NIH Public Access
Author Manuscript
Proteins. Author manuscript; available in PMC 2012 March 20.

Published in final edited form as:
Proteins. 2011 March ; 79(3): 735–751. doi:10.1002/prot.22913.

FINDSITE-metal: Integrating evolutionary information and machine learning for structure-based metal binding site prediction at the proteome level

Michal Brylinski and **Jeffrey Skolnick***
Center for the Study of Systems Biology, Georgia Institute of Technology, Atlanta, GA 30318

Abstract

The rapid accumulation of gene sequences, many of which are hypothetical proteins with unknown function, has stimulated the development of accurate computational tools for protein function prediction with evolution/structure-based approaches showing considerable promise. In this paper, we present FINDSITE-metal, a new threading-based method designed specifically to detect metal binding sites in modeled protein structures. Comprehensive benchmarks using different quality protein structures show that weakly homologous protein models provide sufficient structural information for quite accurate annotation by FINDSITE-metal. Combining structure/evolutionary information with machine learning results in highly accurate metal binding annotations; for protein models constructed by TASSER, whose average Cα RMSD from the native structure is 8.9 Å, 59.5% (71.9%) of the best of top five predicted metal locations are within 4 Å (8 Å) from a bound metal in the crystal structure. For most of the targets, multiple metal binding sites are detected with the best predicted binding site at rank 1 and within the top 2 ranks in 65.6% and 83.1% of the cases, respectively. Furthermore, for iron, copper, zinc, calcium and magnesium ions, the binding metal can be predicted with high, typically 70-90%, accuracy. FINDSITE-metal also provides a set of confidence indexes that help assess the reliability of predictions. Finally, we describe the proteome-wide application of FINDSITE-metal that quantifies the metal binding complement of the human proteome. FINDSITE-metal is freely available to the academic community at http://cssb.biology.gatech.edu/findsite-metal/.

Keywords

metalloproteins; metal binding residue prediction; protein threading; protein structure prediction; human proteome; machine learning

Introduction

With the continuing advances in genome sequencing, there has been a rapid accumulation of protein sequences, whose molecular functions are yet to be annotated 1›2. Consequently, the meticulous functional characterization of all gene products in a given proteome has become one of the greatest challenges in the post-genomic era. This ambitious goal can be achieved by combining experimental and computational efforts 3. In this spirit, a number of sequence- and structure-based methods for function inference by computational means have been

*Corresponding author .

Availability

FINDSITE-metal source code as well as the benchmarking results are freely available to the academic community from http://cssb.biology.gatech.edu/findsitemetal/. Moreover, we set up a FINDSITE-metal web server that can be accessed at the same URL. Modeling results for the human proteome are available from http://cssb.biology.gatech.edu/human/.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

developed 4–6. One particular group of highly efficient and broadly applicable algorithms that show a considerable promise for proteome-scale functional annotation consists of evolution/structure-based approaches 7–10, whose common underlying principle is that protein function is transferable between evolutionarily related proteins. Of course, protein function is multifaceted, ranging from biochemical processes to phenotypical responses 11. The critical functional aspects giving rise to life emerge from interactions among molecular species present in a cell, such as proteins, small organic molecules, nucleic acids, and metal ions. The latter bind to a broad spectrum of proteins to facilitate many important biological functions and fundamental chemical processes 12–14.

The metal binding complement of a typical proteome comprises about one-quarter to one-third of all gene products 12,15. Metalloproteins belong to many different functional classes 16; the most important are enzymes, transport and storage proteins, gene expression regulators and signal transduction proteins 17–19. The presence of metal ions is critical not only for many specific molecular functions that cannot be easily performed by a relatively limited repertoire of chemical groups in naturally occurring amino acids, but also for the folding and the stability of protein structures 20. Recognition of the importance of metal binding in numerous cellular processes has stimulated the development of computational methods aimed specifically at the prediction of metal binding sites in proteins. Global sequence similarity methods that use BLAST searches 21 are generally applicable in the high sequence identity regime 22,23, but their ability to detect functional relationships falls off dramatically in the twilight zone of sequence identity 24,25. In the absence of closely homologous sequences, another group of methods that employ short sequence motifs searches can be employed 26–28. However, local sequence matching approaches suffer from low coverage of metal binding sites, since many are non-local in sequence without any distinct spacing patterns 29,30. To overcome these problems, a number of structure-based approaches have been developed. Since global structure similarity between proteins may lead to a very high false positive rate due to the complex and ambiguous relationships between protein structure and function 31, most metal binding site predictors utilize highly conserved local structural patterns 32–34 and focus on the local physicochemical environment around a metal binding site 35. Despite their high accuracy in benchmark simulations carried out using known metal binding protein structures, the ability to detect novel metal binding sites may be somewhat limited 30.

Another complicating fact is that the local geometrical matching typically requires high quality structures, preferably solved by X-ray crystallography or constructed from very close homology. As demonstrated for 653 structures modeled at different resolutions, the precise recognition of the functional site location typically requires high-resolution structures whose root-mean-square deviation, RMSD, from the native structures is 1-2 Å 36. Similar limitations apply to purely structure-based metal binding site detectors that use all-atom force fields 37. While such high accuracy of modeled protein structures is generally achievable in template-based structure prediction using closely homologous templates 38,39, most models constructed from remote homology, despite having the correct global topology, have an RMSD far above 2 Å resolution 40,41, with significant structural inaccuracies in the binding regions 42,43. If the goal is function inference at the level of entire proteomes (where very high quality models are present for only a small fraction of proteins), effective structure-based approaches that cope well with structural inaccuracies in modeled protein structures are required. Combined evolution/structure-based function inference was previously demonstrated to be quite successful in the detection of binding sites for small organic molecules 7,44,45 and DNA 46,47 in the presence of only remotely related templates.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Here, we extend the application of the FINDSITE algorithm, originally designed to identify ligand-binding sites 7,48, to predict metal-binding sites in weakly homologous protein models using distantly related templates. We begin with a statistical analysis of the conservation of metal binding patterns in remotely related proteins followed by comprehensive large-scale benchmarks using different quality protein models as the structures used for binding site prediction. The results for proteins that bind to transition metals (cobalt, copper, iron, manganese, nickel and zinc) as well as to hard metals (calcium and magnesium) are assessed in terms of the predicted binding site location, the accuracy of identified binding residues and the precision of the binding metal prediction. Furthermore, we demonstrate that the performance of FINDSITE-metal is notably improved by integrating structure/evolutionary information and machine learning. The important feature of FINDSITE-metal is that it offers a set of confidence indexes, which help assess the reliability of its predictions. Finally, we describe a proteome-wide application of FINDSITE-metal that provides a detailed functional characterization of the metal binding complement of the human proteome. FINDSITE-metal is freely available to the academic community at http://cssb.biology.gatech.edu/findsite-metal/.

Materials and Methods

Dataset

The metal binding proteins used in this study were obtained from the Metalloprotein site Database and Browser (MDB) 49, which provides quantitative information on all metal-containing sites available from structures in the PDB 50. Only proteins bound to the following eight metal ions were included in the dataset: Ca, Co, Cu, Fe, Mg, Mn, Ni and Zn. For each binding metal, a non-redundant set was compiled using PISCES 51. For proteins 50-600 residues in length, redundancy was removed at the 35% pairwise sequence identity level. The final non-redundant dataset comprises 860 proteins, of which 201, 29, 35, 117, 152, 87, 21, and 251 bind to Ca, Co, Cu, Fe, Mg, Mn, Ni and Zn, respectively. For each metal binding site, a set of binding residues was identified using the interatomic contacts provided by the LPCsoftware 52 with the remaining residues classified as non-binding. The list of proteins and associated metal binding ions can be found at http://cssb.biology.gatech.edu/findsite-metal/.

Protein structure modeling

For each target protein, we have constructed several models of different quality in terms of their RMSD 53 from the native structure. In addition to the crystal structures, we use three sets of uniformly distorted structures with an average RMSD of 2, 4 and 6 Å from native. The distorted structures were generated starting from the crystal structures by a simple Monte Carlo procedure that deforms protein structures to a desired resolution 54. Moreover, we apply a state-of-the-art template-based structure prediction algorithm 55 to construct a set of weakly homologous protein models. First, for each target protein, distantly related template structures (<35% sequence identity to the target) were identified in a non-redundant PDB library by our meta-threading procedure that employs the SP3 56, SPARKS2 57 and PROSPECTOR_3 58,59 algorithms. Subsequently, full-length models were assembled by chunk-TASSER 55. Finally, all-atom models from the top ranked chunk-TASSER structures were constructed by Pulchra 60 and energy minimized in the CHARMM22 force field 61 using the Jackal modeling package 62. The benchmark dataset in terms of the average quality of protein structures as assessed by RMSD 53, TM-score 63 and MaxSub 64 is summarized in Table I.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Template identification

FINDSITE-metal is a template-based procedure for metal binding site prediction. Here, template proteins are identified in a non-redundant PDB library using meta-threading that employs three threading procedures: SP3 56, SPARKS2 57 and PROSPECTOR_3 58,59. Only weakly homologous (<35% sequence identity to the target) template structures that have a Z-score of ≥4 reported by at least one threading method are included. The initial set of templates provided by threading is used to retrieve all metal binding protein structures from the PDB that are homologous to at least one threading-identified template. Multiple instances of a metal binding template protein (>90% sequence identity) are only retained if they bind either to a different metal ion or to the same metal but in a different location, with a distance of >4 Å. Otherwise, only one PDB structure is included. Again, we remove all PDB templates with >35% sequence identity to the target. Finally, only those template structures that have a TM-score to the provided target structure of ≥0.4 are retained. If distorted or modeled proteins are used as the targets for metal binding site prediction, the TM-score is calculated vs. these structures. This structure similarity threshold ensures that the template-to-target structure alignments generated by fr-TM-align 65,66 are statistically significant.

Metal binding site prediction

Similar to the original FINDSITE approach 7,48, FINDSITE-metal employs structure alignments provided by fr-TM-align 65,66 to superimpose metal-binding templates detected by threading onto the target structure (either crystal or modeled). Subsequently, upon global superposition of the template structures, template-bound metal ions are clustered using an average linkage clustering procedure and the resulting clusters are ranked by the number of binding metals. Each cluster represents a putative metal binding site with the predicted metal location at the cluster geometrical center (averaged coordinates of all templatebound metals).

Binding residue prediction

For each metal binding cluster, the initial set of binding residues is calculated as follows: Each target residue is assigned a probability that corresponds to the fraction of templates that have a residue in equivalent position in contact with a metal, including pseudo counts:

$$p_i^B = \frac{c + f_i \sqrt{n}}{n + \sqrt{n}} \quad \text{Eq. 1}$$

where p_i^B is the metal binding probability, c is the number of templates that have the equivalent residue in contact with a metal, n is the total number of templates and f_i is the frequency of occurrence of residue i in UniProtKB/Swiss-Prot 67 (see Release notes for UniProtKB/Swiss-Prot release 56.0). Residue equivalences are calculated from structure alignments generated by fr-TM-align.

Prediction of binding metal preferences

Similarly, for each putative binding site, we calculate the preferences toward different metal ions using a fraction of templates that bind particular type of metal:

$$p_j^M = \frac{c + f_i \sqrt{n}}{n + \sqrt{n}} \quad \text{Eq. 2}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where p_j^M is the probability of binding metal *j* (we use eight different metal types: Ca, Co, Cu, Fe, Mg, Mn, Ni and Zn), *c* is the number of templates that bind metal *j*, *n* is the total number of templates and f_j is the frequency of occurrence of metal *j* in a non-redundant subset of the MDB 49.

The uncertainty of metal binding (*ME*) is quantified by the Shannon's entropy 68:

$$ME = -\sum_{j=1}^{8} p_j^M \log_2 p_j^M \quad \text{Eq. 3}$$

Low *ME* values are indicative of relatively homogenous metal binding sites, i.e. similar locations in evolutionarily related proteins tend to bind the same type of metal. We use *ME* to construct a reliable confidence index for binding metal prediction.

Machine learning

The accuracy of metal binding residue prediction is further improved by machine learning using classification-based Support Vector Machines (SVM). Here, we use libSVM 2.9 69 to build a C-SVC model with a radial basis function. To avoid memorization of the dataset, we use a 2-fold cross validation protocol. The complete dataset of the target complexes was randomly divided into two subsets with < 35% sequence identity between any two proteins that belong to the different subsets. Subsequently, each subset was used to train the model, and then predictions were made for the remaining targets, excluded from the training procedure. The constructed SVM model employs the set of 25 features summarized in Table II. We use the trained SVM classifier to assign each residue with a probability to bind a metal ion (probability being a positive). One of the features used in SVM is the TM-score to native estimated for the target structure. Using crystal structures, this value is 1.0. For structures distorted to 2, 4 and 6Å RMSD, we train the model on the TM-score values calculated vs. native structures; however, in validation, we use random TM-score values sampled from a normal distribution calculated using the mean and standard deviation for a given dataset of distorted structures (see Table I). For example, the estimated TM-score for each target from the set of structures distorted to 4 Å RMSD is selected randomly from a normal distribution with a mean of 0.75 and a standard deviation of 0.07. For TASSER models, the TM-score is estimated from the C-score, the confidence score calculated from TASSER simulations 70. This is described in the following section.

