Journal of Biomolecular NMR, **26:** 25–37, 2003.
KLUWER/ESCOM

PROSHIFT: Protein chemical shift prediction using artificial neural networks

Jens Meiler
University of Washington, Department of Biochemistry, Box 357350, Seattle, Washington 98195-7350, U.S.A.

Received 4 November 2003; Accepted 30 January 2003

Key words: chemical shift prediction, neural networks, NMR, proteins

Abstract

The importance of protein chemical shift values for the determination of three-dimensional protein structure has increased in recent years because of the large databases of protein structures with assigned chemical shift data. These databases have allowed the investigation of the quantitative relationship between chemical shift values obtained by liquid state NMR spectroscopy and the three-dimensional structure of proteins. A neural network was trained to predict the ^{1}H, ^{13}C, and ^{15}N of proteins using their three-dimensional structure as well as experimental conditions as input parameters. It achieves root mean square deviations of 0.3 ppm for hydrogen, 1.3 ppm for carbon, and 2.6 ppm for nitrogen chemical shifts. The model reflects important influences of the covalent structure as well as of the conformation not only for backbone atoms (as, e.g., the chemical shift index) but also for side-chain nuclei. For protein models with a RMSD smaller than 5 Å a correlation of the RMSD and the r.m.s. deviation between the predicted and the experimental chemical shift is obtained. Thus the method has the potential to not only support the assignment process of proteins but also help with the validation and the refinement of three-dimensional structural proposals. It is freely available for academic users at the PROSHIFT server: www.jens-meiler.de/proshift.html

Introduction

The chemical shift value for nuclei in liquid state NMR spectroscopy is not only determined by the covalent structure of the molecule but also depends to a less degree on through-space interactions. Thus the three-dimensional structure of the molecule itself, its dynamics and interactions with other molecules (e.g., the solvent) are of importance for predicting chemical shifts. Empirical methods for chemical shift prediction of organic molecules suffer if these effects are not considered. Most of these empirical methods rely on small molecule chemical shift databases that do not store the three-dimensional structure of the molecule, since often only the covalent structure is known and of interest (Meiler et al., 2002). In contrast, for proteins an increasing number of high resolution three-dimensional structures with assigned chemical shift information are available.

Many attempts have been undertaken to derive correlations between protein chemical shift values and their three-dimensional structure. The dependence of protein chemical shift values on covalent structure only was summarized by Wüthrich (1986) for single random coil amino acids and later investigated in more detail by including the effects of the adjacent amino acids in the chain (Braun et al., 1994; Wishart et al., 1995). Modern approaches try to use sequence dependent chemical shifts of known proteins to predict the chemical shift for unknown sequences (Gronwald et al., 1997; Wishart et al., 1997; Iwadate et al., 1999). Such methods consider secondary structure at least partially in an indirect way, because secondary structure can be derived in large part from the protein sequence (Rost, 1996; Jones, 1999).

Several attempts have been undertaken to predict chemical shift values from a three-dimensional protein structure. In 1991, Osapay and Case described

*To whom correspondence should be addressed. E-mail: jens@jens-meiler.de

an empirical approach to compute ^{1}H chemical shift values from three-dimensional protein structures using a database of 17 x-ray structures with assigned ^{1}H chemical shifts (Osapay and Case, 1991). Also quantum chemical calculations of shift values have become more and more reliable and powerful in recent years (Oldfield, 1995; Pearson et al., 1997; Luman et al., 2001; Xu and Case, 2001).

Following the opposite direction, a second group of algorithms have been developed to estimate either secondary structure or the protein backbone ϕ- and ψ-angles from experimentally determined chemical shifts. The most widely used approach to derive three-dimensional structural information from chemical shift values is the chemical shift index (CSI), which describes a systematic change of backbone chemical shift values in the presence of α-helix or β-sheet relative to the values seen for unstructured peptides. First introduced for the chemical shifts of C^α atoms (Wishart et al., 1992), the concept was soon enlarged to cover C, and C^β nuclei in the protein backbone (Wishart and Sykes, 1994) and now includes also H^α, H^N, and N nuclei (e.g., Le and Oldfield (1994)). Wang and Jardetzky describe an optimized way to combine these six chemical shift values into one consensus prediction of secondary structure (Wang and Jardetzky, 2002). The CSI and related methods are currently widely used to predict the backbone conformation of an amino acid from its chemical shift values (Spera and Bax, 1991) and to restrict the possible ranges for the protein backbone angles Φ and Ψ (Cornilescu et al., 1999) in structure determination protocols.

Artificial neural networks have become a common methodology in chemistry and biochemistry in recent years (Zupan and Gasteiger, 1993). In the context of proteins, they are widely used for secondary structure prediction (Rost, 1996; Qian and Sejnowski, 1988; Kneller et al., 1990; Stolorz et al., 1992; Rost and Sander, 1993; Rost et al., 1994; Meiler et al., 2001; Petersen et al., 2000; Chandonia and Karplus, 1999; Salamov and Solovyev, 1997; Meiler, 2002a) and for the assignment of NMR spectra (Pons and Delsuc, 1999; Choy et al., 1997). Neural networks are further intensively used for the prediction and analysis of NMR spectra obtained from organic substances (Kvasnicka et al., 1992; Meusinger and Moros, 1995; Thomas and Kleinpeter, 1995; Meiler et al., 2000; Meiler and Will, 2001; Ivanciuc et al., 1996; Clouser and Jurs, 1996; Robien, 1998).

The well understood dependence of the chemical shift value on the conformation of the protein backbone combined with the steadily growing number of three-dimensional protein structures with assigned chemical shifts suggest that these data are suitable to establish a general empirical relationship between the chemical shift values and the three-dimensional structure of a protein. In this article, a neural network is trained using $\sim$ 69,000 chemical shift values from 322 BMRB entries and corresponding three dimensional structures of the proteins from the PDB to generate a general model to predict ^{1}H, ^{13}C, and ^{15}N chemical shifts in proteins.

Materials and methods

Preparation of data

The complete BMRB database was searched for references to entries in the PDB. All possible BMRB-PDB matches were collected yielding 2,839 pairs comprised of a total of 333 different BMRB entries and 1,699 different PDB entries. The total number of chemical shift entries in this set of BMRB files was $\sim$220,000. For every BMRB-PDB pair, the longest exact protein chain match between the two entries was identified and as many chemical shift values as possible were assigned to atoms in the PDB entry.

If more than one PDB file fit a BMRB sequence, the quality of the structure as well as the completeness of the assignment was used as selection criteria. Thus, out of the 2,839 assignments 322 BMRB-PDB pairs were selected to provide a training data set using the following criteria: (1) Every BMRB entry was used only once; (2) the sequence of BMRB and PDB file should be identical (if possible missing amino acids at the beginning or the end of the chain were avoided); (3) no missing atom coordinates and no alternate positions should occur in the PDB file; (4) the resolution of the structure should be as high as possible (three-dimensional structures solved by NMR spectroscopy were assumed to have a resolution of 2.5 Å); and (5) as many as possible chemical shift values should be assigned. These criteria were combined into a quality measure and, if more than one PDB entry was assigned to one BMRB entry, the highest ranked PDB match was generally chosen. For 11 out of the original 333 BMRB entries, no PDB entry of sufficient quality was identified.

