Journal of Biomolecular NMR (2006) 34: 259–269
DOI 10.1007/s10858-006-0027-5

Article

Identification of zinc-ligated cysteine residues based on $^{13}C\alpha$ and $^{13}C\beta$ chemical shift data

Gregory J. Kornhaber[a,b,c], David Snyder[a,c], Hunter N.B. Moseley[a,c] & Gaetano T. Montelione[a,b,c,*]

[a]*Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854, USA;* [b]*Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA;* [c]*Northeast Structural Genomics Consortium, Piscataway, NJ 08854, USA*

Received 14 November 2005; Accepted 27 February 2006

Key words: chemical shift distribution analysis, logistic regression analysis, Zn-ligated cysteine

Abstract

Although a significant number of proteins include bound metals as part of their structure, the identification of amino acid residues coordinated to non-paramagnetic metals by NMR remains a challenge. Metal ligands can stabilize the native structure and/or play critical catalytic roles in the underlying biochemistry. An atom's chemical shift is exquisitely sensitive to its electronic environment. Chemical shift data can provide valuable insights into structural features, including metal ligation. In this study, we demonstrate that overlapped $^{13}C\beta$ chemical shift distributions of Zn-ligated and non-metal-ligated cysteine residues are largely resolved by the inclusion of the corresponding $^{13}C\alpha$ chemical shift information, together with secondary structural information. We demonstrate this with a bivariate distribution plot, and statistically with a multivariate analysis of variance (MANOVA) and hierarchical logistic regression analysis. Using 287 $^{13}C\alpha/^{13}C\beta$ shift pairs from 79 proteins with known three-dimensional structures, including 86 $^{13}C\alpha$ and $^{13}C\beta$ shifts for 43 Zn-ligated cysteine residues, along with corresponding oxidation state and secondary structure information, we have built a logistic regression model that distinguishes between oxidized cystines, reduced (non-metal ligated) cysteines, and Zn-ligated cysteines. Classifying cysteines/cystines with a statisical model incorporating all three phenomena resulted in a predictor of Zn ligation with a recall, precision and *F*-measure of 83.7%, and an accuracy of 95.1%. This model was applied in the analysis of *Bacillus subtilis* IscU, a protein involved in iron–sulfur cluster assembly. The model predicts that all three cysteines of IscU are metal ligands. We confirmed these results by (i) examining the effect of metal chelation on the NMR spectrum of IscU, and (ii) inductively coupled plasma mass spectrometry analysis. To gain further insight into the frequency of occurrence of non-cysteine Zn ligands, we analyzed the Protein Data Bank and found that 78% of the Zn ligands are histidine and cysteine (with nearly identical frequencies), and 18% are acidic residues aspartate and glutamate.

Introduction

A large number of proteins include metal atoms, such as zinc (Zn), as part of their structures. For example, at least 3% of the protein sequences inferred from the *Caenorhabditis elegans* and human genomes are predicted to contain Zn binding motifs (Clarke and Berg, 1998; Lu et al., 2003). In fact, Zn-fingers are the most commonly observed structural motifs in transcription factors, and

*To whom correspondence should be addressed. E-mail: guy@cabm.rutgers.edu

among the most common protein structural motifs in the human proteome (Venter et al., 2001; Clamp et al., 2003; Bateman et al., 2004). Zinc's electronic configuration allows for a flexible coordinate geometry and enhances the nucleophilicity of thiol substrates (Zhou et al., 1999). It can participate in enzymatic reactions and regulatory functions (Berg and Shi, 1996; Lipscomb and Strater, 1996; Hernick and Fierke, 2005). Since Zn is commonly coordinated by four or more ligands, it can also bridge together and stabilize disparate regions of protein structure (Klug and Rhodes, 1987; Krishna et al., 2003; Kwon et al., 2003).

^{64}Zn, ^{66}Zn and ^{68}Zn are the three most naturally abundant isotopes of Zn (Coplen et al., 2002). They have an even number of protons and an even number of neutrons and are therefore non-NMR active (spin $I = 0$). As a consequence, Zn ligands may be undetected in the course of determining a protein's structure using solution NMR methods. Histidine and cysteine are the two most common Zn ligands. Once the metal ligation status of a protein has been established, the metal-coordinating residues can sometimes be identified with techniques involving metal displacement with ^{113}Cd (spin $I = 1/2$) followed by NMR experiments that measure ^{113}Cd–^{1}H scalar couplings (Neuhaus et al., 1984; Vasak et al., 1987). Nitrogen-15 NMR experiments that resolve the histidine tautomer present in some Zn binding sites can also be used to deduce the nitrogen coordinating a Zn (Pelton et al., 1993; Ramelot et al., 2004). In addition, there are several examples of histidine ^{13}Cδ2 resonances undergoing characteristic downfield shifts upon Zn binding (Ramelot et al., 2004), but this phenomenon has not been systematically analyzed.

Sharma and Rajarathnam (2000) have reported that the ^{13}Cα and ^{13}Cβ(Cα/Cβ) chemical shifts of metal-ligated cysteines of metalloproteins are within the chemical shift distribution of reduced non-metal-ligated cysteine Cα/Cβ resonances. While providing some indication that the bivariant Cα/Cβ chemical shift plots can provide some discrimination of reduced and metal-ligated cysteine residues, this study was based on a limited database of metalloprotein chemical shift data, and did not identify the metalloprotein nor the associated metal. Two recent advances warrant an updated analysis of metal ligated cysteine Cα/Cβ shift distributions. First, in 2003, David Wishart's group published SHIFTCOR and the Reference-Corrected Database (RefDB) of chemical shifts (Zhang et al., 2003), providing refined chemical shift data for proteins using self-consistent referencing procedures. Secondly, the number of metal complexed protein NMR structures with both Protein Data Bank (PDB) and BioMagResBank (BMRB) depositions, have significantly increased since 2000; over the last five years, Zn-complexed NMR structures with BMRB chemical shift depositions have increased by more than three-fold.