TM-score estimation

Previously, the C-score was introduced as a structure prediction confidence index 70. A positive C-score indicates that the modeled structure is very likely to be topologically similar to native at the structurally significant level. Here, C-score values are used to estimate the TM-score vs. native for protein structures modeled by TASSER. To find the correlation between the C-score and the TM-score, we use the results of large-scale benchmark simulations carried out for a standard, non-redundant dataset of proteins that cover the PDB at the 35% sequence identity. The dataset consists of 1,489 single protein chains up to 100 residues in length, 2,494 proteins between 100-200 residues and 1,203 larger proteins between 200-300 residues. Structure models are constructed by TASSER from weakly homologous template structures (<35% sequence identity) identified by threading 58,59. The regression analysis is performed using the C-score values calculated from TASSER simulations and the TM-score values calculated vs. crystal structures of the targets.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Binding site re-ranking

The identity of metal-binding residues is typically highly conserved across a set of evolutionarily related proteins. We use this observation to re-rank binding sites by a sequence profile score, which involves the summation over all twenty amino acid types of the product of the probabilities that a given amino acid occupies the equivalent position in the target and template and is derived from structure alignments generated by fr-TM-align for a set of weakly homologous metal-binding templates. Putative binding sites, initially ranked by the number of binding metals, are re-ranked by the total sequence profile score calculated over binding residues predicted by the SVM.

Prediction confidence

FINDSITE-metal uses three confidence indexes that estimate the chances of i) the metal position to be predicted within a distance of 4 Å, ii) the Matthew's correlation coefficient for the binding residues to be at least 0.5 and iii) the binding metal type to be correctly predicted. The prediction confidence is estimated by a Naïve Bayes classifier 71 from a set of features listed in Table II. A separate classifier is trained for each confidence index. Similar to the binding residue prediction by SVM, we use a 2-fold cross validation protocol. We find that the raw instance scores returned by the Bayesian classifier make rather poor confidence estimates, because they are grouped around the extreme values (0 and 1). Therefore, we apply a calibration procedure to normalize confidence estimates generated by the classifier. Here, we use the Pool Adjacent Violators (PAV) algorithm 72 that transforms the raw scores into well-calibrated posterior probabilities, which are further used as the confidence estimates.

Comparison to a sequence-based method

We compare the performance of FINDSITE-metal to SVM-Prot, an SVM-based method that predicts the functional class of metal-binding proteins from sequence derived physicochemical properties 73. Here, we use the same non-redundant MDB dataset, which is described in the previous sections. To make results comparable between SVM-Prot and FINDSITE-metal, for the latter, we employ weakly homologous TASSER models as the target structures. Moreover, we use only distantly related (<35% sequence identity to target) metal-bound template structures identified by meta-threading. In this manner, the predictions are made by both approaches solely using sequence information as the input. The SVM-Prot server at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi was queried automatically and the results were parsed for metal-binding assignments. The class of binding metal is selected based on the p-value reported by SVM-Prot. For FINDSITE-metal, the metal type is selected based on the highest preference calculated using Eq. 2. The results are assessed separately for each metal type (Ca, Co, Cu, Fe, Mg, Mn, Ni and Zn) using a standard receiver operating characteristic (ROC) analysis.

Prediction of metal-binding sites in the human proteome

Amino acid sequences of all gene products identified in the human proteome were obtained from the Ensembl genome database 74. Here, we use 56,376 protein sequences 50-600 residues in length selected from the human assembly GRCh37, release 55. For each sequence, a Cα backbone model was built by TASSER 70,75 from template structures identified by SP3 56, SPARKS2 57 and PROSPECTOR_3 58,59. Subsequently, all-atom models reconstructed by Pulchra 60 from low-resolution TASSER structures were subject to short energy minimization using Jackal 62. The set of meta-threading identified templates and the modeled structures were then used by FINDSITE-metal to detect putative metal-binding sites in the human proteome. Each predicted binding site was assigned a confidence

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and further characterized by the prediction of metal-binding residues, the class of binding metal and the molecular function.

Results

Metal-binding templates

FINDSITE-metal employs a set of evolutionarily related metal-binding templates selected by threading. The average number of templates per target is 18. Upon the global superposition onto the target structure, putative metal-binding sites are detected by a clustering procedure. First, for a given set of templates, we analyze what is the optimal clustering cutoff in terms of the average distance from the metal position in the target structure and the ranking ability by the cluster multiplicity. In Figure 1, we show that the average distance of the best binding site increases with the clustering threshold. Small cutoff values result in many puny clusters, one of which is typically close to the native metal binding site; however, this is at the expense of a poor ranking capability (Figure 1, inset). On the other hand, if a large cutoff is used, the ranking becomes more efficient, but the average distance from the natively bound metal position increases. As a trade-off between the accuracy of the site prediction and the ranking ability, we use a clustering cutoff of 8 Å in further calculations.

Analysis of metal binding residues

Metal binding in proteins typically requires a specific geometrical arrangement of relatively few residues, whose identity strongly depends on the type of binding metal 13,76-78. This is shown in Figure 2, where we compare the amino acid preferences to bind different metals. The electron donors in the side chains are mainly the carboxyl oxygen atoms of Asp and Glu, the imidazole nitrogen of His and the thiol group of Cys. Moreover, the amide nitrogen and oxygen of Asn and the thioether group of Met also coordinate divalent metal ions. Calcium, magnesium and manganese ions preferentially bind to the acidic chain of Asp. Zinc, with a lower coordination number preference, is typically chelated by Cys and His. Histidine residues also have strong preference toward binding of cobalt, copper, iron, nickel and zinc atoms. These results correlate very well with the recent statistical analysis of the chemical environment of metal binding in proteins 79. Binding patterns are potentially important for the prediction of metal binding residues. If the type of metal is correctly predicted, the differential metal binding preferences of amino acid side chains can be used to increase the accuracy of binding residue prediction.

Binding metal preferences

In this section, we show that evolutionarily remotely related proteins tend to bind similar metals in equivalent locations. For each target protein and the corresponding set of threading identified templates, we calculate the preferences toward binding native and non-native metal ions. The native metal preference is equivalent to the fraction of templates that bind the same metal as the target structure. Likewise, the non-native preference is an averaged fraction of templates that bind different metal types in similar locations. Figure 3 shows that the native metal preference is strongly correlated with the distance between the target- and template-bound metal ions upon the global superposition of their structures. For site distances less than 2 Å (4 Å), more than 70% (50%) of the templates bind the same metal type as the target structure. This fraction drops dramatically for sites >4 Å away from each other. Encouragingly, this tendency is observed not only for crystal structures (Figure 3A), but also for the distorted target structures (Figure 3 B-D). As we demonstrate in the following sections, combining information on the predicted binding metal with the differential metal binding preferences of amino acids improves the accuracy of the prediction of binding residues particularly against modeled protein structures.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TM-score estimation

Before we discuss the performance of FINDSITE-metal in metal binding site prediction, we shortly describe the results of an analysis that focuses on the relationship between the C-score and the TM-score. The TM-score provides a length-independent measure of the structural similarity between two proteins 63. A significant similarity is indicated by a TM-score of >0.4. FINDSITE-metal uses the estimated TM-score to native as one of the SVM features to accurately predict metalbinding residues and to estimate the prediction confidence. For a given protein model, its TM-score can be directly calculated against the native crystal structure, if known. However, in a real prediction scenario when the experimental structure of a target is unavailable, the TM-score needs to be estimated. Most contemporary structure prediction algorithms estimate the reliability of structure modeling using some score 80-82. TASSER, a template-based structure assembly/refinement approach 70,75, calculates a confidence score, called the C-score. To estimate the TM-score for a target structure modeled by TASSER, we carried out a regression analysis using the results of large-scale benchmarks on a non-redundant and representative dataset of 5,186 proteins. For each modeled target structure, we calculate the TM-score vs. its crystal structure and plot it against the C-score obtained from TASSER simulations; this is shown in Figure 4. Next, we calculate the regression line, which in this case is:

$$TM-score = 0.13173 \times C-score + 0.42895 \quad \text{Eq. 4}$$

with a Pearson correlation coefficient of 0.81. We use this equation to estimate the TM-score for target structures modeled by TASSER. A high correlation between the TM-score and a modified version of the C-score was also reported for I-TASSER 83. We note that FINDSITE-metal does not require the exact TM-score; all benchmark results reported here were obtained using TM-score estimates rather than the exact values (see Materials and Methods). In principle, any other model quality assessment that correlates with the TM-score can also be used.

Accuracy of the predicted metal position

FINDSITE-metal predicts putative metal binding locations by the clustering of template-bound metal ions upon the global superposition of the template structures. In Figure 5, we assess how far is the predicted site center from a metal position in the crystal structure, when the crystal structures themselves as well as structures distorted to a desired RMSD are used as the targets. Using a distance of 4 Å as the hit criterion, the fraction of correctly predicted sites at the top five ranks is 69.5%, 67.2%, 58.9% and 50.8% for the crystal structures and structures distorted to 2, 4 and 6 Å RMSD, respectively. The difference between crystal and distorted structures diminishes with more promiscuous distance thresholds; for 6 Å (8 Å), the difference between the fraction of correctly predicted metal binding sites using crystal and the most distorted structures is only 8.1% (3.9%). This is a common feature of many template-based predictors that employ structure alignments, which are sensitive to the global topology rather than to the local structural features 7,31,48.

The ability of FINDSITE-metal to correctly rank the predicted binding sites is assessed in Figure 5 (inset) for two ranking protocols. Similarly to the original FINDSITE algorithm, predicted binding locations could be ranked by the fraction of templates that share a common site. Using ranking by fraction, in 67.8%, 67.0%, 66.8% and 66.5% of the cases, the best predicted site is at rank 1 for the crystal structures and structures distorted to 2, 4 and 6 Å RMSD, respectively. As demonstrated above, binding metal types as well as residue profiles are remarkably strongly conserved across a set of evolutionarily related proteins. Here, we use this observation to re-rank the predicted sites by a sequence profile score calculated vs. metal-bound templates. Figure 5 (inset) demonstrates that ranking by a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

sequence profile score calculated for binding residues predicted by SVM (see Materials and Methods) yields 3-4% improvement with respect to ranking by fraction. Furthermore, ranking ability is hardly affected by the distortion of the target structure; using the crystal structures and structures distorted to 2, 4 and 6Å RMSD, the correct ranking by the sequence score is found in 70.7%, 70.6%, 70.3% and 69.3% of the cases, respectively.

Binding residue prediction

Many algorithms for metal binding residue prediction use the fact that metal binding typically involves a limited set of strongly conserved binding patterns 30,33,84. Here, we analyze the performance of FINDSITE-metal in the prediction of binding residues using two classifiers: a probability-based residue selection and machine learning that employs a set of sequence and structure derived features. A probability-based approach simply assigns a residue with a binding probability that corresponds to the fraction of templates that have a residue harboring a metal ion in the equivalent position. Figure 6 demonstrates that the overall accuracy of binding residue prediction is significantly improved when the SVM classifier is applied. For example, at the cost of a 1% false positive rate, machine learning increases the true positive rate by 7% (from 84% to 91%), 9% (from 81% to 90%), 11% (from 75% to 86%) and 12% (from 69% to 81%) for the crystal structures and structures distorted to 2, 4 and 6 Å RMSD, respectively. Similar improvement is observed in the Recall-Precision graphs shown as the inset plots in Figure 6.

Prediction of binding metal

As demonstrated in the previous sections, similar binding sites across a set of evolutionarily related proteins bind similar types of metal ions. We use this observation to predict a metal that likely binds to the detected binding sites. The fraction of dataset targets for which the binding metal was correctly predicted is shown in Figure 7, separately for each metal class. The highest accuracy, 70-90%, is observed for Fe, Cu, Zn, Ca and Mg binding proteins. The fraction of proteins correctly predicted to bind Mn, Co and particularly Ni is significantly lower; this is caused by the underrepresentation of these proteins in the PDB 50,79. Moreover, only very distant homologues are used in this study that might have evolved to bind a different class of metal ions. It is noteworthy that the accuracy of binding metal prediction is highly insensitive to the structural distortions of the target structures.

