WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Exploiting 3D structural templates for detection of metal-binding sites in protein structures

Kshama Goyal and Shekhar C. Mande*

Laboratory of Structural Biology, Center for DNA Fingerprinting and Diagnostics, Nacharam, Hyderabad 500076, Andhra Pradesh, India

ABSTRACT

High throughput structural genomics efforts have been making the structures of proteins available even before their function has been fully characterized. Therefore, methods that exploit the structural knowledge to provide evidence about the functions of proteins would be useful. Such methods would be needed to complement the sequence-based function annotation approaches. The current study describes generation of 3D-structural motifs for metal-binding sites from the known metalloproteins. It then scans all the available protein structures in the PDB database for putative metal-binding sites. Our analysis predicted more than 1000 novel metal-binding sites in proteins using three-residue templates, and more than 150 novel metal-binding sites using four-residue templates. Prediction of metal-binding site in a yeast protein YDR533c led to the hypothesis that it might function as metal-dependent amidopeptidase. The structural motifs identified by our method present novel metal-binding sites that reveal newer mechanisms for a few well-known proteins.

Proteins 2008;70:1206–1218.

Key words: *metal-binding proteins; function prediction; structural signature; metal–ligands.*

INTRODUCTION

The biochemical function of a protein is usually dictated by the topology of the polypeptide chain, the placement of key functional residues in the three-dimensional space, and the nature of interacting cofactors. Cofactors often determine the functional properties of proteins and therefore, one of the important steps in characterizing the function of a protein is to identify the cofactor and the cofactor-binding site in the polypeptide. Proteins bind to varied types of cofactors such as ATP, NADP, halides, pyridine, cyanate, and so forth. The most widely used cofactors that display diverse functions are, however, metals. Binding of metal ions is known to have a profound effect on the overall protein conformation, where metals play a wide range of roles such as stabilizing the active conformation of a protein, playing a structural role or acting as catalysts in several catalytic and regulatory processes. For example, metalloproteases harbor a Zn^{2+} ion that is in most cases held by two histidines of an HEXXH motif and one glutamate or histidine and a water molecule.[1] The water molecule held by Zn^{2+} ion attacks and hydrolyzes the peptide bond in the substrate, whereas the glutamate residue of the HEXXH motif donates a proton to the leaving substrate and restores the original conformation. The Zn^{2+} ion in this manner displays a catalytic function in metalloproteases. Thus, in the era of large-scale genome sequencing and structural genomics, recognition of metal-binding sites in proteins could be a significant step in identifying their respective functions.

Metal-binding sites have been studied[2–11] extensively in the structures available in Cambridge structural database (CSD) or protein data bank (PDB). It has been observed that metals vary in their coordination number as determined by their structural and functional roles. Also, metals display different geometries on the basis of their coordination number in different metalloproteins.[12,13] The metal-binding sites in polypeptides are of diverse nature such as those comprising backbone oxygen and nitrogen atoms, or those of side chain oxygen, nitrogen, and sulfur atoms. Yet each metal has specific ligand preferences.[9] For example, calcium and magnesium ions prefer side chain oxygen atoms of aspartate and glutamate residues. On the other hand the d-block metals such as zinc and iron mainly bind to nitrogen- and sulfur-containing residues such as histidine and cysteine.

Experimental techniques such as X-ray crystallography are often encountered with difficulty in the correct identification of a metal ligand, unless prior identity of the metal is known. Experimental evidence of the presence of a metal ion can be more reliably obtained by techniques such as atomic absorption spectroscopy,[14] extended X-ray

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

*Correspondence to: Shekhar C. Mande, Center for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad 500076, India. E-mail: shekhar@cdfd.org.in

Received 5 January 2007; Revised 9 April 2007; Accepted 30 April 2007
Published online 10 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21601

absorption fine structure (EXAFS), or X-ray absorption near edge structure (XANES).[15] Although these techniques provide definite evidence about the presence of a metal in a protein, it is difficult to perform such experiments for all the structural genomics proteins. Further, proteins might not always retain the bound metal during purification or crystallizations. Computational tools therefore can complement these experimental limitations and thereby aid functional annotations.

In the present work, we have reanalyzed the metal-binding sites in the structures available in the PDB database. This analysis was carried out to judge the consistency of the knowledge of variation in the metal-binding site observed earlier with smaller number of structures. Not surprisingly, earlier inferences still hold true with small variations in certain parameters.[16] Information so obtained was used to design 3D motifs for metal-binding sites. All the structures in PDB database were then analyzed to validate these motifs. Interestingly, the structural motifs matched binding sites in structures where the role of metal was known but the metal-binding site was not known. We discuss some of these interesting findings in detail here.

RESULTS

Analysis of the known metal-binding sites

All the structures in the PDB, which possess biologically important metal ions bound in them, were analyzed for their characteristics of metal-binding sites. As mentioned in the Materials and Methods section, PDBselect90 was used instead of the entire PDB database so as to avoid redundancy. The following metal-bound structures from the PDBselect90 list were obtained: Ca (430 structures), Cu (52 structures), Fe (57 structures), Mg (379 structures), Mn (118 structures), and Zn (585 structures). Structures containing Ca, Mg, and Zn metal are significantly more populated over other metals such as Ni or Co, which perhaps might reflect their relative use in biological systems. Likewise, large numbers of structures are available with bound Ca, Mg, and Zn atoms as compared with structures harboring Fe, Cu, and Mn atoms.

The number of ligand atoms for different metals normalized by the total number of ligands in all the metal-binding PDB structures is shown in Figure 1(a). In certain cases, the number of the protein ligand atoms for a metal was less than three, which meant that the remaining coordination bonds were satisfied by water molecules. Interestingly, some metal chelation bonds were observed to be longer than anticipated, explanation for which can be sought as follows: in transition metals (d-block metals) there are two groups of d-orbitals. One set of d-orbitals (e_g) has higher energy than the other set (t_{2g}). In low-spin complexes, that is, complexes with high coordination number, due to electron–electron repulsion distortion of the orbitals occurs to remove degeneracy known as Jahn-Teller effect or Jahn-Teller distortion. This causes distortion of the electron cloud around the metal and hence weak bondage with ligand atoms with long bond lengths. Thus, the long bond distances in the cases of metals could be due to the Jahn-Teller effect[9] thereby explaining the seemingly unreasonable observation of low coordination number.

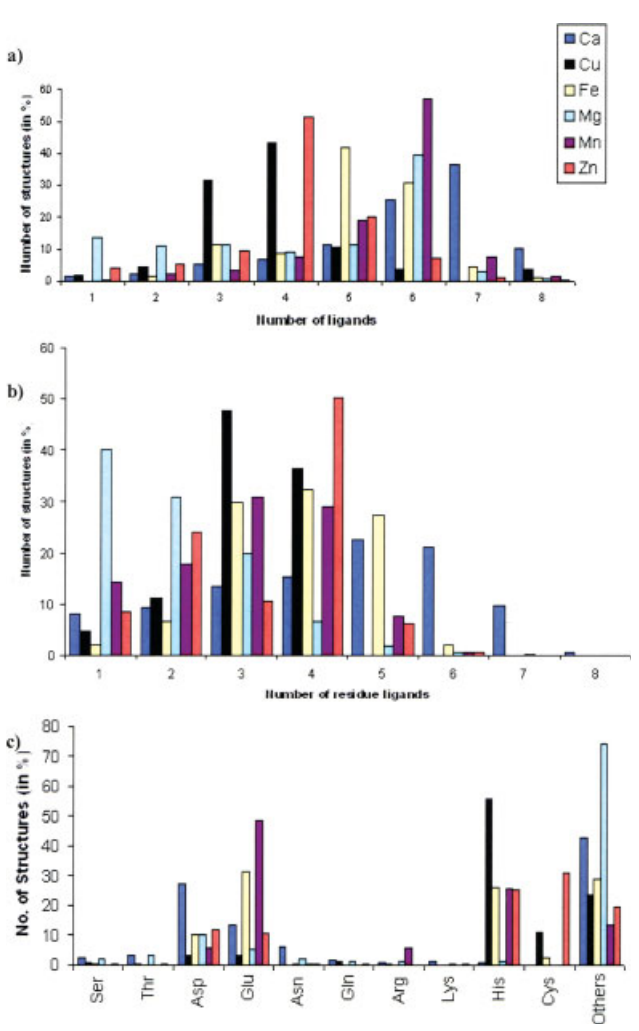

Figure 1

*Analysis of metal-binding sites in nonredundant structures. (**a**) The number of ligands, including water molecules and inhibitors for different metals as observed in PDBselect90 list. (**b**) The number of amino acid residue ligands for each metal in PDBselect90 list. (**c**) Frequency of different residues as ligands for different metals.*

