Predicting the redox state and secondary structure of cysteine residues in proteins using NMR chemical shifts

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · APRIL 2006

Impact Factor: 2.63 · DOI: 10.1002/prot.20875 · Source: PubMed

4 AUTHORS, INCLUDING:

Woei-Jer Chuang
National Cheng Kung University
91 PUBLICATIONS **1,256** CITATIONS

Predicting the Redox State and Secondary Structure of Cysteine Residues in Proteins Using NMR Chemical Shifts

Ching-Cheng Wang,[1] Jui-Hung Chen,[1] Shih-His Yin,[1] and Woei-Jer Chuang[2]*

[1]*Institute of Manufacturing Engineering, National Cheng Kung University College of Electrical Engineering and Computer Science, Tainan, Taiwan*

[2]*Department of Biochemistry and Molecular Biology, National Cheng Kung University College of Medicine, Tainan, Taiwan*

ABSTRACT **We report 2D cluster analyses of $^{1}H^{\alpha}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, and $^{13}C'$ versus $^{13}C^{\beta}$ NMR chemical shifts (CSs) that can be used to predict the redox state and secondary structure of cysteine residues in proteins. A database of cysteine $^{1}H^{\alpha}$, $^{1}H^{\beta 2}$, $^{1}H^{\beta 3}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{15}N^{H}$ CSs as a function of secondary structure and redox state was constructed from BioMagResBank entries. One-dimensional statistical analysis showed that cysteine $^{1}H^{\alpha}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C'$, and $^{15}N^{H}$ CSs reflected the secondary structure, and that cysteine C^{β} CS is extremely sensitive to the redox state. In contrast, cysteine $^{1}H^{\beta}$ CS was not correlated with its redox state or secondary structure. Two-dimensional cluster analysis revealed that 2D C^{α}/C^{β}, C'/C^{β}, H^{N}/C^{β}, and H^{α}/C^{β} clusters were helpful in distinguishing both the redox state and secondary structure of cysteine residues. Based on these results, we derived rules using a score matrix to predict the redox state and secondary structure of cysteines using their CSs. The score matrix predicts the redox state and secondary structure of cysteine residues in proteins with ~90% accuracy. This suggests that the redox state and secondary structure of cysteine residues in peptides and proteins can be obtained from their CSs without recourse to nuclear Overhauser effect measurements. Proteins 2006;63:219–226.**

Key words: chemical shifts; disulfide-bonded proteins; NMR; prediction; redox state; secondary structure

INTRODUCTION

Cysteine residues and disulfide bonds are important for protein structure and function.[1] Disulfide bond formation is a posttranslational modification; it is also a reversible process via thiol disulfide exchange. Both intramolecular and intermolecular disulfide bonds help stabilize the tertiary and quaternary structures of many proteins.[2] Free cysteine residues can also be essential for protein function, participating in substrate binding and enzyme catalysis.[3] Therefore, identification of the redox state and secondary structure of cysteine residues and disulfide bonds is critical for understanding protein structure and function.[4,5]

Chemical shifts (CSs) of amino acids in proteins may be the most sensitive and easily obtainable NMR parameters that reflect the primary, secondary, and tertiary structure of the protein.[6–8] Currently, NMR CSs are routinely used to determine protein secondary structure by comparing the observed $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, or $^{13}C'$ CSs with their random coil values.[9–19] Sharma and Rajarathnam[20] showed that the C^{β} CS of cysteine amino acid is extremely sensitive to the redox state, and it can be used to distinguish the redox state of cysteine residues. They also showed that C^{α} and C^{β} CSs of cysteine amino acid in different redox states occupy distinct clusters as a function of secondary structure in the C^{α}/C^{β} CS map. Because the databases of CSs and three-dimensional (3D) protein structures have been rapidly accumulated over the past years, it is worth re-evaluating the correlation of the redox state and secondary structure of cysteine residues with new cysteine CSs. Not only the $^{13}C^{\alpha}$ CS, but also the $^{1}H^{\alpha}$, $^{1}H^{N}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{15}N^{H}$ CSs are sensitive to the secondary structure of proteins. Therefore, we included more cysteine CSs and incorporated $^{1}H^{\alpha}$, $^{1}H^{\beta 2}$, $^{1}H^{\beta 3}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{15}N^{H}$ for 2D cluster analysis in this study. On the basis of the results of CS analysis, we derived three rules using a score matrix to predict the redox state and secondary structure of cysteines. The probability analysis of 2D C^{α}/C^{β}, C'/C^{β}, H^{N}/C^{β}, and H^{α}/C^{β} suggest that prediction of both the redox state and secondary structure of cysteines in proteins using their NMR CSs is feasible.

MATERIALS AND METHODS

Data Sets of Cysteine Chemical Shifts and Protein Secondary Structures

Two data sets of cysteine CSs and secondary structure assignments were generated for the analysis. A data set of cysteine CSs was obtained from the NMR CSs of 3207

Abbreviations: BMRB, BioMagResBank; B, β-strand; CS, chemical shift; CSI, chemical shift index; H, α-helix; NMR, nuclear magnetic resonance; PDB, protein data bank.

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585suppmat/

Grant sponsor: National Science Council of ROC; Grant numbers: NSC-93-2311-B006-004, NSC-93-2212-E-006. Grant sponsor: Program for Promoting University Academic Excellence; Grant number: 91-B-FA09-1-4.

