J Biomol NMR (2011) 51:411–424
DOI 10.1007/s10858-011-9579-0

ARTICLE

Uncovering symmetry-breaking vector and reliability order for assigning secondary structures of proteins from atomic NMR chemical shifts in amino acids

Wookyung Yu · Woonghee Lee · Weontae Lee · Suhkmann Kim · Iksoo Chang

Received: 27 April 2011 / Accepted: 10 June 2011 / Published online: 30 October 2011

Abstract Unravelling the complex correlation between chemical shifts of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{\alpha}$, ^{15}N, $^{1}H^{N}$ atoms in amino acids of proteins from NMR experiment and local structural environments of amino acids facilitates the assignment of secondary structures of proteins. This is an important impetus for both determining the three-dimensional structure and understanding the biological function of proteins. The previous empirical correlation scores which relate chemical shifts of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{\alpha}$, ^{15}N, $^{1}H^{N}$ atoms to secondary structures resulted in progresses toward assigning secondary structures of proteins. However, the physical-mathematical framework for these was elusive partly due to both the limited and orthogonal exploration of higher-dimensional chemical shifts of hetero-nucleus and the lack of physical-mathematical understanding underlying those correlation scores. Here we present a simple multi-dimensional hetero-nuclear chemical shift score function (MDHN-CSSF) which captures systematically the salient feature of such complex correlations without any references to a random coil state of proteins. We uncover the symmetry-breaking vector and its reliability order not only for distinguishing different secondary structures of proteins but also for capturing the delicate sensitivity interplayed among chemical shifts of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, $^{1}H^{\alpha}$, ^{15}N, $^{1}H^{N}$ atoms simultaneously, which then provides a straightforward framework toward assigning secondary structures of proteins. MDHN-CSSF could correctly assign secondary structures of training (validating) proteins with the favourable (comparable) Q3 scores in comparison with those from the previous correlation scores. MDHN-CSSF provides a simple and robust strategy for the systematic assignment of secondary structures of proteins and would facilitate the de novo determination of three-dimensional structures of proteins.

Keywords Assigning secondary structures of proteins · NMR chemical shift · Complex correlation between chemical shifts and secondary structures of proteins · Singular value decomposition analysis

Electronic supplementary material The online version of this article (doi:10.1007/s10858-011-9579-0) contains supplementary material, which is available to authorized users.

W. Yu · I. Chang (✉)
Department of Physics, Center for Proteome Biophysics, Pusan National University, Busan 609-735, Korea
e-mail: iksoochang@pusan.ac.kr

W. Lee · W. Lee
Department of Biochemistry, Structural Biochemistry and Molecular Biophysics Laboratory, Yonsei University, Seoul 120-749, Korea

S. Kim
Department of Chemistry, Biochemistry and Bio-NMR Laboratory, Pusan National University, Busan 609-735, Korea

Introduction

The determination of secondary structure of a protein is an important step toward the experimental determination of a three-dimensional structure of a protein. The assignment of secondary structure based on chemical shifts of atoms in a protein from NMR experiments exploits the fact that chemical shifts are very sensitive to the local structure of protein conformation. (Wagner et al. 1983; Szilagyi and Jardetzky 1989; Pastore and Saudek 1990; Spera and Bax 1991; Wishart et al. 1991, 1992; Wishart and Sykes 1994; Luginbuhl et al. 1995; Cornilescu et al. 1999) Unravelling the complex correlations between chemical shifts of

$^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ atoms from hetero-nuclear NMR experiment and the local environment of amino acids greatly facilitates the successful assignment of secondary structures of bigger proteins.(Bowers et al. 2000; Cavalli et al. 2007; Gong et al. 2007; Shen et al. 2008; Wishart et al. 2008) An empirical chemical shift index (CSI) method was put forward by Wishart et al. (Wishart et al. 1991, 1992; Wishart and Sykes 1994), which compared each chemical shift value of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$ atoms of amino acids in α-helix or β-strand with those in a random coil structure for several training proteins, and provided the digital measure of the propensity of chemical shifts for secondary structures via considering four orthogonal one-dimensional CSIs(1DCSI) separately. Jardetzky et al. (Wang and Jardetzky 2002) (PSSI method) enumerated the probability for each of three secondary structures from the occurrence distribution of chemical shift values of six atoms in twenty kinds of amino acids, and assigned secondary structures based on the simple orthogonal product of six independent probabilities. Samudrala et al. (Hung and Samudrala 2003) developed the PsiCSI method by taking both amino acid sequences and chemical shifts into account and applied a neural network technique.

While CSI and PSSI basically relied on the information of an individual chemical shift value of a given amino acid and combined several chemical shift values by the orthogonal manner, PsiCSI considered the information of amino acids sequence and chemical shift values for three consecutive amino acids. Markley et al. (Eghbalnia et al. 2005) (PECAN method) utilized either a pair or a triplet of chemical shift values of amino acids after optimizing a combination of sequence information and residue-specific statistical energy function. Chung et al. (Wang et al. 2007) (2DCSI), in particular, used pairs of chemical shift values and showed a success ratio(so-called Q3 score) of 88.1% (86.7%) for the correct assignment of secondary structures of amino acids with six (more than three) chemical shift values from 165 (336) proteins. For 45 new validating proteins, the Q3 scores of CSI, PsiCSI, and 2DCSI were 84.3, 87.7, and 87.7%, respectively. (The error bars of these Q3 scores for the previous approaches were not presented in their corresponding references.) Instead of using chemical shift values of a single amino acid, Bax and co-workers (Cornilescu et al. 1999) considered chemical shifts of the adjacent triplet amino acids along a sequence in 200 training proteins and developed TALOS+ (Shen et al. 2009) in conjunction with the two-layer artificial neural network, which predicts backbone torsion angles and secondary structures of proteins. The leave-one-out validation test of TALOS+ provided the overall Q3 score of 88.9% which compared favourably with that of previous approaches. (Spera and Bax 1991; Wishart et al. 1991, 1992; Wishart and Sykes 1994; Wang and Jardetzky 2002; Hung and Samudrala 2003; Eghbalnia et al. 2005; Wang et al. 2007). Although progresses were made in assigning secondary structures of proteins by previous empirical methods (Spera and Bax 1991; Wishart et al. 1991, 1992; Wishart and Sykes 1994; Wang and Jardetzky 2002; Hung and Samudrala 2003; Eghbalnia et al. 2005; Wang et al. 2007; Shen et al. 2009), the fundamental physical-mathematical framework, pertaining for assigning secondary structures of a protein from chemical shift values of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ atoms in the protein NMR experiment, has been neither explored nor the parameter space of hetero-nuclear NMR chemical shifts was fully explored. These hindered us from fully understanding the inherent structural characters embedded in that complex correlation between chemical shifts of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ atoms and the local structural environment of amino acids.

In this paper, we present a simple multi-dimensional hetero-nuclear chemical shift score function (MDHN-CSSF) which captures systematically and quantitatively the salient feature of the complex correlation between chemical shifts and secondary structures of proteins. MDHN-CSSF is robust in its simplest form of score parameters for the complex correlation without a priori assumption or adjustable objective parameters, therefore it can be applied to any set of proteins without the loss of generality. Score parameters in MDHN-CSSF were constructed by either the statistical approach or the neural perceptron learning approach (Bowie et al. 1991; Chang et al. 2001; Heo et al. 2005). The singular value decomposition (SVD) analysis (Leon 1998) of MDHN-CSSF uncovers the symmetry-breaking vector not only for distinguishing and assigning different secondary structures but also for capturing its reliability order which provides the straightforward physical-mathematical basis for explaining the delicate yet orchestrated sensitivity of chemical shift values of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ atoms, either alone or in their hetero-combinations, toward determination of secondary structures. Such a physical-mathematical framework uncovered herein encompasses the digital filtering concept of CSI (Wishart et al. 1991, 1992; Wishart and Sykes 1994) and its applications (Wang and Jardetzky 2002; Hung and Samudrala 2003; Eghbalnia et al. 2005; Wang et al. 2007).

