WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Prediction of transition metal-binding sites from apo protein structures

Mariana Babor,[1] Sergey Gerzon,[1] Barak Raveh,[2,3] Vladimir Sobolev,[1]* and Marvin Edelman[1]*

[1] Department of Plant Sciences, Weizmann Institute of Science, Rehovot, Israel

[2] Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, Israel

[3] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel

ABSTRACT

Metal ions are crucial for protein function. They participate in enzyme catalysis, play regulatory roles, and help maintain protein structure. Current tools for predicting metal–protein interactions are based on proteins crystallized with their metal ions present (holo forms). However, a majority of resolved structures are free of metal ions (apo forms). Moreover, metal binding is a dynamic process, often involving conformational rearrangement of the binding pocket. Thus, effective predictions need to be based on the structure of the apo state. Here, we report an approach that identifies transition metal-binding sites in apo forms with a resulting selectivity >95%. Applying the approach to apo forms in the Protein Data Bank and structural genomics initiative identifies a large number of previously unknown, putative metal-binding sites, and their amino acid residues, in some cases providing a first clue to the function of the protein.

Proteins 2008; 70:208–217.

Key words: *structural bioinformatics; structural genomics; molecular modeling; structure prediction; protein function.*

INTRODUCTION

Currently, about 20 novel protein structures are resolved each week by the structural genomics initiative (SGI), a worldwide effort having as one of its goals the creation of a catalog of all protein folds. Functional information for SGI targets is often limited or nonexistent; thus, there is a growing need for procedures to deduce the information directly from the resolved structure.[1] In such instances, initial clues to biochemical function can be sought from ligands and cofactors that often accompany the protein during crystallization. No cofactor group is more prevalent than metal ions, which play crucial roles in enzyme catalysis, molecular regulation, and structure stability. The problem is that a large fraction of metal-binding proteins are resolved in the Protein Data Bank (PDB[2]) in a prebound (or "apo") state with respect to their metal ion cofactors.

Surprisingly, among the large number of web servers and databases developed to assist in the annotation process, only a few are devoted to analysis and prediction of metal–protein interactions: MDB[3] and MSDsite,[4] which describe metal-binding geometries and ligand preferences of metalloprotein structures; a sequence comparison approach for identifying metalloprotein-binding site residues;[5–7] MetSite,[8] which assesses the likelihood of a given protein residue to be a metal ligand by considering the extent of its conservation among homologous proteins; and an extension of FoldX to predict the 3D coordinates of a bound metal and the free energy of binding.[9] All of these tools are based, and statistically validated, on information derived from holo forms containing the metal.

However, proteins are highly dynamic molecules. Recently, we analyzed the frequency and type of structural rearrangements that occur upon metal binding on a database scale.[10] We found that the rearrangements are mostly restricted to first and second shell residues, with backbone shifts occurring in less than 15% of cases while side chain rearrangements occur in more than 40% of first shell residues. Coordination numbers for the common metal ions in PDB mostly range from 3 to 10, with the coordinating groups being amino acids and water molecules. We observed that the vast majority of metal ion-binding sites with two or fewer amino acid residues are located on the surface of the protein and have no assigned function.[10] On the other hand, zinc[11] and other transition metal ions are mostly bound by three or more amino acid residues at catalytic, cocatalytic, or structural sites. Simplifying matters, the bound amino acids are almost always limited to four amino acid types—Cys, His, Glu, and Asp[4,11] (referred to hereafter as "CHED") that ligate the metal ion through polar side chain atoms.[12] Our findings showed that in more than 90% of such binding sites, at most one side

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

*Correspondence to: Vladimir Sobolev or Marvin Edelman, Plant Sciences Department, Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: vladimir.sobolev@weizmann.ac.il; marvin.edelman@weizmann.ac.il

Received 9 January 2007; Revised 27 March 2007; Accepted 24 April 2007

Published online 26 July 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21587

chain changed rotamer position as a result of metal binding.[10] Thus, while significant rearrangements occur upon metal binding, part of the binding site is already structured in the apo state, and transitions can be approximated by the rearrangement of a single ligand.

With this as background, we set out to develop an algorithm for predicting transition metal-binding sites in apo proteins, the structural state for which prediction is functionally useful. A major feature of the algorithm is that, in the first step (geometric search), structural rearrangements upon metal binding are taken into account. This is crucial for attaining a high level of sensitivity with apo structures. The zinc ion is the most abundant metal ion in PDB structures; indeed, it interacts with about 10% of human proteins.[13] Therefore, the zinc ion was employed as template during work up of the algorithm, and the results proved applicable for transition metals in general. Here, we demonstrate a selectivity of the algorithm greater than 95% with a sensitivity approaching 80% and show its applicability to SGI-resolved structures.

METHODS

Extracting metal-binding sites from the PDB

All X-ray files with a resolution better than 2.5 Å (about 80% of all the X-ray structures), and having proteins that bind metal ions, were extracted from the PDB. Distance cutoffs for ligands binding metal ions were defined as in the MDB metallodatabase[3] and set at 2.7 Å between protein and metal atoms. Backbone nitrogen atoms and nitrogen atoms from basic side chains were not considered as ligands, as their role in metal ion binding is controversial.[14] Carboxylysines were excluded as well as they ligate transition metals infrequently in the PDB. Only sites containing three or more amino acid ligands were scored.

Holo protein sets

Nonredundant datasets of polypeptide chains binding Zn, Co, Ni, Fe, Cu, or Mn ions were created. For each transition metal type, chains free of other metal types were extracted (i.e., PDB entries that include several bound transition metal ions in the structure, such as entry 2occ for cytochrome C oxidase, were not included). Chains sharing more than 30% sequence identity were grouped and a representative retained using the PISCES server.[15] Binuclear sites have two metal ions with two or more ligands in contact. Contact is defined according to CSU,[16] with the distance cutoff set to 4 Å. Separate binding sites within the same chain lack contacting ligands. Lists of chains comprising the holo protein datasets are provided in Dataset 1, Supplementary Material. The sets consist of 197, 19, 20, 33, 17, and 63 chains for Zn, Co, Ni, Fe, Cu and Mn, respectively.