*Correspondence to: Woei-Jer Chuang, Department of Biochemistry, National Cheng Kung University College of Medicine, Tainan 701, Taiwan. E-mail: wjcnmr@mail.ncku.edu.tw

Received 7 May 2005; Revised 27 October 2005; Accepted 4 November 2005

Published online 27 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20875

TABLE I. Summary of Cysteine Chemical Shifts

	No. of proteins	C^{α}	C^{β}	C'	N^{H}	H^{N}	H^{α}	$H^{\beta 2}$	$H^{\beta 3}$
Oxidized	191[a]	379	297	280	549	987	919	887	873
Reduced	190[b]	335	294	208	350	423	384	359	345
Total	377[c]	714	591	488	899	1410	1303	1246	1218

[a]Number of proteins contained cysteines in the oxidized state.
[b]Number of proteins contained cysteines in the reduced state.
[c]Total number of cysteine/cystine containing proteins. A protein may contain both oxidized and reduced cysteines. Thus, the sum of 191 and 190, in comparison to 377, revealed that 4 proteins contained both oxidized and reduced cysteines.

entries collected from the BioMagResBank (BMRB) at the URL http://www.BMRB.wisc.edu in the October 27, 2004 release. To avoid the CS referencing problem, we used either the RefDB database, a secondary database of reference-corrected protein CSs derived from the BMRB, or PSSI software to adjust their CSs.[21,22] Paramagnetic proteins, which were identified through checking the paramagnetic property as reported in each BMRB entry, were not included in the database. We also performed a sequence homology search to prevent redundancy in protein sequences; proteins with less than 60% sequence homology were included in the database. Using an automatic pattern-matching procedure coded in ANSI C programming language (Supplementary Material), we performed a blanket search for cysteines with linkages to residue atoms of the 20 naturally occurring amino acids. Eight cysteine CSs ($^{1}H^{\alpha}$, $^{1}H^{\beta 2}$, $^{1}H^{\beta 3}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{15}N^{H}$) were collected for the analysis. Cysteines with linkages to metal or nonprotein ligands were excluded from the data sets.

A data set of secondary structures was collected from X-ray and NMR protein structures in the protein data bank (PDB) that contained structural information of cysteine residues in the corresponding BMRB entries. The automatic definitions produced by the method of Kabsch and Sander (dictionary of protein secondary structure; DSSP) were applied to assign the secondary structure of the published coordinates.[23]

Assignments of Secondary Structure and Redox State of Cysteine Residues in Proteins

Data sets of cysteine CSs and secondary structure were cross-referenced, and all CSs were classified into three categories: α-helix (H), β-strand (E), and others (O). Cysteine and cystine were referred to as the reduced and oxidized form of the cysteine amino acid. The redox state of cysteine residues was assigned according to the "SSBOND" record deposited in the connectivity annotation section of the PDB identification code and was verified by checking the coordinate section along with the corresponding BMRB entry. If multiple PDB codes had been deposited for a BMRB entry, the BMRB-entry-referred PDB code was applied. Cysteine CSs as a function of α-helix (H) and β-strand (E) secondary structure in reduced (Red) and oxidized (S-S) states were collected into a data set, named the target data set.

The target data set contained 377 BMRB entries, of which 187 entries had only oxidized cysteines, 186 entries had only reduced cysteines, and 4 entries had both oxidized and reduced cysteines. Therefore, 191 entries had oxidized cysteines and 190 entries had reduced cysteines, and their 1303 $^{1}H^{\alpha}$, 1246 $^{1}H^{\beta 2}$, 1218 $^{1}H^{\beta 3}$, 1410 $^{1}H^{N}$, 714 $^{13}C^{\alpha}$, 591 $^{13}C^{\beta}$, 488 $^{13}C'$, and 899 $^{15}N^{H}$ CSs are summarized in Table I. The target data set was tabulated in the supplementary material that included protein name, the BMRB entries, and their corresponding PDB entries.

Cluster Analysis of Cysteine Chemical Shifts

One- (1D) and two-dimensional (2D) clusters of cysteine CSs as a function of secondary structure and redox state were plotted and analyzed using a Microsoft Excel spreadsheet program and the Matlab program (MathWorks, Inc). Histograms of $^{1}H^{\alpha}$, $^{1}H^{\beta 2}$, $^{1}H^{\beta 3}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{15}N^{H}$ versus frequency and 2D clusters of 15 paired CSs of $^{1}H^{\alpha}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C'$, $^{13}C^{\beta}$, and $^{15}N^{H}$ were examined using the target data sets. The data points of cysteine CSs defined as the α-helix structure in reduced state (Red, H), α-helix structure in oxidized state (Oxi, H), β strand in reduced state (Red, E), and β strand in oxidized state (Oxi, E) are shown in red, black, blue, and green, respectively (Figs. 1 and 2). 2D ellipses marked cluster boundaries that were searched using the Matlab program. Each ellipse contained 90% of the CSs, which were used due to joints between the ellipses. Thus, the [(Red, H), (Oxi, H)] and [(Red, E), (Oxi, E)] ellipses were disjointed in the C'/C^{β} plot and the [(Red, E), (Oxi, H)] was disjointed in the H^{α}/C^{β} plot.

Score Matrix and Prediction Rules

Four of 15 paired CSs of $^{1}H^{\alpha}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C'$, $^{13}C^{\beta}$, and $^{15}N^{H}$ as a function of secondary structure and redox state occupied in distinct clusters. 2D cluster analyses of C^{α}/C^{β}, C'/C^{β}, H^{α}/C^{β}, and N^{H}/C^{β} were used to derive three rules using the score matrix to predict the redox state and secondary structure of cysteines using their CSs. The score matrix is shown in Table SI. $\hat{P}r(\zeta|\chi_1\chi_2)$ represented the probability of a ζ-state for observed CSs of χ_1 and χ_2, and $\tau(\zeta)$ was the sum of those four probability scores. It should be noted that (χ_1, χ_2) was the value of (c^{α}, c^{β}), (c', c^{β}), (h^{α}, c^{β}) and (n^{h}, c^{β}) states, and c^{α}, c^{β} c^{o}, h^{α}, and n^{h} was the value of C^{α}, C^{β}, C', H^{α}, and N^{H} CSs, respectively. In addition, ζ took one value of four states: (Red, H), (Red, E), (Oxi, H), and (Oxi, E).

