

2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts

Ching-Cheng Wang · Jui-Hung Chen ·
Wen-Chung Lai · Woei-Jer Chuang

Received: 15 September 2006 / Revised: 3 January 2007 / Accepted: 15 January 2007 / Published online: 27 February 2007
© Springer Science+Business Media B.V. 2007

Abstract Chemical shifts of amino acids in proteins are the most sensitive and easily obtainable NMR parameters that reflect the primary, secondary, and tertiary structures of the protein. In recent years, chemical shifts have been used to identify secondary structure in peptides and proteins, and it has been confirmed that $^1\text{H}^\alpha$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, and $^{13}\text{C}'$ NMR chemical shifts for all 20 amino acids are sensitive to their secondary structure. Currently, most of the methods are purely based on one-dimensional statistical analyses of various chemical shifts for each residue to identify protein secondary structure. However, it is possible to achieve an increased accuracy from the two-dimensional analyses of these chemical shifts. The 2DCSi approach performs two-dimension cluster analyses of $^1\text{H}^\alpha$, $^1\text{H}^\text{N}$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, and $^{15}\text{N}^\text{H}$ chemical shifts to identify protein secondary structure and the redox state of cysteine residue. For the analysis of paired chemical shifts of 6 data sets, each of the 20 amino acids has its own 15 two-dimension cluster scattering diagrams.

Grant sponsor: National Science Council of ROC; Grant numbers: NSC-94-2323-B006-001, NSC-93-2212-E-006.

Electronic Supplementary Material The online version of this article (doi:10.1007/s10858-007-9146-x) contains supplementary material, which is available to authorized users.

C.-C. Wang · J.-H. Chen (✉) · W.-C. Lai
Institute of Manufacturing Engineering, National Cheng Kung University College of Electrical Engineering and Computer Science, Tainan 701, Taiwan
e-mail: second@mail.ncku.edu.tw

W.-J. Chuang
Department of Biochemistry and Molecular Biology,
National Cheng Kung University College of Medicine,
Tainan 701, Taiwan

Accordingly, the probabilities for identifying helix and extended structure were calculated by using our scoring matrix. Compared with existing the chemical shift-based methods, it appears to improve the prediction accuracy of secondary structure identification, particularly in the extended structure. In addition, the probability of the given residue to be helix or extended structure is displayed, allows the users to make decisions by themselves.

Keywords Chemical shift · Protein secondary structure identification · Two-dimension cluster

Abbreviations

2D	Two-dimension
BMRB	BioMagResBank
H	α -helix
G	3_{10} -helix
I	π -helix
B	β -strand
E	Extended structure
C	Random coil structure
NMR	Nuclear magnetic resonance
PDB	Protein data bank

Introduction

Early secondary structure prediction methods rely upon amino-acid sequence alone, and have been reported to be as high as 78% on selected datasets (Albrecht et al. 2003). It seems that there is an upper limit for prediction accuracy by using amino-acid sequence information only (Rost 2001). Over the past 15 years, the strong

correlation between chemical shift and secondary structure has been presented (Szilagyí and Jardetzky 1989; Pastore and Saudek 1990; Spera and Bax 1991; Wishart et al. 1992; Le and Oldfield 1994; Luginbuhl et al. 1995; Wishart and Nip 1998; Iwadate et al. 1999; Sibley et al. 2003). Obviously, chemical shift data provide valuable insights into identification of protein secondary structure. Wishart and Sykes (1994) proposed first automatic protocol (CSI) based on statistical analyses of $^1\text{H}^\alpha$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, and $^{13}\text{C}'$ NMR chemical shifts to identify the protein secondary structure. Since then various automatic programs to assign secondary structures have been presented along with a variety of nuclei chemical shifts and machine learning approach. Basically, two strategies of the secondary structure identification using chemical shifts were adopted. The first strategy was solely on the basis of chemical shift information, as used in CSI approach. In addition, the probability-based protein secondary-structure identification (PSSI) is also a chemical shift-based method (Wang and Jardetzky 2002). The second strategy was a combined algorithm using both chemical shifts and the sequence information. For instance, PsiCSI combines information from chemical shifts and protein sequences by using three layers of neural network approach (Hung and Samudrala 2003). A recent study by Eghbalian et al. (2005) shows that an energetic conformational analysis model (PECAN) that presents a framework of combining sequence and chemical shift yielded the most favorable energetic description to identify secondary structure.

