BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 9 2011, pages 1262–1268
doi:10.1093/bioinformatics/btr133

Structural bioinformatics

Advance Access publication March 16, 2011

Structure-based *de novo* prediction of zinc-binding sites in proteins of unknown function

Wei Zhao[1,2], Meng Xu[1,2], Zhi Liang[1,2], Bo Ding[1,2], Liwen Niu[1,2], Haiyan Liu[1,2,*] and Maikun Teng[1,2,*]

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China and [2]Key Laboratory of Structural Biology, Chinese Academy of Sciences, 96 Jinzhai Road, Hefei, Anhui 230027, China

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Zinc-binding proteins are the most abundant metalloproteins in Protein Data Bank (PDB). Accurate prediction of zinc-binding sites in proteins of unknown function may provide important clues for the inference of protein function. As zinc binding is often associated with characteristic 3D arrangements of zinc ligand residues, its prediction may benefit from using not only the sequence information but also the structure information of proteins.

Results: In this work, we present a structure-based method, TEMSP (3D TEmplate-based Metal Site Prediction), to predict zinc-binding sites. TEMSP significantly improves over previously reported best methods in predicting as many as possible true ligand residues for zinc with minimum overpredictions: if only those results in which all zinc ligand residues have been correctly predicted are defined as true positives, our method improves sensitivity from less than 30% to above 60%, and selectivity from around 25% to 80%. These results are for predictions based on apo state structures. In addition, the method can predict the zinc-bound local structures reliably, generating predictions useful for function inference. We applied TEMSP to 1888 protein structures of the 'Unknown Function' class in the PDB database. A number of zinc-binding sites have been discovered *de novo*, i.e. based solely on the protein structures. Using the predicted local structures of these sites, possible functional roles were analyzed.

Availability: TEMSP is freely available from http://netalign.ustc.edu.cn/temsp/.

Contact: hyliu@ustc.edu.cn; mkteng@ustc.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 18, 2010; revised on January 25, 2011; accepted on January 26, 2011

1 INTRODUCTION

About one-third of all proteins in nature require metal ions for their normal structures and functions (Holm *et al.*, 1996; Matthews *et al.*, 2008). In the protein structure database PDB (Bernstein *et al.*, 1977), zinc is the most abundant metal ion (Babor *et al.*, 2008). It has been suggested that zinc-binding proteins may comprise up to 10% of the entire human proteome (Andreini *et al.*, 2006). In zinc-binding proteins, the zinc ions or their binding motifs often play important functional roles. For example, about 40% of the zinc-binding proteins in the human proteome are transcription factors, in which the zinc-containing motifs are involved in DNA binding (Andreini *et al.*, 2006). The remaining 60% are primarily enzymes, in many of them zinc plays important catalytic roles (Andreini *et al.*, 2006). Thus, the prediction of zinc-binding sites in proteins is of general interest for the inference of protein function.

Metal binding sites, especially zinc sites, can be predicted based on the sequences of target proteins. Many reasonably successful methods are available for this purpose (Andreini *et al.*, 2004; Lin *et al.*, 2005, 2006; Lippi *et al.*, 2008; Passerini *et al.*, 2006; Shu *et al.*, 2008). In some methods, comparative modeling of the structures of the target proteins are attempted to verify or improve the predictions (Levy *et al.*, 2009; Passerini *et al.*, 2007). Recently, a method using the sequence profiles of known zinc-binding domains in PDB assisted by the so-called 'metal-binding patterns' has been proposed and tested (Andreini *et al.*, 2009). Because it needs only the sequence information of the target proteins, the method can be applied to annotate full zinc proteomes from genome sequences (Bertini *et al.*, 2010). However, as metal binding is very sensitive to the local arrangements of ligand residues in the 3D space, if the 3D structure of a target protein is available, it can be expected that much higher accuracy can be achieved by making predictions directly based on the structure .

Several methods were recently proposed to predict metal-binding sites, especially zinc sites, directly based on the 3D structures. Among them, the method MetSite uses neural networks trained on sequence profiles, secondary structure states, the solvent accessible surface areas (SASAs) as well as the inter-residue distance matrices (Sodhi *et al.*, 2004). The empirical Fold-X force field has been developed to predict binding sites and binding affinities for water and for metal ions (Schymkowitz *et al.*, 2005). Goyal and Mande reported a method based on statistically derived geometrical constraints on zinc-binding sites (Goyal and Mande, 2008). The method CHED makes use of the properties of triads of zinc-coordinating ligand residues to make predictions based on apoprotein structures (Babor *et al.*, 2008). The method FEATURE uses a Bayesian classifier based on analyses of the properties of amino acid residues in concentric shells around zinc ions (Ebert and Altman, 2008). The method SitePredict uses Random Forest classifiers trained on diverse ligand residue-based site properties to

[*]To whom correspondence should be addressed.

predict binding sites for metal ions or small molecules (Bordner, 2008).

For some of these methods, apparently excellent performance, including high sensitivity and selectivity, have been reported (Babor *et al.*, 2008; Bordner, 2008; Ebert and Altman, 2008). By definition, sensitivity is the ratio of the number of true positive (TP) predictions over the total number of positive cases (TP+FN, FN being the number of false negative predictions), while selectivity is the ratio of TP over the total number of positive predictions (TP+FP, FP being the number of false positive predictions). Obviously, these measures depend on the exact definition of what consists of TP predictions.

