.

Neural network training

The weight and bias terms were determined by training the artificial neural network on the 580-protein structural database with associated chemical shifts, described above. To prevent over-training, a three-fold training and validation procedure was employed for the neural network model by dividing the input–output training dataset into three separate subsets, followed by separate training of the corresponding neural networks. For each of these three network optimizations, one-third of the database was excluded from the training but then used to evaluate the training performance of the neural network on the other two input–output subsets during the training. This subset, referred as the validation dataset, was not used to calculate the weight changes in this network. Training of the network was terminated when the

Table 1 Statistics of influence of various factors on SPARTA+ chemical shift prediction

	SPARTA	SPARTA+[a]					
		Full	Test I	Test II	Test III	Test IV	Test V
Training input[b]							
Residue type	●	●	●	●	●	●	●
$\phi/\psi/\chi_1/\chi_2$	●/●/●/×	●/●/●/●	●/●/●/●	●/●/●/●	●/●/●/●	●/●/●/×	●/●/×/×
H-bond	×	●	●	●	×	×	×
S^2	×	●	●	×	×	×	×
Training target $\delta-\delta_{rc}^{c}$							
$-\delta_{neighbor}$	●	×	×	×	×	×	×
$-\delta_{ring}$	●	●	●	●	●	●	●
$-\delta_{EF}$	×	●	×	×	×	×	×
Output $\Delta\delta^{pred} + \delta_{rc}^{d}$							
$+\delta_{neighbor}$	●	×	×	×	×	×	×
$+\delta_{ring}$	●	●	●	●	●	●	●
$+\delta_{EF}$	×	●	×	×	×	×	×
$+\Delta_{HB}$	●	×	×	×	×	×	×
RMSD(δ^{pred}, δ^{obs})[e] [ppm]							
$\delta^{15}N$	2.56 (2.56)	2.45 (2.48)	2.46 (2.48)	2.47 (2.49)	2.52 (2.52)	2.50 (2.51)	2.62 (2.64)
$\delta^{1}H^{\alpha}$	0.29 (0.27)	0.25 (0.25)	0.27 (0.25)	0.27 (0.25)	0.29 (0.29)	0.29 (0.29)	0.29 (0.29)
$\delta^{13}C'$	1.14 (1.13)	1.09 (1.11)	1.09 (1.11)	1.09 (1.11)	1.13 (1.14)	1.13 (1.14)	1.16 (1.16)
$\delta^{13}C^{\alpha}$	1.04 (1.01)	0.94 (0.98)	0.94 (0.98)	0.97 (0.99)	0.99 (1.00)	1.02 (1.03)	1.02 (1.05)
$\delta^{13}C^{\beta}$	1.16 (1.06)	1.14 (1.11)	1.14 (1.11)	1.14 (1.11)	1.14 (1.11)	1.15 (1.12)	1.16 (1.14)
$\delta^{1}H^{N}$	0.54 (0.51)	0.49 (0.47)	0.50 (0.48)	0.50 (0.48)	0.58 (0.54)	0.58 (0.54)	0.58 (0.54)

[a] See text ("Results and discussion") for the description of each testing neural network

[b] Structural and dynamic factors used as inputs for the database search (SPARTA) or neural network (SPARTA+) training procedure. All factors are for all three residues of a given tripeptide (see Fig. 1a). Parameters included and omitted in each input set are marked ● and ×, respectively

[c] NMR (secondary) chemical shifts used as the targets (outputs) of the database search (SPARTA) or neural network (SPARTA+) training procedure. The (secondary) NMR chemical shifts were obtained from the difference between the chemical shift δ and the random coil chemical shift δ_{rc}, after subtracting the corrections from neighboring residues ($\delta_{neighbor}$), the contributions from ring current effects (δ_{ring}), or electric fields (δ_{EF})

[d] Offsets and corrections, in addition to the random coil chemical shift δ_{rc}, applied to SPARTA or SPARTA+ predicted secondary chemical shift ($\Delta\delta^{pred}$), i.e., the final SPARTA/SPARTA+ predicted chemical shifts

[e] RMS deviation between the predicted and experimental (obs) chemical shifts for eleven proteins which are not present in the SPARTA+ training database. For SPARTA+, the prediction performances for the validation datasets (see "Methods") in the training database are provided in parentheses. For the SPARTA predictions, performances listed between brackets are those obtained for the 580-protein training database, but with the protein predicted excluded from this database

performance of the network on the validation dataset, represented by the mean squared errors between the predicted values and targets, began to degrade. This procedure was repeated three times, each time with a different one-third of the database proteins assigned to the validation set.