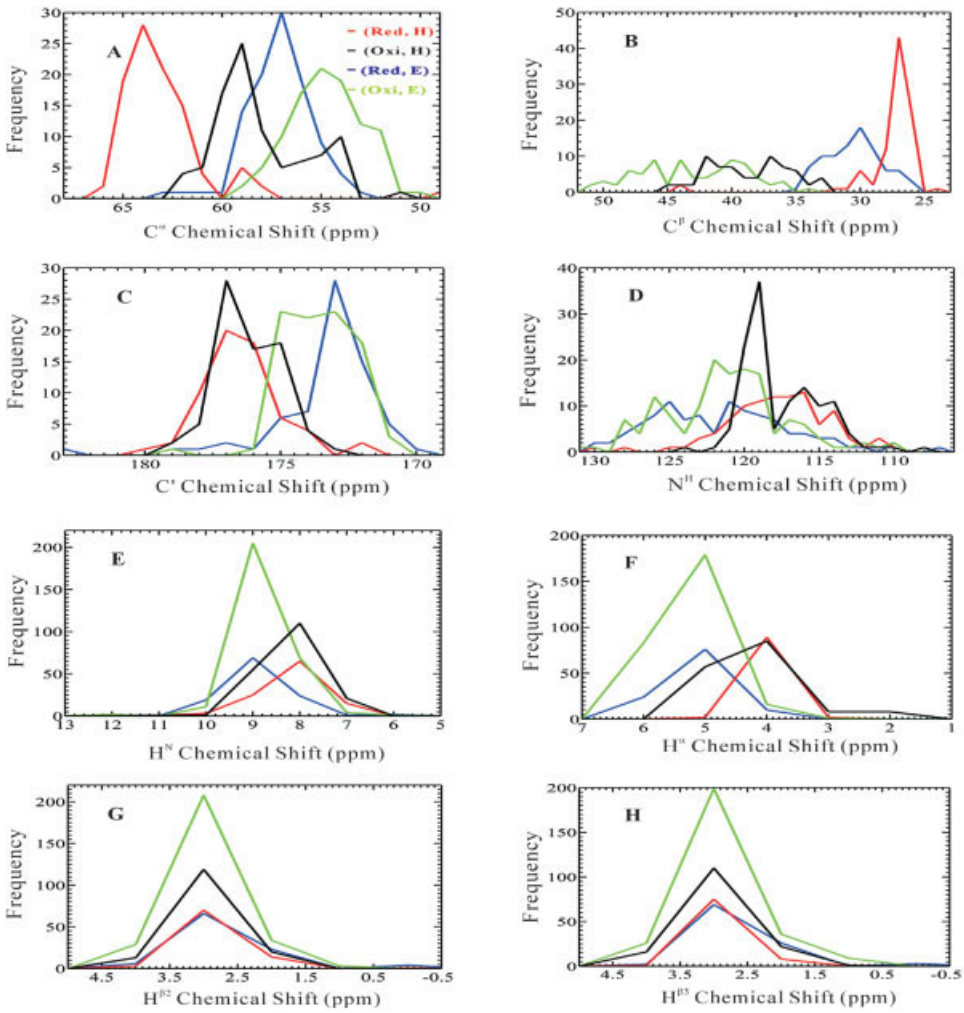

Fig. 1. Distribution of cysteine C^{α} (**A**), C^{β} (**B**), C′ (**C**), N^{H} (**D**), H^{N} (**E**), H^{α} (**F**), $H^{\beta 2}$(**G**), and $H^{\beta 3}$ (H) CSs as function of both redox state and secondary structure. The α-helix structure in reduced state, α-helix structure in oxidized state, β-strand in reduced state, and β-strand in oxidized state are shown in red, black, blue, and green, respectively. NMR C^{α}, C^{β}, C′, and N^{H} CSs are shown in 1 ppm intervals, and the H^{N}, H^{α}, $H^{\beta 2}$, and $H^{\beta 3}$ CSs are shown in 0.25 ppm intervals.

RESULTS

1D Cluster Analysis of Cysteine Chemical Shifts

NMR CSs of $^{1}H^{\alpha}$, $^{1}H^{\beta 2}$, $^{1}H^{\beta 3}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{15}N^{H}$ were analyzed as a function of secondary structure and redox state, and frequency plots of NMR CSs are individually presented in Figure 1(A–H). The statistics of eight CSs are summarized in Tables SII–SIX. The average C^{α} CSs of oxidized cysteines in α-helix and β-strand structures were shifted to the downfield with higher ppm values and to the upfield with lower ppm values, respectively (Fig. 1A and Table SII). The mean ppm values of C^{α} CSs as a function of secondary structure and redox state were (54.7 ppm; Oxi, E) < (57.1 ppm; Red, E) < (58.1 ppm; Oxi, H) < (62 ppm; Red, H). They were consistent with the results of C^{α} CSs of the other 19 amino acids, which were sensitive to their secondary structure. In contrast, the C^{β} CS of cysteine was sensitive to the redox state but not to the secondary structure. The C^{β} CSs of oxidized and reduced cysteines were shifted to the downfield and upfield, respectively (Fig. 1B and Table SIII). The mean ppm values of C^{β} CSs as a function of secondary structure and redox state were (27.6 ppm; Red, H) < (30.5 ppm; Red, E) < (38.6 ppm; Oxi, H) < (43.1 ppm; Oxi, E). Their values were similar to the reported values of C^{β} CSs that were 26.5, 29.7, 38.4, and 43 ppm, respectively.[20] The mean ppm values of C′ CSs as a function of secondary structure and redox state were (173.3 ppm; Red, E) ≈ (173.6 ppm; Oxi, E) < (176.2 ppm; Oxi, H) ≈ (176.5 ppm; Red, H), indicating that the correlation of C′ CSs with their secondary structure was distinguishable, but the correlation of these with their redox state was indistinguishable (Fig. 1C and Table SIV). Similar to the trends for C′, the mean ppm values of N^{H}, H^{N}, and H^{α} CSs as a function of secondary structure and redox state were (Oxi, H) ≈ (Red, H) < (Oxi, E) ≈ (Red, E). In contrast,