Predicted amino acid/metal composition

Analyzing all predicted binding residues and metal ions across the non-redundant dataset, we can estimate the rate of over- and under-prediction of certain amino acids and metal types. Overall, we find that the predicted residue as well as metal composition is in an excellent agreement with these calculated directly from the PDB complexes; this is shown in Table III. Typically, the amino acid composition difference is less than a couple of percent, with the exception of Cys that is slightly under-predicted by 2.6-3.7%. For low quality protein structures, particularly these that are distorted to 4 Å and 6 Å RMSD, we observe a moderate composition excess of Asp, Glu and His residues of 1.0-5.5%. Considering the metal composition across the dataset, most binding metal types are correctly predicted with two exceptions: magnesium-binding sites that are over-predicted by 7.9-8.6% and manganese-binding sites are under-predicted by 6.3-7.0%. Nevertheless, the residue/metal composition of metal binding sites is fairly well reproduced by FINDSITE-metal predictions; this feature is important for large-scale functional assignments at the level of entire proteomes.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TASSER models as targets for FINDSITE-metal

Artificially distorted structures provide some notion about the performance of a method using different quality target structures. However, from the point of view of real applications, the most interesting results are these obtained using protein structures modeled by a state-of-the-art protein structure prediction approach. In this study, using TASSER, we constructed protein models from only weakly related template structures. In Figure 8, the performance of FINDSITE-metal using TASSER models is compared in terms of the site location accuracy and ranking capability to that using the crystal structures of the targets. Unlike other structure-based approaches to metal binding site prediction that strongly rely on the quality of the target structure 85,86, FINDSITE-metal is insensitive to some extent to the structural distortions in modeled proteins structures. Using a 4 Å (8 Å) distance as a hit criterion, the accuracy drops by 10% (5%) if weakly homologous TASSER models are used instead of the crystal structures, with the ranking ability reduced by 4.5%. Furthermore, the high accuracy of binding residue identification is retained; this is shown in Figure 9 as a ROC plot and a Recall-Precision graph (inset). We also find a good agreement between the number of binding residues per site in the crystal structures of the complexes (3.13 ±0.99) and the number of residues predicted for TASSER models (3.69 ±1.93), see Table I.

Confidence indexes

Due to the inherent limitations of many template-based approaches to functional annotation, such as the unavailability of suitable templates or the possibility of severe topological inaccuracies in the structures modeled using remote homology, confidence indexes are required to assess the prediction reliability. FINDSITE-metal employs three confidence estimates for the predicted site distance, the set of identified binding residues and the class of binding metals. As described in the Materials and Methods, these indexes are calculated using calibrated Bayesian classifiers. In Figure 10, we present the performance of FINDSITE-metal, as assessed by the fraction of templates for which a correct prediction was made, for targets assigned with different confidence values: the binding site distance ≤4Å, the Matthew's correlation coefficient, MCC, ≥0.5 and the confidence that the binding metal is correctly identified. Our confidence indexes correlate well with the prediction accuracy not only for the target crystal structures, but also when distorted structures and TASSER models are used as the targets. We note that the confidence is estimated for a given target without using any information on the experimental structure of the metal-protein complex.

Comparison to SVM-Prot

We compare the performance of FINDSITE-metal to SVM-Prot, a sequence-based predictor of the functional class of metal-binding proteins. In this analysis, the focus is on the prediction of a metal-binding type. The results for Ca, Co, Cu, Fe, Mg, Mn, Ni and Zn binding proteins are presented in Figure 11 and Table IV. FINDSITE-metal achieves a relatively high sensitivity of >50% for five out of eight functional classes (Fe, Cu, Zn, Ca and Mg) at a moderate false positive rate below 20%. For the remaining metal-binding proteins, the assignments are marginally better than random. The accuracy of SVM-Prot is notably better than random only for targets that bind to Fe, Zn and Mn. The sensitivity of SVM-Prot for zinc is higher than using FINDSITE-metal; however, with a much higher false positive rate, which indicates a significant over-prediction of zinc-binding proteins. SVM-Prot is also more sensitive in detecting manganese-binding targets (29% sensitivity at 5% false positive rate). Nevertheless, considering the overall accuracy of the functional class assignment, FINDSITE-metal represents a significant improvement over SVM-Prot. Since the benchmarking results reported here seem encouraging, below we describe the application of FINDSITE-metal to the entire human proteome.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Metal binding complement of the human proteome

In this study, we constructed structural models for 56,376 gene products in the human proteome, 50-600 residues in length. 34,808 of these were assigned with at least one metal binding site. The distribution of the estimated TM-score for putative metal-binding proteins in the human genome is shown in Figure 12. A TMscore of ≥0.4 was assigned to 70.7% of the targets (24,617 gene products). Structural models for these sequences are very likely to be correct, at least at the topological level. Since FINDSITE-metal tolerates to some extent structural inaccuracies in modeled structures, these results suggest that reliable predictions can be made for the majority of proteins.

The prediction confidence for the metal binding site prediction in the human proteome is presented in Figure 13. For roughly one third of the gene products, the estimated confidence that the distance of the top-ranked site is predicted within 4 Å and the binding residues are identified with a Matthew's correlation coefficient of at least 0.5 is higher than 40-50%. For three quarters of the targets, the confidence of binding metal prediction is >50%. Considering the large number of targets, FINDSITE-metal provides confident metal binding information for thousands of gene products in the human proteome.

Figure 14 shows the statistics on the assigned metal binding class to the human metalloproteome. Nearly one third of putative metal binding proteins in the human proteome were predicted as being calcium-binding and another 30% as magnesium-binding. The third largest class (one-quarter of putative metalloproteins) consists of proteins that bind to zinc. Fe, Mn, Co, Cu and Ni were assigned to 5.4%, 3.6%, 1.9%, 1.4% and 1.2% of the targets, respectively. Benchmark results reported in the previous sections suggest that the number of proteins that bind to Mn, Co and Ni might be somewhat underestimated. Nevertheless, the composition of the metal binding complement of the human proteome identified by FINDSITE-metal is consistent with other studies 23,79.

Discussion

In this study, we describe the development of FINDSITE-metal, a new threading-based approach to metal binding site prediction from remote homology. FINDSITE-metal is essentially an extension of FINDSITE, which was designed to identify binding sites for small organic molecules 7,31,48. In large-scale benchmarking, we demonstrate that FINDSITE-metal performs satisfactorily in the presence of only weakly related template structures that are detectable by sequence profile-driven threading 87. Moreover, it is highly insensitive to the deformation of the target structure; thus, it can be applied to approximate protein structures modeled by state-of-the-art structure prediction approaches. Highly conserved binding patterns observed across the interactions between metal ions and proteins constitute a perfect set of attributes for machine learning applications. Indeed, many metal binding site predictors routinely use Support Vector Machines 73,88 and Neural Networks 30,34. Here, we show that integrating structure/evolution information from threading and machine learning significantly improves the accuracy of metal-binding residue prediction. Similar to the use of local filters, such as the clustering of molecular entities (ligands, DNA or metal ions) bound to proteins, machine learning improves the prediction accuracy by reducing the false positive rate.

Many existing approaches to metal binding site prediction focus either on a specific binding metal, e.g. zinc 88, copper 89 or iron 90 or selected amino acids, typically cysteine, histidine, glutamic and aspartic acid residues 30,33. The statistical analysis of the crystal structures of protein-metal complexes shows that although these four amino acids dominate the metal binding environment, other residue types also contribute to the metal coordination spheres through the interactions with the backbone carbonyl 79. FINDSITE-metal

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

concomitantly considers eight types of commonly occurring metal ion sites in proteins; there are also no explicit restrictions imposed on the identity of binding residues. This feature is important for proteome-scale applications, where the emphasis is on the quantification of interactions between proteins and metal ions. In this study, we describe the application of FINDSITE-metal to the human proteome and provide the detailed structural characterization of its metal binding complement. Such knowledge is important not only for helping to elucidate the molecular function, but by providing information about which metals bind, it may assist in the determining suitable crystallization environments for use in structural genomics 91. In the near future, we will next apply FINDSITE-metal to other important eukaryotic as well as prokaryotic proteomes.

As any other computational method for functional inference, FINDSITE-metal has several limitations. The most prominent is the availability and detectability of metal-bound template structures. Here, we show that only evolutionarily distantly related templates are required; however, in some cases, these may be absent in the PDB or the template identification procedure may fail to detect them. We can expect the accuracy of FINDSITE-metal to gradually improve with the advances in the development of sensitive threading algorithms as well as with the continuous growth of the structural databases and the progress of Structural Genomics projects 91. Regarding binding metal prediction, FINDSITE-metal neglects the mechanisms that control how metalloproteins acquire their metals from the cellular pools. For instance, it has been demonstrated for cupin A that the compartment in which a protein folds may override its binding preference to control its metal content 92. This is a very challenging problem from the point of view of fully automated function annotation and, to the best of our knowledge, no effective algorithms have been developed so far to address this phenomenon.

Sensitive sequence profile driven threading detects evolutionarily related homologues with respect to many aspects of protein function. Binding of small organic molecules, nucleic acids or metal ions are only a few examples that can be extended to ultimately cover all aspects of protein molecular function. Thus, combined structure/evolution function annotation emerges as a powerful technique for the large-scale functional screening of the available genomic information.

Acknowledgments

We thank Dr Shashi B. Pandit for his TASSER benchmark dataset used here to calculate the correlation between C-score and TM-score. This work was supported in part by grant No. GM-48835 and GM-37408 of the Division of General Medical Sciences of the National Institutes of Health.

References

1. Aury JM, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F, Wincker P. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics. 2008; 9:603. [PubMed: 19087275]
2. Tettelin H, Feldblyum T. Bacterial genome sequencing. Methods Mol Biol. 2009; 551:231–247. [PubMed: 19521879]
3. Roberts RJ. Identifying protein function--a call for community action. PLoS Biol. 2004; 2(3):E42. [PubMed: 15024411]
4. Juncker AS, Jensen LJ, Pierleoni A, Bernsel A, Tress ML, Bork P, von Heijne G, Valencia A, Ouzounis CA, Casadio R, Brunak S. Sequence-based feature prediction and annotation of proteins. Genome Biol. 2009; 10(2):206. [PubMed: 19226438]
5. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A. Protein function annotation by homology-based inference. Genome Biol. 2009; 10(2):207. [PubMed: 19226439]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

6. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci. 2003; 60(12):2637–2650. [PubMed: 14685688]
7. Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A. 2008; 105(1):129–134. [PubMed: 18165317]
8. Friedberg I. Automated protein function prediction--the genomic challenge. Brief Bioinform. 2006; 7(3):225–242. [PubMed: 16772267]
9. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys. 2003; 36(3):307–340. [PubMed: 15029827]
10. Erdin S, Ward RM, Venner E, Lichtarge O. Evolutionary trace annotation of protein function in the structural proteome. J Mol Biol. 2009; 396(5):1451–1473. [PubMed: 20036248]
11. Betz SF, Baxter SM, Fetrow JS. Function first: a powerful approach to post-genomic drug discovery. Drug Discov Today. 2002; 7(16):865–871. [PubMed: 12546953]
12. Barondeau DP, Getzoff ED. Structural insights into protein-metal ion partnerships. Curr Opin Struct Biol. 2004; 14(6):765–774. [PubMed: 15582401]
13. Sarkar B. Metal protein interactions. Prog Food Nutr Sci. 1987; 11(3-4):363–400. [PubMed: 3328221]
14. Yamashita MM, Wesson L, Eisenman G, Eisenberg D. Where metal ions bind in proteins. Proc Natl Acad Sci U S A. 1990; 87(15):5648–5652. [PubMed: 2377604]
15. Holm RH, Kennepohl P, Solomon EI. Structural and Functional Aspects of Metal Sites in Biology. Chem Rev. 1996; 96(7):2239–2314. [PubMed: 11848828]
16. Finkelstein J. Metalloproteins. Nature. 2009; 460(7257):813. [PubMed: 19675640]
17. Caradonna JP, Stassinopoulos A. Functional non-heme iron metalloenzyme model systems. Adv Inorg Biochem. 1994; 9:245–315. [PubMed: 8140949]
18. Karlin KD. Metalloenzymes, structural motifs, and inorganic models. Science. 1993; 261(5122): 701–708. [PubMed: 7688141]
19. Lill R. Function and biogenesis of iron-sulphur proteins. Nature. 2009; 460(7257):831–838. [PubMed: 19675643]
20. Cox EH, McLendon GL. Zinc-dependent protein folding. Curr Opin Chem Biol. 2000; 4(2):162–165. [PubMed: 10742185]
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–410. [PubMed: 2231712]
22. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003; 333(4):863–882. [PubMed: 14568541]
23. Andreini C, Bertini I, Rosato A. Metalloproteomes: a bioinformatic approach. Acc Chem Res. 2009; 42(10):1471–1479. [PubMed: 19697929]
24. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999; 12(2):85–94. [PubMed: 10195279]
25. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A. 1998; 95(11): 6073–6078. [PubMed: 9600919]
26. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. Nucleic Acids Res. 2006; 34:D227–230. Database issue. [PubMed: 16381852]
27. Passerini A, Andreini C, Menchetti S, Rosato A, Frasconi P. Predicting zinc binding at the proteome level. BMC Bioinformatics. 2007; 8:39. [PubMed: 17280606]
28. Thilakaraj R, Raghunathan K, Anishetty S, Pennathur G. In silico identification of putative metal binding motifs. Bioinformatics. 2007; 23(3):267–271. [PubMed: 17148509]
29. Harding MM. The architecture of metal coordination groups in proteins. Acta Crystallogr D Biol Crystallogr. 2004; 60(Pt 5):849–859. [PubMed: 15103130]
30. 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(2):305–316. [PubMed: 16927295]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