The total number of amino acid ligands for each metal was also determined [Fig. 1(b)]. It was observed that for metals with higher coordination numbers, the amino acid residues alone do not satisfy all the valencies. This could be attributed to increase in ligand–ligand repulsion as the coordination number increases. Thus, in the

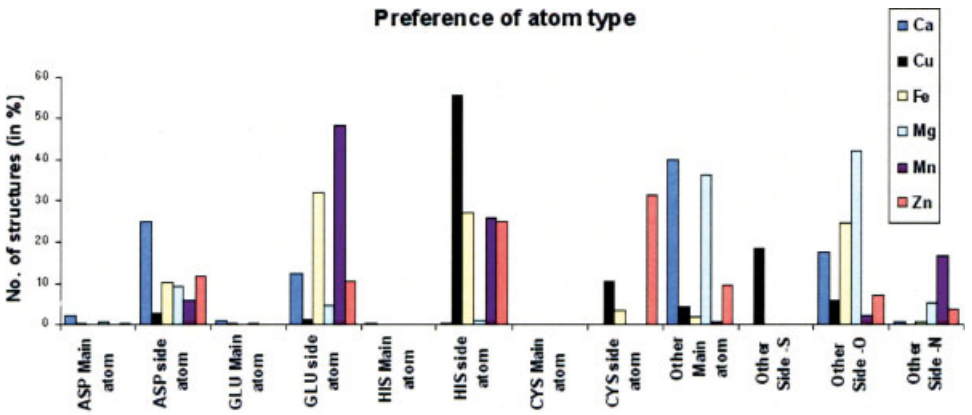

Figure 2

Preference of different metals for side-chain or main-chain ligand atoms. Statistical analysis clearly suggested that side chains of Asp, Glu, His, and Cys are preferred ligands for metal ions.

higher coordination number complexes, to reduce steric hindrance between the side chains of the chelating residues, water molecules satisfy the remaining valencies. Besides, certain residues such as aspartate and glutamate are bidentate and can provide two ligand atoms thereby reducing the noticeable number of individual ligands.

The carbonyl oxygen and the peptide nitrogen atoms of all the 20 amino acids are capable of donating electrons to the metals. The other electron donors in amino acids are the side-chain —O, —N, and —S atoms. Different residues that possess side-chain donor atoms are hydroxyl oxygens of Ser and Thr, carboxyl oxygens of Asp and Glu, amide oxygen and nitrogen of Asn and Gln, amine nitrogen of Arg and Lys, imidazole nitrogen of His, and thiol group of Cys. Analysis showed that in spite of the use of all these atoms as ligands, divalent metals prefer Asp, Glu, His, and Cys [Fig. 1(c)]. As shown in the Figure 1(c), main chain carbonyl oxygen and nitrogen of the residues depicted as "others" also coordinate with metal ions. However, no side chain preference in this category is observed as anticipated from the participation of only the main chain atoms in metal chelation. Metals such as Ca^{2+} and Mg^{2+} being strong electron quenchers might prefer more nucleophilic side-chain electron donors than main chain atoms (Fig. 2).

Analysis of metal–ligand (M–L) distances showed that on an average all the metals form coordinate bonds of the order of 2.0–2.5 Å. The observed mean and standard deviation in different metals is listed in Table I. The M–L distances were longer than 2.5 Å for Mn^{2+}/Mn^{3+} and possess a mean of 2.64 Å. This, as explained earlier, might be due to the Jahn-Teller effect, which is more prominent in Mn^{2+}/Mn^{3+}.

Interestingly, it was observed that the metal ligands need not necessarily be solvent accessible. Residues binding to metals possess solvent accessibility ranging from 0 to 100% (data not shown); yet calculation of per-atom solvent accessibility (accessibility of the liganding atom) showed that most of the chelating atoms possess solvent accessibility in the 0–20% accessibility range (see Materials and Methods section for details) (Fig. 3). The obser-

Table I

The Mean and Standard Deviation of the M–L Distances for Different Metals

Metal atom	Mean of M–L distance	Standard deviation of M–L distance
Ca^{2+}	2.45	0.2
Cu^{+}/Cu^{2+}	2.18	0.26
Fe^{2+}/Fe^{3+}	2.23	0.25
Mg^{2+}	2.3	0.28
Mn^{2+}/Mn^{3+}	2.64	0.26
Zn^{2+}	2.27	0.24

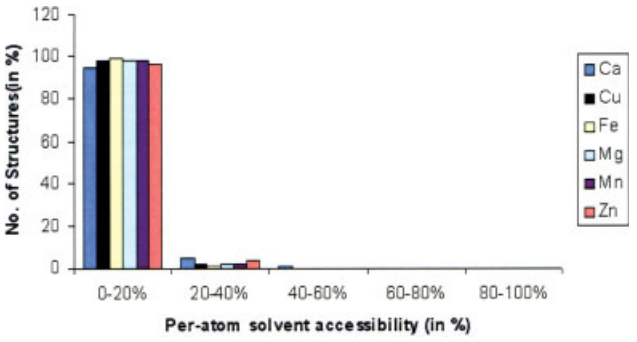

Figure 3

Variation in the per-atom solvent accessibility of residues chelating metals.

DOI 10.1002/prot

vations, although unusual, could be ascribed to the structural role of the bound metal ions. Most of the metal ions that play structural roles are acquired during the folding of the protein. Hence, residues interacting with the structural metal ions are less than 20% accessible.

Structural template generation

The range of distance variation between the C^{α}, C^{β}, and pseudo atoms of the residue ligands was calculated for all the different metals. It was observed in our analysis that in spite of metal specificity of each site, all the metal-binding sites are similar in terms of inter-C^{α} or inter-C^{β} distances. Consequently, only the zinc-bound structures were considered for template generation. The Zn-list was selected due to availability of a larger number of structures with bound zinc than any other metal. Thus, the final training set consisted of all the 585 Zn-containing structures in the PDBselect90 list. The complete list of training set structures was screened for the structures having a three-residue Zn-binding site and those that have four-residue Zn-binding site. Hundred structures were used as the training set for the three-residue template and 400 structures were used for the four-residue 3D template generation. Different geometrical parameters were derived by training the three-residue (triad) and four-residue (tetrad) structural templates (Table II). Similar parameters were used in the two cases, the only difference being the training set structures. The C^{α}—C^{α} and C^{β}—C^{β} distances showed large variation from ~3 to 12 Å than pseudo atom distance variation. The pseudo atom distances varied from 2 to 9 Å.

Template validation

A positive data set containing 571 structures and a negative data set containing 1000 structures were used for the validation. Metal-binding sites could successfully be identified in 268 structures among the 571 of the positive data set. On the other hand, no metal-binding site could be found in 881 structures of the negative data set. Thus, the analysis suggests that the structural templates described in the current work possess 46.9% sensitivity and 88.1% specificity, where, sensitivity is defined as the ratio of predicted true positive with respect to all positives. On the other hand, specificity is defined as the ratio of predicted true negatives with respect to all negatives. A relatively large number of false negatives were either due to involvement of an inhibitor or a water molecule as M–L, or because of less than three amino acid ligands in the binding site. False negatives were also due to involvement of residues other than Asp, Glu, His, and Cys as metal ligands. However, other residues were not considered during prediction of metal-binding site, so as to obtain high specificity of prediction.