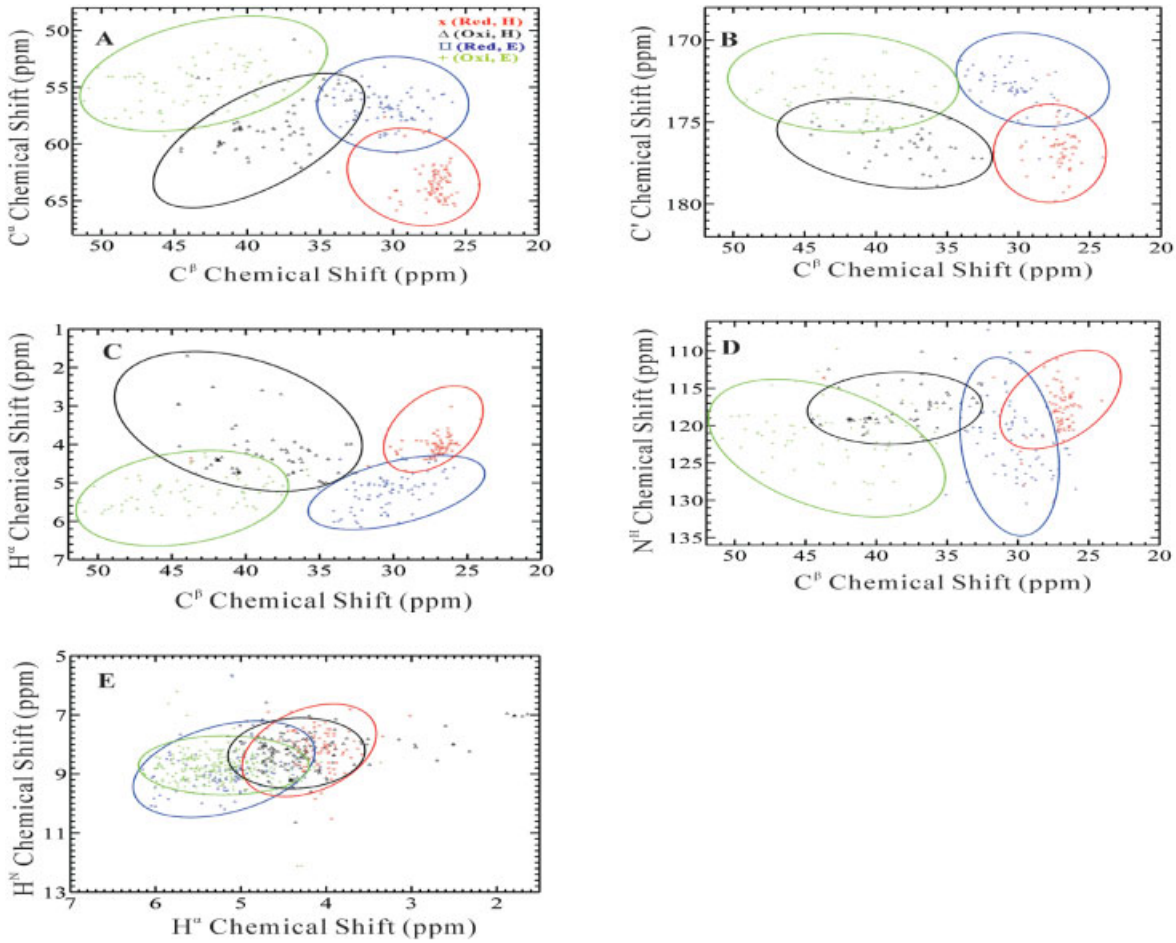

Fig. 2. 2D C^{α}/C^{β} (**A**), C'/C^{β} (**B**), H^{α}/C^{β} (**C**), N^{H}/C^{β} (**D**), and H^{N}/H^{α} (**E**) CS plots as a function of both redox state and secondary structure. The CSs of α-helix structure in reduced state (×), α-helix structure in oxidized state (△), β-strand in reduced state (□), and β-strand in oxidized state (+) are shown. The ellipses of the α-helix structure in reduced state (Red, H), α-helix structure in oxidized state (Oxi, H), β-strand in reduced state (Red, E), and β-strand in oxidized state (Oxi, E) are shown in red, black, blue, and green, respectively.

Figure 1(G,H) and Tables SVIII–SIX revealed that the CSs of $H^{\beta 2}$ and $H^{\beta 3}$ as a function of secondary structure and redox state were indistinguishable. Therefore, $H^{\beta 2}$ and $H^{\beta 3}$ CSs were excluded from 2D cluster analysis.