In reported studies, TPs were often defined using relatively loose criteria (Babor *et al.*, 2008; Bordner, 2008; Ebert and Altman, 2008) such as containing as few as only one correctly predicted zinc-coordinating residue (Babor *et al.*, 2008). With such type of TP definitions, an inaccurate prediction that mistakenly predicts all but one of the several native zinc-coordinating residues would be counted a TP, the same as an accurate prediction that correctly predicts all the zinc-coordinating residues. Thus, results that are suboptimal at least for the purpose of function inference may produce apparently good performance. For example, although a zinc ion is normally chelated with only three or four ligand residues, the number of predicted ligand residues per zinc ion by some methods reported with good performance measures is often larger than five. In addition, many of the predicted zinc-binding sites by these methods consist of mostly incorrect ligand residues.

We believe that for the development of zinc-binding prediction as a tool for function inference, it is necessary to use more stringent criteria to judge the results. First, the inference of a protein's function based on its zinc-binding site does need as many as possible correctly predicted zinc-chelating residues. In addition, it also needs as few as possible falsely predicted ligand residues because they are potentially misleading. If a new method shows clear improvements over previous ones by a criterion taking into account of these factors, it should also improve function inference.

Another limitation of currently available methods is that they do not produce structural models of the zinc-bound states. Upon zinc binding, the protein backbone conformation may mostly stay unchanged or only change slightly (Babor *et al.*, 2008). However, in more than 40% of cases, the side chain conformation of ligand residue has been found to alter significantly (Babor *et al.*, 2008). Thus, given a prediction comprising a cluster of likely ligand residues in their apo state conformations, it is difficult to conceive how the zinc ion would coordinate with its ligand residues, especially when many of the predicted ligand residues are incorrect.

None of the current methods has been designed to provide a structural model (i.e. atomic coordinates of the zinc ion and its coordinating protein side chains) for the zinc-coordinated state. One arguable exception is the Fold-X force field method (Schymkowitz *et al.*, 2005). However, the method uses local optimizations starting from an input structure, and it is not clear to what extent the method can still make correct predictions if the local conformation in the holo state significantly differs from the input conformation (for example, the input may consist of an apo state in which some or all of the ligand residues are in different side chain conformations as compared with the holo state).

In this report, we developed a new structure-based approach to predict zinc-binding sites. In our approach, a binding site must consist of a certain number of appropriately arranged ligand residues. To detect such residues, we compare the relative positions of C_α and C_β atoms of potential ligand residues in a target protein with those in known zinc-binding sites. For a group of potential ligand residues predicted to form a zinc-binding site by this approach, atomic positions of the bound zinc and the coordinating side chains are also predicted. By considering the positions of the C_α and C_β atoms, changes of side chain conformations of ligand residues from the apo to the holo state are allowed, but of backbone conformations are not. The method has been trained using a selected set of zinc-binding proteins to give optimum performance by stringent TP definitions. It was then tested on an independent dataset. Finally, we applied the method to predict zinc-binding sites and their local structures in proteins classified as of 'Unknown Function' in PDB. The *de novo* predicted sites were analyzed and possible functional roles were assigned based on the predicted local structures.

2 METHODS

Figure 1 shows an overview of the workflow. The method is briefly summarized below and details are given in the method section of Supplementary Materials. In summary of the method, known zinc-binding sites have been extracted from PDB structures and decomposed into pairs of zinc-chelating residues. Each pair forms a template and all such pairs constituted a pre-compiled template library. During the prediction, this library is searched for templates that 'match' the configuration of a pair of candidate ligand residues in the target protein. The candidate ligand residue pairs that can be matched by templates are again combined in pairs, forming the so-called 'pairs-of-pairs'. Each possible pair-of-pairs is filtered by using a range of geometric criteria. A pair-of-pairs passing all the filters comprises a predicted zinc-binding site. For a predicted site, atomic coordinates of the zinc ion and its ligand residues in the zinc-bound state are proposed based on the local configurations of the matching templates. The method has been optimized on a training set of target proteins (Supplementary Table S1). Details of parameters and their optimization are given in Supplementary Materials. It has been tested in an unbiased way using another set of target proteins (the independent test set) that are independent from the training set as well as from the template library (Supplementary Table S1).

In both training and testing, the following 'intersection over union ratio' (IoUR) was used to quantify the accuracy of results.

$$\mathrm{IoUR}=\frac{N\left(\text{predicted ligand residues}\cap\text{actual ligand residues}\right)}{N\left(\text{predicted ligand residues}\cup\text{actual ligand residues}\right)} \qquad (1)$$