In this study we demonstrate that cysteine residues with coordinated metal, in this case Zn, can be reliably identified given (i) a well-referenced chemical shift data base, and (ii) the secondary-structure of the cysteine residue. Once identified, the presence of ligated metal can be validated by examining the effect of metal chelation on NMR spectra of the subject protein, and confirmed by quantitative inductively coupled plasma mass spectrometry. We demonstrate the utility of this approach for identifying and validating Zn-ligated cysteines on the iron–sulfur complex assembly protein U (IscU), a metalloprotein from *Bacillus subtilis* involved in iron–sulfur cluster (Fe–S) biosynthesis.

Methods

Non-metal-ligated reduced cysteine and oxidized cystine Cα/β chemical shift datamining

The RefDB database (http://redpoll.pharmacy.ualberta.ca/RefDB/ref.data) used in this study was downloaded on December 6, 2004. Two Perl scripts facilitated collation and analysis of cysteine/cystine Cα/Cβ shifts. The first Perl script searches each RefDB entry for a metal ligand by examining values reported for the following definitions: "_PDB_code", "_Abbreviation_common", and "metal ligand". It then creates a list of RefDB entries to exclude from further analysis entries for which "Ca", "Zn", "Cu", "Mg", "Cd", or "Fe" are reported in any of the above definitions. The second Perl script gathers information for each entry, including: BMRB ID, protein sequence, cysteine/cystine Cα/Cβ shifts, PDB files and "_Mol_thiol_state" definitions. Entries with their "_Mol_thiol_state" defined as "all free", "all disulfide bound", or "fully oxidized" are then classified as reduced or

oxidized. The script then retrieves the PDB entry associated with each RefDB entry, extracts the protein sequence and secondary structure annotation, compares the RefDB sequence to the PDB sequence, and if verified to be the same in numbering and residue, assigns the secondary structure to each residue as defined in the PDB structure file. Numbering inconsistencies were resolved manually, allowing proper assignment of secondary structure information provided in the PDB structure file. If the secondary structure was not defined in the header of the PDB structure file, it was determined using the DSSP algorithm (Kabsch and Sander, 1983) functionality of the program MOLMOL (Koradi et al., 1996). These ^{13}C chemical shifts were then binned in 1.0 ppm increments, and statistics compiled as a final automated step of this second Perl script. Histograms based on these populated shift classes were then constructed with Kaleidagraph v.3.5 and Microsoft Excel.

Zn-ligated cysteine Cα/Cβ shift datamining

All NMR structures containing both (i) BMRB cross references and (ii) Zn HET atom annotations were downloaded from the PDB on December 8, 2004. These protein coordinate sets were then screened for cysteine Sγ atoms within 3 Å of a Zn atom, and the corresponding Cα/Cβ chemical shifts extracted from the BMRB. Numbering inconsistencies between BMRB and PDB data were resolved manually. The Zn Cα/Cβ shifts from BMRB data sets of Zn-containing proteins were referenced with a procedure analogous to that used in creating the RefDB chemical shift database, using the SHIFTCOR 1.0 procedure (http://redpoll.pharmacy.ualberta.ca/shiftcor/) (Zhang et al., 2003), so that these values could be compared with chemical shifts of non-metal-coordinated Cys residues reported for proteins in the RefDB data set. Cα and Cβ chemical shift mean and standard deviations were calculated, and the data binned in 1.0 ppm increments.

Zn ligand datamining

The entire PDB was downloaded on May 12, 2005 and analyzed for (i) all atoms within 3 Å of a Zn atom, (ii) the corresponding amino acid residue, (iii) the secondary structure of the corresponding residue specified in the PDB file, (iv) the distance to the Zn atom, and (v) the name of the PDB structure file. This information-rich output enabled us to determine the frequencies of the various residue ligands, by tallying the residues containing atoms within 3 Å of a Zn atom. Cysteine ligands were then further classified based on their corresponding secondary structure.

Linear modeling and logistic regression analysis

The multivariate analysis of variance (MANOVA) functionality within the General Linear Modeling (PROC GLM) procedure of SAS 8.2 was used to assess the ability of Cα and Cβ chemical shifts to distinguish reduced cysteine from Zn-ligated cysteine residues, as well as to distinguish between oxidized and non-oxidized cysteines. We employed a logistic regression model (PROC LOGISTIC in SAS 8.2) with our Cα/Cβ chemical shift pair data (summarized in Supplemental Table S1) and corresponding secondary structure data as input. To more faithfully reflect the nature of the distinction between oxidized, reduced and Zn-ligated cysteines (as the Zn-ligated cysteines are a special class of reduced cysteines), as well as to ensure a numerically robust result, two logits were modeled sequentially: (i) the log odds ratio of a cysteine being oxidized as opposed to not oxidized, and (ii) the log odds ratio of a non-oxidized cysteine coordinating Zn as opposed to being in an unligated, reduced state. The MANOVA analysis, and fit to the hierarchical logistic regression model, included all data obtained from the data mining procedure described above, with the exception of data with outlying chemical shifts identified by Chauvenet's criterion (Taylor, 1997). We also excluded data from five cysteines in proteins with PDB IDs 1G03 and 1GH9, in which Zn binding was suspected but could not be confirmed with confidence.

Cloning, expression and purification of IscU

DNA coding the full length iron sulfur cluster metallochaperone protein IscU from *Bacillus subtilis* (Swissprot locus: NIFU_BACSU) was cloned into a pET21d (Novagen) derivative possessing a C-terminal hexa-His tag (Acton et al., 2005). The resulting plasmid was transformed into *Escherichia coli* strain BL21 (DE3) pMGK (Novagen) cells, grown up in 1 l of uniformly ^{15}N-enriched

MJ9 minimal media (Jansson et al., 1996), induced, expressed, and purified using standard procedures described elsewhere (Acton et al., 2005). The resulting sample was buffer exchanged into NMR Buffer (50 mM sodium phosphate, 50 mM NaCl, 10 mM DTT, 5% D_2O, 0.02% NaN_3, pH 6.5) at a protein concentration of ~0.7 mM.