2D Cluster Analysis of Paired Chemical Shifts

Fifteen scatter diagrams were analyzed for paired CSs of $^{1}H^{\alpha}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C'$, $^{15}N^{H}$, and $^{13}C^{\beta}$. Clusters as a function of secondary structure and redox state were examined. Five of 15, including C^{α}/C^{β}, C'/C^{β}, H^{α}/C^{β}, N^{H}/C^{β}, and H^{N}/H^{α}, exhibited distinct clusters with paired CSs, and their 2D plots are shown in Figure 2(A–E). The ellipses of (Red, H), (Oxi, H), (Red, E), and (Oxi, E) are shown in red, black, blue, and green, respectively. These ellipses marked cluster boundaries, and each ellipse contained 90% of the CSs. Depending on the 2D plots, these ellipses were joined to or disjointed from each other with six possibilities [Fig. 2(A–D)]. For example, the (Red, H)-ellipse is externally disjointed from the (Oxi, H)- and (Oxi, E)-ellipses, and the (Red, E)- and (Oxi, E)-ellipses are also externally disjointed. In contrast, three pairs [(Red, E), (Red, H)], [(Red, E), (Oxi, H)], and [(Oxi, E), (Oxi, H)] of ellipses are jointed [Fig. 2(A)]. Because the disjointed-ellipse pairs of 2D plots were useful for distinguishing the redox state and secondary structure, we summarized them in Table II. Three disjointed-ellipse pairs of [(Red, H), (Oxi, H)], [(Red, H), (Oxi, E)], and [(Red, E), (Oxi, E)] were found after analyzing 2D C^{α}/C^{β}, C'/C^{β}, H^{α}/C^{β}, and N^{H}/C^{β} plots, which indicated that they are useful for distinguishing the redox state and secondary structure of cysteine residues using paired CSs of C^{α}, C', H^{α}, and N^{H} versus C^{β}.

TABLE II. Summary of the Disjointed-Ellipses Pairs of 2D C^{α}/C^{β}, C'/C^{β}, H^{α}/C^{β}, and N^{H}/C^{β} Plots

	(Red, H)	(Red, E)	(Oxi, H)	(Oxi, E)
(Red, H)	—			
(Red, E)		—		
(Oxi, H)	C^{α}/C^{β}, C'/C^{β} H^{α}/C^{β}, N^{H}/C^{β}	C'/C^{β} H^{α}/C^{β}	—	
(Oxi, E)	C^{α}/C^{β}, C'/C^{β} H^{α}/C^{β}, N^{H}/C^{β}	C^{α}/C^{β}, C'/C^{β} H^{α}/C^{β}, N^{H}/C^{β}		—

Specifically, 2D plots of C'/C^{β} and H^{α}/C^{β} were unique for distinguishing the state of [(Red, E), (Oxi, H)]. In contrast, none of 2D plots can be used to distinguish the states of [(Red, H) and (Red, E)] and [(Oxi, H) and (Oxi, E)].

Score Matrix for Prediction of Secondary Structure and Redox State

Depending on 2D C^{α}/C^{β}, C'/C^{β}, H^{α}/C^{β}, and N^{H}/C^{β} cluster plots, the four ellipses of (Red, H), (Red, E), (Oxi, H), and (Oxi, E) were joined to or disjointed from each other. Thus, different plots can provide helpful information to distinguish the redox state and secondary structure of cysteine residues. We therefore hypothesized that the probability scores calculated from the four 2D ellipses of C^{α}/C^{β}, C'/C^{β}, H^{α}/C^{β}, and N^{H}/C^{β} clusters can be used to predict the redox state and secondary structure of cysteine residues. The score matrix, which was produced from the probability scores of C^{α}/C^{β}, C'/C^{β}, H^{α}/C^{β}, and N^{H}/C^{β} clusters, is shown in Table SI. Three situations were distinguished by applying 2D C^{α}/C^{β}, C'/C^{β}, H^{α}/C^{β}, and N^{H}/C^{β} clusters to estimate $\hat{P}r(\zeta|\chi_1\chi_2)$: (1) (χ_1, χ_2) fell outside all elliptical areas; (2) (χ_1, χ_2) fell onto one and only one elliptical area; and (3) (χ_1, χ_2) fell onto an intersection area of two ellipses. Because the probability within 1 standard deviation of the mean is 0.683, we used $\hat{P}r(\zeta|\chi_1\chi_2) \geq 0.683$ as minimal threshold. In all three situations, f_ζ was the relative frequency of the ζ state in the area where (χ_1, χ_2) resided. Thus, $\hat{P}r(\zeta|\chi_1\chi_2) = f_\zeta$. Three rules were proposed to predict the redox state and secondary structure of cysteines using the total score $\tau(\zeta)$ and $\hat{P}r^{max}(\zeta)$, the maximum of the four $\hat{P}r(\zeta|\chi_1\chi_2)$ from the score matrix (Table SI).

Rule 1. Add the probabilities of each column in Table II to obtain the $\tau(\zeta)$ for every secondary structure and redox state. The predicted state ζ if and only if $\tau(\zeta) \geq 2.732$; otherwise apply Rule 2.

Rule 2. The predicted state ζ if and only if $\hat{P}r^{max}(\zeta)$ is the only value greater than 0.683 in $\hat{P}r^{max}$; otherwise apply Rule 3.

Rule 3. Let $\hat{P}r^{max}(\zeta) \geq 0.683$ for every member of ζ in set S_ζ^{max}, where S_ζ^{max} is a subset of {(Red, H), (Red, E), (Oxi, H), (Oxi, E)} and contain no fewer than 2 elements. Then, the predicted state ζ is in any state of S_ζ^{max} and rules out the possibility of any state that is not in S_ζ^{max}; otherwise, no prediction is made.