Neural network testing and validation

In addition to the above threefold training and validation, a second validation procedure was performed for a set of eleven additional proteins, with also nearly complete chemical shifts, a good quality reference structure, and no homologous protein (≥30% sequence identity) in the 580-protein database. This set of eleven proteins was identified after the original 580-protein database had been assembled and used for training of the ANN.

The final predicted NMR chemical shifts are obtained from:

$$\delta X = \Delta\delta X_{pred} + \delta X_{rc} + \delta X_{ring} + \delta X_{EF} \quad (3)$$

where $\Delta\delta X_{pred}$ is the ANN-predicted secondary chemical shift (Eq. 2) using the weights and biases obtained from the above training steps, after averaging over the outputs from the three separately trained networks.

Estimated errors for the predicted NMR chemical shifts

The original SPARTA program estimates the chemical shift prediction errors on the basis of an empirical

correlation between this error and the spread in chemical shifts among the 20 best matched tripeptides (Shen and Bax 2007). In the present study, an estimate for the chemical shift prediction error, σ, can be obtained by using an empirical $\Delta\delta(\phi,\psi)$ error surface (Spera and Bax 1991), which is calculated by:

$$\Delta(\phi,\psi) = \sqrt{\frac{\sum\left(\delta(\phi_k,\psi_k)^{\text{pred}} - \delta(\phi_k,\psi_k)^{\text{obs}}\right)^2 \exp\left(-\frac{(\phi-\phi_k)^2+(\psi-\psi_k)^2}{450}\right)}{\sum \exp\left(-\frac{(\phi-\phi_k)^2+(\psi-\psi_k)^2}{450}\right)}} \quad (4)$$

where the prediction errors between ANN-predicted $\delta(\phi_k,\psi_k)^{\text{pred}}$ and experimental $\delta(\phi_k,\psi_k)^{\text{obs}}$ chemical shifts are convoluted with a Gaussian function and then summed over all residues (k) of the validation subsets in the training database, followed by normalization.

The SPARTA+ chemical shift prediction, accomplished by the above described ANN procedure, is carried out by a program largely written in C++, which is ten times faster than the original SPARTA method. On a PC with a single 2.4 GHz CPU, the SPARTA+ chemical shift prediction takes *ca* 2 s for a 100-residue protein, the majority of which is actually attributed to loading of the error surfaces.

Results and discussion

Neural network chemical shift prediction

For each type of nucleus (^{15}N, ^{13}C', $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$ and $^{1}H^{N}$), three artificial neural networks were trained separately to predict the chemical shift, using a three-fold training and validation procedure. The trained weights and biases obtained for each network are then used to calculate the chemical shifts for each of a protein's backbone and $^{13}C^{\beta}$ atoms (except for the N- and C-terminal residues), using Eqs. 2 and 3. The low rms difference between the predicted and observed NMR chemical shifts, evaluated over the validation datasets (Table 1), indicates that the networks are well-trained.