Apo protein sets

Sequence comparisons were performed between each holo form binding a metal ion and all members of its SCOP family[17] to search for an apo protein partner. Apo/holo pairs sharing 95% or more sequence identity within the SCOP domain containing the metal-binding site were retained. The residues in a putative apo partner corresponding to those binding the metal ion in the holo form were then identified. If the residues in the putative apo form lack metal ion contacts (as defined by CSU[16]) the apo structure was accepted as a partner for the holo protein, and the pair retained. In this nontrivial step, it was necessary to account for modified amino acids, atoms not resolved in the X-ray structures, differences in PDB format, and any gaps in alignment. Finally, pairs with holo or apo chains sharing more than 30% sequence identity were grouped, and a representative retained using the PISCES server.[15] Lists of chains comprising the apo protein datasets are provided in Dataset 2, Supplementary Material. The sets consist of 27, 8, 8, 6, 7, and 26 chains for Zn, Co, Ni, Fe, Cu, and Mn, respectively.

Structural genomics set

The Protein Structure Initiative dataset (http://targetdb.pdb.org/target_files) contained 2673 protein chains. Chains of resolution better than 2.5 Å, and having no bound hetero groups, were extracted (614 chains). Among these, chains with sequence identity >30% were grouped and a representative retained using the PISCES server.[15] This resulted in a final nonredundant dataset of 230 apo chains (Dataset 3, Supplementary Material).

Geometric search

For a given structure, all sets of three amino acid residues (*triads*) from the four common ligands for zinc (Cys, His, Glu, and Asp) were retrieved whose atomic distances d_1, d_2, and d_3 from separate residues satisfy the cutoff criterion,

$$d_1 < 4.7\text{Å}, d_2 < 5.1\text{Å}, d_3 < 5.4\text{Å} \quad (1)$$

These distances were chosen following the inspection of the retained apo and holo protein structures. As ligating atoms, we considered side chain oxygen from Asp and Glu, sulfur from Cys, and ring atoms ND, NE, CD, CE from His (the carbon atoms were included due to possible mislabeling[18]). If one or two of the interligand distances were not initially satisfied, side chains for these residues were built, one at a time, using a backbone-independent rotamer library.[19] If no clashes between residues ("external" clashing) were observed, and the

interatomic distances now satisfied the cutoffs, then the built-up triad was retained as well. External clashing is considered to occur if the distance between an atom of a built-up residue and any other atom of the protein is, $<0.7 \times (R_a + R_b)$, where R_a and R_b are Van der Waals radii of atoms *a* and *b*, as defined in Bondi.[20] The value 0.7 was determined by statistically analyzing interatomic distances in PDB structures. When both atoms were Cys sulfurs, then clashing was set for $\text{dist}_{a,b} < 1.8$ Å. Internal clashing within a built-up residue was checked for all pairs of atoms separated by four or more bonds. The distance criteria were the same as those used for external clashing. A *site* is defined as a single triad or multiple triads that share at least one residue between two or more triads. A *true site* is one that includes at least a *partially true triad* (see below). All other sites are defined as false.

Decision tree and support vector machine classifiers

The training set (Dataset 1, Supplementary Material) consists of 125 chains that include 139 binding sites and 367 triads. *True*, *partially true*, and *false* triads contain, respectively, all three, some or no correct (i.e., experimentally determined) ligand residues. Machine learning was performed using only true and false triads. A decision tree was created based on a CART tree algorithm,[21] using Matlab 7.0.4 implementation, Treefit, and Gini's diversity index as a split criterion. The tree was pruned using a 10-fold cross-validation test within the tree test function.

A second classifier was based on support vector machines.[22] For implementation, mySVM Version 2.1.4 was used with the radial basis function as kernel.[23] The values of γ and the complexity parameter *C* were optimized using 10-fold cross-validation tests and were eventually set to 0.2 and 1, respectively.

Sensitivity and selectivity

Results at the site and residue levels were evaluated for the following parameters:

Sensitivity = No. true positives/(No. true positives + No. false negatives);
Selectivity = No. true positives/(No. true positives + No. false positives).

RESULTS AND DISCUSSION

Zinc-binding sites

The search procedure is based on a geometric definition of the prebound state, followed by a machine-learning filtering step to reduce false positives. For the geometric step, we searched for a 3D constellation of three CHED residues (similar to searching for a catalytic triad[24]), whose metal-ligating atoms satisfy distance criteria large enough to allow for small backbone shifts, with at most one side chain rotating. Applying this geometric step to a high-resolution, nonredundant dataset of apo proteins for which corresponding Zn-binding holo partners exist in the PDB, we identified 27 of 28 sites (i.e., 96% *sensitivity*). However, along with these correct identifications, an additional 20 putatively false-positive sites were scored (i.e., *selectivity* of 57%).

Filtering procedures were therefore added to maximize selectivity. A simple *mild filter* was created based on our observation that sites composed of a relatively large number of triads tend to be true (defined as have at least one partially true triad). Specifically, in cases where a putative site contains five or more triads, all other putative sites with three or fewer triads are discarded. The mild filter retained the 96% sensitivity level for Zn apo sites while increasing selectivity to 71%. These percentages were mirrored in a larger set of 197 holo proteins containing 215 binding sites, where the mild filter essentially retained the prefiltration sensitivity (99% before, 98% after) and upped selectivity from 49% to 70%.

A more elaborate *stringent filter* made use of two machine-learning techniques: a decision-tree classifier and a support vector machine. The decision tree (Fig. 1, Supplementary Material) considers: *Number of predicted sites*—number of sites predicted by the geometric search; *Minimum residue-position frequency*—the frequencies of all positions within a site are first tallied, then the position with the lowest frequency is scored for each triad; *Amino acid composition*—the number of acidic, His and Cys residues present in a given triad; *Conservation score*—the HSSP database[25] was used to obtain a multiple sequence alignment. The conservation score was then calculated by a modification of Mirny and Shakhnovich,[26] in which CHED residues replaced the acidic group. Median and maximum conservation scores were defined as the intermediate and highest conservation score values among the three residues of the triad; *Hydrogen bond contact surface area*—contact surface area between each residue in a triad and its neighbors was calculated using CSU software[16] and a distance cutoff of 3.5 Å, and summed. Median and maximum areas were defined as the intermediate and highest hydrogen bonding contacting surface values among the three residues of the triad. The support vector machine includes in addition the number of triads per predicted site and relative solvent accessible surface.