Metal analysis

NMR spectra were recorded using 350 μl samples in 5-mm Shigemi susceptibility-matched NMR tubes, at 20 °C, on a 600 MHz Varian Inova spectrometer. 2D ^{1}H–^{15}N HSQC NMR spectra were recorded prior to and after the addition of 10 mM EDTA. Following these NMR measurements, the EDTA-treated sample was buffer exchanged (Amicon Ultra-4 concentrator, Millipore) to remove dissociated metal. These EDTA-treated and untreated samples were then analyzed for Zn metal content by inductively coupled plasma mass spectrometry (ICP-MS) (Rea, 2003) using an Agilent Technologies 4500 ICP-MS system at the Pacific Northwest National Laboratory, Richland, WA.

Results

BMRB and RefDB chemical shift datamining summary

Chemical shift data for non-metal-coordinated cystine/cysteine residues were extracted from RefDB chemical shift lists. Of the 601 protein chemical shift files in RefDB at the time of this analysis, 85 identified as possibly containing bound metals were excluded from this analysis. The resulting 516 protein chemical shift files provided data for 557 Cα or Cβ cysteine/cystine chemical shifts. The number was reduced to 554 shifts of non-metal-ligated cysteine/cystine residues, when one cysteine Cα and two cysteine Cβ shifts were omitted because they exhibit unusual values that failed Chauvenet's criterion.

Following outlier detection (and identification of chemical shifts associated with zinc ligated cysteines, as explained below) the Cα/Cβ shift distributions of Cys residues were plotted to visualize the data at hand (Figures 1 and 2). Analysis of the initial Cα/Cβ shift distributions of non-metal-ligated Cys residues, when compared with Cα/Cβ shifts of documented Zn-ligated Cys residues, allowed identification of an incorrectly annotated protein structure, 3-methyladenine DNA glycosylase (1LMZ) (Drohat et al., 2002), for which chemical shift data for two Cys residues indicated metal ligation that was not evident in the atomic coordinate data. We further recognized that the same lab that deposited 1LMZ subsequently redeposited the protein coordinates in a Zn-ligated form (1NKU) (Kwon et al., 2003), confirming our suspicion that 1LMZ does in fact include two Zn-ligated cysteine residues. As our recognition of zinc binding in 1LMZ/1NKU occurred before finishing the compilation of the training data, we included the data associated with this protein (which chemical shift data and secondary structure results did not change between the 1LMZ structure and the 1NKU structure). This correct identification of 1LMZ as "Zn-binding", which was not initially annotated as a zinc-binding protein, supported our proposal that Cys Cα/Cβ chemical shift values are indeed useful in identifying metal-ligated Cys residues.

Chemical shift data for Zn-ligated cysteines were obtained by cross-referencing Zn-containing protein structures listed in the PDB with BMRB NMR chemical shift files. The resulting cysteine carbon-13 chemical shifts were then refined by the SHIFTCOR procedure, as described in Methods. This analysis provided eight cross-referenced proteins, from six different SCOP families (summarized in Supplementary Table S2) and ranging in size from 51 to 182 residues long, containing 82 Zn-ligated cysteine Cα or Cβ shifts (41 Cα/Cβ pairs). RefDB data for the two cysteines of 3-methyladenine DNA glycosylase (1LMZ), described above, were added to this set, bringing the total of Zn-coordinated cysteine Cα or Cβ shifts to 86 (43 Cα/Cβ pairs) and increasing the SCOP fold families (Murzin et al., 1995) represented to seven. These chemical shift data include 38 Zn-coordinated cysteine residues in loops or turns, and five residues in helices. None of these Zn-coordinated residues occur in β-strands.

This data mining and refinement process thus provided a total of 325 Cα shifts and 311 Cβ shifts, collected from 215 RefDB and eight BMRB files. These include 287 Cα/Cβ shift pairs for 79 proteins for which a 3D structure is available in the

PDB (Supplemental Table S1), along with corresponding secondary structure information. Statistics for these Cα and Cβ chemical shift distributions for oxidized, reduced and Zn-coordinated Cys residues are tabulated in Table 1. The mean of oxidized and reduced Cα/Cβ shifts resemble (approx. ±0.5 ppm) previously reported values (Sharma and Rajarathnam, 2000; Zhang et al., 2003). Cβ shift distributions for oxidized, Zn-coordinated, and non-metal-ligated reduced Cys residues are shown in Figure 1. The Cβ shift distribution for Zn-coordinated cysteines is slightly downfield of the reduced non-metal-ligated distribution, and upfield of the oxidized shift distribution (Figure 1a–c). Consistent with previous results, the Zn-coordinated cysteine Cβ shift is significantly overlapped with the reduced non-metal-ligated distribution (Figure 1d–e). However, as shown in Figure 2, when the corresponding Cα chemical shift data for Cα/Cβ pairs are plot together with these Cβ shift data, the Zn-ligated distribution forms a largely distinct cluster. This analysis (Figure 2) demonstrates that combined analysis of Cα/Cβ pair chemical shift data can distinguish reduced from Zn-ligated cysteine residues.

PDB Zn ligand datamining summary

One thousand six hundred and fifty nine out of 30,763 PDB entries (5.4%) are annotated with Zn "HET" atoms and 7484 residues possess atoms within 3 Å of a Zn atom. Statistics for the distributions across residue types for these potential Zn ligands are summarized in Table 2. Histidine and cysteines are the most commonly found potential Zn ligands, collectively occurring ~78% of the time with nearly equal frequencies. The acidic residues aspartate and glutamate are the second most frequent set, occuring ~ 18% of the time. The remaining residue types of Zn ligands constitute less than 4%. In the set of Zn-containing proteins used as training data in this paper (Supplementary Table S2), Zn was ligated either to cysteine alone or to cysteine and histidine, with the exception of a single case in which the zinc was ligated to both cysteine and aspartic acid.