Coding the three-dimensional structure

To train a neural network on the correlation between structure and chemical shift, both must be represented by numeric values. Several formats for the neural network were considered: (1) Coding a single atom environment and predicting its chemical shift; (2) coding an amino acid environment and predicting all related chemical shifts in a parallel manner; (3) training specialized neural networks for every amino acid. As the specialization of the network increases the accuracy of the predicted shift might be expected to increase but the amount of available training data decreases.

Setup (3) requires the training of as many as 20 individual networks. The overall general information about chemical shift – structure correlation is not available for the network during the training but instead only the information for the respective amino acid is available. An amino acid specialized network however, might lead to an improved prediction in comparison with other setups. The amount of data available for some amino acids (e.g., C, H, M, P, W) and therefore for some side-chain atoms in particular, was insufficient to stabilize the network connections.

Setup (2) combines all data allowing the overall shift-structure correlation to be used by the network in training. The chemical shift prediction for the backbone atoms C^{α}, C^{β}, C, H^{α}, H^{N}, and N became sufficiently better than in setup (3). The prediction of side-chain atoms, however, did not improve significantly, likely because the amount of available data did not increase for many side-chain atoms. A network predicting only C^{α}, C^{β}, C, H^{α}, H^{N}, and N using this setup, however, achieves results that are only slightly worse than the final optimized setup.

Setup (1) is atom focused and therefore the most general of the considered versions. Since the general dependence of chemical shift from atom environment is trained, the frequently sampled backbone atoms can help to supplement information of sparsely sampled side-chain atoms.

The chemical shift values of a hydrogen atom and its covalently linked carbon or nitrogen atom are highly correlated, because both nuclei have a similar chemical environment. Neural networks can use such correlations between designated output values to improve prediction accuracy. Thus the network is built to always predict the shift of a linked hydrogen atom in parallel to the shift of the heavy atom if it is a carbon or nitrogen atom. Figure 1 visualizes the architecture of the neural network. The possible fragments in proteins are C_{quart}, C–H, N_{tert}, N–H, O–H, S–H.

The network has three output neurons for the hydrogen, carbon, and nitrogen shift value. In turn not for every output neuron a corresponding atom exists in every of the possible fragments. The weights that belong to these non-assigned output neurons are not modified during the training process and the output is also not evaluated when testing the network. The non-hydrogen atoms in these fragments are the focus for deriving the constitutional as well as the spatial description of the protein structure at this three-dimensional position. The input consists of four different groups describing (1) the atom in the focus (which is always non-hydrogen), (2) all atoms (up to 16) that are less than three bonds away from the focus in the covalent structure, (3) the 16 atoms that are closest in space, (4) protein and sample-dependent parameters.

The composition of the input parameters was optimized by (1) varying the number and kind of atom descriptors used in the parameter groups 1, 2, and 3, (2) varying the number of atoms in groups 2 and 3 from 4 to 32, (3) varying parameters from in the protein and sample-dependent group 4. The presented setup was found to be the optimal trade-off between the accuracy of the description (number of input parameters and in result the number of weights in the network) and the amount of available data for stabilizing these connections. The network had 350 input units, 64 hidden neurons and three output neurons (compare Figure 1) which resulted in the overall number of 22,659 connections in a standard feed-forward three-layer neural network.

Fewer input parameters would have been insufficient to describe the protein structure. In turn the prediction would become worse due to under-fitting the available information. More input parameters would have caused more connections that cannot be stabilized by this amount of training data. The result would be an over-fitted network perfect in predicting the training data but much worse for unknown proteins.

Generating training, monitoring and independent set of data

Of the 322 BMRB-PDB pairs 15 were randomly selected to form a monitoring set and another 15 were selected as the independent test set. The remaining 292 form the training set. The ^{15}N and ^{13}C chemical shifts were recalibrated as previously described (Cornilescu et al., 1999). The following data were