Database for chemical shifts: secondary structures of proteins and design of conformational state of amino acid

Database relating chemical shift values to secondary structures of proteins

In order to construct, train, and validate MDHN-CSSF, we first prepare the database for chemical shift-secondary structure information of proteins from BMRB entries

(Ulrich et al. 2007) (Biological Magnetic Resonance Bank, http://www.bmrb.wisc.edu) and PDB entries (Berman et al. 2000) (Protein Data Bank, http://www.pdb.org) in conjunction with RefDB (Zhang et al. 2003) (http://redpoll.pharmacy.ualberta.ca/RefDB/). BMRB provides chemical shift values of atoms in amino acids' sequence from NMR experiment. PDB provides three-dimensional structures of proteins, from which the one-to-one correspondence among amino acids' sequence, experimental chemical shift values and structures of proteins are correctly established. Since there are differences in the reference values of chemical shifts due to different experimental conditions, the structural information of proteins are rigorously cross-checked, and chemical shift values are re-referenced by RefDB. The careful selection process results in a database for chemical shift values-secondary structures of 324 proteins with the sequence identity of less than 30% among them, in which there are 36,289 amino acids with the smallest (biggest) protein containing 20 (994) residues. For an alternate evaluation of MDHN-CSSF, we also used a database ASTRAL SCOP (Chandonia et al. 2004) (http://astral.berkeley.edu) for proteins whose structures were resolved by X-ray crystallography. We selected 6,812 proteins with the sequence identity of less than 25% among them, in which there are 1,156,412 amino acids with the smallest (biggest) protein containing 21 (937) residues, and chemical shift values of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ atoms were calculated by SHIFTX (Neal et al. 2003), SPARTA (Shen and Bax 2007) and SHIFTX2 (Han et al. 2011). Since the chemical shift values calculated via SHIFTX and SPARTA are not as accurate as those from NMR experiments, one might concern that the correlations of chemical shift values with secondary structures of amino acids are not high so as to result in the slightly biased assignment. Recently Wishart and co-workers (Han et al. 2011), however, developed SHFITX2 which significantly improved protein chemical shift prediction. One may, however, note that SHIFTX2, an updated version of the original SHIFTX with a homology module, works better for proteins with homology in the protein database. Otherwise, its performance is similar to the original SHIFTX and other similar programs. The correlation coefficients between predicted chemical shift values from SHIFTX2 and observed ones from NMR experiment are very high, and they are 0.9800 (^{15}N), 0.9959 ($^{13}C^{\alpha}$), 0.9992 ($^{13}C^{\beta}$), 0.9676 ($^{13}C'$), 0.9714 ($^{1}H^{N}$), 0.9744 ($^{1}H^{\alpha}$). Having SHIFTX2 available in the public domain, which achieved a high level of accuracy for predicting chemical shift values from protein coordinate date, we decide to employ SHIFTX2 for calculating chemical shift values of backbone atoms in 6,812 proteins from ASTAL SCOP. We will call the RefDB as "RDB" and the ASTRAL SCOP as "ADB". Within RDB (ADB), 254 (6,152) out of 324 (6,814) proteins belong to the classes of α, β, $\alpha + \beta$, α/β protein, and remaining proteins belong to the other miscellaneous classes. The secondary structures of amino acids in both RDB and ADB were evaluated by DSSP, STRIDE, and VADAR (Kabsch and Sander 1983; Frishman and Argos 1995; Willard et al. 2003) from which 30,228 (949,541) amino acids in RDB (ADB) possessed the full consensus of three secondary structures and thus participated into the construction of MDHN-CSSF (see Supplementary Table 1 for PDB codes of 324 and 6,812 training proteins).

Parameterization of local environments of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ atom

Prior to the multi-dimensional construction of the correlation between chemical shift values and secondary structures, we first set up a minimalistic parameterization for local environments of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ ($j = 1, 2, \ldots, 6$) atoms by the amino acid type ($i = 1, 2, \ldots, 20$) where atoms belong to and its secondary structure ($l = 1, 2, 3$) out of α-helix, β-strand and random coil, respectively. Then, χ_{ij}^{l} denotes a chemical shift value of an atom of a type j in an amino acid of a type i when its secondary structure is of a type l. $\mathrm{Min}(\chi_{ij}^{l})$ ($\mathrm{Max}(\chi_{ij}^{l})$) is the minimum (maximum) value out of $\overline{\chi_{ij}^{l}} - 3\sigma_{ij}^{l}$, ($\overline{\chi_{ij}^{l}} + 3\sigma_{ij}^{l}$) for each of 20×6 cases where $\overline{\chi_{ij}^{l}}$ denotes an averaged chemical shift value for an atom j of an amino acid i in lth secondary structure, and σ_{ij}^{l} is its standard deviation. Then, we divide the range between χ_{ij}^{min} and χ_{ij}^{max} into n_b bins so that chemical shift values between $\mathrm{Min}(\chi_{ij}^{l}) + (k_{ij} - 1)(\mathrm{Max}(\chi_{ij}^{l}) - \mathrm{Min}(\chi_{ij}^{l}))/n_b$ and $\mathrm{Min}(\chi_{ij}^{l} + k_{ij}(\mathrm{Max}(\chi_{ij}^{l})$- $\mathrm{Min}(\chi_{ij}^{l}))/n_b$ are parameterized by an index $k_{ij} = 1, 2, \ldots, n_b$ where n_b can be 8, 10, …, 20 subject to the resolution of binning for chemical shift values.

Designing the multi-dimensional conformational state of amino acid

Using RDB, the occurrence statistics $N(i, j, k_{ij}, l)$ for the number of each of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ ($j = 1, 2, \ldots, 6$) atoms with chemical shift type k_{ij} in amino acid i with the secondary structure type l is constructed. Figure 1, for instance, shows a probability distribution $N(i, j, k_{ij}, l)$ for each of three secondary structures for six atoms in alanine amino acid. It illustrates that three probability distribution curves for α-helix, β-strand, and random coil could be distinguished from each other to some degree, and that the same qualitative criterion applies for other

atoms and amino acids (Wang and Jardetzky 2002; Wang et al. 2007) (see Supplementary Figure 1 for remaining 19 amino acids). This kind of observational and putative criterion, however, has its drawback with regard to the systematic and quantitative assignment of secondary structures due to the limited and orthogonal exploration of multi-dimensional conformational states of amino acids. In order to overcome such ambiguities, we develop one framework which captures the salient quantitative measure to assign secondary structures of amino acids from their chemical shift values. We define the multi-dimensional conformational state of a given amino acid i by indices k_{ij} and l so that the conformational state of a given amino acid could belong to one state out of $3(n_b)^d$ states, where d is the number of hetero-nucleus whose chemical shift values are utilized. Since there are ${}_6C_d$ ways to choose d atoms out of ${}^{13}C^{\alpha}, {}^{13}C^{\beta}, {}^{13}C', {}^{1}H^{\alpha}, {}^{15}N, {}^{1}H^{N}$ atoms, the correlation matrix for each amino acid becomes a ${}_6C_d$ by $3(n_b)^d$ matrix. Provided with all or few chemical shift values of ${}^{13}C^{\alpha}, {}^{13}C^{\beta}, {}^{13}C', {}^{1}H^{\alpha}, {}^{15}N, {}^{1}H^{N}$ atoms for a given amino acid from the protein NMR experiments, our main aim is to uncover the complex yet coherent correlation through the quantitative and systematic manner between chemical shift values and secondary structures, which can then result in the successful assignment of the secondary structure from the simultaneous orchestration of heterogeneous chemical shift values in the d-dimensional phase space where d can be 1, 2, 3, 4, 5, and 6.

Construction of chemical shift score function and strategy for assigning secondary structure

Statistical approach

For each ensemble of ${}_6C_d$ ways of simultaneously considering chemical shift values of d atoms in an amino acid i, the multi-dimensional conformational state of an amino acid is parameterized in terms of an environmental index $m = 1 \sim (n_b)^d$, $(n_b)^d + 1 \sim 2(n_b)^d$, $2(n_b)^d + 1 \sim 3(n_b)^d$ for an amino acid to reside in α-helix, β-strand, and random coil, respectively. The environmental index m parameterizes a local structural configuration around an amino acid in terms of its secondary structure and conformational state of the combination of chemical shifts for $d-$ hetero atoms. We adopt a well-known analytic form of a score function, so-called log-odd ratio, for the statistical analysis of bioinformatic data, which measures the relative importance for an occurrence of a particular conformational state of amino acid with respect to an average occurrence of that (Bowie et al. 1991; Chang et al. 2001; Heo et al. 2005). We define a simple score function:

$$S_i^{stat}(q,m) = -ln[P_i(q,m)/P_i(q)], \quad (q = 1,2,\ldots,{}_6C_d) \tag{1}$$

which represents the relative propensity for an amino acid i to be at the multi-dimensional conformational state (q, m) in the systematic and quantitative manner. Here $P_i(q) = N_i(q)/N_i^{total}$ and $P_i(q, m) = N_i(q, m)/N_i(q)$. $N_i(q, m)$ is the number of an amino acid i found in an environment m with the qth ensemble of configuring d atoms. $P_i(q)$ is a probability to find an amino acid i in qth ensemble of configuring d-hetero atoms. $P_i(q, m)$ is a probability to find an amino acid i in qth ensemble of configuring d-hetero atoms at the conformational state m of the combination of chemical shifts. The conditions $N_i(q) = \sum_{m=1}^{3(n_b)^d} N_i(q,m)$ and $N_i^{total} = \sum_{q=1}^{{}_6C_d} N_i(q)$ hold. $N_i(q, m)$ could be readily obtained from 270 (6,812) proteins from RDB (ADB) since it contains the multi-dimensional occurrence statistics for chemical shift values of atoms in a given amino acid at a particular secondary structure.