Table II
Range of Values Obtained for Different Geometrical Parameters

Parameter	Maximum	Minimum
Values obtained for triad sites		
Cα—Cα distance between res1 and res2	12.87	3.79
Cβ—Cβ distance between res1 and res2	11.15	3.52
Cα—Cα distance between res1 and res3	12.85	3.90
Cβ—Cβ distance between res1 and res3	11.22	3.25
Cα—Cα distance between res2 and res3	12.48	3.71
Cβ—Cβ distance between res2 and res3	10.63	3.45
Angle between Cα and Cβ plane	35.07	0.07
Volume of the sphere within Cβ triangle	133.20	5.70
Distance between pseudo atom1 and 2	6.63	2.04
Distance between pseudo atom1 and 3	9.31	2.03
Distance between pseudo atom2 and 3	7.93	1.75
Values obtained for tetrad sites		
Cα—Cα distance between res1 and res2	11.50	3.78
Cβ—Cβ distance between res1 and res2	10.52	0.01
Cα—Cα distance between res1 and res3	12.85	3.59
Cβ—Cβ distance between res1 and res3	11.32	3.09
Cα—Cα distance between res1 and res4	13.12	4.10
Cβ—Cβ distance between res1 and res4	10.69	3.32
Cα—Cα distance between res2 and res3	12.37	3.76
Cβ—Cβ distance between res2 and res3	10.69	3.23
Cα—Cα distance between res2 and res4	12.40	3.70
Cβ—Cβ distance between res2 and res4	10.97	2.73
Cα—Cα distance between res3 and res4	12.30	3.70
Cβ—Cβ distance between res3 and res4	10.08	3.08
Angle between Cα and Cβ plane	34.56	0.34
Volume of the sphere within Cβ triangle	113.95	5.45
Distance between pseudo atom1 and 2	8.27	0.94
Distance between pseudo atom1 and 3	7.38	1.13
Distance between pseudo atom1 and 4	7.52	1.70
Distance between pseudo atom2 and 3	6.91	1.15
Distance between pseudo atom2 and 4	8.23	1.15
Distance between pseudo atom3 and 4	6.35	0.92

Template matching

Tetrad metal-binding sites

Each structure in the complete PDB (April, 2005 release) containing 31,059 entries was scanned for the structural fragments that match the three- or four-residue templates. The algorithm for four-residue template matching predicted 9767 putative tetrad sites belonging to 2525 structures (Supplementary Table I). Out of the 2525 structures, prior knowledge of metal-binding property was well known in 2338 structures. However, among these, 2272 are holo structures and possessed bound metal atom at the site predicted by the program, whereas 66 structures included apo structures. Interestingly, our analysis predicted 187 additional structures for which either the metal-binding site is not known or the role of metal itself is not known. Further analysis of the predicted tetrad metal-binding site in these 187 structures led to interesting findings, few of which are discussed below.

The current algorithm identified a putative metal-binding site in an antifungal protein (pdb code 1AFP).[17] AFP, protein from ascomycete *Aspergillus giganteus*, inhibits

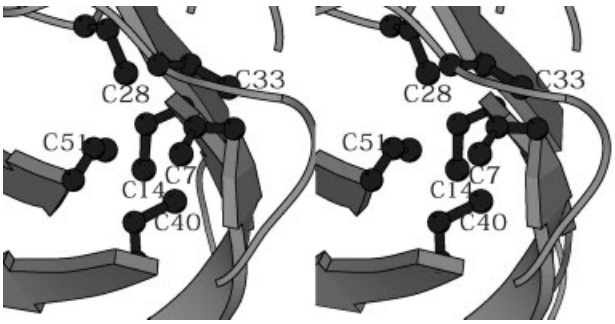

Figure 4

The thiolate cluster in antifungal protein (1AFP) in stereo. Our method predicted that this cluster is involved in metal binding including six cysteine residues. We suggest (see text in details) that metal-binding property of the protein might be relevant in the antifungal properties of this protein.

growth of broad-spectrum filamentous fungi. Unlike other antifungal proteins, AFP does not permeabilize fungal cell wall. Therefore, its mechanism of action is not well understood. We identified a putative metal-binding thiolate cluster in this protein, which we believe might help in elucidating its mechanism of action. The thiolate cluster in AFP resembles the one that occurs in metallothioneins (MTs). The conformational arrangement of cysteine residues in AFP is also similar to the cysteine cluster in MTs (Fig. 4). Moreover, the S^{γ} atom of each of the cysteines predicted to be involved in metal binding is too distant (3.2–3.9 Å) to be involved in disulfide bond formation. Thus, it is probable that the cysteine residues in AFP are metal ligands.

MTs are proteins that possess metals coordinated by a group of Cys residues forming a metal–cysteine thiolate cluster. In such cases, 7–12 di- or monovalent metal ions are held by 8–10 Cys residues.[18] Bound divalent metal ions display tetrahedral geometry, while monovalent ions bind in trigonal bipyramidal geometry. Although the actual numbers of sites are likely to be less than the total number of predicted sites in 1AFP, the algorithm predicted a total of 10–16 sites for these structures. It was observed that the different permutations of ligand cluster qualify the geometry criterion due to proximity and thus additional sites are predicted. Manual inspection of these sites showed that all the combinations satisfy geometrical criteria for binding and therefore are theoretically possible. MTs are known to be involved in varied functions such as metal storage, transport and detoxification. Besides, MTs play important role in host defense processes and metal metabolism.[19] It is therefore tempting for us to suggest that 1AFP might exerts its antifungal action through metal-binding clusters.

The current algorithm predicted a metal-binding site in diiron carboxylate protein from *Thermotoga maritima* (pdb code 1VJX) (unpublished). This protein is a known diiron protein, but the structure does not contain any bound iron atoms. Since, the site record is not available for this protein, we presume that the binding site of the iron is not known in this case. The algorithm predicted three metal-binding sites in it: (a) E14, E43, H46, E130; (b) E43, E100, E130 E133; (c) E43, H46, E130, H133. Diiron carboxylate sites, as the name suggest, possess two iron atoms bridged by carboxylates of two glutamate residues. Such sites occur in the active sites of several enzymes such as methane monooxygenase, ribonuclease reductase, stearoyl ACP-Δ^9 desaturase, and ferritins. Apart from the bridging glutamates, two additional glutamates and one histidine also coordinate with each metal ion and give rise to a five-coordination geometry.[20] Two of the predicted sites (a, b) in 1VJX possess all the geometric features observed in other di-metal carboxylate proteins [Fig. 5(a)]. The third site is probably predicted due to the structural proximity of the two

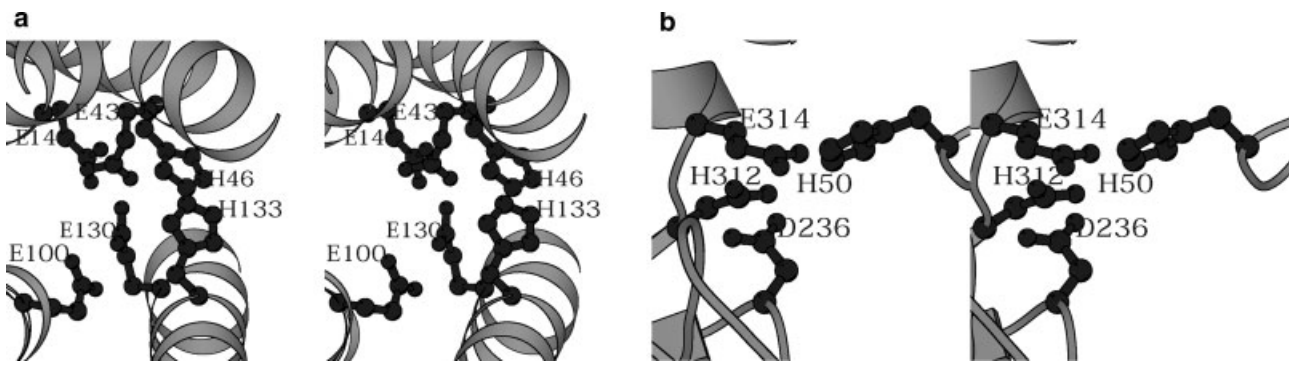

Figure 5

Metal-binding site in T. maritima *proteins in stereo. (**a**) Diiron carboxylate protein (1VJX), (**b**) glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (1HDG). The former is known to be a metal-binding protein, while the latter is a novel site predicted by our program.*

DOI 10.1002/prot

binding sites, as it is a hybrid of the actual metal-binding sites. 1VJX also possesses the EXXH motif known to bridge the two metals in the known examples. Moreover, the motifs are packed in four-helix bundles. Therefore, the prediction of similar metal-binding site further supports that 1VJX is a metal-binding protein with characteristics similar to ferritin-like proteins.