The defining feature of the algorithm is in the geometric search, which is based on an analysis of apo forms and structural rearrangements of the metal-binding sites. This permits a high level of sensitivity. Once this is in place, a larger holo dataset can be used for priming the stringent filter to reduce the number of false positives. The filter was trained using about two thirds of the Zn

 DOI 10.1002/prot

holo dataset (125 proteins, containing 139 binding sites) and subjected to a 10-fold cross-validation test. The resulting filter was tested on the remaining one-third of the dataset (72 proteins, containing 76 binding sites). A triad was retained if accepted by either the decision tree or support vector machine classifier. This approach led to a moderate reduction in sensitivity to 91% but a marked increase in selectivity to 97%. Application to the apo dataset mirrored these results (sensitivity, 79%; selectivity, 100%).

The final search procedure averaged ~3 correct residues for every false one, with the average number of residues per correctly predicted binding site being 4.1 (95% confidence level, 3.6–4.6) for mononuclear sites and 7.3 (95% confidence level, 6.8–7.8) for binuclear ones. Thus, predicted sites are realistically compact and the approach can distinguish between mono- and multinuclear sites in a large majority of cases ($P < 0.0001$).

Extension to other transition metals

Dominance of side chain carboxylates, sulfhydryls, and imidazole groups in the coordination of transition metals is a general characteristic. Accordingly, the performance of our Zn-modeled approach was evaluated for apo and holo sets of additional transition metal ions (Fe, Cu, Co, Ni, Mn). Figure 1 summarizes the statistics for apo and holo structures. It can be seen that the high level of selectivity obtained with the stringent filter for Zn extends in general to the other transition metals. A comparison of mild and stringent filter results show that selectivity comes at the expense of sensitivity. The pattern and percentages for apo structures are mirrored in the larger datasets for holo structures.

Importantly, for the transition metals analyzed, the approach recognized on average >90% of the correct residues in both apo and holo sets, with averages of 1.5 and 0.9 extra residues, respectively, in correctly predicted sites (Table I). We define these extra residues as false positives; however, we note that catalytic residues sometimes contribute to the extra residue "noise."

Comparison to other methods

Two recent studies relate to metal-binding site prediction.[8,9] Although both achieve sensitivity levels comparable to ours, neither matches our selectivity level, nor are they geared to, or statistically tested for, prediction of metal-binding sites from apo structures. MetSite[8] shows selectivity values less than 50% of ours for each of the transition metals analyzed. This may be partially attributed to our approach that searches for a binding site motif consisting of a set of three residues with a particular spatial relationship, while MetSite searches for a single residue. Schymkowitz *et al.*,[9] using FoldX, report an average binding site sensitivity ("percentage predicted") of

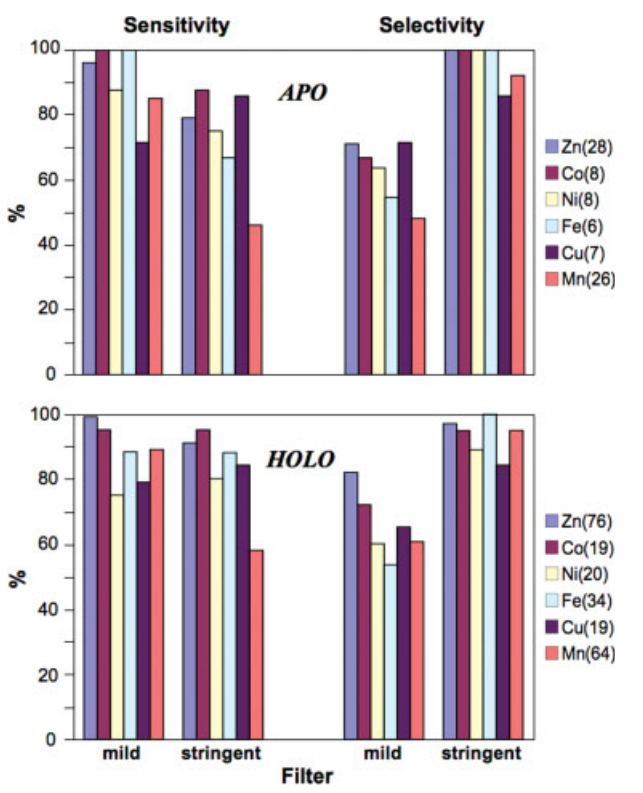

Figure 1

Sensitivity and selectivity predictions for transition metal ions. Nonredundant datasets of paired apo and holo proteins binding transition metals were created. A subset of 125 chains of Zn mononuclear-binding sites was used for training. Shown are the search results for test sets (numbers of protein chains in parentheses) following application of the mild and stringent filters. Sensitivity indicates the fraction of experimental-binding sites predicted; selectivity indicates the fraction of all predicted sites that are correct. Mild filtration allows high sensitivity while stringent filtration allows high selectivity. For all metal types, the average stringent selectivity for apo and holo structures is 96 and 95%, respectively.

91% and selectivity (100 minus "percent overpredicted") of 82% for holo structures containing Zn, Cu, or Mn (calculated from the data in Table I of Ref. [9]). Using these values, we can calculate the number of false positive sites that were obtained ((No. crystallographic sites) (percentage predicted/100) (100/(100 − percent overpredicted) − 1): (277) (91/100) (100/(100 − 18) − 1) = 55. Thus, 55 false positive sites were predicted for the 124 protein files. On the other hand, using the results shown in Figure 1 obtained with the CHED algorithm: (159) (77/100) (100/100 − 5) − 1) = 6. Thus, 6 false positive sites are predicted for 152 proteins containing Zn, Cu, or Mn. The authors provide two apo cases in support of their view on "overprediction;" however, as their force field optimization step is very sensitive to small changes of position due to the electrostatic nature of the interactions,[9] and as it has been shown that apo-binding sites often undergo partial rearrangement upon

