J Biomol NMR (2014) 60:131–146
DOI 10.1007/s10858-014-9863-x

ARTICLE

CSI 2.0: a significantly improved version of the Chemical Shift Index

Noor E. Hafsa · David S. Wishart

Received: 27 June 2014 / Accepted: 17 September 2014 / Published online: 2 October 2014

Abstract Protein chemical shifts have long been used by NMR spectroscopists to assist with secondary structure assignment and to provide useful distance and torsion angle constraint data for structure determination. One of the most widely used methods for secondary structure identification is called the Chemical Shift Index (CSI). The CSI method uses a simple digital chemical shift filter to locate secondary structures along the protein chain using backbone ^{13}C and ^{1}H chemical shifts. While the CSI method is simple to use and easy to implement, it is only about 75–80 % accurate. Here we describe a significantly improved version of the CSI (2.0) that uses machine-learning techniques to combine all six backbone chemical shifts ($^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, ^{15}N, ^{1}HN, $^{1}H_{\alpha}$) with sequence-derived features to perform far more accurate secondary structure identification. Our tests indicate that CSI 2.0 achieved an average identification accuracy (Q3) of 90.56 % for a training set of 181 proteins in a repeated tenfold cross-validation and 89.35 % for a test set of 59 proteins. This represents a significant improvement over other state-of-the-art chemical shift-based methods. In particular, the level of performance of CSI 2.0 is equal to that of standard methods, such as DSSP and STRIDE, used to identify secondary structures via 3D coordinate data. This suggests that CSI 2.0 could be used both in providing accurate NMR constraint data in the early stages of protein structure determination as well as in defining secondary structure locations in the final protein model(s). A CSI 2.0 web server (http://csi.wishartlab.com) is available for submitting the input queries for secondary structure identification.

Keywords Nuclear magnetic resonance · Chemical shifts · Secondary structure multi-class support-vector machine · Markov model

Introduction

Secondary structures are considered fundamental to both the description and the understanding of protein tertiary structures. Indeed, secondary structure maps and secondary structure ribbon diagrams are standardly used in almost all structural biology books, journals and databases (Wuthrich 1986; Berman et al. 2000). It is also notable that secondary structure assignments or predictions are still widely used as the basis to many protein fold recognition algorithms (Soding et al. 2005), protein threading methods (Jones et al. 1999), 3-D protein structure prediction algorithms (Wishart 2011; Soding and Remmert 2011) and intrinsically disordered protein (IDP) identification methods (He et al. 2009). Secondary structure is also key to many heuristic energy functions that are designed to assess, fold and/or refine protein structures (Wishart et al. 2008; Berjanskii et al. 2009; Adams et al. 2013). Furthermore, secondary structure provides not only approximate torsion angle and qualitative backbone flexibility data, it also provides hydrogen bonding information (for α-helices and β-strands), implied contact information (for β-strands) and important topological information (through β-turns). While increasing interest is turning to extracting or predicting more quantitative measures of

N. E. Hafsa · D. S. Wishart (✉)
Department of Computing Science, University of Alberta, Edmonton, Canada
e-mail: david.wishart@ualberta.ca

D. S. Wishart
Department of Biological Sciences, University of Alberta, Edmonton, Canada

protein structure (i.e. torsion angles, backbone order parameters, accessible surface area) it is important to note that the accuracy of these methods is not yet sufficient to permit their widespread use in 3D protein structure prediction or 3D structure calculation algorithms (Wishart 2011). Consequently the identification and delineation of secondary structure elements continues to be of interest to protein chemists, bioinformaticians, X-ray crystallographers and, of course, NMR spectroscopists (Wuthrich 1986, 1990; Wishart 2011).

In the field of protein NMR, NOE-based methods are widely used to identify and assign secondary structures. Indeed, they continue to be the predominant method for identifying or delineating secondary structures in peptides and proteins (Wuthrich 1986). Less well known is the fact that NMR chemical shifts can also be used to identify secondary structures and that they are remarkably accurate and far easier to use than NOEs (Wishart et al. 1992; Wishart and Sykes 1994a, b). The idea of using chemical shifts to identify secondary structures was first exploited with the development of the Chemical Shift Index or CSI (Wishart et al. 1992). The CSI method applies a "digital filter" to backbone ^{1}H and ^{13}C chemical shifts to precisely identify the type and location of protein secondary structure elements (helices, β-strands, coils) along a protein chain (Wishart and Sykes 1994b). The CSI method is particularly popular because it is easy to implement and surprisingly accurate with the reported agreement between X-ray-defined secondary structures and CSI-identified secondary structure being about 75–85 % (Wishart and Sykes 1994b; Wishart and Case 2002; Mielke and Krishnan 2004, 2009).

However, the CSI method is not without some shortcomings. For instance, it requires near complete backbone assignments, it is sensitive to the choice of random coil shifts used to calculate the secondary shifts, and it identifies α-helices (>90 % accuracy) more accurately than β-strands (<75 %). Because of these limitations, a number of alternative CSI-like approaches have been developed over the past decade, including PSSI (Wang and Jardetzky 2002a), PsiCSI (Hung and Samudrala 2003), PLATON (Labudde et al. 2003), PECAN (Eghbalnia et al. 2005), and 2DCSi (Wang et al. 2007a). These methods typically extend the CSI concept by incorporating more advanced chemical shift models or additional statistical information. For instance, PSSI replaced CSI's simplistic digital filter with a more sophisticated joint probability model to improve its secondary structure identification accuracy. On the other hand, PsiCSI combined the basic CSI concept with a sequence-based secondary structure routine called PSIPRED (Jones 1999) to boost its performance. PLATON used a database consisting of reference chemical shift patterns from previously assigned proteins to improve its secondary structure calls, while PECAN employed a pseudo-energy model that combined sequence data with chemical shift data to more accurately identify secondary structure elements. Finally, 2DCSi used two-dimensional cluster analysis to analyze paired scattering diagrams of all six backbone chemical shifts to obtain improved secondary structure identification. All of these methods appear to achieve three-state secondary structure (Q3) accuracies better than 80 %.

More recently, sophisticated chemical shift-based secondary structure assignment approaches that exploit machine-learning techniques, torsion angle estimates, sequence homology and far more extensive chemical shift-structure databases have appeared. These include TALOS+ (Shen et al. 2009a), TALOS-N (Shen and Bax 2013), DANGLE (Cheung et al. 2010) and Delta2D (Camilloni et al. 2012). Both TALOS+ and TALOS-N predict backbone torsion angles, as well as secondary structure locations by using neural networks to match chemical shift patterns over a five-residue window against a large database of previously assigned proteins with high-resolution structures. DANGLE exploits some of the same ideas as TALOS+ but employs Bayesian-inference techniques instead of neural nets to perform its analyses. Delta2D identifies secondary structure elements and secondary structure populations in both disordered and native-state proteins by analyzing the probability distribution of a very large database of backbone chemical shifts. In general, these newer approaches have average Q3 prediction accuracies between 83 and 86 %.

With ongoing advances in machine learning and with continued improvements of our understanding of protein chemical shifts (Wishart 2011; Shen and Bax 2012; Fesinmeyer et al. 2005), we believe that further improvements in shift-based secondary structure identification accuracy are possible. In particular, by making use of chemical shift information, sequence information and predicted backbone flexibility and then integrating this information using a multi-class Support Vector Machine (SVM) model we found that it was possible to make statistically significant improvements (3–8 %) in the accuracy of shift-based secondary structure assignments. Since this concept builds from our previous work on the CSI, we decided to call the new method CSI 2.0. The level of accuracy achieved by CSI 2.0 suggests that it could be used to assist with the initial stages of conventional NMR structure generation (i.e. fold identification via threading or providing useful torsion angle and distance restraints) as well as a robust alternative to standard coordinate-based methods for secondary structure identification.

Methods and materials

Data set preparation

Training and testing data set

To construct the database needed to train and test our CSI 2.0 method, we chose a local, manually curated data set that we previously used to train and test the SHIFTX2 program (Han et al. 2011). An initial data set of ~300 X-ray protein structures with good quality NMR assignments was filtered based on following criteria: (1) a resolution <2.1 Å, (2) largely monomeric, (3) free of bound DNA, RNA or large cofactors, (4) an average pairwise sequence identity <33 % to any other protein in the data set, (5) nearly-complete (>90 %) sequential assignment of ^{1}H, ^{13}C and/or ^{15}N backbone chemical shifts, and (6) must be a BMRB (Ulrich et al. 2008) entry. Several measures were taken to eliminate chemical shift re-referencing problems, check chemical shift quality and detect chemical shift outliers. A more detailed accounting of the data preparation protocol is provided in the SHIFTX2 paper (Han et al. 2011). The above selection and filtering process reduced the data set to 240 proteins. This data set was then divided into a training set and an independent test set. The training dataset consisted of 181 proteins (25,205 residues) whereas the test dataset contained 59 entries (8,078 residues). Among the training proteins, 146 proteins belonged to the α + β folding class, 15 proteins to the all-α, 18 proteins to the all-β and two proteins to the all-coil folding class. For the test proteins, 52 proteins had an α + β architecture, three were all-α and four were all-β. Note that there were no disordered proteins in the test set. The free parameters for the secondary structure assignment model were optimized on the training data set while the test set was used to perform an independent validation of the program's performance.