In Table 3, these Zn-ligated cysteine residues observed in the PDB are segregated into groups based on their secondary structure. This analysis shows that the distribution of secondary structure annotations for Zn-coordinated cysteine residues observed in the complete PDB is similar to that of subset of Zn-coordinated proteins cross referenced in both the PDB and BMRB and used in the datamining study described above (summarized in Table 1); most Zn-ligated cysteines (~72%) occur in non-regular structures, fewer (~22%) in helical structures, and a very small number (~6%) in β-strands. Visual inspection of several representative structures from this latter group revealed that Zn-ligated cysteines in β-strands generally occur at the ends of β-sheets, immediately adjacent to loop or turn regions.

Factors predicting cysteine oxidation/ligation state

MANOVA (Everitt and Dunn, 2001) provides statistical quantification of the results depicted graphically in Figures 1 and 2. The Cα/Cβ chemical shift vectors of oxidized cystines are significantly different from those of non-oxidized cystines (Wilks' Lambda, described, e.g., in Everitt and Dunn, 2001, equals 0.22, $p < 0.0001$). Treated separately, the Cα and Cβ shift distributions of oxidized cystines are still significantly different from non-oxidized Cα and Cβ shift distributions (0.01% level by the appropriate univariate *F*-tests as performed in standard Analysis of Variance, ANOVA – Miller, 1997.). While there is considerable overlap between the Cβ shifts of Zn-ligated cysteines and reduced non-metal-ligated cysteines, there is none-the-less significant difference in Cβ shift values between these two populations (again

Table 1. Statistics for ^{13}Cα and ^{13}Cβ chemical shifts of oxidized, reduced and Zn-coordinated Cys residues

Thiol state	Cα shift (ppm) mean (st. dev.)	No. of Cα shifts	Cβ shift (ppm) mean (st. dev.)	No. of Cβ shifts
Oxidized	55.57 (2.46)	179	41.17 (3.93)	166
Reduced	59.25 (3.06)	103	28.92 (2.11)	102
Zn coord.	59.27 (2.12)	43	30.89 (1.01)	43
Total		325		311

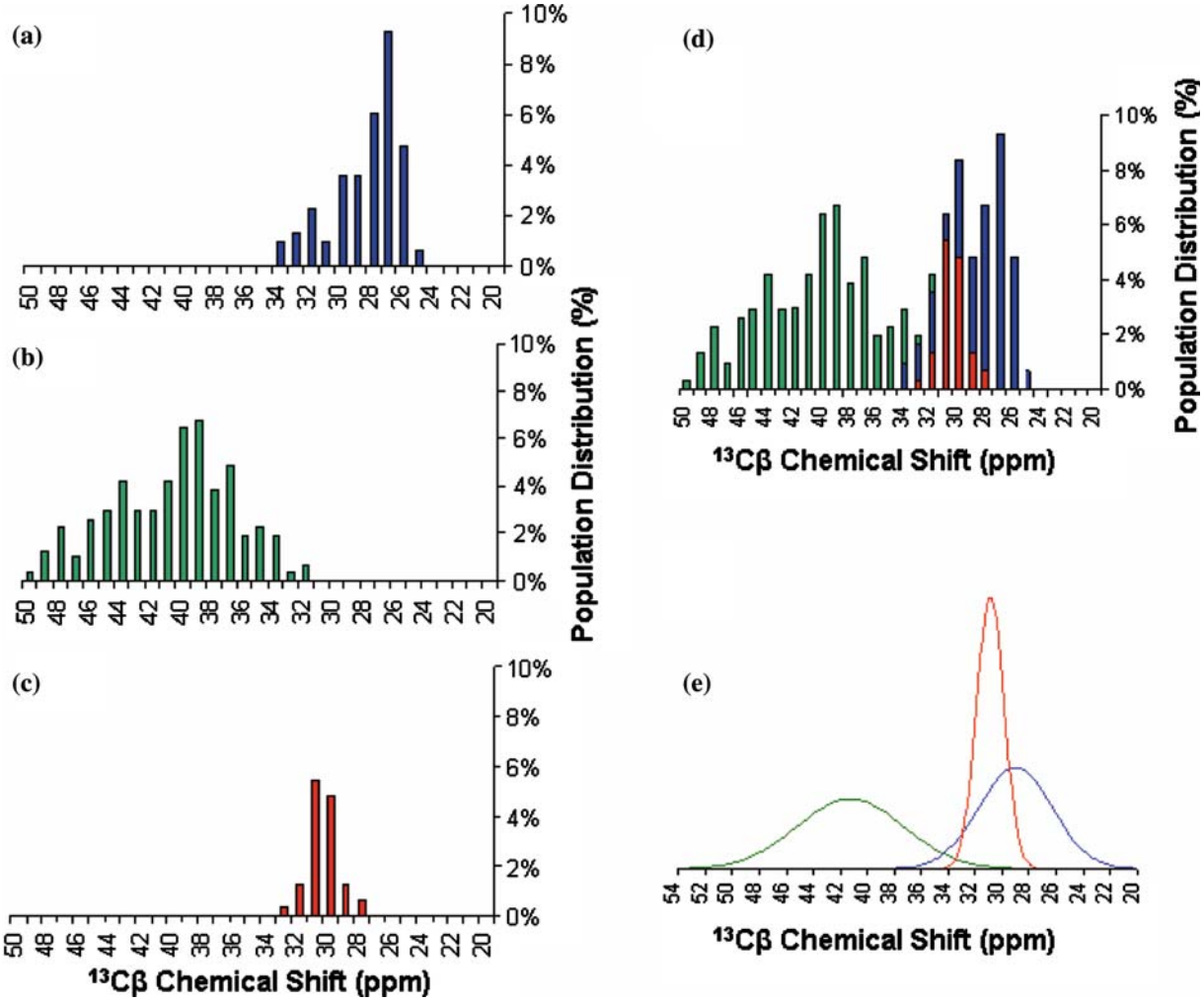

Figure 1. (a–d) Histograms and (e) normal distribution curves for 311 cysteine/cystine Cβ shifts grouped into three categories based on three states of the thiol; (a) 102 from reduced non-metal-ligated cysteines, (b) 166 shifts from oxidized cystines, and (c) 43 from Zn-ligated cysteines.

at the 0.01% level by the appropriate *F*-test). The Cα chemical shift of Zn-ligated cysteines, when considered independently of any other variable, is not significantly different from those of reduced, non-metal ligated cysteines (*F*-test probability of no difference is 0.73). However, Zn-ligated cysteine Cα/Cβ chemical *shift pairs* are indeed significantly downfield shifted relative to the Cα/Cβ chemical shift pairs of reduced non-metal ligated chemical shifts (Wilks' Lambda = 0.66, $p < 0.0001$). In other words, where their Cβ shift distributions overlap (27 ppm $< \delta_{C^{13}\beta} <$ 34 ppm, Figure 1), Zn-ligated and reduced non-metal ligated Cys residues can be distinguished by Cα chemical shift data; in this Cβ chemical shift range, Cα chemical shifts for Zn-ligated Cys residues are usually downfield of Cα chemical shifts for reduced non-metal ligated Cys residues (as shown in Figure 2).