One of the most interesting outcomes of the current study was the prediction of a metal-binding site in D-glyceraldehyde-3-phosphate dehydrogenase (GAPDH). GAPDH carries out a key step in the glycolytic pathway where it converts D-glyceraldehyde-3-phosphate to 1,3 bisphosphoglycerate in a NAD-dependent manner. Several structures of GAPDH are known from different organisms. The enzyme plays a crucial role in organisms such as *Plasmodium falciparum* where glycolysis is the sole pathway for energy generation.[21] Besides, GAPDH is known to perform functions other than energy generation. For example, GAPDH is involved in membrane fusion, microtubule bundling, nuclear RNA export, DNA replication, and repair and apoptosis.[22] This suggests that in spite of being ubiquitously present, GAPDH shows species-specific variations in its function and mechanism of action. Our analysis predicted a well-placed metal-binding site in GAPDH structures. The organisms include *T. maritima* (1HDG),[23] *Leishmania mexicana* (1GYP, 1GYQ, 1I32, 1I33),[24] *Trypanosoma cruzi* (1K3T, 1ML3, 1QXS),[25] *Trypanosoma brucei* (1GGA),[26] and *Homarus americanus* (4GPD).[27] In all these cases the predicted metal-binding site includes one histidine, two aspartates, and one glutamate residue. Such a metal-binding site has been observed in certain hydrolases that are functional under different stress conditions and some other metal-binding proteins. Interestingly, the last two residues of the predicted site form a DXXE pattern and occur in all the GAPDH structures mentioned above. Further, it has been demonstrated earlier that the mammalian and yeast GAPDH are zinc-binding proteins,[28] where the Zn^{2+} ion is essential for the activity of the enzyme and stabilizes its structure. However, it is not understood if the metal ion plays a structural or a functional role in bacterial GAPDH structures. The metal-binding site as shown for 1HDG [Fig. 5(b)] is in close proximity of the bound NAD. It is likely that the bound metal atom might be required for the proper conformation of active site or for an appropriate geometry for binding NAD. Alternatively, metal might play a role in stabilizing the protein conformation under stress conditions such as temperature stress in the case of *T. maritima*. Considering that identical metal-binding sites have been predicted for the all the parasitic organisms listed above, this observation might be useful to the development of parasite specific inhibitors.

A metal-binding site was also proposed in a PhoU-like transcription regulator from *Aquifex aeolicus*, (pdb code 1T72).[29] PhoU regulon is involved in the uptake of an important nutrient in the bacterial physiology, inorganic phosphate (P_i). The only other structure known for the members of this large family is the PhoU homologue from *T. maritima* (pdb code 1SUM).[30] The two proteins are only 23% identical in their amino acid sequence. The *T. maritima* PhoU contains two multinuclear metal-binding sites with three iron atoms and one nickel atom bound to N-terminal region and three iron atoms bound to its C-terminal region. Both the sites harbor metals bound to E(D)XXXD motif.[30] It has been shown that the arrangement of multinuclear iron cluster in *T. maritima* is similar to the diiron metal assembly observed in ferritin-like proteins that are involved in iron uptake. The biological role of the multinuclear metal cluster in the PhoU-transcription regulators is not known. However, uptake of iron ions during cell growth, as demonstrated for *T. maritima* PhoU clearly suggests functional association of the metal cluster. On the basis of multiple sequence alignment of PhoU protein sequences from different organisms, it has been suggested that the multinuclear iron cluster binds to a total of four repeats of E(D)XXXD motif. The 1T72 being an apo-structure, the metal-binding residues were not reported.[29] Encouragingly, the current algorithm identified the metal-binding site for this protein, which was exactly similar to the site obtained upon structural alignment with 1SUM [Fig. 6(a,b)]. In the predicted metal-binding site, all the four repeats are well in agreement with the E(D)XXXD motif. This suggests that the multinuclear metal-binding motif in PhoU-like regulator proteins consists of E(D)XXXD motif. Further, it also supports the suggested involvement of metal in the function of PhoU protein.

Another interesting prediction of the algorithm described here is that of the putative metal-binding site in 1UC2, a hypothetical extein protein from *Pyrococcus horikoshii* OT3. 1UC2 shows 33% sequence identity with RtcB protein from *Escherichia coli*. *rtcB* gene occurs on an operon with the cyclase gene, *rtcA*.[31] Homologues of RtcB have been observed to occur across all lineages, bacteria, archaea, and eukarya. However, its function is still not understood. Six conserved histidine residues have been observed in RtcB and hence, involvement of a metal ion has been proposed for it. 1UC2 is the first structure of this family of proteins.[32] As described, 1UC2 does not show structural similarity to any of the known structures or fold. The binding site of the proposed metal involved in RtcB function was stated as C98, H203, H234, and H404 on the basis of proximity of these residues [Fig. 7(a)]. However, the current algorithm predicted the involvement of H329 rather than H203 as the metal ligand. The predicted site [Fig. 7(b)] is well in agreement with a square planar geometry observed for metal atoms. As shown in Figure 7, the predicted site possesses better geometry than the proposed site. Thus, the algorithm could successfully suggest the unknown metal-binding site in a protein structure.

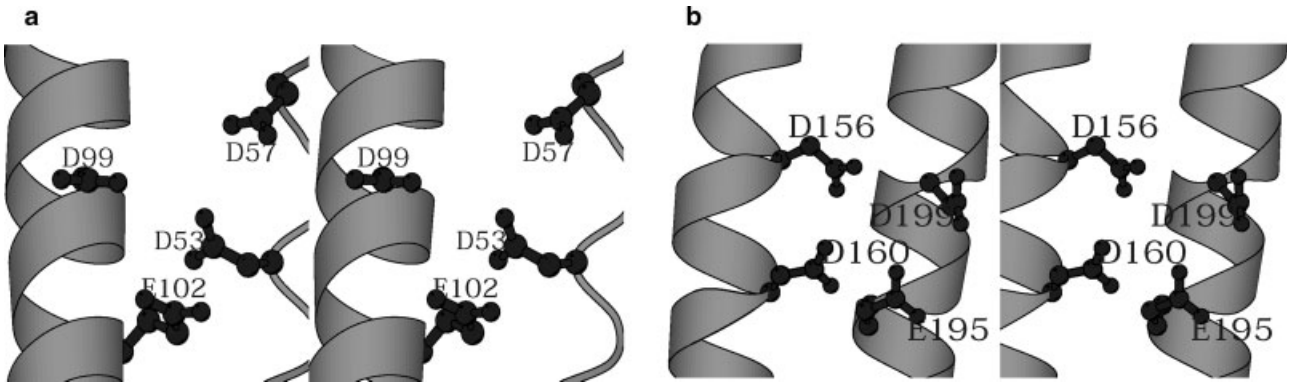

Figure 6

Two multinuclear metal-binding sites in the PhoU protein (1T72) from A. aeolicus *in stereo. (**a**) N-terminal site D53-D57-D99-E102, (**b**) C-terminal site D156-D160-D199-E195. All the residues in both the sites appear to be well poised for metal binding.*

Metal-binding site was also predicted in *E. coli* colicin protein E3 (pdb code 1E44). *E. coli* and related bacteria produce colicin-like antibiotics that act against bacteria by competing for limited nutrient resources. Colicins act either by depolarizing cell membrane or by specifically cleaving 16SrRNA or tRNA. Colicin E3 is a ribonuclease that specifically cleaves 16srRNA. The host bacterial cells protect themselves against toxic effects of E3 by producing an immunity protein called as Im3. Im3 inhibits E3 activity by binding to its C-terminal domain that displays ribonuclease activity. 1E44 is a structure of Im3 complexed with C-terminal domain of E3.[33] The current algorithm predicted metal-binding site in colicin E3 chain of the structure [Fig. 8(a)]. The prediction was not surprising because metal requirement is known for ribonucleotidyl activity of other RNase such as RNaseIII.[34] The *E. coli* RNase III has been shown to require Mg^{2+} or Mn^{2+} ion for its activity. The metal ion in RNase III binds to a combination of aspartate and glutamate residues. However, metal requirement has not been demonstrated for E3 colicin. The metal-binding site predicted by our algorithm, as shown [Fig. 8(a)], has a nearly perfect geometry for binding. Further, an extensive study has been carried out to identify the active site of E3 colicin by generating 27 independent point mutations in this 96-residue long protein. Five mutants were observed to be completely inactive and four displayed <1% activity.[35] Interestingly, four of these nine positions are the residues that have been predicted by current algorithm as metal ligands. On the basis of similarity with barnase, H58 and E62 of E3 colicin have been proposed to form the acid–base pair required for catalysis[35] and thus explained the inactivity of H58A and E62A mutants. However, no role could be described for the inhibitory effects of D55A and E60A mutations. On the basis of the prediction by the current algorithm, we propose that E3