As mentioned above, existing methods (PSSI, CSI, and PsiCSI) are one-dimensional analysis of chemical shift data, i.e., each residue has n types of decision indices to identify secondary structures if n kinds of nuclei chemical shifts were used. For example, for six data sets of chemical shifts, i.e., $^1\text{H}^\alpha$, $^1\text{H}^\text{N}$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, and $^{15}\text{N}^\text{H}$, then six independent chemical shift indices or probability values for each residue will be obtained. Thus, it is possible to estimate a consensus secondary structure by combining these indices or probabilities. Unlike the one-dimensional analysis, $n!/((n-2)! \times 2)$ decision indices with paired chemical shifts of n data sets could be calculated by using the two-dimensional analysis. Herein, paired chemical shifts of 6 data sets were analyzed, and then 15 probability values of being in helix or extended structure from their two-dimension cluster analyses for each of 20 amino acids were calculated. Therefore, it is reasonable to assume that global accuracy for identifying secondary structure obtained by two-dimensional analysis would be more accurate and reliable than those from one-dimensional.

In this paper, we applied the two-dimensional statistical analysis method to identify secondary structure in protein by using only chemical shift information. To avoid an artifactual increase in calculated average global accuracy, a dataset containing ~ 40,706 residues from 336 non-redundant proteins were used in evaluating the global accuracy of our 2DCSi method. In addition, we analyze the performance of the 2DCSi, CSI, and PsiCSI methods using a set of 45 reference-corrected novel proteins, and compare individual prediction accuracy of three secondary structure states. A 2DCSi web server is also established which enables users to submit NMR chemical-shift data and return protein secondary structure identifications and the redox states of cysteine residues in both simple graphic and tabular formats. In addition, the actual probability values of being in helix or extended structure states were presented.

Materials and methods

Description of the 2DCSi method

2DCSi is based on analyzing paired two-dimensional scattering diagrams of six chemical shift data sets, i.e., $^1\text{H}^\alpha$, $^1\text{H}^\text{N}$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, and $^{15}\text{N}^\text{H}$ to identify the secondary structure and redox state of amino-acid residues in proteins. A three-step method was employed herein: (i) data sets of chemical shifts and protein secondary structures were collected for analysis, and then cross-referenced; (ii) 15 cluster scattering diagrams were plotted for paired chemical shifts of 6 data sets, and the clusters as a function of the secondary structure were examined; (iii) the score matrices created for each of 20 amino acids were used to determine the secondary structures and redox state of cysteine residues.

Data sets of chemical shifts and protein secondary structures

Two preliminary data sets that separately containing chemical shifts of 20 amino acids and secondary structure assignments were generated. The Oct. 27, 2005 release of RefDB (Zhang et al. 2003) at the URL <http://www.redpoll.pharmacy.ualberta.ca> contains NMR chemical shifts of 601 proteins. We performed an automatic pattern matching procedure coded in ANSI C programming language to collect six chemical shifts. At the same time, data set of secondary structures was collected by following three different assignment programs. In order to reduce the bias to a particular

assignment, DSSP, STRIDE and VADAR (Kabsch and Sander 1983; Frishman and Argos 1995; Willard et al. 2003) were used to define the secondary structure. All definitions were reduced to three state models as follows: (i) DSSP: G and I to H, B to E, all other states to C; and (ii) STRIDE: G to H, b to E, all others states to C, where G is 3_{10} -helix, I is π -helix, B and b are isolated β -bridge. A simple majority rule (two out of three) was applied to look for the consensus of the three secondary structure assignments.

Then the above two primary data sets, i.e., data sets of chemical shifts and secondary structures, were cross-referenced and all chemical shifts were classified into three categories: helix (H), extended structure (E), and random coil structure (C). Six chemical shifts in three secondary structure states were collected into a data set, call the target data set. The target data set contained 601 BMRB entries—44377 $^1\text{H}^\alpha$, 55338 $^1\text{H}^\text{N}$, 44203 $^{13}\text{C}^\alpha$, 35317 $^{13}\text{C}^\beta$, 28099 $^{13}\text{C}'$, and 47986 $^{15}\text{N}^\text{H}$ chemical shifts—of which 377 entries contained cysteine residues are summarized in Table 1.