DSSP (Kabsch and Sander 1983), STRIDE (Frishman and Argos 1995) and VADAR (Willard et al. 2003) served as the three programs used to assign reference secondary structures ("α-helix", "β-strand", "coil") in both the training and test set proteins. These methods assign secondary structures based on the coordinates of the 3D structures as well as inferred H-bonds and torsion angles derived from those coordinates. The normal eight-state DSSP assignments were transformed into a three-state (helix, sheet, coil) assignment using the EVA convention (Eyrich et al. 2001). The same procedure was applied to the STRIDE output. No such transformation was required for the VADAR output. According to DSSP, there were a total of 2,335 β-strand residues (29 %), 2,186 residues in α-helices (27 %) and 3,557 coil assignments (44 %) in the test set. STRIDE determined 2,499 residues as β-strands (31 %), 2,677 as α-helices (33 %) and 2,902 residues as coil structures (36 %) in the test set. Finally, VADAR found 2,489 β-strands (31 %), 2,720 α-helices (34 %), and 2,869 coil structures (35 %). According to DSSP, the training set had a total of 6,837 β-strand residues (28 %), 7,588 residues in α-helices (29 %) and 10,780 coil assignments (43 %). STRIDE identified 7,368 residues in β-strands (29 %), 8,857 residues in α-helices (35 %), and 8,980 coil residues (36 %). VADAR identified 7,196 β-strand residues (28 %), 8,910 α-helical residues (36 %), and 9,099 coil residues (36 %).

Missing chemical shifts handling and neighbor residue correction

The completeness of a given protein's chemical shift assignments plays a crucial role in determining the performance of any chemical shift-based secondary structure assignment method (Shen et al. 2009a, b). The current model is no exception. We assessed the performance of our CSI 2.0 program using both complete and incomplete shift assignments. Incomplete shift assignments were found to negatively affect the accuracy of the secondary structure assignments by up to 3 %.

As mentioned in the previous section, because a small, but significant number (<10 %) of chemical shift assignments were missing in some entries in our protein data set, we needed to take appropriate measures to handle the assignment gaps. This was done by searching through a sequence-chemical shift triplet database to fill in any missing assignments in a manner similar to that described by Shen et al. (2009a, b). More specifically, each entry in our database was converted to an amino acid triplet and each had six backbone ($^{13}C_{\alpha}$, $^{13}C_{\beta}$, ^{13}C, ^{15}N, ^{1}HN, $^{1}H_{\alpha}$) experimental chemical shifts associated with it (except for Gly and Pro). To fill in the missing data, the query sequence triplet was compared with each triplet entry in the database and scored in terms of sequence and chemical shift similarity. The ten best scoring triplets were selected and the average of the ten central residue shifts was used as a proxy for the missing assignment. This process was repeated for all missing assignments (except $^{13}C_{\beta}$ for Glycine, ^{15}N and ^{1}HN for Proline).

Several studies have reported on the significant influence of the nearest neighbor residues on random coil chemical shifts (Wishart and Nip 1998; Wang and Jardetzky 2002b; Wang et al. 2007b). In particular, it has long been noted that the preceding amino acid type significantly affects the ^{15}N and amide proton chemical shift, while the ^{13}C and ^{1}H proton chemical shifts are largely affected by the identity of the following amino acid. Proper accounting for these nearest-neighbor effects is critical to accurately determining protein secondary and tertiary structures from chemical shift data (Wishart 2011). Hence, the random coil chemical shifts for all 20 amino acids were corrected by neighboring residue correction factors provided in Schwarzinger et al. (2001). Finally the secondary chemical shifts for all six-backbone atoms were calculated by subtracting the sequence-corrected random coil shift from the observed shift.

Feature set

In developing any kind of machine-learning algorithm it is necessary to extract a set of input features from the training data that will be used to infer or calculate the desired output (in this case, the secondary structure). Features can either be the raw data (i.e. sequence, chemical shifts, etc.) or derived data (i.e. estimated accessible surface area) that is calculated from the raw data. In developing CSI 2.0 we derived a set of eleven different features from our chemical shift and sequence data. These features included: (1) shift-derived beta strand propensity; (2) shift-derived helix propensity; (3) shift-derived coil propensity; (4) sequence-derived beta strand propensity; (5) sequence-derived helix propensity; (6) sequence-derived coil propensity; (7) Random Coil Index (RCI) (Berjanskii and Wishart 2005); (8) real-valued fractional accessible surface area; (9) two-state relative accessibility classification; (10) multi-sequence alignment-derived residue conservation score and (11) PSIPRED (Jones 1999) predicted secondary structure. Furthermore, for each data point in the protein sequence, a five-residue window was evaluated, with the central residue being the residue of interest. This translates to a total of 55 features for each data point within the five-residue window, as each residue had 11 features. Note that all of the input features were derived from only the sequence and the backbone chemical shifts.

Secondary chemical shift-based probability of three-state secondary structure

The shift-based secondary structure probability of a residue is derived from the secondary chemical shift value of its constituent atoms. The secondary chemical shift (Δδ) is defined as the difference between the absolute chemical shift (δ_{abs}) and the corresponding (neighbor-adjusted) random coil (δ_{rc}) shift (Wishart 2011).

$$\Delta\delta = \delta_{abs} - \delta_{rc}$$

The probability of a residue being in one of the three secondary structure classes "α-helix", "β-strand" or "coil", is derived from its six backbone atom secondary chemical shifts, as described in (Wang and Jardetzky 2002a). For each backbone atom, a Gaussian probability distribution is assumed, where the two parameters for the distribution (μ and σ) correspond to the average (μ) chemical shift value (for each of the three different secondary structure states) and the standard deviation (σ) of the chemical shift distribution respectively. These statistical parameters were derived from the "RefDB" database (Zhang et al. 2003). Therefore, given ($\Delta\delta_{\mathrm{n}}$) {n = $\Delta\delta_{\mathrm{CA}}$, $\Delta\delta_{\mathrm{CB}}$, $\Delta\delta_{\mathrm{C}}$, $\Delta\delta_{\mathrm{HA}}$, $\Delta\delta_{\mathrm{HN}}$, $\Delta\delta_{\mathrm{N}}$}, the six experimental backbone secondary chemical shifts for a given residue *i*, the joint probability of being in one of three secondary structure states can be calculated from the Gaussian distributions of the six backbone atom types of non-Gly/Pro residues (five in case of Gly and four in case of Pro). The joint probability equation is formulated as:

$$P_i^s(\Delta\delta\mathrm{n}) = \Pi \prod_n G_i^s(\Delta\delta\mathrm{n})$$

where π represents the probability or likelihood for an amino acid of type *i* being in the secondary structure type *s* (s = ("α-helix", "β-strand", "coil")). Note that this probability or likelihood π can also be described by amino acid conformational preference and is calculated using the same method described in the next paragraph. G_i^s represents the Gaussian distribution of a particular atom for amino acid type *i* and secondary structure type *s*.

$$G_i^s = \frac{1}{\sqrt{2\pi\sigma_{i,n}^s}} \exp\left(-\frac{(\Delta\delta\mathrm{n} - \overline{\Delta\delta\mathrm{n}}_{i,n}^s)^2}{2\sigma_{n,s,i}^2}\right)$$

The joint probability P^s for each residue is normalized so that its sum of three secondary structure types is equal to 1.0.

Sequence based probability of three-state secondary structure

The conformational preference for an amino acid is taken into account using this feature. Each amino acid has a predisposition to assume a specific secondary structure type, which is referred to as its conformational preference. We derived the secondary structure conformational preferences for all 20 amino acids using an in-house high-resolution sequence-structure database (the sequences and secondary structures in FASTA format are available on the CSI 2.0 website). This database contains 2100 X-ray structures that share no more than 33 % sequence identity with each other, have an R-value ≤0.2 and a resolution ≤1.5 Å. These proteins were extracted using the PISCES server (Wang and Dunbrack 2003) via the PDB (Berman et al. 2000) and the secondary structures were assigned via DSSP (Kabsch and Sander 1983). The conformational preference statistic was calculated as follows: given a residue *i*, and the available secondary structure conformation *s* (*s* = ("α-helix", "β-strand", "coil")) that it can adopt, then the equation to calculate the conformational preference is given as (Levitt 1978):

$$C_i^s = \frac{T_i^s / T^s}{T_i / T}$$

where T_i^s denotes the total number of residues i adopting conformation S, while T^s is the total number conformation S observed in the database, T_i is total number of residues of type i. T represents the total number of different residue types in the database. The conformational preference of each residue for three secondary structure types is then normalized so that its sum is equal to 1.0.

Random Coil Index (RCI) for backbone atoms

The RCI for protein backbone atoms is an easily calculated measure that corresponds to the flexibility of an amino acid on a residue-level as derived from backbone chemical shifts (Berjanskii and Wishart 2005). The backbone RCI quantitatively traces the relative amount to which a protein backbone's chemical shifts match with the random coil values. Those that are closer to random coil values are the most flexible, while those that are most different from random coil values are least flexible. This feature was calculated using the RCI equation provided in the original RCI paper.