Neural perceptron learning approach

We also construct alternative score parameters by the neural perceptron learning approach (Krauth and Mezard 1987) which discriminates a wrong assignment of secondary structure from a correct one. The main idea is that the propensity score for a given amino acid to be at a correct secondary structure must be lower than that at the remaining two-wrong secondary structures. For example, if alanine residue with a particular set of chemical shift values is located at α-helix in a given protein, the score function for alanine must have a lower score at α-helix than at β-strand or random coil with the same set of chemical shift values. Knowing a priori chemical shift values and correct secondary structures for 270 training proteins in RDB, we build the following inequality for each of 19,887 amino acids:

$$\sum_{q=1}^{{}_6C_d} \sum_{m=1}^{3(n_b)^d} [n_i^{wrong}(q,m) - n_i^{correct}(q,m)] S_i^{neur}(q,m) > 0 \tag{2}$$

where $n_i^{correct}(q,m)(n_i^{wrong}(q,m))$ is 1 for the correct (wrong) assignment of secondary structure taking place at (q, m), and otherwise 0. The score parameters $S_i^{neur}(q,m)$, which satisfy simultaneously 39,774 inequalities constructed from 19,887 amino acids in RDB, are determined by the neural perceptron learning method (Krauth and Mezard 1987). The validity and effectiveness of neural perceptron learning method, applied to design protein energy function for protein fold recognition, was demonstrated before (Chang et al. 2001; Heo et al. 2005). Note however, that there are few cases

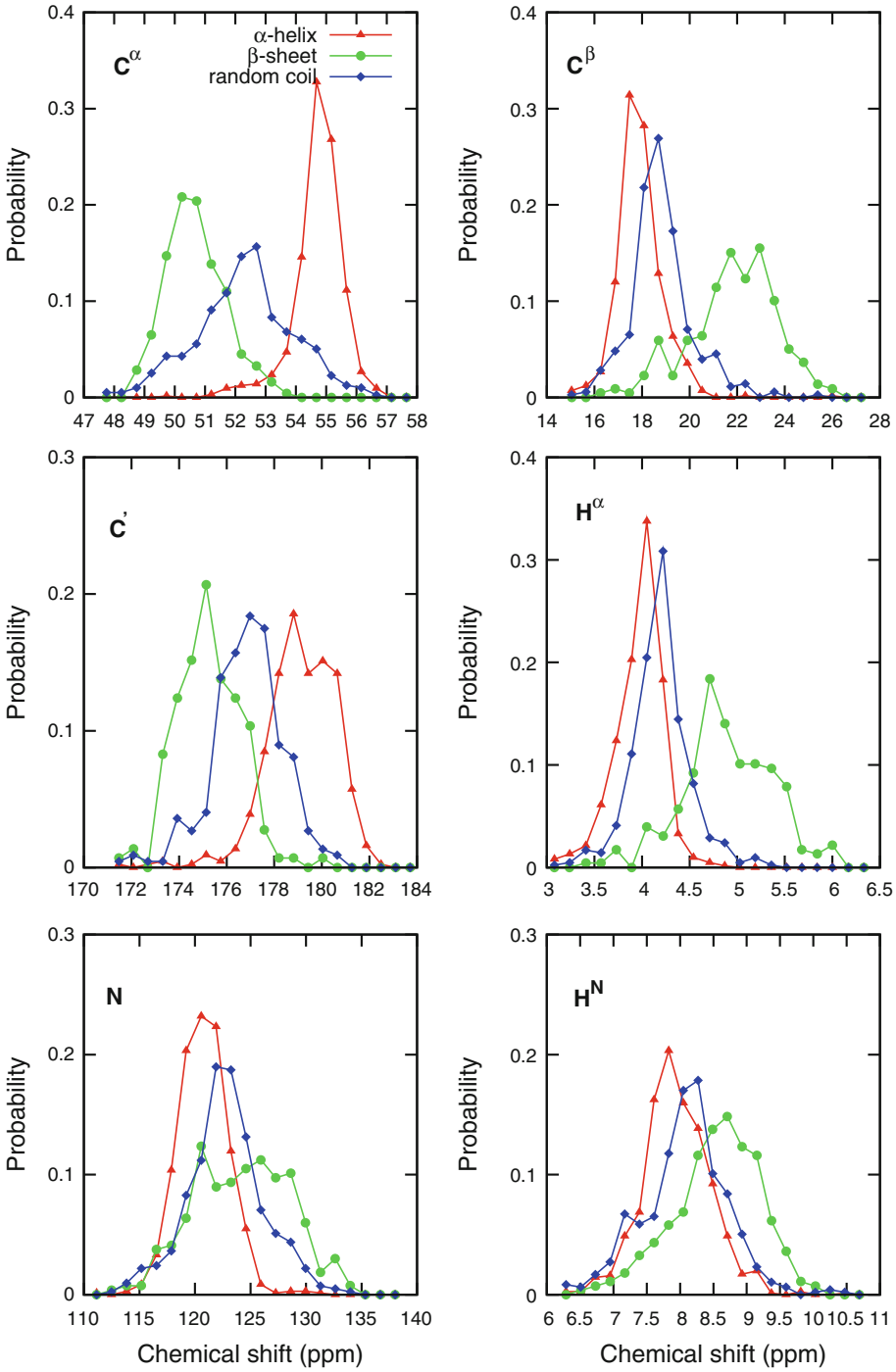

Fig. 1 Probability distribution curves for chemical shift values of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$, ^{15}N,$^{1}H^{N}$ atoms in α-helix, β-strand, and random coil for alanine amino acid. The occurrence statistics for these distributions were calculated based on 19,887 amino acids in 270 training proteins in RDB

which do not satisfy the inequality intrinsically, when two-same amino acids with the same chemical shift environment reside in two-different secondary structures. Therefore, after subtracting few unsolvable inequalities of such intrinsic cases from 39,774 inequalities, we are able to make the neural perceptron learning of score parameters learnable, and produce optimal score parameters $S_i^{neur}(q,m)$ for $d = 2$ and 3. The general strategy to solve (2) for each amino acid i is to obtain the unknown $S_i^{neur}(q,m)$ which satisfies all inequalities simultaneously in the ${}_6C_d$ by $3(n_b)^d$ dimensional space of score parameters. Here, $\left[n_i^{wrong}(q,m) - n_i^{correct}(q,m)\right] \equiv \overrightarrow{n}_i$ is fixed once a set of 19,887 amino acids from RDB is chosen. We started from an initial value of $S_i^{neur}(q,m)_{t=0}$ and calculated the value of

$$\overrightarrow{n}_i \cdot \overrightarrow{S}_i^{neur} = \sum_{q=1}^{{}_6C_d} \sum_{m=1}^{3(n_b)^d} \left[n_i^{wrong}(q,m) - n_i^{correct}(q,m)\right] \cdot S_i^{neur}(q,m)_{t=0} \quad (3)$$

for all inequalities. The vectors $\vec{n}_i$, whose score gap $\vec{n}_i \cdot \vec{S}_i^{neur}$ is negative, are the ones which do not satisfy the inequality, and the corresponding assignment of secondary structure becomes a failed one. We choose the worst vector $\vec{n}_i^{worst}$, among the failed assignments, which have the lowest value of score gap, and update $\vec{S}_{i,t+1}^{neur}$ by $\vec{S}_{i,t+1}^{neur} = \vec{S}_{i,t}^{neur} + \lambda \vec{n}_i^{worst} / |\vec{S}_{i,t}^{neur} + \lambda \vec{n}_i^{worst}|$, $(0 < \lambda < 1)$ so that the score gap for the worst assignment can increase. We calculate the scalar product $\vec{n}_i \cdot \vec{S}_i^{neur}$ again with the new $\vec{S}_{i,t+1}^{neur}$, and the set of failed assignments and the worst assignment are identified in order to update $\vec{S}_{i,t+1}^{neur}$ again. We repeat this procedure until the failed assignment does not appear. As we repeat this iteration procedure with the updated $\vec{S}_{i,t+1}^{neur}$, the score gap $\vec{n}_i \cdot \vec{S}_{i,t+1}^{neur}$ increases monotonically from negative values to positive values. The fact that the score gap becomes positive, however, does not mean that we attain the optimal value for $\vec{S}_i^{neur}$. The main purpose of this update is to optimize $\vec{S}_{i,final}^{neur}$, which can stabilize the score of a correct assignment against those of wrong assignments of secondary structures so that a correct assignment can be maximally recognized. We observe that the score gap increases as the iteration procedure goes on, and we stop the perceptron learning process when the score gap converges and saturates to the maximum positive value within a finite number of iteration. If a solution of (2) exists, the vector $\vec{S}_{i,final}^{neur}$ converges to an optimal region of points in the ${}_6C_d$ by $3(n_b)^d$ dimensional space of score parameters, and the worst score gap $\vec{n}_i^{worst} \cdot \vec{S}_{i,final}^{neur}$ become positive finite and maximal within a finite number of iterations.