Two-dimension cluster analysis of chemical shifts

The successful identification of secondary structure and redox state of cysteine residue was already discussed in an earlier paper (Wang et al. 2006). Hence, it has been widely believed that it should be possible to apply similar cluster analysis to identify remaining 19 amino acids. A total of 267 scattering diagrams were plotted as shown in Supplementary material III. The 15 paired chemical shifts of 6 data sets plots were available for each of remaining 19 amino acids, except Gly and Pro which are short of 2 kinds of chemical shift information (Gly: C^β and H^α , Pro: N^H and H^N). Because the significance level of 10%, which corresponds to the selected 90% inclusion rate, has been one of the popular options tabulated for statistical analysis, a direct simplex search algorithm²¹ of the Matlab functions (MathWorks, Inc.) was employed to find the minimum of ellipse area and to ensure that each ellipse contains 90% of the chemical shifts in the same color. In addition, the cluster boundaries of all ellipses in 2D plots contained 90% of the chemical shifts suggesting that this method provides a 90% level of confidence of the prediction.

For example, the two-dimension $\text{N}^\text{H}/\text{C}'$, $\text{H}^\text{N}/\text{C}'$, and $\text{H}^\alpha/\text{C}'$ chemical shift plots of alanine residue exhibited

distinct clusters as shown in Fig. 1a–c, where colored ellipses mark cluster boundaries; helix in red, and extended structure in blue. Furthermore, an eyeball examination reveals that each $\text{H}^\text{N}/\text{N}^\text{H}$ plot of 20 amino acids hardly contributes in the secondary structure recognition. Thus, they are dropped from further consideration, and only 14 scattering diagrams for each residue were employed during our two-dimension cluster analysis.

Scoring matrix and decision ground rule

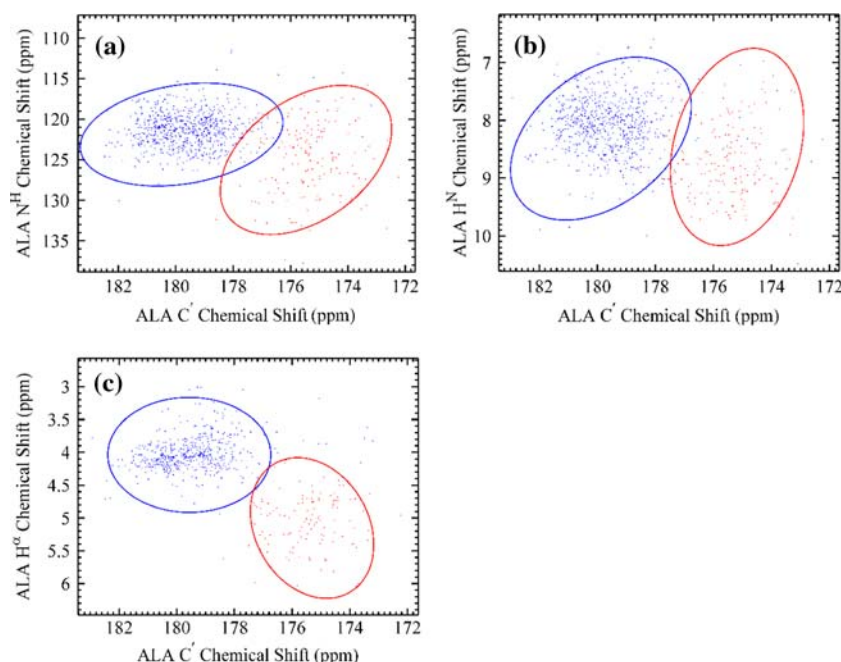
For every one of the 20 amino acids one-dimensional frequency plots for $^{15}\text{N}^\text{H}$, $^1\text{H}^\alpha$, $^1\text{H}^\text{N}$, $^{13}\text{C}^\beta$, $^{13}\text{C}^\alpha$, and $^{13}\text{C}'$ chemical shifts as a function of secondary structure are plotted in Supplementary material II. These plots are consistent with the idea that major difficulty of secondary structure identification with one-dimensional analysis is the overlap region between helix and extended structure. Similarly it is observed that in Fig. 1a, the secondary structure cannot be identified if the C' chemical shift is less than 178.4 ppm or greater than 176.2 ppm. However, an eyeball inspection reveals that the $\text{H}^\text{N}/\text{C}'$, and $\text{H}^\alpha/\text{C}'$ plots of Fig. 1b, c plots are more competent for distinguishing helix from extended structure in this overlap region (176.2–178.4 ppm). In particular, helix and extended structure clearly fall in two distinct clusters in Fig. 1c. All other nuclei share the similar property with various two-dimension cluster plots examined. Therefore, we believe that different plots can provide helpful information to distinguish the secondary structures. As a result, the probability scores were calculated using the score matrix as shown in Supplementary material I Table S1, where $\hat{P}r(\zeta|\chi_1, \chi_2)$ represents the probability of an ζ -state for observed chemical shifts of χ_1 and χ_2 , and $\tau(\zeta)$ is the sum of those 14 probability scores in that column. It is noted that (χ_1, χ_2) might take values of $(\text{c}^\alpha, \text{c}^\beta)$, $(\text{c}', \text{c}^\alpha)$, $(\text{n}^\text{h}, \text{c}^\alpha)$, $(\text{h}^\alpha, \text{c}^\alpha)$, $(\text{h}^\text{n}, \text{c}^\alpha)$, $(\text{c}', \text{c}^\beta)$, $(\text{n}^\text{h}, \text{c}^\beta)$, $(\text{h}^\alpha, \text{c}^\beta)$, $(\text{h}^\text{n}, \text{c}^\beta)$, $(\text{n}^\text{h}, \text{c}')$, $(\text{h}^\alpha, \text{c}')$, $(\text{h}^\text{n}, \text{c}')$, $(\text{h}^\alpha, \text{n}^\text{h})$, and $(\text{h}^\text{n}, \text{h}^\alpha)$, and c^α , c^β , c' , n^h , h^α , and h^n are values of C^α , C^β , C' , N^H , H^α , and H^N chemical shifts, respectively. In addition, ζ can be in either the helix (H) or extended structure (E). The random coil structure (C) is defined simply as neither the helix nor extended structure.