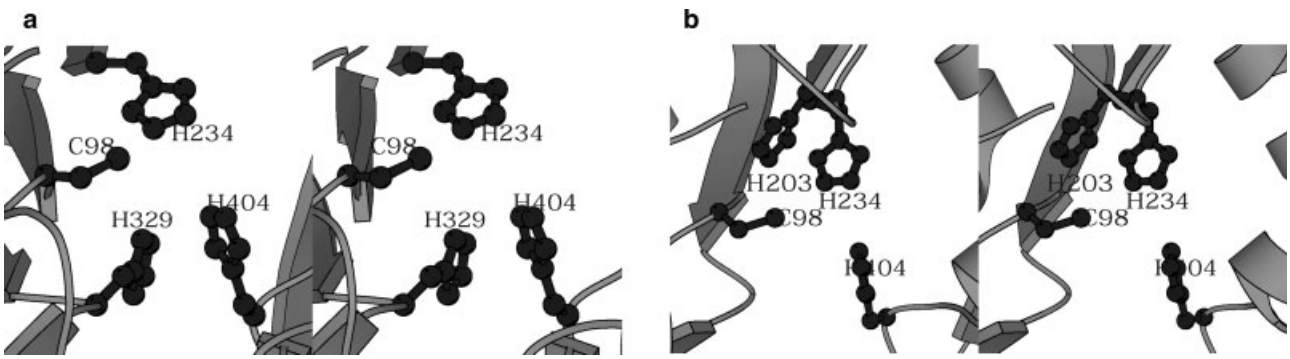

Figure 7

*Stereo view of the metal-binding sites in extein protein (1UC2). (**a**) Site proposed on the basis of alignment. (**b**) Site predicted by the current algorithm. Our method predicts a different metal-binding site than the published proposal.*

DOI 10.1002/prot

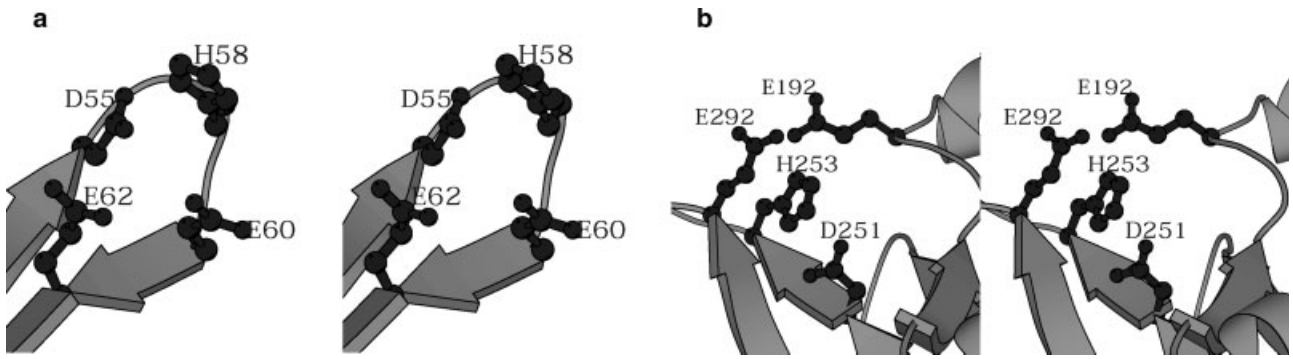

Figure 8

*Stereo view of the predicted metal-binding site in (**a**) colicin E3 protein (1E44) from* E. coli *and (**b**) Exo-β-1,4-glucanase (1EQC) from* C. albicans.

colicin displays metal-dependent ribonucleotidyl activity and inability of D55A, H58A, E60A, and E66A mutants to bind to the metal ion renders them inactive.

The current algorithm predicted a metal-binding site in *Candida albicans* exo-β(1,3)-glucanase (pdb code 1EQC). Exo-β-(1,3)-glucanases belong to a large array of glycosyl hydrolases that hydrolyze a diverse array of polysaccharides. In spite of wide sequence variation, exo-β-(1,3)-glucanases have been grouped together with bacterial cellulases (endo-β-1,4-glucanases) into glycosyl hydrolase family 5.[36] All the family members of family 5 possess eight conserved residues and share common reaction mechanism. All the eight conserved residues occur around the catalytic site and share similar spatial arrangement.[37] Two of the eight invariant residues are the glutamates that act as a nucleophile and a proton donor, respectively, during catalysis. Interestingly, the current method proposed a metal-binding site including these two glutamates (E192 and E292) along with H226 and H253 [Fig. 8(b)]. The structure of *C. albicans* exo-β-(1,3)-glucanase[38] shows that eight residues conserved in the family are required for maintaining the orientation of E192 and E292. The structure suggested that Y255 helps maintaining the orientation of E292, while H253 influences E192 conformation in a similar manner. Further, D251 interacts electrostatically with H253. Thus, all the four predicted metal-binding residues are involved in generating the active site conformation. This implies that the predicted metal atom might play a structural role and might keep the catalytically important residues in proper conformation. Further, it has been proposed earlier[39] that 1ECE (*E. coli* endocellulase E1) possesses active site conformation similar to DNase I of *E. coli*. E39 of DNase I binds to a divalent metal ion and is equivalent to the proton donor (E168) of endocellulase E1. Thus, the active site similarity of Endocellulase E1 to DNase I and the metal requirement of the latter, together suggest that E1 might also bind to a metal. We propose that the glycosyl hydrolase family 5 members might bind to a metal ion that is not involved in catalysis, but might be essential to maintain the conformation of the active site.

Triad metal-binding sites

Scanning the PDB database for triad metal-binding sites resulted into 24,328 sites belonging to 7439 structures (Supplementary Table II). Out of 7439 structures, 4473 structures were observed to be holo structures with bound metal ions and 1481 were related apo structures. Remaining 1485 structures included several oxidoreductases, haloperoxidases, transferases that are known to require a metal ion for their activity. Interestingly, a few examples were identified by our analysis where the protein is known to bind a metal ion; however, the metal-binding site has not been identified. One of the most interesting findings is discussed below.

One of the most interesting predictions of the current algorithm was the prediction of metal-binding site in *YDR533c* ORF from *Saccharomyces cerevisiae* (pdb code 1QVV). 1QVV is a 25.5-kDa protein of unknown function. On the basis of sequence similarity to ThiJ/PfPI superfamily, it was proposed that YDR533c could be involved in stress response of yeast. Hypothesis that YDR533c acts as molecular chaperone was further strengthened by its observed similarity to *E. coli* Hsp31 and human DJ-1 proteins. Crystal structure of YDR533c was solved to establish its role as a molecular chaperone protein.[40] As expected from the sequence similarity, the core structure of YDR533c is similar to that of Hsp31 with an rmsd of 1.84 Å for 196 C^{α} positions.[41] Structural comparisons showed that YDR533c also possesses a Cys–His–Glu catalytic triad as that occurs in Hsp31, which is similar to the cysteine protease-like catalytic triad. However, proteolytic activity could not be obtained for YDR533c in spite of the presence of catalytic triad and elements of oxyanion hole. It was proposed that lack of activity was