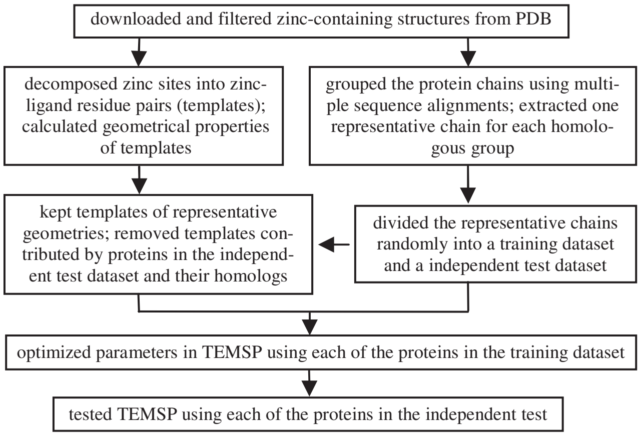
downloaded and filtered zinc-containing structures from PDB
decomposed zinc sites into zinc-ligand residue pairs (templates); calculated geometrical properties of templates
grouped the protein chains using multiple sequence alignments; extracted one representative chain for each homologous group
kept templates of representative geometries; removed templates contributed by proteins in the independent test dataset and their homologs
divided the representative chains randomly into a training dataset and a independent test dataset
optimized parameters in TEMSP using each of the proteins in the training dataset
tested TEMSP using each of the proteins in the independent test

Fig. 1. A schematic representation of the workflow.

This ratio balances between the numbers of correctly and wrongly predicted ligand residues for a particular binding site. If unspecified, we have considered prediction results with IoUR ≥ 0.5 to be TPs. In additions, performance measures under the most stringent criterion of IoUR $= 1$ are also reported.

3 RESULTS AND DISCUSSION

3.1 Performance on the training dataset

Key geometric parameters used in TEMSP have been optimized to achieve the best performance for the 338 target proteins in the training set. Among the various geometric parameters, we found that the performance of the methods, i.e. sensitivity and selectivity, is most responsive to the distance cutoff between the two zinc positions, each predicted for one of the candidate ligand residue pairs in the pair-of-pairs (see also 'S1.6 Training and testing' section of Supplementary Material). The prediction results on the training set by the final model containing the optimized parameters are summarized in Table 1.

By the chosen TP definition, TEMSP attains a sensitivity of 89.4% for the training sets, i.e. 422 out of the 472 zinc-binding sites in the training set have been predicted correctly. The average IoUR value is 0.95 for the TPs (Table 1), which is close to 1.0 and indicates that most of the TP predictions consist of correct predictions of all ligand residues. In addition, the average distance between the predicted and the actual positions of the zinc ions ('Zinc-deviation' in Table 1) is only 0.39 Å. If TP is defined with the most stringent criterion IoUR $= 1$, namely, all actual ligand residues must be predicted with no overpredictions, the number of TP cases is 352, corresponding to a sensitivity of 74.6%. For these more stringent TPs, the average Zinc-deviation decreases to 0.32 Å.

For the training set, TEMSP predicted 10 sites which were not labeled as zinc binding based on the original PDB data or were of IoUR < 0.5 (FP predictions, Supplementary Table S2), resulting in an apparent selectivity of 97.7%. These FP predictions were individually inspected. Three of them (3C37_A, 1ADT_A and 1XRT_A, the notation standing for PDB ID plus chain ID in the PDB entry; See also Supplementary Table S2) correspond to actual zinc-binding sites and each of the respective predicted sites contains two correct zinc-coordinating residues. For another FP case (1T8H_A), although the PDB entry 1T8H contains no zinc ion coordinated with the predicted ligand residues (H80, C125 and H142), we could find two other PDB entries that are highly similar both in sequence and in structure to 1T8H (1U05 and 1XAF) and each contains a zinc ion at the corresponding site. Thus, it is highly possible that this prediction is actually correct. For yet another FP case (3CG7_A), a manganese ion is found to bind at the predicted site (it consisted of D15, E17 and D184) in an isoform PDB structure (3CM5_A). Based on that some of the different transition metal ions, including zinc and manganese, can often bind to the same site and sometimes can be replaced by each other, this prediction may also be actually correct. For the other five FP results, zinc-containing isoforms or homologous proteins were not found. Whether these predictions are correct or not cannot be unambiguously judged based on available evidence. In any case, the actual selectivity of TEMSP for the training dataset should be higher than the results in Table 1.

3.2 Performance on the independent test dataset

Target proteins in the independent test set have not been considered in parameter optimization. They are also independent from the template library. Table 1 shows that TEMSP has more or less similar overall performance on the independent test set as on the training set. With an IoUR cutoff of 0.5 for TP, it identified 117 of all 136 actual zinc-binding sites, giving a sensitivity of 86.0% (Table 1). Average IoUR for the TP cases is 0.96. TEMSP produced five apparent FP predictions (Supplementary Table S2), resulting in a selectivity of 95.9%. Among the FP cases, one prediction, 1ZFD_A, contains two actual zinc ligand residues (Supplementary Table S2). For another FP prediction, 3ISZ_A, a zinc ion was found to bind at the predicted site (D100, E135 and H349) in an isoform structure (3IC1_A, Supplementary Table S2). Only for the remaining three FP predictions (in 3I9F_A and 2JKS_A, Supplementary Table S2), we could not find similar evidence to support the predictions. If they turn out to be truly false predictions, the overprediction rate of TEMSP for the test set is 2.2% (3 FP over 136 TP + FN). This rate is, in fact, likely to be close to the error rates resulting from metal ion lost during protein preparation or from data misinterpretation during structure determination.