Three situations are distinguished while applying two-dimension cluster plots to estimate $\hat{P}r(\zeta|\chi_1, \chi_2)$: (i) (χ_1, χ_2) falls outside all elliptical areas; (ii) (χ_1, χ_2) falls

Table 1 Summary of chemical shifts of amino acids of target data set

	No. of proteins	$^1\text{H}^\alpha$	$^1\text{H}^\text{N}$	$^{13}\text{C}^\alpha$	$^{13}\text{C}^\beta$	$^{13}\text{C}'$	$^{15}\text{N}^\text{H}$
Cysteine only	377	1,303	1,410	714	591	488	899
All amino acid residues	601	44,377	55,338	44,203	35,317	28,099	47,986

Fig. 1 Two-dimension N^H/C' (a), H^N/C' (b), and H^α/C' (c) chemical shift plots of alanine residue. The chemical shifts of helix and extended structure were shown in blue and red, respectively. The ellipse of helix and extended structure contain 90% of the chemical shifts



onto one and only one elliptical area; (iii) (χ_1, χ_2) falls onto an intersection area of two ellipses. The decision ground rule describes as follows:

Rule 1. Add up probability scores of each column in the scoring matrix to obtain the total score $\tau(\zeta)$ for secondary structure states.

Rule 2. Identify state ζ if and only if $\tau(\zeta) \geq 0.8 \times \lambda$, where 0.8 represents the decision threshold we chose after our target data set with 601 entries were tested, and λ is the total number where (χ_1, χ_2) resides at either [i] or [ii] aforementioned situations.

The probability score equals 0 and can not applied to scoring matrix when situation (iii) occurs. Thus, it is reasonable to suggest that the more λ amounts will improve the reliability to estimate the prediction accuracy. In addition to the actual probability values of being in helix or extended structure states, the λ values were displayed for each residue by running our 2DCSi.