Logistic regression models shed additional light on the dependence of the Cα and Cβ chemical shifts of a cysteine and that cysteine's ligation and oxidation state. For example, while the Cα shifts of Zn-ligated cysteines are not significantly different from those of non-metal-ligated reduced cysteines, logistic regression shows that the likelihood of a cysteine being reduced and unligated as opposed to coordinating Zn depends significantly on the Cα chemical shift of that cysteine ($p = 0.0003$ by Wald's test). Indeed, including Cα chemical shift as an independent variable results in

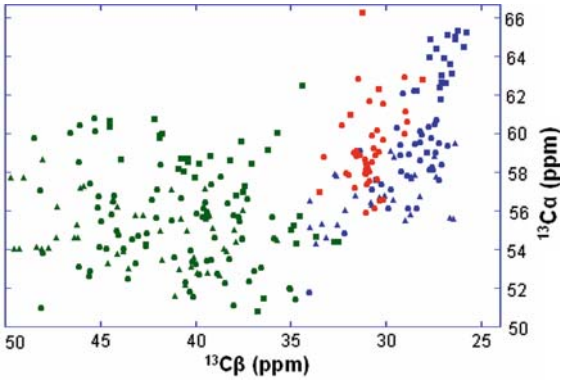

Figure 2. Bivariant plot of 287 Cys Cα/Cβ chemical shift pairs from 79 proteins, consisting of oxidized (green), reduced non-metal-ligated (blue) and Zn-ligated (red) thiols. The corresponding secondary structure context is indicated by boxes (helix), triangles (β-strand), or circles (non-regular structure).

a significantly better model for the prediction of Zn ligation than those models resulting when Cα chemical shift is not included (likelihood ratio χ^2 15.74, $p < 0.0001$). Wald's test and the likelihood ratio χ^2 test show that the odds of Zn ligated cysteine versus reduced non-metal-ligated cysteine are significantly dependent on both Cβ shift and secondary structure.

In distinguishing oxidized from non-oxidized cysteines, logistic regression demonstrates that all three parameters, Cα shift, Cβ shift, and secondary structure are necessary to reliably predict the oxidation state of a given cysteine. We also observe that the linear regression model predicting oxidation state from these variables performs better than the non-linear models tested. Indeed, no unique maximum likelihood main effects model (model considering all independent variables, but without any higher order terms) could be fit as the data were so separated in the space of Cα/Cβ chemical shift values and secondary structure classification that multiple models distinguished oxidized from non-oxidized cysteines equally well. The data provide a clear distinction between oxidized and non-oxidized cystine/cysteines, and only one Zn-ligated cysteine, Cys-161 (in an α-helix) of protein 1IYM, has Cα and Cβ chemical shift values characteristic of an oxidized cysteine.

Table 2. Potential Zn ligand frequencies by residue type

Residue	Percentage	Number
Histidine	42	3140
Cysteine	36	2669
Aspartate	11	859
Glutamate	7	537
Other	4	279
Total		7484

Prediction of cysteine oxidation and Zn binding state from logistic regression models

While the consideration of higher order terms could not improve the ability of logistic regression to predict cysteine oxidation state from Cα shift, Cβ shift and secondary structure data, the consideration of interaction/quadratic terms does improve the ability to predict whether a non-oxidized cysteine bound Zn, or was reduced but uncoordinated by a metal ion. Of all the possible second order terms, only the square of the Cβ chemical shift proved to allow for a model which predicted the binding state of non-oxidized cysteines significantly better than did the model considering only linear terms as predictors. This logistic regression model, given by the equation

$$\begin{aligned}\Lambda_1 &= \log(\text{p\_reduced}/\text{p\_zinc\_bound})\\ &= 569.8 - 0.8018 \times (\text{C}\alpha \text{ chemical shift})\\ &\quad - 32.2621 \times (\text{C}\beta \text{ chemical shift}) + 0.4995\\ &\quad \times (\text{C}\beta \text{ chemical shift})^2 + \{0, \text{ if } \beta\text{-sheet};\\ &\quad - 5.0799, \text{ if } \alpha\text{-helix}; -5.2055, \text{ otherwise}\},\end{aligned}$$

was able to recognize, among the non-oxidized cysteines in the training data set, Zn-ligated as distinguished from reduced non-metal-ligated cysteines with an accuracy of 95.0%, a recall rate of 86.0% and a precision of 84.1%.