To further inspect the chemical shift prediction performance of the trained neural networks, eleven additional proteins were used which were not present in any of the training or validation sets. The chemical shifts predicted for these eleven proteins were obtained by averaging the outputs of the three separately trained neural networks, obtained from the above described threefold training procedure. The predicted chemical shifts show good agreement with the experimental chemical shifts, with standard deviations of 2.45, 1.09, 0.94, 1.14, 0.25, and 0.49 ppm for δ^{15}N, δ^{13}C', $\delta^{13}C^{\alpha}$, $\delta^{13}C^{\beta}$, $\delta^{1}H^{\alpha}$ and $\delta^{1}H^{N}$, respectively, including outliers. The rmsd's for δ^{15}N, δ^{13}C' and $\delta^{13}C^{\alpha}$ in this set of eleven proteins are slightly lower than those for the validation datasets used during the network training (Table 1), most likely the result of the threefold averaging procedure used for this set, which is not applicable for the validation sets (see below). The performance of alternate chemical shift prediction programs was also evaluated on this set of eleven proteins, including SPARTA (Shen and Bax 2007) and webserver versions of ShiftX (Neal et al. 2003), CamShift (Kohlhoff et al. 2009), and PROSHIFT (Meiler 2003).

Comparison of the predicted with experimental chemical shifts (Fig. 2a; Table S1) indicates that SPARTA+ slightly outperforms SPARTA, with rmsd values that are ca 10–15% lower for $\delta^{13}C^{\alpha}$, $\delta^{1}H^{N}$ and $\delta^{1}H^{\alpha}$, 5% for δ^{15}N and δ^{13}C', with the smallest improvement (2%) for $\delta^{13}C^{\beta}$. SPARTA+ outperforms the ShiftX and Camshift programs by slightly larger margins (ca 10–20%) for all six nuclei (Fig. 2a), and the alternate ANN-based PROSHIFT program by somewhat larger margins (Table S1). Interestingly, the fractional improvement in chemical shift prediction accuracy is largest for $^{13}C^{\alpha}$, often used as the most significant indicator of protein secondary structure.

Although with Pearson's correlation coefficients in the 0.7–0.8 range the prediction errors of SPARTA and SPARTA+ are correlated (data not shown), there clearly is considerable scatter. Averaging the predictions made by the original SPARTA program with those of SPARTA+, using weight factors of 0.3 and 0.7, respectively, yields a slight further improvement in prediction accuracy for ^{15}N, ^{13}C', and $^{13}C^{\beta}$ (Fig. 2a; Table S1).

Impact of structural parameters on prediction accuracy

The SPARTA program uses the $\phi/\psi/\chi_1$ torsion angles and residue type information of a query tripeptide to predict the chemical shifts for the atoms of its center residue, followed by applying corrections for the ring-current shift and H-bonding (H-bond distance only). Compared with SPARTA, the SPARTA+ procedure considers more H-bond geometric factors for the H-bonded atoms, as well as additional side-chain χ_2 torsion angle information, electric field effects, and structure-based prediction of backbone flexibility (see "Methods"; Table 1).

In order to investigate the impact of the different structural factors on the prediction accuracy of SPARTA+, multiple neural networks with different input of the protein structural/dynamic parameters and output of the (secondary) chemical shifts are evaluated. The network trained with the full set of the listed input parameters (see "Methods") is named "Full" (Table 1). Five additional testing networks are implemented too and referred to as "Test I" (lacking the electric field effect contribution