Results and discussion

2DCSi web server

A web-based server called 2DCSi (<http://www.ncku.2dcsi.idv.tw/>) which aims to perform the secondary structure identification by chemical shifts was described herein. It can be used via a World Wide Web (WWW) browser or a stand-alone program running under the MS/DOS operating system. The 2DCSi input takes only the chemical shift file in BMRB (NMR-STAR) format,

and the output provides a secondary structure state and the redox states of cysteine residues in both graphic and tabular formats as shown in Fig. 2. Besides, for the user it would be more useful to know the actual probability of being in different secondary structure states. Thus, the values of probability for the given residues to be helix or extended structure states were

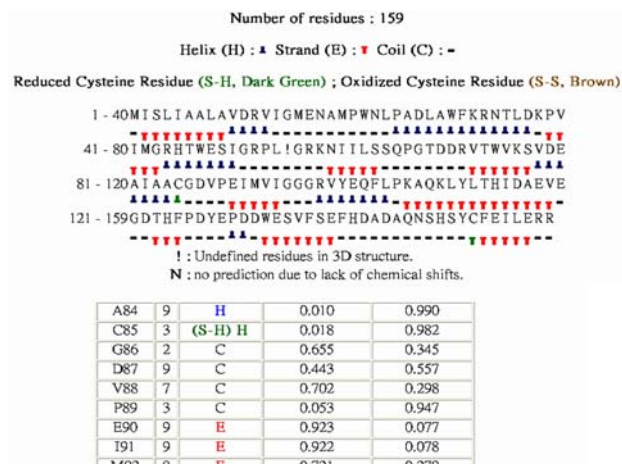


Fig. 2 The 2DCSi web server is shown in both graphic and tabular formats using BMRB5741 as an example. The graphic output shows protein sequences with helix in blue, extended structure in red, oxidized cysteine residue in brown, and reduced cysteine residue in dark green. In tabular format, columns 1–5 show amino-acid residue number, identified secondary structure and redox state, the number of 2D plots applied to score matrix (λ), $p(E)$ probability of extended structure, and $p(H)$ probability of helix, respectively

also presented. The downloadable stand-alone program is available for free at this server.

RefDB analysis

Some algorithms to identify secondary structure claimed high accuracy, but on small datasets those were also used in training the methods. Therefore, we selected RefDB, a secondary database of reference-corrected protein chemical shifts derived from the BMRB, to estimate secondary structure prediction accuracy. The RefDB performed sequence homology search to prevent the redundancy in protein sequences, and the proteins with less than 60% sequence homology were included. Of these 601 proteins for which matching PDB entries could be identified; 101 contain only ^1H assignments; 71 have both ^1H and ^{15}N assignments, and 165 proteins for all H^α , H^N , C^α , C^β , C' , and N^H assignments included. As can be seen from Table 2, there are strong correlations between prediction accuracy and the used nucleus numbers. However, we are often unable to expect that all kinds of nuclei were measured by NMR spectroscopy for each residue. Therefore, we come to a compromise using at least any three chemical shifts of six data sets. The global accuracy of ~87% with the RefDB dataset (336 proteins; 40,706 residues) was observed. Moreover, for 165 proteins with a full 6 data sets of chemical shifts, 2DCSi can achieve an average global accuracy of > 88%. More complete details of these results are available in the Supplementary material IV. The Q_3 global accuracy, which is the same as definition by Chou and Fasman (1974), is given by

$$Q_3 = \frac{(N - \text{total incorrect})}{N},$$

where *total incorrect* is the total number of residues whose secondary structure states are identified incorrectly and *N* is the number of residues in the protein. We emphasize that the accuracy reported in this paper is the global accuracy for full set of residue, not calculated by well-defined or core secondary structure fragments.

Table 2 Average global accuracy of different nucleus combinations

Nucleus types	# Entries	$Q_3(\%)$	H	E	C
($^1\text{H}^\alpha, ^1\text{H}^\text{N}$)	101	76.06	84.47	64.84	85.69
($^{15}\text{N}^\text{H}, ^1\text{H}^\alpha, ^1\text{H}^\text{N}$)	71	80.62	90.89	68.15	82.99
($^{13}\text{C}^\alpha, ^{13}\text{C}^\beta, ^{13}\text{C}'$, $^{15}\text{N}^\text{H}, ^1\text{H}^\alpha, ^1\text{H}^\text{N}$)	165	88.14	91.16	87.78	87.24
At least three nuclei	336	86.71	90.42	84.69	83.73