Relative accessible surface area

The solvent accessibility of a residue is a measure of an amino acid's (especially its side chain) solvent exposure. Generally unstructured coils or other highly hydrophilic regions are more accessible to water than hydrophobic helices or beta-strands. This trend can be exploited to obtain useful information for identifying protein secondary structures. Recent publications suggest that including solvent accessibility along with sequence information can improve secondary structure prediction accuracy (Adamczak et al. 2005; Momen-Roknabadi et al. 2008). In an effort to include solvent accessibility in CSI 2.0 we developed a machine learning regression model that estimates real numerical value of each residue's fractional accessible surface area (fASA). The fASA is equal to the accessible surface area measured for a given residue (X) in a protein divided by the ASA for that residue in a *G-X-G* tripeptide. The fASA varies between 0.0 (fully buried) to 1.0 (fully exposed). The regression model we developed uses two sequence derived features (hydrophobicity and sequence conservation score) and two chemical shift-derived features (3-state structural probability using six backbone chemical shifts and the RCI) to calculate the fASA. The model was trained on a dataset of 28 proteins with known 3D coordinates and near-complete ^{1}H, ^{13}C and ^{15}N chemical shift assignments and validated on a test set of 66 proteins (with known 3D coordinates and near-complete chemical shift assignments). The fASA for all training and test proteins was calculated using VADAR (Willard et al. 2003). The correlation between the observed fASA and the predicted fASA was 0.76. This fASA value was then incorporated into the CSI 2.0 feature set in the same manner as all other features. Additional details regarding this shift-based fASA prediction method, its performance and its potential applications will be described in a forthcoming manuscript.

Two-state buried-exposed class

The two-state buried-exposed classification assignment is simply a transformation of the fractional ASA (fASA) into two discrete classes obtained by applying a 25 % fASA cutoff. In other words, if the fASA is greater than 0.25, the residue is assigned to an "exposed" state, otherwise the residue is said to be "buried". This information was derived from the chemical shift-based fASA calculation described above.

Residue conservation score

Sequence conservation is a measure of how frequently a given residue is seen at an equivalent position, in an equivalent protein, across different species. Generally highly conserved residues are buried within the protein's core, and less conserved residues are more exposed (albeit with some exceptions). The conservation score for each residue position can be calculated as described by Valder (2002). First, a three-iteration PSI-BLAST (Altschul et al. 1997) search is performed on the UniRef90 clustered database (UniProt Consortium 2010). From the identified hits a multiple sequence alignment is then performed using ClustalOmega (Sievers et al. 2011). The conservation score for each column in the alignment (i.e. each residue in the target sequence) is then calculated using Shannon's entropy formula as described below,

$$s(x) = \lambda \sum_a^K p_a \log p_a$$

where p_a is the probability of observing the ath amino acid and λ is the scaling factor and defined as,

$$\lambda = [\log(\min(N, K))]^{-1}$$

where N = number of sequences in the alignment, K = length of the amino acid alphabet. The probability of observing the ath amino acid is the summed weight of sequences having the symbol a in the position x in the sequence which is defined as,

$$p_a = \sum w_i$$

where w_i is the weight of the ith sequence with w_i being defined as,

$$w_i = \frac{1}{L}\sum_x^L \frac{1}{k_x n_x}$$

where L = length of the alignment, k_x = the number of amino-acid types present at the xth position, n_x = the number of times the ath amino acid occurring in the ith sequence at the xth position.

PSIPRED predicted secondary structure

In an effort to boost the performance of CSI 2.0 we supplemented our method with another powerful secondary structure identification tool called PSIPRED (Jones 1999). PSIPRED is a pure sequence-based secondary structure prediction method developed in the 1990s. It has been refined and improved upon over the last decade and is generally considered one of the most accurate sequence-based prediction methods available, with a typical performance of >80 % (Hung and Samudrala 2003). Previous authors have observed a slight boost to the performance of their shift-based secondary structure assignment routines by including this information in their algorithm (Hung and Samudrala 2003). As a result we also added a PSIPRED (sequence-based) prediction as one of the features to CSI 2.0. Therefore, PSIPRED (version 3.3) predicted secondary structure state for each residue is included in the CSI 2.0 feature vector for the training data points.

Feature normalization

A z-score normalization step was done to normalize the features in the training and the test data set. Assuming there are $N_1,\ldots, N_{i\ldots}, N_t$ rows in the training set, with each row containing $M_1\ldots, M_{j\ldots}, M_n$ different features (columns), then the normalized value of element e_i^j at the ith row and jth column is calculated as:

$$Normalized\left(e_i^j\right) = \frac{e_i^j - \bar{N}_i}{std(N_i)}$$

where

$$\bar{N}_i = \frac{1}{t}\sum_{i=1}^t e_i \quad \text{and}$$

$$std(N_i) = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^t (e_i - \bar{N}_i)^2}$$

All test features were normalized using the mean and standard deviation derived from the training feature distribution.

Multi-class SVM training

With a five-residue window, there were total 25,205 data points in our training set. All data points were normalized prior to the training. Two different normalization methods, a "Statistical Z-score" and a "Max/Min" score were assessed, with the "Statistical Z-score" ultimately being selected. In our multi-class SVM model, a Radial Basis Function (RBF) kernel was used to map the features from a higher dimensional space to a lower dimensional space by computing dot product between the features. Using this type of "dimensional reduction", the performance of SVM classification depends on the following two parameters: (1) the regularization parameter "C" (also known as the "cost" factor) and (2) the Gaussian kernel width "σ". The "C" parameter allows one to adjust the trade-off between maximizing the decision-boundary width and minimizing the number of misclassified samples in the training set. The "σ" parameter controls the width of Gaussian kernel and can be adjusted to help minimize the number of misclassified test examples. These two parameters were optimized using a repeated tenfold cross validation (CV). The goal of the parameter optimization was to find the optimal values that maximizes the accuracy or Q3 score of the three-class secondary structure classification. The Multi-class SVM implementation in the R package "*kernlab*" was used to train the classifier (RDC 2009; Karatzoglou et al. 2004). The optimization of "C" and "σ" through the "*repeated-cv*" method was performed using the *train()* function in the "caret" package in R (Kuhn 2008).

A multi-residue Markov model for post-assignment filtering

While the SVM classifier (described above) generally performs very well it is still prone to making confusing, meaningless or "scrambled" secondary structure assignments such as: *CCBHH* or *BBHCC or HCHCH*. This is also a common problem for many other secondary structure prediction/assignment methods such as PSIPRED, TALOS-N or DANGLE. Most programs use heuristic "character smoothing" that employ "if–then-else" ladders or character averaging to correct or eliminate these problem assignments. However, these heuristic methods are not very robust nor are they very accurate. A more robust method to perform character smoothing or character correction is to use a Markov model (Durrett 2010). Markov or hidden Markov models are widely used methods for text filtering, pattern extraction and natural language processing. This also makes them ideally suited to treating the "scrambled" text problem. After assessing character window widths of three, five, seven and nine residues, a seven-residue Markov model was found to be optimal to handle scrambled or discontinuous segments of secondary structure. This Markov filtering involved sliding a trained, seven-residue Markov filter along the protein chain that identified scrambled secondary structure assignments and then corrected them as necessary. According to this multi-residue Markov model, if there are $n(= t_1, t_2, \ldots, t_n)$ residues in a single pattern along the protein chain, then the probability of observing *i*th residue in that pattern depends on the observed probabilities of the preceding $(i - 1)$ residues. This can be expressed by the following equation,

$$P(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} P(t_i | t_1, t_2, \ldots, t_{i-1})$$

The conditional probability of observing a residue in *i*th location given the history of the preceding $(i - 1)$ residues is calculated from $[t_1, t_2, \ldots, t_{i-1})$ and $[t_1, t_2, \ldots, t_i)$ pattern frequency counts.

$$P(t_i | t_1, t_2, \ldots, t_{i-1}) = \frac{count(t_1, \ldots, t_{i-1}, t_i)}{count(t_1, \ldots, t_{i-1})}$$

To calculate the probability of a seven-residue pattern, the frequencies of smaller patterns consisting of one, two, three, etc. up to six residues are extracted from the training database of reference structures. An example formula to calculate the probability of a five-residue pattern *HHBCC* is as follows:

$$P(HHBCC) = P(C|HHBC) * P(C|HHB) * P(B|HH) \\ * P(H|H) * P(H)$$

and a probability value like $P(C|HHBC)$ can be calculated by following equation,

$$P(C|HHBC) = \frac{count(HHBCC)}{count(HHBC)}$$

The probability cutoff to validate a pattern is chosen as 0.00 (i.e. if the probability of a multi-residue pattern, along with its two preceding and following patterns is found to be equal to the cut-off value, then the central pattern is considered to be “scrambled”). For a scrambled secondary structure pattern to be identified, the outlier must be either in the middle, or any of the two adjacent positions. The outlier is then corrected by looking at the secondary structure assignments of the four surrounding residues.

Evaluation metrics

Q3-accuracy

Q3-accuracy is the most widely used metric to evaluate three-state secondary structure predictions or assignments. It is the ratio of correctly predicted or identified states divided by the total number of amino acids or residues in the dataset. Q3-accuracy is simply defined as:

$$Q3 = \frac{N_P}{N}$$

where N_p is the total number of residues for which secondary structure state is predicted correctly by the model and N is the total number of residues in the example set.