31. Brylinski M, Skolnick J. Comparison of structure-based and threading-based approaches to protein functional annotation. Proteins. 2009; 78(1):18–134.
32. Gregory DS, Martin AC, Cheetham JC, Rees AR. The prediction and characterization of metal binding sites in proteins. Protein Eng. 1993; 6(1):29–35. [PubMed: 8433968]
33. Levy R, Edelman M, Sobolev V. Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. Proteins. 2009; 76(2):365–374. [PubMed: 19173310]
34. 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(1):307–320. [PubMed: 15313626]
35. Ebert JC, Altman RB. Robust recognition of zinc binding sites in proteins. Protein Sci. 2008; 17(1):54–65. [PubMed: 18042678]
36. Wei L, Huang ES, Altman RB. Are predicted structures good enough to preserve functional sites? Structure. 1999; 7(6):643–650. [PubMed: 10404593]
37. Schymkowitz JW, 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 U S A. 2005; 102(29):10147–10152. [PubMed: 16006526]
38. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000; 29:291–325. [PubMed: 10940251]
39. Sanchez R, Sali A. Advances in comparative protein-structure modelling. Curr Opin Struct Biol. 1997; 7(2):206–214. [PubMed: 9094331]
40. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins. 2007; 69(Suppl 8):38–56. [PubMed: 17894352]
41. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins. 2005; 61(Suppl 7):225–236. [PubMed: 16187365]
42. DeWeese-Scott C, Moult J. Molecular modeling of protein function regions. Proteins. 2004; 55(4): 942–961. [PubMed: 15146492]
43. Piedra D, Lois S, de la Cruz X. Preservation of protein clefts in comparative models. BMC Struct Biol. 2008; 8:2. [PubMed: 18199319]
44. Raviscioni M, He Q, Salicru EM, Smith CL, Lichtarge O. Evolutionary identification of a subtype specific functional site in the ligand binding domain of steroid receptors. Proteins. 2006; 64(4): 1046–1057. [PubMed: 16835908]
45. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol. 2009; 5(12):e1000585. [PubMed: 19997483]
46. Gao M, Skolnick J. DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. Nucleic Acids Res. 2008; 36(12):3978–3992. [PubMed: 18515839]
47. Kuznetsov IB, Gou Z, Li R, Hwang S. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins. 2006; 64(1):19–27. [PubMed: 16568445]
48. Skolnick J, Brylinski M. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform. 2009; 10(4):378–391. [PubMed: 19324930]
49. 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(1):379–382. [PubMed: 11752342]
50. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–242. [PubMed: 10592235]
51. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19(12):1589–1591. [PubMed: 12912846]
52. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics. 1999; 15(4):327–332. [PubMed: 10320401]
53. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr A. 1976; 32(Pt 5):922–923.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

54. Bindewald E, Skolnick J. A scoring function for docking ligands to lowresolution protein structures. J Comput Chem. 2005; 26(4):374–383. [PubMed: 15651033]
55. Zhou H, Skolnick J. Ab initio protein structure prediction using chunk-TASSER. Biophys J. 2007; 93(5):1510–1518. [PubMed: 17496016]
56. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins. 2005; 58(2):321–328. [PubMed: 15523666]
57. Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins. 2004; 55(4): 1005–1013. [PubMed: 15146497]
58. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR--a new approach to threading. Proteins. 2001; 42(3):319–331. [PubMed: 11151004]
59. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins. 2004; 56(3):502–518. [PubMed: 15229883]
60. Rotkiewicz P, Skolnick J. Fast procedure for reconstruction of full-atom protein models from reduced representations. J Comput Chem. 2008; 29(9):1460–1465. [PubMed: 18196502]
61. MacKerell AD, Bashford D, Bellott, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. The Journal of Physical Chemistry B. 1998; 102(18):3586–3616.
62. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol. 2001; 311(2):421–430. [PubMed: 11478870]
63. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004; 57(4):702–710. [PubMed: 15476259]
64. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. 2000; 16(9):776–785. [PubMed: 11108700]
65. Pandit SB, Skolnick J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics. 2008; 9:531. [PubMed: 19077267]
66. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005; 33(7):2302–2309. [PubMed: 15849316]
67. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2005; 33:D154–159. Database issue. [PubMed: 15608167]
68. Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal. 1948; 27:379–423.
69. Chang, C-C.; Lin, C-J. LIBSVM: a library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
70. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A. 2004; 101(20):7594–7599. [PubMed: 15126668]
71. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning. 1997; 29(2):131–163.
72. Fawcett T, Niculesu-Mizil A. Technical Note: PAV and the ROC Convex Hull. PAV-ROCCHtex. 2007; 9(52):1–12.
73. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, Cao ZW, Chen YZ. Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. BMC Bioinformatics. 2006; 7(Suppl 5):S13. [PubMed: 17254297]
74. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M. The Ensembl genome database project. Nucleic Acids Res. 2002; 30(1):38–41. [PubMed: 11752248]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

75. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins. 2005; 61(Suppl 7):91–98. [PubMed: 16187349]
76. Chakrabarti P. Geometry of interaction of metal ions with sulfur-containing ligands in protein structures. Biochemistry. 1989; 28(14):6081–6085. [PubMed: 2775752]
77. Chakrabarti P. Geometry of interaction of metal ions with histidine residues in protein structures. Protein Eng. 1990; 4(1):57–63. [PubMed: 2290835]
78. Harding MM. Geometry of metal-ligand interactions in proteins. Acta Crystallogr D Biol Crystallogr. 2001; 57(Pt 3):401–411. [PubMed: 11223517]
79. Dokmanic I, Sikic M, Tomic S. Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. Acta Crystallogr D Biol Crystallogr. 2008; 64(Pt 3):257–263. [PubMed: 18323620]
80. Gopal SM, Klenin K, Wenzel W. Template-free protein structure prediction and quality assessment with an all-atom free-energy model. Proteins. 2009; 77(2):330–341. [PubMed: 19422063]
81. Kryshtafovych A, Fidelis K. Protein structure prediction and model quality assessment. Drug Discov Today. 2009; 14(7-8):386–393. [PubMed: 19100336]
82. Randall A, Baldi P. SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. BMC Struct Biol. 2008; 8:52. [PubMed: 19055744]
83. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008; 9:40. [PubMed: 18215316]
84. Goyal K, Mande SC. Exploiting 3D structural templates for detection of metalbinding sites in protein structures. Proteins. 2008; 70(4):1206–1218. [PubMed: 17847089]
85. Dudev M, Lim C. Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. BMC Bioinformatics. 2007; 8:106. [PubMed: 17389049]
86. Ferre F, Ausiello G, Zanzoni A, Helmer-Citterich M. Functional annotation by identification of local surface similarities: a novel tool for structural genomics. BMC Bioinformatics. 2005; 6:194. [PubMed: 16076399]
87. Jones, DT.; Hadley, C. Threading methods for protein structure prediction. In: Higgins, D.; Taylor, WR., editors. Bioinformatics: Sequence, structure and databanks. Springer-Verlag; Heidelberg: 2000. p. 1-13.
88. Shu N, Zhou T, Hovmoller S. Prediction of zinc-binding sites in proteins from sequence. Bioinformatics. 2008; 24(6):775–782. [PubMed: 18245129]
89. Andreini C, Banci L, Bertini I, Rosato A. Occurrence of copper proteins through the three domains of life: a bioinformatic approach. J Proteome Res. 2008; 7(1):209–216. [PubMed: 17988086]
90. Andreini C, Banci L, Bertini I, Elmi S, Rosato A. Non-heme iron through the three domains of life. Proteins. 2007; 67(2):317–324. [PubMed: 17286284]
91. Todd AE, Marsden RL, Thornton JM, Orengo CA. Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol. 2005; 348(5):1235–1260. [PubMed: 15854658]
92. Tottey S, Waldron KJ, Firbank SJ, Reale B, Bessant C, Sato K, Cheek TR, Gray J, Banfield MJ, Dennison C, Robinson NJ. Protein-folding location can regulate manganese-binding versus copper- or zinc-binding. Nature. 2008; 455(7216):1138–1142. [PubMed: 18948958]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

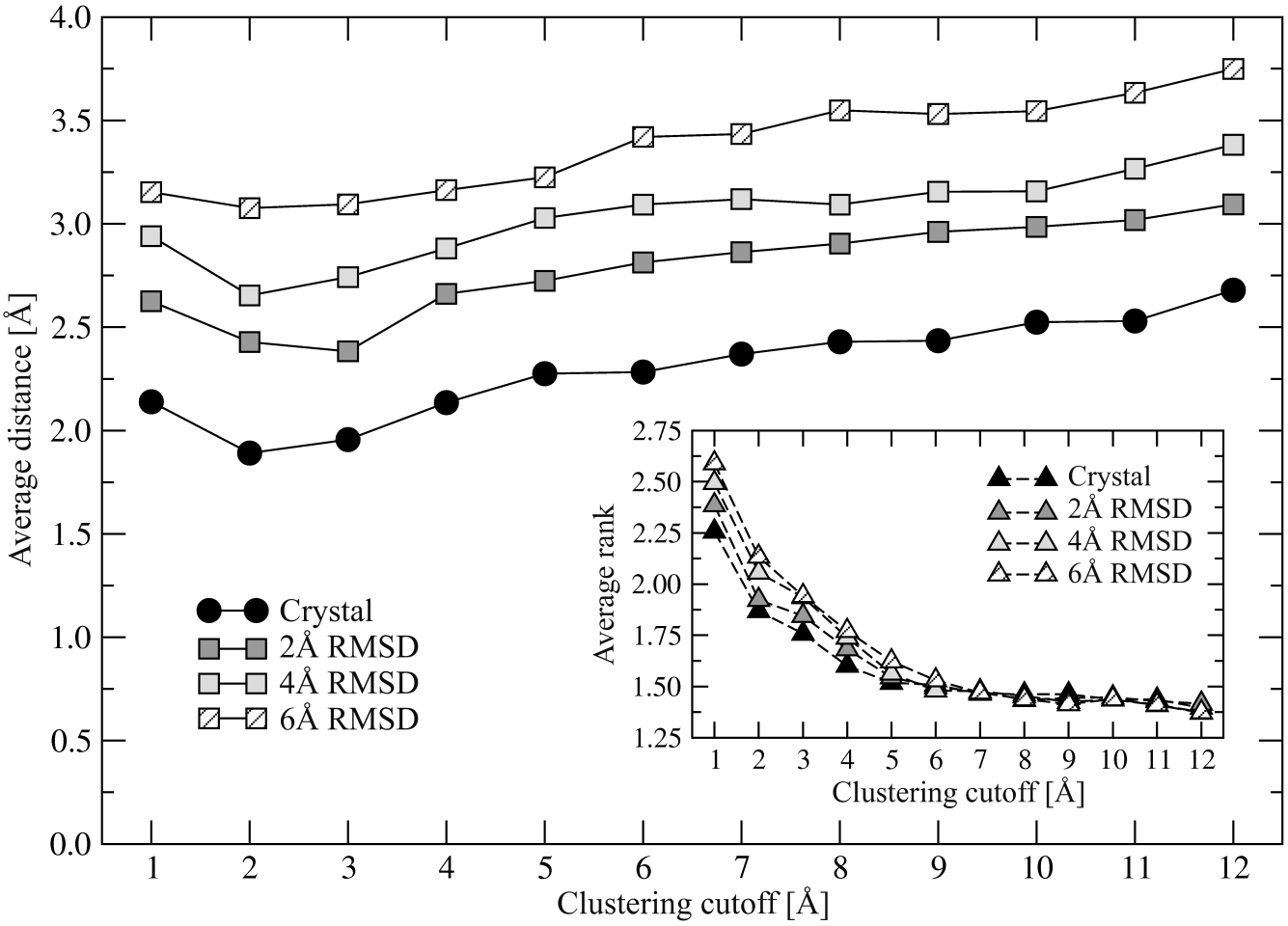

Figure 1.
Average distance (inset: average cluster rank) of the best template-bound metal location from the position of a metal in the crystal structure as a function of the clustering cutoff.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

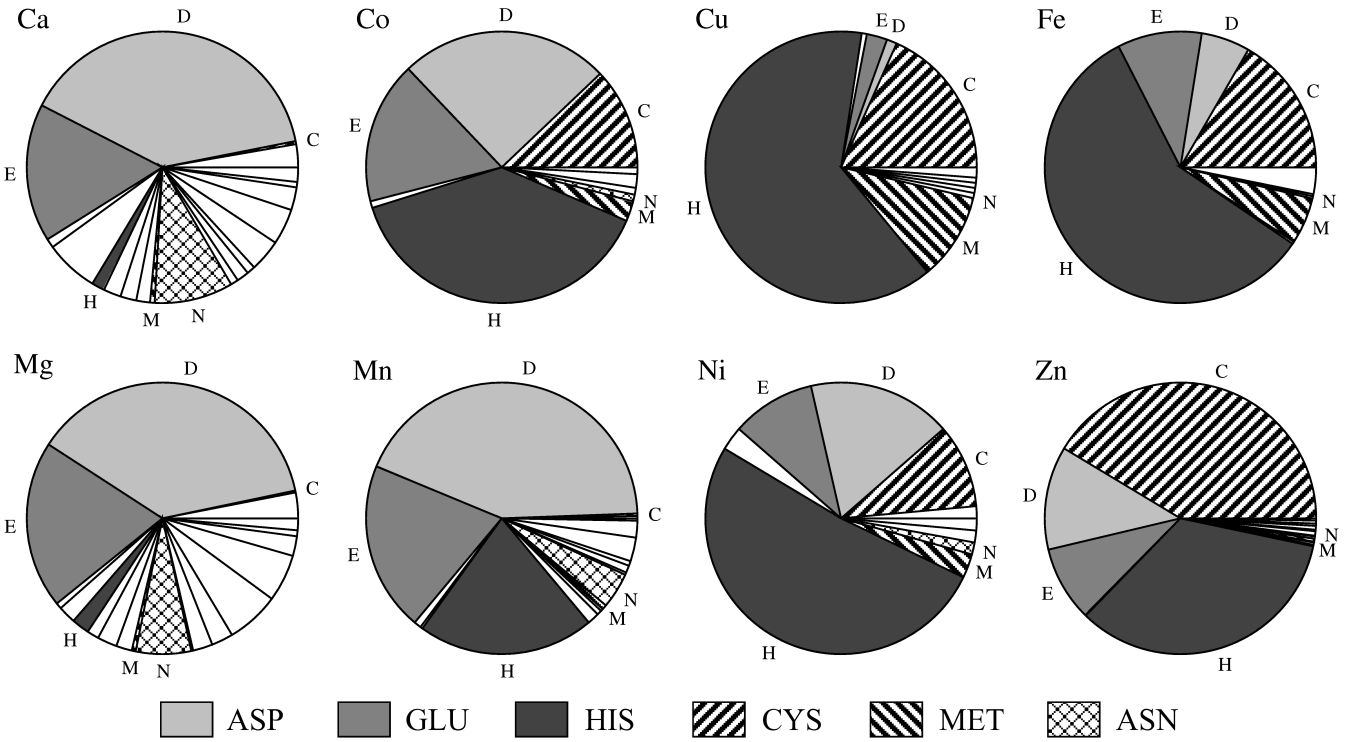

Figure 2.
Amino acid preferences toward binding different metal ions.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