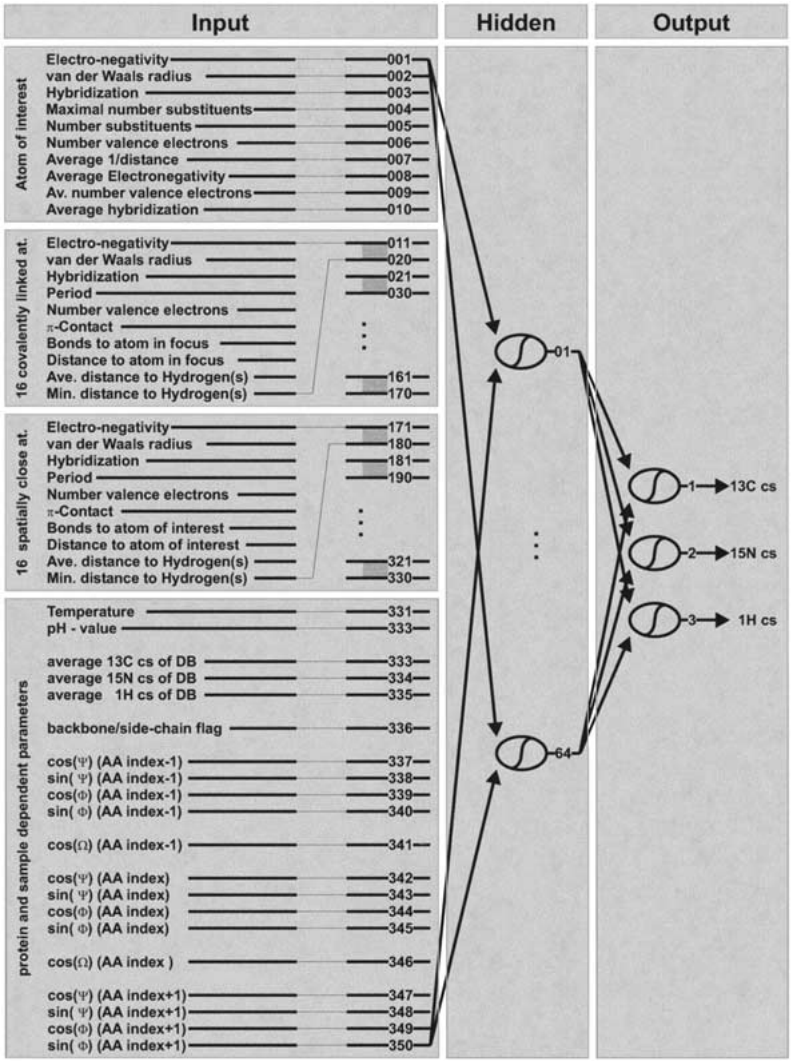

Figure 1. Structure of the artificial neural network. 350 input units can be subdivided into four groups: 10 parameters describe the atom in the focus; 16·10 parameters describe the up to 16 atoms in the first two covalent spheres around the atom in focus; another 16·10 parameter for 16 atoms closest in space (excluding atoms considered in the first group); and another 20 parameters which are derived from sample conditions or protein specific structural features. Table 1 summarizes the physical and chemical constants used to describe an atom. The value for 'hybridization' is defined to be 0 if the atom has sp^3-hybridization and 1 if the atom has sp^2-hybridization. The maximal number of substituents is defined to be the coordination given in Table 1 minus the hybridization. The actual number of substituents is determined by counting all non-hydrogen substituents of an atom. The average reciprocal distance is computed over all atoms out of group (2) and (3): $\overline{R} = \frac{1}{N}\sum_{i=1}^{N} r_i^{-1}$. It is large if these atoms are close but small if they are distant and therefore a density measure of the atom environment. The average electro-negativity, average number of valence electrons, and average hybridization are weighted by the inverse distance of the individual atoms, so that atoms close in space have a higher influence on the resulting parameter: $\overline{P} = \frac{1}{N \cdot R}\sum_{i=1}^{N} p_i \cdot r_i^{-1}$ if p is the respective property. The second block can hold up to four atoms directly linked to the atom in focus and up to 12 further atoms linked over two covalent bonds. Two atoms have 'π-contact' if there is one and the same conjugated π-electronical system which contains either both atoms or directly linked neighbors of the atoms (Meiler et al., 2000). This value is therefore 1, if the atom is in π-contact with the atom in focus and zero otherwise. The number of bonds to the atom in focus is defined by counting the bonds on the shortest path between the two atoms in the covalent structure. The average distance and the minimal distance to the hydrogen(s) linked to the atom in focus are only distinguishable if the chemical shifts of a CH_3 group are predicted. In this case one and the same value is predicted for all three hydrogen atoms in a single run, since for a CH_3 group fast rotation must be assumed which results in only one obtainable shift value. In contrast, two separate runs are carried out in order to predict the eventually different shift values of the hydrogen atoms in a CH_2 group. The 16 spatial closest atoms that are not considered in the second block form the third block. The average database values for the respective atom in the actual amino acid are derived from the BMRB. The backbone/side-chain flag is set to be 1, if the atom in focus is C, C^{α}, C^{β}, or N and otherwise to be 0. The last 14 input units reflect the backbone structure of the actual, the previous and the next amino acid. The cos and sin of the angles were used rather than the angles themselves to achieve a stetic function.

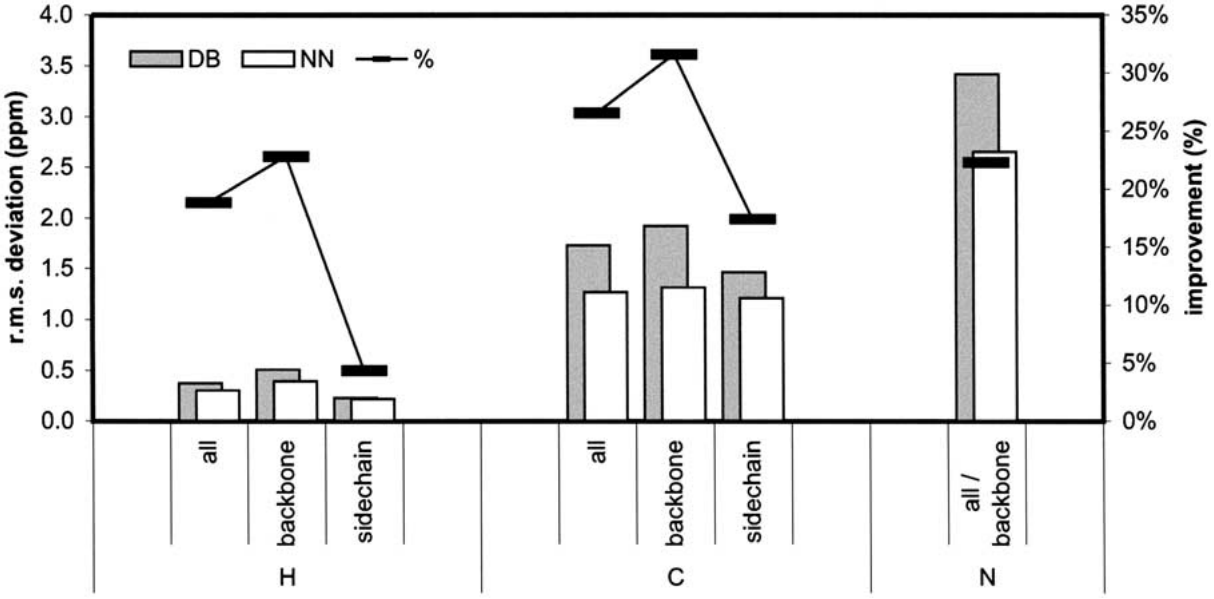

Figure 2. Root mean square deviations between experimental and predicted chemical shift value for hydrogen, carbon and nitrogen nuclei, using the average database values (gray) or the neural network (white) computed for the independent set of data. For hydrogen and carbon nuclei the plot distinguishes further between backbone and side-chain shift prediction. The improvement over the database estimation in percent is marked by horizontal black bars.

Table 1. Atom properties

Element	Electro-negativity[a]	van der Waals radius[a]	Period	Number valence electrons	Coordination
H	2.20	1.20	1	1	1
C	2.50	1.70	2	4	4
N	3.10	1.50	2	5	3
O	3.50	1.40	2	6	2
S	2.40	2.20	3	6	2,(4,6)[b]

[a]Taken from: HOLLEMANN-WIBERG, 'Inorganic Chemistry'.
[b]Values of 4 and 6 do not occur in proteins.

excluded from the set of data: (1) All ^{15}N or ^{13}C chemical shifts were ignored if less than ten data points were available for this calibration or if the calibration offset was larger than 4.0 ppm, respectively. (2) All chemical shifts with an error code different from '1' (not uniquely assigned) in the BMRB were excluded. (3) Further all chemical shift values for the first and the last two amino acids of every sequence were excluded. (4) All ^{15}N chemical shifts available for side-chain atoms were excluded (they were rather rare in the database so that their complex dependence on the three-dimensional structure could not be trained). (5) The complete set of data was excluded if the temperature given in the BMRB entry lay outside the range of [270K,330K] or the pH-value was smaller than 3. With these restrictions, the overall number of ~65,000 ^{13}C, ~16,000 ^{15}N, and ~88,000 ^{1}H chemical shifts are coded out of the earlier mentioned ~220,000 detected entries (~77%). However, most of the data (~19%) were excluded for having an error code different from '1'.

Neural network training

The weights were trained using back-propagation of errors. The transfer function was a sigmoid function $y = (1 + e^{-x})^{-1}$. Since this function only allows output values between 0 and 1 and, moreover, also the input data applied on a neural network should lie in the same order of magnitude, all input parameters were linearly scaled to lie between 0 and 1 before

entering the network. The output values were also linearly scaled after passing the neural network from the range of 0 to 1 to the range of 0 ppm to 200 ppm for carbon, 90 ppm to 150 ppm for nitrogen and −3 ppm to 13 ppm for hydrogen. These ranges were chosen to be somewhat larger than the ranges actually spanned by protein chemical shifts to (1) be flexible in predicting chemical shifts for new proteins that might lay slightly outside the ranges found in the used database, and (2) to avoid the network predicting values close to the extreme points 0 and 1, since therefore also extremely large absolute weights in the output layer become necessary, which make the network rather unstable and harder to train. The weights were set to random values between −0.1 and 0.1 before the training was started. The learning rate was set to 0.01 and decreased to 0.0001 at the end of the training procedure. The momentum was kept constant at 0.5. The network connections were trained with the training set of data until the r.m.s. of the monitoring set of data was minimized, which took 2,125 iterations and ∼96 h on a Pentium III 1GHz. The Software 'Smart' (Meiler, 1996–2002) was used to train and analyze the neural network.