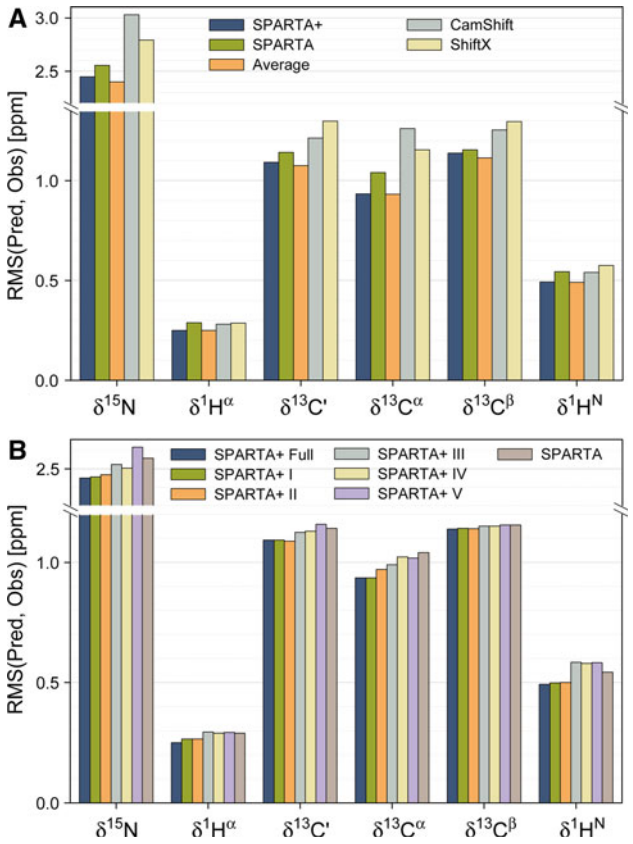

Fig. 2 Chemical shift prediction performance of various methods, evaluated over a set of eleven proteins not included in the neural network training database. The prediction performance (vertical axis) for the ^{15}N, ^{13}C', $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$ and $^{1}H^{N}$ chemical shifts is represented by the rms difference between the experimental and the predicted chemical shifts. *Colors* of the *bars* indicate the program used for predicting the chemical shifts, as marked in the panel. The *orange bar* corresponds to the weighted average (70%/30%) of the SPARTA+ and SPARTA predicted chemical shifts. **b** Impact of various structural and dynamic input parameters on SPARTA+ chemical shift predictions. *Dark blue* columns correspond to using the full set of SPARTA+ input parameters; the adjacent 5 bars correspond to input parameters defined in Table 1. The most right hand bar in each set corresponds to the original SPARTA prediction method

relative to "Full"), "Test II" (additionally lacking the predicted backbone order parameter), "Test III" (additionally lacking H-bonding information), "Test IV" (additionally lacking χ_2 torsion angles), and finally "Test V" (additionally lacking χ_1 torsion angles). All five testing networks have 30 and 1 neurons in their hidden and output layers, respectively; the number of input neurons are 113, 110, 90, 81 and 72, respectively (Table 1; see "Methods" for details on the number of neurons/nodes used for each individual structural/dynamic parameter). All testing networks are trained in the same threefold training and validation procedure, and using the same training database, as used for the network "Full". The accuracy of the chemical shift predictions performed by the trained testing networks is used to evaluate the importance of the various parameters for chemical shift prediction (Fig. 2b).

When only the residue type, backbone ϕ/ψ and side-chain χ_1 torsion angles, and ring-current effects are considered (network "Test IV"), the ANN remains capable of capturing the relation between NMR chemical shifts and protein structure reasonably well for all six types of nuclei (Table 1). Compared with the original SPARTA method, the overall prediction accuracy for the validation datasets is 1–2% worse for ^{13}C' and $^{13}C^{\alpha}$ predictions, 5–7% worse for $^{13}C^{\beta}$, $^{1}H^{\alpha}$ and $^{1}H^{N}$, and about 2% better for the ^{15}N (Table 1). Considering that the H-bond correction applied by SPARTA after its initial database search contributes a *ca* 5% improvement to its chemical shift prediction performance for $^{1}H^{\alpha}$ and $^{1}H^{N}$, the accuracy of the chemical shifts predicted by the Test IV network actually is quite close to that of the database search component of the original SPARTA method, with the exception of the ca 5% lower prediction accuracy for $^{13}C^{\beta}$. This result applies for both the validation datasets in the training database and for the eleven test proteins which are absent in the training database (Table 1). Moreover, the threefold training and validation procedure results in three networks that are trained separately with "half-independent" training datasets, making the contribution to chemical shift prediction errors from imperfect training data somewhat uncorrelated. As a result, averaging the chemical shifts predicted by the three separately trained networks then further improves the accuracy of the predicted chemical shifts by 2–4% (Table S2), making it slightly better than that of the SPARTA predicted shifts (except for ^{1}H predictions).