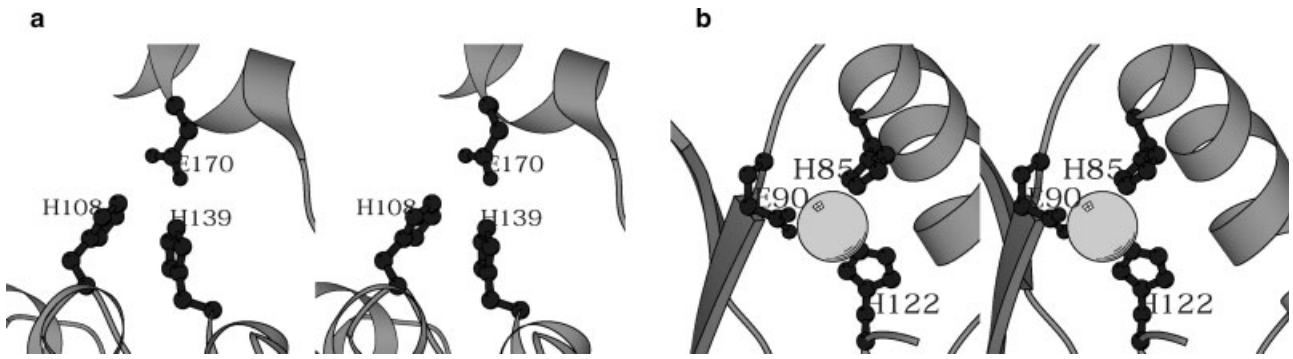

Figure 9

*The triad metal-binding site in (**a**) YDR533c (1QVV) from* S. cerevisiae. *(**b**) Hsp31 (1ONS) from* E. coli *is shown in stereo. The predicted site in 1QVV is similar to the known metal-binding site in 1ONS.*

due to inaccessible Cys138, the probable nucleophile. Surprisingly, chaperonic activity also could not be observed for YDR533c. Hsp31, the closest homologue of YDR533c, is 31-kDa heat shock protein of *E. coli yedU* gene which is overexpressed during heat stress. Hsp31 was shown to possess chaperonic[42] activity, but its proteolytic activity could not be demonstrated. However, a recent report[43] demonstrated that the protease-like catalytic triad in Hsp31 is not involved in proteolysis, rather it displays metal-dependent amidopeptidase activity in agreement with its narrow groove.[42] In addition to the narrow groove catalytic triad, the crystal structure of Hsp31 (pdb code 1ONS) showed the presence of a Zn^{2+}-ion coordinated with 2-His-1-carboxylate motif. The metal-binding motif in Hsp31 is similar to Zn^{2+}-binding motif of carboxypeptidase A[44] and iron-binding motif of several hydrolases.[45] Yet interestingly, no metal was found in the YDR533c structure. Our current algorithm predicted a 2-His-1-carboxylate motif metal-binding site in YDR533c [Fig. 9(a)]. The predicted metal-binding motif of YDR533c is similar to the one observed in Hsp31 [Fig. 9(a,b)]. Therefore, we propose that YDR533c might be a stress related amidopeptidase like Hsp31. Further, we also propose that similar to Hsp31, YDR533c might be involved in metal dependent clearance of 8–12-mer peptides that accumulate in the cell during heat or other stresses.

DISCUSSION

A new sensitive and fast method has been described in the current study for the identification of metal-binding sites in the query structures. The method takes into account the information obtained from the analysis of known metal-binding sites. Algorithm searches only the electron donors as probable ligands and thus, it is highly efficient. Geometrical parameters are used to detect structurally similar metal-binding sites in the query structures. Thus, the algorithm enables functional annotation of new structures on the basis of the predicted metal-binding site. The most advantageous feature of the current method is that it does not require sequence information for identifying the site. Therefore, the method can be used to characterize even the remote sequence homologues without any loss of accuracy.

The significance of recognizing metal-binding sites is best exemplified by sulfatases. Sulfatases are the enzymes that hydrolyze sulfate ester bonds in wide variety of substrates. It was known that a conserved cysteine in sulfatases is oxidized to formylglycine and is crucial for the activity of this enzyme. However, the role of formylglycine and mechanism of action of sulfatases could not be understood. Interestingly, an Mg^{2+} ion was observed as bound to the formylglycine in the crystal structure of human arylsulfatase A.[46] It was then deciphered that formylglycine gets hydrated by a water molecule coordinated to the metal atom and is involved in novel transesterification reaction. Thus, metal-binding site identification enabled understanding of the reaction mechanism. Similarly, the role of PhoU protein in uptake of inorganic phosphate was not known. Identification of dicarboxylate iron cluster in the structure of PhoU protein[30] suggested that the PhoU-protein might be involved in inorganic phosphate metabolism using iron cluster as a cofactor. This implies that identification of metal-binding site can give clues about the function and mechanism of action of a protein. These examples therefore suggest that the ability of metal-binding site prediction makes the current algorithm a useful tool for function annotation of structural genomics proteins.

Availability of a large number of metal bound structures and significance of spotting metal-binding sites have led to approaches that exploit structural features to generate 3D-motif for prediction of metal-binding sites

DOI 10.1002/prot

in uncharacterized structures. For example, using position specific scoring matrices, metal-binding sites were identified in uncharacterized proteins.[47] Secondary structure and solvent accessibility were used as additional features in an artificial neural network to distinguish metal sites from nonmetal sites. Comparison of this artificial neural network method with the one reported in this communication showed that the current method is more sensitive. The artificial neural network based method possesses an average of 39.2% true positive rate at 5% false positive rate. However, the current method as described here exhibits a true positive rate of 46.9% for all the metals.

In another method, structural motifs were generated for sequence specific metal-binding patterns in DRESPAT.[48] This approach used a geometric subgraph to identify residue patterns in a group of sequences. These subgraphs were then used for the identification of binding sites in the query structures. However, DRESPAT method requires sequence information for all the different metal-binding site patterns. The availability of homologous sequences for all the metal-binding residue patterns therefore limits the scope of DRESPAT.

Kleywegt designed an algorithm SPASM[49] that not only predicts the metal-binding site but also design new motifs from a set of structures. This algorithm extracts residues binding to a ligand or a heterocompound and generates structural motifs using C^{α} and side-chain pseudo atoms of the ligand-binding residues. The sequence pattern of the ligand-binding residues is also considered. Similar to DRESPAT this method also requires a training set for all the metal-binding patterns. Further, the authors suggested that the motifs are too abstract for sensitive prediction. Russell[39] used similar approach as SPASM for prediction. However, to improve sensitivity C^{β} atoms were also used as a part of structural motif. The method however still searches for residue patterns as Cys_2–His_2 and not all the permutations of probable electron donors. One, another method uses a clustering procedure to recognize the cloud of protein atoms coordinating a metal-ion in the center.[50] The approach uses a set of high-resolution crystal structures for identifying metal-ion position to use it as a center and then uses fold-X force field for prediction of the sites in the query structures. Though highly sensitive the method requires optimum conformation of coordinating atoms. Therefore, it might not be suitable for unbound structures. This suggests that current algorithm might prove to be more useful for function annotation especially in low resolution and structural genomics structures.

Numbers of other methods also exist that use structural templates.[51–54] Polacco and Babbitt[51] used a method similar to SPASM for generating the structural motifs. However, evolutionary conservation was used as the criterion to identify the probable ligand-binding residues. The method has been observed to be highly accurate for four enzyme families. Nonetheless, the efficiency of the method for metal-binding site prediction is not known. Similarly, fuzzy functional forms[52,53] also use geometrical parameters and have been shown to perform proficiently for glutaredoxin/thioredoxin like active sites. Fuzzy functional forms predict well even in low-resolution protein models. However, their usability for metal-binding site prediction is not known.

Sequence based approaches have also been used for prediction of metal-binding sites. A machine learning method was trained on a set of metal-bound proteins.[55] The features used for training included occurrence probability, secondary structure propensity, and metal-binding character of all the 20 amino acids in the training set. However, the method was observed to possess low sensitivity and specificity. The current method is therefore advantageous as it might perform better than the existing methods. The current method can be accessed at http://sunserver.cdfd.org.in:8080/protease/PAR_3D/index.html.

There are certain limitations of the current method. Since, the minimum number of ligand-binding residues is set to three or four, the metal-binding sites involving less than three residues will not be identified. However, encouragingly the sites involving more than four residues would be predicted accurately. The method predicts a five residues site, as a combination of two, four-residue sites. Another limitation of the method is that it can identify only those sites that involve Asp, Glu, Cys, and His. Since, no significant signature has been observed for other residues as the metal ligands [Fig. 1(c)], we believe that current method does not lack generality by considering only the four residues mentioned above.

A large number of structures are known prior to the functional knowledge due to the increased efforts of structure genomics consortia. Thus, the methods based on structural motifs such as the current method and described by others would enable characterization of hypothetical proteins and structures. Also, these methods would aid in uncovering less understood evolutionary aspects. Identification of unknown ligand-binding site would also increase the extent of drug designing against known targets.

MATERIALS AND METHODS

The design and analysis of structural motifs in proteins for metal-binding sites involved three steps: (i) analysis of known metal-binding sites, (ii) generation of structural template on the basis of training set structures, and (iii) prediction of sites using the derived structural templates as in step (ii).

Analysis of known metal-binding sites

PDBselect90 was used to obtain the list of structures containing bound metal ions. PDB90 was considered to

avoid analyzing more than one crystal structure of the proteins with more than 90% sequence similarity. PDBselect90 list was obtained from http://bioinfo.tg.fh-giessen.de/pdbselect/ site and consist of structures solved at resolution higher than 3.0 Å and R-factor below 30%. PDBselect90 does not contain theoretical models. For the current analysis, only one subunit of oligomeric proteins was considered. For NMR structures, only the first model was considered for analysis.