Strategy for assigning the most probable secondary structure using MDHN-CSSF

Provided with a set of chemical shift values for d atoms in a given amino acid i, which are parameterized by (q, m_q), and score parameters $S_i(q, m)$ by either the statistical approach or the perceptron learning approach, we assign the secondary structure of a given amino acid i using the following strategy. We calculate the propensity for a given amino acid i to reside in α-helix, β-strand, and random coil, respectively:

$$\begin{aligned} S_{i,\alpha} &= \sum_{\{q\}} S_i(q, m_q), S_{i,\beta} = \sum_{\{q\}} S_i(q, m_q + (n_b)^d), \\ S_{i,coil} &= \sum_{\{q\}} S_i(q, m_q + 2(n_b)^d). \end{aligned} \quad (4)$$

We employ the ground state rule to assign the most probable secondary structure of an amino acid i as the one possessing the lowest score among $S_{i,\alpha}$, and $S_{i,\beta}$, $S_{i,coil}$. Our strategy for assigning the secondary structure does not resort to assumptions and adjustable parameters a priori in the shape of occurrence statistics $N(i, j, k_{ij}, l)$ for the heuristic, observational criteria, and neither a reference with respect to chemical shift values of amino acids in a random coil. Therefore, our MDHN-CSSF method is robust for an assignment of the most probable secondary structure of amino acid, and we examine how the Q3 score for the correct assignment of secondary structures improves progressively with values of d and n_b. Since we assign the secondary structure based on chemical shift values of an individual amino acid, this might cause a fragmented assignment of secondary structures. In order to consider the connectivity of the amino acids' sequence, we apply the smoothing procedure to assign the most probable secondary structure in accordance with their statistics of the occurrence of three consecutive secondary structures. Whenever secondary structures of two adjacent (i.e. $(k-1, k+1)$th, $(k-2, k-1)$th, or $(k+1, k+2)$th) amino acids are different from that of the kth amino acid. We construct the adjacency, forward, and backward smoothing matrix around a given kth amino acid by the frequency counting of secondary structures for three consecutive amino acids. Therefore, given secondary structures of $(k-1, k+1)$th, $(k-2, k-1)$th, or $(k+1, k+2)$th amino acids, we assign the secondary structure of the kth amino acid as the one which has the highest occurrence. Each smoothing matrix is a $(20 \times 3) \times (20 \times 3) \times (20 \times 3)$ dimensional matrix, and is constructed from 20,922 proteins in ASTRAL SCOP database (v. 1.67).

Results and discussion

Training of MDHN-CSSF score parameters and re-assignment of secondary structures for training proteins

From 324 proteins in RDB we employ 270 proteins for training MDHN-CSSF score parameters whereas the remaining 54 proteins will be used later for the validation test of MDHN-CSSF (see Supplementary Table 1 for PDB codes of 270 training proteins). Score parameters $S_i^{stat}(q,m)(S_i^{neur}(q,m))$ are constructed based on (1) (Eq. (2)) for several values of d and n_b (see "Statistical approach", "Neural perceptron learning approach"). It is important to test how well $S_i^{stat}(q,m)(S_i^{neur}(q,m))$ reflects the complex and heterogeneous correlations between chemical shift values and secondary structures by examining how successfully it reproduces correct secondary structures of 19,887 amino acids in 270 training proteins. Given both these score parameters and parameterized

indices of the multi-dimensional conformational state of 19,887 amino acids in 270 training proteins as the input information and pretending that correct secondary structures of amino acids are not known, we re-assign the most probable secondary structures of amino acids by the strategy described in "Strategy for assigning the most probable secondary structure using MDHN-CSSF". Figure 2a illustrates the progressively improving Q3 scores of $S_i^{stat}(q,m)(S_i^{neur}(q,m))$ for the correct assignment of secondary structures for $d = 1, 2, 3, 4, 5$ $(d = 2, 3)$ with $n_b = 8, 10, \ldots, 20$ after considering 16,579 amino acids having more than three chemical shift values $n_{cs} \geq 4$. As we consider simultaneously chemical shift values of more hetero-nucleus in a given amino acid, and as the resolution of the parameterization of chemical shift values increases, the multi-dimensional exploration for capturing the salient feature of complex correlations between chemical shift values and secondary structures becomes more precise. Therefore, the Q3 score of $S_i^{stat}(q,m)(S_i^{neur}(q,m))$ for the correct assignment of secondary structure improves dramatically from 80% for $d = 1$, $n_b = 8$ to 97% for $d = 3$, $n_b = 20$ or $d = 4$, $n_b = 14$ (from 96% for $d = 2$, $n_b = 8$ to 98% for $d = 3$, $n_b = 14$), as shown in Fig. 2a. Table 1a presents the Q3 scores averaged over 270 training proteins for $d = 1, 2, 3, 4, 5$ $(d = 2, 3)$ with $n_b = 10, 10, 10, 10, 8$ $(n_b = 10, 10)$. For each protein, both the assignment and the smoothing of secondary structures are performed as described in "Strategy for assigning the most probable secondary structure using MDHN-CSSF". Note the Q3 score of 95.4 ± 4.0% for $d = 5$, $n_b = 8$ (97.1 ± 2.7% for $d = 3$, $n_b = 10$). The results indicate that a simple and straightforward MDHN-CSSF captures the essential complex correlations between chemical shift values and secondary structures. We repeat the same calculations for both constructing $S_i^{stat}(q,m)$ from 6,812 proteins (see Supplementary Table 1) in ADB and re-assigning secondary structures of themselves. Table 1b lists the Q3 scores of 91.3 ± 4.1% for $d = 5$, $n_{cs} \geq 5$ when averaged over 6,812 proteins. The Q3 scores averaged over the number of amino acids are also listed in the parenthesis in Table 1. Previous approaches such as CSI, PSSI, PsiCSI, PECAN, 2DCSI, and TALOS+ (Wishart et al. 1991, 1992; Wishart and Sykes 1994; Wang and Jardetzky 2002; Hung and Samudrala 2003; Eghbalnia et al. 2005; Wang et al. 2007; Shen et al. 2009) employed certain numbers of training proteins to produce score parameters. From these parameters, secondary structures of amino acids in training proteins themselves were re-assigned, and then compared with their correct secondary structures. This is, in fact, the recognition rather than the prediction of secondary structure. The Q3 scores for the correct recognition of secondary structures by the CSI, PSSI, PsiCSI, PECAN, 2DCSI, TALOS+ methods were 92, 88, 85.9, 83, 86.7 and 88.9% based on 20, 36, 92, 310, 336 (with $n_{cs} \geq 4$), and 200 training proteins, respectively. (The error bars of these Q3 scores for the previous approaches were not presented in their corresponding references.)

Symmetry-breaking vector and reliability order from singular value decomposition(SVD) analysis of MDHN-CSSF

Having constructed score parameters $S(q, m)$ of MDHN-CSSF for all 20 amino acids based on (1) using 6,812 proteins in ADB. We performed a singular value decomposition (SVD) analysis of $S(q, m)$. The purpose is to re-express $S(q, m)$ in terms of a linear combination of new orthogonal eigenmodes in the lower-dimensional space so that the biological and chemical characteristics of $S(q, m)$ can emerge naturally in the few most dominant eigenmodes (Chang et al. 2001; Heo et al. 2005; Leon 1998). The SVD theorem allows us to express S as $S = YV^T = U\Sigma V^T$ where Y, U, Σ, and V is ${}_6C_d$ by $3(n_b)^d$, ${}_6C_d$ by ${}_6C_d$, ${}_6C_d$ by $3(n_b)^d$, and $3(n_b)^d$ by $3(n_b)^d$ dimensional matrix, respectively, and T denotes a transpose matrix. Matrix elements of Σ are all zero except diagonal terms σ_k, where $k = 1, 2, \ldots, 6C_d$. For each amino acid i, the SVD $S = YV^T = U\Sigma V^T$ resents it by a new set of orthogonal eigenvectors $(V^T)_k$ such that $S(q,m) = \sum_{k=1}^{6C_d} y_k^q (V^T)_k^m$, where $q = 1, 2, \ldots, 6C_d$ and $m = 1, 2, \ldots, 3(n_b)^d$. The $V_k^{'}s$ are eigenvectors corresponding to rank ordered eigenvalue of S^TS, and the y_k^q are its expansion coefficients.