Segment-Overlap (SOV) score

The Segment-OVerlap score (SOV) is based on the average overlap between the observed and predicted segments. It is designed to evaluate the correctness of segment prediction with respect to a reference assignment (Rost et al. 1994; Zemla et al. 1999). The SOV score measures how much the predicted segments deviate from experimental segment length distributions. The definition of the SOV score for a secondary structure i, where $i \in (H, B, C)$,

$$SOV_i = \frac{1}{N_i} \sum_{S_i} \frac{minOV(s_1, s_2) + \delta(s_1, s_2)}{maxOV(s_1, s_2)} \times len(s_1)$$

here, s_1 and s_2 are the observed and predicted secondary structure segments in one of the three states; S_i is the number of all segment pairs (s_1, s_2), where $s_1$and s_2 contains at least one residue in *i*th state in common, $minOV(s_1, s_2)$ is the length of actual overlap of $s_1$and s_2 and $maxOV(s_1, s_2)$ is the length of the total extent for which either of the segments $s_1$or s_2 has a residue in *i*th state. N_i is the total number of residues observed in the *i*th conformation. $\delta(s_1, s_2)$ is defined as,

$$\delta(s_1, s_2) = \min \left\{ \begin{array}{c} maxOV(s_1, s_2) - minOV(s_1, s_2) \\ minOV(s_1, s_2) \\ int(0.5 \times len(s_1)) \\ int(0.5 \times len(s_2)) \end{array} \right\}$$

where $len(s_1)$ is the number of residues in the segment s_1. The SOV measure for all three states, SOV_{all} is defined as,

$$SOV_{all} = \frac{1}{N} \left(\sum_{i \in H,B,C} \sum_{S_i} \frac{minOV(s_1, s_2) + \delta(s_1, s_2)}{maxOV(s_1, s_2)} \times len(s_1) \right) \\ \times 100$$

where s_1 and s_2 are the observed and predicted secondary structure segments in *i*th state. N is the total length of proteins under consideration.

Results and discussion

As described earlier in the “Methods and materials” section, the “C” parameter (the “cost” value) in the SVM classifier and the kernel parameter, “σ” in the Gaussian RBF kernel were optimized using tenfold Cross Validation (CV). After achieving an optimal value of 0.0157, “σ” was

Table 1 Weighting coefficients (|**w**|) of chemical-shift and sequence-derived features for CSI 2.0's SVM model

Feature	Weight Coeff.	Feature	Weight Coeff.
ProbBCS(i − 2)	19.88598	ProbCAA(i + 1)	14.47068
ProbBCS(i − 1)	47.83482	ProbCAA(i + 2)	12.65604
ProbBCS(i)	53.96416	RCI(i − 2)	11.71170
ProbBCS(i + 1)	21.52516	RCI(i − 1)	29.60009
ProbBCS(i + 2)	9.230448	RCI(i)	19.74642
ProbHCS(i − 2)	8.925556	RCI(i + 1)	14.32193
ProbHCS(i − 1)	29.11158	RCI(i + 2)	7.806794
ProbHCS(i)	25.97708	RSA(i − 2)	4.939002
ProbHCS(i + 1)	5.456130	RSA(i − 1)	30.64792
ProbHCS(i + 2)	14.17828	RSA(i)	22.54183
ProbCCS(i − 2)	12.56523	RSA(i + 1)	14.22241
ProbCCS(i − 1)	19.28457	RSA(i + 2)	10.04285
ProbCCS(i)	31.72981	BuriedExposed(i − 2)	0.931595
ProbCCS(i + 1)	19.65260	BuriedExposed(i − 1)	14.87281
ProbCCS(i + 2)	8.465084	BuriedExposed(i)	1.417325
ProbBAA(i − 2)	2.680083	BuriedExposed(i + 1)	15.17127
ProbBAA(i − 1)	21.07358	BuriedExposed(i + 2)	1.813477
ProbBAA(i)	9.595489	Scon(i − 2)	10.84571
ProbBAA(i + 1)	23.39969	Scon(i − 1)	10.29277
ProbBAA(i + 2)	7.901136	Scon(i)	18.86649
ProbHAA(i − 2)	3.968213	Scon(i + 1)	8.360907
ProbHAA(i − 1)	6.390549	Scon(i + 2)	22.17609
ProbHAA(i)	3.372710	PSIPRED(i − 2)	17.87049
ProbHAA(i + 1)	8.024158	PSIPRED(i − 1)	28.21943
ProbHAA(i + 2)	8.566371	PSIPRED(i)	57.99818
ProbCAA(i − 2)	5.022616	PSIPRED(i + 1)	20.72087
ProbCAA(i − 1)	22.44050	PSIPRED(i + 2)	6.860726
ProbCAA(i)	5.881907		

The position of the feature over a five-residue window is given using standard indices in parentheses. The feature name abbreviations are as follows: ProbBCS = β-strand probability using chemical shift, ProbHCS = α-helix probability using chemical shift, ProbCCS = coil probability using chemical shift, ProbBAA = β-strand probability using amino acids, ProbHBAA = α-helix probability using amino acids, ProbCAA = coil probability using amino acids, RCI = Random Coil Index (protein flexibility), RSA = fractional or real-valued solvent accessibility, BuriedExposed = 2-state (Buried/Exposed) solvent accessibility, Scon = residue conservation score, and PSIPRED = PSIPRED predicted secondary structure

held constant while "C" was iteratively changed to optimize its value. To achieve an unbiased training result, the whole process was repeated five times. For each repetition, the accuracy of the three-state assignment of the training classes was measured. The optimal "cost" and "σ" values that were found to maximize the Q3 accuracy using this repeated training were 2.0 and 0.0157 respectively. The training accuracy was averaged over five repetitions of the tenfold CV process. A training accuracy of Q3 = 90.56 % on 181 training proteins was observed with the aforementioned optimized parameter values. A test accuracy of Q3 = 89.35 % was achieved on an independent test set of 59 proteins.

The final set of weighting coefficients for the sequence and chemical shift-based features in our multi-class SVM model are listed in Table 1. The sum of all the weights (over the five residue positions) for chemical shift-derived features was 683 while the sum of all the weights for the sequence-derived features was 202 (a difference of 3.4X). Among individual features, the PSIPRED predicted secondary structure for the central residue (residue *i*) was found to have the largest single weight in the SVM formulation (|w| = 58.0). The second largest weighted feature (|w| = 54.0) was the β-strand propensity calculated from backbone chemical shifts at the central residue location. Chemical shift derived α-helix, β-strand and coil probability scores in the central residue or immediate neighbor locations were found to be moderately relevant in terms of their weighting. Both protein flexibility (RCI) and solvent accessibility (fASA) at the (*i* − 1) location had larger weights than the same feature values at other residue positions. Interestingly the RCI and fASA weightings also proved to be more important than the sequence conservation scores. Given the 3.4X greater weight attached to shift-derived features in CSI 2.0's final SVM model, we believe it is fair to claim that CSI 2.0 is essentially a chemical shift based method that incorporates a small amount of sequence information. This assertion is also borne out by the fact that the performance of CSI 2.0 (without the sequence-based prediction) was only 2 % worse than the version with sequence-based prediction (see below).

CSI 2.0 comparative performance

In Table 2, we compare the performance of our CSI 2.0 method with seven hybrid (chemical shift and sequence-based) and one pure sequence-based secondary structure identification/prediction programs. The eight programs are: TALOS+ (Shen et al. 2009a), TALOS-N (Shen and Bax 2013), DANGLE (Cheung et al. 2010), CSI (Wishart et al. 1992), PSSI (Wang and Jardetzky 2002a), Delta2D (Camilloni et al. 2012), Psi-CSI (Hung and Samudrala 2003) and PSIPRED (Jones 1999). The performance of all eight programs was evaluated on the basis of: (1) Q3-accuracy of predicting three different structure states; (2) individual structural state ("α-helix", "β-strand", "coil") prediction accuracy; (3) Segment-Overlap or SOV score; and (4) coverage (proportion of residues in the test set that were predicted). For Table 2, the first column indicates the name of the prediction model, while the second, third, and fourth columns indicate the accuracy for each category of secondary structure. The fifth column presents the overall Q3-

Table 2 Performance of CSI 2.0 and eight other chemical shift and sequence-based methods on an independent test set of 59 proteins (total 8,078 residues) when using "DSSP" (Kabsch and Sander 1983) secondary structure assignments as the reference structure. Columns 2–5 correspond to Q3 scores while columns 6–9 correspond to SOV scores

Methods	Helix	Beta	Coil	Q3-score	Helix	Beta	Coil	SOV-score	% coverage
TALOS+	93.39	77.93	80.34	83.89	80.09	80.73	83.58	84.83	97.80
TALOS-N	95.54	82.65	79.08	86.39	88.78	85.71	83.08	87.85	97.70
DANGLE	95.88	80.0	76.44	83.0	80.61	80.58	81.46	83.66	98.60
CSI	84.17	67.40	84.23	80.33	76.01	69.34	71.53	75.18	100
PSSI	62.85	70.77	62.58	67.33	59.49	73.39	72.62	71.37	96.80
δ2D	43.29	33.17	36.73	42.24	42.58	38.20	42.28	42.82	48.24
Psi-CSI	92.88	80.0	85.53	86.20	89.02	81.43	83.06	86.94	100
PSIPRED	85.95	79.17	88.11	85.36	72.08	63.74	63.00	79.00	100
CSI 2.0	93.41	86.50	87.80	89.35	90.76	85.34	82.75	88.45	100

accuracy, while the last four columns indicate the individual and overall SOV-scores. The last column shows the percent coverage (proportion of residues of test data that were identified or predicted) by each method. As seen in this table, CSI 2.0 achieves the best overall Q3 and SOV scores while Psi-CSI and TALOS-N are essentially tied for second in their overall performance. With regard to the performance for individual secondary structure state (helix, sheet, coil) identification, CSI 2.0 also shows superior accuracy for all three-structure states. In particular, for DSSP-referenced structures, CSI 2.0's performance was an average of 10.87 % better in case of β-sheet identification, and 8.59 % better for coil identification, than the eight other chemical shift and sequence-based methods (see Table 2). For helix identification, the CSI 2.0 shows a comparable performance with respect to other methods. In terms of the SOV measure, the same trend is observed. Although the Q3 accuracy of CSI 2.0's residue-specific helix assignments was not much better than existing programs, its higher average SOV-score indicates a better agreement for helical segments. The same is true for the overall SOV-score for all three-secondary structure types. In terms of SOV-scores, the next best performance was seen for the most recent program, TALOS-N (Shen and Bax 2013). CSI 2.0's assignments, unlike most of other programs, covers the full fraction (≈100 %) of the test data points.