The effects of side-chain conformation on backbone chemical shifts have been well recognized (de Dios et al. 1993; Wang and Jardetzky 2004; Villegas et al. 2007; London et al. 2008; Mulder 2009). As indicated by the results of the Test V network, which lacks χ_1 torsion angle input information relative to network Test IV, the accuracy of the predicted chemical shifts decreases by 5% for ^{15}N and by about 1–2% for the other nuclei. When additionally considering the impact of the χ_2 torsion angle by comparing the difference in prediction accuracy of networks Test III and Test IV, a small improvement ($\sim$3%) of the $\delta^{13}C^{\alpha}$ prediction is observed (Fig. 2b; Table 1), but with the other nuclei virtually unaffected. Further inspection indicates that the observed improvement in $\delta^{13}C^{\alpha}$ prediction is almost entirely accounted for by the aromatic amino acids (Phe, His, Tyr and Trp) and Met (Fig. S2).

When H-bonding parameters are additionally included as input parameters when training the network (Test II), accuracy of the predicted chemical shifts further increases, both for the validation datasets in the training database and the set of eleven test proteins (Fig. 2b; Table 1). The

improvement in prediction accuracy upon of inclusion of H-bond input parameters is largest for proton chemical shifts (10–13%), but an improvement of 1–3% is also seen for $^{13}C'$, ^{15}N, and $^{13}C^{\alpha}$. A small further improvement (2–3%) in chemical shift prediction accuracy of the network is observed for $^{13}C^{\alpha}$ chemical shifts when the predicted backbone flexibility, as represented by the structure-predicted S^2 order parameter of Zhang and Brüschweiler (2002), is included with the input parameters (network Test I). Finally, the accuracy of the network-predicted $^{1}H^{\alpha}$ and $^{1}H^{N}$ chemical shifts is improved by several percentage points, when the electric field contribution to the $^{1}H^{\alpha}$ and $^{1}H^{N}$ chemical shifts is excluded prior to the network training and added back later to the predicted chemical shifts (as present by the network Full).

Application of SPARTA+ to CS-rosetta

Recently introduced procedures to generate protein structures using NMR chemical shifts as the only experimental input data have been quite successful in generating good quality models for small to medium-sized proteins (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008). Here, we evaluate the impact of improved chemical shift prediction on the effectiveness of one such protocol, CS-Rosetta (Shen et al. 2008).

CS-Rosetta utilizes NMR chemical shifts at two distinct steps of its protocol: fragment selection, and selection of its final models. The impact of improved chemical shift prediction on these two stages will be discussed below.

CS-Rosetta relies on the existence of a large database of protein structures from which fragments are selected to function as building blocks for the query protein. Similarity between the experimental chemical shifts of short segments in the query protein and chemical shifts of fragments in the protein database is used to guide the selection of the most suitable fragments. As the procedure requires a large database of high quality structures with known chemical shifts, and the database of experimentally determined NMR structures remains relatively small, CS-Rosetta utilizes a much larger database of X-ray structures, to which chemical shift values are added by prediction methods. A considerable improvement was found when the program SPARTA was used for adding chemical shifts to the protein database compared to predictions obtained using a less advanced program, known as DC, even though the accuracy of chemical shift predictions by SPARTA is only 10–20% better than those obtained by DC (Shen et al. 2008).

Considering that SPARTA+ offers a similar level of improvement over SPARTA, a comparable improvement in fragment quality might be expected when using the database with more accurately predicted chemical shifts, where fragment quality is measured by the backbone coordinate rms difference between the query segment and selected database fragments that most closely match the experimental secondary chemical shifts. However, on average, we find no improvement in fragment quality when using the protein structural database to which chemical shifts have been added by SPARTA+ over the database where these chemical shifts were added by SPARTA (data not shown). A likely reason for the lack of improvement is that the Rosetta structure generation procedure only utilizes the backbone torsion angles ($\phi/\psi/\omega$) from the selected fragments, whereas the improved chemical shift prediction above was shown to be dominated by sidechain and hydrogen bonding contributions (Fig. 2b; Table 1).