Depending on the purpose of the prediction, the geometric parameters used in TEMSP can be adapted to aim at either higher sensitivity or better selectivity. Figure 2 shows the receiver operating characteristic (ROC) curve, i.e. the plot of sensitivity versus (1 − selectivity). Strictly speaking, the ROC curve should be the plot of sensitivity versus (1 − specificity), where specificity is defined as the total number of FP over the total number of actual negative cases. However, it is difficult to define the total number of

Table 1. Statistics of prediction results of TEMSP on the training dataset and of TEMSP and CHED on the independent test dataset and the CHED test set

	Method	Number of TP	Number of FN	Number of FP	Sensitivity (%)	Selectivity (%)	Average IoUR	Average Zinc-deviation[a] (Å)
In training set	TEMSP	422 (352)[b]	50 (120)	10 (82)	89.4 (74.6)	97.7 (81.1)	0.953 (1)	0.39 (0.32)
In independent test set	TEMSP	117 (100)	19 (36)	5 (22)	86.0 (73.5)	95.9 (82.0)	0.956 (1)	0.38 (0.32)
	CHED	112 (65)	24 (71)	11 (58)	82.4 (47.8)	84.2 (52.8)	0.878 (1)	–
In CHED test set[c]	TEMSP	19 (16)	6 (9)	1 (4)	76.0 (64.0)	95.0 (80.0)	0.961 (1)	–
	CHED	18 (5)	7 (20)	2 (15)	72.0 (20.0)	90.0 (25.0)	0.734 (1)	–

[a] The average distance deviation between the predicted and the actual zinc ion.

[b] Values in parentheses were obtained with a tightened TP definition of IoUR $= 1$, i.e. the predicted sites contained only and all actual zinc ligand residues.

[c] CHED test set given in Supplementary Table S1 were derived from the dataset 2 of CHED method [dataset 2, Supplementary Material of reference (Babor *et al.*, 2008)]. Three apo structures were excluded for reasons described in the text.

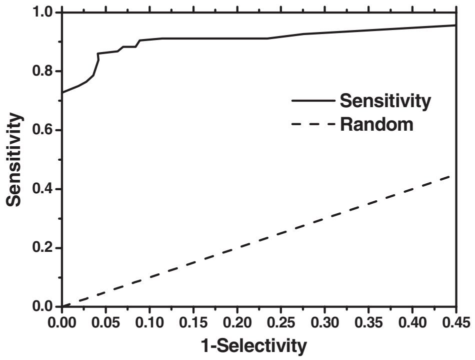

Fig. 2. The plot of sensitivity versus 1 − selectivity of TEMSP prediction results for the independent test dataset. The dashed line represents random predictions.

actual negative cases contained in our datasets except that it must be much larger than the number of positive cases or the number of actual zinc-binding sites. Thus, selectivity has been used in the place of specificity in our discussions. For TEMSP applied to the independent test dataset, the ROC curve is significantly above the line associated with random predictions, suggesting good performance of the model. The area under the ROC curve (AUC) attains a value of 0.945, whereas a value of 1.0 represents perfect predictions.

3.3 Comparisons with other methods

Among previously reported methods, several have been reported (Goyal and Mande, 2008; Schymkowitz *et al.*, 2005) or have been shown (Babor *et al.*, 2008) to give much lower sensitivity and/or selectivity as compared with TEMSP. The CHED algorithm developed by Babor *et al.* (2008) have been reported with comparable performance, i.e. with a sensitivity of 79% at 100% selectivity using a test set consisting of apo structures. However, their results corresponded to a relaxed definition of TP predictions: a predicted site containing as few as only one correct zinc-coordinating residue was considered as a TP. On the other hand, the independent test dataset used here consists of holo instead of apo structures. It is thus interesting to compare TEMSP with CHED based on the same TP criteria, using both the independent test dataset of holo structures here and the test set of apo structures used in reference (Babor *et al.*, 2008) (noted as CHED test set).

The results of the comparisons are also summarized in Table 1. When CHED was applied to the independent test set, 112 TP predictions were made with the TP criterion IoUR ≥ 0.5, corresponding to a sensitivity of 82.4%. It also generated 11 FP predictions, which corresponded to a selectivity of 84.2%. Both results are only slightly worse than TEMSP. However, the average IoUR of the TP predictions by CHED is only 0.88, suggesting that it performed definitely worse than TEMSP in correctly predicting all the true ligand residues without over predictions. More importantly, if the criterion for TP is tightened to IoUR $= 1$, the sensitivity and selectivity of CHED drop significantly to 47.8 and 52.8%, respectively. These are much lower than the corresponding values of 73.5 and 82.0%, respectively, for TEMSP.

TEMSP was then applied to the CHED test set. With the TP criterion set at IoUR ≥ 0.5, it achieved a sensitivity of 76% and a selectivity of 95% (Table 1). These results are also only slightly higher than the CHED values of 72 and 90%, respectively, calculated with the same TP criterion. However, the average IoUR of TPs by TEMSP is 0.96, again much higher than the value of 0.73 by CHED. If the TP criterion is tightened to IoUR $= 1$, TEMSP still retains a sensitivity of 64% and a selectivity of 80%, while the corresponding results for CHED drop to 20 and 25%, respectively.