Application of the Prediction Rules Using the Score Matrix

To demonstrate the feasibility of using these three rules of the score matrix to predict the redox state and secondary structure of cysteines using their CSs, we collected new entries between October 27, 2004, and April 9, 2005. New entries with the related PDB entries were selected. Twelve new entries contained 21 cysteine CSs, which were used to test the three prediction rules using the score matrix. Table III shows the results of applying the three score-matrix prediction rules. Rule 1 was applied to 18 of 21 cysteine CSs. For example, the Cys40 residue of BMRB6299 is an (Oxi, E) cysteine.[24] The τ(Oxi, E) of Cys40-BMRB6299 was 3.889, which was greater than 2.732. Thus, an (Oxi, E) cysteine was predicted for the Cys40 residue of BMRB6299 by applying Rule 1. Rule 2 was applied to 1 of 21 cysteine CSs. For example, the Cys11 residue of BMRB6336 is a (Red, E) cysteine.[25] Because none of the $\tau(\zeta)$ of BMRB6336 was greater than 2.732 and the (Red, E) cysteine was the only $\hat{P}r^{max} \geq 0.683$ (Table III), a (Red, E) cysteine was predicted for the Cys11 residue of BMRB6336 using Rule 2. Rule 3 was applied to 2 of 21 cysteine CSs. The Cys29 residue of BMRB5933 is a (Red, H) cysteine.[26] None of the $\tau(\zeta)$ of BMRB5933 was greater than 2.732, and the $\hat{P}r^{max}$ values of both (Red, H) and (Red, E) were greater than 0.683 (Table III). Therefore, Rule 3 predicted that it was either a (Red, H) or (Red, E) cysteine. This is because all 2D plots of CSs show a significant overlap between the two ellipses corresponding to the (Red, H) and (Red, E) states (Fig. 2). In other words, no combination of CS clearly distinguishes between the helical and extended state of a reduced cysteine. Thus, the prediction rules accurately predicted its redox state but not secondary structure. In summary, 100% accuracy was achieved by predicting the redox states of all cysteine CSs from new entries using

TABLE III. Score Tables for BMRB6299-40, BMRB6336-11, and BMRB5933-29[†]

Shifts Plot	(Red, H)	(Red, E)	(Oxi, H)	(Oxi, E)
BMRB6299-40 (Oxi, E)				
C^{α}/C^{β}	0	0	4/68[a]	64/68[b]
C'/C^{β}	0	0	1/34	33/34[b]
H^{α}/C^{β}	0	0	0	1[a]
N^{H}/C^{β}	0	0	1/44	43/44[b]
$\hat{P}r^{max}$	0	0	0.059	**1**[c]
Total $\tau(\zeta)$	0	0	0.111	**3.889**[d]
BMRB6336-11 (Red, E)				
C^{α}/C^{β}	0	21/27	6/27	0
C'/C^{β}	0	0	0	0
H^{α}/C^{β}	0	73/75[b]	1/75	1/75
N^{H}/C^{β}	0	4/6	2/6	0
$\hat{P}r^{max}$	0	**0.973**[c]	0.333	0.013
Total $\tau(\zeta)$	0	2.418	0.569	0.013
BMRB5933-29 (Red, H)				
C^{α}/C^{β}	1/59	57/59[b]	1/59	0
C'/C^{β}	5/11	6/11[b]	0	0
H^{α}/C^{β}	10/14	4/14	0	0
N^{H}/C^{β}	72/73[b]	1/73	0	0
$\hat{P}r^{max}$	**0.986**[c]	**0.966**[c]	0.017	0
Total $\tau(\zeta)$	2.172	1.811	0.017	0

[†]BMRB6299-40, BMRB6336-11, and BMRB5933-29 represent the Cys40 residue of BMRB6299, the Cys11 residue of BMRB6336, and the Cys29 residue of BMRB5933. Boldface values represent the predicted state from the score matrix.

[a]The value of 4/68 represented the probabilities of the (Oxi, H) state for observed C^{α} and C^{β} CS of BMRB6299-40. Thus, the BMRB6299-40 with C^{α} (54.97ppm) and C^{β} (48.01ppm) is in the region that 4 of 68 CSs have a (Oxi, H) state.

[b]$\hat{P}r(\zeta|\chi_1,\chi_2) \geq 0.683$.

[c]$\hat{P}r^{max}(\zeta) \geq 0.683$.

[d]$\tau(\zeta) \geq 2.732$.

TABLE IV. Summary of the Predicted Secondary Structures Using the Score Matrix and CSI[†]

Protein	PDB Code	BMRB code-residue	Score matrix[a]	CSI[b]	Structure[c]
The *Saccharomyces cerevisiae* SAM (Sterile Alpha Motif) domain	IOW5	5933-29	Red, ND[d]	H	Red, H
The E2 DNA-Binding domain from human papillomavirus Type-16	IBY9	5952-16	Red, H	H	Red, H
The BRCT-c domain from human BRCA1	IOQA	6114-14	Red, E	E	Red, E
		6114-15	Red, E	**Coil**	Red, E
The carnobacteriocin B2 immunity protein	ITDP	6211-34	Red, H	H	Red, H
The UBCH5B-CNOT4 complex	IUR6	6277-107	Red, H	H	Red, H
The gene product from *Arabidopsis thaliana* AT1G77540	IXMT	6338-58	Red, H	H	Red, H
The *Arabidopsis* ortholog of the C-terminal domain of human thioredoxin-like protein	IXOY	6341-27	Red, E	E	Red, E
The Partially Disordered Protein At2g23090 from *A. thaliana*	IWVK	6432-48	Red, ND[d]	**Coil**	Red, E
		6432-64	Red, H	H	Red, H
The complex of porcine pancreatic elastase in complex with cadmium	IUV0	6231-13	Oxi, E	E	Oxi, E
		6231-30	Oxi, H	H	Oxi, H
		6231-108	Oxi, E	E	Oxi, E
		6231-133	Oxi, E	E	Oxi, E
The chordin-like cysteine-rich repeat (VWC module) from collagen IIA	IU5M	6299-10	Oxi, E	**H**	Oxi, E
		6299-31	Oxi-E	E	Oxi, E
		6299-33	Oxi-E	E	Oxi, E
		6299-40	Oxi-E	E	Oxi, E
The murine T cell receptor (TCR) Valpha2.6Jalpha38 (TCRAV2S6J38) domain	IB88	6406-23	Oxi-E	E	Oxi, E
		6406-90	Oxi-E	E	Oxi, E
The human nudix enzyme AP4A hydrolase	1XSA	6336-11	Red, E	E	Red, E

[†]The predicted secondary structure was determined by the CSI program which is available at the URL http://www.bionmr.ualberta.ca/bds/software/csi/index.html.[16] The structures in boldface represent mistakenly predicted results.
[a]The redox state and secondary structure were predicted using our method.
[b]The secondary structure was predicted using the CSI method.
[c]The secondary structure was obtained from X-ray or NMR structure.
[d]ND, not determined.