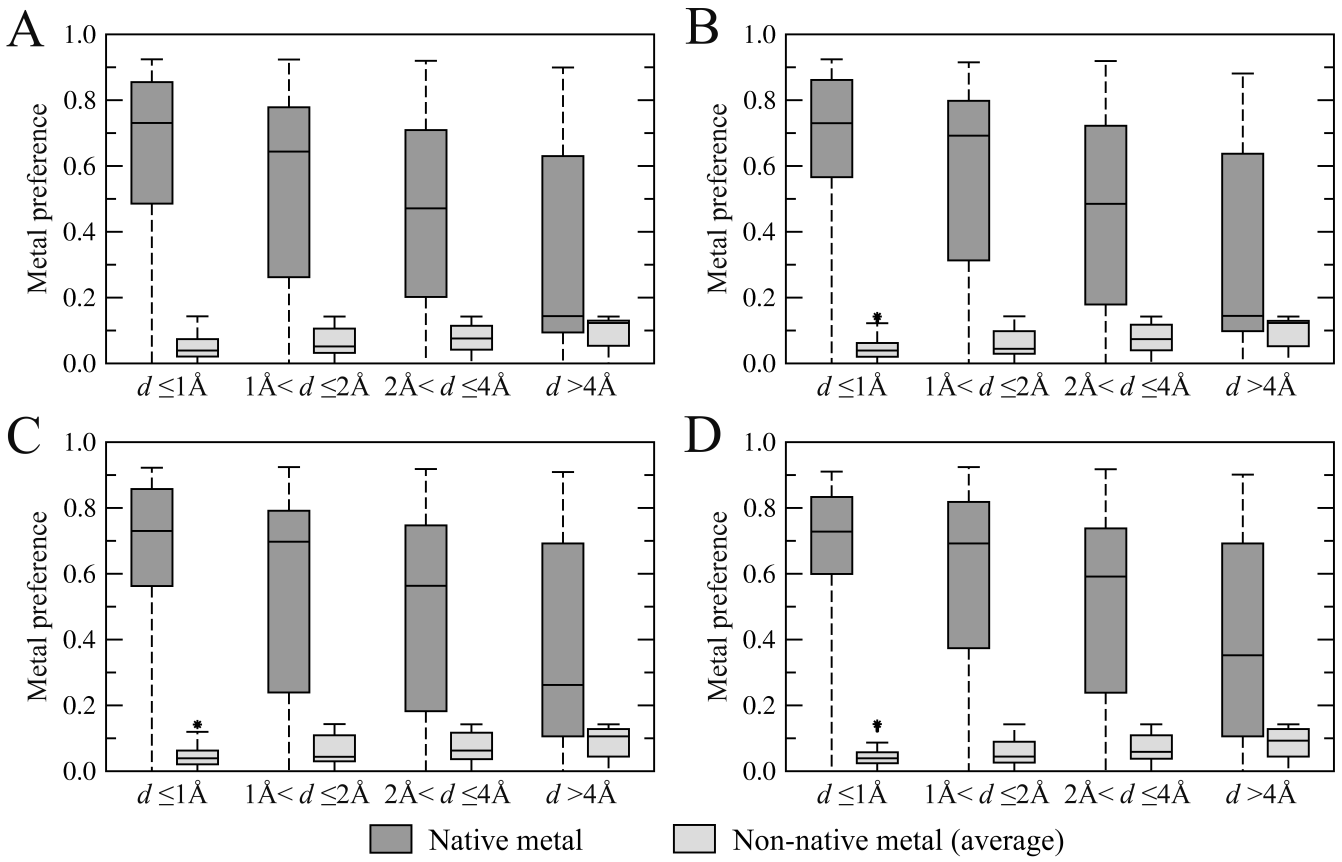

Figure 3.
Binding preferences for native and non-native metal ions for the template sites located within a distance *d* from a metal position in the target structure. Target crystal structures, structures distorted to 2, 4 and 6 Å Cα RMSD are shown in A, B, C and D, respectively.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

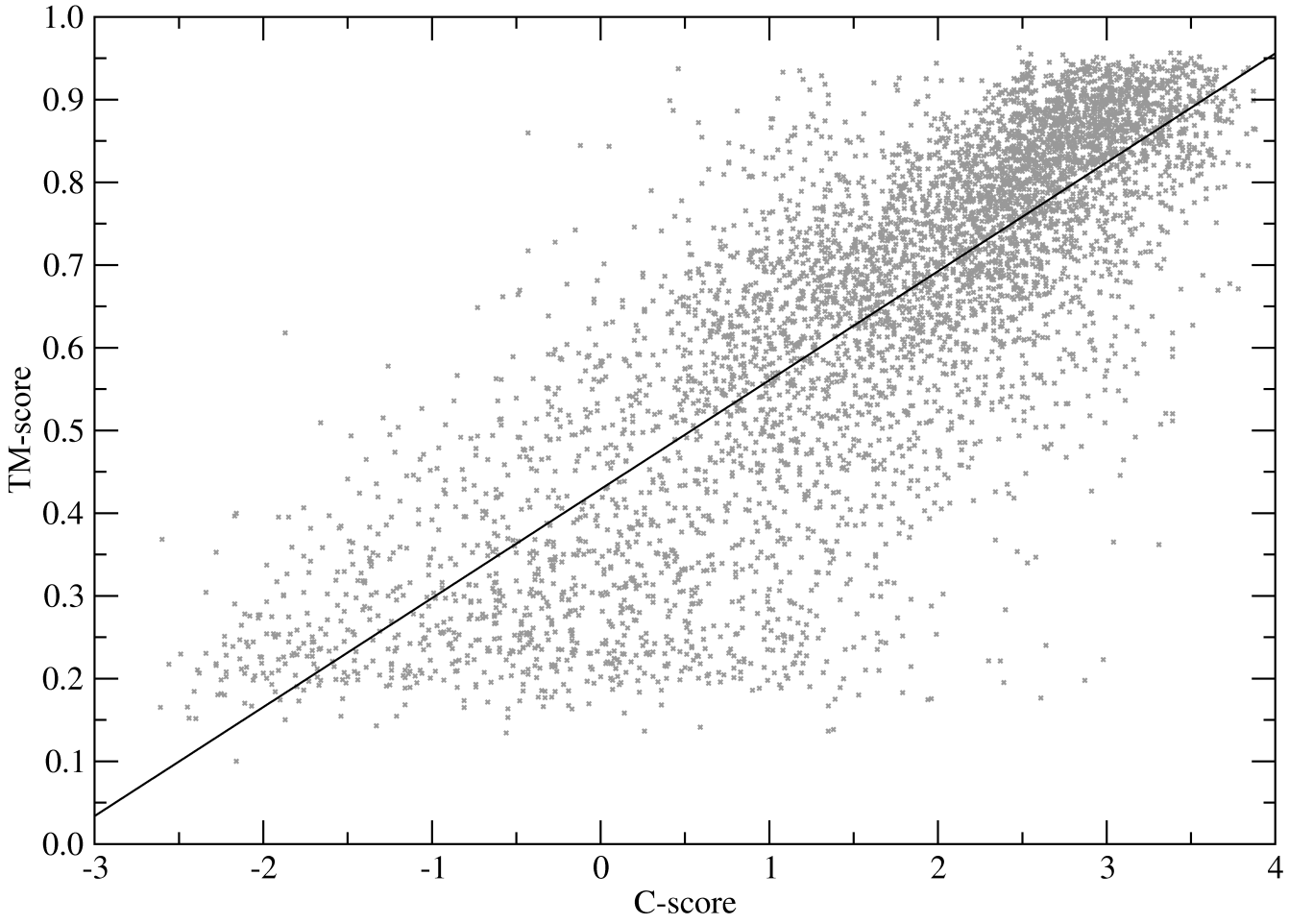

Figure 4.
Correlation between C-score and TM-score to native for a large dataset of protein models constructed by TASSER.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

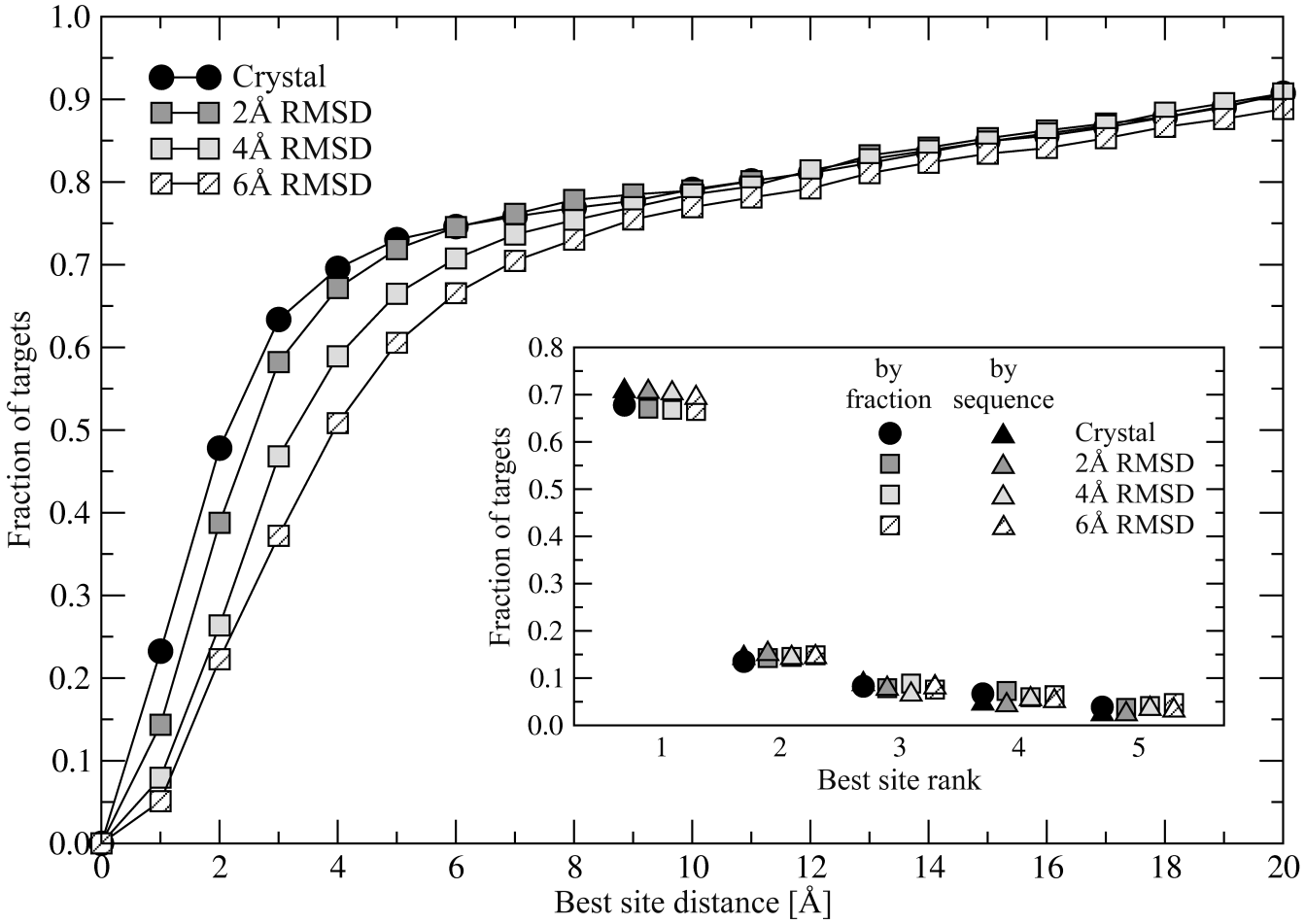

Figure 5.