Thus, irrespective of whether apo or holo structures are used for prediction, TEMSP gives slightly better results than CHED if only the presence or absence of zinc-binding sites is predicted, and significantly better results than CHED if the exact zinc-coordinating residues are predicted. The latter improvement of TEMSP over CHED is especially significant for the CHED test set that consists of apo structures: for 16 out of the 25 zinc-binding sites in total, TEMSP exactly predicted all the actual zinc-coordinating residues, while CHED only made such exact predictions for five sites.

Table 1 also indicates that the performances of both TEMSP and CHED declined using apo structures instead of holo structures as input. This is expected, and the declination of TEMSP results may suggest that the assumption of unaltered backbone conformations upon zinc binding does not always hold. On the other hand, the declination associated with TEMSP is much less significant as compared with CHED. It suggests that the flexible side chain model in TEMSP can account for many of the conformational differences between the apo and the holo states. One illustrative example is the successful prediction by TEMSP of the zinc-binding site in the apo structure, 1RDZ_A, contained in the CHED test set. CHED failed to predict this site as zinc binding. Comparisons with the corresponding holo structure (1FRP_A) indicate that zinc binding at this site involves somewhat significant changes in the side chain conformations of the ligand residues (Supplementary Fig. S1). TEMSP correctly predicted the three ligand residues from the apo structure. In addition, Supplementary Figure S1 shows that the predicted side chain conformations are much closer to the true holo state structure than the apo state input structure.

It is also interesting to note that the FN predictions produced by TEMSP and CHED do not completely overlap. For example, for the CHED test set, out of the seven binding sites missed by CHED, three binding sites (in proteins 1ET9_A, 1RDZ_A and 1C3P_A, respectively) were correctly predicted by TEMSP, with all zinc-coordinating residues correctly assigned. On the other hand, out of the six FN binding sites of TEMSP, two binding sites (in proteins 1EMV_B and 1E65_A, respectively) were predicted by CHED with IoUR ≥ 0.5. Given that both methods are highly selective, in real applications, it may be possible to apply both methods to the same target to improve the sensitivity further.

Another method, FEATURE, has also been reported with apparently good performances (a selectivity of 73.6% and a sensitivity of 75.5% on a dataset of apo structures) (Ebert and Altman, 2008). However, FEATURE predicts explicit positions of the metal ion instead of the ligand residues, and TP was defined as a predicted metal position within 5 Å from any of the actual ligand residues. Thus, it is difficult to make solid comparisons between FEATURE and TEMSP or CHED. However, Table 1 shows that the sensitivity and selectivity given by TEMSP using either holo or apo structures as input are all higher than that reported for FEATURE, even though the definition of TP here seems to be much more stringent than in the reference (Ebert and Altman, 2008).

Still another method, SitePredict, have reported an excellent AUC value of 0.964 (Bordner, 2008). However, it is not explicitly stated how TP prediction results have been defined in calculating the AUC. In reference (Bordner, 2008), a typical prediction by SitePredict was illustrated [Figure 1 in the reference of Bordner (2008)]. At least for this prediction, several non-ligand residues have been mistakenly predicted as metal binding and some true ligand residues have been predicted only with low certainty. Such predictions would have a low IoUR.

TEMSP predicts not only the zinc-coordinating residues but also the binding position of the zinc ion as well as the side chain conformations of the zinc-coordinating residues. Given the ligand residues, the Fold-X force field can also predict the position of the zinc ion (Schymkowitz *et al.*, 2005). However, the reported average deviation of predicted zinc positions based on holo structures by Fold-X, 0.43 Å, is larger than the average deviation of 0.38 Å by TEMSP.

Another advantageous feature of TEMSP is that predictions can be made with only the coordinates of main chain atoms (the needed positions of C_β atoms can be easily derived from them). As side chain conformation is not required, the method can be applied to target proteins with only low-resolution structures, which often contain dubious side chain conformations but somewhat well-defined main chains.

3.4 Predict zinc binding in proteins of ‘Unknown Function’

TEMSP was applied to 1888 protein structures (Supplementary Table S1) from PDB classified as having ‘Unknown Function’ and determined by X-ray diffraction with resolutions higher than 2.5 Å. It predicted 186 zinc-binding sites in 145 protein structures. Among the predicted sites, 87 sites can be confirmed by the actual presence of zinc at the predicted sites in the structures themselves (82 sites from 65 structures, Supplementary Table S4a) or in respective isoform structures (5 sites from 4 structures, Supplementary Table S4b). For other 36 sites, metal ions other than zinc, including manganese, iron, cobalt, nickel, copper or cadmium ions were found in the 31 structure entries (Supplementary Table S4c). Sites binding these transition metals are often similar to zinc-binding site and should also have the potential to bind zinc ions.

For the remaining predicted sites, three (from three structures 3E61_A, 3IPF_A and 1MZG_A, respectively) were located at the artificially fused His-tag tails (Supplementary Table S4d). They are probably true zinc-binding sites but are irrelevant for protein function. Another four sites (from four structures 1XV2_A, 3F3B_A, 2D7V_A and 2O95_A, respectively) may be actually wrong predictions (for example, with less than four predicted ligand atoms while the structure lacked space for any additional small molecule ligand such as water) after close inspections (Supplementary Table S4e). The remaining 54 predicted sites are distributed in 45 structures. They are listed in Supplementary Table S5.