Statistical significance of CSI 2.0's improvement

As indicated in Table 2 and Fig. 1, the best-performing methods all achieve Q3 accuracies above 80 % and the difference between CSI 2.0 and the other top performing programs is only 3–4 %. One may ask is this performance improvement statistically significant? To address this question we performed a Student's *t* test to assess the *p* value between CSI 2.0 and TALOS+, TALOS-N, DANGLE and Psi-CSI. The results are shown in Table 3. These data confirm that the performance improvement seen in CSI 2.0 is, indeed, highly significant with most *p* values being ≪0.001.

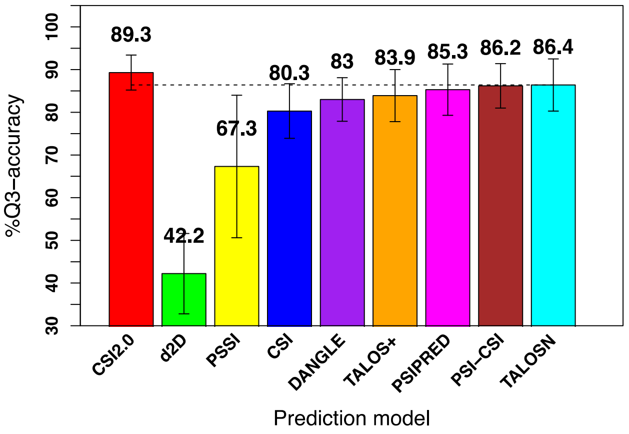


Fig. 1 A *bar* graph comparing CSI 2.0's Q3 accuracies with eight other chemical shift and sequence-based protocols over an independent test set of 59 proteins. The *error bar* (i.e. standard deviation in Q3-accuracy of each method) appears on *top* of *each bar* plot

Table 3 The *p* values or probabilities of Student's two sample *t* tests between CSI 2.0 and four other best performing methods are shown

Method1 versus Method2	*p* value
CSI 2.0 versus TALOS+	6.575e−08
CSI 2.0 versus TALOS-N	0.0016
CSI 2.0 versus DANGLE	6.347e−08
CSI 2.0 versus Psi-CSI	0.00087

Here the null hypothesis is that the difference between sample1 mean (mean accuracy of method1) and sample2 mean (mean accuracy of method2) is equal to zero. Alternative hypothesis indicates that the sample1 mean is greater than the sample2 mean

CSI 2.0 performance using selected and partial shift assignments

It is not particularly common for a protein to have all ^{1}H, ^{13}C and ^{15}N backbone shifts fully assigned. Indeed, many shorter peptides and proteins will only have their ^{1}H assignments completed, while larger proteins may only have their ^{1}H and ^{15}N shifts, ^{15}N and ^{13}C shifts or ^{1}H and ^{13}C shifts assigned. Given that only certain nuclei may be

Table 4 CSI 2.0 performance with selected chemical shift assignments and combinations of shift assignments

Shift assignment	Helix	Beta	Coil	All
$^{13}C_{\alpha}$	84.38	93.36	87.90	88.72
^{13}C	83.41	90.22	87.95	87.37
$^{13}C_{\beta}$	84.37	87.03	86.59	86.50
$^{1}H_{\alpha}$	83.59	87.22	87.21	86.83
^{15}N	82.63	85.74	86.68	85.68
$^{13}C_{\alpha}$, ^{13}C, $^{13}C_{\beta}$, $^{1}H_{\alpha}$	81.46	92.89	98.98	90.92
$^{1}H_{\alpha}$, ^{15}N	81.47	92.45	98.99	90.77

measured we decided to evaluate CSI 2.0's performance using only selected sets of chemical shifts or selected nuclei. The results are listed in Table 4 for five individual backbone nuclei ($^{13}C_{\alpha}$, ^{13}C, $^{13}C_{\beta}$, $^{1}H_{\alpha}$, ^{15}N) along with other common assignment combinations ($^{13}C_{\alpha}$, ^{13}C, $^{13}C_{\beta}$, $^{1}H_{\alpha}$ and $^{1}H_{\alpha}$, ^{15}N). As can be seen from this table, combinations of multiple nuclei give the best performance, but the performance for any single nucleus is surprisingly good (Q3 >85 %). This is because CSI 2.0 also uses sequence information (i.e. PSIPRED predicted secondary structures) to supplement its chemical shift-derived estimates. As was noted in the original CSI papers (Wishart et al. 1992; Wishart and Sykes 1994b) certain nuclei carry more information about secondary structures than others. In particular, the ranking of nuclei for secondary structure information content, from most informative to least informative, is: $^{13}C_{\alpha} > {}^{13}C > {}^{13}C_{\beta} \approx {}^{1}H_{\alpha} > {}^{15}N$.

Because it is often difficult to obtain complete chemical shift assignments for a protein (due to signal broadening from intermediate exchange events, signal overlap, solvent suppression, etc.) we were also interested to see how well CSI 2.0 performed with partial or incomplete chemical shift assignments. To do so we evaluated the performance of CSI 2.0 relative to the percentage of missing chemical shift assignments and compared its results to several other software packages. We analyzed a subset of 21 proteins (from our test set of 59 proteins) with a fraction of missing assignments >15 %. In particular, for this set the percentage of incomplete or missing backbone ^{1}H, ^{13}C or ^{15}N assignments ranged from 16.7 to 37.0 % (based on the total number of expected NMR signals from the protein's amino acid sequence). The secondary structures of these proteins were then determined using five different methods (including CSI 2.0) and evaluated against the observed secondary structures as determined by DSSP. The results are shown in Table 5. As can be seen from this table, CSI 2.0 does significantly better (~7–10 %) in terms of Q3 accuracy than any of the other methods in terms of handling missing shift data. Furthermore, for all of the methods (except CSI 2.0) there is a general trend (r <0.5) showing a degradation in their performance with an increasing fraction of missing chemical shifts. Interestingly, CSI 2.0 seems to be largely immune to any detectable performance degradation with respect to missing chemical shifts (at least up to a level of ~35 % missing shifts). This appears to be due to its robust handling of missing shift data (described earlier) as well as its use of sequence-based secondary structure prediction from PSIPRED.

Different definitions of secondary structure

Secondary structure is not an absolute quantity nor is it universally defined. In other words, there is no gold standard for secondary structure. Different definitions exist of helices, β-strands, β-turns and coils (Zhang et al. 2008). As a result, no two individuals and no two coordinate-based secondary structure assignment programs will agree on the exact start and end locations of many secondary structure elements (Tyagi et al. 2009; Shen et al. 2009a). Likewise some programs (or some individuals) will invariably classify short helices and short beta-strands as coil structures and vice versa. Given the variation in secondary structure "calling" from well-defined 3D structures and the fact that there are several different secondary structure identification algorithms that are widely used by structural biologists, we decided to investigate the performance of CSI 2.0 and the other eight programs against three of the most commonly used coordinate-based secondary structure assignment algorithms: DSSP (Kabsch and Sander 1983), STRIDE (Frishman and Argos 1995) and VADAR (Willard et al. 2003). Table 6 lists the Q3-accuracies of the eight secondary structure prediction/identification programs when compared against the calls made by locally installed versions of DSSP, STRIDE and VADAR. As can be seen in this table, CSI 2.0 agrees best with the DSSP secondary structure assignments while its performance drops slightly with the STRIDE or VADAR calls. The same trend is seen with the other eight programs as well. This is largely due to the fact that essentially all of these programs were trained using DSSP data, as opposed to STRIDE or VADAR data. It is worth noting that PSIPRED (which is used by both Psi-CSI and CSI 2.0) was also trained exclusively on DSSP data. Attempts to train CSI 2.0 with STRIDE or VADAR secondary structure calls yielded no overall improvement in the performance.