The results from the SVD analysis of $S^{stat}(q, m)$ for histidine amino acid, for example, are presented in Figs. 3 and 4 for $d = 1, 2, 3$ with $n_b = 10$. Figure 3a shows the first dominant eigenvector V_1^m of the largest eigenvalue σ_1 for $d = 1$, and Fig. 3e gives the reliability y_1^q of each atom to this first mode when considering a single chemical shift value. The element of V_1^m represents the overall shape for the occurrence distribution of chemical shift values as a function of the local environmental parameter m, where m = 1–10, 11–20 and 21–30 corresponding to chemical shift values of atoms in α-helix, β-strand, and random coil, respectively. It shows that they are mostly centered around the median chemical shift value $\chi_{mid}(q)$ irrespective of secondary structures, and this tendency is strongest (weakest) for α-helix (random coil). The value of $y_1^q V_1^m$ gives the likelihood of finding q atom $(q = {}^{13}C^{\alpha}, {}^{13}C^{\beta}, {}^{13}C', {}^{1}H^{\alpha}, {}^{15}N, {}^{1}H^{N})$ at the mth local environment. Therefore, V_1^m is an eigenmode which prefers to assign random coil as the secondary structure of the amino acid since $y_1^q < 0$ for all q and $y_1^q\ V_1^{m=1-10}$, $y_1^q\ V_1^{m=11-20} > y_1^q\ V_1^{m=21-30}$. However, this over-tendency is corrected in a

Table 1 Q3 scores for correctly assigning secondary structures of amino acids based on their chemical shift values and the smoothing process. (a) Q3 scores averaged over 270 training proteins, 54 validating proteins with $n_{cs} \geq 4$ by MDHN-CSSF score parameters $S_i^{stat}(q,m)$ and $S_i^{neur}(q,m)$, and other existing methods such as CSI, PsiCSI, PSSI, PECAN, (b) The same for 6,812 training proteins and 1,026 validating proteins in ADB with $n_{cs} \geq 5$. The same table with different values of $n_b = 20, 20, 20, 14, 8$ (20, 14) for $d = 1, 2, 3, 4, 5$ ($d = 2, 3$) is presented as supplementary Table 2 in the supplementary information

(a)

Database (♯ of proteins)		Method	d	n_b	Q3 score(%) ($n_{cs} \geq 4$) training proteins (270)	validating proteins (54)
RDB (324)	MDHN-CSSF	Statistical Counting	1	10	80.0±9.1 (80.6)	79.0±8.4 (79.9)
			2	10	86.6±8.3 (87.2)	83.6±6.8 (83.2)
			3	10	91.6±7.9 (92.3)	84.1±7.3 (84.0)
			4	10	95.2±3.4 (95.5)	84.0±7.8 (83.8)
			5	8	95.4±4.0 (95.8)	82.8±9.6 (83.0)
		Neural Perceptron Learning	2	10	97.1±2.7 (97.2)	79.3±8.3 (78.4)
			3	10	97.1±2.7 (97.2)	81.6±7.4 (81.4)
	Existing Method	CSI				82.9± 9.5 (82.8)
		PsiCSI				84.2±13.0 (85.0)
		PSSI				80.8± 8.9 (81.2)
		PECAN				83.6±10.3 (84.7)

(b)

					Q3 score(%) ($n_{cs} \geq 5$) (6,812)	(1,026)	(54)
ADB (7,838)	MDHN-CSSF	Statistical Counting	1	10	86.0±6.1 (86.4)	85.7±5.8 (86.3)	80.6±8.0 (80.5)
			2	10	89.0±5.1 (89.4)	88.8±4.8 (89.2)	84.3±6.8 (84.2)
			3	10	90.4±4.5 (90.7)	90.0±4.4 (90.4)	85.4±6.5 (85.0)
			4	10	91.4±4.1 (91.7)	90.6±4.1 (91.0)	85.6±6.7 (85.2)
			5	8	91.3±4.1 (91.6)	90.2±4.3 (90.6)	86.2±7.4 (85.6)

self-consistent manner by accumulating the contributions from higher eigenmodes for α-helix and β-strand. Thus, the first mode V_1^m not only distinguishes chemical shift values of atoms around $\chi_{mid}(q)$ from those away from $\chi_{mid}(q)$ but also is an eigenmode for assigning random coil. And y_1^q characterizes its reliability in the order of $^{13}C^\beta$,$^{13}C^\alpha$, ^{15}N,$^{13}C'$,$^{1}H^\alpha$,$^{1}H^N$. Figure 3b (Fig. 4a) presents the first dominant eigenvector $V_1^m(d=2)$ ($V_1^m(d=3)$) for $d = 2$ (3), and Fig. 3f (Fig. 4c) illustrates its reliability $y_1^q(d=2)$ ($y_1^q(d=3)$) in the order of $C^\alpha C^\beta, C^\beta C', \ldots (C^\alpha C^\beta C', C^\alpha C^\beta N, \ldots)$ among 15 (20) ensembles when configuring pair (triplets) of chemical shift value of histidine amino acid simultaneously. Figure 3b resembles to 3a and demonstrates a diffraction-like pattern with the global/local maximum and local minimum depending upon whether two, one, or none of chemical shift values of two atoms coincide with $\chi_{mid}(q)$. Therefore, the SVD analysis of MDHN-CSSF score parameters systematically capture the delicate sensitivity interplayed by chemical shift values of d atoms simultaneously in that the small change in a chemical shift value of one atom within a multiplet of d atoms may cause a large change in score parameter $S(q, m)$, which in turn necessitates the use of higher-dimensional hetero-nuclear chemical shifts and facilitates the assignment of secondary structures of proteins.

Figures 3c, d and 4b are plots of the second dominant eigenvectors V_2^m for $d = 1$, 2 and 3, respectively, which clearly distinguish α-helix from β-strand. And Figs. 3g, h and 4d present the reliability y_2^q of a single atom ($d = 1$), pair of atoms ($d = 2$), and triplet of atoms ($d = 3$) to the second mode, respectively. For instance, the secondary structure of a histidine amino acid according to the chemical shift value of its $^{13}C^\beta$ atom favours α-helix for $m \leq 5$ (β-strand for $m \geq 5$) since $y_2^q\, V_2^m < 0$ for $d = 1$. On the other hand, the opposite holds for $^{13}C^\alpha$ atom because the more negative $y_2^q\, V_2^m$ ($d = 1$) is for one secondary structure, the higher is the likelihood for the amino acid to be in that secondary structure. We interpreted $V_2^m(d = 1, 2, 3)$ of Figs. 3c, d and 4b as *a symmetry-breaking eigenvector* which distinguishes α-helix from β-strand, and the corresponding $y_2^q(d = 1, 2, 3)$ of Figs. 3, h and 4d entails *the reliability order* in the order of the magnitude of $y_2^q(d = 1, 2, 3)$. Note that for random coil $V_2^m \simeq 0$ and it barely has a preference. The message from this algebraic analysis of the second mode is that it provides a fundamental physical-mathematical framework of the digital filtration concept behind the original CSI (Wishart et al. 1991, 1992; Wishart and Sykes 1994) to assign a ternary index of −1, 0, or 1 depending on the measured chemical shift value of an atom with respect to that in a random coil as the reference value. What is important in the consequence of the SVD analysis of MDHN-CSSF herein is that it not only encompasses the digital filtration concept behind CSI but also generalizes toward its continuous filtration concept manifested in $y_2^q V_2^m$ uncovering the symmetry-breaking vector and its reliability order for general d. Furthermore, it provides a physical-mathematical understanding for the interpretation of the second dominant eigenvector. Note also that the eigenvector components $V_2^m(d = 1, 2, 3)$ for α-helix, β-strand are