The local configurations predicted by TEMSP for each of these 54 sites were closely inspected. In most cases, the predicted ligands were found to form reasonable zinc-coordinating structures with no additional likely ligand except water. In only a few cases, a likely ligand residue was not predicted. These are all indicated in Supplementary Table S5. Then the predicted zinc sites were empirically judged whether it could be potential ‘catalytic sites’ or ‘structural sites’, based on geometry criterion (see details in Supplementary ‘S1 METHODS’).

In what follows, three examples from Supplementary Table S5 are discussed in details to illustrate functional inference based on predicted zinc-binding sites. To our knowledge, metal-binding-related functions of these three target proteins have not been suggested before, and the inferences here may be subjected to future experimental tests.

2QQY_A: this target protein from *Bacillus anthracis str. Ames* folds into a ferritin-like four-helix bundle (Fig. 3a). TEMSP predicts a binuclear zinc site with four ligand residues per zinc ion in this target. Based on the criterion described in Supplementary ‘S1 METHODS’, the potential bizinc site is inferred as a structural site rather than a catalytic site. Similar binuclear sites binding zinc or iron ions can be found in some ferritin structures. In Figure 3b, the predicted site in 2QQY_A is compared with the natural bizinc site in mycobacterial bacterioferritin (3BKN_A). The two sites contain the same types of ligand residues (EEEH for each zinc ion) in similar side chain conformations. And after the ligand residues are superimposed, the positions of the two zinc ions in the predicted site are similar to the experimentally determined zinc positions in 3BKN_A. In some other ferritins, iron instead of zinc have been found to occupy the corresponding sites, and iron has been speculated to be the native substrate for the bizinc site of 3BKN_A (Janowski *et al.*, 2008). Thus, we may safely infer that the biological function of this target protein may involve the binding to zinc or to other transition metal ions.

2DT4_A: in this example, a metal-containing catalytic center may be inferred. The protein, *Ph*PPC, is a Plants and Prokaryotes Conserved (PPC) protein from *Pyrococcus horikoshii*. Sequence search using *Ph*PPC against the non-redundant GeneBank database detected a large protein family constituting a conserved protein domain. However, all members of the family are so far of unknown functions except that several members have been hypothesized to be related to DNA binding. The crystal structure of *Ph*PPC had been solved by Lin *et al.* (2007; PDB ID 2DT4) in the hope to obtain more insights into its function. However, based on the metal-free structure solved with 1.6 Å resolution, few conclusions regarding possible functions could be drawn (Lin *et al.*, 2007). The structure of *Ph*PPC is shown in Figure 3c. TEMSP predicted that three histidine residues, H89, H91 and H105, form a zinc-binding site (Supplementary Table S5). Structure database search detected six 2DT4-like structures with Dali Z-scores > 16 and RMSD < 1.9 Å, including 3HWU, 2H6L, 3HTN, 2HX0, 2NMU and 2P6Y. Their sequence identities with *Ph*PPC range from 25% to 44%. They have all been labeled either as having unknown function or as putative DNA-binding proteins. In all these six structures, the histidine residues forming the predicted zinc-binding site in 2DT4 are conserved, and can also be predicted to form zinc sites by TEMSP. Besides this, the structure 2H6L and 2P6Y actually contain a zinc ion at this position, and in 3HTN, the site is occupied by iron or nickel. 2DT4_A and 2H6L_A superimposed (Fig. 3e); the distance between the predicted zinc ion in 2DT4_A and the native zinc ion in 2H6L_A is only 0.93 Å. In addition, all these seven proteins form homotrimers as the biological unit. In the trimer form, the predicted zinc site for 2DT4 would locate in a pocket formed at the interface of two monomers (Fig. 3d). A glutamic acid residue (E71) from the other monomer is located at a distance of 4.73 Å from the predicted zinc ion (Fig. 3d). Similar to the