It is also interesting to note that the pairwise agreement between the three different secondary structure assignment methods (DSSP, STRIDE and VADAR) in our independent test dataset ranged from 85 to 90 % with an average pairwise agreement of 87.63 %. Furthermore, the overall agreement between all three methods was only 82 %. This suggests that secondary structure identification is

Table 5 Comparison of the performance of CSI 2.0 versus other four methods (Psi-CSI, TALOSN, TALOS+, DANGLE) relative to the percentage of missing backbone ^{1}H, ^{13}C or ^{15}N chemical shift assignments

PDB	BMRB	Percent missing shifts	CSI 2.0	Psi-CSI	TALOSN	TALOS+	DANGLE
1HQ2	4300	27.60	92.76	83.55	71.71	69.08	77.68
1T8L	5358	21.88	96.36	83.64	83.64	83.64	80.00
1JTG	6357	20.81	85.27	78.29	75.97	77.13	81.01
1UDR	4083	18.23	95.04	86.78	90.08	90.91	81.82
1ODV	6321	18.14	84.00	83.00	87.00	83.00	84.00
1W80	6034	23.24	88.74	76.62	75.76	73.16	74.03
1V9T	4037	16.68	86.50	82.82	81.60	79.75	76.07
2AOJ	5967	34.65	85.26	71.58	85.26	82.11	82.11
1CWC	2208	17.64	91.98	76.54	82.10	77.16	81.48
1YKY	4831	35.14	92.97	82.81	81.25	68.75	80.47
1KDB	6250	21.98	84.78	83.70	67.39	65.22	77.17
2A38	5760	37.01	91.62	85.86	74.35	79.06	79.06
256B	6560	17.68	93.07	92.08	94.06	95.05	93.07
1BT5	6024	19.32	88.03	81.85	84.17	81.08	78.38
1SGZ	6016	23.40	78.43	69.68	57.14	53.94	65.01
1SYD	15232	22.17	91.38	82.76	82.76	81.90	78.45
1JR2	7242	18.30	88.85	83.46	86.15	83.85	83.85
1U7B	15501	20.12	90.73	83.47	84.27	80.24	79.44
2A0 N	15741	20.45	89.60	83.60	88.40	84.00	82.00
2DYI	10139	24.69	86.84	73.03	78.29	72.37	68.42
1B1H	10053	18.02	86.62	81.10	81.95	77.92	79.83
Average		22.72	88.99	81.25	80.63	78.06	79.21

Table 6 Percentage Q3-accuracies of the CSI 2.0 protocol and eight other methods over an independent test set of 59 proteins using three different reference [DSSP (Kabsch and Sander 1983), STRIDE (Frishman and Argos 1995) and VADAR (Willard et al. 2003)] structures

Assignment method	DSSP	STRIDE	VADAR
TALOS+	83.89	82.07	81.47
TALOS-N	86.36	85.62	83.89
DANGLE	83.0	81.40	81.0
CSI	80.33	74.64	76.29
PSSI	67.33	65.15	64.0
δ2D	42.24	40.66	41.16
Psi-CSI	86.20	83.53	82.17
PSIPRED	85.36	79.81	78.0
CSI 2.0	89.35	86.72	86.10

inherently imprecise and that the best possible performance that a secondary structure identifier (or predictor) could attain is probably no better than 90 %. Given that all of the proteins we studied had both X-ray structures and NMR structures, we also investigated the level of agreement between the secondary structure assigned via more conventional NMR approaches (NOEs, J-couplings) or via author-assigned secondary structure assignments with those generated from the coordinate data (determined by DSSP, VADAR or STRIDE). Among the coordinate–based assignment methods, STRIDE showed the highest level of agreement (90.05 %) with the author assignments, while DSSP and VADAR had slightly lower levels of agreement (88.54 and 84.54 % respectively). Again, this level of agreement between secondary structure assignment methods (human vs. computer) suggests that CSI 2.0 is performing near the maximum level of accuracy achievable for secondary structure assignment.

Local interaction effects

Regular secondary structure is formed when the local environment induces nearby residues to interact and adopt a specific pattern such as an "α-helix" or a "β-strand". Hence, local interactions and nearest neighbor data (such as nearby shifts and amino acids) can provide important information about the secondary structure propensity of a certain region. To capture these local interaction effects, we assessed CSI 2.0's performance using several different residue window lengths (three, five and seven residues). Our data indicated that CSI 2.0 achieved its best performance, in terms of Q3-accuracy, when using a five-residue window (data not shown for other windows). No significant improvement was achieved by including more than four neighbors (two preceding and two following). This indicates that the features of immediately nearby residues provide the most useful secondary structure information.

Mis-assigned secondary structures

As accurate as CSI 2.0 appears to be, it still exhibits less-than-ideal performance with regard to distinguishing between β-strands and coil regions. In our test data set, there were a total of 2,335 residues in β-strands, in which CSI 2.0 correctly identified 2,019 of them (see Table 7). However, it also mis-identified 316 residues as "coils", or about 13.5 % of the β-strand population. On the other hand, a somewhat smaller percentage of coil residues (7.9 %) were also incorrectly identified as being in β-strands. The probable reason for this is the high degree of chemical shift and amino acid compositional similarity between these two structure types. Indeed, the chemical shifts in β-strands and coil regions tend to exhibit more similarity to each other than to helices. Furthermore, as we discovered on further inspection, many of the mis-identifications occurred at the borders or edges of β-strands and coil regions. While some ambiguity or mis-identification would be expected between the borders of secondary structure elements or short β-strands and extended coil regions, one would hope that there would be no ambiguity between β-strands and helical regions. Therefore it is worth noting that CSI 2.0 did not confuse any β-strands with α-helices and vice versa. In a few cases (6.5 %), CSI 2.0 failed to recognize α-helical residues and identified them as "coil". Likewise, about 4 % of coil residues were mis-identified as α-helices. Once again, many of the mis-identifications occurred at the borders or edges of α-helices and coil regions. In all likelihood these misidentified helices were somewhat flexible or only partially helical under the solution conditions that were originally used to collect the NMR data. The fact that protein structures do sometimes differ between crystal forms (solved by X-ray methods) and in solution (solved by NMR) has been noted for many years. Indeed, there are many examples showing these discrepancies (Andrec et al. 2007; Ratnaparkhi et al. 1998). It is also important to remember the agreement between the secondary structures determined by conventional NMR methods and those determined using X-ray data typically differ by 5–10 %.

Identification of 3_{10} helices and β-bridges

3_{10}-Helices are short helical structures with an average length of three residues and a distorted hydrogen-bonding network, whereas β-bridges are single-residue β-strands. Only the DSSP program identifies these structures and consolidates them into α-helices and β-strands. On the other hand, STRIDE and VADAR often characterize them simply as "coil". In looking more closely at our results, we found that CSI 2.0, regardless of its training set, would identify isolated 3_{10} helices and β-bridges as simple "coil" structures. This underscores one of the challenges with secondary structure identification, namely the fact that different programs (and different structural biologists) have different opinions or different definitions of what secondary structures are. Interestingly CSI 2.0 still performed best when it was working with DSSP assigned secondary structures (as opposed to VADAR or STRIDE assignments)—even with the presence of these "hard-to-identify" 3_{10} helices and β-bridges.

Table 7 Confusion matrix of secondary structure assignments generated by CSI 2.0 on the independent test set of 59 proteins

Secondary structure	H (pred)	B (pred)	C (pred)	Total
H (obs)	2,043	0	143	2,186
B (obs)	0	2,019	316	2,335
C (obs)	281	153	3,123	3,557

PSIPRED improves performance

CSI 2.0 was originally intended to be a chemical shift-only method. However, the exceptional performance of Psi-CSI (Hung and Samudrala 2003) led us to reconsider the use of sequence information. Indeed, the inclusion of PSIPRED (Jones 1999) into the CSI 2.0 algorithm improved the Q3-accuracy from 87.3 % (chemical shift only) to 89.35 %. This improvement is statistically significant ($p < 0.001$). More specifically, the inclusion of PSIPRED was found to improve the "β-strand" accuracy by 4 % and the "coil" accuracy by 2.3 %. On the other hand, the identification accuracy of α-helices was not improved in any substantial way. Given that chemical shift-based methods tend to confuse some β-strand residues with coil residues (and vice versa), it appears that PSIPRED helps to remove this chemical shift ambiguity.

CSI 2.0 accurately identifies secondary structure with "trick" proteins

Proteins with high sequence identity but very different folds pose special challenges for sequence-based structure prediction methods (Shen et al. 2010). One example of note is the protein G pair known as GA (95) and GB (95) (Alexander et al. 2009). Protein GA (95) is a specially designed, mostly helical protein that shares a high degree of sequence identity (95 %) with the native, β-rich protein G. Here, we investigated how CSI 2.0 performed in distinguishing the local structures of these two proteins when compared to other methods [TALOS-N (Shen and Bax 2013) and PSIPRED (Jones 1999)]. As seen in Fig. 2, and as expected, PSIPRED did quite well with its secondary structure prediction for GB but not so well with GA. On the

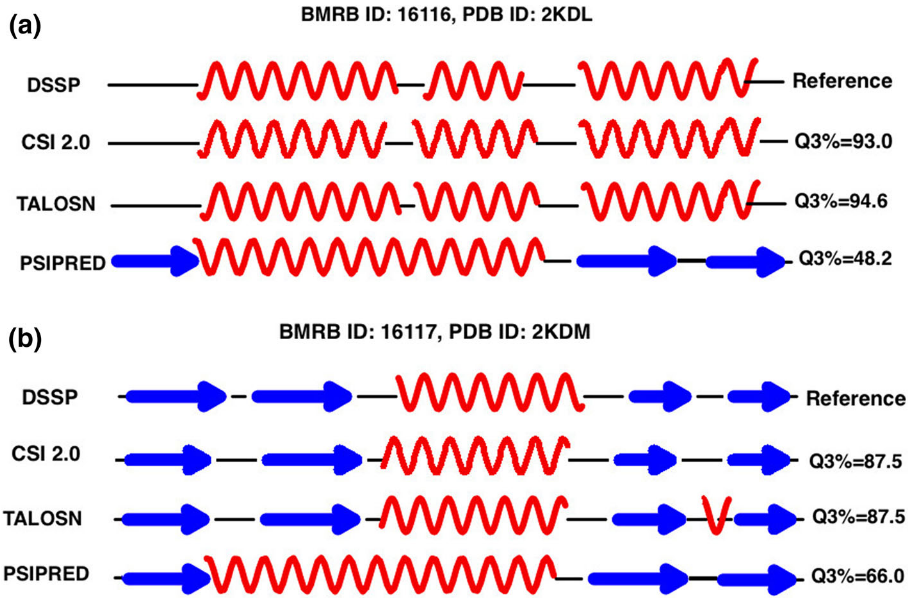


Fig. 2 Secondary structure prediction/assignment for BMRB 16116 (PDB ID: 2KDL) and BMRB 16117 (PDB ID: 2KDM) by CSI 2.0, TALOS-N and PSIPRED

other hand, CSI 2.0 and TALOS-N performed comparably well and were able to correctly identify the secondary structures in both proteins. The fact that CSI 2.0 uses PSIPRED in its determination of secondary structure, but its performance was not compromised in this "GA versus GB test" illustrates how CSI 2.0 is able to appropriately balance experimental chemical shift information with sequence/PSIPRED information.