Comparison with existing methods

In order to compare our method with others, a data set of 102 new proteins that does not include proteins from the RefDB has been downloaded from the “New Entries” page of BioMagResBank (BMRB; <http://www.bmrwisc.edu>) since October 27, 2005, and those meeting the following criteria were removed: (i) Only one kind of nucleus was measured, (ii) No any corresponding PDB entries, (iii) Many missing chemical shifts were observed, and (iv) Paramagnetic proteins. Thus, this left us with 45 proteins (~5329 residues) for which we could perform the secondary structure identification with 3 different programs. To obtain the reference corrections in chemical shifts, we ran SHIFTCOR program (Zhang et al. 2003) on this 45 proteins dataset at <http://www.redpoll.pharmacy.ualberta.ca/shiftco>. The average global accuracies (Q_3) of 2DCSi, CSI, and PsiCSI over this dataset are shown in Table 3. The CSI Q_3 accuracy was ~84% with a standard deviation of 6.8%; both 2DCSi and PsiCSI achieved an average global accuracy of at least 87% with a standard deviation of 5.3% and 5.7%, respectively. Obviously, 2DCSi provides the smallest standard deviation and it might suggest that the secondary structure prediction is more reliable than others.

In general, a major weakness of existing methods is in distinguishing extended from random coil structure. However, our 2DCSi achieves the best global accuracy for identifying extended structure from all the data sets used in this paper. Using BMR7004 as an example, CSI and PsiCSI are partially correct on the identification of longer extended structure fragment (Val82–Thr97). It has been shown that the shorter extended structure (Val82–Ala88), and incorrect random coil structure (Asp89–Thr97) are identified. In contrast, 2DCSi is able to identify this longer extended structure fragment completely as shown in Supplementary material V. These results demonstrate that 2DCSi is more accurate for identifying extended structure than other methods. The more detailed results and the frequencies of the different types of errors are tabulated in Supplementary material I Table S2, and Table S3, respectively.

Pecan method was not included here because it does not assign secondary structures from their web server report. It just gives the probabilities and allows users to make decision by themselves. However, Pecan reported a Q_3 accuracy of 90% for 36 proteins with ~ 6100 residues from testing dataset of PSSI reporting. In order to compare Pecan, we used same dataset and achieved 88% Q_3 accuracy. More complete details of these results are available in the Supplementary material VI. Again, we emphasize

Table 3 Average global accuracy of 2DCSi, CSI, and PsiCSI^a

Program	$Q_3(\%)$	H	E	C
2DCSi	87.71	89.92	85.25	86.51
CSI	84.35	82.85	76.77	88.66
PsiCSI	87.78	90.15	76.54	89.76

^a Forty-five non-overlapping protein were used for the test. The test contained 5329 residues with 2193 H, 861 E, and 2275 C

that the 88% accuracy obtained in this dataset is the global accuracy for full set of residue. However, the 90% accuracy obtained by the Pecan method used core scoring protocol, it has been believed that it should be possible to increase the accuracy of prediction by apply well-defined or core secondary structure fragments.

Only ¹H nucleus chemical shift Identification

As we showed earlier, there is a positive relationship between the usable nuclei and global accuracy. However, it is not reasonable to assume that the each protein would be measured by all three different atoms, i.e., ¹H, ¹³C, and ¹⁵N. Thus, we make no secondary structure identification for each residue which is less than three chemical shifts of six data sets; instead it gives the probability and allows users to make decision by themselves. It indicates the probability for the given residue to be in the given state, and user could set a threshold based on their needs. For example, the bmr5207 protein only included ¹H nucleus chemical shift assignment, the probability values for each residue output by our 2DCSi program was shown in Supplementary material I Table S4. In general, a residue can be readily distinguished from the other state when it has a probability of that secondary structure type above 0.8. With the decision threshold of 0.8, we obtained a ~86% accuracy.