Results and Discussion

Accuracy of the method

To evaluate the quality of the prediction method, the chemical shifts for the independent data set were predicted using the trained neural network. Figure 2 compares the overall achieved r.m.s. deviation between experimental and predicted chemical shift for the individual nuclei. Compared to the use of the average database values the neural network shows a significant improvement of 20% for hydrogen, 28% for carbon, and 23% for nitrogen. The resulting r.m.s. deviations are 0.3 ppm for hydrogen, 1.3 ppm for carbon, and 3.6 ppm for nitrogen nuclei. For all three nuclei, the improvement is mainly achieved in the prediction of backbone chemical shift values (C^{α}, C^{β}, C, H^{α}, H^{N}, and N) rather in prediction of the side-chain chemical shift values. Here the improvement for hydrogen and carbon nuclei is significantly smaller, 16% and 15%, respectively. This result is not surprising because more training data exist for backbone chemical shifts and because the backbone atoms are, on average, more perturbed from the random coil values by protein structure than side-chain shifts. Figure 2 suggests that in contrast to the averaged database values, the network makes equally accurate predictions for both backbone and side-chain atoms. The remaining uncertainty in the chemical shift values can likely be attributed to solvent effects, uncertainties in the three-dimensional structure itself, dynamics, an insufficient description of the three-dimensional structure, and uncertainty in the database used for training the neural networks (e.g., mis-assigned chemical shift values, calibration bias). As the database of chemical shifts and known structures grows, uncertainties in the training data are expected to decrease, improving the prediction accuracy of the model.

The prediction accuracy of the model is assessed in Figure 3. The correlation coefficients between predicted and observed chemical shifts are 1.000 for carbon, 0.828 for nitrogen, and 0.994 for hydrogen. Carbon chemical shifts can be predicted much more accurately than hydrogen and nitrogen chemical shifts, because they are less influenced by spatial interactions and the solvent, and are probably less often mis-assigned due to the better resolution of the carbon dimension. This logic applies also to hydrogen nuclei that are directly linked to a carbon atom. Carbon-linked hydrogen atoms are predicted with an r.m.s. deviation of 0.25 ppm while nitrogen-linked hydrogen atoms are predicted with an r.m.s. accuracy of 0.55 ppm. This suggests in combination with the comparably large r.m.s. deviation for nitrogen nuclei itself (3.6 ppm) that these chemical shifts are in general harder to predict accurately. The free electron pair of the nitrogen makes it more likely to be incorporated into spatial interactions with other parts of the protein or the solvent. Moreover the incorporation of the lone pair into this interaction will influence the chemical shift value of the nitrogen atom as well as the chemical shift value of linked hydrogen nuclei. The variety of such through-space interactions is much broader and geometrically less well defined than the number of possible local covalent structures in a protein and therefore less completely sampled in the database and consequently harder to predict accurately.

Table S1 summarizes the r.m.s. deviations achieved for the individual nuclei in the 20 amino acids. In some cases, the number of reported chemical shifts in the database was less than 30 (most side-chain nitrogen nuclei and some other side-chain atoms). Here the neural net predicted chemical shift value is no longer reliable and the average database value replaces the neural network predicted chemical shift value in the prediction. The only exception to this

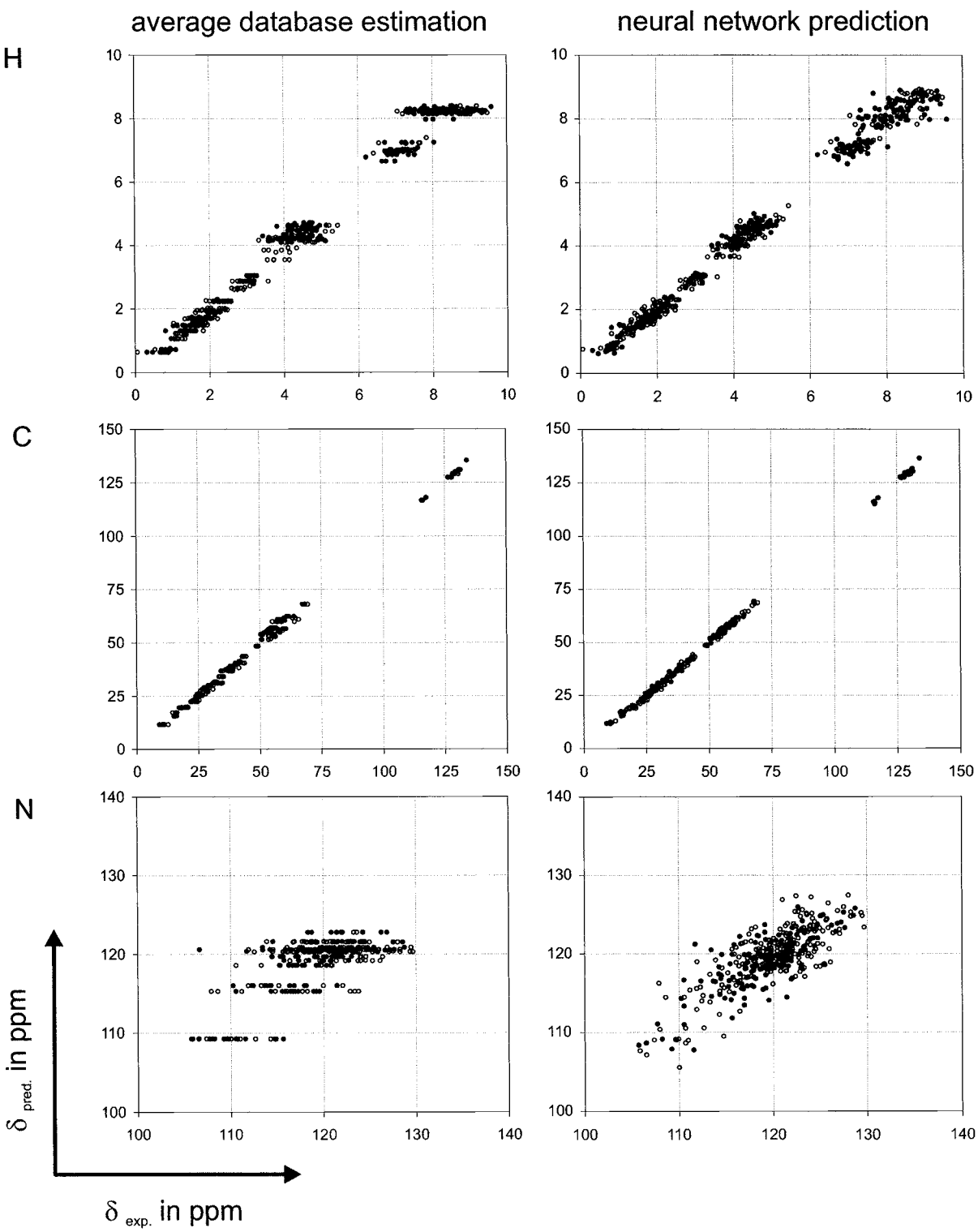

Figure 3. The correlation plots show the experimental chemical shift value versus the neural network predicted shift value for all three nuclei types. The open circles are randomly selected out of the training data set while the closed circles represent the independent data set.

rule is the nitrogen backbone chemical shift in proline, which is only represented 10 times in the training database. Thus the r.m.s. deviation of the predicted chemical shift values is higher than for most of the other backbone nitrogen chemical shifts. The network, in this case, still has some predictive power (the r.m.s. deviation is significantly smaller than achieved by the average database value). The chemical shift prediction for the proline nitrogen profits from the large number of available nitrogen backbone chemical shifts, although the covalent structure is slightly different.