Accuracy of metal binding site prediction by FINDSITE-metal using crystal structures as well as structures distorted to 2, 4 and 6 Å Cα RMSD. Main plot: the cumulative fraction of proteins with a distance between the metal position in the crystal structure and the closest of the top five predicted binding sites displayed on the *x*-axis, abbreviated as the “Fraction of targets”. Inset: the rank of the predicted site closest to the metal-binding site in the crystal structure using two different ranking procedures: by the fraction of templates that share a common site (“by fraction”) and by a sequence profile score (“by sequence”).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

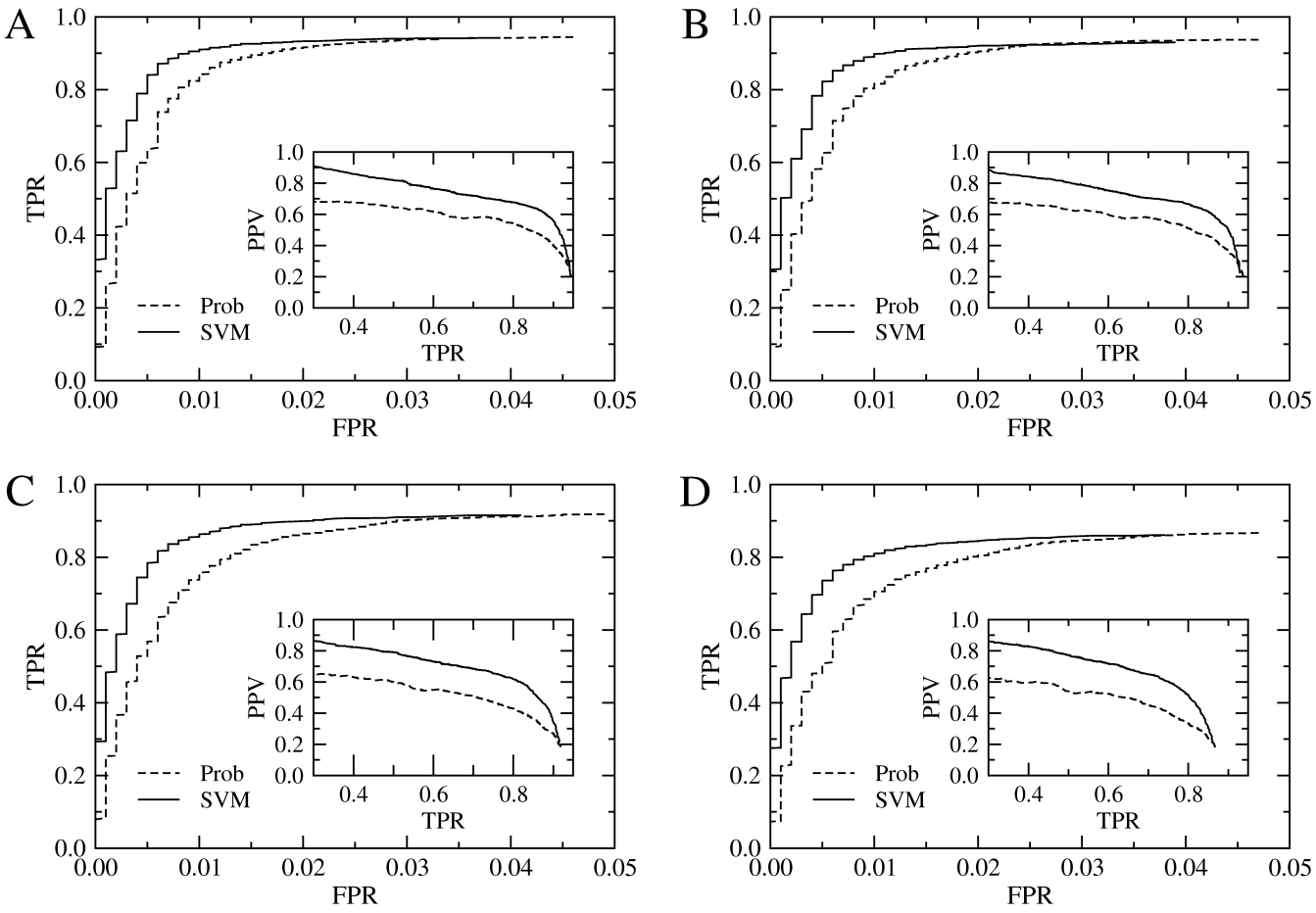

Figure 6.

ROC plots (FPR – false positive rate, TPR – true positive rate) for metal-binding residue prediction by FINDSITE-metal using crystal structures (A) and structures distorted to 2 Å (B), 4 Å (C) and 6 Å (D) Cα RMSD. Inset plots show corresponding Recall-Precision graphs (TPR – recall, PPV – precision). Metal-binding residues are identified based on the probability estimation (Prob) as well as by machine learning (SVM).

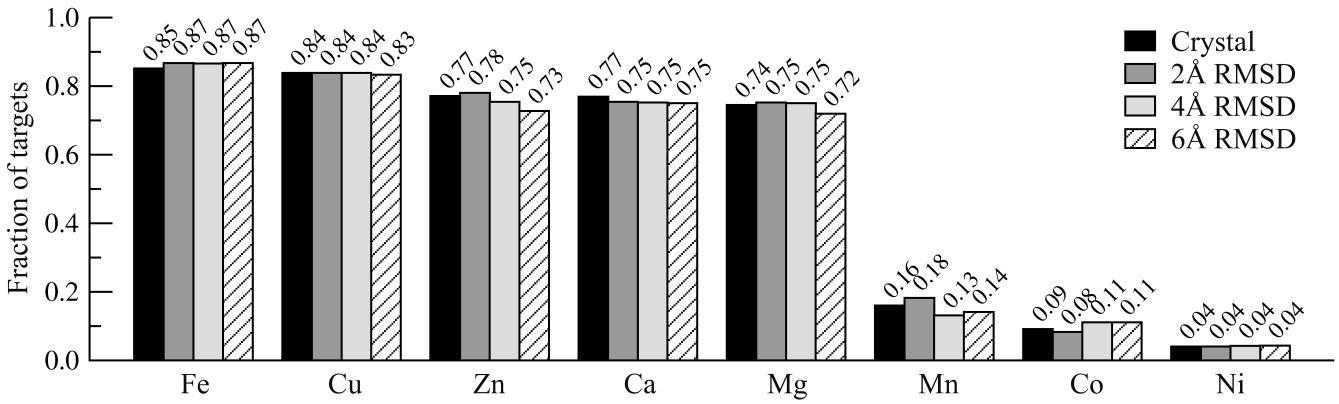

Figure 7.
Accuracy of binding metal prediction in terms of the fraction of targets correctly assigned with the native binding metal.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

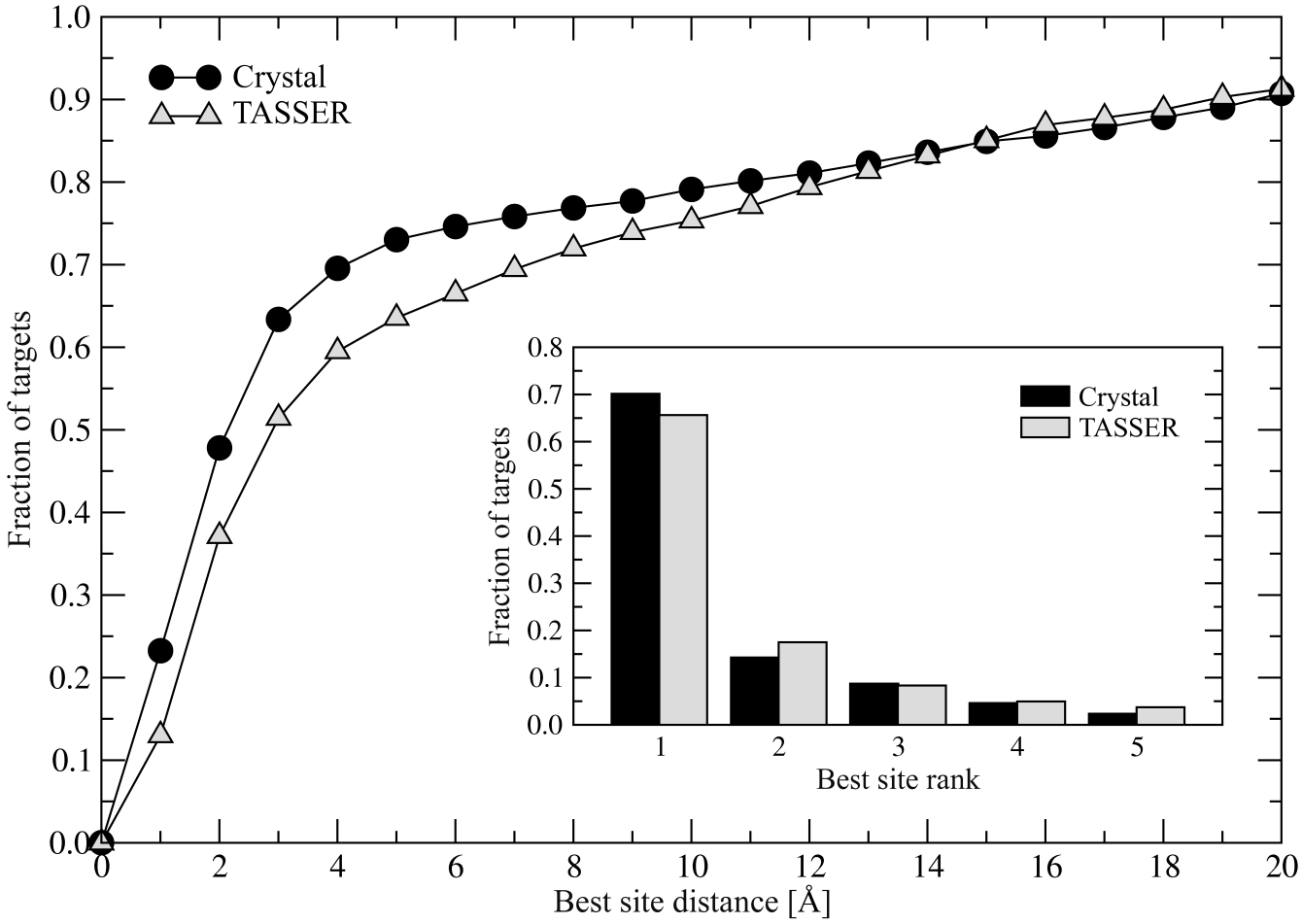

Figure 8.

Comparison of the metal binding site prediction accuracy of FINDSITE-metal for crystal structures and TASSER models. Main plot: the cumulative fraction of proteins with a distance between the metal position in the crystal structure and the closest of the top five predicted binding sites displayed on the *x*-axis, abbreviated as the “Fraction of targets”. Inset: the rank of the predicted site closest to the metal-binding site in the crystal structure using ranking by a sequence profile score.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

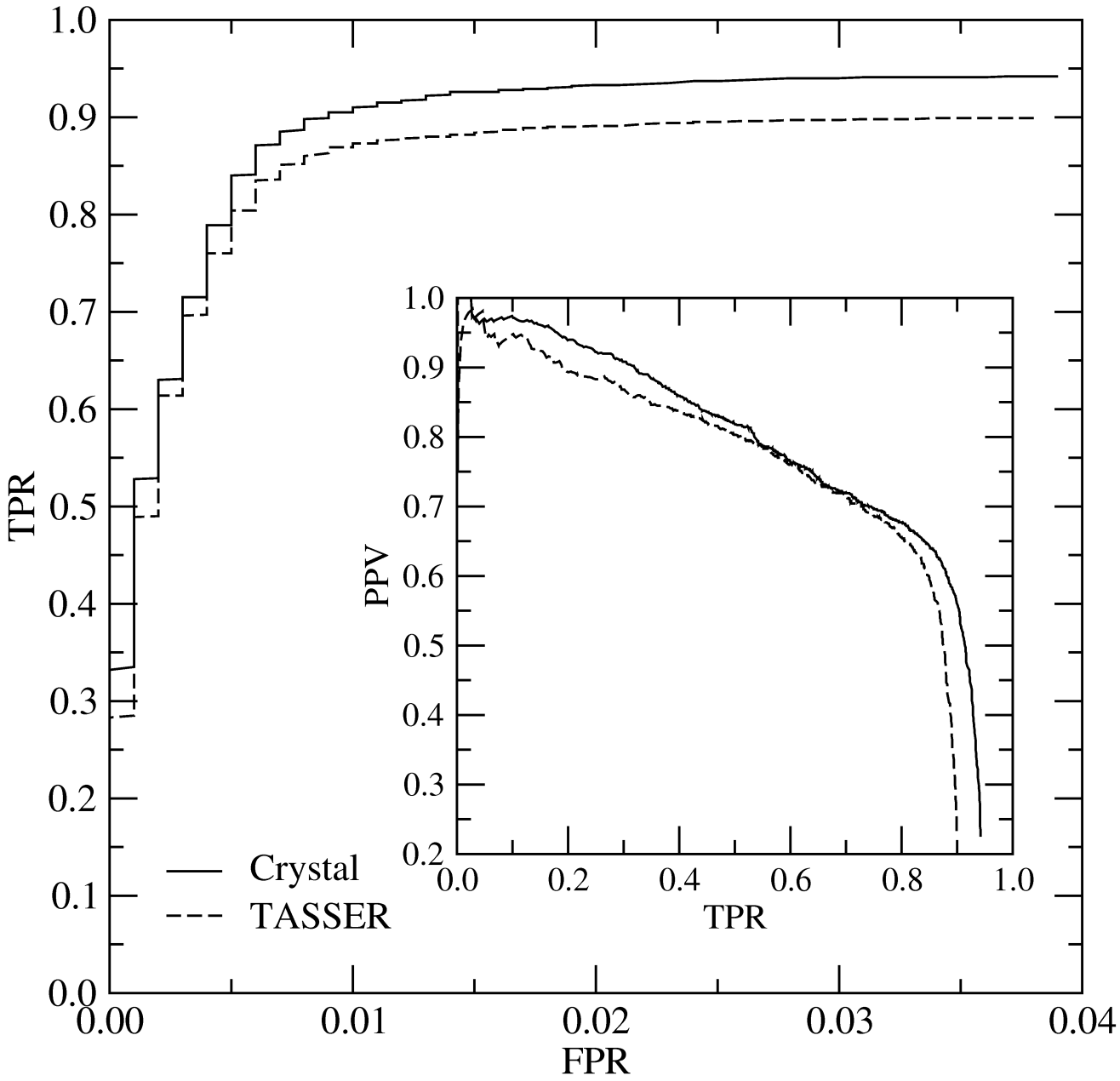

Figure 9.
ROC plot and Recall-Precision graph (inset) for metal binding residue prediction by FINDSITE-metal using crystal structures as well as TASSER models. Binding residues are identified by the SVM.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