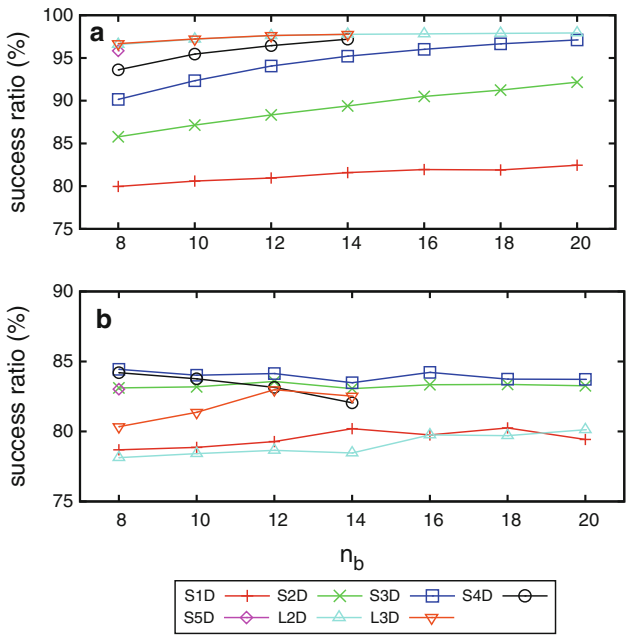

Fig. 2 Q3 scores for correctly assigning secondary structures of amino acids based on their chemical shift values and the smoothing process. **a** Q3 scores for 19,887 amino acids with $n_{cs} \geq 4$ in 270 training proteins by MDHN-CSSF score parameters $S_i^{stat}(q,m)$ and $S_i^{neur}(q,m)$, **b** Q3 scores for 3,800 amino acids with $n_{cs} \geq 4$ in 54 validating proteins by MDHN-CSSF score parameters $S_i^{stat}(q,m)$ and $S_i^{neur}(q,m)$. Q3 scores are drawn by S1D–S5D for $d = 1$–5 using $S_i^{stat}(q,m)$, and by L2D, L3D for $d = 2$, 3 using $S_i^{neur}(q,m)$

anti-correlated from Figs. 3c, d and 4b; namely the symmetry-breaking character distinguishing α-helix from β-strand manifested in the second mode is of the universal character, which generally prevails in d-dimensional phase space of chemical shift values for d atoms. Here, $y_2^q V_2^m(d=2)$ ($y_2^q V_2^m(d=3)$) again provides the second dominant propensity of finding the qth ensemble of combining chemical shift values of d atoms at its mth local environment. Figure 3h (Fig. 4d) presents the reliability order for assigning α-helix and β-strand, for which the second mode is reliable in the order of $C^\alpha H^N, C'H^N, C^\alpha H^\alpha, C'H^\alpha, C^\alpha N, \ldots$ ($C'H^\alpha H^N, C'H^\alpha N, C^\alpha C^\beta H^N, C^\alpha H^\alpha H^N, \ldots$) based on the magnitude of $y_2^q(d=2)$ ($y_2^q(d=3)$). We know that $^1H^N$ and N atoms play little role in assigning secondary structure when considering only one atom at a time as shown in Fig. 3c, g that $^1H^N$ and ^{15}N atoms bear the weakest propensity in $y_2^q V_2^m(d=1)$. Nevertheless, we recognize that $^1H^N$ and N atoms can play a significant role in the multi-dimensional phase space of hetero-nuclear by forming pairs such as $C'H^N$ and $C^\alpha N$ (triplets like $C^\alpha C^\beta H^N$ and $C'H^\alpha N$); namely, these pairs (triplets) have the highest reliability order to exercise the dominant role in assigning secondary structures with the strongest propensity in $y_2^q V_2^m(d=2)$. On the other hand, note from Fig. 3h (Fig. 4d) that pairs NC^β, NH^α (triplet $C^\alpha C'H^\alpha$) possess a negligible contribution to $y_2^q V_2^m(d=2)$ ($y_2^q V_2^m(d=3)$. The delicate-yet-rigorous role played by $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C'$, $^1H^\alpha$, ^{15}N, $^1H^N$ atoms, either alone or in their combination of d atoms, in assigning secondary structures, is precisely captured by the SVD analysis of MDHN-CSSF. The SVD analysis results for the remaining 19 amino acids bear the similar eigenmode interpretation (see Supplementary Figure 2 for $d = 1$, 2 and Figure 3 for $d = 3$). The same SVD analysis are performed on score parameters $S(q, m)$ from 270 proteins in RDB. We observe the similar characters of the symmetry-breaking vectors and reliability orders, but the overall behaviors of them are less clear that those from 6,812 proteins in ADB because the statistics of data for chemical shift values-secondary structures from RDB is worse than that from ADB.

Validation test of MDHN-CSSF for assigning secondary structures of new validating proteins

54 new proteins that do not overlap with the 270 training proteins of this work are subjected to the validation test of $S_i^{stat}(q,m)(S_i^{neur}(q,m))$. We plot the Q3 scores for the correct assignment of secondary structures of 3,800 amino acids with $n_{cs} \geq 4$ in 54 new proteins in Fig. 2b for $d = 1$, 2, 3, 4, 5 ($d = 2$, 3) with $n_b = 8$, 10, …, 20. It demonstrates the improving Q3 scores, from 79% for $d = 1$,$n_b = 8$ to 84% for $d = 3$, $n_b = 20$ or $d = 4$, $n_b = 8$ (from 78% for $d = 2$, $n_b = 8$ to 83% for $d = 3$, $n_b = 12$) as d and n_b increases. After performing both the assignment by $S_i^{stat}(q,m)(S_i^{neur}(q,m))$ and the smoothing of secondary structures for each of 54 proteins, Table 1a lists the averaged Q3 scores for $d = 1$, 2, 3, 4, 5 with $n_b = 10$, 10, 10, 10, 8 for $n_{cs} \geq 4$. The best result is 84.1 ± 7.3% for $d = 3$, $n_b = 10$ (81.6 ± 7.5% for $d = 3$, $n_b = 10$). Note also Supplementary Table 2 in the supplementary information in which the averaged Q3 scores for $d = 1$, 2, 3, 4, 5 ($d = 2$, 3) with $n_b = 20$, 20, 20, 14, 8 ($n_b = 20$, 14) are listed. CSI, PsiCSI, PSSI, and PECAN provide the web-service or an executable program to predict secondary structures upon the input information, namely sequence and chemical shift value information. For 54 new validating proteins we employ in our manuscript the Q3 score for CSI, PsiCSI, PSSI, and PECAN predicting secondary structures correctly is 82.9 ± 9.5%, 84.2 ± 13.0%, 80.8 ± 8.9%, and 83.6 ± 10.3%, respectively, where the error bar is one standard deviation. These results are listed in Table 1a, and it turns out that the prediction performance of MDHN-CSSF for $d > 1$ is comparable to those of these existing methods within the error bar for the same set of validating proteins. For the validation test of $S_i^{stat}(q,m)$ constructed from 6,812 training proteins with a sequence identity of less than 25% in the ADB, we selected again