Table I
Percent of Correct Binding-Site Residues Recognized

	Experimental		Predicted (ligands/ion)			
Metal	No. of ions	Ligands/ion	Total	Correct	Extra	% of Experimental
(a) Holo sets						
Zn	106	3.3	3.8	3.1	0.7	94
Co	23	3.1	3.4	2.8	0.6	90
Ni	19	3.1	3.8	2.8	1.0	90
Fe	35	3.4	4.4	3.2	1.2	94
Cu	19	3.2	4.0	2.9	1.1	91
Mn	54	3.2	3.7	2.7	1.0	84
(b) Apo sets						
Zn	23	3.3	4.4	3.2	1.2	97
Co	7	3.9	5.4	3.6	1.8	92
Ni	6	3.5	5.5	3.3	2.2	94
Fe	4	3.0	3.3	2.8	0.5	93
Cu	6	3.5	4.3	3.0	1.3	86
Mn	12	3.4	5.0	2.8	2.2	82

All data shown are for true site residues. *Experimental* refers to information derived from the PDB as analyzed by LPC software,[16] using a distance cutoff of 2.7 Å. *Predicted (ligands/ion)* was calculated using the CHED algorithm following stringent filtration. *Correct* refers to the average number of ligands matching the experimental data in the PDB, while *extra* is the average number of false positive ligands/ion. *% of Experimental* is derived by dividing the *correct* column by the experimental *ligands/ion* column. (a) Holo sets: Datasets contain both mono and binuclear metal binding sites; therefore, data are normalized per ion. (b) Apo sets: Average values for ligands/ion were derived from the holo partner of the apo/holo pair.

metal binding,[10] it is not clear that these two apo cases are typical for their algorithm.

Promiscuous binding sites?

Stringent filtration resulted in a substantially reduced sensitivity for Mn, identifying only about half of the total Mn-binding sites. However, among the sites identified, selectivity was greater than 90% (Fig. 1). The low level of sensitivity may be due to the fact that Mn, the hardest of the transition metals analyzed, can often be accommodated in Mg-like-binding sites.[27] Amino acids that bind hard metal ions such as Mg and Ca are not limited to the four CHED residue types, and are poorly captured (~30 and ~10%, respectively) by the stringent filter. To study this, we analyzed a nonredundant dataset of 127 proteins containing 138-mg ion-binding sites (Dataset 4, Supplementary Material) and found that metal-binding sites predicted by our algorithm participate disproportionately (22 of 46 sites) in transition metal exchange in homologous PBD structures (defined by the VAST server[28]), in comparison to nonpredicted sites (20 of 92 sites). Thus, the stringent filter appears to be geared to identifying transition metal ion sites. The hard metal sites predicted by our approach may represent a promiscuous subset that accommodates transition metal substitutions.

It is well known that *in vitro*, the same binding site can be occupied by different metal ions under different conditions. In fact, it is unclear to what degree metal ions found *in vivo* are intrinsic to individual enzymes or are determined by the ambient levels of ions within the organism.[29] As such, the "physiological state" may be regarded as a dynamic feature that encompasses a number of static snapshots, represented by a binding site in a resolved PDB structure either empty or occupied by different metal ions. In line with this, our procedure, modeled on Zn ion characteristics, is suitable for transition metals in general. Essentially, our procedure relates to two questions: Does a protein structure have a transition metal-binding site; and if it does, where is the binding site location (set of ligating residues). However, our approach is unable to specify which transition metal ion is bound, nor its exact coordinates. Applications using force fields[9] can assist in resolving these issues.

Application to the structural genomics initiative

We applied our approach to protein structures from the SGI project. Frequently, little is known about the function of the protein targets chosen for high-throughput crystallization. Starting from a full list of SGI polypeptide chains, we analyzed a high resolution, nonredundant set of 230 crystallized proteins lacking heterogroups for potential transition metal-binding sites (Dataset 3, Supplementary Material). Thirty-three binding sites in 31 proteins were predicted. Structure, sequence, and literature searches provide support for the validity of about half the cases (Table II).

One of these, a putative glycerophosphoryl diester phosphodiesterase (GDPD; 1v8e) is illustrated in Figure 2. Even though no metal requirement is annotated for the GDPD family, some members (1xx1, 1ydy) contain a hard metal (Mg or Ca) in the active site. Structural alignment reveals that the three acidic residues ligating the bound metal in each case correspond to the CHED-predicted residues Glu40, Asp42, Glu100 for PDB file 1v8e (Table II). Thus, the GDPD family may require a hard metal as cofactor. However, significantly, divalent Fe ion inhibits GDPD in *Arabidopsis thaliana*,[45] indicating that the transition metals are also able to bind GDPDs, and suggesting that the active site of the GDPD family contains a promiscuous metal-ion-binding site.

A second case is that of Iaa-amino acid hydrolase from *Arabidopsis thaliana* (PDB entry 1xmb) illustrated in Figure 3. Supporting evidence for a transition metal ion site in 1xmb can be adduced from biochemical data, which shows a Mn-ion dependency for the *Arabidopsis* enzyme.[4] The CHED algorithm predicts eight residues ligating a metal ion (Asp91, Cys116, His118, His121, Glu151, Glu152, His176, Glu 347), suggesting a binuclear site. Structural alignment indicates that four of these residues (Cys116, His118, Glu152, His176) superimpose with the binuclear Ni-binding site of Yxep peptidase of *Bacillus subtilis* (1ysj), and three (Asp91, Glu152, His176)

DOI 10.1002/prot

Table II
Predicted Metal-Binding Sites in Apo Structures from the Structural Genomics Initiative