Comparison with existing methods

Already in 1997 an empirical chemical shift prediction method for proteins on the basis of the BMRB database was published (Wishart et al., 1997). The prediction is done on the basis of sequence and secondary structure similarity of the unknown protein to database entries. Their program SHIFTY predicts ^{1}H and ^{13}C chemical shifts with an r.m.s of 0.2–0.3 ppm and 0.9–1.2 ppm, respectively, if a homologous sequence can be found in the database. Iwadate et al. showed that C^{α} and C^{β} chemical shifts are accurately predictable (r.m.s = 0.96 ppm) if backbone and side-chain conformation as well as hydrogen bonding is included (Iwadate et al., 1999). Finally Xu and Case (2001) published an algorithm based on a density functional database that predicts all ^{13}C and ^{15}N backbone chemical shifts at an accuracy of 1.0 and 1.9 ppm, respectively, if they are represented in the underlying database.

The neural network presented achieves here similar accuracies having the advantage of predicting all chemical shifts, backbone and side-chain, ^{1}H, ^{13}C, and ^{15}N at the same time independent of their representation in any database at very high speed (several 1,000 shifts per sec).

Chemical shift index

Figure S1 compares the result of secondary structure estimation using the CSI (Wishart et al., 1992; Wishart and Sykes, 1994) applied on either the experimental or the predicted chemical shift values. In ∼60% of all cases actual and predicted secondary structure are identical for both experimental and predicted chemical shift values. In ∼40% of all cases the minor failure of exchanging a sheet or a helix with a coil region occurs, and in only 3% of all predictions sheet and helix are exchanged. As described in the literature, this overall success rate can be pushed above 80% using a post analysis focusing on the consensus of the predicted secondary structure for a single amino acid itself and also between neighboring amino acids (Wang and Jardetzky, 2002).

The neural network accurately represents the well-defined relationship between chemical shifts and (Φ,Ψ)-angles for C^{α}, C, and H^{α} as well as the less defined relationship for C^{β}, H^{N}, and N as shown in Figure 4 for H^{α}, C and N.

Analysis of the neural network

Figure 5 provides an analysis of the influence of each input unit in the network on an output neuron. Diagram a) shows the influence of the average database value for the particular atom. As expected the database value of the carbon atom has an enormous influence for every computed carbon chemical shift value. The influence of the respective hydrogen and nitrogen shift values is smaller, because these shifts are more dependent on the three-dimensional structure. The database value of the covalently linked hydrogen has nearly the same influence on the nitrogen shift as the nitrogen database value. Looking at the influence of atoms in certain distances to the atom in focus either in the covalent structure or in space (panel b), the dominating influence of the covalent structure for the carbon chemical shift is easily seen. In contrast, the very similar covalent structure around the amide nitrogen atoms has, as expected, a lower influence and spatial interactions become somewhat more important. For the hydrogen chemical shift, the relevant 'atom in focus' is the heavy atom directly linked to the hydrogen and therefore highly predictive of the chemical shift value. Atoms close in covalent structure and spatially close atoms have similar smaller effects. Diagram c) compares the sensitivity of the model to the ten parameters used for describing a nucleus in the environment of the atom in focus. These values were summarized over all 32 atoms so that the actual value reflects the maximal change that could be caused by one of these individual parameters. Hybridization and valence electrons of the atoms close in covalent structure have the biggest influence on predicted carbon chemical shifts. In contrast, spatially close atoms (noted by the high influence of the input unit 'sphere') have a larger influence on nitrogen chemical shift values. The profile obtained for hydrogen atoms lies between these two extremes and reflects an average of hydrogen atoms having a covalent bond to a carbon and hydrogen atoms having a covalent bond to a nitrogen atom. Diagram d) and