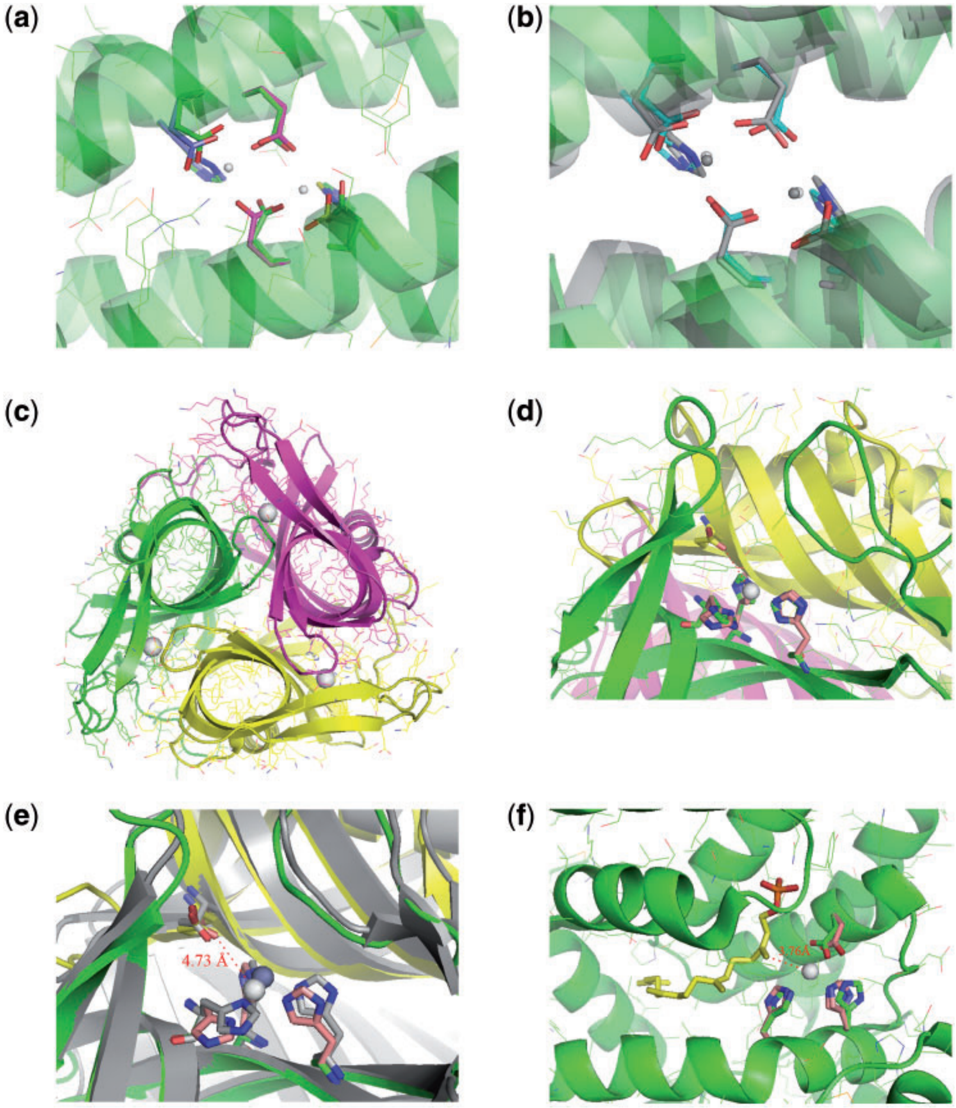

Fig. 3. Examples for the prediction of zinc binding sites in proteins of unknown function from PDB database. Oxygen atoms are shown in red and nitrogen atoms are in blue by default, while carbon atoms are shown in different colors as described below. (**a** and **b**) The predicted binuclear site in 2QQY_A. (a) The native structure is shown in green. Side chains of six zinc ligand residues are shown as thicker sticks. Three templates that matched different pairs of these ligand residues are shown in slate, magenta and yellow, respectively. Two predicted zinc ions are shown as white spheres. (b) A superposition of 2QQY_A (green) and mycobacterial bacterioferritin 3BKN_A (gray), showing ligand residues at the binuclear center in thicker sticks. For 2QQY_A, the predicted side chain conformations are shown in cyan and two predicted zinc ions are shown as white spheres. For 3BKN_A, the native zinc ions are shown in gray. (**c**) The biological trimer unit of 2DT4 with three predicted zinc ions. The three composing monomers are shown in green, yellow and magenta, respectively. The predicted zinc ions are shown as white spheres. (**d**) An enlarged view of the predicted zinc sites in 2DT4_A. The original side chains are in green or yellow, and the predicted ones in salmon. (**e**) A superposition of a predicted zinc site in 2DT4_A and a native site in 2H6L_A. The native structure from 2H6L_A is shown in gray. The predicted side chains and a zinc ion are shown in salmon and white, respectively. In (d) and (e), thicker sticks represent ligand residues. (**f**) The predicted zinc-binding site and the GRO-binding pocket 3KB4_A. The zinc ligand residues and GRO (geranyl monophosphate, yellow) are shown as thicker sticks. The input structure is in green and the predicted conformations of ligand residues and a zinc ion are shown in salmon and while, respectively. TEMSP outputs PDB files which contain the atom coordinates of the predicted zinc-coordinating residues and of the corresponding templates, respectively. Images of protein structures were generated using the program PyMOL (DeLano, 2002) using the atomic coordinates of zinc and zinc ligands predicted by TEMSP.

three histidine residues, this glutamic acid residue is also conserved in all the seven structures, despite their low or moderate overall sequence identities. Thus, the three histidine residues coordinating zinc together with E71 may form a conserved functional motif. Empirically, such a motif typically appears in the catalytic center of a metallohydrolase, such as matrix metalloproteinases and other members of the metalloendopeptidase superfamily (Gomis-Ruth, 2003), in which a water coordinated to the histidine-coordinated zinc assisted by a nearby carboxyl from a glutamate or aspartate side chain acts as a nucleophile. Therefore, our results strongly suggest that the *Ph*PPC homotrimer unit may have metallohydrolase function with the predicted zinc site comprising a catalytic center. This suggestion may be generalized to other members of this PPC family that contain this conserved site. We note that although supporting evidence has been discussed here, this novel inference can be made based on applying TEMSP to the (apo state) structure of 2DT4 alone, demonstrating the usefulness of TEMSP in functional inference.