Combining two logistic regression models, one modeling the log odds ratio of a cysteine being oxidized *versus* not oxidized –

$$\begin{aligned}\Lambda_2 &= \log(\text{p\_oxidized}/\text{p\_non- oxidized})\\ &= -271.2 - 2.623 \times (\text{C}\alpha \text{ chemical shift})\\ &\quad + 12.1211 \times (\text{C}\beta \text{ chemical shift})\\ &\quad + \{0, \text{ if } \beta\text{-sheet}; 20.0356, \text{ if } \alpha\text{-helix};\\ &\quad - 10.7391, \text{ otherwise}\}\end{aligned}$$

– and the other, whose optimally parameterized formula is given above, modeling the log odds ratio of a cysteine being Zn-ligated versus reduced non-metal ligated, into a single hierarchical

Table 3. Cysteine Zn ligand frequencies segregated by secondary structure

Secondary structure	Percentage	Number
Helix	22	590
Sheet	6	168
Other	72	1911
Total		2669

logistic regression model allows for the prediction of whether a given cysteine is (i) oxidized, (ii) coordinating Zn or (iii) reduced but not bound to a metal ion. Considering these possibilities as the only three, and the requirement that probabilities of mutually exclusive and exhaustive states sum to unity, the two logistic regression models provide a system of three linear equations in three unknowns which can be solved to provide expressions for the probabilities that a given Cys residue is oxidized (p_oxidized), Zn-ligated (p_zinc_ligated), or reduced non-metal-ligated (p_reduced) as a function of Cα shift, Cβ shift and secondary structure. In terms of the logits presented above, Λ_1 and Λ_2, these probabilities are

$$\text{p.oxidized} = (e^{\Lambda 1}e^{\Lambda 2} + e^{\Lambda 1})/(1 + e^{\Lambda 1} + e^{\Lambda 2} + e^{\Lambda 1}e^{\Lambda 2})$$

$$\text{p.reduced} = e^{\Lambda 2}/(1 + e^{\Lambda 1} + e^{\Lambda 2} + e^{\Lambda 1}e^{\Lambda 2})$$

$$\text{p.zinc\_ligated} = 1/(1 + e^{\Lambda 1} + e^{\Lambda 2} + e^{\Lambda 1}e^{\Lambda 2})$$

These probability countours are plotted in Figure 3a–c.

Our modeling assumes (i) three Cys states: oxidized, non-metal ligated reduced, and ''other'', and (ii) similar positions of these states with respect to one another in the Cα/Cβ plots for each secondary structure type (alpha, beta, and non-regular). To allow the relative positions of the states to be different would require a much more complicated model, which is not indicated by the available data. For alpha and non-regular secondary structure the data fit well taking the ''other'' state as the Zn-ligated reduced state (Figure 3b,c). The corresponding region of the β-strand plot (Figure 3a) with low probability of being either oxidized or reduced non-metal ligated is *predicted* to be the region characterizing the Zn-ligated reduced state. Inspection of Figure 2 or Figure 3c (non-regular secondary structure, for which the most Zn-ligated Cys chemical shift data

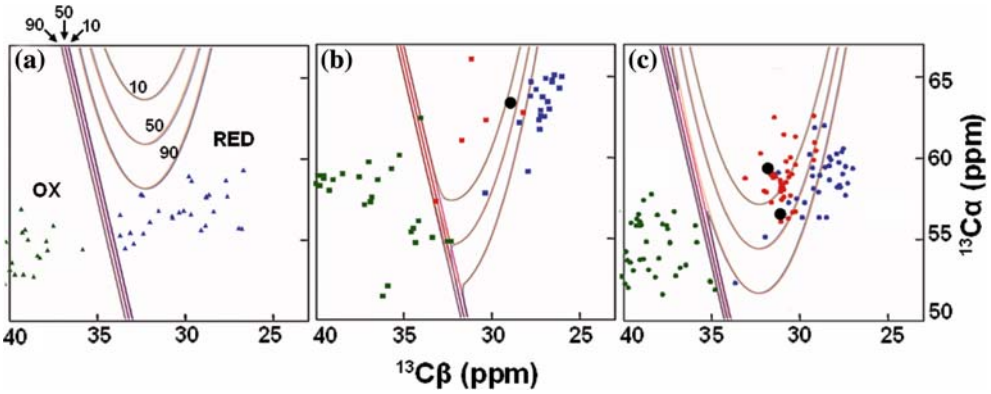

Figure 3. Cys oxidation/ligation state probabilities as a function of Cα and Cβ chemical shift and secondary structure. Each contour plot depicts oxidized (OX), reduced non-metal-ligated (RED), and Zn-ligated probabilities from the optimally fit logistic regression models. The contours were derived from Cα and Cβ cysteine chemical shifts residing within (a) β-strand, (b) α-helical and (c) non-regular structures. The training data (same data as Figure 2) are superposed. Also superimposed, in the appropriate panel by secondary structure, are the data points indicated by ●, which comprise Cys41 (59.9, 32.2 ppm), Cys66 (56.9, 31.1 ppm) and Cys128 (63.7, 29.6 ppm) from the test protein, *B. subtilis* IscU. These probability surfaces infer Cys residues with Cα/Cβ shifts on the left side of the plot to be oxidized, those on the right side as reduced non-metal-ligated, and those toward the upper-center as Zn-ligated cysteines. The three linear contours just left of center on each plot denote, from left to right, Cα/Cβ pair values having 90%, 50% and 10% probability, respectively, of arising from oxidized cystines. The other contours mark the transition, going from bottom to top, from 90% probability of being reduced non-metal-ligated through an equal likelihood of being non-metal-ligated or Zn-ligated, to a 10% probability of being non-metal ligated.

are available), indicates that in the Cβ chemical shift range characteristic of Zn-ligated Cys residues (27 ppm $< \delta_{C13\beta} <$ 34 ppm, Figure 1) the corresponding Cα values of Zn-ligated Cys are downfield relative to Cα values in reduced non-metal ligated Cys residues. Thus, characteristic Cβ shifts (27–34 ppm) and *anticipated* downfield (relative to reduced non-metal ligated Cys values) Cα shifts for Zn-ligated Cys residues in β-strands are completely consistent with the available data. However, this is indeed just a prediction based on data available for Cys residues in alpha and non-regular secondary structures, which is yet to be verified by experimental data. None-the-less, the probability surfaces Cys residues in β-strands (Figure 3a) can be estimated based on the data for reduced (non-metal-coordinated) and oxidized β-strand Cys residues, as well as all of the data for the three classes of Cys residues in helical and non-regular structures. Classifying cysteines with the model incorporating all three parameters (Cα shift, Cβ shift, and secondary structure class) resulted in a predictor of Zn ligation and cysteine/cystine oxidation state with a recall, precision and *F*-measure of 83.7% and an accuracy of 95.1%.