The second stage where accuracy of the chemical shift prediction plays a role during the CS-Rosetta protocol is during selection of the final models, from the very large ensemble of structures generated by its Monte Carlo procedure. Model selection is based on a combination of lowest empirical energy, as scored by the classic Rosetta program (Rohl et al. 2004), combined with a weighted chemical shift error score, χ^2, that accounts for the agreement between experimental chemical shifts and values predicted for each model. These latter models are full atom structures, including sidechains, H-bonds, etc., and improved ability to predict the chemical shifts for such structures is therefore expected to somewhat increase the ability to distinguish between accurate and less accurate models. We evaluate the impact of SPARTA+ on model selection for two proteins, DinI and Vc0424, neither of which is included in the SPARTA+ training database. For both proteins, a standard CS-Rosetta procedure (Shen et al. 2008) is performed, using a SPARTA+ assigned protein structural database. For each protein, the 10,000 structures generated by CS-Rosetta are then evaluated by calculating the total χ^2 score between the experimental chemical shifts and values predicted either by SPARTA+ or by SPARTA. For both proteins, models with the lowest total chemical shift χ^2 value are closer to the experimental reference structure (Fig. 3a, b, e, f) when using SPARTA+ chemical shifts. This small advantage remains when combining the χ^2 value with the Rosetta empirical energy function in the standard manner (Shen et al. 2008), again yielding slightly lower backbone rms differences between the models with the lowest total score and the corresponding reference structures (Fig. 3c, d, g, h; Table S2).

Concluding remarks

By using the artificial neural network approach, including a more complete consideration of various structural/dynamic parameters in proteins, SPARTA+ is able to predict

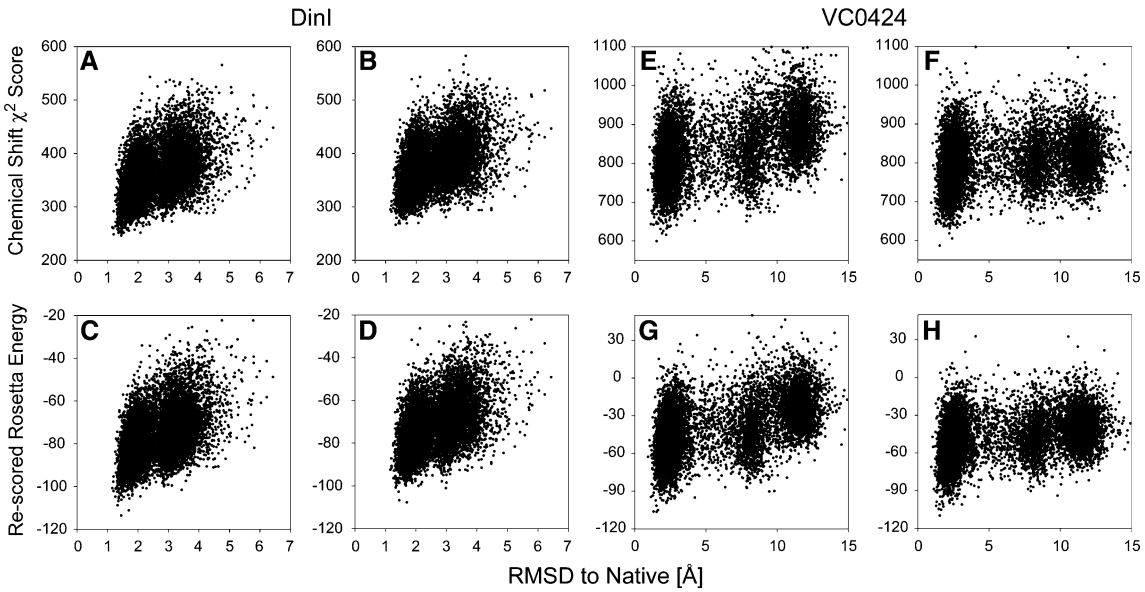

Fig. 3 CS-Rosetta model selection using either SPARTA+ or SPARTA chemical shift predictions. For proteins DinI (A-D; PDB entry 1GHH (Ramirez et al. 2000)) and VC0424 (E–H; PDB entry 1NXI (Ramelot et al. 2003)), 10,000 structures each were generated by a standard CS-Rosetta protocol, using a protein structural database with chemical shifts added by SPARTA+. For each CS-Rosetta model, the total χ^2 error function between the experimental chemical shifts and values predicted by SPARTA+ (**a**, **e**) or SPARTA (**b**, **f**) are plotted against the C^{α} coordinate rmsd relative to the experimental PDB structure. The re-scored Rosetta energy, calculated by adding the scaled SPARTA+ (**c**, **g**) or SPARTA (**d**, **h**) chemical shift χ^2 score to the raw Rosetta energy, is also plotted and used to select the final models (Table S3)