			Supporting data			
			Structural[b]			
PDB ID & chain	Description	Predicted binding site residues[a]	PDB ID	% homol.	Bound metal	Biochemical
1ilw_A	Hypothetical pyrazinamidase	**D_{52} H_{54} H_{71}** C_{72}	1im5	100	Zn	Zn activates[30]
1iv1_A	MECDP synthase	**D_8 H_{10} H_{42}**	1iv3	98	Mg	Mg, Mn, Zn
		D_8 H_{10} H_{42}	1knk	41	Mn	Required[31]
		D_8 H_{10} H_{42}	1h48	40	Zn	
1vr6_A	DAHP synthase	**C_{102} H_{272} E_{298} D_{309}**	1rzm	96	Cd	Transition metals activate[32]
		C_{102} H_{272} E_{298} D_{309}	1zco	48	Mn	
1wvi_A	Putative phosphatase	**D_{1009} D_{1206}** D_{1211}	1ydf	74	Mg	—
1o0x_A	Methionine aminopeptidase	**D_{100} H_{174}** H_{181} **E_{207}** H_{236} **E_{238}**	1mat	39	Co	Transition metals activate[33]
		D_{100} H_{174} H_{181} **E_{207}** H_{236} **E_{238}**	1r58	20	Mn	
		D_{100} H_{174} H_{181} **E_{207}** H_{236} **E_{238}**	1kq0	16	Zn	
1xmb_A	Iaa-aminoacid hydrolase	D_{91} **C_{116} H_{118}** H_{121} E_{151} **E_{152} H_{176}** E_{347}	1ysj	35	Ni	Mn activates[34]
		D_{91} C_{116} H_{118} H_{121} E_{151} **E_{152} H_{176}** E_{347}	1cg2	19	Zn	
1ujn_A	Dehydroquinate synthase	D_{125} **E_{173} H_{231}** H_{235} E_{238} **H_{247}**	1xag	34	Zn	Zn, Co, Mn
		D_{125} **E_{173} H_{231}** H_{235} E_{238} **H_{247}**	2bi4	21	Fe	Activate[35]
1oz9_A	Hypothetical protein Aq_1354	**H_{115} H_{119}** D_{124} **H_{125}** E_{126}	1xm5	30	Ni	—
		H_{115} H_{119} D_{124} **H_{125}** E_{126}	1qua	15	Zn	
1uan_A	Protein Tt1542	**H_{10} D_{13}** D_{74} D_{106} H_{108} D_{110} **H_{111}**	1q74	24	Zn	Zn metalloenzyme[36]
1xfk_A	Forminoglutamase	**H_{127} D_{157} H_{159} D_{161} D_{254}** E_{300}	1cev	22	Mn	Transition metals & Mg activate[37]
1v8e_A	Putative glycerophosphoryl diester phosphodiesterase	H_{13} **E_{40} D_{42}** H_{55} **E_{100}**	1ydy	20	Ca	Ca activates[38], Fe inhibits[45]
		H_{13} **E_{40} D_{42}** H_{55} **E_{100}**	1xx1	19	Mg	
1xm7_A	Hypothetical protein Aq_1665	**D_7 H_9 D_{50} H_{111}**	1g5b	20	Mn	—
		D_7 H_9 D_{50} H_{111}	1su1	16	Zn	
		D_7 H_9 D_{50} H_{111}	1v73	15	Ca	
		D_7 H_9 **D_{50} H_{111}**	1war	15	Fe	
		H_{143} C_{155} C_{157} C_{164}[c]	—	—	—	
1t70_A	Novel phosphatase	**D_8 E_{37}** H_{66} **H_{148} H_{173} H_{175}**	1t71	34	Fe	—
		H_{46} H_{70} D_{72}	—	—	—	
1vcm_A	CTP synthetase	C_{391} H_{469} H_{471} E_{506} H_{522}	—	—	—	Mg required; Zn, Cu inhibit[39]
1ybe_A	Nicotinate phosphoribosyl transferase	H_{249} D_{290} D_{314}	—	—	—	Mg activates[41]
1wkj_A	Nucleoside diphosphate kinase	H_{48} H_{115} D_{122} E_{126}	—	—	—	Metal activates[41]
1f89_A	Putative CN hydrolasee	H_{205} H_{237} E_{250}	—	—	—	—
1mk4_A	Hypothetical protein Yqjy	H_0 E_{48} H_{49}	—	—	—	—
1mkf_A	Chemokine binding protein	E_{31} H_{35} E_{41} E_{46}	—	—	—	—
1nye_A	Salt shock induced protein C	H_{76} C_{79} C_{145}	—	—	—	—
1ri6_A	Putative isomerase	H_{182} D_{235} H_{237}	—	—	—	—
1rz2_A	Hypothetical protein Ba4783	H_{140} C_{233} D_{234}	—	—	—	—
1s7j_A	Phenazine biosynthesis protein	E_{46} C_{71} H_{73} D_{200}	—	—	—	—
1sef_A	Hypothetical protein 1sef	E_{199} H_{201} H_{205}	—	—	—	—
1ufi_A	Cenp-B dimerization domain	H_{38} H_{41} D_{42} H_{45}	—	—	—	—
1v6t_A	Lactam utilization protein	D_8 E_{11} H_{38} H_{62} H_{111}	—	—	—	—
1x9g_A	Putative Mar1 ribonuclease	D_{19} E_{108} C_{112}	—	—	—	—
1y9w_A	Acetyltransferase	H_{86} E_{89} H_{116}	—	—	—	—
1olz_A	Semaphorin 4d	C_{482} H_{485} C_{491} D_{496} C_{499} C_{508}[c]	—	—	—	—
1xak_A	SARS Orf7a accessory protein	C_{20} C_{52} H_{58}[c]	—	—	—	—
1s4y_B	Erytroid differentiation protein	C_{11} C_{40} C_{44} C_{81} C_{113} C_{115}[c]	—	—	—	—

[a]Only stringent filter results are shown. Predicted ligand positions in structurally related metalloproteins within 2.7 Å of the metal ion are in bold. Pairwise superimposition was performed using MSD Secondary Structure Matching (http://www.ebi.ac.uk/msd-srv/ssm/).
[b]Related structures were identified using the following web servers: OCA (http://bip.weizmann.ac.il/oca-bin/ocamain); MSD (http://www.ebi.ac.uk/msd-srv/ssm/); VAST (http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml); BRENDA (http://www.brenda.uni-koeln.de). Ligand residues were determined using LPCCSU (http://ligin.weizmann.ac.il/lpccsu/), and % homology using LALIGN (http://www.ch.embnet.org/software/LALIGN_form.html).
[c]Binding sites contain Cys ligands that form S—S bonds in the apo state.