Test protein

The predictive value of the probability surfaces shown in Figure 3 were tested using the *Bacillus subtilis* protein IscU (Swissprot locus: NIFU_BACSU). This protein provides a molecular scaffold for the construction and delivery of Fe–S clusters (Zheng et al., 1998; Agar et al., 2000). It is widely known that many iron ligated proteins are capable of binding Zn (Becker et al., 1998; Dauter et al., 1996; Fujii et al., 1997; Lipscomb and Strater, 1996). Indeed it has been shown that this is the case with *Haemophilus influenzae* IscU (Liu et al., 2005; Ramelot et al., 2004). *B. subtilis* IscU has three cysteines: Cys128, Cys66 and Cys41. Cys128 is in an α-helical secondary structural element and the other two cysteines are found in loop/turn regions. Our final hierarchical logits model, and the probability surfaces shown in Figure 3, predicted all three cysteines of IscU to be Zn-ligated. The one helical cysteine, Cys128, has a Cα/Cβ chemical shift vector residing within the > 50% Zn-ligated probability region (Figure 3e). The two remaining cysteines, Cys41 and Cys66 which correspond to loop/turn regions, reside in the > 90% and > 50% Zn-ligated probability regions, respectively (Figure 3f). The fact that all three cysteine residues are predicted to be "Zn-ligated" indicates a very high likelihood of metal ligation by IscU. Moreover, preliminary IscU structure calculations (carried out without any metal-binding constraints) place all three cysteines within 5 Å of each other.

To further validate the presence of a metal ligand in IscU, we added EDTA to the NMR sample (10 mM final concentration). If IscU possessed a metal which was susceptible to chelation by EDTA, then addition of EDTA could lead to structural destabilization. Figure 4a shows a well-dispersed 2D ^{1}H–^{15}N HSQC NMR spectrum collected on native IscU, indicative of a well-folded structure. After adding EDTA, the spectrum collapses (Figure 4b) in a manner indicative of global destabilization due to metal chelation. Finally, inductively coupled-plasma mass spectrometry (ICP-MS) identified stoichiometric amounts of Zn (0.9 mol Zn: mol IscU) and trace amounts of Fe (< 0.12 mol Fe: mol IscU). No appreciable amounts of other metals (Mg, Al, Ca, Ti, Mn, Co, Ni, Cu and Cr) were detected.

Discussion

Zn increases the nucleophilicity of thiol ligands by functioning as a Lewis acid, biasing the thiol:thiolate equilibrium towards thiolate (Zhou et al, 1999). Greater nucleophilicity implies increasing electron density on the sulfur atom, due to the formation of a coordinate covalent bond to the bound metal ion (Glusker et al., 1999). It therefore

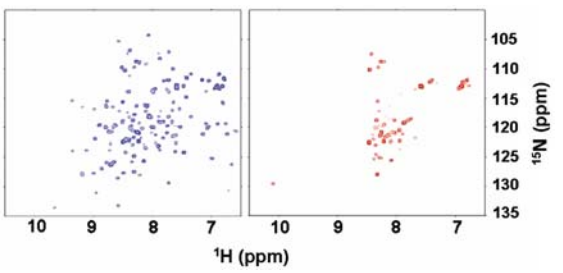

Figure 4. ^{1}H–^{15}N HSQC spectra of (a) Zn-bound [U-^{15}N] IscU (~0.7 mM) in 20 mM Na_2PO_4, 50 mM NaCl, 10 mM DTT, 0.02% NaN_3 at pH 6.5 and temperature of 20 °C, and (b) after addition of 10 mM EDTA.

seems plausible that the electron density around both the adjacent Cβ and Cα atoms of zinc-ligated cysteine is reduced relative to a fully reduced non-metal ligated cysteine. This de-shielding effect would account for the observed downfield shift of *both* Cα and Cβ resonances of Zn-ligated Cys residues.

Zn-coordinated cysteine residues can be identified by analysis of a combination of NMR data including ^{13}C chemical shift, secondary structure and initial fold information. If a protein possesses more than one Cys Cα/Cβ chemical shift vector within the >50% Zn distribution probability contours shown in Figure 3, it is very likely that the protein contains a cysteine-bound metal. In this case, additional steps are required to validate metal-ligation and to determine the nature of the bound metal. First, based on our experience with *H. influenza* IscU (Ramalot et al. 2004) and *B.subtilis* IscU, we suggest that metal coordination can often be validated by adding EDTA to chelate out the divalent cation. If the 2D ^{1}H–^{15}N HSQC NMR spectrum changes and/or the protein precipitates, then there is probably a divalent cation functioning in a structural capacity. Second, if an initial fold can be obtained (without using metal-binding distance constraints), the potential metal-binding site can sometimes be identified as three to four potential ligands converging on one area of the structure. In the case of Zn binding, the metal binding site is generally anionic, and the Zn-coordinating ligands are commonly cysteine, histidine, and (less commonly) acidic residues (see Table 2) and water (Lipscomb and Strater, 1996). Finally, if these results support the identification of a bound metal, ICP-MS can be used to determine the stoichiometry and type of metal present.

While the data available in the BMRB are sufficient for the analysis of chemical shift data characteristic of Zn-coordinated cysteine residues, there are yet not sufficient data to characterize effects on chemical shift of ligation by other metals. In particular, "Zn-ligated" Cys residues identified by carbon-13 chemical shift data, as described here, may actually ligate metals other than Zn. *Definitive identification of the coordinating metal must be made by follow-up analytical methods, such as ICP-MS.*

As the quantity, quality, and variety of NMR structures increase, so will the opportunity to datamine structurally characteristic chemical shift trends. Databases and tools such the RefDB (Zhang et al., 2003), SHIFTCOR (Zhang et al., 2003), and resonance assignment validation methods (Moseley et al., 2004) are necessary to address referencing and assignment errors. Using these tools, together with the invaluable data resources of the PDB and BMRB, additional chemical-shift-based methods for characterizing protein structure features can be expected in the coming years.

Electronic supplementary material is available at http://dx.doi.org/10.1007/s10858-006-0027-5.