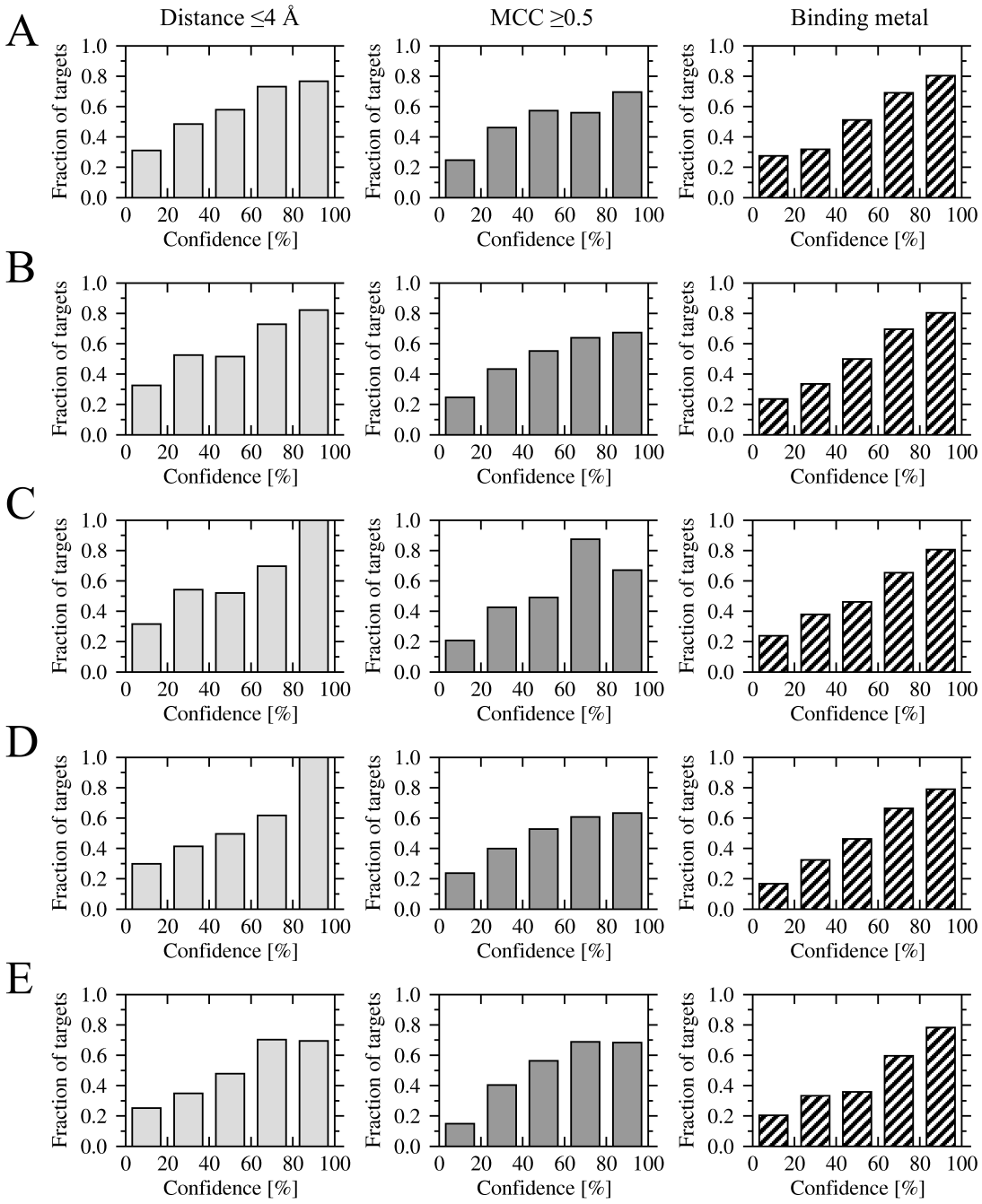

Figure 10.

Performance of FINDSITE-metal in terms of the fraction of targets whose binding site distance is ≤4Å, whose Matthew's correlation coefficient (MCC) for the binding residues is ≥0.4 and whose binding metal is correctly identified as a function of the confidence index. Accuracy is reported for crystal structures (A), structures distorted to 2 Å (B), 4 Å (C) and 6 Å (D) Cα and TASSER models (E) assigned with different confidence values.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

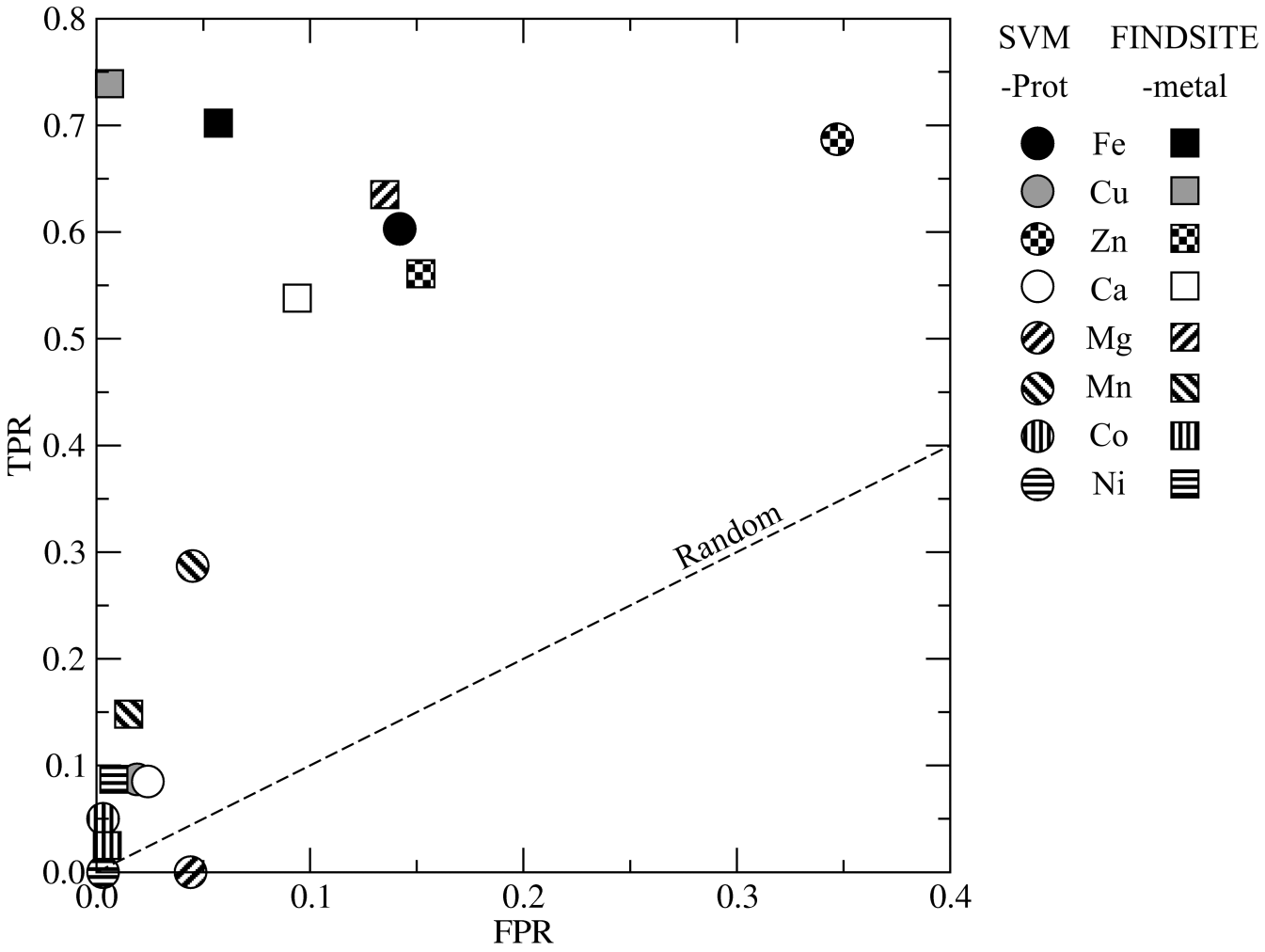

Figure 11.
Comparison of the performance of SVM-Prot and FINDSITE-metal in metal-binding protein prediction. FPR – false positive rate, TPR – true positive rate, dashed line represents a random classifier.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

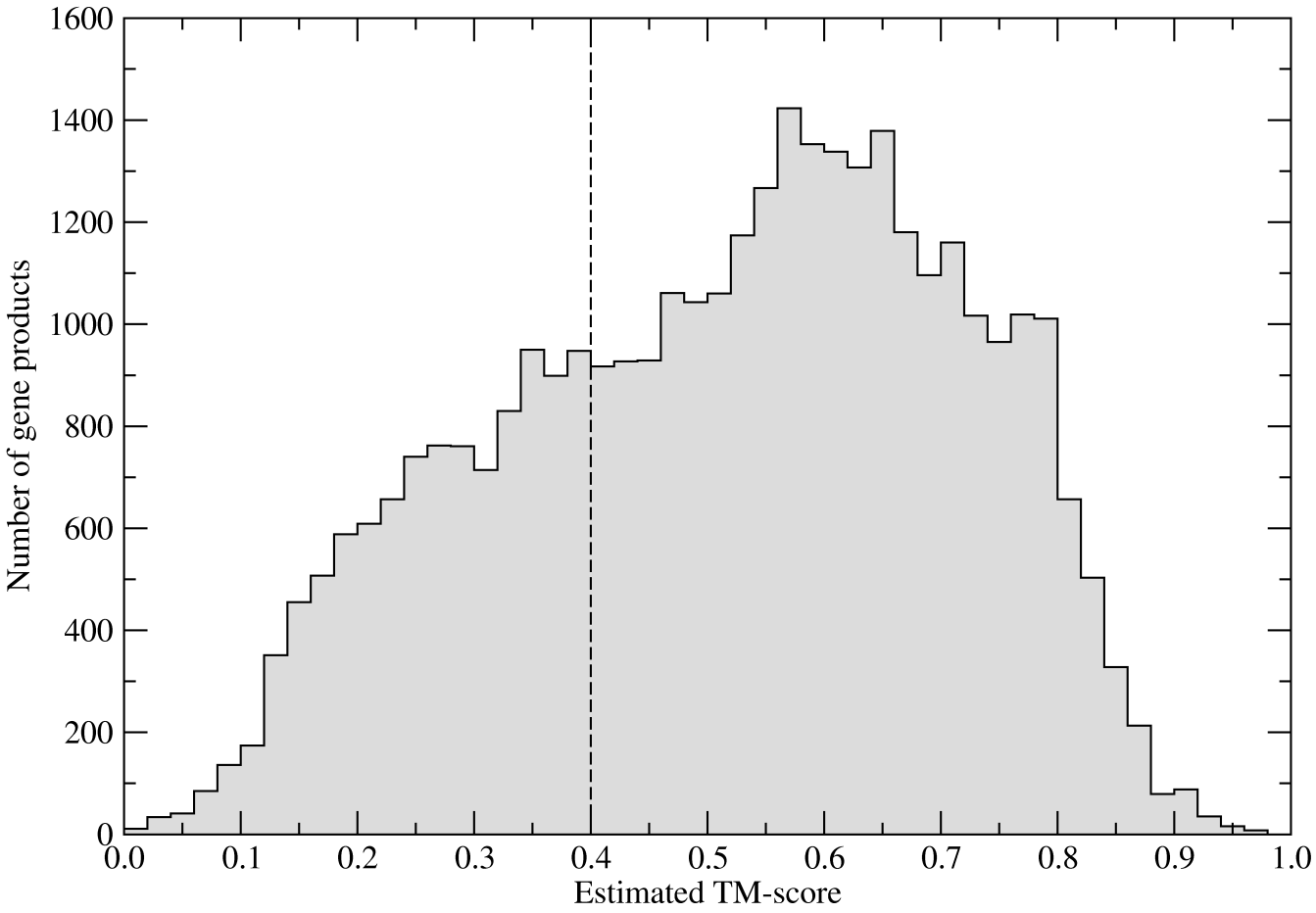

Figure 12.
Histogram of the coverage of the human metalloproteome by TASSER models. TM-score is estimated from the C-score. Dashed line delineates confidently predicted models with an estimated TM-score of ≥0.4.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