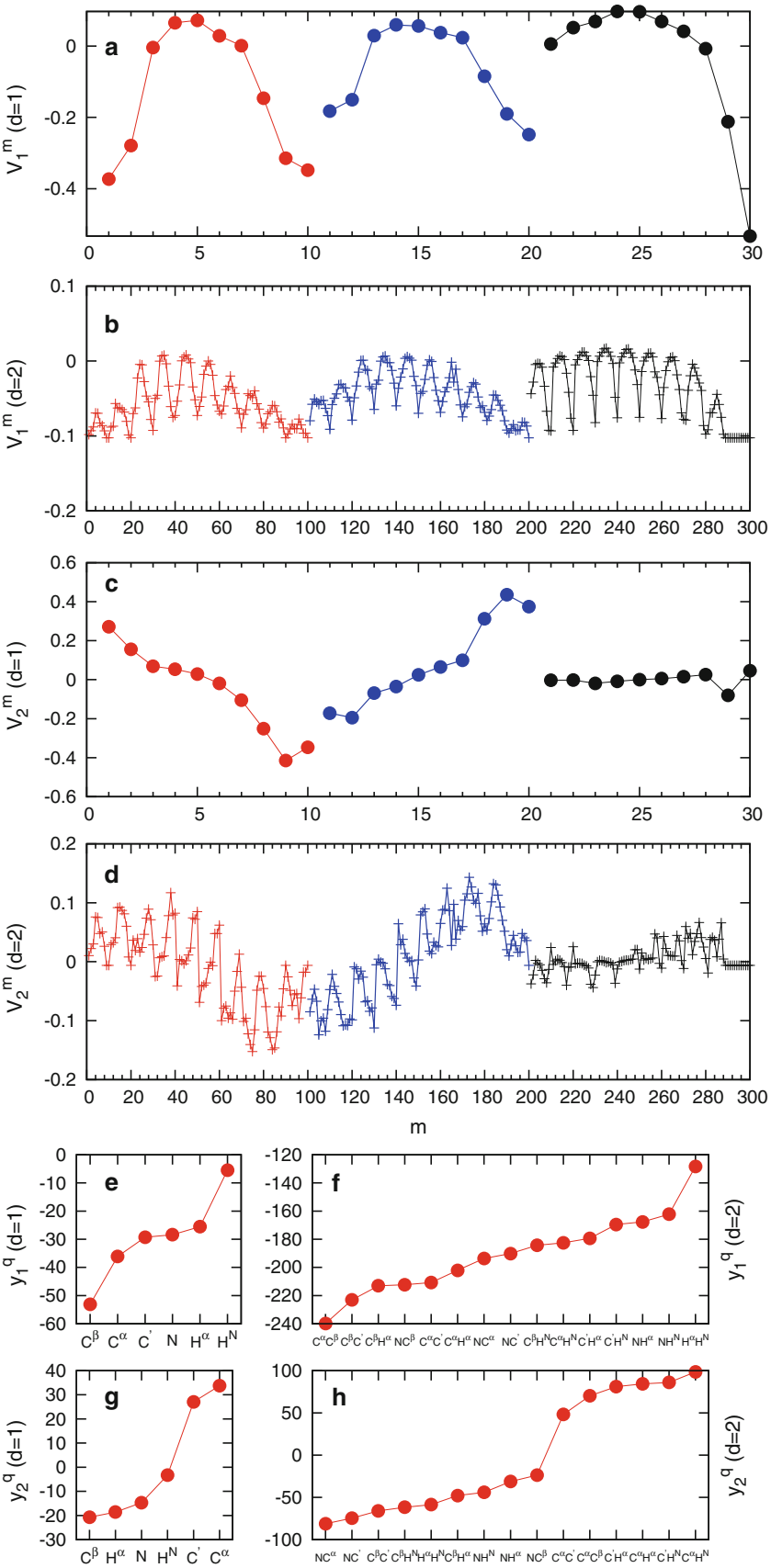

Fig. 3 Results of the SVD analysis of $S_i^{stat}(q,m)$ constructed from 6,812 training proteins in ADB for histidine amino acid. **a** (**c**) The first (second) eigenmode from considering chemical shift value of a single atom, **b** (**d**) the first (second) eigenmode from considering chemical shift values of pair atoms. The first eigenmode favours a random coil, whereas the second eigenmode possesses the symmetry for α-helix, β-strand. **e** (**g**) The reliability order of the first (second) eigenmode for considering a single chemical shift value of $^{13}C^\alpha, ^{13}C^\beta, ^{13}C', ^{1}H^\alpha, ^{15}N, ^{1}H^N$ atoms. **f** (**h**) The reliability order of the first (second) eigenmode for considering chemical shift values of pair atoms

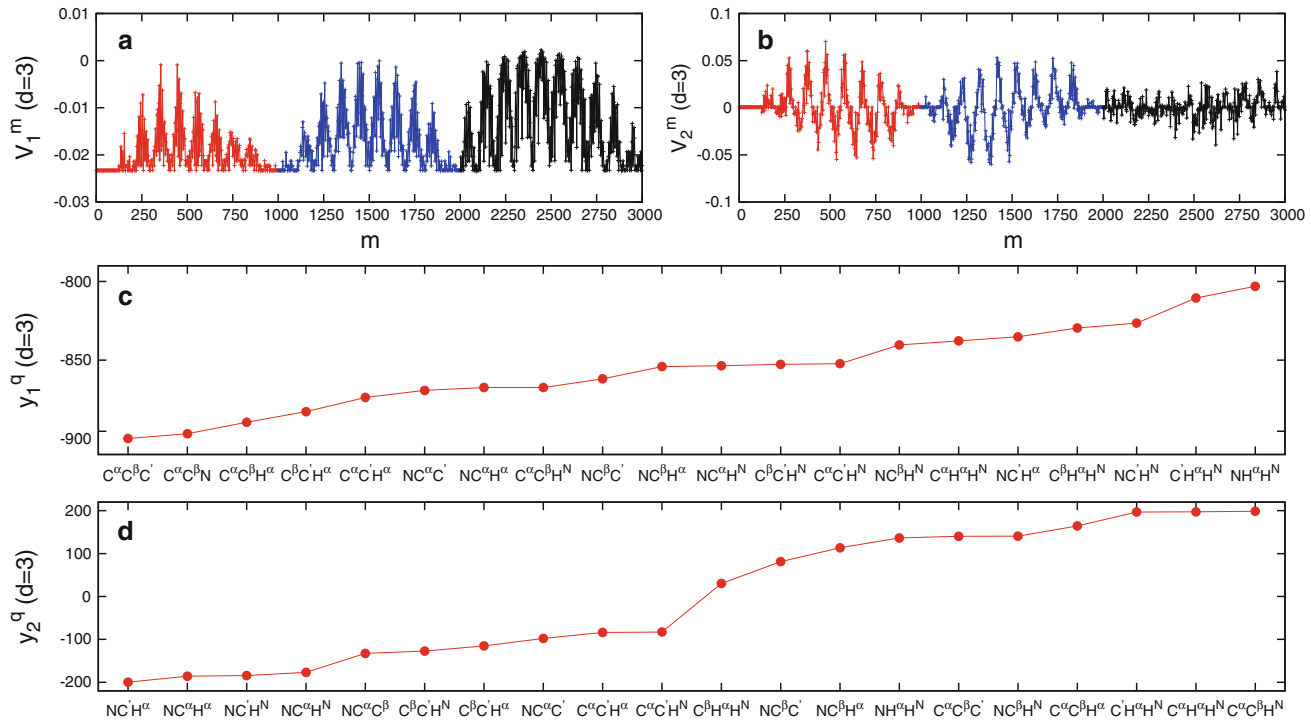

Fig. 4 Results of the SVD analysis of $S_i^{stat}(q,m)$ constructed from 6,812 training proteins in ADB for histidine amino acid. **a** (**b**) The first (second) eigenmode from considering chemical shift value of three hetero-atoms ($d = 3$). **c** (**d**) The reliability order of the first (second) eigenmode for considering chemical shift values of triplet atoms

1,026 new proteins with $n_{cs} \geq 5$ from ADB whose sequence identity is between 25 and 30% among them. We predict secondary structures of 154,076 validating amino acids in 1,026 new proteins using $S_i^{stat}(q,m)$ as previously described. The Q3 scores for the correct assignment of secondary structure improved from 85.7% for $d = 1$, $n_b = 10$ to 90.6% for $d = 4$, $n_b = 10$. Table 1b lists the averaged Q3 scores over 1,026 proteins, and the best results is $90.6 \pm 4.1\%$ for $d = 4$, $n_b = 10$. The validation test of MDHN-CSSF on sets of new proteins shows the comparable Q3 scores for the correct assignment of secondary structures in comparison with those of the previous correlation scores (Wishart et al. 1991, 1992; Wishart and Sykes 1994; Hung and Samudrala 2003; Eghbalnia et al. 2005; Wang et al. 2007; Shen et al. 2009) and demonstrates that MDHN-CSSF can capture most of the essential heterogeneous correlation between chemical shift values and secondary structures of amino acids. (see Supplementary Table 1 for the PDB codes of 54, 1026 validating proteins)

7,838(324) proteins in ADB(RDB) are a set of nonredundant proteins which serves as the structural representatives of the known protein structures by X-ray (NMR). It is worthwhile to note that in Table 1 the Q3 scores of correctly assigning secondary structures for 1,026(54) validating proteins of ADB(RDB) is compatible with (a little less than) those for 6,812(270) training proteins of ADB(RDB). This implies that the structural characters represented via 6,812 training proteins of ADB are good in such a way to maintain its Q3 scores for 1,026 validating proteins at the level of those for 6,812 training proteins. Although it looks like that score parameters $S(q, m)$ from ADB correlate the predicted chemical shifts with the secondary structures of validating proteins better for ADB compared to what score parameters from RDB do for validating proteins of RDB, it does not mean that score parameters are biased to have better correlations for assigning secondary structures of validating proteins in ADB because the same holds for training proteins of RDB with $d > 2$ when using score parameters constructed from RDB. Therefore, the lesson from these results is that the more representative training proteins ADB or RDB has, the better score parameters describe the structural characters of validating proteins, and also reduce the possibility of overtraining. This is why we decide to construct and use ADB for recognizing and predicting secondary structure of proteins from chemical shift values. It is also instructive to see how well the score parameters constructed from ADB performs in predicting secondary structures of 54 validating proteins of RDB. Table 1b lists such Q3 scores which are better overall by a few percent than those in Table 1a predicted by score parameters constructed from RDB. It

illustrates that score parameters from ADB are not biased since they also correlate chemical shift values with secondary structures of 54 validating proteins of RDB better than what score parameters from RDB themselves do.

An issue of sparse distribution and over-training of correlation score parameters

The over-training problem of score parameters may arise when (1) the number of nonredundant proteins in the protein database is not large enough to manifest the representative structural characters of all protein families or (2) the number of score parameters exceeds much more than the number of amino acids, hence backbone atoms in the protein database. Either of both cases may result in the over-training of score parameters.