Acknowledgements

This work was supported by NIH Protein Structure Initiative Grants P50 GM62413 and U54 GM074958. The authors would like to thank T.W. Weitsma and M.A. Kennedy of the Pacific Northwest National Laboratory (PNNL) for the ICP-MS analysis of *B. subtilis* IscU, and T.A. Ramelot (also from PNNL) for useful discussions pertaining to the identification of chemical shifts of metal-ligated residues.

References

Acton, T.B., Gunsalus, K.C., Xiao, R., Ma, L.C., Aramini, J., Baran, M.C., Chiang, Y.W., Climent, T., Cooper, B., Denissova, N.G., Douglas, S.M., Everett, J.K., Ho, C.K., Macapagal, D., Rajan, P.K., Shastry, R., Shih, L.Y., Swapna, G.V.T., Wilson, M., Wu, M., Gerstein, M., Inouye, M., Hunt, J.F. and Montelione, G.T. (2005) *Meth. Enzymol.*, **394**, 210–243.

Agar, J.N., Krebs, C., Frazzon, J., Huynh, B.H., Dean, D.R. and Johnson, M.K. (2000) *Biochemistry*, **39**, 7856–7862.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) *Nucleic Acids Res.*, **32**, 138–141.

Becker, A., Schlichting, I., Kabsch, W., Groche, D., Schultz, S. and Wagner, A.F. (1998) *Nat. Struct. Biol.*, **5**, 1053–1058.

Berg, J.M. and Shi, Y. (1996) *Science*, **271**, 1081–1085.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., Iyer, V., Melsopp, C., Mongin, E., Pettett, R., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Birney, E. (2003) *Nucleic Acids Res.*, **31**, 38–42.

Clarke, N.D. and Berg, J.M. (1998) *Science*, **282**, 2018–2022.

Coplen, T.B., Bohlke, J.K., De Bievre, P., Ding, T., Holden, N.E., Hopple, J.A., Krouse, H.R., Lamberty, A., Peiser,

H.S., Revesz, K., Rieder, S.E., Rosman, K.J.R., Roth, E., Taylor, P.D.P., Vocke, R.D. and Xiao, Y.K. (2002) *Pure Appl. Chem*., **74**, 1987–2017.

Dauter, Z., Wilson, K.S., Sieker, L.C., Moulis, J.M. and Meyer, J. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 8836–8840.

Drohat, A.C., Kwon, K., Krosky, D.J. and Stivers, J.T. (2002) *Nat. Struct. Biol*., **9**, 659–664.

Everitt, B.S. and Dunn, G. (2001) *Applied Multivariate Data Analysis*. Arnold, London.

Fujii, T., Hata, Y., Oozeki, M., Moriyama, H., Wakagi, T., Tanaka, N. and Oshima, T. (1997) *Biochemistry*, **36**, 1505–1513.

Glusker, J.P., Katz, A.K. and Bock, C.W. (1999) *Rigaku*, **16**, 8–16.

Hernick, M. and Fierke, C.A. (2005) *Arch. Biochem. Biophys*., **433**, 71–84.

Jansson, M., Li, Y.C., Jendeberg, L., Anderson, S., Montelione, B.T. and Nilsson, B. (1996) *J. Biomol. NMR*, **7**, 131–141.

Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577–2637.

Klug, A. and Rhodes, D. (1987) Zinc fingers: a novel protein fold for nucleic acid recognition. *Cold Spring Harb. Symp. Quant. Biol* **52**, 473–482.

Koradi, R., Billeter, M. and Wuthrich, K. (1996) *J. Mol. Graph*., **14**, 29–32.

Krishna, S.S., Majumdar, I. and Grishin, N.V. (2003) *Nucleic Acids Res*., **31**, 532–550.

Kwon, K., Cao, C. and Stivers, J.T. (2003) *J. Biol. Chem*., **278**, 19442–19446.

Lipscomb, W.N. and Strater, N. (1996) *Chem. Rev*., **96**, 2375–2433.

Liu, J., Oganesyan, N., Shin, D.H., Jancarik, J., Yokota, H., Kim, R. and Kim, S.H. (2005) *Proteins*, **59**, 875–881.

Lu, D., Searles, M.A. and Klug, A. (2003) *Nature*, **426**, 96–100.

Miller, R.G. (1997) *Beyond ANOVA: Basics of Applied Statistics*. Chapman & Hall, Boca Raton, FL.

Moseley, H.N., Sahota, G. and Montelione, G.T. (2004) *J. Biomol. NMR*, **28**, 341–355.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) *J. Mol. Biol*., **247**, 536–540 .

Neuhaus, D., Wagner, G., Vasak, M., Kagi, J.H. and Wuthrich, K. (1984) *Eur. J. Biochem*., **143**, 659–667.

Pelton, J.G., Torchia, D.A., Meadow, N.D. and Roseman, S. (1993) *Protein Sci*., **2**, 543–558.

Ramelot, T.A., Cort, J.R., Goldsmith-Fischman, S., Kornhaber, G.J., Xiao, R., Shastry, R., Acton, T.B., Honig, B., Montelione, G.T. and Kennedy, M.A. (2004) *J. Mol. Biol*., **344**, 567–583.

Rea, P.A. (2003) *Nat. Biotechnol*., **21**, 1149–1151.

Sharma, D. and Rajarathnam, K. (2000) *J. Biomol. NMR*, **18**, 165–171.

Taylor, J.R. (1997) *An introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, Sausalito, CA.

Vasak, M., Worgotter, E., Wagner, G., Kagi, J.H. and Wuthrich, K. (1987) *J. Mol. Biol*., **196**, 711–719.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.Q.H., Chen, L. and Skupski, M. (2001) *Science,*, **291**, 1304–1351.

Zhang, H.Y., Neal, S. and Wishart, D.S. (2003) *J. Biomol. NMR*, **25**, 173–195.

Zheng, L., Cash, V.L., Flint, D.H. and Dean, D.R. (1998) *J. Biol. Chem*., **273**, 13264–13272.

Zhou, Z.S., Peariso, K., Penner-Hahn, J.E. and Matthews, R.G. (1999) *Biochemistry*, **38**, 15915–15926.