chemical shifts for backbone and $^{13}C^{\beta}$ atoms with modestly improved accuracy, compared with other similar chemical shift prediction approaches. The improvement of the accuracy in the SPARTA+ predicted chemical shifts is mostly credited to the additional structural/dynamic factors, i.e., χ_2 torsion angle, H-bonding and electric fields, as well as an averaging procedure over the outputs from three separated neural networks. Of all predicted chemical shifts, $\delta^{13}C^{\alpha}$ appears to benefit most from incorporation of the structure-predicted effect of backbone dynamics, used as an input parameter by SPARTA+. Conceivably, further improvements in this regard could be obtained by recording very extended ($\sim$1 μs) molecular dynamics trajectories, and averaging predicted chemical shifts over such a trajectory (Li and Brüschweiler 2010). However, from a practical perspective, such a computationally demanding approach is not yet practical.

Two interesting questions remain: Have we reached the limit of how well empirical methods can predict chemical shifts from known structure, and what is the reason for such a limit? Indeed the finding that only small increments in prediction accuracy are obtained when including additional input parameters suggests that we are asymptotically approaching the limit at which empirical approaches can predict chemical shifts. One may wonder whether the accuracy of the coordinates plays a role in prediction accuracy, for example. For the program ShiftX, a correlation between the accuracy of the prediction and the quality of the structure was reported (Neal et al. 2003). However, the SPARTA+ database uses far more stringent criteria for its database, including a crystallographic resolution threshold of 2.4 Å. Comparing the prediction accuracy for the 10 highest resolution structures (all $\leq$1 Å) with those of the lowest resolution structures (all at $\sim$2.4 Å) also shows a modest improvement for the higher resolution structure, although the effect is much smaller than found for ShiftX (Table S4). When evaluating proteins of even lower crystallographic resolution, the SPARTA+ accuracy further deteriorates (Table S4). However, with structures solved at a crystallographic resolution of 1 Å representing the most favorable case, and prediction errors remaining rather large, further progress by using a better reference database will not substantially improve results any further.

At a crystallographic resolution of 1 Å, atom positions are defined very well, and errors in backbone torsion angles are small compared to the gradient of the chemical shift surface with respect to these angles. However, two important sources of potential error remain. First, many sidechains are highly disordered in solution as judged, for example, by NMR relaxation measurements (Palmer 1997; Kay 1998; Yang et al. 1998; Lee and Wand 2001), an effect not easily accounted for by an empirical approach such as SPARTA+. Second, ab initio calculations indicate chemical shifts to be extremely sensitive to relatively small

deviations from ideal geometry and small steric clashes. Even at the highest level of resolution, the atomic coordinate precision is usually insufficient to accurately account for such distortions (Karplus 1996), and empirical characterization by an approach such as SPARTA+ appears beyond reach. Even if we were to add corrections for specific geometry distortions to the SPARTA+ values, predicted by density functional theory (DFT) computations, this would not be of immediate practical use, as the precise magnitude of a local geometric distortion almost invariably remains subject to high experimental uncertainty.

Although the improvement of the chemical shifts prediction performance is modest, chemical shift prediction by SPARTA+, using Eq. 2 with its trained weights and biases, is more than an order of magnitude faster than SPARTA. Moreover, the neural network equation (Eq. 2) used by SPARTA+ is differentiable with respect to the torsion angles, making it potentially possible to be used (on the fly) by the protein structure calculation and refinement procedures in combination with other, standard input restraints, in a manner similar to that proposed for CamShift (Kohlhoff et al. 2009).