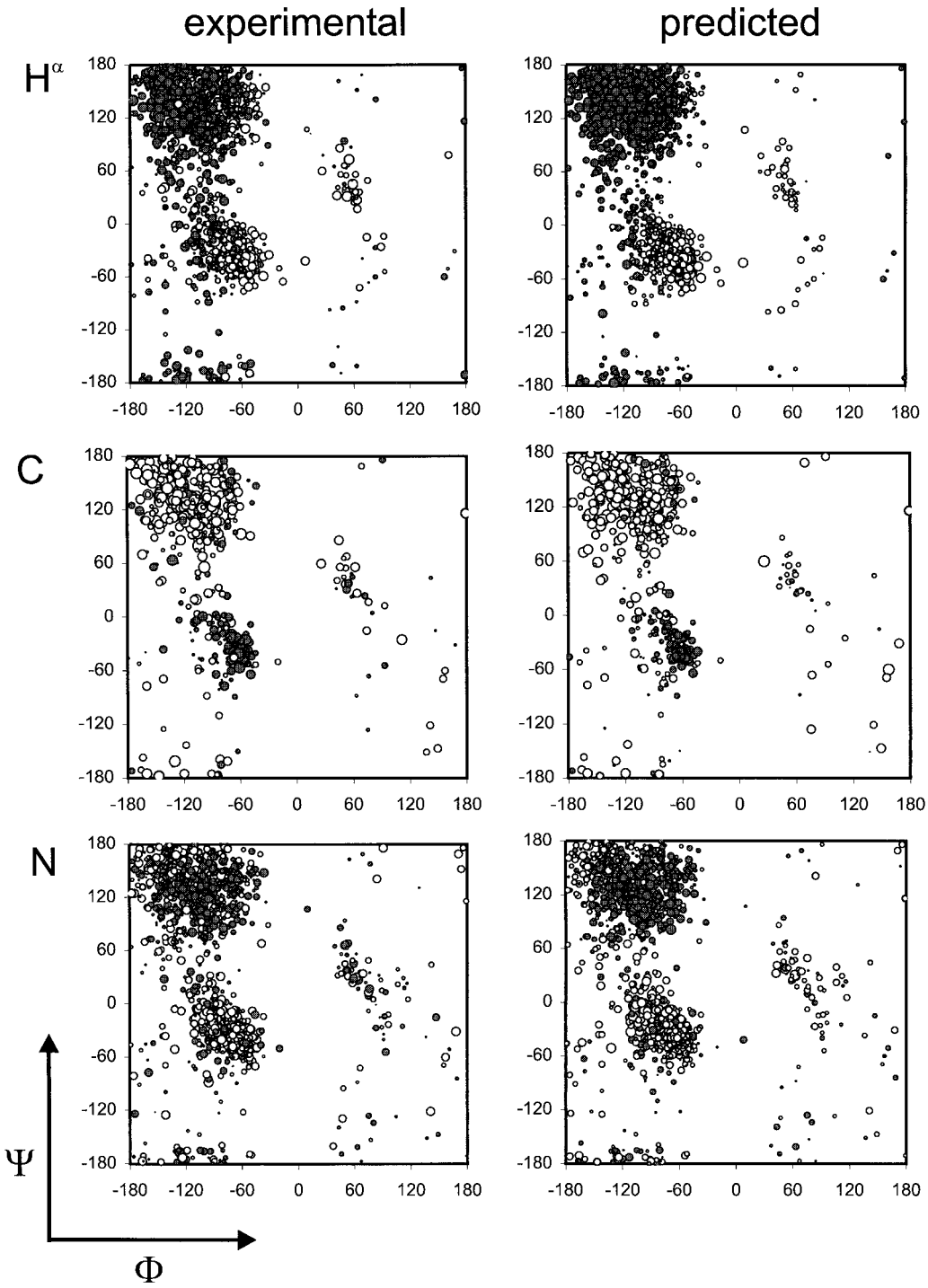

Figure 4. For each of the backbone atoms H^{α}, C and N the difference between the experimental chemical shift and the average database value on the left side is compared with the difference between the predicted chemical shift and the average database value on the right sight. Those values are plotted for all amino acids in the independent data set as Ramachandran diagrams. The area of the circles represents the absolute difference while a positive sign is indicated by a filled circle and a negative sign is indicated by an open circle.

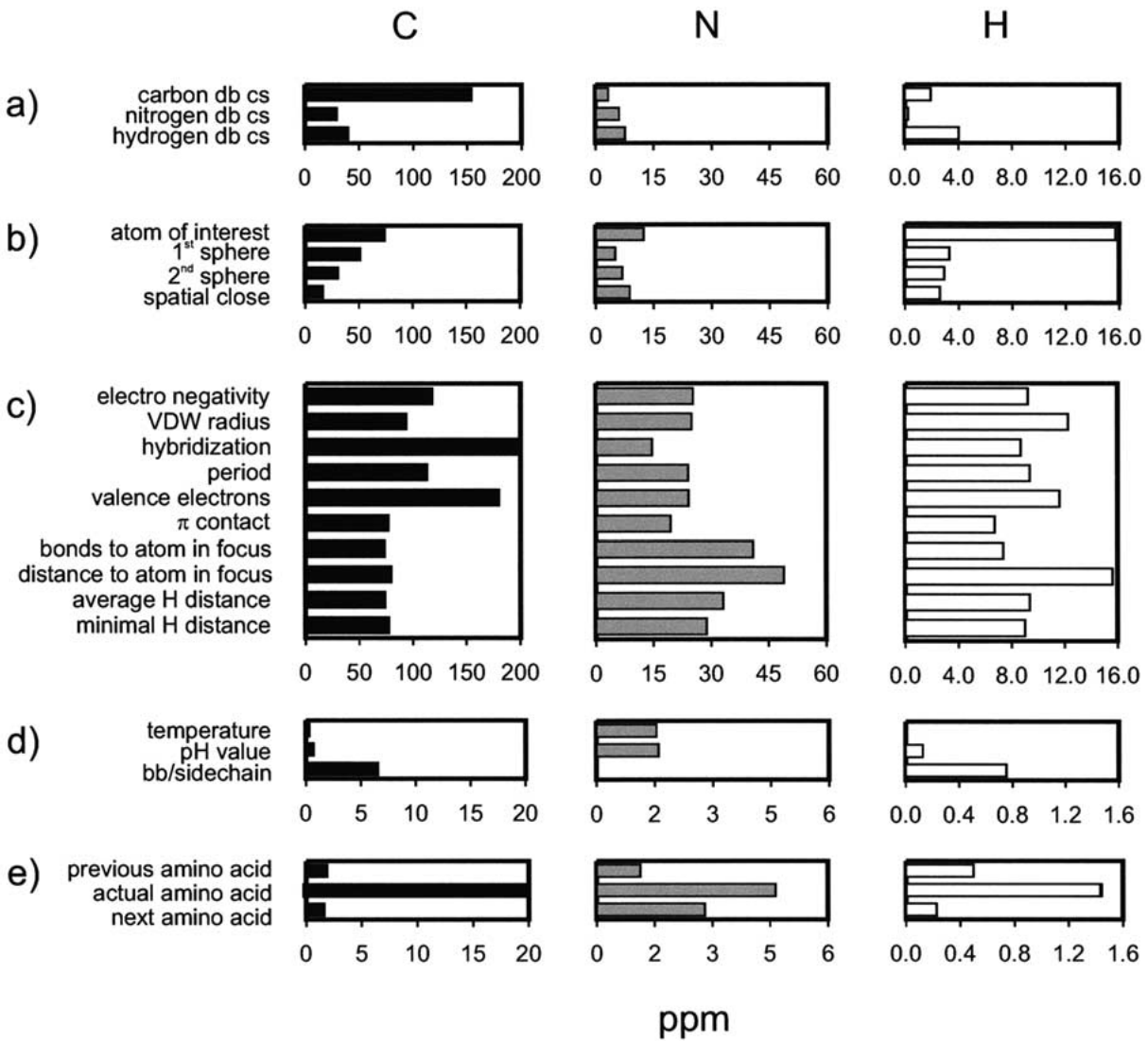

Figure 5. Sensitivity analysis of the finally trained neural network. Changes at input units which are likely to cause a large change of an output value are reflected by a large 'input sensitivity' for the corresponding input unit. The sensitivity can be given in the same units as the output value and is a semi-quantitative measure for the maximal change that can be caused by a certain input neuron. The measure is only semi-quantitative, since in most cases the assumption that every input value can be changed independently from every other input value does not hold. The absolute values given for the sensitivity tend to be higher than expected. In diagram a) the influence of the average database value on the actually computed shift value is plotted. Diagram b) discusses the average influence of the atom in focus itself compared to an atom out of the first sphere, the second sphere, and a spatial close nucleus. Diagram c) reports the input sensitivity with respect to the individual chemical, physical, and geometrical properties chosen to describe an atom close to the atom in focus (compare Figure 1) and diagram d) discusses the influence of temperature, pH-value, and the location of the atom (backbone ore side-chain) on the output. Diagram e) shows the influence of the backbone conformation on the computed chemical shift value. All values are given in ppm and report therefore the theoretical maximal effect that can be caused by a single change of the particular input value. The ranges covered from the plots of 200 ppm and 20 ppm for carbon, 60 ppm and 6 ppm for nitrogen, and 16.0 ppm and 4.0 ppm for hydrogen reflect either the maximal range covered by the output value (a, b, c) or 10% of this value (d, e).

e) are scaled by a factor of 10 to show the smaller absolute influences of the remaining input parameters on the prediction. It is striking that for carbon, the difference between the effect of the actual and adjacent amino acids is significantly larger than it is for nitrogen atoms.