The number of data points in the multi-dimensional parameter space of our correlation score parameters is very large, which is ${}_6C_d \times 3 \times n_b^d$. For example, for $d = 3$, $n_b = 10$ which is a case for the best parameter set there are ${}_6C_3 \times 3 \times (10)^3 = 450{,}000$ data points in the multi-dimensional parameter space. However, the number of amino acids is 19,887 in 270 training proteins of RDB. The distribution of a set of chemical shift values of amino acid on the multi-dimensional parameter space is sparse, and therefore the statistics of the occurrence of the chemical shift value-secondary structure is not that good [see (1)]. In order to circumvent this situation we need much more training proteins, but currently the number of available proteins from NMR experiment for this purpose is limited. Since we recently know that the calculated chemical shift values via SHIFTX2 using X-ray structures is accurate and compatible to that from NMR-structures of proteins, we decided to build up the better statistics for the occurrence of chemical shift value-secondary structure on the multi-dimensional parameter space. There are 6,812 training proteins in ADB, and the number of amino acids n there is 1,156,412 which is about 580 times larger than 19,887. Since one of our aim is to test how good our approach is and hence correlation score parameters are for representing the heterogeneous correlation between chemical shift values and secondary structures, the data set with more proteins like ADB help us to pursue this purpose.

The strategy here is to evaluate (or learn) the correlation score parameters which correlate chemical shift values with secondary structures of amino acids based on the known propensity between these in training proteins. One tests how good the correlation score parameters are by re-predicting secondary structures of training proteins by themselves. Therefore, the prediction capability for training proteins themselves is indeed good if the parameterization for the score function was done properly. And then one applies these correlation score parameters to predict secondary structures of new validating proteins which do not participate in evaluating the correlation score parameters. Therefore the prediction capability for validating proteins is naturally not as good as that for training proteins because the correlation score parameters are more prone to describe easily the structural character of training proteins than validating proteins. It is worthwhile to note that secondary structure identification has an intrinsic error rate and that the agreement between any two secondary structure identification methods/programs is usually no better than 90%. Even two experts looking at the same protein will identify different secondary structures or different start/end points of secondary structure elements. The fact that Q3 scores using score parameters constructed from RDB for correctly re-assigning secondary structures of training proteins are mostly more than 90% where as those for validating proteins are less than 90% might mean that there is an over-training problem in our approaches In fact, one can't avoid certain degree of over-training problem in designing the correlation score parameters in this kind of knowledge-based bioinformatics approach because the number of proteins employed in the training set of RDB is not large enough to manifest the representative structural characters of all protein families. But what is important is how inherent the correlation score parameters evaluated from training proteins are for predicting secondary structures of new validating proteins. In view of this aspect, although there might be an over-training problem, our correlation score parameters captures good enough structural characters in predicting secondary structures of new validating proteins with the comparable Q3 score to that of existing methods.

Availability of MDHN-CSSF

Score parameters of MDHN-CSSF $S_i^{stat}(q,m)(S_i^{neur}(q,m))$ as a function of q, m for $d = 1, 2, 3, 4, 5$ ($d = 2, 3$) for all 20 kinds of amino acids ($i = 1, 2, \ldots, 20$) are available at http://protein.phys.pusan.ac.kr/MDHN-CSSF/index.htm. Among several parameter sets derived in our approaches, the best parameter set is the one from $d = 4$ with $n_b = 10$ from ADB. And we provide the full information such as PDB code, BMRB index, sequence, secondary structures, chemical shift values for not only RDB and ADB datasets but also correlation score parameters at our web page. More importantly we also provide a downloadable computer program as well as a web-server where one can use it for predicting secondary structures upon providing sequence/chemical shift value information. For facilitating of the easy use of all programs and data files, README file in there will guide one how to use all these programs and data files. Provided with amino acids' sequence of a

protein and chemical shift values, re-referenced by RefDB, of atoms there, the secondary structure assignment of a given protein could be evaluated.

Summary

We present a simple multi-dimensional hetero-nuclear chemical shift score function (MDHN-CSSF) which captures the salient features of the complex correlation between chemical shifts and secondary structures of proteins. MDHN-CSSF is robust in its simplest form of score parameters for the complex correlation without assumptions or adjustable objective parameters a priori so that it can be applied to any set of proteins without the loss of generality. Score parameters in MDHN-CSSF are constructed by either the statistical approach or the neural perceptron learning approach. The aim is not to show merely whether our approach is better or worse than other approaches (Wishart et al. 1991, 1992; Wishart and Sykes 1994; Hung and Samudrala 2003; Eghbalnia et al. 2005; Wang et al. 2007; Shen et al. 2009) by a few percentages in the Q3 score for the correct assignment of secondary structures, but to understand and provide the physical-mathematical basis underlying such complex correlations between chemical shift values and secondary structures of proteins. The singular value decomposition analysis of MDHN-CSSF uncovers not only the symmetry-breaking vector for distinguishing different secondary structures but also its reliability order which provide the straightforward physical-mathematical understanding for the delicate yet orchestrated sensitivity of chemical shift values of $^{13}C^{\alpha}$,$^{13}C^{\beta}$,$^{13}C'$,$^{1}H^{\alpha}$,^{15}N,$^{1}H^{N}$ atoms, either alone or in their hetero-combinations, toward determination of secondary structures. Such a physical-mathematical framework uncovered herein encompasses the digital filtration concept of CSI and its applications, and generalizes towards its continuous filtration concept providing the symmetry-breaking vector and its reliability order for the general d. MDHN-CSSF correctly assign secondary structures of training (validating) proteins with the favourable (comparable) Q3 scores in comparison with those from the previous correlation scores. The coherent and quantitative construction of MDHN-CSSF together with its SVD eigenmode analysis, and the assignment of secondary structures without any references to a random coil state provide a robust strategy for the assignment of secondary structures of a protein from its atomic chemical shifts. We hope that the result of this work would facilitate the de novo determination of three-dimensional structures of proteins (Cavalli et al. 2007; Shen et al. 2008).

Acknowledgments This work was supported by the Creative Research Initiative (Center for Proteome Biophysics, Grant No. 2011-0000041 to W.Y. and I.C.) of National Research Foundation/ Ministry of Education, Science and Technology, Korea. This research was also supported by World Class University (WCU) program (R33-2009-000-10123-0 to W.L.). Authors would like to thank Prof. Kurt Wuthrich for fruitful discussion and Weonjoong Kim for constructing the web-server for MDHN-CSSF.

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

Bowers PM, Strauss CE, Baker D (2000) De novo protein structure determination using sparse NMR data. J Biomol NMR 18(4):311–318

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Sci Agric 253:164–170

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

Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE (2004) The ASTRAL compendium in 2004. Nucleic Acids Res 32:D189–D192

Chang I, Cieplak M, Dima RI, Maritan A, Banavar JR (2001) Protein threading by learning. Proc Natl Acad Sci USA 98:14350–14355

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(3):289–302

Eghbalnia HR, Wang L, Bahrami A, Assadi A, Markley JL (2005) Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. J Biomol NMR 32:71–81

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins Struct Funct Genet 23:566–579

Gong H, Shen Y, Rose GD (2007) Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly. Protein Sci 16:1515–1521

Han B, Liu Y, Ginzinger S, Wishart D (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57

Heo M, Kim S, Moon EJ, Cheon M, Chung K, Chang I (2005) Perceptron learning of pairwise contact energies for proteins incorporating the amino acid environment. Phys Rev E 72:11906–11915

Hung LH, Samudrala R (2003) Accurate and automated classification of protein secondary structure with PsiCSI. Protein Sci 12:288–295

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Krauth W, Mezard M (1987) Learning algorithms with optimal stability in neural networks. J Phys A 20:L745–L752

Leon SJ (1998) Linear algebra with applications. Prentice Hall, New Jersey

Luginbuhl P, Szyperski T, Wuthrich K (1995) Statistical basis for the use of 13 C alpha chemical shift in protein structure determination. J Magn Reson B 1009:229–233

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

Pastore A, Saudek V (1990) The relationship between chemical shift and secondary structure in proteins. J Magn Reson 90:165–176

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

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 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 C-ALPHA and C-BETA 13C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

Szilagyi L, Jardetzky O (1989) ALPHA-proton chemical shift and secondary structure in proteins. J Magn Reson 83:441–449

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL (2007) BioMagResBank. Nucleic Acids Res 36:D402–D408

Wagner G, Pardi A, Wuthrich K (1983) Hydrogen-bond length and H-1-NMR chemical shifts in proteins. J Am Chem Soc 105:5948–5949

Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11:852–861

Wang CC, Chen JH, Lai WC, Chuang WJ (2007) 2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts. J Biomol NMR 38:57–63

Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, Wishart DS (2003) VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic Acids Res 31:3316–3319

Wishart DS, Sykes BD (1994) The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. J Biomol NMR 4(2):171–180

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, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochem Cell Biol 31(6):1647–1651

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

Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195