We also investigated the performance of CSI 2.0 for assigning the secondary structure for a completely unfolded protein (i.e. unfolded ubiquitin in 8 M urea—BMRB 4375). As seen in Fig. 3, CSI 2.0 was able to accurately identify the disordered structure of this protein, whereas PSIPRED and TALOS-N proved to be somewhat less accurate than CSI 2.0. In the case of TALOS-N, nine "coil" regions were incorrectly predicted as "β-strands". In the case of PSIPRED most of the protein was predicted to contain a high proportion of helices and β-strands. Because PSIPRED predicts the secondary structure from sequence, it just reported the folded ubiquitin structure retrieved by a PSI-BLAST search. However, because CSI 2.0 weighs both the chemical shift information with PSIPRED predictions, its performance was not compromised.

Potential improvements

J-coupling constants and NOE data can obviously aid in inferring the existence or delineation of secondary structure. This is why conventional NMR methods have traditionally depended so heavily on these NMR-derived parameters to identify secondary structures. Potentially some improvement in CSI 2.0's performance could be achieved if these parameters were also included in the model, particularly in cases when chemical shift data is missing or ambiguous. However, our focus has primarily on developing a simple approach that requires only sequence data and backbone chemical shift information to accurately identify protein secondary structures. The advantage of using chemical shifts is that these are the first pieces of experimental data that one obtains when studying proteins by NMR. Chemical shifts are also far easier to measure and far more accurately measured than NOEs and J-coupling data.

Instead of adding more experimental data, another approach that could potentially improve the performance of CSI 2.0 is to include sequence homology information from previously solved protein structures. With more than 100,000 protein structures in the PDB, this represents a significant and largely untapped information resource on secondary structure. The use of sequence homology from solved structures has been shown to substantially improve the performance of sequence-only secondary structure prediction methods (Montgomerie et al. 2006; Cole et al. 2008). However, it is not clear whether the same level of improvement could be achieved when working with data that already has some experimental information concerning the secondary structure (i.e. chemical shifts).

The CSI 2.0 web server

A web server (http://csi.wishartlab.com) has been developed that accepts a BMRB (NMR-Star 2.1 or NMR-Star 3.1) or SHIFTY-formatted chemical shift file and generates secondary structure assignments along with a colorful CSI bar graph plot with secondary structure icons marked above the bar graph. The server supports a number of user-

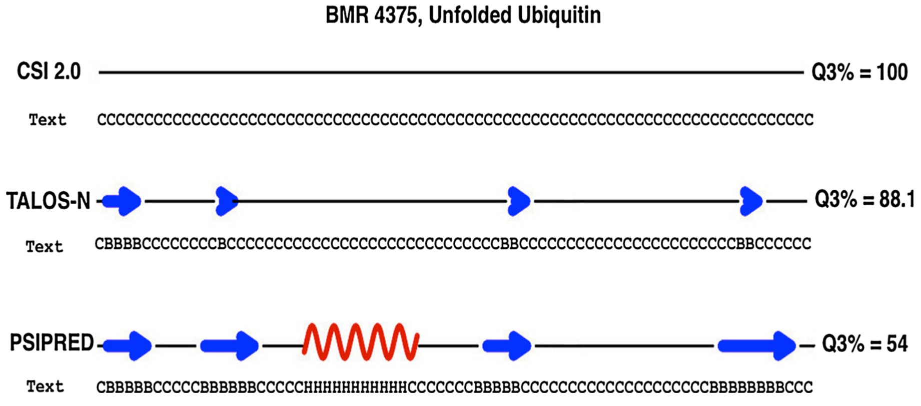
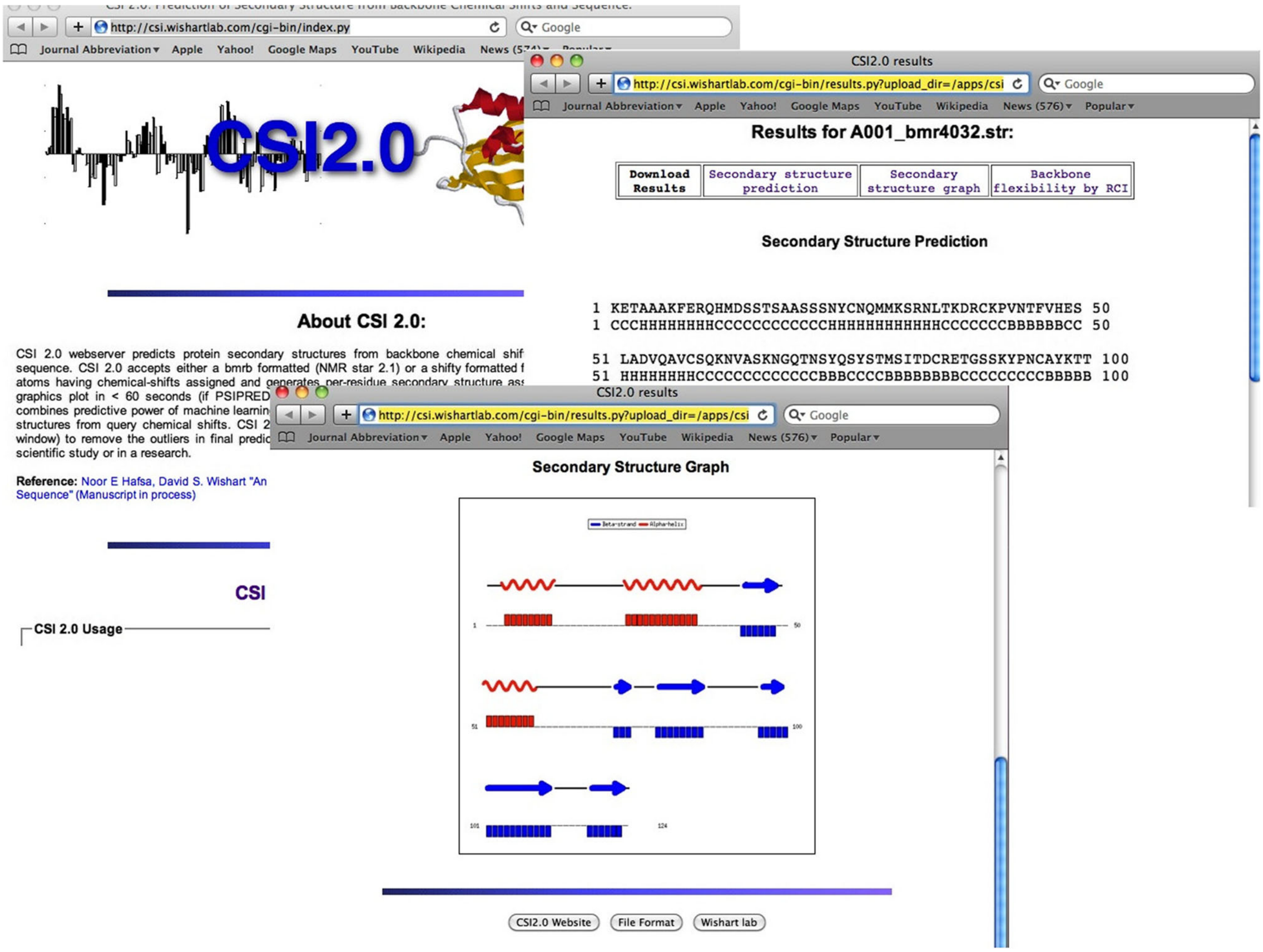

BMR 4375, Unfolded Ubiquitin
CSI 2.0
Q3% = 100
Text
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
TALOS-N
Q3% = 88.1
Text
CBBBBCCCCCCCCBCCCCCCCCCCCCCCCCCCCCCCCCCCBBCCCCCCCCCCCCCCCCCCCCCCCCBBCCCCCC
PSIPRED
Q3% = 54
Text
CBBBBBCCCCCBBBBBBCCCCHHHHHHHHHHHCCCCCCCBBBBBCCCCCCCCCCCCCCCCCCCCBBBBBBBBCCC

Fig. 3 Secondary structure assignment/prediction for BMRB 4375 (unfolded ubiquitin) as determined by the CSI 2.0, TALOS-N and PSIPRED programs


CSI 2.0: Prediction of Secondary Structure from Backbone Chemical Shifts and Sequence.

http://csi.wishartlab.com/cgi-bin/index.py

About CSI 2.0:

CSI 2.0 webserver predicts protein secondary structures from backbone chemical shif
sequence. CSI 2.0 accepts either a bmrb formatted (NMR star 2.1) or a shifty formatted f
atoms having chemical-shifts assigned and generates per-residue secondary structure ass
graphics plot in < 60 seconds (if PSIPRED
combines predictive power of machine learnin
structures from query chemical shifts. CSI 2
window) to remove the outliers in final predic
scientific study or in a research.