Comparing the quality of the prediction for carbon and nitrogen and especially for amide hydrogen and other hydrogen atoms, it appears that the dependence of the chemical shift value on covalent structure is significantly easier to describe than through-space interactions. This weakness in predicting through- space interactions may be attributed to several points: (1) Interactions with solvent molecules cannot be reflected, since their exact position is unknown, (2) dynamics is not considered in the approach and will affect through-space interactions most, (3) the variety of possible through-space interactions is much larger and consequently much less sampled than the well defined and restricted covalent structure of proteins, and (4) uncertainties in the three-dimensional structures will affect the dependence of the chemical shift on geometry only and not its dependence on covalent structure.

Nevertheless, the neural network picks up a large part of the relation between structure and chemical shift and forms an overall uniform approach to predict chemical shifts in proteins. The fast growing database of available chemical shifts assigned to high-resolution three-dimensional protein structures will allow the rapid refinement of the network connections and cause an improvement of the prediction, in particular for side-chain nuclei. Also a chemical shift prediction for the side-chain atoms excluded so far will then become possible. Moreover the incorporation of structural noise/dynamic information via experimental order parameters, B-factors or by analyzing the usually available set of NMR structures might help in the improvement of the method.

Possible application in protein structure elucidation

Since the neural network based chemical shift prediction is extremely fast (∼5,000 chemical shifts per second on a 1.5 GHz Pentium 4 processor), it can serve as an additional restraint in structure elucidation process. While at the moment, only the backbone angles are restrained on the basis of the well-known dependencies of the backbone conformation, this much more general approach could be used to predict the chemical shifts during the structure elucidation on the fly and compare it to the experimental values. Moreover the quality of three-dimensional models for a protein structure could be investigated on the basis of the chemical shift prediction. The speed of the prediction would also allow a rapid search for similar proteins or protein fragments in the PDB, which includes in contrast to existing approaches not only static backbone chemical shifts but using dynamically defined shift values for both, backbone and side-chain atoms.

To support the potential of the method for the mentioned applications, the following experiment was performed: The protein fold prediction algorithm 'Rosetta' (Bonneau et al., 2001; Rohl and Baker, 2002) was used to create 2000 structures for the 81 amino acid DNA-Damage-Inducible protein I (1dinI). The solution structure for this protein as solved by NMR spectroscopy was published by Ramirez et al. (2000). The protein was not part of the 322 BMRB-PDB pairs used for training and testing the neural network.

For all 2000 models as well as for the native fold the chemical shifts were computed using the artificial neural network. Models with a RMSD smaller 5.0 Å to the native fold have a significant lower r.m.s. deviation in the predicted chemical shifts for all three groups of nuclei (compare Figure 6). The best discrimination of wrong models is seen for hydrogen chemical shift values followed by carbon, and nitrogen. A combination of all three values should be most successful. Since the Rosetta algorithm uses secondary structure prediction, the large majority of the models do have a reasonable secondary structure and the differences in the r.m.s. deviations of the chemical shift are caused by the three-dimensional arrangement of the secondary structure elements and loop conformations.

This result suggests that a selection or refinement of protein models to minimize the r.m.s. deviation or experimental and predicted chemical shift can improve the models. A consensus score was defined to become $\sqrt[3]{r.m.s_N \cdot r.m.s_C \cdot r.m.s_N}$ and achieves a correlation coefficient of 0.81 with the RMSD of the model. The best scoring model achieves a value of 0.83 ppm while the native scores at 0.82 ppm. In this particular case ∼80% of all models can be excluded if only structures with a reasonable small r.m.s. deviation for all three nuclei are considered (score < 0.90 ppm). The remaining ∼400 structures models have an average RMSD of 3.0 Å instead of 7.7 Å obtained for the complete set of 2,000 structures. Although the method alone will be insufficient because low r.m.s deviations are obtained for some models structurally very different from the native fold, it has the ability to be applied in combi-

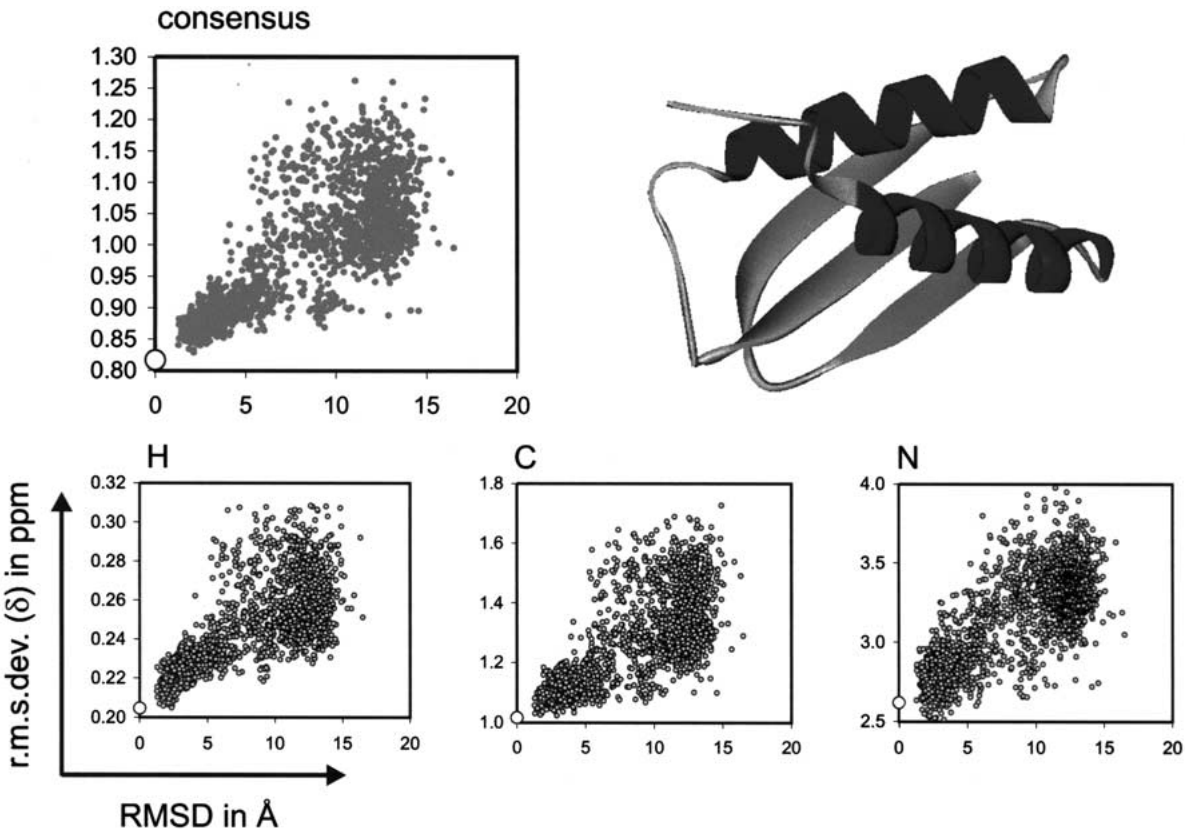

Figure 6. Ribbon model for the native 1dinI fold. Correlation diagrams for the r.m.s. deviation of experimental and neural network predicted chemical shift values versus RMSD for the native structure (large open circle) and 2,000 models generated with the Rosetta algorithm (small gray circles). The diagrams are separately plotted for H, C, and N nuclei including all predicted backbone and side- chain atoms. Models with a RMSD smaller 5.0 Å to the native fold have a significant lower r.m.s. deviation in the predicted chemical shifts for all three groups of nuclei. A consensus value is yielded by calculating $\sqrt[3]{r.m.s_N \cdot r.m.s_C \cdot r.m.s_N}$ and plotted in the top diagram.

nation with other parameters such as residual dipolar couplings or NOE intensities.