Software availability

SPARTA+ and detailed instructions on its use can be downloaded from http://spin.niddk.nih.gov/bax/software/SPARTA+. Source code is available upon request.

Acknowledgments This work was supported by the Intramural Research Program of the NIDDK, NIH, and by the Intramural AIDS-Targeted Antiviral Program of the Office of the Director of the NIH.

References

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

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

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 U S A 104:9615–9620

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

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

Doreleijers JF, Vriend G, Raves ML, Kaptein R (1999) Validation of nuclear magnetic resonance structures of proteins and nucleic acids: hydrogen geometry and nomenclature. Proteins-Struct Funct Genet 37:404–416

Doreleijers JF, Nederveen AJ, Vranken W, Lin JD, Bonvin A, Kaptein R, Markley JL, Ulrich EL (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. J Biomol NMR 32:1–12

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

Iwadate M, Asakura T, Williamson MP (1999) C-alpha and C-beta carbon-13 chemical shifts in proteins from an empirical database. J Biomol NMR 13:199–211

Karplus PA (1996) Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci 5:1406–1420

Kay LE (1998) Protein dynamics from NMR. Nat Struct Biol 5: 513–517

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

Lee AL, Wand AJ (2001) Microscopic origins of entropy, heat capacity and the glass transition in proteins. Nature 411:501–504

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

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

Luginbühl P, Szyperski T, Wüthrich K (1995) Statistical basis for the use of 13Cα chemical shifts in protein structure determination. J Magn Reson Ser B 109:229–233

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

Morozov AV, Kortemme T, Tsemekhman K, Baker D (2004) Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. Proc Natl Acad Sci U S A 101:6946–6951

Mulder FAA (2009) Leucine side-chain conformation and dynamics in proteins from C-13 NMR chemical shifts. Chembiochem 10:1477–1479

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

Palmer AG (1997) Probing molecular motion by NMR. Curr Opin Struct Biol 7:732–737

Ramelot TA, Ni SS, Goldsmith-Fischman S, Cort JR, Honig B, Kennedy MA (2003) Solution structure of Vibrio cholerae protein VC0424: a variation of the ferredoxin-like fold. Protein Sci 12:1556–1561

Ramirez BE, Voloshin ON, Camerini-Otero RD, Bax A (2000) Solution structure of DinI provides insight into its mode of RecA inactivation. Protein Sci 9:2161–2169

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using rosetta. Meth Enzymol 383:66–93

Saito H (1986) Conformation-dependent C13 chemical shifts—a new means of conformational characterization as obtained by high resolution solid state C13 NMR. Magn Reson Chem 24:835–852

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

Shen Y, Bax A (2010) Prediction of Xaa-Pro peptide bond conformation from sequence and chemical shifts. J Biomol NMR 46:199–204

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu GH, Eletsky A, Wu YB, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008)

Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A 105:4685–4690

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and Ca and Cb 13C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

Vila JA, Aramini JM, Rossi P, Kuzin A, Su M, Seetharaman J, Xiao R, Tong L, Montelione GT, Scheraga HA (2008) Quantum chemical C-13(alpha) chemical shift calculations for protein NMR structure determination, refinement, and validation. Proc Natl Acad Sci U S A 105:14389–14394

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

Villegas ME, Vila JA, Scheraga HA (2007) Effects of side-chain orientation on the C-13 chemical shifts of antiparallel beta-sheet model peptides. J Biomol NMR 37:137–146

Wang YJ, Jardetzky O (2004) Predicting N-15 chemical shifts in proteins using the preceding residue-specific individual shielding surfaces from phi, psi(i-1), and chi1 torsion angles. J Biomol NMR 28:327–340

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

Wishart DS, Sykes BD, Richards FM (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J Mol Biol 222:311–333

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: 496–502

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

Yang DW, Mittermaier A, Mok YK, Kay LE (1998) A study of protein side-chain dynamics from new H-2 auto-correlation and C-13 cross-correlation NMR experiments: application to the N-terminal SH3 domain from drk. J Mol Biol 276:939–954

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