Reference: Noor E Hafsa, David S. Wishart "An
Sequence" (Manuscript in process)

CSI

CSI 2.0 Usage

CSI2.0 results

http://csi.wishartlab.com/cgi-bin/results.py?upload_dir=/apps/csi

Results for A001_bmr4032.str:

Secondary Structure Prediction

```
1 KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHES 50
1 CCCHHHHHHHHCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCCBBBBBBCC 50

51 LADVQAVCSQKNVASKNGQTNSYQSYSTMSITDCRETGSSKYPNCAYKTT 100
51 HHHHHHHCCCCCCCCCCCCBBBCCCCBBBBBBBBCCCCCCCCCBBBBB 100
```

CSI2.0 results

http://csi.wishartlab.com/cgi-bin/results.py?upload_dir=/apps/csi

Secondary Structure Graph

Fig. 4 The CSI 2.0 (http://csi.wishartlab.com) web server showing screen shots of the home page and result pages

selectable options including the choice of running with or without PSIPRED. The web server is implemented as Python CGI-script. In general, the web server takes <60 s (if PSIPRED is off) or >140 s (if PSIPRED is on). A screen shot of the CSI 2.0 web server and its output is shown in Fig. 4.

Conclusion

CSI 2.0 represents a substantial improvement over the original CSI concept. In particular it uses an extended feature set derived from chemical shift and sequence data. It also replaces the simple digital filtering used in the original CSI algorithm with a more powerful "feature filter" that uses machine learning. Using the standard 3-state criteria (α-helix, β-strand and coil) and standard evaluation method such as Q3-accuracy, CSI 2.0 shows a significantly improved performance over the original CSI (89 vs. 80 %) as well as significantly improved performance over other available state-of-the-art secondary structure identification methods (89 vs. ~86 %). This performance improvement was statistically significant for the most common secondary structure assignment method, DSSP. Based on data presented here concerning the level of agreement between different secondary structure identification methods (NMR vs. X-ray vs. different programs), we suspect that we are at or near the maximum performance that secondary structure assignment methods can achieve. In addition to the performance improvement seen with CSI 2.0, we also showed that CSI 2.0 successfully detected different secondary structures in structurally dissimilar proteins sharing high sequence identity—something that commonly fools other programs. We also showed that CSI 2.0 is able to identify the (lack of) secondary structure in unfolded proteins.

To make this method publicly accessible, a CSI 2.0 webserver (http://csi.wishartlab.com) has been developed. It accepts chemical shift assignments in a variety of formats and generates colorful graphical output describing the identity and location of all secondary structure elements. We believe that CSI 2.0, with its superior performance will be a useful contribution to the field of biomolecular NMR. It should be particularly useful in the initial stages of conventional NMR structure generation (i.e. identifying homologous folds or providing useful torsion and distance restraints) as well as serving as a robust alternative to standard coordinate-based methods for secondary structure identification. CSI 2.0 is currently being used in the development of improved chemical shift-only 3D structure determination methods.

Acknowledgments The authors would like to thank Yongjie Liang for his help in preparing the CSI 2.0 web server. Financial support from the Natural Sciences and Engineering Research Council (NSERC), the Alberta Prion Research Institute (APRI) and PrioNet is gratefully acknowledged.

References

Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins Struct Funct Bioinform 59(3):467–475

Adams PD, Baker D, Brunger AT, Das R, DiMaio F, Read RJ, Richardson DC, Richardson JS, Terwilliger TC (2013) Advances, interactions, and future developments in the CNS, Phenix and Rosetta structural biology software systems. Annu Rev Biophys 43:265–287

Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci 106(50):21149–21154

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Andrec M, Snyder DA, Zhou Z, Young J, Montelione GT, Levy RM (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. Proteins Struct Funct Bioinform 69(3):449–465

Berjanskii MV, Wishart DS (2005) A simple method to predict protein flexibility using secondary chemical shifts. J Am Chem Soc 127(43):14970–14971

Berjanskii M, Tang P, Liang J, Cruz JA, Zhou J, Zhou Y, Bassett E, MacDonell C, Lu P, Wishart DS (2009) GeNMR: a web server for rapid NMR-based protein structure determination. Nucleic Acids Res 37((Web server issue)):W670–W677

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(1):235–242

Camilloni C, De Simone A, Vranken WF, Vendruscolo M (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. Biochemistry 51(11):2224–2231

Cheung MS, Maguire ML, Stevens TJ, Broadhurst RW (2010) DANGLE: a Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. J Magn Reson 202(2):223–233

Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36(suppl 2):W197–W201

Development Core Team R (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Durrett R (2010) Probability: theory and examples, vol 3. Cambridge University Press, London

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(1):71–81

Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B (2001) EVA: Continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17:1242–1243

Fesinmeyer RM, Hudson FM, Olsen KA, White GW, Euser A, Andersen NH (2005) Chemical shifts provide fold populations and register of β-hairpins and β-sheets. J Biomol NMR 33(4):213–231

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins Struct Funct Bioinform 23(4):566–579

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50(1):43–57

He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK (2009) Predicting intrinsic disorder in proteins: an overview. Cell Res 19(8):929–949

Hung LH, Samudrala R (2003) Accurate and automated classification of protein secondary structure with PsiCSI. Protein Sci 12(2):288–295

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202

Jones DT, Tress M, Bryson K, Hadley C (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins Suppl 3:104–111

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab-an S4 package for kernel methods in R. J Stat Softw 11:1–20

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(5):1–26

Labudde D, Leitner D, Krüger M, Oschkinat H (2003) Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts. J Biomol NMR 25(1):41–53

Levitt M (1978) Conformational preferences for globular proteins. J Am Chem Soc 17(20):4277–4284

Mielke SP, Krishnan VV (2004) An evaluation of chemical shift index-based secondary structure determination in proteins: influence of random coil chemical shifts. J Biomol NMR 30(2):143–153

Mielke SP, Krishnan VV (2009) Characterization of protein secondary structure from NMR chemical shifts. Prog Nucl Magn Reson Spectrosc 54(3–4):141–165

Momen-Roknabadi A, Sadeghi M, Pezeshk H, Marashi SA (2008) Impact of residue accessible surface area on the prediction of protein secondary structures. BMC Bioinform 9(1):357

Montgomerie S, Sundraraj S, Gallin WJ, Wishart DS (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinform 7:301

Ratnaparkhi GS, Ramachandran S, Udgaonkar JB, Varadarajan R (1998) Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. Biochemistry 37(19):6958–6966

Rost B, Sander C, Schneider R (1994) Redefining the goals of protein secondary structure prediction. J Mol Biol 235:13–26

Schwarzinger S, Kroon GJ, Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123(13):2970–2978

Shen Y, Bax A (2012) Identification of helix capping and β-turn motifs from NMR chemical shifts. J Biomol NMR 52(3):211–232

Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR 56(3):227–241

Shen Y, Delaglio F, Cornilescu G, Bax A (2009a) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44(4):213–223

Shen Y, Vernon R, Baker D, Bax A (2009b) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43(2):63–78

Shen Y, Bryan PN, He Y, Orban J, Baker D, Bax A (2010) De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. Protein Sci 19(2):349–356

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7(1):539

Soding J, Remmert M (2011) Protein sequence comparison and fold recognition: progress and good practice benchmarking. Curr Opin Struct Biol 21(3):404–411

Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33((Web server issue)):W244–W248

Tyagi M, Bornot A, Offmann B, de Brevern AG (2009) Analysis of loop boundaries using different local structure assignment methods. Prot Sci 18(9):1869–1881

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J et al (2008) BioMagResBank. Nucleic Acids Res 36(Suppl 1):D402–D408

UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38(Suppl 1):D142–D148

Valdar WSJ (2002) Scoring residue conservation. Proteins Struct Funct Bioinform 48(2):227–241

Wang G, Dunbrack RLJ (2003) PISCES: a protein culling server. Bioinformatics 19(12):1589–1591

Wang Y, Jardetzky O (2002a) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11(4):852–861

Wang Y, Jardetzky O (2002b) Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc 124(47):14075–14084

Wang CC, Chen JH, Lai WC, Chuang WJ (2007a) 2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts. J Biomol NMR 38(1):57–63

Wang L, Eghbalnia HR, Markley JL (2007b) Nearest-neighbor effects on backbone alpha and beta carbon chemical shifts in proteins. J Biomol NMR 39(3):247–257

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(13):3316–3319

Wishart DS (2011) Interpreting protein chemical shift data. Prog Nucl Magn Reson Spectrosc 58(1):62–87

Wishart DS, Case DA (2002) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338:3–34

Wishart DS, Nip AM (1998) Protein chemical shift analysis: a practical guide. Biochm Cell Biol 76(2–3):153–163

Wishart DS, Sykes BD (1994a) Chemical shifts as a tool for structure determination. Methods Enzymol 239:363–392

Wishart DS, Sykes BD (1994b) 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 (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 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((Web server issue)):W496–W502

Wuthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York

Wuthrich K (1990) Protein structure determination in solution by NMR spectroscopy. J Bio Chem 265(36):22059–22062

Zemla A, Venclovas C, Fidelis K, Rost B (1999) A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. Proteins 34:220–223

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

Zhang W, Dunker AK, Zhou Y (2008) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. Proteins Struct Funct Bioinform 71(1):61–67