Conclusion

In 1994 Wishart and Sykes carried out a one-dimensional statistical analysis of chemical shift-based for the identification of protein secondary structure. Since then there are two different strategies for identifying secondary structures either by chemical shift alone or upon combination with sequence information. All of these methods based on a one-dimensional analysis to provide some indices of identifiable secondary structures from various nuclei chemical shifts. However, each of the 15 paired chemical shifts, except for

¹⁵N/¹H plot, has its characteristic cluster to separate helix from extended structure on specific amino acids by analyzing the two-dimension cluster scattering diagrams. For example, ¹⁵N/¹H/¹³C^α plots of Ala, Leu, Met, Ser, and Thr are very precise in distinguishing the helix from extended structure. Similarly, ¹³C^α/¹³C^β plots are good in Ala, Gln, His, Ile, Phe, Ser, and Thr. However, the most significant scattering diagram of 15 paired chemical shifts is the ¹³C^α/¹³C^β plot. It is widely useful and clearly distinguishing between helix from extended structure for each of 19 amino acids. This result differ slightly from the earlier study applied four ¹³C^α/¹³C^β, ¹³C^α/¹³C^γ, ¹³C^β/¹³C^γ, and ¹³C^α/¹³C^γ plots to identify secondary structure and redox state of cysteine residue (Wang et al. 2006).

In this work, we have shown that the 2DCSi method, only based on chemical shift information, produced a better average global accuracy and the smallest standard deviation than the existing chemical shift-based approaches, especially for the extended structure. In addition, the cluster boundaries of all ellipses in 2D plots contained 90% of the chemical shifts, i.e., a significance level of 10%, suggesting that 2DCSi provides a 90% level of confidence of the prediction. For proteins with a full six data sets of chemical shifts, i.e., ¹H^α, ¹H^N, ¹³C^α, ¹³C^β, ¹³C^γ, and ¹⁵N^H, 2DCSi can achieve an average global accuracy of > 88%. This is in part due to the decrease of overlap problem of one-dimensional analysis. This result suggests that the further improvement will be obtained from an *n*-dimensional analysis (*n* > 2).

Acknowledgements We are indebted to Dr. Wenya Huang for valuable comments. This work was supported by grants NSC-94-2323-B006-001 and NSC-93-2212-E-006 from the National Science Council, ROC, and by grant 91-B-FA09-1-4 from the Ministry of Education's Program for Promoting Academic Excellence in Universities.

References

- Albrecht M, Tosatto SC, Lengauer T, Valle G (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng* 16:459–462
- Chou PY, Fasman GD (1974) Prediction of protein conformation. *Biochemistry* 13:222–245
- Eghbali HR, Wang L, Bahrami A, Assadi A, Markley JL (2005) H-1, N-15 and C-13 resonance assignments of a protein involved in the autophagy process, At4g21980.1 from Arabidopsis. *J Biomol NMR* 32:71–81
- Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins* 23:566–579
- Hung LH, Samudrala R (2003) Bioverse: functional, structural and contextual annotation of proteins and proteomes. *Protein Sci* 12:288–295

- Iwadate M, Asakura T, Williamson MP (1999) C-alpha and C-beta carbon-13 chemical shifts in proteins from an empirical database. *J Biomol NMR* 13:199–211
- Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
- Le H, Oldfield E (1994) Correlation between ¹⁵N NMR chemical shifts in proteins and secondary structure. *J Biomol NMR* 4:341–348
- Luginbuhl P, Szyperski T, Wuthrich K (1995) Statistical Basis for the Use of ¹³C alpha Chemical Shifts in Protein Structure Determination. *J Magn Reson B* 109:229–233
- Pastore A, Saudek V (1990) The relationship between chemical shift and secondary structure in proteins. *J Magn Reson* 90:165–176
- Rost B (2001) Review: protein secondary structure prediction continues to rise. *J of Struct Biol* 134:204–218
- Sibley AB, Cosman M, Krishnan VV (2003) An empirical correlation between secondary structure content and averaged chemical shifts in proteins. *Biophys J* 84:1223–1227
- 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) α -Proton chemical shifts and secondary structure in proteins. *J Magn Reson* 83:441–449
- Wang CC, Chen JH, Yin SH, Chuang WJ (2006) Predicting the redox state and secondary structure of cysteine residues in proteins using NMR chemical shifts. *Proteins* 63:219–226
- Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci* 11:852–861
- 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, Nip AM (1998) Protein chemical shift analysis: a practical guide. *Biochem Cell Biol* 76:153–163
- 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* 4:171–180
- Wishart DS, Richard FM, Sykes BD (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* 31:1647–1651
- Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. *J Biomol NMR* 25:173–195