All the structures in the PDBselect90 list were searched for the presence of Ca, Cu, Fe, Mg, Mn, and Zn metal atoms and a sublist was generated for each metal. Oxidation states and coordination number of the metals were not considered for the generation of the list. For each metal, ligands were identified in the listed structures by searching the structure file for all the atoms occurring at a distance of $\leq$3 Å from the metal atom. Metals are known to coordinate with —O, —N, and —S atoms, therefore only these atoms were retained as metal ligands and other atoms such as main-chain or side-chain —C atoms were discarded. Water and other ligands bound to the metal containing protein were also explored as metal ligands. The coordination number of the metal and its preference for different residues were analyzed using local programs. Preference of each metal among main-chain or side-chain atoms in the case of different residues was also assessed. Variations in the M–L distances were checked to understand the flexibility of each metal. Solvent accessibility of ligand atom of all the residues coordinating with the metals was determined using the NACCESS software.[56] As defined in the NACCESS documentation, accessibility (relative accessibility) is the ratio between absolute solvent accessibility in the structure with respect to the peptide Gly–Gly–X–Gly–Gly.

Structure template generation

Metals can have varied coordination number based on their oxidation state. However, analysis of known metal-binding sites suggested that the most preferred site consist of three or four amino acid residue ligands, whereas water molecules satisfy additional coordination bonds. Thus, 3D-templates were generated consisting of either three or four residues. Analysis of the known metal-binding sites showed the preference of certain residues as M–L. Consequently, the structures having Asp, Glu, Cys, and His as the M–L were considered for template generation. Single template was generated for all the metals due to the similar nature of the binding sites of different metals. Each ligand was represented by its C^{α}, C^{β}, and a pseudo atom coordinate. The pseudo atoms were S^{γ} in the case of Cys, midpoint of carboxylate oxygens for Asp and Glu and midpoint of $N^{\delta 1}$ and $C^{\delta 2}$ for His. Structural templates were defined by certain geometrical parameters as follows.

Geometrical parameters that were considered included:

(a) distances between C^{α} atoms of the ligand residues;
(b) distances between their C^{β} atoms;
(c) angle between the C^{α} and C^{β} planes;
(d) distances between the pseudo atoms of the ligand residues;
(e) volume of the largest sphere that can be inscribed within the C^{β} atoms triangle (triads) or quadrilateral (tetrads); and
(f) distance of the proposed position of the metal from the possible ligand atoms.

The PDBselect90 list of Zn^{2+} atoms was taken as the training set for calculating the variations of above parameters in the metal-binding sites. The minimum and maximum values of distances, angle and volume as mentioned in (a)–(f) that constituted the ranges of values (Table II). To calculate angle between the C^{α} and C^{β} planes for tetrad sites, a least square plane was determined for four C^{α} (or C^{β}) atoms of the ligands.

For prediction of metal-binding site, every structure was scanned for the presence of Asp, Glu, Cys, and His residues. All the permutations of these residues were then assessed for calculating the values of the six parameters. A combination of three or four residues for which values of all the six parameters were within the range obtained from training set was flagged as a putative metal-binding site.

Template validation

Having derived templates for metal-binding sites, testing was carried out using structures from PDBselect90 list obtained from http://swift.cmbi.kun.nl/whatif/select. The list consists of 3165 structures that are less than 90% identical to each other and derived from the Aug. 17, 2006 release of the PDB. The structures in the list possessed *R*-factor $\leq$0.21 and resolution $\leq$2.0 Å. A total of 571 structures that contained bound Ca, Cu, Fe, Mg, and Mn atoms constituted the test set structures, and did not have any structure common to the training set. A subset of PDBselect90 was derived that did not contain any apo- or holo- metal-binding proteins. From this list, 1000 structures were randomly chosen to form the negative data set for cross-validation.

Template matching

The structural templates generated from the training dataset were used to predict metal-binding sites in all the 31,059 structures in the PDB database (April 2005 release). Matched residue groups were checked for the volume encompassed by their chelating side-chains. The three-residue (triads) or four-residue (tetrads) templates that meet the volume range were then checked for the

DOI 10.1002/prot

orientation of the ligand atoms. Since in the case of Asp and Glu, both the carboxylate oxygen can donate electrons to metal, both the atoms were explored as possible ligands. Similarly, in the case of His both of the imidazole —N atoms were examined for possible ligand. The prospective ligands were identified as the ones that matched the allowed M–L distances. To calculate M–L distance metal was placed at each grid point in a 2/2 grid around the center of the pseudo atom triangle. To get the coordinates of the points in the 2/2 grid, the incenter of the pseudo-atom triangle (or quadrilateral) was first determined. Coordinates of the grid points were then determined by adding or subtracting 1 from *x*, *y*, and *z* coordinates (one at a time) of the incenter. The sites that qualified this criterion were then predicted as metal-binding site. In the present study, each of the predicted sites, obtained upon scanning complete PDB database was analyzed manually to identify true and false positive sites.

ACKNOWLEDGMENTS

We acknowledge the overall support of S.E. Hasnain and J. Gowrishankar in this work. We would like to thank N. Srinivasan and D. Mohanty for their valuable suggestions and encouraging discussions. Yashaswini Kandan is acknowledged for useful ideas. K.B. is a CSIR Senior Research Fellow. S.C.M. is a Wellcome Trust International Senior Research Fellow.