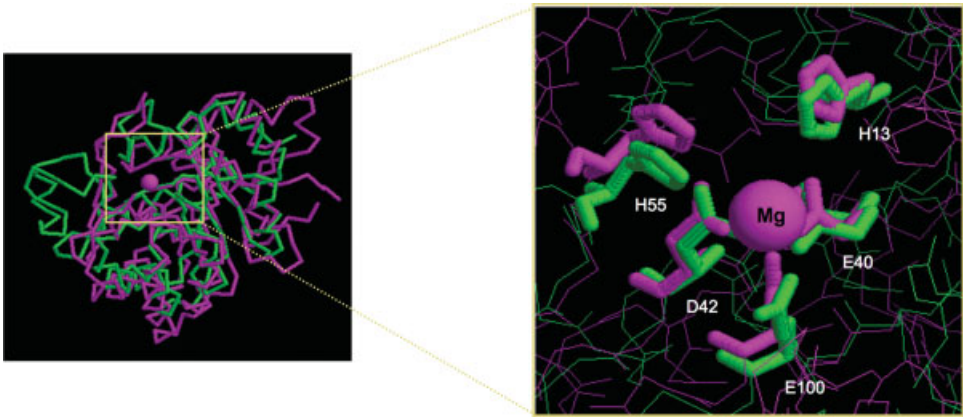

Figure 2

Prediction and structural validation for a putative glycerophosphoryl diester phosphodiesterase (GDPD). Left panel, structural alignment of the apo structure from Thermus thermophilus *Hb8 (1v8e, green) and spider toxin sphingomyelinase (SMaseD) (1xx1, magenta). Right panel, enlargement showing the five residues predicted by the stringent filter (H13, E40, D42, H55, E100; see Table II), and the Mg ion of SMaseD with its structurally aligned ligand residues and catalytic histidines.[42] Pfam analysis[43] indicates that 1v8e belongs to the GDPD family. SmaseD (1xx1), which shares 19% sequence identity with GDPDs,[44] possesses a conserved catalytic and metal-binding core.[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

with the binuclear Zn site of carboxypeptidase G2 (1cg2) of *Pseudomonas sp.* (Table II).

We can speculate on the role of Cys116 in the predicted binuclear site of Iaa-amino acid hydrolase. A blast search against the Uniprot database identified a related Zn-dependent protein from *Pyrococcus horikoshii* that has both carboxypeptidase and aminoacylase activity, with an active site Cys[46] corresponding to Cys116 of Figure 3. Subsequent research on D-aminoacylases revealed a subclass of enzymes that bind two Zn ions with widely different affinities.[47] Enzymatic activity is coordinated by an unusual bridging, active site cysteine that controls an activation/attenuation mechanism. The enzyme is activated by the metal ion at the high affinity site but inhibited by subsequent binding of the second metal ion at the low affinity site.[48] We analyzed the atomic interactions of the cysteine in peptidase 1ysj by LPC software[16] and found that this cysteine (Fig. 3) bridges the two Ni

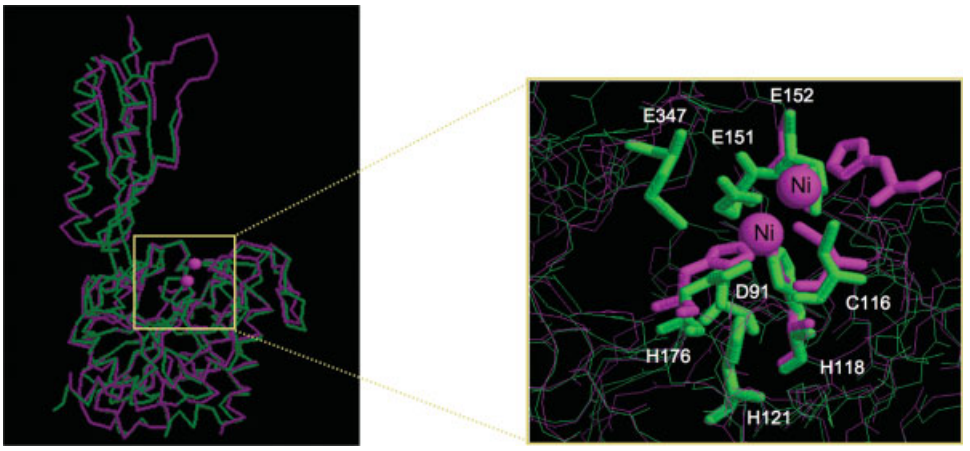

Figure 3

Structural alignment of Iaa-amino acid hydrolase from Arabidopsis thaliana *(1xmb, green) and Yxep peptidase from* Bacillus subtillis *(1ysj, magenta). Panel to the right shows an enlargement of the predicted binding site residues and the experimentally determined metals ions and their ligand residues. The CHED algorithm predicts a binuclear metal-binding site for 1xmb composed of eight residues (Asp91, Cys116, His118, His121, Glu151, Glu152, His176, Glu 347) (Table II). Structural alignment indicates that four of these residues (Cys116, His118, Glu152, His176) superimpose with the binuclear Ni-binding residues of PDB file 1ysj. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

DOI 10.1002/prot

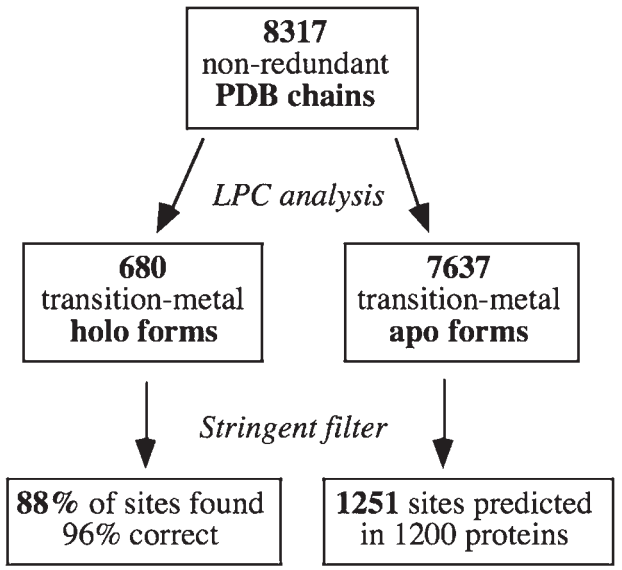

Figure 4

Prediction of transition metal-binding sites from apo structures in the PDB. The search procedure was applied to the full PDB, using a precompiled non-redundant list of 8317 polypeptide chains with resolution better than 2.5 Å and sequence identity <90%.[15] In this list, there are 680 proteins with 900 binding sites ligating transition metals (Zn, Co, Ni, Fe, Cu, Mn). The stringent filter found 88% of these sites with a selectivity of 96% (Table I, Supplementary Material). The remaining 7637 polypeptide chains were apo with respect to transition metals. The search procedure found 1251 putative binding sites distributed among 1200 proteins for this apo subset; i.e., about 16%. Searching all 8317 PDB entries took ~3 h of CPU on a 3 Ghz PC with 2 Gb of RAM.

ions in the active site. Having come full circle, we now wonder whether peptidase 1ysj and, by analogy, Iaa-amino acid hydrolase attenuate their metal center activity(ies) in a fashion similar to the D-aminoacylases. Further experimentation is needed.