Rules 1 and 2, and no faulty predictions were made using the prediction rules of the score matrix (Table IV).

Comparison of the Predicted Secondary Structures Using the Score Matrix and CSI

We also compared the accuracy of predicting the secondary structures of 12 new BMRB entries using the score matrix and CSI (Table IV).[16] The score matrix method accurately predicted the secondary structures for 19 of 21cysteine CSs, and 2 of 21 cysteine CSs cannot be determined using this method (Table IV). Therefore, this method has a predictive accuracy of ~90% in this study. Similar to the results of the score matrix, the CSI method accurately predicted the secondary structures for 18 of 21 cysteine CSs. It produced three mistaken predictions: two CSs with a coil structure were predicted to have a β structure, and one CS with α helix structure was predicted to have a β structure. Overall, the CSI method had a predictive accuracy of ~86%. The CSI method did correctly predict the secondary structure of one CS, the Cys29 residue of BMRB5933, which cannot be predicted using the score matrix method.

DISCUSSION

Our statistical analysis of 3401 $^{1}H^{\alpha}$, 3605 $^{1}H^{N}$, 1512 $^{13}C^{\alpha}$, 1324 $^{13}C^{\beta}$, 983 $^{13}C'$, and 3605 $^{15}N^{H}$ of cysteine CSs collected from the BMRB database was consistent with a previous report that cysteine C^{α} and C^{β} CSs are extremely sensitive to their secondary structure and redox state, respectively.[20] 1D statistical analysis showed that cysteine $^{1}H^{\alpha}$, $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C'$, and $^{15}N^{H}$ CSs reflected the secondary structure, which is consistent with previous reports.[9–19] 2D C^{α}/C^{β} cluster analysis of cysteine CSs has already been found helpful in distinguishing both the redox state and secondary structure of cysteine residues.[20] Our statistical analysis of cysteine CSs revealed that 2D clusters of *x*-coordinate coded redox states and *y*-coordinate coded secondary structures showed distinct clusters. Based on our analysis, 2D C'/C^{β}, H^{N}/C^{β}, and H^{α}/C^{β} clusters were also found to be helpful in distinguishing both redox states and secondary structures of cysteine residues. Specifically, the 2D C'/C^{β} and H^{α}/C^{β} clusters were found to be helpful in distinguishing the overlapping (Red, E) and (Oxi, H) states of the cysteine amino acid in 2D C^{α}/C^{β} and H^{N}/C^{β} clusters. It was found that 1D histograms of CSs did not follow an exact Gaussian distribution, and 2D clusters have the outliers (Figs. 1 and 2). By analyzing the structures of the outliers, we found that ~85% cysteine residues are located in N- and C-termini of protein secondary structures. Therefore, the secondary structure assignment of boundary residues in proteins may be responsible for most outlines.

The use of protein CS values for the determination of 3D protein structure has increased in recent years because of the large databases of protein structures with assigned CS data.[7,8,27] The BMRB database has allowed the investigation of the quantitative relationship between CS values obtained by solution NMR spectroscopy and 3D structures of proteins.[28,29] For example, the amino acid CSs obtained from NMR measurement were routinely used to assign the secondary structure in proteins.[6] The CSs of 20 amino acids are in the following order: $H^N < N^H < C^\beta < H^\alpha < C' < C^\alpha$ to distinguish an α-helix from a random coil; and $H^N \approx C^\alpha \approx C' \approx N^H < C^\beta < H^\alpha$ to distinguish a β-strand from a random coil.[19] Similar to this result, our data showed that C^α, C', and H^α CSs of cysteine residue were sensitive to their secondary structure. In contrast to the other 19 amino acids, the cysteine N^H CS was more sensitive to its secondary structure.

It is shown that by combining both ^1H and ^{13}C chemical-shift indices to produce an estimate of secondary structure, it is possible to achieve a predictive accuracy greater than 85%.[9–19] Similar to the predicted results using the CSI, our score matrix method gave an estimated accuracy of 90% in predicting protein secondary structures. Because the cluster boundaries of all ellipses in 2D plots contained 90% of the CSs, it is reasonable to suggest that this method may provide a 90% level of confidence of the prediction. No faulty predictions of protein secondary structures were made using the prediction rules of the score matrix. In contrast, we found that predictions of protein secondary structures for three of 21 cysteine CSs using CSI method were mistakenly predicted. Compared to the results predicted using the CSI method, our results suggested that the score matrix using 2D cluster analysis of cysteines $^1H^\alpha$, $^1H^N$, $^{13}C^\alpha$, and $^{13}C'$ versus $^{13}C^\beta$ CSs may produce more accurate prediction of their secondary structures. Thus, combinational 2D analysis of ^1H, ^{13}C, and $^{15}N^H$ CSs may provide more accurate prediction; however, more CS data and data analysis are required to verify this.

It has been shown that redox state of cysteines in proteins except those that are paramagnetic can be simply determined using the rules provided by Sharma and Rajarathnam. On the basis of their three ground rules, the prediction was made by analyzing the locations of C^α and C^β CSs in Figures 1(A,B) and 2(B). Our method used a score matrix by analyzing cysteines $^1H^\alpha$, $^1H^N$, $^{13}C^\alpha$, and $^{13}C'$ versus $^{13}C^\beta$ CSs. The score matrix method has several predictive advantages: it provides additional analysis of 2D C'/C^β, H^N/C^β, and H^α/C^β clusters so that it can make the prediction without C^α CS; it predicts not only the redox state but also the secondary structure, and the 2D C'/C^β and H^α/C^β clusters are helpful in distinguishing the overlapping (Red, E) and (Oxi, H) states in 2D C^α/C^β. It does, however, require more CS data and data calculation to make the prediction.