REFERENCES

1. Rawlings ND, Barrett AJ. Evolutionary families of metallopeptidases. Methods Enzymol 1995;248:183–228.
2. Chakrabarti P. Geometry of interaction of metal ions with sulfur-containing ligands in protein structures. Biochemistry 1989;28: 6081–6085.
3. Chakrabarti P. Geometry of interaction of metal ions with histidine residues in protein structures. Protein Eng 1990;4:57–63.
4. Chakrabarti P. Interaction of metal ions with carboxylic and carboxamide groups in protein structures. Protein Eng 1990;4:49–56.
5. Chakrabarti P. Systematics in the interaction of metal ions with the main-chain carbonyl group in protein structures. Biochemistry 1990;29:651–658.
6. Chakrabarti P. Conformational analysis of carboxylate and carboxamide side-chains bound to cations. J Mol Biol 1994;239:306–314.
7. Harding MM. The geometry of metal–ligand interactions relevant to proteins. Acta Crystallogr 1999;D55:1432–1443.
8. Harding MM. The geometry of metal–ligand interactions relevant to proteins. II. Angles at the metal atom, additional weak metal–donor interactions. Acta Crystallogr 2000;D56:857–867.
9. Harding MM. Geometry of metal–ligand interactions in proteins. Acta Crystallogr 2001;D57:401–411.
10. Harding MM. Metal–ligand geometry relevant to proteins and in proteins: sodium and potassium. Acta Crystallogr 2002;D58:872–874.
11. Harding MM. The architecture of metal coordination groups in proteins. Acta Crystallogr 2004;D60:849–859.
12. Babor M, Greenblatt HM, Edelman M, Sobolev V. Flexibility of metal binding sites in proteins on a database scale. Proteins 2005;59:221–230.
13. Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer JA, Pique ME. MDB: the metalloprotein database and browser at the Scripps research institute. Nucl Acid Res 2002;30:379–382.
14. Smith SB, Hieftje GM. A new background-correction method for atomic absorption spectrometry. Appl Spectrosc 1983;37:419–424.
15. Koningsberger DC, Prins R. X-ray absorption: principles, Applications, techniques of EXAFS, SEXAFS and XANES, New York: Wiley; 1988.
16. Harding MM. Small revisions to predicted distances around metal sites in proteins. Acta Crystallogr 2006;D62:678–682.
17. Campos-Olivas R, Bruix M, Santoro J, Lacadena J, Martinez del Pozo A, Gavilanes JG, Rico M. NMR solution structure of the antifungal protein from *Aspergillus giganteus*: evidence for cysteine pairing isomerism. Biochemistry 1995;34:3009–3021.
18. Nielson KB, Atkin CL, Winge DR. Distinct metal-binding configurations in metallothionein. J Biol Chem 1985;260:5342–5350.
19. Peterson CW, Narula SS, Armitage IM. 3D solution structure of copper and silver-substituted yeast metallothioneins. FEBS Lett 1996;379:85–93.
20. Pasternak A, Kaplan J, Lear JD, Degrado WF. Proton and metal ion-dependent assembly of a model diiron protein. Protein Sci 2001;10:958–969.
21. Robien MA, Bosch J, Buckner FS, Van Voorhis WC, Worthey EA, Myler P, Mehlin C, Boni EE, Kalyuzhniy O, Anderson L, Lauricella A, Gulde S, Luft JR, DeTitta G, Caruthers JM, Hodgson KO, Soltis M, Zucker F, Verlinde CL, Merritt EA, Schoenfeld LW, Hol WG. Crystal structure of glyceraldehyde-3-phosphate dehydrogenase from *Plasmodium falciparum* at 2.25 Å resolution reveals intriguing extra electron density in the active site. Proteins 2006;62:570–577.
22. Sirover MA. New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase Biochim Biophys Acta 1999;1432:159–184.
23. Korndorfer I, Steipe B, Huber R, Tomschy A, Jaenicke R. The crystal structure of holo-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima* at 2.5 A resolution. J Mol Biol 1995;246:511–521.
24. Suresh S, Bressi JC, Kennedy KJ, Verlinde CL, Gelb MH, Hol WG. Conformational changes in *Leishmania mexicana* glyceraldehyde-3-phosphate dehydrogenase induced by designed inhibitors. J Mol Biol 2001;309:423–435.
25. Pavao F, Castilho MS, Pupo MT, Dias RL, Correa AG, Fernandes JB, da Silva MF, Mafezoli J, Vieira PC, Oliva G. Structure of *Trypanosoma cruzi* glycosomal glyceraldehyde-3-phosphate dehydrogenase complexed with chalepin, a natural product inhibitor, at 1.95 A resolution. FEBS Lett 2002;520:13–17.
26. Vellieux FM, Hajdu J, Verlinde CL, Groendijk H, Read RJ, Greenhough TJ, Campbell JW, Kalk KH, Littlechild JA, Watson HC, Hol WGJ. Structure of glycosomal glyceraldehyde-3-phosphate dehydrogenase from *Trypanosoma brucei* determined from Laue data. Proc Natl Acad Sci USA 1993;90:2355–2359.
27. Murthy MR, Garavito RM, Johnson JE, Rossmann MG. Structure of lobster apo-D-glyceraldehyde-3-phosphate dehydrogenase at 3.0 Å resolution. J Mol Biol 1980;138:859–872.
28. Keleti T. Zn in yeast D-glyceraldehyde-3-phosphate dehydrogenase. Biochem Biophys Res Commun 1966;22:640–643.
29. Oganesyan V, Oganesyan N, Adams PD, Jancarik J, Yokota HA, Kim R, Kim SH. Crystal structure of the "PhoU-like" phosphate uptake regulator from *Aquifex aeolicus*. J Bacteriol 2005;187:4238–4244.
30. Liu J, Lou Y, Yokota H, Adams PD, Kim R, Kim SH. Crystal structure of a PhoU protein homologue: a new class of metalloprotein containing multinuclear iron clusters. J Biol Chem 2005;280:15960–15966.

31. Genschik P, Drabikowski K, Filipowicz W. Characterization of the *Escherichia coli* RNA 3′-terminal phosphate cyclase and its sigma54-regulated operon. J Biol Chem 1998;273:25516–25526.
32. Okada C, Maegawa Y, Yao M, Tanaka I. Crystal structure of an RtcB homolog protein (PH1602-extein protein) from *Pyrococcus horikoshii* reveals a novel fold. Proteins 2006;63:1119–1122.
33. Carr S, Walker D, James R, Kleanthous C, Hemmings AM. Inhibition of a ribosome-inactivating ribonuclease: the crystal structure of the cytotoxic domain of colicin E3 in complex with its immunity protein. Structure 2000;8:949–960.
34. Sun W, Li G, Nicholson AW. Mutational analysis of the nuclease domain of *Escherichia coli* ribonuclease III. Identification of conserved acidic residues that are important for catalytic function in vitro. Biochemistry 2004;43:13054–13062.
35. Walker D, Lancaster L, James R, Kleanthous C. Identification of the catalytic motif of the microbial ribosome inactivating cytotoxin colicin E3. Protein Sci 2004;13:1603–1611.
36. Stubbs HJ, Brasch DJ, Emerson GW, Sullivan PA. Hydrolase and transferase activities of the beta-1,3-exoglucanase of *Candida albicans*. Eur J Biochem 1999;263:889–895.
37. Sakon J, Adney WS, Himmel ME, Thomas SR, Karplus PA. Crystal structure of thermostable family 5 endocellulase E1 from *Acidothermus cellulolyticus* in complex with cellotetraose. Biochemistry 1996;35:10648–10660.
38. Cutfield SM, Davies GJ, Murshudov G, Anderson BF, Moody PC, Sullivan PA, Cutfield JF. The structure of the exo-beta-(1,3)-glucanase from *Candida albicans* in native and bound forms: relationship between a pocket and groove in family 5 glycosyl hydrolases. J Mol Biol 1999;294:771–783.
39. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1227.
40. Graille M, Quevillon-Cheruel S, Leulliot N, Zhou CZ, de la Sierra Gallay IL, Jacquamet L, Ferrer JL, Liger D, Poupon A, Janin J, van Tilbeurgh H. Crystal structure of the YDR533c *S. cerevisiae* protein, a class II member of the Hsp31 family. Structure 2004;12:839–847.
41. Quigley PM, Korotkov K, Baneyx F, Hol WG. The 1.6-Å crystal structure of the class of chaperones represented by *Escherichia coli* Hsp31 reveals a putative catalytic triad. Proc Natl Acad Sci USA 2003;100:3137–3142.
42. Zhao Y, Liu D, Kaluarachchi WD, Bellamy HD, White MA, Fox RO. The crystal structure of *Escherichia coli* heat shock protein YedU reveals three potential catalytic active sites. Protein Sci 2003;12:2303–2311.
43. Malki A, Caldas T, Abdallah J, Kern R, Eckey V, Kim SJ, Cha SS, Mori H, Richarme G. Peptidase activity of the *Escherichia coli* Hsp31 chaperone. J Biol Chem 2005;280:14420–14426.
44. Christianson DW, Mangani S, Shoham G, Lipscomb WN. Binding of D-phenylalanine and D-tyrosine to carboxypeptidase A. J Biol Chem 1989;264:12849–12853.
45. Que L Jr. One motif—many different reactions. Nat Struct Biol 2000;7:182–184.
46. Lukatela G, Krauss N, Theis K, Selmer T, Gieselmann V, von Figura K, Saenger W. Crystal structure of human arylsulfatase A. The aldehyde function and the metal ion at the active site suggest a novel mechanism for sulfate ester hydrolysis. Biochemistry 1998;37:3654–3664.
47. Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, Jones DT. Predicting metal-binding site residues in low-resolution structural models. J Mol Biol 2004;342:307–320.
48. Wangikar PP, Tendulkar AV, Ramya S, Mali DN, Sarawagi S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. J Mol Biol 2003;326:955–978.
49. Kleywegt GJ. Recognition of spatial motifs in protein structures. J Mol Biol 1999;285:1887–1897.
50. Schymkowitz JWH, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. Prediction of water and metal-binding sites and their affinities by using the Fold-X force field. Proc Natl Acad Sci USA 2005;102:10147–10152.
51. Polacco BJ, Babbitt PC. Automated discovery of 3D motifs for protein function annotation. Bioinformatics 2006;22:723–730.
52. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using sequence-to-structure-to-function paradigm with application to glutaredoxin/thioredoxin and T_1 ribonucleases. J Mol Biol 1998;281:949–968.
53. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of the proteins exhibiting the glutaredoxin/thioredoxin disulphide oxidoreductase activity. J Mol Biol 1998;282:703–711.
54. Kuhn D, Weskamp N, Schmitt S, Hullermeier E, Klebe G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. J Mol Biol 2006;359:1023–1044.
55. Lin K-L, Yang C-H, Chung I-F, Huang C-D, Yang Y-S. Protein metal binding site residue prediction based on neural networks. Int J Neural Syst 2005;15:71–8.
56. Hubbard SL, Thornton JM. 'NACCESS', computer program. Department of Biochemistry and Molecular Biology, University College London; 1993.

 DOI 10.1002/prot