Furthermore, CSU analysis[16] identifies two water molecules (atoms O1 and O2) that were seen to occupy the predicted binuclear site in the structure model of Iaa-amino acid hydrolase (PDB file 1xmb). The O2 atom is located 2.0, 2.5, and 2.9 Å from Cys116, Glu152, and His376, respectively. Electron density analysis indicated that this oxygen atom is actually part of oxidized Cys116. The O1 atom is located at 2.3 and 2.8 Å from His118 and His176, respectively. Even though the latter interaction distance is slightly larger than our cut off of 2.7 Å, others have accepted 3.0 Å for metal-protein interactions in crystallized structures.[40] The O1 atom is well ordered, with a B-factor (18.5) substantially lower than the backbone average of the protein (28.7) and of the ligating atoms His118NE2 (21.8) and His176NE2 (35.8). Moreover, inspection of the electron density map indicated a higher electron density peak for the O1 atom than expected for a water oxygen atom. Combined, these observations suggest that the O1 atom represents a metal ion present in the crystal structure but not identified as such. Thus, electron density data may provide a largely untapped source for analysis of metal ion binding in protein structures.

In summary, following direction by CHED prediction to a specific 3D location, homology searching in the structural and linear databases and evaluation of electron density data can sometimes reveal a wealth of information to decorate an otherwise bare apo protein structure.

However, sometimes no correlative or supporting information is available. This is the case for about half of the SGI predictions in Table II. These cases represent completely novel metal-binding site predictions. Based on the high level of selectivity of the CHED algorithm, we propose that our approach can indicate the likely presence and location of transition metal-binding sites even when there are neither related protein structures nor biochemical information about metal-binding capabilities available.

There are four predicted metal-binding sites in Table II with putative Cys ligands in disulfide-bonded form in the apo state. Several examples of proteins that can either coordinate metals or form disulfides have been described. For example, switching between Zn-bound versus disulfide-bonded forms controls the activation of Hsp33 chaperone[49] (1vzy). In addition, expulsion of Zn leads to disulfide bond formation in betaine-homocysteine methyltransferase[50] (1lt7, 1lt8). Since disulfides need to be reduced before they can coordinate metal, the level of surface-exposure needed to encounter a reductant may be a factor in the switch to metal-binding mode.

Application to the full PDB

Finally, we applied our search procedure for metal-binding sites to a high resolution, nonredundant set of crystallographic structures from the full PDB (Fig. 4). The stringent filter found 88% of the holo protein sites and predicted that 16% of the apo proteins are transition metal binders. This level of predicted binders is similar to that found for the SGI dataset (Table II; see also Ref. [51]), and is clearly distinguished from the low level (4%) of false positives found for apo proteins (Fig. 1). The list of 1251 predicted novel transition-metal-binding sites in apo structures from a nonredundant list of 8317 PDB polypeptide chains[15] is provided in Table I, Supplementary Material.

CHED web server

An interactive web server for predicting transition metal-binding sites in protein structures based on our approach has been created and is available at http://ligin.weizmann.ac.il/ched. The server produces a graphical presentation of the predicted binding site(s) for the geomet-

ric search, and following either mild or stringent filtration. PDB or user-generated structures can be submitted.

ACKNOWLEDGMENTS

We thank Felix Frolow for guiding us in the electron density analysis, and Deborah Fass for valuable discussions. We also thank Harry Greenblatt, Eran Eyal, Vladimir Potapov and Ronen Levy for assistance during the study. Supported in part by the Avron-Wilstatter Minerva Center for Research in Photosynthesis.