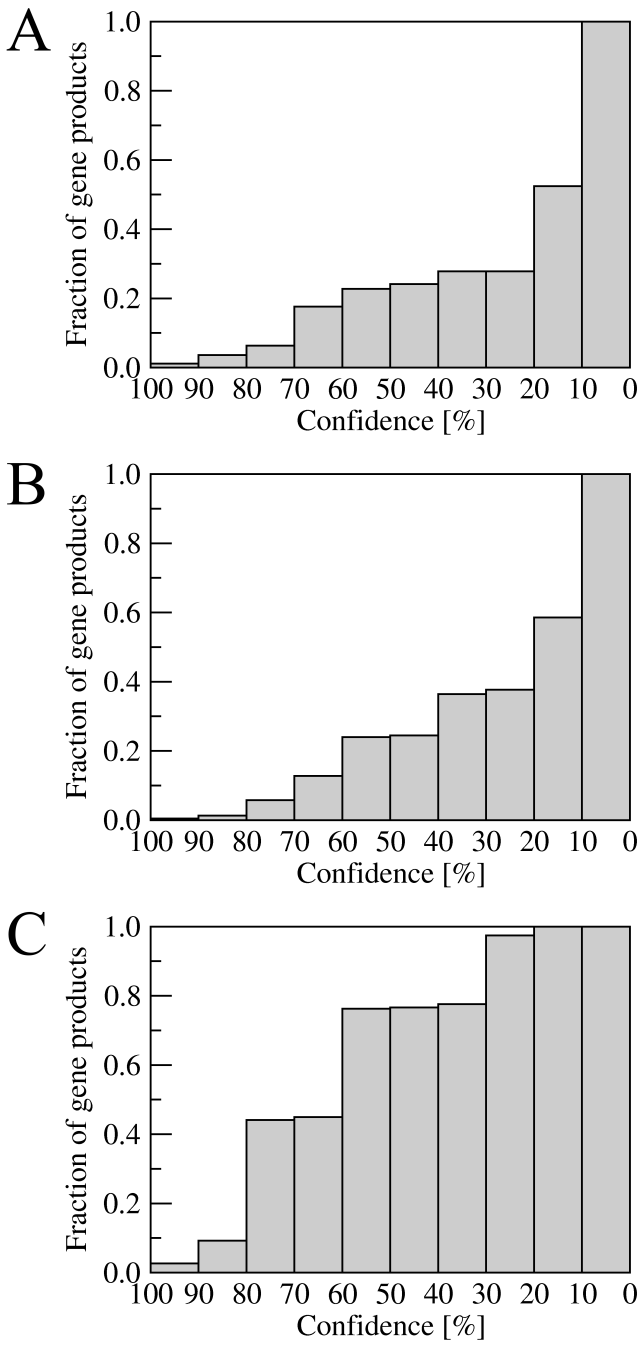

Figure 13.
Cumulative distribution of confidence indexes for metal-binding proteins identified in the human proteome by FINDSITE-metal. Three confidence indexes are reported for (A) site distance ≤4 Å, (B) Matthew's correlation coefficient for binding residues of ≥0.5 and (C) binding metal.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

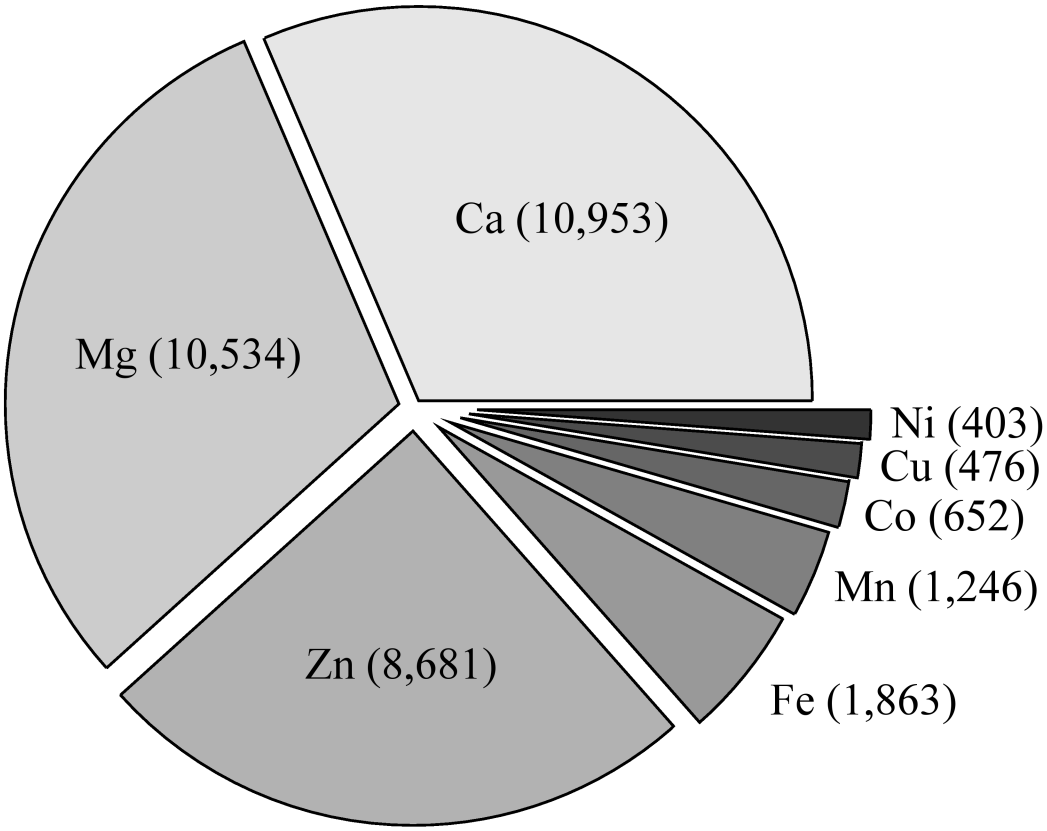

Figure 14.
Metal binding complement of the human proteome. Number of proteins predicted to bind a given metal is given in parenthesis.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table I

Structure quality of the target proteins used in this study

	Crystal	2Å RMSD	4Å RMSD	6Å RMSD	TASSER
Number of proteins	761	756	728	721	747
Number of metal binding sites	1038	1036	1000	986	1034
Target protein Cα RMSD[a]	-	1.99 ±0.06	3.99 ±0.07	5.98 ±0.10	8.93 ±6.58
Target protein TM-score[a]	-	0.89 ±0.05	0.75 ±0.07	0.66 ±0.07	0.68 ±0.19
Target protein Maxsub[a]	-	0.78 ±0.02	0.55 ±0.06	0.47 ±0.07	0.52 ±0.20
Binding site RMSD[a] [Å]	-	0.58 ±0.51	1.05 ±1.04	1.50 ±1.64	2.74 ±2.42
Number of observed binding residues[a]	3.14 ±0.98	3.14 ±0.98	3.14 ±0.98	3.13 ±0.98	3.13 ±0.9
Number of predicted binding residues[a]	3.83 ±1.91	3.76 ±1.90	3.59 ±1.89	3.25 ±1.82	3.69 ±1.93

[a]Average value ± one standard deviation.

Proteins. Author manuscript; available in PMC 2012 March 20.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table II

Set of features used in metal binding residue prediction as well as in the confidence estimation by machine learning

Binding residue prediction by SVM		Confidence index by Bayesian classifier	
Feature #	***Description***	***Feature #***	***Description***
1	Estimated TM-score to native[a]	1	Estimated TM-score to native[a]
2	Distance between the Cα atom and the predicted site center	2	Template fraction[c]
3	Binding probability[b]	3, 4	Average TM-score of templates to the target structure and its standard deviation
4	Template fraction[c]	5	Average deviation of template-bound metals from predicted site center
5, 6	Average TM-score of templates to the target structure and its standard deviation	6	Binding metal entropy[d]
7	Average deviation of template-bound metals from predicted site center	7	Sequence profile score[e]
8	Binding metal entropy[d]	8	Number of metal-bound templates
9	Sequence profile score[e]	9	Number of putative binding residues[h]
10-17	Binding metal preferences[f]	10	Average metal-binding probability for binding residues[h]
18-25	Residue preferences for different metal types[g]	11	Fraction of templates that bind the top-ranked predicted metal

[a] TM-score to native for the target structure estimated from modeling procedure or by model quality assessment.

[b] Calculated from Eq. 1.

[c] Fraction of templates that have a residue in equivalent position in contact with a metal.

[d] Calculated from Eq. 3.

[e] Sequence profile score calculated for alignments constructed by fr-TMalign.

[f] Predicted binding metal preferences, Eq. 2.

[g] Generic residue preferences to bind each type of metal ions.

[h] Predicted by SVM.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table III

Observed and predicted residue and binding metal composition for the dataset of protein targets with different quality structures.

	Crystal			2Å RMSD			4Å RMSD			6Å RMSD		
Residue composition[b] [%]												
Asp	27.1	*26.6*	**−0.5**	27.1	*26.9*	**−0.2**	27.2	*30.0*	**2.8**	27.6	*32.4*	**4.8**
Glu	14.4	*15.4*	**1.0**	14.4	*16.0*	**1.6**	14.4	*17.9*	**3.5**	14.6	*20.1*	**5.5**
His	23.7	*22.8*	**−0.9**	23.5	*24.1*	**0.6**	24.0	*25.0*	**1.0**	24.1	*27.2*	**3.1**
Cys	12.7	*9.1*	**−3.6**	13.2	*9.5*	**−3.7**	12.2	*8.6*	**−3.6**	11.2	*8.6*	**−2.6**
Met	1.6	*1.1*	**−0.5**	1.5	*1.2*	**−0.3**	1.6	*1.1*	**−0.5**	1.6	*1.0*	**−0.6**
Asn	4.4	*3.5*	**−0.9**	4.4	*3.3*	**−1.1**	4.4	*3.2*	**−1.2**	4.5	*2.6*	**−1.9**
Other	16.1	*21.5*	**5.4**	15.9	*19.0*	**3.1**	16.2	*14.2*	**−2.0**	16.4	*8.1*	**−8.3**
Metal composition[b] [%]												
Ca	25.9	*23.9*	**−2.0**	25.8	*23.4*	**−2.4**	25.6	*23.9*	**−1.7**	26.1	*24.3*	**−1.8**
Co	3.0	*0.7*	**−2.3**	3.0	*0.6*	**−2.4**	2.8	*0.6*	**−2.2**	2.9	*0.4*	**−2.5**
Cu	4.3	*4.3*	**0.0**	4.3	*4.3*	**0.0**	4.4	*4.3*	**−0.1**	4.5	*4.4*	**−0.1**
Fe	12.7	*14.8*	**2.1**	12.8	*15.0*	**2.2**	13.2	*15.0*	**1.8**	13.2	*15.7*	**2.5**
Mg	15.0	*22.9*	**7.9**	15.0	*23.0*	**8.0**	15.3	*23.9*	**8.6**	15.2	*23.4*	**8.2**
Mn	10.3	*4.0*	**−6.3**	10.5	*3.9*	**−6.6**	10.6	*3.9*	**−6.7**	10.8	*3.8*	**−7.0**
Ni	2.0	*0.4*	**−1.6**	1.8	*0.1*	**−1.7**	1.9	*0.2*	**−1.7**	1.6	*0.2*	**−1.4**
Zn	26.8	*29.0*	**2.2**	26.8	*29.7*	**2.9**	26.2	*28.2*	**2.0**	25.7	*27.8*	**2.1**

[b] Observed, predicted values and the differences are given in regular font, italics and bold, respectively.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table IV

Comparison of the performance of SVM-Prot and FINDSITE-metal in metal-binding protein prediction

Binding metal	SVM-Prot			FINDSITE-metal		
	ACC	*SPC*	*PPV*	*ACC*	*SPC*	*PPV*
Fe	0.826	0.858	0.381	0.913	0.943	0.642
Cu	0.946	0.981	0.154	0.984	0.994	0.829
Zn	0.663	0.653	0.443	0.766	0.848	0.598
Ca	0.748	0.976	0.553	0.812	0.906	0.664
Mg	0.820	0.956	0.000	0.832	0.865	0.439
Mn	0.891	0.955	0.407	0.904	0.985	0.515
Co	0.966	0.997	0.400	0.962	0.995	0.143
Ni	0.977	0.997	0.000	0.975	0.992	0.182

ACC – accuracy, *SPC* – specificity, *PPV* – precision.