3KB4_A: the high selectivity or low FP rate of TEMSP suggests that even without further supporting evidence, functional inference made based solely on the predictions may still be substantially credible. 3KB4_A is such an example, which corresponds to the Alr8543 protein from *Nostoc sp. PCC 7120*. No information about its function could be retrieved using conventional sequence, structure or literature searches. A potential zinc-binding site consisting of H120, H124 and E136 was predicted by TEMSP (Supplementary Table S5). The predicted configuration for zinc binding suggested that a water molecule together with the predicted ligands may form a reasonable tetrahedron (Fig. 3f). Empirically, this site is more likely a catalytic center as in different metalloenzymes than a purely structural site. Besides this, the zinc site is located at the bottom of a pocket in which a geranyl monophosphate (GRO) molecule was found bound to the protein. The predicted zinc position is 3.76 Å from the bound GRO. In addition, the orientation and geometry of the GRO relative to the predicted zinc ion is analogous to that of substrates bound in the catalytic centers of metalloenzymes (Fig. 3f). These strongly suggest that Alr8543 may be a metalloenzyme with the predicted zinc site as a catalytic center located at the bottom of a substrate-binding pocket.

4 CONCLUSIONS

We have developed and tested TEMSP as a method that significantly improves over existing method in predicting zinc binding from protein structures. The method is *de novo*, namely, it does not require any information other than the main chain structure of the target protein itself. Thus, it can greatly improve functional inference based on zinc-binding site predictions, as demonstrated by the examples. From a methodological point of view, the strategy adopted by TEMSP may be extended to the analysis and prediction of binding sites for other types of metal ions or small molecules. As this method does not depend on side chain conformations, it may also be applied in protein design to identify where to introduce new zinc-binding sites in a target protein.

ACKNOWLEDGEMENTS

We thank Dr Xiaohua Lou and Dr Jian Zhan for helpful discussions.

Funding: National Natural Science Foundation of China (grant 30970560, 10979039); Chinese Academy of Sciences (grant KSCX2-YW-R-127, INFO-115-D01-2009); China Postdoctoral Science Foundation (Grant 20080430768).

Conflict of Interest: none declared.

REFERENCES

Andreini,C. *et al.* (2004) A hint to search for metalloproteins in gene banks. *Bioinformatics*, **20**, 1373–1380.

Andreini,C. *et al.* (2006) Counting the zinc-proteins encoded in the human genome. *J. Proteome Res.*, **5**, 196–201.

Andreini,C. *et al.* (2009) Metalloproteomes: a bioinformatic approach. *Acc. Chem. Res.*, **42**, 1471–1479.

Babor,M. *et al.* (2008) Prediction of transition metal-binding sites from apo protein structures. *Proteins*, **70**, 208–217.

Bernstein,F.C. *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Bertini,I. *et al.* (2010) The annotation of full zinc proteomes. *J. Biol. Inorg. Chem.*, **15**, 1071–1078.

Bordner,A.J. (2008) Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics*, **24**, 2865–2871.

DeLano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.

Ebert,J.C. and Altman,R.B. (2008) Robust recognition of zinc binding sites in proteins. *Protein Sci.*, **17**, 54–65.

Gomis-Ruth,F.X. (2003) Structural aspects of the metzincin clan of metalloendopeptidases. *Mol. Biotechnol.*, **24**, 157–202.

Goyal,K. and Mande,S.C. (2008) Exploiting 3D structural templates for detection of metal-binding sites in protein structures. *Proteins*, **70**, 1206–1218.

Holm,R.H. *et al.* (1996) Structural and functional aspects of metal sites in biology. *Chem. Rev.*, **96**, 2239–2314.

Janowski,R. *et al.* (2008) Bacterioferritin from Mycobacterium smegmatis contains zinc in its di-nuclear site. *Protein Sci.*, **17**, 1138–1150.

Levy,R. *et al.* (2009) Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins*, **76**, 365–374.

Lin,C.T. *et al.* (2005) Protein metal binding residue prediction based on neural networks. *Int. J. Neural Syst.*, **15**, 71–84.

Lin,H.H. *et al.* (2006) Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. *BMC Bioinformatics*, **7** (Suppl. 5), S13.

Lin,L. *et al.* (2007) Crystal structure of Pyrococcus horikoshii PPC protein at 1.60 A resolution. *Proteins*, **67**, 505–507.

Lippi,M. *et al.* (2008) MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics*, **24**, 2094–2095.

Matthews,J.M. *et al.* (2008) Designed metal-binding sites in biomolecular and bioinorganic interactions. *Curr. Opin. Struct. Biol.*, **18**, 484–490.

Passerini,A. *et al.* (2006) Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins*, **65**, 305–316.

Passerini,A. *et al.* (2007) Predicting zinc binding at the proteome level. *BMC Bioinformatics*, **8**, 39–51.

Schymkowitz,J.W.H. *et al.* (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl Acad. Sci. USA*, **102**, 10147–10152.

Shu,N. *et al.* (2008) Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics*, **24**, 775–782.

Sodhi,J.S. *et al.* (2004) Predicting metal-binding site residues in low-resolution structural models. *J. Mol. Biol.*, **342**, 307–320.