In conclusion, we report 2D cluster analyses of cysteines $^1H^\alpha$, $^1H^N$, $^{13}C^\alpha$, and $^{13}C'$ versus $^{13}C^\beta$ CSs. On the basis of the analysis of these 2D clusters, we derived three rules using a score matrix to predict the redox state and secondary structure of cysteine residues using their CSs. Applying the three rules, we were able to predict the redox state and secondary structure of 21 cysteine CSs with greater than 90% accuracy. This analysis suggests that the redox state and secondary structure of cysteine residues in peptides and proteins can be obtained from their CSs, without recourse to nuclear Overhauser effect measurements.

ACKNOWLEDGMENTS

We are indebted to Dr. Wenya Huang for valuable comments and to Dr. Yunjun Wang for providing the PSSI software.

REFERENCES

1. Gilbert HF. Thiol-disulfide exchange of divalent sulfur. In: Comprehensive Biological Catalysis, vol. 1, Sinnott M, editor. San Diego, CA: Academic Press; 1998. p 609–625.
2. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
3. Chen CY, Luo SC, Kuo CF, Lin YS, Wu JJ, Lin MT, Liu CC, Jeng WY, Chuang WJ. Maturation processing and characterization of streptopain. J Biol Chem 2003;278:17336–17343.
4. Ferre F, Clote P. Disulfide connectivity prediction using secondary structure information and diresidue frequencies. Bioinformatics 2005;21:2336–2346.
5. O'Connor BD, Yeates TO. GDAP: a web tool for genome-wide protein disulfide bond prediction. Nucleic Acids Res 2004;32 (Web Server issue):W360–W364.
6. Wishart DS, Nip AM. Protein chemical shift analysis: a practical guide. Biochem Cell Biol 1998;76:1–10.
7. Bowers PM, Strauss CE, Baker D. De novo protein structure determination using sparse NMR data. J Biomol NMR 2000;18: 311–318.
8. Hung LH, Samudrala R. PROTINFO: secondary and tertiary protein structure prediction. Nucleic Acids Res 2003;31:3296–3299.
9. Le H, Oldfield E. Correlation between ^{15}N NMR chemical shifts in proteins and secondary structure. J Biomol NMR 1994;4:341–348.
10. Nakamura A, Jardetzky O. Systematic analysis of chemical shifts in the nuclear magnetic resonance spectra of peptide chains. II. Oligoglycines. Biochemistry 1968;7:1226–1230.
11. Pastore A, Saudek V. The relationship between chemical shifts and secondary structure in proteins. J Magn Reson 1990;90:165–176.
12. Schwarzinger S, Kroon GJA, Foss TR, Wright PE, Dyson HJ. Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView. J Biomol NMR 2000;18:43–48.
13. Spera S, Bax A. Empirical correlation between protein backbone conformation and Cα and Cβ ^{13}C nuclear magnetic resonance chemical shifts. J Am Chem Soc 1991;113:5490–5492.
14. Szilagyi L, Jardetzky O. α-Proton chemical-shifts and secondary structure in proteins. J Magn Reson 1989;83:441–449.
15. Williamson MP, Asakura T, Nakamura E, Demura M. A method for the calculation of protein α-CH chemical-shifts. J Biomol NMR 1992;2:93–98.
16. Wishart DS, Sykes BD. The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. J Biomol NMR 1994;4:171–180.
17. Wishart DS, Sykes BD, Richards FM. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J Mol Biol 1991;222:311–333.
18. Wishart DS, Sykes BD, Richards FM. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 1992;31:1647–1651.
19. Wang Y, Jardetzky O. Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 2002;11:852–861.
20. Sharma D, Rajarathnam K. ^{13}C NMR chemical shifts can predict disulfide bond formation. J Biomol NMR 2000;18:165–171.
21. Zhang H, Neal S, Wishart DS. RefDB: a database of uniformly

referenced protein chemical shifts. J Biomol NMR 2003;25:173–195.
22. Wang Y, Wishart DS. A simple method to adjust inconsistently referenced ^{13}C and ^{15}N chemical shift assignments of proteins. J Biomol NMR 2005;31:143–148.
23. Kabsch W, Sander CA. dictionary of protein secondary structure. Biopolymers 1983;22:2577–2637.
24. O'Leary JM, Hamilton JM, Deane CM, Valeyev NV, Sandell LJ, Downing AK. Solution structure and dynamics of a prototypical chordin-like cysteine-rich repeat (von Willebrand Factor type C module) from collagen IIA. J Biol Chem 2004;279:53857–53866.
25. Gaiser OJ, Oschkinat H, Heinemann U, Ball LJ. ^{1}H, ^{13}C and ^{15}N resonance assignments of the C-terminal BRCT domain from human BRCA1. J Biomol NMR 2004;30:221–222.
26. Bhattacharjya S, Xu P, Gingras R, Shaykhutdinov R, Wu C, Whiteway M, Ni F. Solution structure of the dimeric SAM domain of MAPKKK Ste11 and its interactions with the adaptor protein Ste50 from the budding yeast: implications for Ste11 activation and signal transmission through the Ste50-Ste11 complex. J Mol Biol 2004;344:1071–1087.
27. Nilges M. Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. J Mol Biol 1995;245:645–660.
28. Seavey BR, Farr EA, Westler WM, Markley JL. A relational database for sequence-specific protein NMR data. J Biomol NMR 1991;1:217–236.
29. Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Markley JL, Ulrich EL. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. J Biomol NMR 2003;26:139–146.