Conclusion

This approach represents the first empirical method to predict all relevant chemical shifts of a protein using one uniform model – an artificial neural network. Using a detailed description of the three-dimensional structure, an accurate and rapid prediction of protein chemical shifts is possible. An r.m.s. deviation of 1.3 ppm for carbon, 3.6 ppm for nitrogen, and 0.3 ppm for hydrogen nuclei is achieved. The model predicts the well-known dependencies between chemical shift and secondary structure as well as experimental chemical shifts. In particular, the dependency of the backbone angles Φ and Ψ is mapped correctly. For side-chain atoms, the improvement with respect to the average chemical shift values in the database is smaller, but the r.m.s. deviations of the neural network prediction are about the same for side-chain and backbone nuclei, which suggests that the decreased improvement results primarily from the higher prediction accuracy for side-chain atoms using database averages. The existence of a fast and accurate uniform chemical shift prediction method for proteins has the potential to supplement and accelerate the structure elucidation process for proteins. The method is made available for academic use (Meiler, 2002b).

Acknowledgements

I would like to thank Dr David Baker for useful discussions and providing computer facilities in his laboratories. I thank Carol Rohl for reading the manuscript and making valuable suggestions. The assigned

chemical shift data for 1dinI were kindly provided by Ben Ramirez and Georg Kontaxis. Further I would like to thank the Human Frontier Science Program (HFSP) for financial support.

References

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M. and Baker, D. (2001) *Proteins*, **45** (Suppl.), 119–126.

Braun, D., Wider, G. and Wuethrich, K. (1994) *J. Am. Chem. Soc.*, **116**, 8466–8469.

Chandonia, J.-M. and Karplus, M. (1999) *Proteins Struct. Funct. Genet.*, **35**, 293–306.

Choy, W.Y., Sanctuary, B.C. and Zhu, G. (1997) *J. Chem. Inf. Comput. Sci.*, **37**, 1086–1094.

Clouser, D.L. and Jurs, P.C. (1996) *Anal. Chim. Acta*, **321**, 127–135.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.

Gronwald, W., Boyko, R.F., Sonnichsen, F.D., Wishart, D.S. and Sykes, B.D. (1997) *J. Biomol. NMR*, **10**, 165–179.

Ivanciuc, O., Rabine, J.P., Cabrol-Bass, D., Panaye, A. and Doucet, J.-P. (1996) *J. Chem. Inf. Comput. Sci.*, **36**, 644–653.

Iwadate, M., Asakura, T. and Williamson, M.P. (1999) *J. Biomol. NMR*, **13**, 199–211.

Jones, D.T. (1999) *J. Mol. Biol.*, **292**, 195–202.

Kneller, D.G., Cohen, F.E. and Langridge, R. (1990) *J. Mol. Biol.*, **214**, 171–182.

Kvasnicka, V., Sklenak, S. and Pospichal, J. (1992) *J. Chem. Inf. Comput. Sci.*, **32**, 742–747.

Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341–348.

Luman, N.R., King, M.P. and Augspurger, J.D. (2001) *J. Comput. Chem.*, **22**, 366–372.

Meiler, J. (1996–2002) www.jens-meiler.de

Meiler, J. (2002a) www.jens-meiler.de/jufo.html

Meiler, J. (2002b) www.jens-meiler.de/proshift.html

Meiler, J. and Will, M. (2001) *J. Chem. Inf. Comput. Sci.*, **41**, 1535–1546.

Meiler, J., Maier, W., Will, M. and Meusinger, R. (2002) *J. Magn. Reson.*, **157**, 242–252.

Meiler, J., Müller, M., Zeidler, A. and Schmäschke, F. (2001) *J. Mol. Model.*, **7**, 360–369.

Meiler, J., Will, M. and Meusinger, R. (2000) *J. Chem. Inf. Comput. Sci.*, **40**, 1169–1176.

Meusinger, R. and Moros, R. (1995) In *Software – Entwicklung in der Chemie*, Vol. 10, Gasteiger, J., Ed. Gesellschaft Deutscher Chemiker, Frankfurt am Main, pp. 209–216.

Oldfield, E. (1995) *J. Biomol. NMR*, **5**, 217–225.

Osapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436–9444.

Pearson, J.G., Le, H., Sanders, L.K., Godbout, N., Havlin, R.H. and Oldfield, E. (1997) *J. Am. Chem. Soc.*, **119**, 11941–11950.

Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P. and Lund, O. (2000) *Proteins Struct. Funct. Genet.*, **41**, 17–20.

Pons, J.L. and Delsuc, M.A. (1999) *J. Biomol. NMR*, **15**, 15–26.

Qian, N. and Sejnowski, T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.

Ramirez, B.E., Voloshin, O.N., Camerini-Otero, R.D. and Bax, A. (2000) *Protein Sci.*, **9**, 2161.

Robien, W. (1998) *Nachr. Chem. Tech. Lab.*, **46**, 74–77.

Rohl, C. and Baker, D. (2002) *J. Am. Chem. Soc.*, **124**, 2723–2729.

Rost, B. (1996) *Meth. Enzymol.*, **266**, 525–539.

Rost, B. and Sander, C. (1993) *J. Mol. Biol.*, **232**, 584–599.

Rost, B., Sander, C. and Schneider, R. (1994) *J. Mol. Biol.*, **235**, 13–26.

Salamov, A.A. and Solovyev, V.V. (1997) *J. Mol. Biol.*, **268**, 31–36.

Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.

Stolorz, P., Lapedes, A. and Xia, Y. (1992) *J. Mol. Biol.*, **225**, 363–377.

Thomas, S. and Kleinpeter, E. (1995) *J. Prakt. Chem./Chem.-Ztg.*, **337**, 504–507.

Wang, Y. and Jardetzky, O. (2002) *Protein Sci.*, **11**, 852–861.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995) *J. Biomol. NMR*, **5**, 67–81.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.

Wishart, D.S., Watson, M.S., Boyko, R.F. and Sykes, B.D. (1997) *J. Biomol. NMR*, **10**, 329–336.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids (^{1}H-NMR Shifts of Amino Acids)*, John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore.

Xu, X.-P. and Case, D.A. (2001) *J. Biomol. NMR*, **21**, 321–333.

Zupan, J. and Gasteiger, J. (1993) *Neural Networks for Chemists*, VCH Verlagsgesellschaft mbH, Weinheim.