REFERENCES

1. Friedberg I, Jambon M, Godzik A. New avenues in protein function prediction. Protein Sci 2006;15:1527–1529.
2. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. Protein data bank—computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
3. Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer JA, Pique ME. MDB: the metalloprotein database and browser at the scripps research institute. Nucleic Acids Res 2002;30:379–382.
4. Golovin A, Dimitropoulos D, Oldfield T, Rachedi A, Henrick K. MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. Proteins 2005;58:190–199.
5. Andreini C, Bertini I, Rosato A. A hint to search for metalloproteins in gene banks. Bioinformatics 2004;20:1373–1380.
6. Lin CT, Lin KL, Yang CH, Chung IF, Huang CD, Yang YS. Protein metal binding residue prediction based on neural networks. Int J Neural Syst 2005;15:71–84.
7. Passerini A, Punta M, Ceroni A, Rost B, Frasconi P. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. Proteins 2006;65:305–316.
8. 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.
9. 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.
10. Babor M, Greenblatt HM, Edelman M, Sobolev V. Flexibility of metal binding sites in proteins on a database scale. Proteins 2005; 59:221–230.
11. Auld DS. Zinc coordination sphere in biochemical zinc sites. Biometals 2001;14:271–313.
12. Alberts IL, Nadassy K, Wodak SJ. Analysis of zinc binding sites in protein crystal structures. Protein Sci 1998;7:1700–1716.
13. Andreini C, Banci L, Bertini I, Rosato A. Counting the zinc-proteins encoded in the human genome. J Proteome Res 2006;5:196–201.
14. Vallee BL, Auld DS. Zinc coordination, function, and structure of zinc enzymes and other proteins. Biochem 1990;29:5647–5659.
15. Wang GL, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
16. Sobolev V, Eyal E, Gerzon S, Potapov V, Babor M, Prilusky J, Edelman M. SPACE: a suite of tools for protein structure prediction and analysis based on complementarity and environment. Nucleic Acids Res 2005;33:W39–W43.
17. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP—a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
18. McDonald IK, Thornton JM. The application of hydrogen-bonding analysis in X-ray crystallography to help orientate asparagines, glutamine and histidine side-chains. Protein Eng 1995;8:217–224.
19. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
20. Bondi A. Van der Waals volumes and radii. J Phys Chem 1964;68: 441–451.
21. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Boca Raton: CRC Press; 1998.
22. Chapelle O, Haffner P, Vapnik VN. Support vector machines for histogram-based image classification. IEEE Trans Neural Network 1999;10:1055–1064.
23. Scholkopf B, Smola AJ, Williamson RC, Barlett PL. New support vector algorithms. Neural Comput 2000;12:1207–1245.
24. Wallace AC, Laskowski RA, Thornton JM. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 1996;5:1001–1013.
25. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
26. Mirny L, Shakhnovich E. Evolutionary conservation of the folding nucleus. J Mol Biol 2001;308:123–129.
27. Bock CW, Kaufman-Katz A, Markham GD, Glusker JP. Manganese as a replacement for magnesium and zinc: functional comparison of the divalent ions. J Am Chem Soc 1999;121:7360–7372.
28. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
29. Oefner C, Douangamath A, D'Arcy A, Hafeli S, Mareque D, Mac Sweeney A, Padilla J, Pierau S, Schulz H, Thormann M, Wadman S, Dale GE. The 1.15 Å crystal structure of the *Staphylococcus aureus* methionyl-aminopeptidase and complexes with triazole based inhibitors. J Mol Biol 2003;332:13–21.
30. Du XL, Wang WR, Kim R, Yakota H, Nguyen H, Kim SH. Crystal structure and mechanism of catalysis of a pyrazinamidase from *Pyrococcus horikoshii*. Biochemistry 2001;40:14166–14172.
31. Kemp LE, Bond CS, Hunter WN. Structure of 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase: an essential enzyme for isoprenoid biosynthesis and target for antimicrobial drug development. Proc Natl Acad Sci USA 2002;99:6591–6596.
32. Wu J, Howe DL, Woodard RW. Thermotoga maritima 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase—the ancestral eubacterial DAHP synthase? J Biol Chem 2003;278:27525–27531.
33. Oefner C, Douangamath A, D'Arcy A, Hafeli S, Mareque D, Mac Sweeney A, Padilla J, Pierau S, Schulz H, Thormann M, Wadman S, Dale Ge. The 1.15 angstrom crystal structure of the *Staphylococcus aureus* methionyl-aminopeptidase and complexes with triazole based inhibitors. J Mol Biol 2003;332:13–21.
34. LeClere S, Tellez R, Rampey RA, Matsuda SPT, Bartel B. Characterization of a family of IAA-amino acid conjugate hydrolases from Arabidopsis. J Biol Chem 2002;277:20446–20452.
35. Lambert JM, Boocock MR, Coggins JR. The 3-dehydroquinate synthase activity of the pentafunctional arom enzyme complex of neurospora-crassa is Zn^{2+}-dependent. Biochem J 1985;226:817–829.
36. Urbaniak MD, Crossman A, Chang TH, Smith TK, van Aalten DMF, Ferguson MAJ. The *N*-acetyl-D-glucosaminylphosphatidylinositol De-*N*-acetylase of glycosylphosphatidylinositol biosynthesis is a zinc metalloenzyme. J Biol Chem 2005;280:22831–22838.
37. Arakawa N, Igarashi M, Kazuoka T, Oikawa T, Soda K. D-arginase of *Arthrobacter* sp KUJ 8602: characterization and its identity with Zn^{2+}-guanidinobutyrase. J Biochem (Tokyo) 2003;133:33–42.
38. Larson TJ, Ehrmann M, Boos W. Periplasmic glycerophosphodiester phosphodiesterase of *Escherichia coli*, a new enzyme of the GLP region. J Biol Chem 1983;258:5428–5432.
39. Robertson JG, Villafranca JJ. Characterization of metal-ion activation and inhibition of CTP synthetase. Biochem 1993;32:3769–3777.
40. Dudev T, Lin YL, Dudev M, Lim C. First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations. J Am Chem Soc 2003;125:3168–3180.

 DOI 10.1002/prot

41. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 2004;32:D431–D433.
42. Murakami MT, Fernandes-Pedrosa MF, Tambourgi DV, Arni RK. Structural basis for metal ion coordination and the catalytic mechanism of sphingomyelinases D. J Biol Chem 2005;280:13658–13664.
43. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34:D247–D251.
44. Cordes MHJ, Binford GJ. Lateral gene transfer of a dermonecrotic toxin between spiders and bacteria. Bioinformatics 2006;22:264–268.
45. van der Rest B, Rolland N, Boisson AM, Ferro M, Bligny R, Douce R. Identification and characterization of plant glycerophosphodiester phosphodiesterase. Biochem J 2004;379:601–607.
46. Ishikawa K, Ishida H, Matsui I, Kawarabayasi Y, Kikuchi H. Novel bifunctional hyperthermostable carboxypeptidase/aminoacylase from *Pyrococcus horikoshii* OT3. Appl Environ Microbiol 2001;67:673–679.
47. Liaw SH, Chen SJ, Ko TP, Hsu CS, Chen CJ, Wang AHJ, Tsai YC. Crystal structure of D-aminoacylase from *Alcaligenes faecalis* DA1—a novel subset of amidohydrolases and insights into the enzyme mechanism. J Biol Chem 2003;278:4857–4962.
48. Lai WL, Chou LY, Ting CY, Kirby R, Tsai YC, Wang AHJ, Liaw SH. The functional role of the binuclear metal center in D-aminoacylase—one-metal activation and second-metal attenuation. J Biol Chem 2004;279:13962–13967.
49. Janda I, Devedjiev Y, Derewenda U, Dauter Z, Bielnicki J, Cooper DR, Graf PCF, Joachimiak A, Jakob U, Derewenda ZS. The crystal structure of the reduced, Zn^{2+}-bound form of the *B. subtilis* Hsp33 chaperone and its implications for the activation mechanism. Structure 2004;12:1901–1907.
50. Evans JC, Huddler DP, Jiracek J, Castro C, Millian NS, Garrow TA, Ludwig ML. Betaine-homocysteine methyltransferase: zinc in a distorted barrel. Structure 2002;10:1159–1171.
51. Shi WX, Zhan CY, Ignatov A, Manjasetty BA, Marinkovic N, Sullivan M, Huang R, Chance MR. Metalloproteomics: high-throughput structural and functional annotation of proteins in structural genomics. Structure 2005;13:1473–1486.